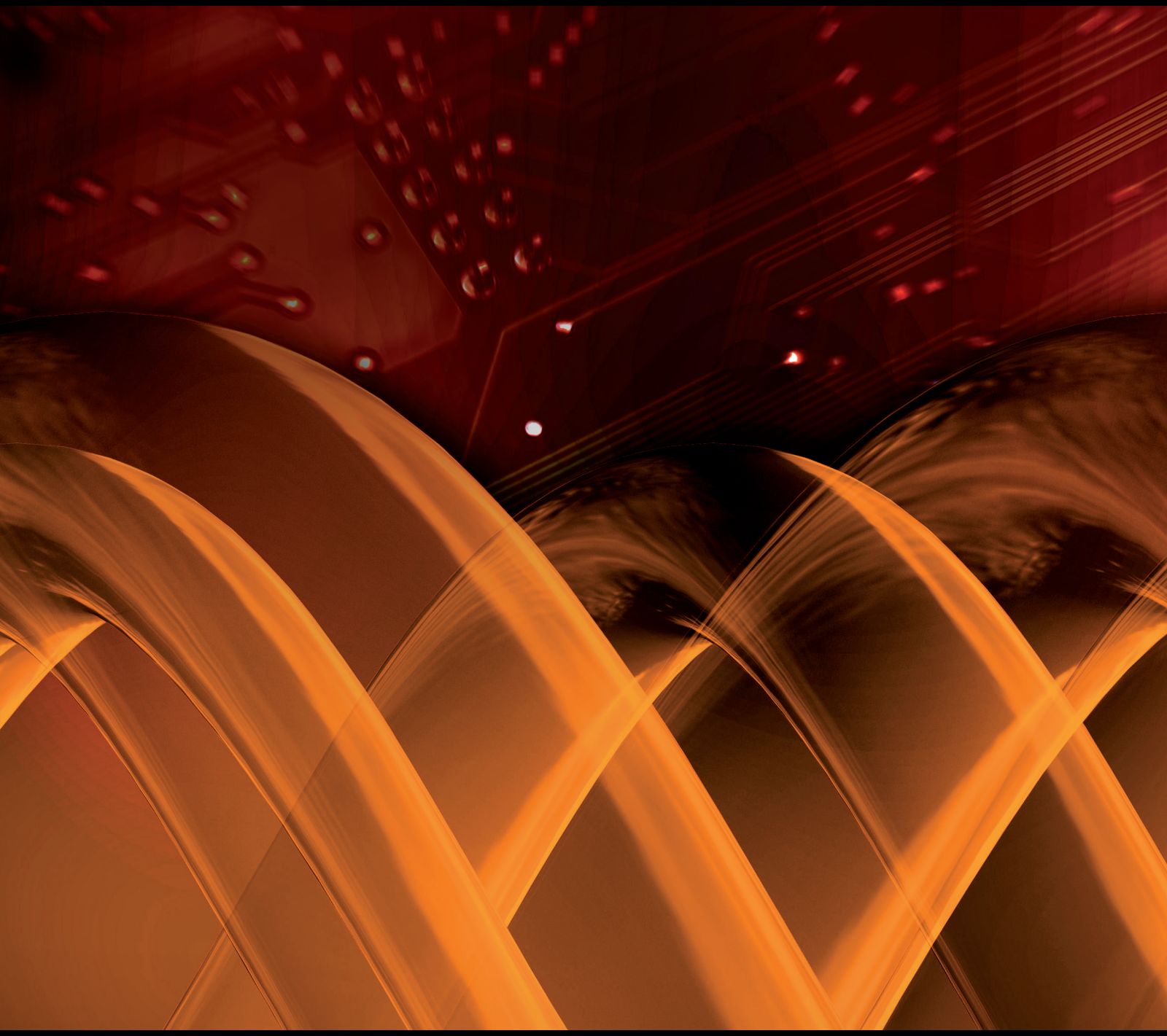
International Journal of Genomics

Plant Comparative and Functional Genomics

Guest Editors: Xiaohan Yang, Jim Leebens-Mack, Feng Chen, and Yanbin Yin



Plant Comparative and Functional Genomics

International Journal of Genomics

Plant Comparative and Functional Genomics

Guest Editors: Xiaohan Yang, Jim Leebens-Mack, Feng Chen, and Yanbin Yin

Copyright © 2015 Hindawi Publishing Corporation. All rights reserved.

This is a special issue published in "International Journal of Genomics." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Editorial Board

Jacques Camonis, France
Shen Liang Chen, Taiwan
Prabhakara V. Choudary, USA
Martine A. Collart, Switzerland
Ian Dunham, United Kingdom
Soraya E. Gutierrez, Chile
M. Hadzopoulou-Cladaras, Greece
Sylvia Hagemann, Austria
Henry Heng, USA
Eivind Hovig, Norway
Peter Little, Australia
Shalima Nair, Australia
Giuliana Napolitano, Italy
Ferenc Olasz, Hungary
Elena Pasyukova, Russia
Graziano Pesole, Italy
Giulia Piaggio, Italy
Mohamed Salem, USA
Brian Wigdahl, USA
Jinfa Zhang, USA

Contents

Plant Comparative and Functional Genomics, Xiaohan Yang, Jim Leebens-Mack, Feng Chen, and Yanbin Yin
Volume 2015, Article ID 924369, 2 pages

Quantification and Gene Expression Analysis of Histone Deacetylases in Common Bean during Rust Fungal Inoculation, Kalpalatha Melmaiee, Venu (Kal) Kalavacharla, Adrianne Brown, Antonette Todd, Yaqoob Thurston, and Sathya Elavarthi
Volume 2015, Article ID 153243, 10 pages

Divergence of the bZIP Gene Family in Strawberry, Peach, and Apple Suggests Multiple Modes of Gene Evolution after Duplication, Xiao-Long Wang, Yan Zhong, Zong-Ming Cheng, and Jin-Song Xiong
Volume 2015, Article ID 536943, 11 pages

Expressed Sequence Tags Analysis and Design of Simple Sequence Repeats Markers from a Full-Length cDNA Library in *Perilla frutescens* (L.), Eun Soo Seong, Ji Hye Yoo, Jae Hoo Choi, Chang Heum Kim, Mi Ran Jeon, Byeong Ju Kang, Jae Geun Lee, Seon Kang Choi, Bimal Kumar Ghimire, and Chang Yeon Yu
Volume 2015, Article ID 679548, 7 pages

***De Novo* Transcriptome Sequencing of the Orange-Fleshed Sweet Potato and Analysis of Differentially Expressed Genes Related to Carotenoid Biosynthesis**, Ruijie Li, Hong Zhai, Chen Kang, Degao Liu, Shaozhen He, and Qingchang Liu
Volume 2015, Article ID 843802, 10 pages

Genome-Wide Identification of Genes Probably Relevant to the Uniqueness of Tea Plant (*Camellia sinensis*) and Its Cultivars, Yan Wei, Wang Jing, Zhou Youxiang, Zhao Mingming, Gong Yan, Ding Hua, Peng Lijun, and Hu Dingjin
Volume 2015, Article ID 527054, 7 pages

Analysis of *Polygala tenuifolia* Transcriptome and Description of Secondary Metabolite Biosynthetic Pathways by Illumina Sequencing, Hongling Tian, Xiaoshuang Xu, Fusheng Zhang, Yaoqin Wang, Shuhong Guo, Xuemei Qin, and Guanhua Du
Volume 2015, Article ID 782635, 11 pages

PPCM: Combing Multiple Classifiers to Improve Protein-Protein Interaction Prediction, Jianzhuang Yao, Hong Guo, and Xiaohan Yang
Volume 2015, Article ID 608042, 7 pages

Significant Microsynteny with New Evolutionary Highlights Is Detected through Comparative Genomic Sequence Analysis of Maize CCCH IX Gene Subfamily, Wei-Jun Chen, Yang Zhao, Xiao-Jian Peng, Qing Dong, Jing Jin, Wei Zhou, Bei-Jiu Cheng, and Qing Ma
Volume 2015, Article ID 824287, 12 pages

Characterization and Development of EST-SSRs by Deep Transcriptome Sequencing in Chinese Cabbage (*Brassica rapa* L. ssp. *pekinensis*), Qian Ding, Jingjuan Li, Fengde Wang, Yihui Zhang, Huayin Li, Jiannong Zhang, and Jianwei Gao
Volume 2015, Article ID 473028, 11 pages

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 924369, 2 pages
http://dx.doi.org/10.1155/2015/924369

Editorial

Plant Comparative and Functional Genomics

Xiaohan Yang,[1] Jim Leebens-Mack,[2] Feng Chen,[3] and Yanbin Yin[4]

[1]*Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[2]*Department of Plant Biology, University of Georgia, Athens, GA 30602, USA*
[3]*Department of Plant Sciences, University of Tennessee, Knoxville, TN 37996, USA*
[4]*Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115, USA*

Correspondence should be addressed to Xiaohan Yang; yangx@ornl.gov

Received 23 November 2015; Accepted 23 November 2015

Copyright © 2015 Xiaohan Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Plants form the foundation for our global ecosystem and are essential for environmental and human health. With an increasing number of available plant genomes and tractable experimental systems, comparative and functional plant genomics research is greatly expanding our knowledge of the molecular basis of economically and nutritionally important traits in crop plants. Inferences drawn from comparative genomics are motivating experimental investigations of gene function and gene interactions. This special issue aims to highlight recent advances made in comparative and functional genomics research in plants. Nine original research articles in this special issue cover five important topics: (1) transcription factor gene families relevant to abiotic stress tolerance; (2) plant secondary metabolism; (3) transcriptome-based markers for quantitative trait locus; (4) epigenetic modifications in plant-microbe interactions; and (5) computational prediction of protein-protein interactions. The plant species studied in these articles include model species as well as nonmodel plant species of economic importance (e.g., food crops and medicinal plants).

Evolution of Transcription Factor Gene Families Relevant to Abiotic Stress Tolerance. The extant flowering plants have experienced multiple rounds of genome duplication and gene duplication is the primary source of gene family evolution. X.-L. Wang et al. in "Divergence of the bZIP Gene Family in Strawberry, Peach, and Apple Suggests Multiple Modes of Gene Evolution after Duplication" explored the evolutionary dynamics of the bZIP family, which contains transcription factors associated with plant tolerance to abiotic stress. They performed evolutionary analysis of the bZIP family in three rosaceous species in multiple aspects such as selection pressure on protein-coding sequences and genomic synteny. Another transcription factor gene family relevant to abiotic stress, the CCCH zinc finger family, was studied by W.-J. Chen et al. in "Significant Microsynteny with New Evolutionary Highlights Is Detected through Comparative Genomic Sequence Analysis of Maize CCCH IX Gene Subfamily." They performed comparative analysis of the CCCH IX subfamily in three cereal grain species (i.e., *Zea mays*, *Oryza sativa*, and *Sorghum bicolor*) and found that segmental duplication has played an important role in the expansion of this gene family. Their analysis also indicates that deletions, multiplications, inversions, and purifying selection have contributed to the evolution of the CCCH IX subfamily.

Plant Secondary Metabolism. Plants produce a wide range of secondary metabolites that underpin functional diversity in plants. Gene expression profiling through transcriptome sequencing is a powerful approach for understanding the molecular basis of plant secondary metabolism. H. Tian et al. in "Analysis of *Polygala tenuifolia* Transcriptome and Description of Secondary Metabolite Biosynthetic Pathways by Illumina Sequencing" analyzed expression of secondary metabolite biosynthetic genes in *P. tenuifolia*, a well-known medicinal plant, using RNA-seq approach. Their analysis revealed candidate genes that are potentially involved in biosynthesis of several important secondary metabolites such

as triterpene saponins and phenylpropanoid. Similarly, R. Li et al. in "*De Novo* Transcriptome Sequencing of the Orange-Fleshed Sweet Potato and Analysis of Differentially Expressed Genes Related to Carotenoid Biosynthesis" performed RNA-seq analysis of secondary metabolism in *Ipomoea batatas*, an important food crop. Through comparing the global gene expression profile in relation to the differences in the carotenoid content of two *I. batatas* cultivars, they identified more than 50 genes potentially involved in carotenoid biosynthesis. Also, Y. Wei et al. in "Genome-Wide Identification of Genes Probably Relevant to the Uniqueness of Tea Plant (*Camellia sinensis*) and Its Cultivars" performed comparative analysis of RNA-seq data in several species of the genus *Camellia*, an important source for tea production. They identified differentially expressed genes relevant to the biosynthesis of flavonoid, theanine, and caffeine. Furthermore, their sequence comparison revealed nonsynonymous mutations that are potentially related to the diversity between the two cultivars of *C. sinensis*.

Transcriptome-Based Markers for Quantitative Trait Locus. Quantitative trait locus (QTL) analysis has been widely used for elucidating the genetic basis of complex traits in plants. Molecular makers are prerequisites for QTL analysis. QTL makers can be developed from either genome sequences or transcriptome sequences. Development of genome-based markers requires genome-sequencing data, which are available only in model plant species or major crop species. For nonmodel plant (crop) systems, transcriptome-based markers can be a better choice for QTL analysis with a limited budget. E. S. Seong et al. in "Expressed Sequence Tags Analysis and Design of Simple Sequence Repeats Markers from a Full-Length cDNA Library in *Perilla frutescens* (L.)" developed simple sequence repeats (SSR) markers based on approximately 1,000 expressed sequence tags (ESTs) derived from cDNA libraries for this member of the mint family used in traditional Asian medicine. They identified 18 SSR makers that could be very useful for understanding of genomic basis of medicinal function in *P. frutescens*. Recent advance in next-generation sequencing technology greatly enhances the capability for molecular marker development. Q. Ding et al. in "Characterization and Development of EST-SSRs by Deep Transcriptome Sequencing in Chinese Cabbage (*Brassica rapa* L. ssp. *pekinensis*)" identified 10,420 SSR markers from 51,694 nonredundant unigenes assembled from RNA-seq data. This large set of SSR makers could facilitate genome-wide discovery of QTLs in Chinese cabbage. Also, R. Li et al. in "*De Novo* Transcriptome Sequencing of the Orange-Fleshed Sweet Potato and Analysis of Differentially Expressed Genes Related to Carotenoid Biosynthesis" identified 1,725 SSR markers in the transcriptome data for sweet potato.

Epigenetic Modifications in Plant-Microbe Interactions. While it is widely accepted that genetics governs plant growth, development, and response to environment, an increasing number of studies have showed that epigenetics also plays an important regulatory role in plants. The paper by K. Melmaiee et al. entitled "Quantification and Gene Expression Analysis of Histone Deacetylases in Common Bean during Rust Fungal Inoculation" revealed that epigenetic modification via histone deacetylases is involved in the response of common bean to rust fungal inoculation. The results from this paper provide new insight into the molecular mechanism underlying plant-microbe interactions.

Computational Prediction of Protein-Protein Interactions. Protein-protein interaction (PPI) is an important molecular mechanism underlying various biological processes. Computational prediction of protein-protein interactions based on protein sequences is a straightforward approach to the utilization of whole-genome gene annotation for the global view of protein-protein interaction network in an organism. Various algorithms have been developed for protein sequence-based PPI prediction, though with limited success. J. Yao et al. in "PPCM: Combing Multiple Classifiers to Improve Protein-Protein Interaction Prediction" present a machine learning approach for PPI prediction based on various features derived from protein sequences. Their results demonstrated that integration of multiple features could significantly improve the PPI prediction accuracy as compared with prediction classifiers based on individual features. This novel approach has a great potential for PPI prediction in nonmodel organisms, including plant species.

Acknowledgments

This work is supported by the Department of Energy (DOE), Office of Science, Genomic Science Program under Award no. DE-SC0008834. Thanks are due to the authors for contributing original research articles in a timely manner. Special thanks go to referees for their careful and critical evaluation of the manuscripts. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US DOE under Contract no. DE–AC05–00OR22725.

Xiaohan Yang
Jim Leebens-Mack
Feng Chen
Yanbin Yin

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 153243, 10 pages
http://dx.doi.org/10.1155/2015/153243

Research Article

Quantification and Gene Expression Analysis of Histone Deacetylases in Common Bean during Rust Fungal Inoculation

Kalpalatha Melmaiee,[1] Venu (Kal) Kalavacharla,[1,2] Adrianne Brown,[1] Antonette Todd,[1] Yaqoob Thurston,[1,3] and Sathya Elavarthi[1]

[1]*College of Agriculture and Related Sciences, Delaware State University, Dover, DE 19901, USA*
[2]*Center for Integrated Biological & Environmental Research (CIBER), Delaware State University, Dover, DE 19901, USA*
[3]*Department of Plant Science, South Dakota State University, Brookings, SD 57007, USA*

Correspondence should be addressed to Venu (Kal) Kalavacharla; vkalavacharla@desu.edu

Received 20 June 2015; Accepted 27 October 2015

Academic Editor: Feng Chen

Copyright © 2015 Kalpalatha Melmaiee et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Histone deacetylases (HDACs) play an important role in plant growth, development, and defense processes and are one of the primary causes of epigenetic modifications in a genome. There was only one study reported on epigenetic modifications of the important legume crop, common bean, and its interaction with the fungal rust pathogen *Uromyces appendiculatus* prior to this project. We measured the total active HDACs levels in leaf tissues and observed expression patterns for the selected HDAC genes at 0, 12, and 84 hours after inoculation in mock inoculated and inoculated plants. Colorimetric analysis showed that the total amount of HDACs present in the leaf tissue decreased at 12 hours in inoculated plants compared to mock inoculated control plants. Gene expression analyses indicated that the expression pattern of gene *PvSRT1* is similar to the trend of total active HDACs in this time course experiment. Gene *PvHDA6* showed increased expression in the inoculated plants during the time points measured. This is one of the first attempts to study expression levels of HDACs in economically important legumes in the context of plant pathogen interactions. Findings from our study will be helpful to understand trends of total active HDACs and expression patterns of these genes under study during biotic stress.

1. Introduction

Histone deacetylases are a family of enzymes that remove acetyl groups from lysine residues present in the N-terminal extension of core histones of nucleosomes [1] and have been found in bacteria, fungi, plants, and animals. Histone acetyltransferases (HATs) and deacetylases (HDACs) play an important role in chromatin structural modifications and epigenetic changes in many organisms. Research on histone deacetylase inhibitors (HDAC) began nearly 30 years ago when studies were laid out to understand why dimethyl sulfoxide (DMSO) caused terminal differentiation of murine erythroleukemia cells [2]. This early observation led to the development of novel pharmacological agents in the field of chromatin remodeling [1]. HDACs catalyze deacetylation reactions, which cause chromatin to coil by removing acetyl groups from lysine residues of histones. This deacetylation increases the positive charge on N-termini of the core histones. As a result, the interaction between core histones and negatively charged DNA increases which causes tight coiling of DNA, which in turn blocks access to the transcriptional machinery. The balance between the actions of HDACs, HATs, and transcriptional elements serves as a key regulatory mechanism for gene expression and in turn governs numerous developmental processes and disease states [3, 4].

HDACs are known to be involved in a myriad of plant physiological and developmental activities and in epigenetic events often for transcriptional repression of genes [5, 6]. Several studies in plants have reported that there is a direct correlation of DNA methylation, histone deacetylation, and gene suppression [5, 7, 8]. In the model plant, *Arabidopsis thaliana*, HDA6 and MET1 interact directly to silence transposable elements by modifying DNA methylation, histone acetylation, and histone methylation status [9, 10]. Genetic

analysis in *Arabidopsis* indicates that HDA6 is a component of RdDM (RNA-directed DNA methylation) pathway [7]. Even in other systems like the African clawed frog *Xenopus laevis*, relaxation of methylated DNA in oocytes by the inhibition of histone deacetylases was observed [8, 11].

Removal of acetyl groups from histones at promoter regions chiefly correlates with gene silencing and transcriptional repression. However, previous studies have also shown gene repression [5, 12] as well as activation of some genes [13]. Hence, the specificity of HDACs for regulation of distinct gene programs depends on cell identity (cell state identified by gene regulation programs) and the scale of available partner proteins along with the signaling networks of the cell [13, 14]. As an example, HDA6 in *Arabidopsis*, by interacting with different proteins, can regulate flowering time [15, 16], leaf development [17], transposon silencing [18], salt and ABA stress [19], ethylene and jasmonate signaling [20], and freezing tolerance [21]. Gene HDA19 can regulate seed maturity [22], flowering time [23], immune response, and seed dormancy by interacting with other proteins [24, 25]. HDA9 has also been reported to regulate flowering in *Arabidopsis* by repressing the floral activator AGL19 [26]. Additionally, HDA6 was shown to be involved in histone modifications by increasing gene expression in *Arabidopsis* during seed germination, salt stress, and abscisic acid treatments [27]. HDACs also showed response to various biotic stresses. In *Arabidopsis*, HDA19 showed induced expression when plants were challenged with *P. syringae* and the stability of induced transcripts was shown to be dependent on the levels of salicylic acid and pathogen-related 1 (NPR-1) gene expression [28].

Recent phylogenetic analysis of sequences from the HDACs superfamily RPD3/HDA1 from *Arabidopsis* enabled further classification into three classes, class I, class II, and class III [29]. Similarly, genome analysis of rice HDACs enabled the identification of an additional class, class IV, [30] indicating the diversity and need for further studies in other commercially important crop plants including legumes. Expression analysis of HDACs from all classes and families showed differential expression during developmental stages, environmental stresses, and hormonal stimuli [31, 32].

The long-term goal of our research is to understand epigenetic modifications in common bean during infection by the rust fungal pathogen. In this study, we report progress on our understanding of the role of HDACs during infection of common bean with the bean rust pathogen, *U. appendiculatus* race 53. We focus on understanding and quantifying total HDAC activity present in mock inoculated (MI) and inoculated (I) leaf tissues at 0, 12, and 84 hours after inoculation (hai) and analyses of the expression profiles of selected genes from each known plant HDAC family.

2. Materials and Methods

2.1. Plant Materials and Pathogen Infection. The common bean cultivar "Sierra" is resistant to common bean rust race 53 (Figure 1(a)) and carries the rust resistant genes *Ur-3* and *Crg* [33, 34]. Sierra exhibits a hypersensitive response upon inoculation with race 53, the cultivar "Olathe," which is recessive at *Ur-3* is susceptible to this pathogen race. The genotype crg, a susceptible mutant derived from Sierra which carries a mutation at the *Crg* locus also develops rust like symptoms (rusty-yellow or bright orange spots) on leaves. Both Olathe and crg were used in this experiment as control for demonstrating successful inoculations.

Plants were grown in the greenhouse as per Melmaiee et al. [35]. When plants were ten days old at the primary leaf stage, half of the seedlings from each genotype were inoculated with *U. appendiculatus* race 53 spores with 1% Tween 20 on the adaxial and abaxial sides of the two leaves and another half of the seedlings were mock inoculated (MI) with only 1% Tween 20 along with Olathe and crg as inoculation experimental controls. After inoculation, plants were placed in a growth chamber with high humidity (approximately 90%) to facilitate the establishment of fungal growth. Sierra leaf samples were collected at 0, 12, and 84 hai along with MI samples for analysis as shown in Figure 2 for nuclear extraction and total RNA isolation. The above time points were chosen based on our previous experiments [35, 36]. For each sampling time, leaves were pooled from three different plants (one leaf from each plant) and utilized for colorimetric assays, and three leaves from another set of three plants were collected and flash frozen for gene expression analyses. The entire experiment was repeated twice yielding two biological replications of the study.

2.2. Scanning Electron Microscopy. Symptomatic leaves were collected from susceptible mutant genotype crg (derived from Sierra) [33], dehydrated with ethanol, and mounted on stubs using carbon filled adhesive. The dehydrated specimens were coated with gold palladium by sputter coater 108 auto (Cressington Scientific Instruments Ltd., Watford, UK) and observed with an analytical scanning electron microscope S-2600N (Hitachi High Technologies America, Inc., Schaumburg, IL) located in the College of Agriculture & Related Sciences at DSU.

2.3. Nuclear Extraction. Nuclear extractions were carried out from MI and I leaf samples using the EpiQuik Nuclear Extraction kit 1 (Epigentek Group Inc., Farmingdale, NY). Approximately one gram of leaf samples (either flash frozen or fresh) was cut into small pieces and submerged in a 1 : 10 diluted nuclear extraction buffer 1 (NE1) with 1x dithiothreitol (DTT) in a mortar and ground thoroughly until all the leaf samples became fine paste. Samples were incubated on ice for 15 minutes and centrifuged for 10 minutes at 11,000 ×g to obtain a nuclear pellet. The supernatant was removed and 500 μL of nuclear extraction buffer 2 (NE2) containing 1x DTT was added to the nuclear pellet and incubated for another 15 minutes on ice. During this time samples were vortexed for 5 sec at three-minute intervals to increase nuclear protein concentration. Samples were then centrifuged at 14,000 g for 10 minutes at 4°C, and the nuclear protein was quantified with Qubit fluorometer (Life Technologies, Grand Island, NY) and stored at −80°C for further analyses.

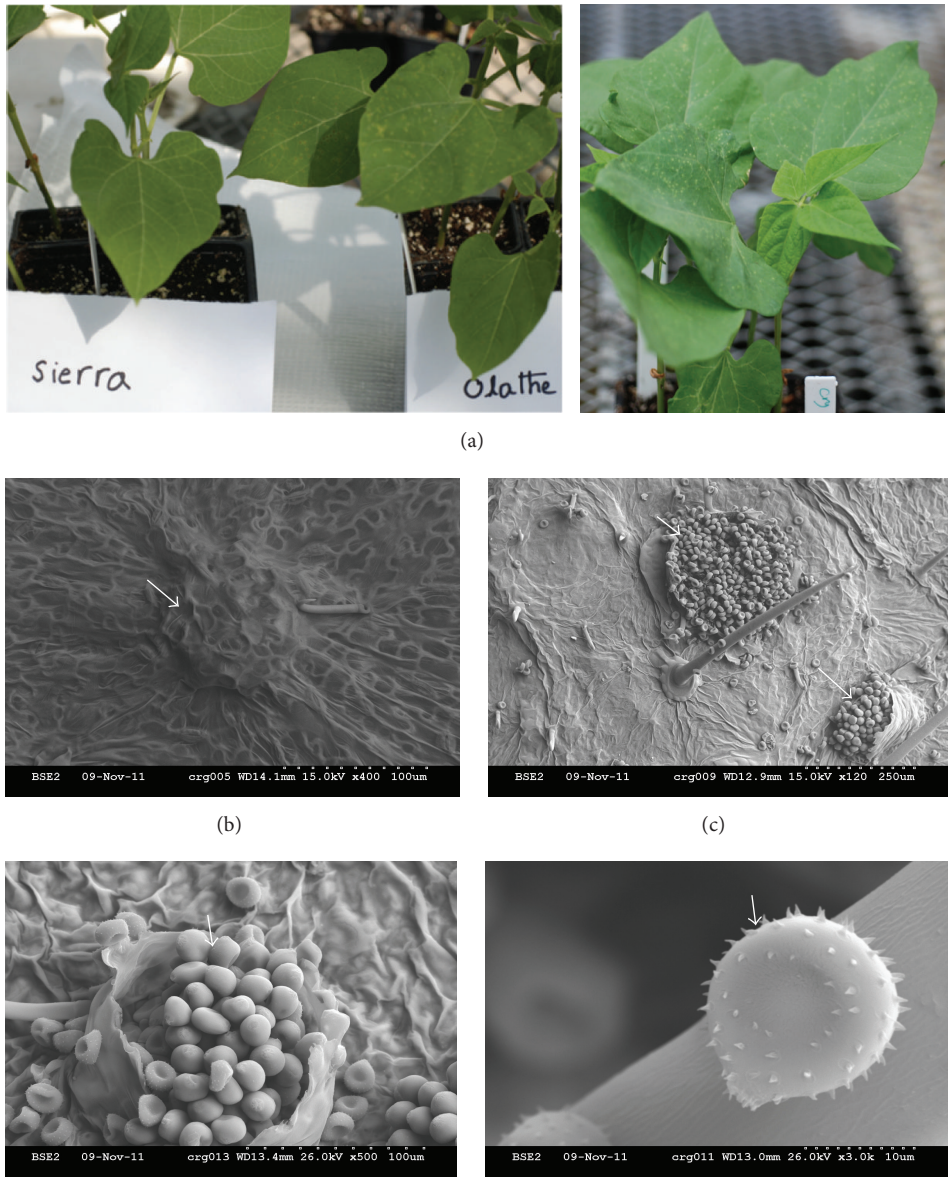


Figure 1: Common bean rust symptoms, rust pustules, and spores. (a) Ten-day-old seedlings at the primary leaf stage were inoculated with rust race 53. The susceptible genotype Olathe and susceptible mutant crg developed visible rust like symptoms after approximately 10 days of inoculation, whereas the resistant genotype Sierra was asymptomatic as expected. The genotypes Olathe and crg were used as an inoculation experimental control only and are shown here. (b–e) Scanning electron microscope pictures from a leaf of a rust susceptible genotype crg after symptoms were developed. Picture (b) is an unopened rust pustule. ((c) and (d)) Pustules which burst open at 120 and 500 times' magnification. (e) Rust spore under 3000 times' magnification. White arrows point to the pustules and fungal spores.

2.4. Quantification of HDACs by Colorimetric Method. HDAC activity was measured with 1.545 μg of nuclear protein extract following the manufacturers protocol (HDAC Assay Kit, colorimetric, Active Motif, Carlsbad, CA). As suggested in the protocol for samples with potentially low HDACs, we extended the initial HDAC reaction time to three hours. Since the kit was developed based on nuclear extracts from mammalian cells, we envisioned that extending the incubation time will help complete the deacetylation reaction. Samples were measured in triplicate (standards were measured in duplicate) using EPOCH colorimetric plate reader (Biotek, Winooski, VT) at 405 nm. In the assay reaction, a short peptide substrate was added along with the nuclear extract and other reagents as per the protocol. This substrate contains acetylated lysine residues and can be deacetylated by most HDAC enzymes. Active HDACs from the experimental samples would then bind to the added substrate by removing acetyl groups from the substrate. This reaction then yielded an HDAC-deacetylated colored product, which was measured by the colorimetric plate

Table 1: Selected representative HDAC genes for expression analysis. GenBank protein accession number and corresponding predicted common bean homologs along with location on the common bean genome.

Gene family/class	GenBank protein accession number	Phytozome common bean CDS number	HDACs names	Location on the bean genome (chromosome)	Corresponding model organism
RPD3/HDA1 family class I	AAC50038 AAK0712.1	Phvul.009G115300.1	*PvRPD3*	9	*Zea mays* *Oryza sativa*
	BAB10553	Phvul.003G203800.1	*PvHDA6*	*6*	*A. thaliana*
Class II	NP_200915 NP 200914	Phvul.003G185200.1	*PvHDA18*	*3*	*A. thaliana*
Class III	AAD40129	Phvul.001G034500.2	*PvHDA2*	*1*	*A. thaliana*
HD2 family	AAB70032	Phvul.001G186300.1	*PvHD2*	sg0.contig 03923: 2769–5777	*A. thaliana*
SIR2 family	BAB09243	Phvul.006G057700.1	*PvSRT1*	*6*	*A. thaliana*

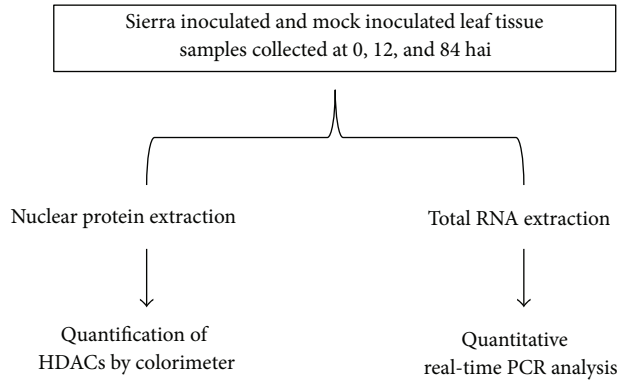


Figure 2: Flowchart outlining the experiment in this study. Same sets of materials from each biological replicate were utilized to perform colorimetric analysis and gene expression studies.

Table 2: Primer sequences utilized for qRT-PCR.

Predicted common bean gene	Primer sequences
Phvul.009G115300.1	ACATGAGCGTGTTCTGTACGTGGA TCAGCACCGCATTGGAGAACTACT
Phvul.003G203800.1	CATCCGCATGGCGCACAATCTTAT ACCCAACCTGTCACCAGACAATGA
Phvul.003G185200.1	TCTGCGGTTAGTGCATTCCAGAGT GGGTCACCAACAGCTGCATCAAAT
Phvul.001G034500.2	TCGGCATAGAGAAACTGCATCCGT ACCACCTGAAGTGAGCATGACGAT
Phvul.001G186300.1	AACTGGTAGCCCTGAACGTGAAGT TCCCATTGGCAGCACTAACTGGAA
Phvul.006G057700.1	CTTGCCAGAAGCATCACTGCCATT GGCAAGTTGCACGCTGGAGTTATT
TC362	GCTCTCCATTTGCTCCCTGTT TGAGCAATTTCAGGCACCAA

reader. The amount of deacetylated product in the reaction is directly proportional to the amount of active HDAC enzymes present in our samples [37].

2.5. Selection of HDAC Gene Sequences and Primer Design. Representative proteins from each HDAC family or class were selected based on the previous reports [29]. The GenBank protein accession numbers AAC50038, AAK0712.1, BAB10553, NP_200914, NP_200915, and AAD40129, from RPD3 (reduced potassium dependency)/HDA1 (histone deacetylase 1) family, AAB70032 from HD2 family, and BAB09243 from SIR2 (Silent Information Regulator 2) were selected for gene expression analysis. All of these sequences were derived from *Arabidopsis thaliana* except AAK01712.1 and AAC50038, which were derived from rice (*Oryza sativa*) and maize (*Zea mays*).

GenBank protein accession numbers were used to extract model organism protein sequences from NCBI database and these sequences were compared against the common bean predicted proteome derived from the common bean genome-sequencing project from http://www.phytozome.org/ [38]. The best match was selected and coding sequences (CDS) were extracted for further analysis (Supplementary File 1) (see Supplementary Material available online at http://dx.doi.org/10.1155/2015/153243). Proteins AAK0712.1 and AAC 50038.1 matched the same common bean CDS Phvul.009G115300.1 and proteins NP_200914 and NP_200915 matched Phvul.003G185200.1 and other proteins matched different common bean sequences as shown in Table 1. For convenience, we named these CDS (referenced in this study as *Phaseolus vulgaris* HDACs) as mentioned in column 4 of Table 1. For gene expression analysis, primers were designed with Primer quest software as in Table 2 and tested with common bean genomic DNA (Figure 4(a)).

2.6. Total RNA Isolation and cDNA Synthesis. Total RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA) from flash frozen pooled leaf tissues (three leaves from three plants) as per manufacturer's protocol and the RNA was digested with the enzyme rDNAse (Life Technologies, Grand Island, NY) to remove any contaminating DNA. Absence

of genomic DNA was confirmed with known primers that can amplify intronic regions as mentioned previously [39]. Total RNA was used to synthesize cDNA with ProtoScript M-MuLV First Strand cDNA synthesis kit (New England BioLabs, Beverly, MA).

2.7. Gene Expression Analysis by Quantitative Real-Time PCR (qRT-PCR). Concentrations of cDNA were equalized for all the samples under consideration and qRT-PCR analysis was carried out on Applied Biosystems 7500 real-time machine (Foster City, CA) using SYBR Green dye. Gene expression was normalized to the housekeeping gene ubiquitin-conjugating enzyme E2 UBC9 (TC362) [40] and included in each PCR run. The whole experiment was replicated twice with three technical replications for each sample analyzed. Gene expression analysis was carried out by comparative $2^{-\Delta\Delta CT}$ method [41] and used to calculate expression values and indicated in fold changes. Student's t-test was performed with a P value cutoff of 0.05.

3. Results and Discussion

3.1. HDACs Activity during Fungal Infection. Activity of total HDAC enzymes was quantified by a colorimetry-based assay. We collected leaf tissue from inoculated and mock inoculated rust resistant genotype Sierra (tissue pooled from 3 leaves for each time point) at 0, 12, and 84 hai, from which nuclear extracts were then isolated and processed. A standard curve was generated using the standards provided with the HDAC Activity Kit and optical density (OD) values of the samples from MI plants and I plants were then extrapolated. The mean values from two independent biological replicates were calculated (Figure 3). Colorimetric analysis revealed that there is a reduction in the amount of active HDACs (37.68 nmol) in the inoculated samples at 12 hai compared to mock inoculated plants (48.97 nmol), whereas at 84 hai the activity was approximately 37.0 nmol in both the samples.

The reduction in overall HDAC activity at 12 hai suggests that there may be less deacetylation reactions at this time point and more uncoiled DNA was available for transcription as there will be a demand for induction of stress resistant genes at this time. However, colorimetric analysis indicates that HDACs activity changes throughout the course of rust infection in common bean plants and differs between plants challenged or not challenged with the fungal pathogen.

3.2. Identification of Common Bean Homologous Sequences. HDAC protein sequences were obtained from *Arabidopsis* and other plant species from GenBank using corresponding protein accession numbers. These protein sequences were searched against the common bean predicted protein database and common bean CDS were obtained (Supplementary File 1) for gene expression analysis. Since common bean CDS were derived by bioinformatics analysis, corresponding primers were initially amplified with genomic DNA of the common bean (Figure 4(a)) and then with cDNA derived from the experimental samples (Figure 4(b)), which were

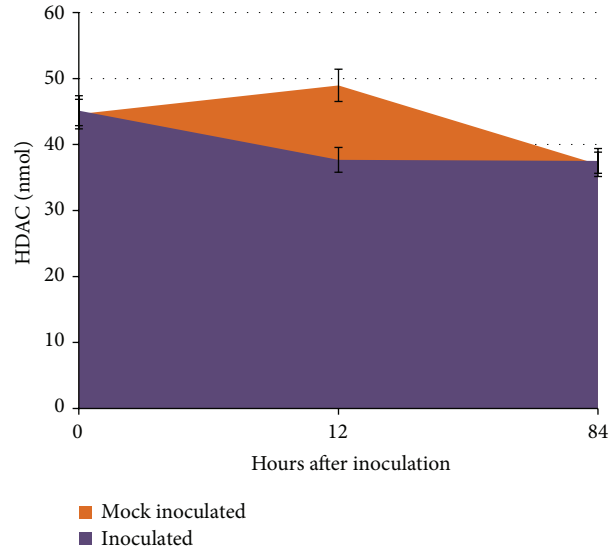


Figure 3: HDAC activity between mock inoculated and inoculated common bean. Total HDAC activity was measured based on the optical density (OD) and the amount of activity was determined based on the standard curve. The experimental values differ significantly with a probability value of 0.05%. The error bars represent standard deviation.

used for qRT-PCR analysis. All the genes tested were amplified in both the genomic DNA and cDNA samples.

3.3. Gene Expression Analysis. Between the two genes that were studied in class 1 (RPD3 family), gene *PvHDA6* showed increased expression at both 12 and 84 hai in the inoculated samples (Figure 5(a)). In the MI samples, the *PvHDA6* expression was seen to be slightly increased at 12 hai and was neutral at 84 hai samples (Figure 5(b)). *PvRPD3* expression was seen to be slightly increased at 12 hai in the I samples. However, both the genes showed slight reduction in expression at 84 hai in the MI samples. Gene *PvHDA18* from class II HDACs showed increased expression at 12 hai and reduced expression at 84 hai in MI and I samples (Figures 5(c) and 5(d)). Class III gene *PvHDA2* showed increased expression at both 12 and 84 hai in I samples whereas its expression was neutral in MI samples at both the time points (Figures 5(e) and 5(f)).

Similar results as observed in this study were seen in a *Pseudomonas syringae* resistant *Arabidopsis* plant with RPD3/HDA class gene *HDA19*. Increased expression levels of *HDA19* were seen when plants were inoculated with the bacterial pathogen pstDC3000, a virulent strain of *P. syringae pv. tomato* [28]. *HDA19* by interacting with the transcription factors *WRKY38* and *WRKY62* was suggested to help fine-tune basal defense responses to pathogen attack in *Arabidopsis* [28]. Contrastingly, Choi et al. [42] showed that *HDA19* played a negative role in basal defense response mediated by salicylic acid-dependent signaling pathway, where they have observed increased expression of pathogen defense genes in *HDA19* mutant plants.

In our analysis, gene *PvHD2* was neutral in its expression at 12 hai and showed decreased expression at 84 hai in both

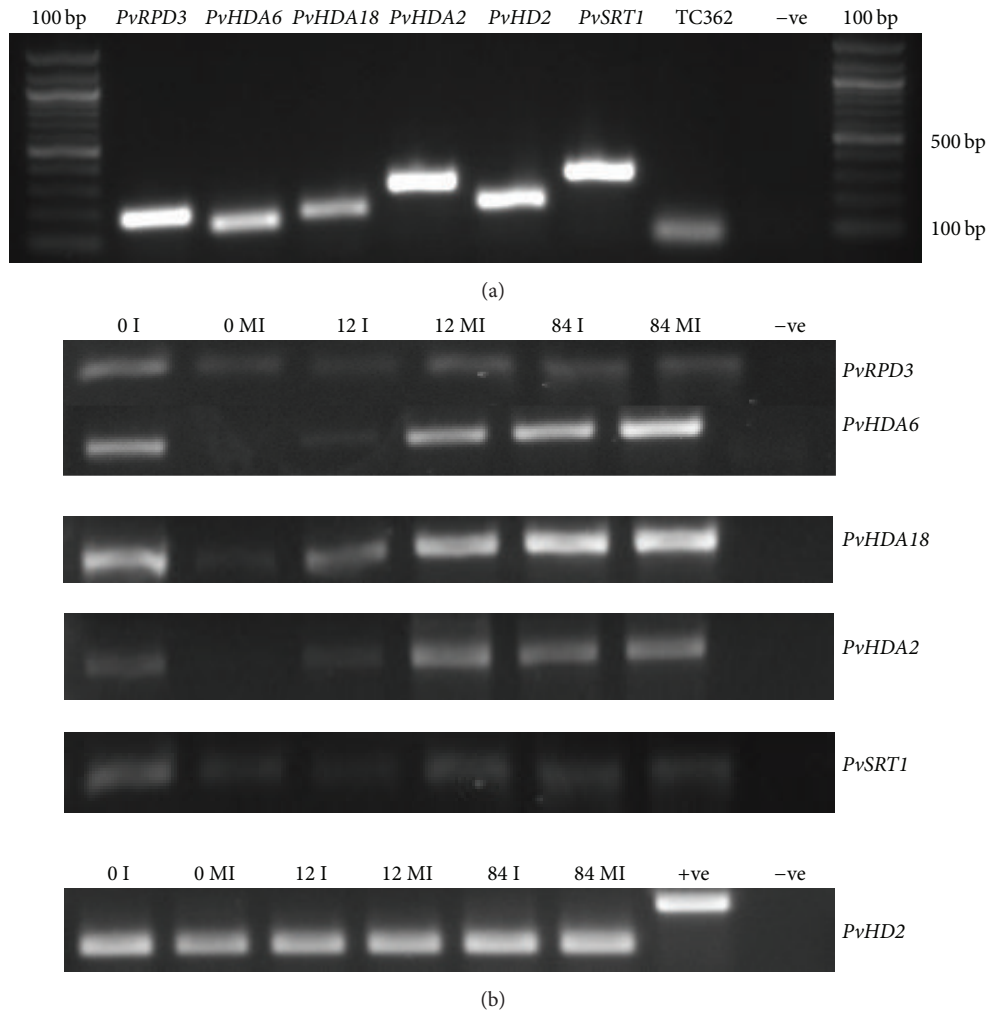


Figure 4: PCR amplification of HDAC genes under study. (a) Primers for the selected genes were amplified by PCR with Sierra genomic DNA and electrophoresed on a 2% agarose gel. In positive (+ve) control, reference gene TC362 primers were used and in negative (−ve) control no primers were added to the PCR reaction. (b) The same HDACs primers were tested with experimental cDNA by PCR amplification and electrophoresed on a 2% agarose gel. In +ve control, genomic DNA was used instead of cDNA and in −ve control no primers were added to the PCR reaction.

MI and I leaf samples (Figures 5(g) and 5(h)). In a recent report, the tobacco *NtHD2a* and *NtHD2b* genes showed a rapid and strong reduction in their expression after treating the tobacco cells with cryptogein, an elicitor of tobacco defense and cell death [43].

Based on earlier findings, the HD2 class is plant specific and found only in plants [29, 44]. Differential expression of the barley HD2 genes (*HvHDAC2* and *HvHDAC2-2*) was observed in different tissues and during seed development in barley [32] and they also exhibited differential expression in barley cultivars with varying seed size. In the same study, these genes responded to plant stress hormones such as jasmonic acid (JA), abscisic acid (ABA), and salicylic acid (SA) suggesting a possible role in epigenetic regulation due to biotic and abiotic stresses and during seed development.

Gene *PvSRT1* from the SIR2 family showed contrasting expression at 12 hai, for which the expression levels were decreased in I samples and increased in MI samples. However we observed that the expression levels were increased at 84 hai in both the samples as in Figures 5(i) and 5(j). *SIRT1* was reported to regulate miRNA in Alzheimer's disease patients [45, 46]. SIR2 genes were found to be highly expressed in highly proliferating stages such as the seedling and developing panicle stages [47]. In the current study, we have used 10-day-old common bean seedlings for inoculation; hence this might be a possible reason for the presence of a higher quantity of SIRT proteins overall. This may be why the trends of active HDACs and the expression profiles of *SIRT* gene are similar. Additionally, we note that in this study we were able to measure one representative gene (*PvSRT1*) from this class, and it will be interesting to measure the second gene. Pandey et al. [29] pointed out that only two genes from the SIRT class are currently known in plants in this class. Another consideration for future quantification experiments would be to determine protein turnover changes.

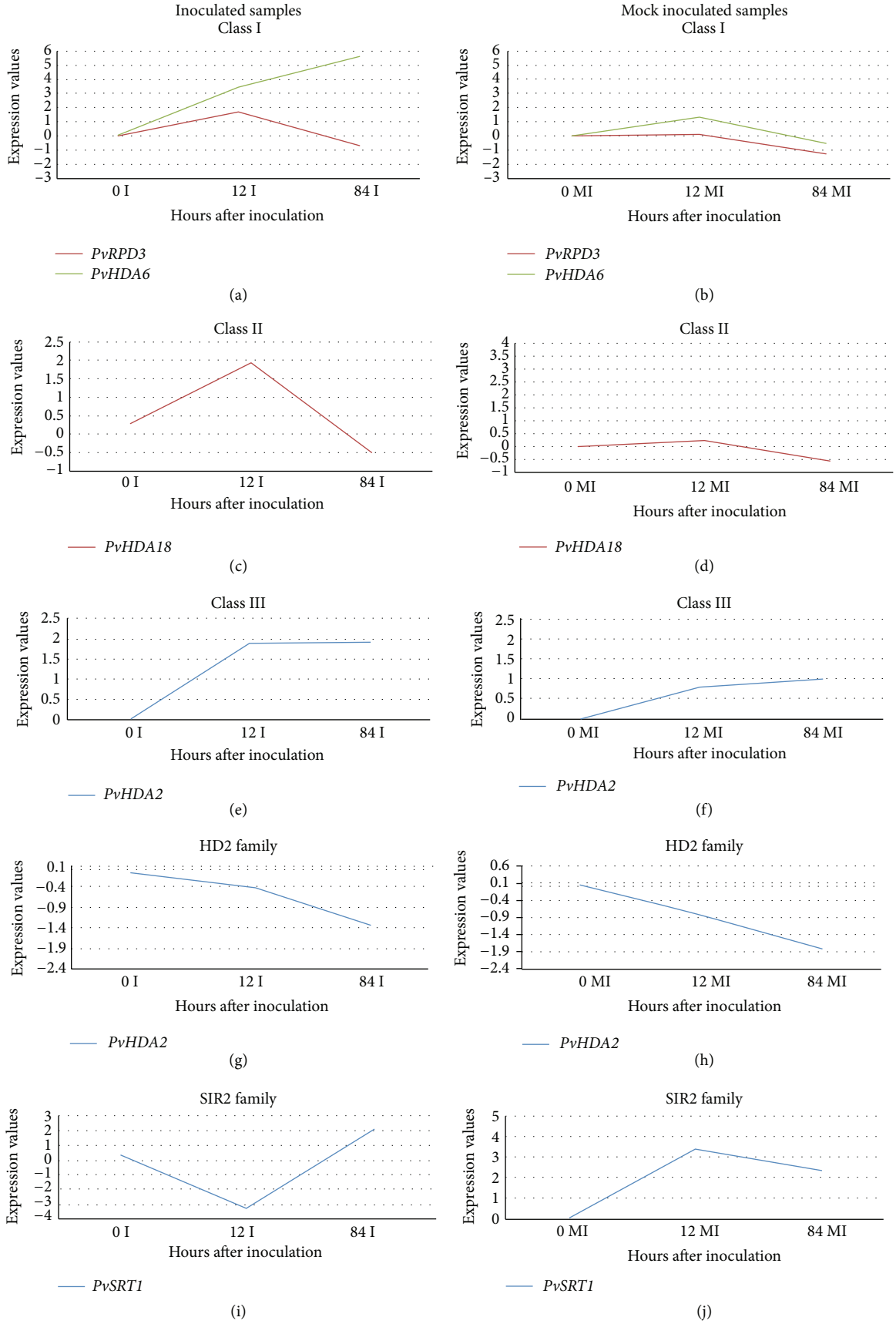


FIGURE 5: qRT-PCR analysis of HDAC genes. Figures on the left-hand side are from inoculated samples, while figures on the right-hand side are from mock inoculated samples. Sampling time points are shown in x-axis and $\Delta\Delta$CT values are shown in y-axis. Sierra 0 hai mock inoculated samples and the endogenous gene TC362 were used for calculating expression values.

Reduced expression of the rice SIR2 family gene *OsSRT1* by specific RNA interference increased histone H3K9 acetylation, decreased H3K9 dimethylation, and also led to the development of cell death and symptoms related to plant hypersensitive response during incompatible interaction with pathogen [47]. Interestingly, in our study, *PvSIRT1* showed decreased gene expression at 12 hai in leaves of inoculated plants in the bean genotype that also exhibits hypersensitive response.

HDACs play an important role in plant growth, development [48], flowering, seed maturity, and defense/tolerance to biotic and abiotic stresses. Each HDAC gene has unique functions and these genes substitute or complement each other's function. A recent observation indicated that rice HDAC genes showed more divergent functions than their homologs in *Arabidopsis* [49] and the same study also showed that their expression is tissue/organ specific in rice.

Based on the available literature, HDAC genes interact with histone and nonhistone proteins as well as other regulatory elements. HDA6 has been reported to interact with small interfering RNAs (siRNAs) that are generated through the RdDM pathway to suppress gene activity [50, 51]. HDA9 has been reported to regulate flowering in *Arabidopsis* by repressing flower-activating gene AGL 19 [26]. Histone deacetylase HDA6 is required for freezing tolerance [7]. HDA19 by interacting with WRKY 38 and WRKY62 showed enhanced basal resistance to bacterial pathogen [28].

In conclusion, reduced total HDACs activity was observed at 12 hai in rust inoculated bean plants compared to mock inoculated plants. Majority of the RPD3/HDA1 family of HDACs studied showed increased expression at least in one time point observed after inoculation. The *PvHD2* gene of plant specific HDACs did not show differential expression with inoculation and may possibly be developmentally regulated. Additionally, the *PvSIRT1* gene showed reduced expression at 12 hai in inoculated samples. Epigenetic analysis in common bean itself is in its infancy. This is one of the first attempts to try to understand HDACs gene regulation in common bean. As HDACs play important roles in chromatin modification, in normal plant developmental process, and in biotic/abiotic responses, our findings can be helpful to study other commercially important legume crops.

Disclosure

The funding agencies USDA or NSF were not involved in designing and implementing the experiments.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors acknowledge USDA funding through Grant nos. 2007-38814-18458 and 2008-38814-04735 to Venu (Kal) Kalavacharla. They also acknowledge National Science Foundation Grant no. DBI-1003917 and Delaware EPSCoR Grant nos. IIA-1301765 and EPS-0814251 and the State of Delaware to Venu (Kal) Kalavacharla. Additionally, they acknowledge members of the Molecular Genetics & Epigenomics Laboratory (MGEL) at DSU and the College of Agriculture & Related Sciences for support of this research activity.

References

[1] M. Paris, M. Porcelloni, M. Binaschi, and D. Fattori, "Histone deacetylase inhibitors: from bench to clinic," *Journal of Medicinal Chemistry*, vol. 51, no. 6, pp. 1505–1529, 2008.

[2] P. A. Marks, "Discovery and development of SAHA as an anticancer agent," *Oncogene*, vol. 26, no. 9, pp. 1351–1356, 2007.

[3] O. Pontes, R. J. Lawrence, M. Silva et al., "Postembryonic establishment of megabase-scale gene silencing in nucleolar dominance," *PLoS ONE*, vol. 2, no. 11, Article ID e1157, 2007.

[4] L. M. Smith, O. Pontes, I. Searle et al., "An SNF2 protein associated with nuclear RNA silencing and the spread of a silencing signal between cells in *Arabidopsis*," *The Plant Cell*, vol. 19, no. 5, pp. 1507–1521, 2007.

[5] X. Liu, S. Yang, M. Zhao et al., "Transcriptional repression by histone deacetylases in plants," *Molecular Plant*, vol. 7, no. 5, pp. 764–772, 2014.

[6] R. L. Momparler, "Cancer epigenetics," *Oncogene*, vol. 22, no. 43, pp. 6479–6483, 2003.

[7] J.-M. Kim, T. K. To, and M. Seki, "An epigenetic integrator: new insights into genome regulation, environmental stress responses and developmental controls by Histone Deacetylase 6," *Plant and Cell Physiology*, vol. 53, no. 5, pp. 794–800, 2012.

[8] P. L. Jones, G. J. C. Veenstra, P. A. Wade et al., "Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription," *Nature Genetics*, vol. 19, no. 2, pp. 187–191, 1998.

[9] S. H. Rangwala and E. J. Richards, "Differential epigenetic regulation within an *Arabidopsis retroposon* family," *Genetics*, vol. 176, no. 1, pp. 151–160, 2007.

[10] B. P. May, Z. B. Lippman, Y. Fang, D. L. Spector, and R. A. Martienssen, "Differential regulation of strand-specific transcripts from *Arabidopsis* centromeric satellite repeats," *PLoS Genetics*, vol. 1, no. 6, article e79, 2005.

[11] P. A. Wade, A. Gegonne, P. L. Jones, E. Ballestar, F. Aubry, and A. P. Wolffe, "Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation," *Nature Genetics*, vol. 23, no. 1, pp. 62–66, 1999.

[12] D. Jiang, W. Yang, Y. He, and R. M. Amasino, "*Arabidopsis* relatives of the human lysine-specific Demethylase1 repress the expression of *FWA* and *Flowering Locus C* and thus promote the floral transition," *The Plant Cell*, vol. 19, no. 10, pp. 2975–2987, 2007.

[13] M. Haberland, R. L. Montgomery, and E. N. Olson, "The many roles of histone deacetylases in development and physiology: implications for disease and therapy," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 32–42, 2009.

[14] D. Hnisz, A. F. Bardet, C. J. Nobile et al., "A histone deacetylase adjusts transcription kinetics at coding sequences during *Candida albicans* morphogenesis," *PLoS Genetics*, vol. 8, no. 12, Article ID e1003118, 2012.

[15] J.-H. Jung, J.-H. Park, S. Lee et al., "The cold signaling attenuator HIGH EXPRESSION OF OSMOTICALLY RESPONSIVE GENE1 activates *FLOWERING LOCUS C* transcription via chromatin remodeling under short-term cold stress in *Arabidopsis*," *The Plant Cell*, vol. 25, no. 11, pp. 4378–4390, 2013.

[16] J. Yun, Y.-S. Kim, J.-H. Jung, P. J. Seo, and C.-M. Park, "The AT-hook motif-containing protein AHL22 regulates flowering initiation by modifying Flowering Locus T chromatin in *Arabidopsis*," *Journal of Biological Chemistry*, vol. 287, no. 19, pp. 15307–15316, 2012.

[17] M. Luo, X. Liu, P. Singh et al., "Chromatin modifications and remodeling in plant abiotic stress responses," *Biochimica et Biophysica Acta (BBA)—Gene Regulatory Mechanisms*, vol. 1819, no. 2, pp. 129–136, 2012.

[18] X. Liu, C.-W. Yu, J. Duan et al., "HDA6 Directly interacts with DNA methyltransferase MET1 and maintains transposable element silencing in *Arabidopsis*," *Plant Physiology*, vol. 158, no. 1, pp. 119–129, 2012.

[19] M. Luo, Y.-Y. Wang, X. Liu et al., "HD2C interacts with HDA6 and is involved in ABA and salt stress response in *Arabidopsis*," *Journal of Experimental Botany*, vol. 63, no. 8, Article ID ers059, pp. 3297–3306, 2012.

[20] Z. Zhu, F. An, Y. Feng et al., "Derepression of ethylene-stabilized transcription factors (EIN3/EIL1) mediates jasmonate and ethylene signaling synergy in *Arabidopsis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 30, pp. 12539–12544, 2011.

[21] T. K. To, K. Nakaminami, J.-M. Kim et al., "*Arabidopsis* HDA6 is required for freezing tolerance," *Biochemical and Biophysical Research Communications*, vol. 406, no. 3, pp. 414–419, 2011.

[22] Y. Zhou, B. Tan, M. Luo et al., "HISTONE DEACETYLASE$_{19}$ interacts with HSL$_1$ and participates in the repression of seed maturation genes in *Arabidopsis* seedlings," *The Plant Cell*, vol. 25, no. 1, pp. 134–148, 2013.

[23] X. Gu, Y. Wang, and Y. He, "Photoperiodic regulation of flowering time through periodic histone deacetylation of the florigen gene FT," *PLoS Biology*, vol. 11, no. 9, Article ID e1001649, 2013.

[24] Z. Wang, H. Cao, Y. Sun et al., "*Arabidopsis* paired amphipathic helix proteins SNL1 and SNL2 redundantly regulate primary seed dormancy via abscisic acid-ethylene antagonism mediated by histone deacetylation," *The Plant Cell*, vol. 25, no. 1, pp. 149–166, 2013.

[25] Z. Zhu, F. Xu, Y. Zhang et al., "*Arabidopsis* resistance protein SNC1 activates immune responses through association with a transcriptional corepressor," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 31, pp. 13960–13965, 2010.

[26] W. Kim, D. Latrasse, C. Servet, and D.-X. Zhou, "Arabidopsis histone deacetylase HDA9 regulates flowering time through repression of AGL19," *Biochemical and Biophysical Research Communications*, vol. 432, no. 2, pp. 394–398, 2013.

[27] L.-T. Chen, M. Luo, Y.-Y. Wang, and K. Wu, "Involvement of *Arabidopsis* histone deacetylase HDA6 in ABA and salt stress response," *Journal of Experimental Botany*, vol. 61, no. 12, pp. 3345–3353, 2010.

[28] K.-C. Kim, Z. Lai, B. Fan, and Z. Chen, "*Arabidopsis* WRKY38 and WRKY62 transcription factors interact with histone deacetylase 19 in basal defense," *The Plant Cell*, vol. 20, no. 9, pp. 2357–2371, 2008.

[29] R. Pandey, A. Müller, C. A. Napoli et al., "Analysis of histone acetyltransferase and histone deacetylase families of *Arabidopsis thaliana* suggests functional diversification of chromatin modification among multicellular eukaryotes," *Nucleic Acids Research*, vol. 30, no. 23, pp. 5036–5055, 2002.

[30] W. Fu, K. Wu, and J. Duan, "Sequence and expression analysis of histone deacetylases in rice," *Biochemical and Biophysical Research Communications*, vol. 356, no. 4, pp. 843–850, 2007.

[31] K. Demetriou, A. Kapazoglou, K. Bladenopoulos, and A. S. Tsaftaris, "Epigenetic chromatin modifiers in barley: II. Characterization and expression analysis of the HDA1 family of barley histone deacetylases during development and in response to jasmonic acid," *Plant Molecular Biology Reporter*, vol. 28, no. 1, pp. 9–21, 2010.

[32] K. Demetriou, A. Kapazoglou, A. Tondelli et al., "Epigenetic chromatin modifiers in barley. I. Cloning, mapping and expression analysis of the plant specific HD2 family of histone deacetylases from barley, during seed development and after hormonal treatment," *Physiologia Plantarum*, vol. 136, no. 3, pp. 358–368, 2009.

[33] V. Kalavacharla, J. R. Stavely, J. R. Myers, and P. E. McClean, "*Crg*, a gene required for Ur-3-mediated rust resistance in common bean, maps to a resistance gene analog cluster," *Molecular Plant-Microbe Interactions*, vol. 13, no. 11, pp. 1237–1242, 2000.

[34] K. Melmaiee, A. Todd, P. McClean et al., "Identification of molecular markers associated with the deleted region in common bean (*Phaseolus vulgaris*) ur-3 mutants," *Australian Journal of Crop Science*, vol. 7, no. 3, pp. 354–360, 2013.

[35] K. Melmaiee, A. Brown, N. Kendall, and V. Kalavacharla, "Expression profiling of wrky transcription factors in common bean during rust fungal infection," Report of the Bean Improvement Cooperative 55, 2012.

[36] V. Ayyappan, V. Kalavacharla, J. Thimmapuram et al., "Genome-wide profiling of histone modifications (H3K9$_{me2}$ and H4K12$_{ac}$) and gene expression in rust (*Uromyces appendiculatus*) inoculated common bean (*Phaseolus vulgaris* L.)," *PLoS ONE*, vol. 10, no. 7, Article ID e0132176, 2015.

[37] X.-J. Yang and S. Grégoire, "Class II histone deacetylases: from sequence to function, regulation, and clinical implication," *Molecular and Cellular Biology*, vol. 25, no. 8, pp. 2873–2884, 2005.

[38] J. Schmutz, P. E. McClean, S. Mamidi et al., "A reference genome for common bean and genome-wide analysis of dual domestications," *Nature Genetics*, vol. 46, no. 7, pp. 707–713, 2014.

[39] V. Kalavacharla, Z. Liu, B. C. Meyers, J. Thimmapuram, and K. Melmaiee, "Identification and analysis of common bean (*Phaseolus vulgaris* L.) transcriptomes by massively parallel pyrosequencing," *BMC Plant Biology*, vol. 11, article 135, 2011.

[40] G. Hernández, M. Ramírez, O. Valdés-López et al., "Phosphorus stress in common bean: root transcript and metabolic responses," *Plant Physiology*, vol. 144, no. 2, pp. 752–767, 2007.

[41] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method," *Methods*, vol. 25, no. 4, pp. 402–408, 2001.

[42] S.-M. Choi, H.-R. Song, S.-K. Han et al., "HDA19 is required for the repression of salicylic acid biosynthesis and salicylic acid-mediated defense responses in *Arabidopsis*," *The Plant Journal*, vol. 71, no. 1, pp. 135–146, 2012.

[43] S. Bourque, A. Dutartre, V. Hammoudi et al., "Type-2 histone deacetylases as new regulators of elicitor-induced cell death in plants," *New Phytologist*, vol. 192, no. 1, pp. 127–139, 2011.

[44] A. Lusser, G. Brosch, A. Loidl, H. Haas, and P. Loidl, "Identification of maize histone deacetylase HD2 as an acidic nucleolar phosphoprotein," *Science*, vol. 277, no. 5322, pp. 88–91, 1997.

[45] A. Zovoilis, H. Y. Agbemenyah, R. C. Agis-Balboa et al., "microRNA–34c is a novel target to treat dementias," *The EMBO Journal*, vol. 30, no. 20, pp. 4299–4308, 2011.

[46] N. Schonrock, Y. D. Ke, D. Humphreys et al., "Neuronal microrna deregulation in response to Alzheimer's disease amyloid-β," *PLoS ONE*, vol. 5, no. 6, Article ID e11070, 2010.

[47] L. Huang, Q. Sun, F. Qin, C. Li, Y. Zhao, and D.-X. Zhou, "Down-regulation of a Silent Information Regulator2-related histone deacetylase gene, OsSRT1, induces DNA fragmentation and cell death in rice," *Plant Physiology*, vol. 144, no. 3, pp. 1508–1519, 2007.

[48] H. H. Tai, G. C. C. Tai, and T. Beardmore, "Dynamic histone acetylation of late embryonic genes during seed germination," *Plant Molecular Biology*, vol. 59, no. 6, pp. 909–925, 2005.

[49] Y. Hu, F. Qin, L. Huang et al., "Rice histone deacetylase genes display specific expression patterns and developmental functions," *Biochemical and Biophysical Research Communications*, vol. 388, no. 2, pp. 266–271, 2009.

[50] W. Aufsatz, M. F. Mette, J. van der Winden, M. Matzke, and A. J. M. Matzke, "HDA6, a putative histone deacetylase needed to enhance DNA methylation induced by double-stranded RNA," *The EMBO Journal*, vol. 21, no. 24, pp. 6832–6841, 2002.

[51] W. Aufsatz, T. Stoiber, B. Rakic, and K. Naumann, "Arabidopsis histone deacetylase 6: a green link to RNA silencing," *Oncogene*, vol. 26, no. 37, pp. 5477–5488, 2007.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 536943, 11 pages
http://dx.doi.org/10.1155/2015/536943

Research Article

Divergence of the bZIP Gene Family in Strawberry, Peach, and Apple Suggests Multiple Modes of Gene Evolution after Duplication

Xiao-Long Wang, Yan Zhong, Zong-Ming Cheng, and Jin-Song Xiong

College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China

Correspondence should be addressed to Jin-Song Xiong; jsxiong@njau.edu.cn

Received 15 September 2015; Revised 10 November 2015; Accepted 11 November 2015

Academic Editor: Yanbin Yin

Copyright © 2015 Xiao-Long Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The basic leucine zipper (bZIP) transcription factors are the most diverse members of dimerizing transcription factors. In the present study, 50, 116, and 47 *bZIP* genes were identified in *Malus domestica* (apple), *Prunus persica* (peach), and *Fragaria vesca* (strawberry), respectively. Species-specific duplication was the main contributor to the large number of *bZIPs* observed in apple. After WGD in apple genome, orthologous *bZIP* genes corresponding to strawberry on duplicated regions in apple genome were retained. However, in peach ancestor, these syntenic regions were quickly lost or deleted. Maybe the positive selection contributed to the expansion of clade S to adapt to the development and environment stresses. In addition, purifying selection was mainly responsible for *bZIP* sequence-specific DNA binding. The analysis of orthologous pairs between chromosomes indicates that these orthologs derived from one gene duplication located on one of the nine ancient chromosomes in the Rosaceae. The comparative analysis of *bZIP* genes in three species provides information on the evolutionary fate of *bZIP* genes in apple and peach after they diverged from strawberry.

1. Introduction

Many of the biological processes in cell or organism, such as responses to the environment and progression through the cell cycle, metabolic and physiological balance are influenced or controlled by regulation of gene expression at the level of transcription. Development is based on the cellular capacity for differential gene expression and is controlled by transcription factors acting as switches of regulatory cascades [1]. Alterations in the expression of genes coding for transcription factors (TFs) are emerging as a major source of the diversity and change that underlie evolution [2]. Presently, at least 64 families of transcription factors have been identified in the plant kingdom [3]. The bZIP proteins represent a large family of TFs with a DNA-binding domain rich in basic amino acid residues, which is adjacent to a leucine zipper dimerization domain (N-x7-R/K-x9) for sequence-specific DNA binding, and a leucine zipper, which is composed of several heptad repeats of Leu or other bulky hydrophobic amino acids, such as Ile, Val, Phe, or Met, for dimerization specificity [4–7]. In addition, the majority of characterized plant *bZIP* genes to date have been associated with enhancing plant tolerance to diverse types of abiotic stress [8–14].

Recent *bZIP* gene sequence analyses in *Arabidopsis* [5], rice [6], castor bean [15], maize [16], sorghum [17], cucumber [18], and grape [19], further indicated illegitimate recombination (IR) as a major source of duplications and deletions [20]. The evidence obtained from these analyses suggests that gene duplications in a common ancestor of those plants gave rise to *bZIP* genes. Therefore, the very earliest origins of the *bZIP* gene family are associated with a series of gene duplications. A total of 75 and 89 *bZIP* genes have been identified in *Arabidopsis* [5] and rice (*Oryza sativa*) [6], respectively. The *bZIP* genes in these two genera have been classified into 10 groups and 11 groups, respectively, based on DNA binding specificity and sequence similarity.

The Rosaceae is one of the most economical plant families [21] composed by some 90 genera with over 3000 distinct species which have $x = 7$ to $x = 17$ chromosomes [22]. According to a phylogenetic treatment based on DNA sequence, data of nuclear and chloroplast genomic regions in Rosaceae reclassified the genus into Dryadoideae, Rosoideae, and Spiraeoideae, each containing a number of distinct supertribes [22]. *Prunus* and *Malus* are included in the Spiraeoideae, supertribe Amygdaleae, and Pyrodae (tribe Pyrinae), respectively, whilst *Fragaria* is included in the Rosoideae, supertribe Rosodae (tribe Fragariinae) [23]. After the rapid evolution of Rosaceae, members of the family display remarkable phenotypic diversity, plant habit, chromosome number, and fruit type which evolved independently on more than one opportunity [24, 25]. A better understanding of how the *bZIP* genes within the Rosaceae arose would provide an insight into how evolution can lead rapidly to diversification. The genomes of three Rosaceous species, woodland strawberry [26], domesticated apple [27], and peach [28], have been recently sequenced, providing an opportunity to conduct a high-resolution comparison of their genomes. In this study, we identified 50, 116, and 47 bZIP transcription factors based on the complete genome sequences of strawberry, apple, and peach. Further, through phylogenetic analysis, *Ka/Ks* ratios of genes and bZIP domains, and orthologous relationships among chromosomes, we explain the evolutionary history of *bZIP* genes in detail.

2. Methods

2.1. Data Resources and the Identification of bZIP Genes. *Fragaria vesca* (strawberry, v1.1), *Malus domestica* (apple, v1.0), and *Prunus persica* (peach, v1.0) genomic and annotation data were downloaded from the Genome Database for Rosaceae (GDR, http://www.Rosaceae.org/) [26–28]. The genome sequences of *Brassica rapa* (v1.3), *Solanum lycopersicum* (iTAG2.3), *Chlamydomonas reinhardtii* (v5.5), *Theobroma cacao* (v1.1), *Selaginella moellendorffii* (v1.0), *Populus trichocarpa* (v3.0), *Medicago truncatula* (Mt4.0v1), *Cucumis sativus* (v1.0), *Carica papaya* (ASGPBv0.4), and *Physcomitrella patens* (v3.0) were downloaded from Phytozome (http://www.phytozome.net/) [29]. Genomic data on *M. acuminate* (v1) (http://banana-genome.cirad.fr/), *Saccharomyces cerevisiae* (v1) (http://www.yeastgenome.org/), and *Cyanidioschyzon merolae* (http://merolae.biol.s.u-tokyo.ac.jp/) were also downloaded for inclusion in the analyses. The *bZIP* genes in the genomes of *Vitis vinifera* [19], *Arabidopsis thaliana* [5], and rice (*Oryza sativa*) [6] were previously identified.

The Hidden Markov Model (HMM) profiles of the bZIP domain (PF00170) were retrieved from Pfam 27.0 [30] and used for identifying the *bZIP* genes from the downloaded database of genomes using HMMER3.0 [31]. All output genes with a default *E*-value (<1.0) were collected and the online software SMART (http://smart.embl-heidelberg.de/) was used to confirm the integrity of the bZIP domain using an *E*-value of <0.1 [32]. Incorrectly predicted genes were removed. Finally, the sequences of nonredundant genes with high confidence were collected and assigned as *bZIP* genes.

2.2. Alignment and Phylogenetic Analysis of bZIP Genes. Based on the location (Table S2 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/536943) predicted in the Pfam 27.0 [30] of conserved domains in complete predicted bZIP protein sequences, the conserved domain sequences of bZIP proteins were extracted and aligned using ClustalX (version 1.83) [33]. The phylogenetic trees were generated with MEGA 5.0 [34] using the Neighbor-Joining (NJ) method and number of difference model [35]. 1,000 bootstraps were used to evaluate the significance of the phylogenetic trees.

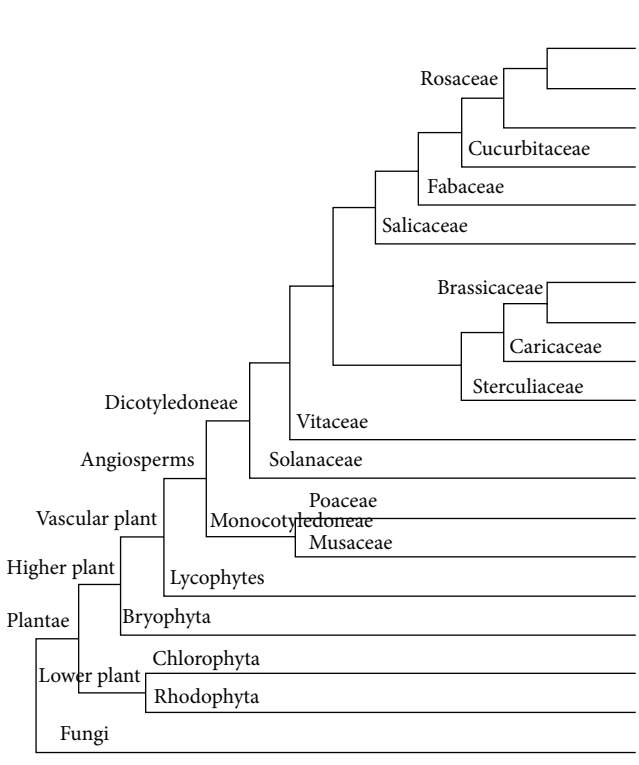
2.3. Synteny Analysis of Strawberry, Apple, and Peach Genomes. For synteny analysis, syntenic genes within the strawberry, apple, and peach genomes, as well as between strawberry and apple, strawberry and peach, peach and apple genomes, were downloaded from the Plant Genome Duplication Database [36] (PGDD, http://chibba.agtec.uga.edu/duplication/) and those containing *bZIP* genes were identified and analyzed. We identified the syntenic gene pairs from the same and different species within the same clade from phylogenetic analysis as paralogous and orthologous genes.

2.4. Estimation of Nonsynonymous Substitutions and Synonymous Substitutions. The nucleotide sequences of *bZIP* gene and bZIP domain in each clade except for UN were aligned by using Clustalw 2.0 [37]. The nonsynonymous substitutions (*Ka*) and synonymous substitutions (*Ks*) and nonsynonymous to synonymous substitution ratios (*Ka*/*Ks*) were estimated in each gene family according to the alignments in MEGA 5.0 [38]. In order to detect selection pressure of different clades of *bZIPs* in phylogenetic tree (A, B, C, D, E, F, G, H, I, and S), *Ka/Ks* ratio greater than 1, less than 1, and equal to 1 represents positive selection, negative or stabilizing selection, and neutral selection, respectively. The software in SPSS version 19.0 (SPSS, Chicago, IL, USA) was used for statistical analysis. The statistical significance of *Ka/Ks* was defined based on Duncan's multiple range test and P value of < 0.05 as statistically significant.

3. Results

3.1. Identification and Comparative Analyses of bZIP Genes in Nineteen Species. The sequences of 1441 *bZIP* sequences in 19 genomes, ranging from fungi to Plantae, including the three Rosaceous species, were used to analyze the evolution of this gene family (Figure 1 and Table S1). In the genome assemblies of strawberry, apple, and peach, 50, 116, and 47 *bZIP* genes were identified, respectively, using the HMM profile from the Pfam database [39] (Table S2). The number of *bZIP* genes varies from 4 (*C. merolae*) to 212 (*P. trichocarpa*) in 19 species with the genome size from 12.2 Mb (*S. cerevisiae*) to 881.3 Mb (*M. domestica*). Furthermore, we found that the number of *bZIP* genes in six higher plant species was more than 100.

The total number of *bZIP* genes in strawberry and peach was very similar. However, it is important to note that the number of *bZIP* genes in these two species was much less than the number in apple (116). The number of *bZIP* genes in strawberry (50) and peach (47) was also much smaller



Species	Number of *bZIP*	Genome size (Mb)	Number (Mb)
Malus domestica	116	881.3	0.13/72.06
Prunus persica	47	227.3	0.21/122.59
Fragaria vesca	50	240	0.21/136.8
Cucumis sativus	118	203	0.58/105.87
Medicago truncatula	66	257.6	0.26/197.57
Populus trichocarpa	212	422.9	0.50/97.74
Arabidopsis thaliana	75	135	0.56/198.86
Brassica rapa	130	283.8	0.46/152.82
Carica papaya	45	135	0.33/202.46
Theobroma cacao	106	346	0.31/85.12
Vitis vinifera	55	487	0.11/61.54
Solanum lycopersicum	69	760	0.09/45.69
Oryza sativa	89	372	0.24/150.5
Musa acuminata	134	331.8	0.40/110.15
Selaginella moellendorffii	23	212.5	0.11/104.81
Physcomitrella patens	64	480	0.13/55.44
Chlamydomonas reinhardtii	24	111	0.22/159.83
Cyanidioschyzon merolae	4	16.5	0.24/406.12
Saccharomyces cerevisiae	14	12.2	1.15/551.31

Figure 1: Phylogenetic relationships, number of *bZIP* genes, genome size, bZIP density, and overall gene density of the nineteen species analyzed. The bZIP density and overall gene density of the nineteen species analyzed were separated by parenthesis. The bZIP density was followed by overall gene density.

than the number in most of the 19 species (Figure 1), which may infer that only a small number of gene duplication events contributed to the *bZIP* members in these two species. The number of apple *bZIP* genes (116) was similar to the number in *C. sativus* (118) but less than the number found in *P. trichocarpa* (212), *B. rapa* (130), and *M. acuminate* (134) (Figure 1). In contrast, the density of *bZIP* genes in the three Rosaceous species was very distinct and not related to the number of *bZIP* genes present. The *bZIP* density in the apple genome (0.13) was lower than that in strawberry (0.21) and peach (0.21) and only exceeded the density observed in *V. vinifera* (0.11), *S. moellendorffii* (0.11), and *S. lycopersicum* (0.09) (Figure 1). By contrast, the apple genome also had a lower overall gene density (72.06), which is probably the reason for low *bZIP* density in the apple genome.

3.2. Phylogenetic Analysis of bZIP Genes in Three Rosaceous Species. A phylogenetic analysis was performed for the *bZIP* genes in the three Rosaceous species using the bZIP domains in strawberry, apple, and peach, as well as *Arabidopsis* [5], in order to further elucidate the evolution of this gene family (Figure S1). Since the *bZIP* genes of *Arabidopsis* have already been clustered, we were able to compare the clustering of the *bZIP* genes of Rosaceous species with the clustering from *Arabidopsis*. Surprisingly, *AtbZIP31*, *AtbZIP33*, and *AtbZIP74* were different from other *bZIP* genes in that they formed individual clades containing only *bZIP* genes of *Arabidopsis*, suggesting that these individual clades may be specific to *Arabidopsis* (Figure S1).

The results indicated that the ten clades (A, B, C, D, E, F, G, H, I, and S) obtained in our phylogenetic tree were in agreement with the clustering and classification of *bZIP* genes in Arabidopsis [5] (Figure 2). However, a few genes formed three small unique clades (UC, Figure 2) in the phylogenetic tree produced from our analyses. This observation supports the hypothesis that these three unique clades may have had independent evolutionary trajectories from the other clades.

All of the clades from Figure 2 include genes from all of the three species. The number of strawberry, apple, and peach *bZIP* genes, respectively, in each of the clades were A (9, 18, 8); B (1, 5, 1); C (3, 6, 4); D (7, 12, 6); E (2, 8, 3); F (2, 6, 2); G (6, 10, 4); H (2, 5, 2); I (6, 17, 6); and S (9, 21, 9). Moreover, the phylogenetic tree of the three Rosaceous species indicated that the *bZIP* genes in strawberry and peach have few paralogs with "one-to-one" topology (two paralogs clustered together in a clade), suggesting that most of them were generated before speciation of strawberry. In contrast, there were many clades with "one-to-one" or "one-to-many" topologies (more than two paralogs clustered together in a clade) in apple, indicating that species-specific duplication events contributed greatly to the large number of apple *bZIPs*.

3.3. Nonsynonymous and Synonymous Substitution of bZIP Genes. Our result indicates that most clades (A, B, C, D, E, F,

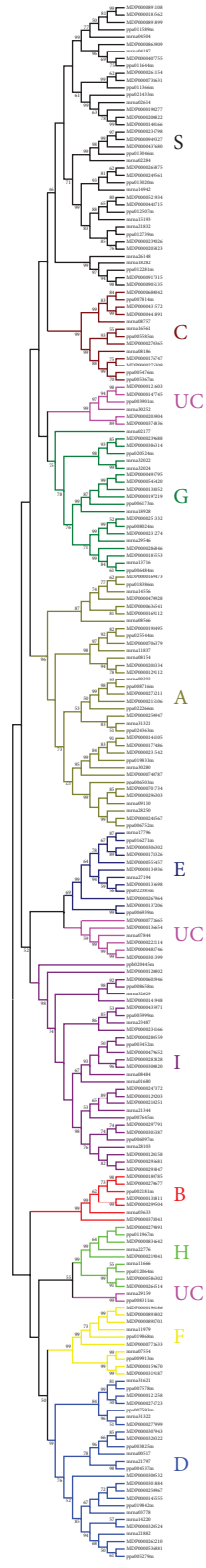


Figure 2: Phylogenetic analysis of *bZIP* members in strawberry, apple, and peach. Phylogenetic analysis of bZIP proteins in strawberry (mrna), apple (MDP), and peach (ppa). Only bootstrap values larger than 50% are indicated. Different colors can be used to distinguish the different subgroups. The names of each subgroup are listed on the right. UC represented "unique clades."

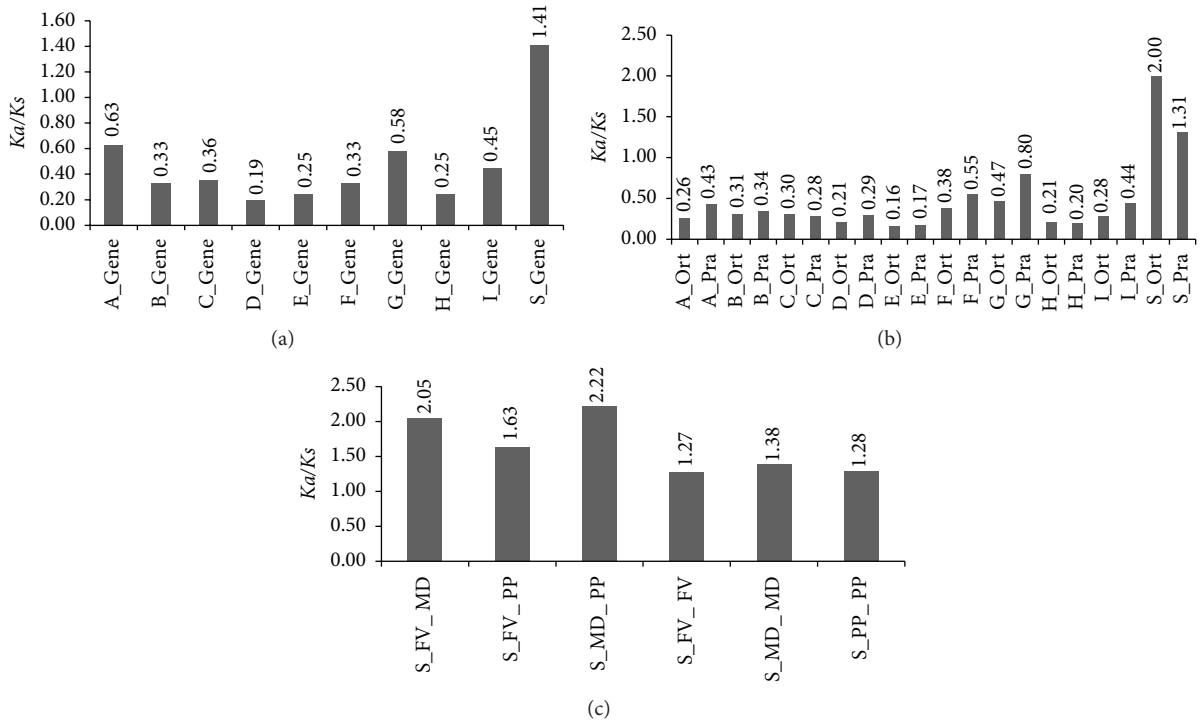


Figure 3: *Ka/Ks* ratios of *bZIP* genes. (a) *Ka/Ks* ratios of genes in clades A–S. (b) *Ka/Ks* ratios of paralogous and orthologous gene pairs in clades A–S. (c) *Ka/Ks* ratios of paralogs (FV_FV, MD_MD, and PP_PP) and orthologs (FV_MD, FV_PP, and MD_PP) in clade S. The *Ka/Ks* ratios are located in the top of the graph.

G, H, and I) had *Ka/Ks* ratios less than 1 (Figure 3(a)), demonstrating that most genes of those clades were undergoing a purifying selection in the three species. Among all the gene pairs in the clades, 25 (7.99% of clade A), 1 (2.13% of clade E), 16 (9.58% of clade G), and 12 (5.33% of clade I) pairs had *Ka/Ks* ratio approximately equal to 1 (*Ka/Ks* ratio = 0.8~1.0) for *bZIP* genes in strawberry, apple, and peach (Table S3). However, 52 (16.61% of clade A), 15 (8.98% of clade G), 1 (4.55% of clade H), and 15 (6.67% of clade I) gene pairs had *Ka/Ks* ratios greater than 1 for *bZIP* genes (Table S3), which indicates that some of *bZIP* genes were under positive selection or relaxed selection for gene pairs with *Ka/Ks* approximately equal to 1. It is worth noting that *Ka/Ks* ratio of gene pairs in clade S is significantly greater than other clades ($P < 0.05$), which illustrated that *bZIP* genes were under strongly positive selection (Figure 3(a)).

In order to explain *Ka/Ks* ratio distribution of gene pairs in each clade, we compared *Ka/Ks* ratio of the orthologous and paralogous gene pairs in strawberry, apple, and peach (Table S4). It is indicated that the *Ka/Ks* ratio of paralogs is bigger than orthologs in each clade except for clades C, H, and S (Figure 3(b)). Most of orthologs and paralogs exhibit a low level *Ka/Ks* ratio (*Ka/Ks* ratio = 0.16~0.80) in different clades (A, B, C, D, E, F, G, H, and I) analyzed (Figure 3(b)). However, the ones of orthologs (*Ka/Ks* ratio = 2.00) and paralogs (*Ka/Ks* ratio = 1.31) in clade S are obviously greater than 1 and significantly higher than orthologs and paralogs in other clades ($P < 0.05$). Orthologs and paralogs in clade S could be further divided into three subgroups separately, FV_PP (between strawberry and peach)/MD_FV (between apple and strawberry)/MD_PP (apple and peach) and FV_FV (within strawberry)/MD_MD (within apple)/PP_PP (within peach). Orthologs in the MD_PP have a highest *Ka/Ks* ratio (2.22) and paralogs in the FV_FV have a lowest *Ka/Ks* ratio (1.27) (Figure 3(c)).

3.4. Nonsynonymous and Synonymous Substitution of bZIP Domains. For getting a more in-depth exploration in selection pressure of *bZIP* genes in different clades during their evolution, we compared the *Ka/Ks* ratio of bZIP domains in each clade (Table S5). We found that all clades with *Ka/Ks* ratios ranging from 0.04 (clade D) to 0.32 (clade G) were less than 0.4 (Figure 4(a)). It is suggested that a strong negative selection plays the leading roles in the evolution of bZIP domains.

Basic leucine zipper (bZIP) proteins, one of the largest families of transcription factors in plants, are characterized by a basic region (BR) responsible for sequence-specific DNA binding, an adjacent heptad leucine repeat, and the leucine zipper (LZ) [40]. It is concluded that all BR domains ranging from 0.02 (BR of clade C) to 0.24 (BR of clade I) and LZ domains ranging from 0.1 (LZ of clade I) to 0.61 (LZ of clade B) were undergoing negative selection (Figure 4(b),Table S5). Interestingly, *Ka/Ks* ratio of BR domain is less than the ones of LZ domain in each clade except for clades H and I (Figure 4(b)).

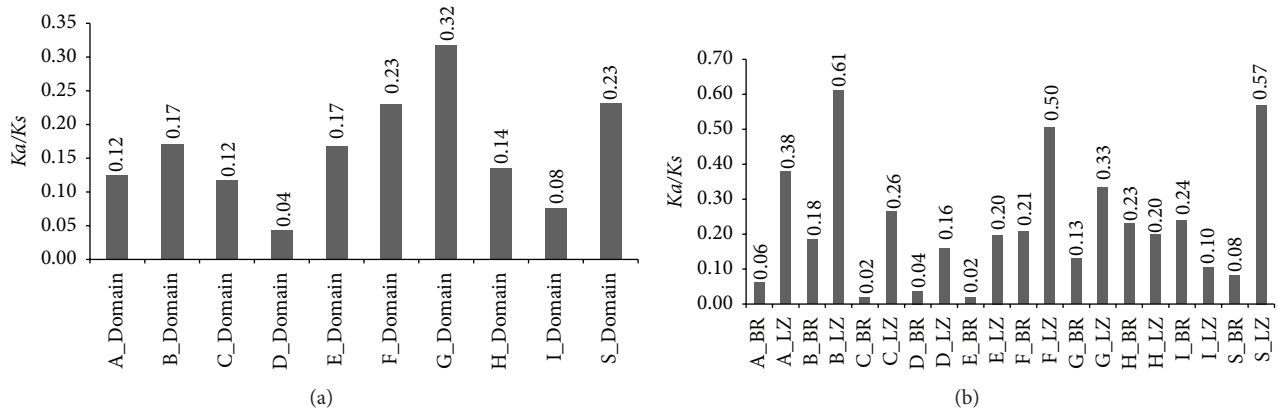


Figure 4: *Ka/Ks* ratios of *bZIP* domains. (a) *Ka/Ks* ratios of domains in clades A–S. (b) *Ka/Ks* ratios of BR and LZ domains in clades A–S. The *Ka/Ks* ratios are located in the top of the graph.

3.5. Evaluation of Orthologous bZIP Genes between Strawberry, Apple, and Peach. In order to trace the evolutionary history of *bZIP* genes among the three Rosaceous species, orthologous regions of *bZIP* genes in the three Rosaceous species were subjected to a comparative analysis in order to ascertain the evolutionary history of *bZIP* genes in the Rosaceae. Using Circos software [41], 57 orthologous gene pairs were identified between strawberry and apple (FV_MD) (Figure 5(a)), 64 between apple and peach (MD_PP) (Figure 5(b)), and 50 between strawberry and peach (FV_PP) (Figure 5(c)). Collectively, these data are presented in Table S6 and Figure 5.

Out of the 57 gene pairs present in the strawberry and apple genomes (Figure 5(a)), 20 strawberry *bZIP* genes correspond to one copy (Type 1), 17 genes correspond to two copies (Type 2), and one gene corresponds to three copies (Type 3) in apple. Therefore, 56 *bZIP* genes in the apple genome have 38 corresponding genes in the strawberry genome. In all three types, some genes have preserved and exhibit the same number of exons (Table S6). Out of 50 gene pairs present in the strawberry and peach genomes (Figure 5(c)), 26 strawberry *bZIP* genes correspond to one copy (Type 1), 9 genes to two copies (Type 2), and 2 genes to three copies (Type 3) in peach. Collectively, 37 strawberry *bZIP* genes corresponded to 38 *bZIP* genes in the peach genome. Genes of all three types in strawberry and peach have preserved similar exon configurations (Table S6). Based on the 30 overlapping *bZIP* strawberry genes, the data collectively indicate that 45 *bZIP* genes, representing 90% of the total number of *bZIP* genes in the strawberry genome, were ancestral and underwent different duplication events after the divergent speciation of apple and peach. Additionally, 56 *bZIP* genes, representing 48.3% of the total number of *bZIP* genes in the apple genome, were retained on duplicated regions. In addition, 38 *bZIP* genes, representing 80.9% of the total number of *bZIP* genes in the peach genome, were retained on syntenic blocks. These data further indicate that most of the *bZIP* genes in strawberry and peach experienced a low level of duplication events compared to the number of duplication events in the apple genome. These findings are consistent with the results of a previous study which reported that a recent whole genome duplication (WGD) event occurred in apple 60–65 million years ago [27].

3.6. Orthologous Relationships among Chromosomes. In order to understand the influence of the WGD in apple on the *bZIP* gene family in the Rosaceae, the major distribution of orthologous chromosomes was identified and compared between paired combinations of strawberry, apple, and peach according to the classification reported by Jung et al. [42] (Table S6, Table S7). The orthologous relationship between chromosomes of peach and strawberry made it evident that the majority of *bZIP* genes on peach chromosomes PC2, PC3, PC5, and PC8 were located on a single homologous FC7, FC6, FC5, and FC2 chromosome in strawberry, respectively. The majority of genes on PC6 and PC7 were also located on strawberry chromosomes, FC1 and FC6. Additionally, 35.71% of the *bZIP* genes on strawberry chromosome FC2 had an orthologous relationship to the PC1. Both ppa016271m and ppa022385m of PC4, however, were located on the FC6 chromosome of strawberry.

The relationship between peach and apple at the chromosome level was more complex than the relationship between peach and strawberry. 66.67%, 66.67%, 50%, 50%, and 50% of *bZIP* genes on five apple chromosomes sets, MC2/MC7, MC9/MC17, MC3/MC11, MC14/MC6, and MC2/MC15, respectively, have their orthologous genes corresponding to the chromosomes PC2, PC3, PC4, PC5, and PC7 of peach. Orthologous genes on PC6 corresponded to major genes on four apple chromosomes, MC3, MC4, MC11, and MC12 (Table S6, Table S7).

4. Discussion

4.1. Evolutionary History of bZIP Family in Three Species of the Rosaceae. The bZIP transcription factor family is one of the largest and most diverse families of transcriptional regulators in eukaryotic organisms [15]. In the present study, the bZIP transcription factor family in 16 species, including 13 higher plants, 2 lower plants, and one fungus, was analyzed, in an

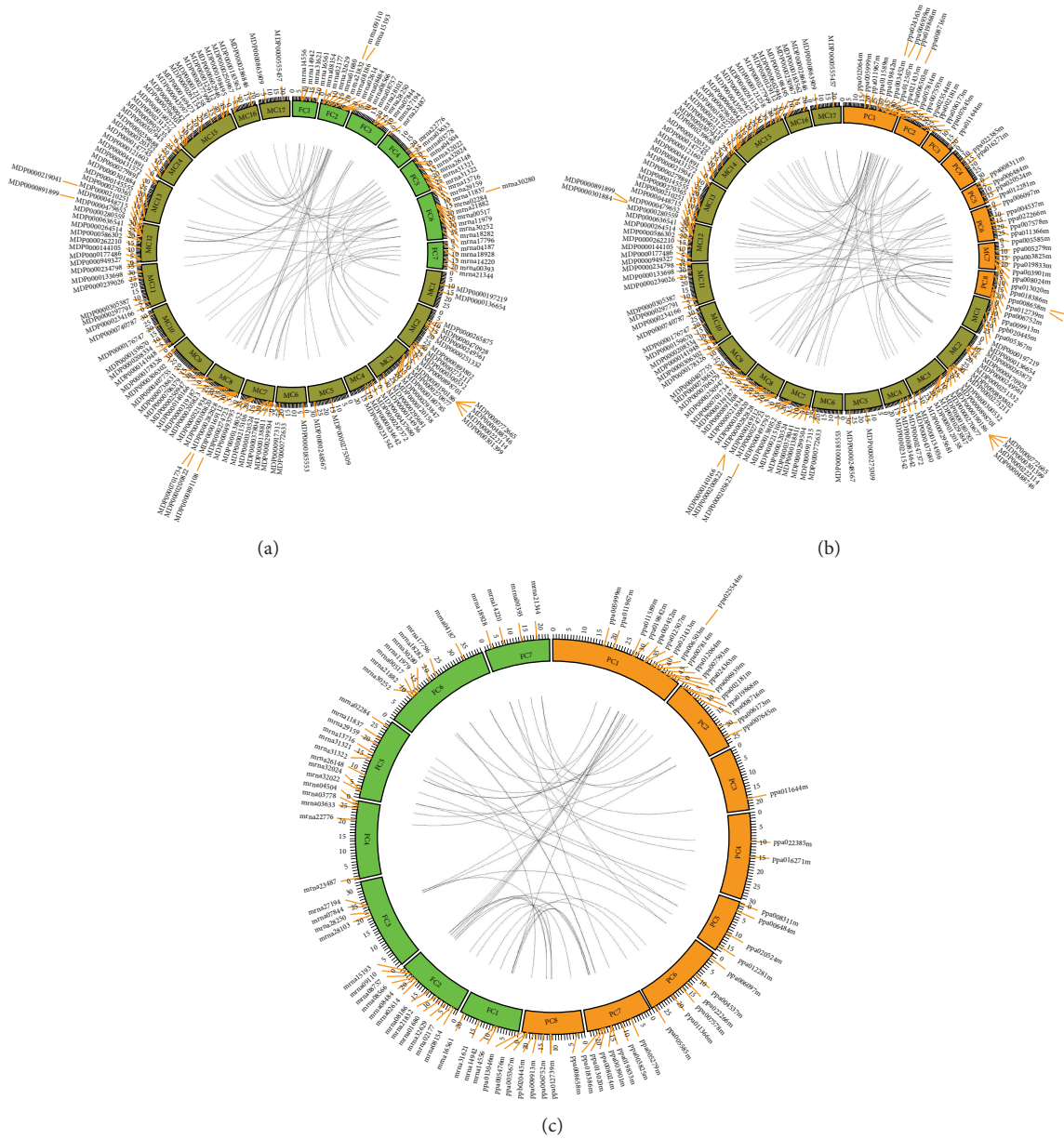


FIGURE 5: Evaluation of orthologous *bZIP* genes between strawberry, apple, and peach. (a) Seven strawberry (FC1 to FC7) and seventeen apple chromosome (MC1 to MC17) maps are based on the orthologous pair position and demonstrate a highly conserved syntenic relationship. (b) Seventeen apple chromosome (MC1 to MC17) and peach (PC1 to PC8) maps are based on the orthologous pair positions and demonstrate highly conserved synteny. (c) Seven strawberry (FC1 to FC7) and peach (PC1 to PC8) maps are based on the orthologous pair positions, and demonstrate highly conserved synteny.

effort to better understand the evolution of this gene family in the Rosaceae. It has been suggested that the *bZIP* gene family existed before the divergence of higher and lower plant species, even in the fungi, which consists with foundation in Wang et al. (2011) research [17]. An uneven distribution of *bZIP* copies among the 19 species was identified, suggesting that the *bZIP* genes within each species had undergone different levels of gene duplication with larger expansion after the divergence of higher and lower plants. For example, the numbers of copies of *bZIP* genes were as follows: *O. sativa* (89), *Cucumis sativus* (118), and *Populus trichocarpa* (212). These observations suggest that specific functional expansion may have resulted from environmental selection pressure or specialization in processes of growth and development, including stress responses [14, 43–46] and abscisic acid (ABA) signaling [10, 11, 47]. As a result of evolutionary pressure and/or environmental selection, critical genes or components of genes were retained, whereas others were deleted or lost [48].

We identified 50 and 47 *bZIP* genes in the genomes of strawberry and peach, respectively. This number is similar with those of previous genome-wide studies on some other

species, indicating the presence of 64 *bZIP* homologs in cucumber [18], 55 in grapevine [19], and 49 in castor bean [15]. The *bZIP* homologs in apple (116) were consistent with the numbers in maize (120) [16] and sorghum (92) [17]. *bZIP* genes in strawberry and peach are much lower than that in apple which has a much larger genome size. These observations support the hypothesis that the WGD [27] event which occurred in apple resulted in significant amplification in the number of apple *bZIP* genes. On the other hand, a low level of gene duplication events may have contributed to the number of *bZIP* genes in strawberry and peach.

4.2. bZIP Genes Expansion in the Rosaceae. The phylogenetic tree of the *bZIP* gene family generated in this study for Rosaceous species is supported by Liu et al. [19], Nijhawan et al. [6], and Wei et al. [16]. Each of the clades included at least 7 *bZIP* genes from the 39 *bZIP* genes identified in the 3 species examined, indicating that many of the *bZIP* genes originated through a process of gene duplication. The widespread existence of paralogs and orthologs with "one-to-one" or "one-to-many" topologies in the Rosaceous species examined suggests that species-specific duplication was the main contributor to the large number of *bZIPs* observed in apple. The number of *bZIP* genes in each of the three species was highly variable, indicating that most of the gene duplication events occurred after evolutionary divergence of each lineage. It is also likely that both WGD and a series of rearrangements occurred during the evolution of certain species.

Extensive genome and EST sequencing of plant species has revealed a substantial history of WGD events [49, 50]. In the Rosaceae, an evolutionary trend toward fruit development and specialization may have been partially based on gene duplication. For example, WGD in apple has resulted in the creation of large families of paralogous genes [27]. In our analysis of the apple genome, 56 (48.3%) *bZIP* genes were retained on duplicated regions. Therefore, the involvement of WGD in the expansion of the *bZIP* gene family in apple is quite evident. Polyploidy provides an excellent genomic resource to study retention and loss of multicopy genes [48, 51]. Following WGD, genes can suffer a variety of fates ranging from massive gene loss to the development of a central role in an essential aspect of the plant [52]. A comparative analysis of *bZIP* genes in strawberry, apple, and peach led us to hypothesize that, after WGD in an apple ancestor, orthologous *bZIP* genes corresponding to strawberry on duplicated regions in apple genome were retained. On the other hand, in the peach ancestor, these syntenic regions were quickly lost or deleted, perhaps due to issues associated with an imbalance in gene dosage [53, 54].

4.3. Selection Pressure of bZIP Genes and bZIP Domains in All Clades. Furthermore, *Ka/Ks* ratios were estimated to detect the diversifying selection pressure on different clades (except for UN clades). The results showed that the *Ka/Ks* ratios for gene pairs in nine clades (A, B, C, D, E, F, G, H, and I) were <1, with most of them being even less than 0.6, suggesting strong purifying selection (Figure 3(a)). However, the other pairs in clade S seemed to be under positive selection, as their *Ka/Ks* ratios were >1. Also, in the phylogenic tree of Rosaceae, we found the biggest clade (S) containing 39 genes (21, 9, and 9 for apple, peach, and strawberry, resp.). Much interest focuses on positive selection (adaptive molecular evolution) associated with adaptation and evolution of new forms or functions in that nonsynonymous mutations offer fitness advantages to the protein [55, 56]. Zhao et al. have concluded that functional gain and divergence of transcription factors were driven by distinct positive selection on their transcription activation domains [57]. Based on the derivative data from monocot and dicot species imply that homologues of S bZIPs are also transcriptionally activated after stress treatment [58], such as drought, cold, and wounding, or are specifically expressed in defined parts of the flower [59, 60]. The positive selection may have contributed to the expansion of clade S to adapt to the development and environment stresses.

The bZIP transcription factors contain a highly conserved bZIP domain composed of two structural features: a basic region (N-X_7-R/K-X_9) for sequence-specific DNA binding and a leucine zipper composed of several heptad repeats of Leu or other bulky hydrophobic amino acids, such as Ile, Val, Phe, or Met, for dimerization specificity [5–7]. Additionally, *bZIP* domains of all clades also appeared as stronger purifying selection. A purifying selection may aid in the detection of regions or residues of functional importance [55]. These results suggested that functions of genes in major clades did not diverge much along with the genome evolution after the duplication events. Possibly because of the rapid evolution, members of the Rosaceae display remarkable phenotypic diversity, with common morphological synapomorphies not readily identifiable [23]. It is worth noting that paralogs were undergoing stronger purifying selection than orthologs in each clade except for clades C, H, and S (Figure 3(b)), which probably accelerates the process of morphological diversity, plant habit, and fruit type within the Rosaceae. From Figure 4(b), we conclude that BR domains were under stronger purifying selection than LZ domains in each clade except for clades H and I, suggesting that purifying selection was mainly responsible for *bZIP* sequence-specific DNA binding.

4.4. Orthologous Pairs between Chromosomes. Peach, at both the macro- and microsyntenic levels, has the most conserved karyotype in relation to the ancestral genome configuration for the Rosaceae [42]. Dirlewanger et al. [61] compared *Malus* and *Prunus* and found strong evidence that single linkage groups in the diploid *Prunus* were homologous to two distinct homologous linkage groups in the amphitetraploid genome of *Malus*. According to orthologous *bZIP* gene pairs analysis, the conserved and syntenic blocks were common to all three genomes analyzed, with a single syntenic block in peach corresponding to one or two syntenic regions in strawberry and two or four syntenic regions in apple. Vilanova et al. [62] compared the diploid reference linkage maps for *Prunus* and *Fragaria* and they identified numerous chromosomal translocations and rearrangements that occurred in the 29 million years since the genera diverged from a common ancestor. Notably, *bZIP* genes on the PC4 peach chromosome corresponded orthologously not to FC6, but rather to FC3. The data indicated that two genes (*ppa016271m* and *ppa022385m*)

located on a nonorthologous chromosome region that had originated from a common ancestor went through some intrachromosomal rearrangements. This interpretation is consistent with the fact that a greater number of small-scale rearrangements occurred in strawberry in comparison to either apple or peach [42]. Whilst an early hypothesis as to the origin of *Malus* implied wide hybridization between ancestral amygdaloid ($x = 8$) and ancestral spiraeoid ($x = 9$) [63], other data suggest that *Malus* may have arisen due to polyploidization of a spiraeoid species [64]. Illa et al. [23] reconstructed a hypothetical ancestral genome for the Rosaceae containing nine chromosomes ($x = 9$), consistent with the report of Vilanova et al. [62]. Based on the analysis of orthologous pairs between chromosomes, we could propose a hypothesis that these orthologs became after one gene duplication located on one of the nine ancient chromosomes in the Rosaceae. An evaluation of the conservation of synteny between *Fragaria*, *Malus*, and *Prunus* based on whole genome sequence data may reveal much about sequence evolution in this closely related family.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Xiao-Long Wang and Yan Zhong contributed equally to this work and should be considered co-first authors.

Acknowledgments

This research was financially supported in part by Natural Science Foundation of China (Grant no. 31401849) and the Priority Academic of Jiangsu Province.

References

[1] M. P. Scott, "Development: the natural history of genes," *Cell*, vol. 100, no. 1, pp. 27–40, 2000.

[2] S. B. Carroll, "Endless forms: the evolution of gene regulation and morphological diversity," *Cell*, vol. 101, no. 6, pp. 577–580, 2000.

[3] P. Pérez-Rodríguez, D. M. Riaño-Pachón, L. G. G. Corrêa, S. A. Rensing, B. Kersten, and B. Mueller-Roeber, "PlnTFDB: updated content and new features of the plant transcription factor database," *Nucleic Acids Research*, vol. 38, supplement 1, Article ID gkp805, pp. D822–D827, 2009.

[4] H. C. Hurst, "Transcription factors. 1: bZIP proteins," *Protein Profile*, vol. 1, no. 2, pp. 123–168, 1993.

[5] M. Jakoby, B. Weisshaar, W. Dröge-Laser et al., "bZIP transcription factors in *Arabidopsis*," *Trends in Plant Science*, vol. 7, no. 3, pp. 106–111, 2002.

[6] A. Nijhawan, M. Jain, A. K. Tyagi, and J. P. Khurana, "Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice," *Plant Physiology*, vol. 146, no. 2, pp. 333–350, 2008.

[7] S. C. Lee, H. W. Choi, I. S. Hwang, D. S. Choi, and B. K. Hwang, "Functional roles of the pepper pathogen-induced bZIP transcription factor, CAbZIP1, in enhanced resistance to pathogen infection and environmental stresses," *Planta*, vol. 224, no. 5, pp. 1209–1225, 2006.

[8] O. Yang, O. V. Popova, U. Süthoff, I. Lüking, K.-J. Dietz, and D. Golldack, "The *Arabidopsis* basic leucine zipper transcription factor AtbZIP24 regulates complex transcriptional networks involved in abiotic stress resistance," *Gene*, vol. 436, no. 1-2, pp. 45–55, 2009.

[9] F. Weltmeier, F. Rahmani, A. Ehlert et al., "Expression patterns within the *Arabidopsis* C/S1 bZIP transcription factor network: availability of heterodimerization partners controls gene expression during stress response and development," *Plant Molecular Biology*, vol. 69, no. 1-2, pp. 107–119, 2009.

[10] Y. Xiang, N. Tang, H. Du, H. Ye, and L. Xiong, "Characterization of OsbZIP23 as a key player of the basic leucine zipper transcription factor family for conferring abscisic acid sensitivity and salinity and drought tolerance in rice," *Plant Physiology*, vol. 148, no. 4, pp. 1938–1952, 2008.

[11] G. Lu, C. Gao, X. Zheng, and B. Han, "Identification of OsbZIP72 as a positive regulator of ABA response and drought tolerance in rice," *Planta*, vol. 229, no. 3, pp. 605–615, 2009.

[12] H. Shimizu, K. Sato, T. Berberich et al., "LIP19, a basic region leucine zipper protein, is a Fos-like molecular switch in the cold signaling of rice plants," *Plant and Cell Physiology*, vol. 46, no. 10, pp. 1623–1634, 2005.

[13] M. Zou, Y. Guan, H. Ren, F. Zhang, and F. Chen, "A bZIP transcription factor, OsABI5, is involved in rice fertility and stress tolerance," *Plant Molecular Biology*, vol. 66, no. 6, pp. 675–683, 2008.

[14] M. A. Hossain, J.-I. Cho, M. Han et al., "The ABRE-binding bZIP transcription factor OsABF2 is a positive regulator of abiotic stress and ABA signaling in rice," *Journal of Plant Physiology*, vol. 167, no. 17, pp. 1512–1520, 2010.

[15] Z. Jin, W. Xu, and A. Liu, "Genomic surveys and expression analysis of bZIP gene family in castor bean (*Ricinus communis* L.)," *Planta*, vol. 239, no. 2, pp. 299–312, 2014.

[16] K. Wei, J. Chen, Y. Wang et al., "Genome-wide analysis of bZIP-encoding genes in maize," *DNA Research*, vol. 19, no. 6, pp. 463–476, 2012.

[17] J. Wang, J. Zhou, B. Zhang, J. Vanitha, S. Ramachandran, and S.-Y. Jiang, "Genome-wide expansion and expression divergence of the basic leucine zipper transcription factors in higher plants with an emphasis on sorghum," *Journal of Integrative Plant Biology*, vol. 53, no. 3, pp. 212–231, 2011.

[18] M. C. Baloglu, V. Eldem, M. Hajyzadeh, and T. Unver, "Genome-wide analysis of the bZIP transcription factors in cucumber," *PLoS ONE*, vol. 9, no. 4, Article ID e96014, 2014.

[19] J. Liu, N. Chen, F. Chen et al., "Genome-wide analysis and expression profile of the bZIP transcription factor gene family in grapevine (*Vitis vinifera*)," *BMC Genomics*, vol. 15, no. 1, article 281, 2014.

[20] T. Wicker, N. Yahiaoui, and B. Keller, "Illegitimate recombination is a major evolutionary mechanism for initiating size variation in plant resistance genes," *The Plant Journal*, vol. 51, no. 4, pp. 631–641, 2007.

[21] E. Dirlewanger, P. Cosson, M. Tavaud et al., "Development of microsatellite markers in peach [*Prunus persica* (L.) Batsch] and their use in genetic diversity analysis in peach and sweet cherry (*Prunus avium* L.)," *Theoretical and Applied Genetics*, vol. 105, no. 1, pp. 127–138, 2002.

[22] D. Potter, T. Eriksson, R. C. Evans et al., "Phylogeny and classification of Rosaceae," *Plant Systematics and Evolution*, vol. 266, no. 1-2, pp. 5–43, 2007.

[23] E. Illa, D. J. Sargent, E. L. Girona et al., "Comparative analysis of rosaceous genomes and the reconstruction of a putative ancestral genome for the family," *BMC Evolutionary Biology*, vol. 11, article 9, 2011.

[24] D. R. Morgan, D. E. Soltis, and K. R. Robertson, "Systematic and evolutionary implications of RBCL sequence variation in Rosaceae," *American Journal of Botany*, vol. 81, no. 7, pp. 890–903, 1994.

[25] D. Potter, F. Gao, P. E. Bortiri, S.-H. Oh, and S. Baggett, "Phylogenetic relationships in Rosaceae inferred from chloroplast matK and trnL-trnF nucleotide sequence data," *Plant Systematics and Evolution*, vol. 231, no. 1–4, pp. 77–89, 2002.

[26] V. Shulaev, D. J. Sargent, R. N. Crowhurst et al., "The genome of woodland strawberry (*Fragaria vesca*)," *Nature Genetics*, vol. 43, no. 2, pp. 109–116, 2011.

[27] R. Velasco, A. Zharkikh, J. Affourtit et al., "The genome of the domesticated apple (*Malus*×*domestica* Borkh.)," *Nature Genetics*, vol. 42, no. 10, pp. 833–839, 2010.

[28] I. Verde, A. G. Abbott, S. Scalabrin et al., "The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution," *Nature Genetics*, vol. 45, no. 5, pp. 487–494, 2013.

[29] D. M. Goodstein, S. Shu, R. Howson et al., "Phytozome: a comparative platform for green plant genomics," *Nucleic Acids Research*, vol. 40, no. 1, pp. D1178–D1186, 2012.

[30] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D290–D301, 2012.

[31] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.

[32] I. Letunic, T. Doerks, and P. Bork, "SMART 7: recent updates to the protein domain annotation resource," *Nucleic Acids Research*, vol. 40, no. 1, pp. D302–D305, 2012.

[33] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins, "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools," *Nucleic Acids Research*, vol. 25, no. 24, pp. 4876–4882, 1997.

[34] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[35] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[36] T.-H. Lee, H. Tang, X. Wang, and A. H. Paterson, "PGDD: a database of gene and genome duplication in plants," *Nucleic Acids Research*, vol. 41, no. 1, pp. D1152–D1158, 2013.

[37] M. A. Larkin, G. Blackshields, N. P. Brown et al., "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[38] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[39] R. D. Finn, A. Bateman, J. Clements et al., "Pfam: the protein families database," *Nucleic Acids Research*, vol. 42, no. 1, pp. D222–D230, 2013.

[40] W. H. Landschulz, P. F. Johnson, and S. L. McKnight, "The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins," *Science*, vol. 240, no. 4860, pp. 1759–1764, 1988.

[41] M. Krzywinski, J. Schein, I. Birol et al., "Circos: an information aesthetic for comparative genomics," *Genome Research*, vol. 19, no. 9, pp. 1639–1645, 2009.

[42] S. Jung, A. Cestaro, M. Troggio et al., "Whole genome comparisons of *Fragaria*, *Prunus* and *Malus* reveal different modes of evolution between Rosaceous subfamilies," *BMC Genomics*, vol. 13, article 129, 2012.

[43] M. A. Hossain, Y. Lee, J.-I. Cho et al., "The bZIP transcription factor OsABF1 is an ABA responsive element binding factor that enhances abiotic stress signaling in rice," *Plant Molecular Biology*, vol. 72, no. 4, pp. 557–566, 2010.

[44] J.-Y. Kang, H.-I. Choi, M.-Y. Im, and Y. K. Soo, "Arabidopsis basic leucine zipper proteins that mediate stress-responsive abscisic acid signaling," *The Plant Cell*, vol. 14, no. 2, pp. 343–357, 2002.

[45] S.-J. Oh, I. S. Sang, S. K. Youn et al., "*Arabidopsis* CBF3/DREB1A and ABF3 in transgenic rice increased tolerance to abiotic stress without stunting growth," *Plant Physiology*, vol. 138, no. 1, pp. 341–351, 2005.

[46] H. Tak and M. Mhatre, "Cloning and molecular characterization of a putative bZIP transcription factor VvbZIP23 from Vitis vinifera," *Protoplasma*, vol. 250, no. 1, pp. 333–345, 2013.

[47] Y. Uno, T. Furihata, H. Abe, R. Yoshida, K. Shinozaki, and K. Yamaguchi-Shinozaki, "Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 21, pp. 11632–11637, 2000.

[48] J. Yu, S. Tehrim, F. Zhang et al., "Genome-wide comparative analysis of NBS-encoding genes between *Brassica* species and *Arabidopsis thaliana*," *BMC Genomics*, vol. 15, no. 1, article 3, 2014.

[49] Y. Jiao, N. J. Wickett, S. Ayyampalayam et al., "Ancestral polyploidy in seed plants and angiosperms," *Nature*, vol. 473, no. 7345, pp. 97–100, 2011.

[50] S. Proost, P. Pattyn, T. Gerats, and Y. Van De Peer, "Journey through the past: 150 million years of plant genome evolution," *Plant Journal*, vol. 66, no. 1, pp. 58–65, 2011.

[51] F. Cheng, J. Wu, and X. Wang, "Genome triplication drove the diversification of *Brassica* plants," *Horticulture Research*, vol. 1, Article ID 14024, 2014.

[52] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae," *Nature*, vol. 428, no. 6983, pp. 617–624, 2004.

[53] M. Freeling, "The evolutionary position of subfunctionalization, downgraded," *Genome Dynamics*, vol. 4, pp. 25–40, 2008.

[54] M. Freeling, "Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition," *Annual Review of Plant Biology*, vol. 60, pp. 433–453, 2009.

[55] J. Yang, Z. L. Wang, X. Q. Zhao et al., "Natural selection and adaptive evolution of leptin in the *Ochotona* family driven by the cold environmental stress," *PLoS ONE*, vol. 3, no. 1, Article ID e1472, 2008.

[56] Z. Yang, R. Nielsen, N. Goldman, and A.-M. K. Pedersen, "Codon-substitution models for heterogeneous selection pressure at amino acid sites," *Genetics*, vol. 155, no. 1, pp. 431–449, 2000.

[57] X. Zhao, Q. Yu, L. Huang, Q. Liu, and A. Asakura, "Patterns of positive selection of the myogenic regulatory factor gene family in vertebrates," *PLoS ONE*, vol. 9, no. 3, Article ID e92873, 2014.

[58] T. Kusano, T. Berberich, M. Harada, N. Suzuki, and K. Sugawara, "A maize DNA-binding factor with a bZIP motif is induced by low temperature," *Molecular & General Genetics*, vol. 248, no. 5, pp. 507–517, 1995.

[59] J. F. Martinez-García, E. Moyano, M. J. C. Alcocer, and C. Martin, "Two bZIP proteins from Antirrhinum flowers preferentially bind a hybrid C-box/G-box motif and help to define a new sub-family of bZIP transcription factors," *Plant Journal*, vol. 13, no. 4, pp. 489–505, 1998.

[60] A. Strathmann, M. Kuhlmann, T. Heinekamp, and W. Dröge-Laser, "BZI-1 specifically heterodimerises with the tobacco bZIP transcription factors BZI-2, BZI-3/TBZF and BZI-4, and is functionally involved in flower development," *The Plant Journal*, vol. 28, no. 4, pp. 397–408, 2001.

[61] E. Dirlewanger, E. Graziano, T. Joobeur et al., "Comparative mapping and marker-assisted selection in Rosaceae fruit crops," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, pp. 9891–9896, 2004.

[62] S. Vilanova, D. J. Sargent, P. Arús, and A. Monfort, "Synteny conservation between two distantly-related Rosaceae genomes: *Prunus* (the stone fruits) and *Fragaria* (the strawberry)," *BMC Plant Biology*, vol. 8, no. 1, article 67, 2008.

[63] K. Sax, "The origin of the Pomoideae," *Proceedings of the American Society for Horticultural Science*, vol. 30, pp. 147–150, 1933.

[64] C. Sterling, "Comparative morphology of the carpel in the Rosaceae. I. Prunoideae: Prunus," *American Journal of Botany*, vol. 51, no. 1, pp. 36–44, 1964.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 679548, 7 pages
http://dx.doi.org/10.1155/2015/679548

Research Article

Expressed Sequence Tags Analysis and Design of Simple Sequence Repeats Markers from a Full-Length cDNA Library in *Perilla frutescens* (L.)

Eun Soo Seong,[1] Ji Hye Yoo,[2] Jae Hoo Choi,[2] Chang Heum Kim,[2] Mi Ran Jeon,[2] Byeong Ju Kang,[2] Jae Geun Lee,[3] Seon Kang Choi,[4] Bimal Kumar Ghimire,[5] and Chang Yeon Yu[2]

[1]*Bioherb Research Institute, Kangwon National University, Chuncheon 200-701, Republic of Korea*
[2]*Department of Bioconvergence Science and Technology, College of Agriculture and Life Sciences, Kangwon National University, Chuncheon 200-701, Republic of Korea*
[3]*Hwajin Cosmetics, Hongcheon 250-807, Republic of Korea*
[4]*Department of Agricultural Life Sciences, Kangwon National University, Chuncheon 200-701, Republic of Korea*
[5]*Department of Applied Bioscience, Konkuk University, Seoul 143-701, Republic of Korea*

Correspondence should be addressed to Chang Yeon Yu; cyyu@kangwon.ac.kr

Received 27 July 2015; Accepted 21 October 2015

Academic Editor: Xiaohan Yang

Copyright © 2015 Eun Soo Seong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Perilla frutescens is valuable as a medicinal plant as well as a natural medicine and functional food. However, comparative genomics analyses of *P. frutescens* are limited due to a lack of gene annotations and characterization. A full-length cDNA library from *P. frutescens* leaves was constructed to identify functional gene clusters and probable EST-SSR markers via analysis of 1,056 expressed sequence tags. Unigene assembly was performed using basic local alignment search tool (BLAST) homology searches and annotated Gene Ontology (GO). A total of 18 simple sequence repeats (SSRs) were designed as primer pairs. This study is the first to report comparative genomics and EST-SSR markers from *P. frutescens* will help gene discovery and provide an important source for functional genomics and molecular genetic research in this interesting medicinal plant.

1. Introduction

Perilla frutescens (L.) is a self-compatible annual herb known as the beefsteak mint plant. It is cultivated in East Asian countries, including Japan, China, and Korea, and is an economical crop in the medicinal herb family, Lamiaceae [1]. Its seeds can be processed into foods and nutritional edible oils, and its leaves can be utilized as a traditional medicinal herb or flavor for vegetables [2, 3]. Perilla oil contains abundant polyunsaturated fatty acids (PUFAs), including linolenic (56.8%) and linoleic (17.6%) acids, which are used in salad oils or cooking [4, 5]. The flavor and odor of perilla are caused by the essential oils of monoterpenoids and sesquiterpenoids, including terpenoids, and they are commercially used as a natural fragrance or for flavoring [6]. The perilla leaf is composed of a number of chemical variants of the volatile essential oil classified as PA-type (mainly in perillaldehyde), EK-type (elsholtziaketone), PK-type (perilla ketone), PL-type (perillene), PP-type (phenylpropanoids), and PT-type (piperitenone) [7]. Perilla has been described as an important pharmaceutical with anti-inflammatory, anti-allergic, and broad antioxidant functions [8, 9].

Expressed sequence tags (ESTs) are fragments of expressed genes occurring from single-pass sequencing of cDNA libraries [10]. EST databases are sources of SSRs that can be developed as ortholog-specific EST-SSR markers and are dependent on genotype applications in many plant species [11–18]. As a molecular tool, EST-SSRs are highly important for studies on genetic populations [19]. They can identify functional markers in the open reading frames (ORFs) or

5′- or 3′-untranslated regions (UTRs) as well as exerting a phenotypic effect [20]. One advantage of the EST-SSR is that it is more transferable across closely related genera compared with unknown SSRs in the UTRs or noncoding sequences. Therefore, EST-SSRs are easy to understand for studying polymorphisms and genetic diversity [21, 22]. EST-derived SSRs have been reported in various plant species, including *Arabidopsis thaliana*, cacao, and sugarcane [23–25]. EST-SSRs also provide a new source for genetic and evolutionary studies based on homology searches of putative SSR functions [26].

In this study, we developed a full-length enriched cDNA library from *P. frutescens* leaves. EST sequence analysis allowed for genome annotation and gene ontologies and the identification of EST-SSR markers for genomic tool development in this less-well-studied medicinal plant species. These results provide useful and multipurpose data for further studies on *P. frutescens*.

2. Materials and Methods

2.1. Plant Materials. Seeds of *P. frutescens* were obtained after harvest in attached farm of Kangwon National University (Republic of Korea) during each year of collected accessions and grown on pot supplemented with commercial soil (GFC, Hongseong, Republic of Korea) in a greenhouse for a photoperiod of 16-hour light/8-hour dark at 25°C under well-water conditions. Leaves were sampled for RNA isolation.

2.2. RNA Extraction, cDNA Library Construction, Plasmid DNA Extraction, and Sequencing. Leaves were ground with a pestle in the presence of liquid nitrogen and ground tissue was used to RNA isolation using the Trizol method [27]. Total RNA was stored at −80°C until use. The full-length cDNA library was constructed using the Creator SMART cDNA Construction Kit (Clontech Laboratories, CA, USA). Concentrations of isolated RNAs were evaluated using a nanodrop spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA) and then used for first-strand cDNA synthesis. Second-strand cDNA was purified by QIAquick (Qiagen, Venlo, Netherlands) and ligated or transformed into the pTripleEX Vector. Plasmid DNA extraction was processed using a Multiscreen Plasmid Extraction Kit (Millipore) and purified. The cDNA library was amplified and using GeneAmp PCR System 9700–384 (Applied Biosystems, CA, USA) and DNA clones was sequenced with single-pass sequencing from the 5′-ends of the cDNA.

2.3. Assembly Annotation. Because genome and gene information were unavailable, assembly was performed without clustering. In the pretreatment process, PHRED was used to transfer peak information into the quality file and trim low-quality bases. Vector sequences were trimmed using Cross_match (http://www.macvector.com/Assembler/trimmingwith-crossmatch.html). Chimeric clones, polyA-tails, and sequences less than 100 bp were removed with Seqclean. Assembly was performed with CAP3 software. Contigs were manually checked and, together with singlet reads, compiled to generate a final unigene file. Finally, 412 unigene sequences were obtained from 1,000 ESTs, which were composed of 69 contigs and 343 singletons.

Unigenes were searched against the NCBI nonredundant nucleotide (NT) and protein (NR) databases (http://www.ncbi.nlm.nih.gov/), the Uniprot_sprot database (http://www.ebi.ac.uk/), and BLAST2GO (https://www.blast2go.com/) for functional annotation using the BLAST alignment tool. A sequence was considered as a significant match when the BLAST probability value (E-value) was less than 1e-5, and the match with the most significant E-value was recognized as the best annotation. A BLASTx search was also conducted against the UniProtKB/Swissprot database (http://www.ebi.ac.uk/) using default parameters. Unigenes were further annotated with GO terms (http://geneontology.org/).

2.4. Primer Design for EST-SSR Markers. A total of 1,000 ESTs of 1,056 samples obtained from a cDNA library were detected and analyzed by TRF version 4.07b online software (http://tandem.bu.edu/trf/trf.html). SSR sequences were then obtained. SSRs that fit the following criteria were considered for primer design: a minimum length of 18 bp with minimum repetitions for di-, tri-, tetra-, penta-, and hexa-4 and 4, respectively. Primers were designed using Primer 3 (http://www.premierbiosoft.com/primerdesign/) according to the following core criteria: a primer length ranging from 18 bp to 22 bp, with 20 bp as the optimum; product size ranging from 100 bp to 400 bp; melting temperature between 50°C and 62°C, with 60°C as the optimum; and GC content between 40% and 60%, with avoidance of mismatch, hairpin structures, and primer dimers that can cause nonspecific amplification.

3. Results

3.1. cDNA Library Quality Check and Reads Assembly. A full-length cDNA library was constructed from a mixture of *P. frutescens* samples. Library quality was evaluated after sequencing 96 randomly selected clones. On average, the insert size was greater than 1.2 kb. Forty-nine clones (51.04%) yielded sequencing reads above 700 bp; 11 clones (11.45%) were less than 500 bp. After confirming clone quality, a mass-scale sequencing approach was used. Construction of the full-length cDNA library was produced from *P. frutescens*. A total of 1,000 randomly selected clones from the cDNA library were subjected to single-orientation sequencing from the 5′-end using an ABI3730xl Platform (BGI). Read lengths ranged from 420 bp to 844 bp, with an average of 632 bp (Figure 1).

3.2. GC Content by Assembly of cDNA Reads. One thousand EST reads were obtained by trimming vector contaminants with Crossmatch and eliminating chimeric clones and short sequences (less than 100 bp). EST reads were then assembled by PHRAP and CAPS software [28, 29]. Results from the CAP3 assembly indicated that the GC content of unigenes varied from 29.46% to 61.32%. Ninety-one percent of the unigenes exhibited GC content between 37.93% and 52.87% (Figure 2).

3.3. Sequence Annotation. Annotation of the EST library was achieved through BLAST (Table 1). The NCBI nonredundant

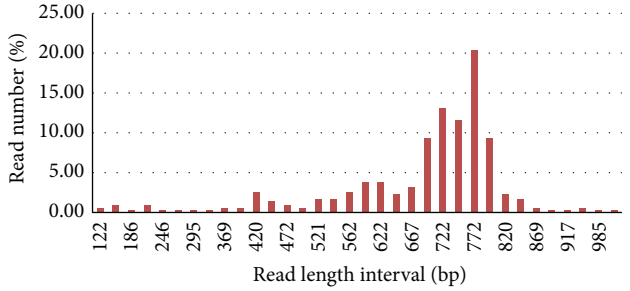


Figure 1: Reads length representation in EST (expressed sequence tags) sequencing of *Perilla frutescens*. Range of read length was indicated from 121 bps to 1051 bps.

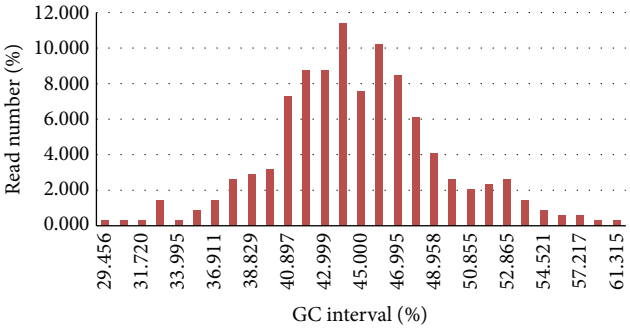


Figure 2: GC content division of unigenes. GC content of unigenes changed from 29.45% to 61.32%.

nucleotide (NT) (BLASTn) database resulted in 312 (90.96%) unigenes, whereas the protein (NR) database (BLASTx) produced 322 (93.88%) annotations. Uniprot/Swissprot (BLASTx) databases revealed the annotation of 317 (92.42%) unigenes. Moreover, the annotation data from COG (BLASTx) classification revealed 111 (32.36%) unigenes. Results from the NR database were determined to match that of the sequence homology with two species, *Sesamum indicum* (185 genes, 57.45) and *Erythranthe guttata* (78 genes, 24.22%). The remaining genes exhibited low levels (less than 1.86%) of sequence homology (Table 2).

3.4. Classification of Annotated Genes by GO Analysis. Gene Ontology (GO) distribution using hierarchy level 2 of the GO program resulted in three major clusters: biological process, cellular component, and molecular function (Figure 3). First, the biological process group was separated into 13 subclasses: signaling (5 genes), response to stimulus (24 genes), growth (1 gene), developmental process (3 genes), multicellular organismal process (3 genes), cellular process (93 genes), biological regulation (16 genes), single-organism process (59 genes), metabolic process (97 genes), localization (15 genes), reproductive process (2 genes), multiorganism process (6 genes), and cellular component organization or biogenesis (21 genes). Organelle (68 genes), cell (94 genes), extracellular region (8 genes), membrane-enclosed lumen (6 genes), cell junction (1 gene), macromolecular complex (44 genes), symplast (1 gene), and membrane (47 genes) genes were distributed from the cellular component cluster. The major components of the molecular function subset consisted of binding (67 genes) and catalytic activity (60 genes) genes.

Table 1: Annotated unigenes from different databases by EST (expressed sequence tags) sequencing of *Perilla frutescens*.

Annotation DB (methods)	Hits	%	No hits	%
NT (BLASTn)	312	90.96%	31	9.04%
NR (BLASTx)	322	93.88%	21	6.12%
Uniprot + Swissprot (BLASTx)	317	92.42%	26	7.58%
COG (BLASTx)	111	32.36%	232	67.64%

Table 2: List of species containing sequence matches to *Perilla frutescens*.

species (total: 38)	genes (total: 322)
Sesamum indicum	185
Erythranthe guttata	78
Salvia miltiorrhiza	6
Coffea canephora	4
Vitis vinifera	4
Nicotiana sylvestris	4
Genlisea aurea	3
Prunus persica	2
Nicotiana tomentosiformis	2
Ricinus communis	2
Brassica napus	2
Gossypium arboreum	2
Phoenix dactylifera	2
Perilla frutescens	2
Prunus mume	1
Medicago truncatula	1
Malus domestica	1
Solanum tuberosum	1
Schiedea haleakalensis	1
Mentha × piperita	1
Ajuga reptans	1
Codiaeum variegatum	1
Scutellaria baicalensis	1
Nicotiana tabacum	1
Morus notabilis	1
Citrus sinensis	1
Eucalyptus grandis	1
Eutrema salsugineum	1
Citrus clementina	1
Elaeis guineensis	1
Tarenaya hassleriana	1
Miscanthus sinensis	1
Arabidopsis thaliana	1
Arachis diogoi	1
Glycine max	1
Populus trichocarpa	1
Lolium perenne	1
Jatropha curcas	1

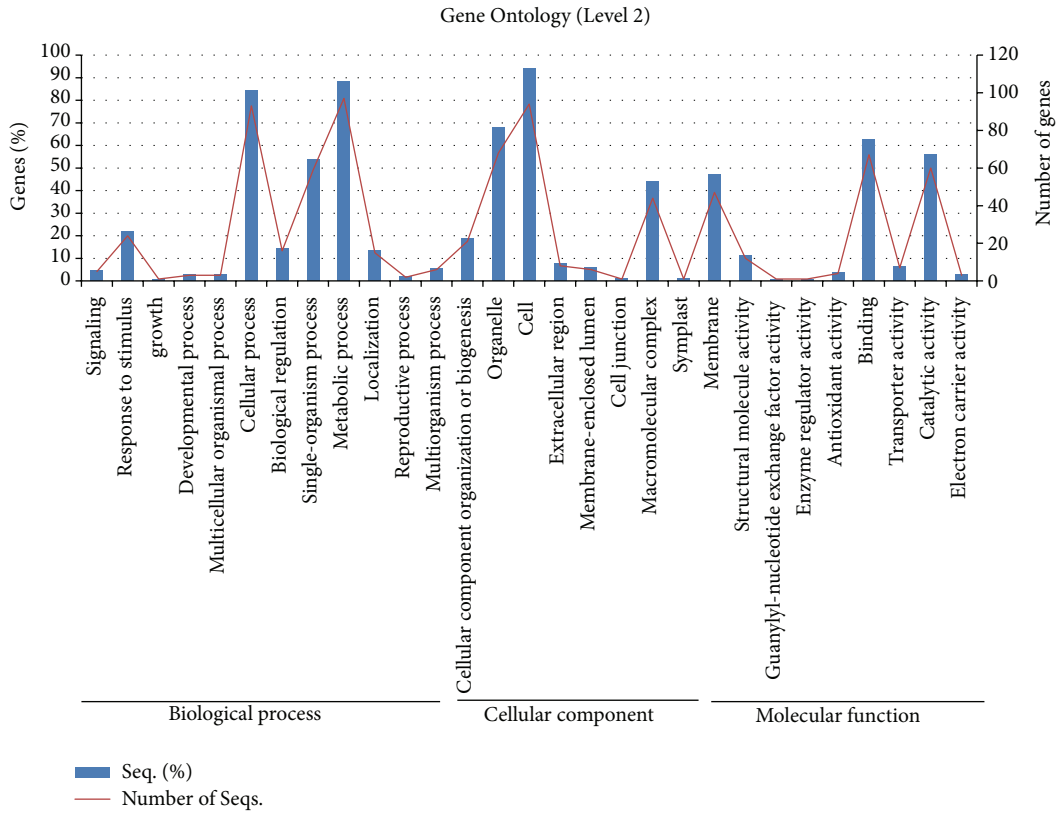


FIGURE 3: Classification of unigenes by Gene Ontology (GO) analysis. Three major clusters were displayed with annotated genes at hierarchy level 2 of GO analysis.

3.5. EST-SSR Traits in P. frutescens. A total of 343 unigene sequences were investigated. SSR sequences were obtained using TRF version 4.07b online software. Eighteen EST-SSR sequences were selected and analyzed following functional annotation (Table 3). Primer pairs were designed using the Primer 3 program. Expected product sizes ranged from 191 bp to 773 bp. In the future, we will perform additional classification studies through gene functions of *P. frutescens* using these EST-SSR primers.

4. Discussion

The major outcomes of this study were the construction of a full-length cDNA library from the important *P. frutescens* L-type (with limonene component) and the preliminary 1,000 ESTs identified (average 632 bp in length). Genome segment quality was affected by many factors. GC content analysis revealed a distribution between 29.46% and 61.32%. Earlier study showed that thirty to fifty percent of GC content influenced genome sequence quality in *Medicago truncatula* and *Lotus japonicas* [30]. GC content increment was related to the ratio of segments with matching EST data [31], consistent with that from the human genome [32].

Gene Ontology (GO) was utilized to obtain functional information and descriptions of gene products by studying domain-specific ontologies [33]. Annotation results consisted of biological data related to stress response genes, which were classified functionally using the GO hierarchy [34]. The corresponding classifications were processed to obtain additional information on the putative functionality for the subject accession number of pepper EST data from the GO databases [35]. GO "biological process" and "molecular function," generated by level 3, were annotated and associated with the number of sequences from each term, which were normalized by labeling with a GO term [36].

We also established 18 EST-SSR primers from the full-length cDNA library of *P. frutescens*. In *Vitis vinifera*, *Artemisia tridentate*, *Panax ginseng*, and *S. miltiorrhiza*, the EST-SSR motifs were generally di- and trinucleotide repeats [37–40]. However, this study revealed various penta-, hexa-, dodeca-, and tetradecanucleotide repeat motifs. This finding is in agreement with that for *Scutellaria baicalensis*, which contains penta- and hexanucleotide repeats [41]. Differences in repeat type may be attributed to the degree of the SSR search criteria for the EST database in various plant species. The development of EST-SSR markers has many advantages compared with other molecular markers and can be used to study genetic diversity, evolution, comparative genomics, and gene-based associations.

Construction of a full-length cDNA library is significant for comparative genomics, genome sequence validation, and design of EST-SSR primers that display entire transcription

Table 3: EST-SSR primer pairs produced in EST (expressed sequence tags) sequencing database of *Perilla frutescens*.

Number	Unigene ID	Repeat motifs	left primer sequence	Tm.	right primer sequence	Tm.	Product size (bp)	Annotation (NR DB)
1	Contig 17	ATCAT(8)	GAGAGTATAAACAAATCCAAAACAGC	58.795	AGCCGGTATATCCAATTCCC	60.006	562	PREDICTED: protein CURVATURE THYLAKOID 1B, chloroplastic [Sesamum indicum]
2	Contig 67	A(29)	AGCAACTGCGGGTAGCTAGA	60.176	CAATCCGACCACAGTTGATG	59.96	172	PREDICTED: photosystem I subunit O [Sesamum indicum]
3	Perilla-1-1a_pTriplEx2-seq_C16	GA(16)	AGCGTACTGTTGAAAGCGTG	59.148	CAGCAAACGTGCTCGAATTA	60.014	247	PREDICTED: uncharacterized protein LOC105172991 isoform X2 [Sesamum indicum]
4	Perilla-1-1a_pTriplEx2-seq_E18	CTT(9)	GCCAATTTGAAGCTTTAGCC	58.969	GAATGTGAAGTGGGAACGCT	60.119	773	PREDICTED: GRF1-interacting factor 3 [Sesamum indicum]
5	Perilla-1-1a_pTriplEx2-seq_M02	AGAATG(4)	TGGAGCAAGTGAAGCAACAG	60.175	CCTTTTCAGTGAGGAGCCAG	59.982	191	
6	Perilla-1-2a_pTriplEx2-seq_A02	TTTTG(7)	AATGATGGGTGTGATGAGCA	59.925	AAAGAATTTGAAGGCGCAGA	59.96	401	PREDICTED: homeobox-leucine zipper protein HAT5-like [Sesamum indicum]
7	Perilla-1-3a_pTriplEx2-seq_B13	ATCAT(8)	GAGAGTATAAACAAATCCAAAACAGC	58.795	CGGTATATCCAATTCCCACG	60.031	559	hypothetical protein MIMGU_mgv1a015066mg [Erythranthe guttata]
8	Perilla-1-3a_pTriplEx2-seq_L05	CT(14)	CCCAAATTCACATCCACTGA	59.343	AACAACTGACATGGCCTTCC	59.973	185	PREDICTED: uncharacterized protein LOC105160440 [Sesamum indicum]
9	Perilla-1-3a_pTriplEx2-seq_L15	TC(15)	CAGTTTTAACTTCGCCTCGC	60.018	CACTCGCAAAAAGGGGTAAG	59.741	619	PREDICTED: annexin D5 [Sesamum indicum]
10	Perilla-2-1a_pTriplEx2-seq_A19	GGA(8)	GCTCCTCGCAGTAACTTTGG	60.015	TCATCTCTTGCTCTGTTTCCA	58.583	107	hypothetical protein MIMGU_mgv1a016040mg [Erythranthe guttata]
11	Perilla-2-2a_pTriplEx2-seq_A06	CT(12)	CATTGGCCTTAAACTTCGGA	60.067	ATAAATGTGGATTGGGGCAA	60.016	341	hypothetical protein MIMGU_mgv1a020048mg [Erythranthe guttata]
12	Perilla-2-2a_pTriplEx2-seq_C18	AG(14)	GGGGGATCATTTCCAGTCTT	60.133	GTGCCCACTGGTTCTTTGTT	60.012	404	hypothetical protein MIMGU_mgv1a012334mg [Erythranthe guttata]
13	Perilla-3-2a_pTriplEx2-seq_E14	GATGACGATGAT(2)	CTTTCCAACCCTCCGAATTT	60.291	CGACGCCTGTCTCATCTACA	60.008	522	NAC transcription factor 1 [Salvia miltiorrhiza]
14	Perilla-3-2a_pTriplEx2-seq_O22	GA(17)	GGGGATATGTTATGTTGCTTGTT	59.179	TCGCCGTACTTGATCCCTAC	60.096	514	PREDICTED: uncharacterized protein LOC105169169 isoform X1 [Sesamum indicum]
15	Perilla-3-3a_pTriplEx2-seq_B03	CT(16)	CGAGTGTGTTCGTATGGGTG	60.025	AACGCGTACGGAACAGAGAC	60.321	184	hypothetical protein MIMGU_mgv1a014836mg [Erythranthe guttata]
16	Perilla-3-3a_pTriplEx2-seq_B23	TCCTCTTCCTCTCC(2)	TAGTGTCGAAGCTCAATGGC	59.028	TGACCAGCATCAGCTTTCAC	59.992	662	PREDICTED: chloroplast stem-loop binding protein of 41 kDa a, chloroplastic [Sesamum indicum]
17	Perilla-3-3a_pTriplEx2-seq_F11	GAG(9)	GAAAGACTGGTTGGCTCTGG	59.844	ATCCAAAATTCGTCCTGTGC	59.939	381	hypothetical protein MIMGU_mgv1a011207mg [Erythranthe guttata]
18	Perilla-3-3a_pTriplEx2-seq_P03	GA(13)	AAAGCTGTTTGCCCTTGCTA	60.018	CTCAAATGGAGTCACGCAGA	59.984	284	hypothetical protein MIMGU_mgv1a016830mg [Erythranthe guttata]

units rather than partial gene sequences [42]. One benefit of constructing a full-length cDNA library is that it allowed us to conduct proper gene modeling while comparing other cDNA sequences in *P. frutescens*. The full-length cDNA sequences will be useful for annotation of the plant genome. Another advantage of EST sequencing is the increased ratio of unigenes with definitive GO categories compared with other libraries. The library built by this method included a high proportion of full-length cDNAs [42], allowing us to have a database of this library available for *P. frutescens* genomics studies.

This full-length cDNA library provides a wealth of knowledge about the unique EST sequences available for the *P. frutescens* genome and, particularly, about the addition of 5′-end sequences that are more unique and valuable for gene identification. These EST tags will be useful for functional gene annotation, analysis of splice site variations, and gene homologies as additional whole-genome sequences become available in *P. frutescens*.

5. Conclusions

Perilla frutescens is valuable as a medicinal plant as well as a natural medicine and functional food. However, comparative genomics analyses of *P. frutescens* are limited due to a lack of gene annotations and characterization. A full-length cDNA library from *P. frutescens* leaves was constructed to identify functional gene clusters and probable EST-SSR markers through 1,056 examples of expressed sequence tag (EST) sequencing data. Unigene assembly was performed using basic local alignment search tool (BLAST) homology searches and annotated Gene Ontology (GO). A total of 18 simple sequence repeats (SSRs) were designed as primer pairs. This study is the first to report comparative genomics and EST-SSR markers from *P. frutescens* to ease gene discovery and provide an important source for functional genomics and molecular genetic research in this interesting medicinal plant.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Eun Soo Seong and Ji Hye Yoo contributed equally to this work.

Acknowledgments

This study was supported by the Ministry of Education (MOE) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation (no. 2014H1C1A1067085) and was partially supported by the Bioherb Research Institute and the Research Institute of Agricultural Science, Kangwon National University, Republic of Korea.

References

[1] Y.-J. Park, A. Dixit, K.-H. Ma et al., "Evaluation of genetic diversity and relationships within an on-farm collection of *Perilla frutescens* (L.) Britt. using microsatellite markers," *Genetic Resources and Crop Evolution*, vol. 55, no. 4, pp. 523–535, 2008.

[2] A. K. Pandey and K. C. Bhatt, "Diversity distribution and collection of genetic resources of cultivated and weedy type in *Perilla frutescens* (L.) Britton var. frutescens and their uses in Indian Himalaya," *Genetic Resources and Crop Evolution*, vol. 55, no. 6, pp. 883–892, 2008.

[3] C. X. You, K. Yang, Y. Wu et al., "Chemical composition and insecticidal activities of the essential oil of *Perilla frutescens* (L.) Britt. aerial parts against two stored product insects," *European Food Research and Technology*, vol. 239, pp. 481–490, 2014.

[4] H.-S. Shin and S.-W. Kim, "Lipid composition of perilla seed," *Journal of the American Oil Chemists' Society*, vol. 71, no. 6, pp. 619–622, 1994.

[5] T. Longvah and Y. G. Deosthale, "Chemical and nutritional studies on Hanshi (*Perilla frutescens*), a traditional oilseed from northeast India," *Journal of the American Oil Chemists Society*, vol. 68, no. 10, pp. 781–784, 1991.

[6] S. J. Kim, E. Y. Kang, S. E. Won et al., "Chemical composition and comparison of essential oil contents of *Perilla frutescens* Britton var. japonica HARA leaves," *Korean Journal of Medicinal Crop Science*, vol. 16, pp. 242–254, 2008.

[7] M. Nitta, H. Kobayashi, M. Ohnishi-Kameyama, T. Nagamine, and M. Yoshida, "Essential oil variation of cultivated and wild Perilla analyzed by GC/MS," *Biochemical Systematics and Ecology*, vol. 34, no. 1, pp. 25–37, 2006.

[8] H. Ueda, C. Yamazaki, and M. Yamazaki, "Luteolin as an anti-inflammatory and anti-allergic constituent of *Perilla frutescens*," *Biological and Pharmaceutical Bulletin*, vol. 25, no. 9, pp. 1197–1202, 2002.

[9] M.-K. Kim, H.-S. Lee, E.-J. Kim et al., "Protective effect of aqueous extract of *Perilla frutescens* on tert-butyl hydroperoxide-induced oxidative hepatotoxicity in rats," *Food and Chemical Toxicology*, vol. 45, no. 9, pp. 1738–1744, 2007.

[10] M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, and J. C. Venter, "Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library," *Nature Genetics*, vol. 4, no. 4, pp. 373–380, 1993.

[11] R. K. Varshney, A. Graner, and M. E. Sorrells, "Genic microsatellite markers in plants: features and applications," *Trends in Biotechnology*, vol. 23, no. 1, pp. 48–55, 2005.

[12] R. K. Varshney, R. Sigmund, A. Börner et al., "Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice," *Plant Science*, vol. 168, no. 1, pp. 195–202, 2005.

[13] R. Peakall, S. Gilmore, W. Keys, M. Morgante, and A. Rafalski, "Cross-species amplification of soybean (Glycine max) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants," *Molecular Biology and Evolution*, vol. 15, no. 10, pp. 1275–1287, 1998.

[14] S. Temnykh, G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch, "Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential," *Genome Research*, vol. 11, no. 8, pp. 1441–1452, 2001.

[15] T. Thiel, W. Michalek, R. K. Varshney, and A. Graner, "Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.)," *Theoretical and Applied Genetics*, vol. 106, no. 3, pp. 411–422, 2003.

[16] J.-K. Yu, T. M. Dake, S. Singh et al., "Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat," *Genome*, vol. 47, no. 5, pp. 805–818, 2004.

[17] J.-K. Yu, M. La Rota, R. V. Kantety, and M. E. Sorrells, "EST derived SSR markers for comparative mapping in wheat and rice," *Molecular Genetics and Genomics*, vol. 271, no. 6, pp. 742–751, 2004.

[18] A. Heesacker, V. K. Kishore, W. Gao et al., "SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility," *Theoretical and Applied Genetics*, vol. 117, no. 7, pp. 1021–1029, 2008.

[19] J. R. Ellis and J. M. Burke, "EST-SSRs as a resource for population genetic analyses," *Heredity*, vol. 99, no. 2, pp. 125–132, 2007.

[20] Y.-C. Li, A. B. Korol, T. Fahima, and E. Nevo, "Microsatellites within genes: structure, function, and evolution," *Molecular Biology and Evolution*, vol. 21, no. 6, pp. 991–1007, 2004.

[21] C. H. Pashley, J. R. Ellis, D. E. McCauley, and J. M. Burke, "EST databases as a source for molecular markers: lessons from *Helianthus*," *Journal of Heredity*, vol. 97, no. 4, pp. 381–388, 2006.

[22] M. A. Chapman, J. Hvala, J. Strever et al., "Development, polymorphism, and cross-taxon utility of EST-SSR markers from safflower (*Carthamus tinctorius* L.)," *Theoretical and Applied Genetics*, vol. 120, no. 1, pp. 85–91, 2009.

[23] A. Depeiges, C. Goubely, A. Lenoir et al., "Identification of the most represented repeated motifs in *Arabidopsis thaliana* microsatellite loci," *Theoretical and Applied Genetics*, vol. 91, no. 1, pp. 160–168, 1995.

[24] G. M. Cordeiro, R. Casu, C. L. McIntyre, J. M. Manners, and R. J. Henry, "Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to *Erianthus* and sorghum," *Plant Science*, vol. 160, no. 6, pp. 1115–1123, 2001.

[25] L. S. Lima, K. P. Gramacho, A. S. Gesteira et al., "Characterization of microsatellites from cacao-*Moniliophthora perniciosa* interaction expressed sequence tags," *Molecular Breeding*, vol. 22, no. 2, pp. 315–318, 2008.

[26] E. De Keyser, J. de Riek, and E. van Bockstaele, "Discovery of species-wide EST-derived markers in *Rhododendron* by intron-flanking primer design," *Molecular Breeding*, vol. 23, no. 1, pp. 171–178, 2009.

[27] E. S. Seong, J. H. Yoo, J. H. Choi et al., "Construction and classification of a cDNA Library from *Miscanthus sinenesis* (Eulalia) treated with UV-B," *Plant Omics Journal*, vol. 8, pp. 264–269, 2015.

[28] F. C. Peixoto and J. M. Ortega, "On the pursuit of optimal sequence trimming parameters for EST projects," in *Proceedings of the 1st Brazilian Symposium/Workshop on Bioinformatics (BSB/WOB '02)*, pp. 48–55, Gramado, Brazil, October 2002.

[29] X. Huang and A. Madan, "CAP3: a DNA sequence assembly program," *Genome Research*, vol. 9, no. 9, pp. 868–877, 1999.

[30] L. Shangguan, J. Han, E. Kayesh et al., "Evaluation of genome sequencing quality in selected plant species using expressed sequence tags," *PLoS ONE*, vol. 8, no. 7, Article ID e69890, 2013.

[31] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, and M. C. Zody, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[32] R. Versteeg, B. D. C. van Schaik, M. F. van Batenburg et al., "The human transcriptome map reveals extremes in gene dentistry, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes," *Genome Research*, vol. 13, no. 9, pp. 1998–2004, 2003.

[33] S. K. Arasan, J.-I. Park, N. U. Ahmed et al., "Gene ontology based characterization of expressed sequence tags (ESTs) of *Brassica rapa* cv. Osome," *Indian Journal of Experimental Biology*, vol. 51, no. 7, pp. 522–530, 2013.

[34] The Gene Ontology Consortium, "Gene ontology annotations and resources," *Nucleic Acids Research*, vol. 41, pp. 530–535, 2013.

[35] H. J. Kim, K. H. Baek, S. W. Lee et al., "Pepper EST database: comprehensive *in silico* tool for analyzing the chili pepper (*Capsicum annuum*) transcriptome," *BMC Plant Biology*, vol. 8, article 101, 7 pages, 2008.

[36] A. Conesa and S. Götz, "Blast2GO: a comprehensive suite for functional analysis in plant genomics," *International Journal of Plant Genomics*, vol. 2008, Article ID 619832, 12 pages, 2008.

[37] K. D. Scott, P. Eggler, G. Seaton et al., "Analysis of SSRs derived from grape ESTs," *Theoretical and Applied Genetics*, vol. 100, no. 5, pp. 723–726, 2000.

[38] P. Bajgain, B. A. Richardson, J. C. Price, R. C. Cronn, and J. A. Udall, "Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentata*)," *BMC Genomics*, vol. 12, article 370, 2011.

[39] T. S. Wu, C. Liang, H. B. Li, and Z. Y. Piao, "Development of unigene derived microsatellite (UGMS) markers in *Panax ginseng*," *Scientia Agricultura Sinica*, vol. 44, pp. 2650–2660, 2011.

[40] K.-J. Deng, Y. Zhang, B.-Q. Xiong et al., "Identification, characterization and utilization of simple sequence repeat markers derived from Salvia miltiorrhiza expressed sequence tags," *Acta Pharmacologica Sinica*, vol. 44, no. 10, pp. 1165–1172, 2009.

[41] Y. Yuan, P. Long, C. Jiang, M. Li, and L. Huang, "Development and characterization of simple sequence repeat (SSR) markers based on a full-length cDNA library of *Scutellaria baicalensis*," *Genomics*, vol. 105, no. 1, pp. 61–67, 2015.

[42] M. Seki, M. Narusaka, A. Kamiya et al., "Functional annotation of a full-length Arabidopsis cDNA collection," *Science*, vol. 296, no. 5565, pp. 141–145, 2002.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 843802, 10 pages
http://dx.doi.org/10.1155/2015/843802

Research Article

De Novo Transcriptome Sequencing of the Orange-Fleshed Sweet Potato and Analysis of Differentially Expressed Genes Related to Carotenoid Biosynthesis

Ruijie Li, Hong Zhai, Chen Kang, Degao Liu, Shaozhen He, and Qingchang Liu

Beijing Key Laboratory of Crop Genetic Improvement/Laboratory of Crop Heterosis and Utilization, Ministry of Education, China Agricultural University, Beijing 100193, China

Correspondence should be addressed to Qingchang Liu; liuqc@cau.edu.cn

Received 30 July 2015; Revised 29 September 2015; Accepted 1 October 2015

Academic Editor: Xiaohan Yang

Copyright © 2015 Ruijie Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sweet potato, *Ipomoea batatas* (L.) Lam., is an important food crop worldwide. The orange-fleshed sweet potato is considered to be an important source of beta-carotene. In this study, the transcriptome profiles of an orange-fleshed sweet potato cultivar "Weiduoli" and its mutant "HVB-3" with high carotenoid content were determined by using the high-throughput sequencing technology. A total of 13,767,387 and 9,837,090 high-quality reads were produced from Weiduoli and HVB-3, respectively. These reads were *de novo* assembled into 58,277 transcripts and 35,909 unigenes with an average length of 596 bp and 533 bp, respectively. In all, 874 differentially expressed genes (DEGs) were obtained between Weiduoli and HVB-3, 401 of which were upregulated and 473 were downregulated in HVB-3 compared to Weiduoli. Of the 697 DEGs annotated, 316 DEGs had GO terms and 62 DEGs were mapped onto 50 pathways. The 22 DEGs and 31 transcription factors involved in carotenoid biosynthesis were identified between Weiduoli and HVB-3. In addition, 1,725 SSR markers were detected. This study provides the genomic resources for discovering the genes involved in carotenoid biosynthesis of sweet potato and other plants.

1. Introduction

Sweet potato, *Ipomoea batatas* (L.) Lam., is an important food crop widely cultivated in the world, especially in the tropics, subtropics, and some temperate zones of the developing countries [1, 2]. This crop is also used to produce alcohol and various antioxidants such as anthocyanin and carotenoids [3, 4]. The storage roots of orange-fleshed sweet potato are rich in beta-carotene, a precursor of vitamin A [5]. High carotenoid content has become one of the most important objectives in sweet potato breeding [6]. Sweet potato is an autohexaploid ($2n = 6x = 90$) and its estimated genome size is 2.4 Gb [7]. The genome data sources for sweet potato are important for gene discovery due to its complex genome.

Carotenoids are widely produced in plants, algae, fungi, and bacteria and provide potent nutritional benefits to humans because their bodies are unable to synthesize carotenoids independently [8, 9]. The necessity of this nutritional component has caused scientists to try to increase the carotenoid content of crops through different approaches. In plants, carotenoids are synthesized through a series of chemical reactions including condensation, dehydrogenation, cyclization, hydroxylation, and epoxidation. To date, a number of genes involved in the carotenoid biosynthesis have been cloned from several plants and their overexpression was found to significantly increase carotenoid levels in canola seeds [10], tomato fruits [11], and rice seeds [12, 13]. Several carotenoid biosynthesis-associated genes have also been isolated from sweet potato [6, 14–17]. However, the molecular mechanisms regulating flux through the pathway are unclear though carotenoid synthesis is well characterized.

Genomic approaches have been used for discovering the important genes involved in plant secondary metabolism pathways. However, the genome of sweet potato is still unavailable. Transcription sequencing is an efficient way for discovering and characterizing novel enzymes and transcription factors from sweet potato. Transcriptome sequencing of sweet potato has provided an important transcriptional data

source for studying storage root formation, flower development, and anthocyanin biosynthesis of this crop [7, 18–22]. Here, we performed *de novo* transcriptome sequencing of the orange-fleshed sweet potato by Illumina paired-end (PE) RNA sequencing technology and analyzed differentially expressed genes related to carotenoid biosynthesis.

2. Materials and Methods

2.1. Plant Materials. The orange-fleshed sweet potato cultivar "Weiduoli" and its high carotenoid mutant "HVB-3" were used in this study. Weiduoli is a commercial cultivar with carotenoid content of 9.02 mg/100 g (FW) and β-carotene content of 7.70 mg/100 g. In HVB-3, the contents of carotenoid and β-carotene are up to 21.42 mg/100 g and 19.95 mg/100 g, respectively. The storage roots of both materials were harvested after 110 days of planting, cleaned with sterilized water, and cut into 5 mm × 5 mm pieces. The collected samples were immediately frozen in liquid nitrogen and stored at −80°C for RNA extraction.

2.2. RNA Extraction. Total RNA from storage roots of Weiduoli and HVB-3 was extracted using the RNAprep Pure Plant Kit (Tiangen Biotech, Beijing, China). To avoid the contamination of genomic DNA, the extracted RNA was treated with DNase I (Takara, Dalian, China) for 4 h at 37°C. The quality of RNA was examined using 1% agarose gel before proceeding. Total RNA was quantified by using a Nanodrop spectrophotometer (Thermo Nanodrop Technologies, Wilmington, DE, USA). Both the A260/280 and A260/230 ratios were checked to ensure the purity of the total RNA and 10 μg RNA was used for Illumina paired-end (PE) sequencing.

2.3. cDNA Library Construction and Illumina Sequencing. Using magnetic beads with oligo (dT), poly-A mRNA was enriched from total RNA to construct a cDNA library for RNA sequencing. The enriched mRNA was broken into short fragments by adding fragment buffer. Using these short fragments, the first-strand of cDNA was synthesized by random hexamer primers. Then, using DNA polymerase I and RNase H (Tiangen Biotech, Beijing, China), the second-strand of cDNA was synthesized. After purification with a QiaQuick PCR extraction kit (Qiagen, Valencia, CA, USA), the cDNA fragments were resolved in elution buffer (EB) for end reparation and the addition of a poly(A) tail. Sequencing adapters were connected to the short fragments. These products were purified by agarose gel electrophoresis and suitable fragments (about 180 bp) were isolated as templates for PCR amplification. The cDNA library was constructed for sequencing by 2 × 100 PE using Illumina HiSeq 2000.

2.4. Raw Sequence Processing and De Novo Assembly. To obtain high quality reads for *de novo* assembly, the raw reads from RNA-seq were cleaned by removing adaptor sequences, empty reads, low quality reads (with ambiguous sequences "*N*"), and reads with more than 10% $Q < 20$ bases ($Q = -10 \times \text{lgE}$). The clean reads from the two libraries were assembled together with the Trinity software [23]. The reads were assembled into the contigs with the Inchworm program. The minimally overlapping contigs were clustered into sets of connected components by the Chrysalis program, and then the transcripts were constructed by the Butterfly program. The transcripts were clustered by similarity of correct match length beyond the 80% of longer transcript or 90% of shorter transcript using multiple sequence alignment tool—BLAST [24]. Taking the longest transcript as the unigene of each cluster, these unigenes formed into the nonredundant unigene database.

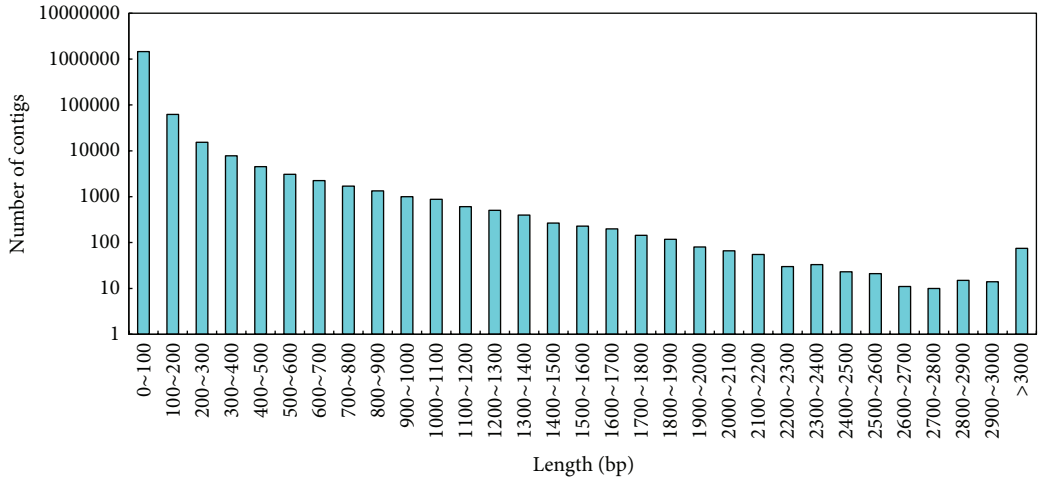
2.5. Analysis of Differentially Expressed Genes (DEGs). The expression of unigenes in Weiduoli and HVB-3 was calculated according to the RPKM method (reads per kb per million reads) described by Mortazavi et al. [25]. The IDEG6 software [26] was used to identify DEGs in the two libraries. The results of all statistical tests were corrected for multiple testing with the Benjamini-Hochberg false discovery rate (FDR < 0.01) and an absolute value of log2 ratio >1 was used to determine significant differences in gene expression.

2.6. Functional Annotation and Classification of DEGs. In order to deduce the correct transcription direction and protein sequences coded by DEGs, a BLASTX search was performed against the National Center for Biotechnology Information (NCBI) nonredundant (Nr) protein database (http://www.ncbi.nlm.nih.gov), the Swiss-Prot protein database, (http://www.expasy.ch/sprot), the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (http://www.genome.jp/kegg), Pfam database, and Cluster of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG) with a typical cut-off E value of 10^{-5}. Gene ontology (GO) was applied with the Blast2GO program to obtain annotation of DEGs [27]. The WEGO software was then used to perform GO functional classification of DEGs. DEGs were annotated with corresponding Enzyme Commission (EC) numbers using BLASTX alignments against KEGG with a cut-off E value of 10^{-5}. Gene names were assigned to each DEG based on the best BLAST hit (highest score). Searches were limited to the first 10 significant hits for each query to increase computational speed.

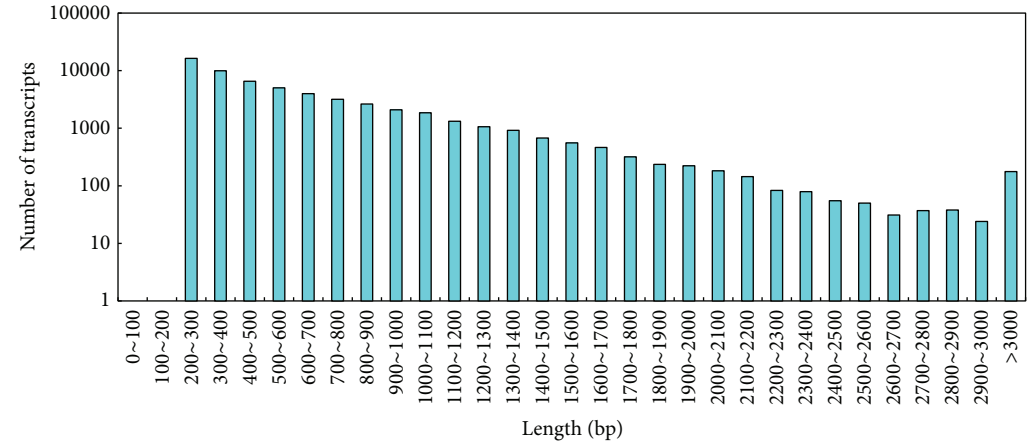
2.7. SSR Detection. SSRs were detected among the unigenes with length >1,000 bp using the software MISA (http://pgrc.ipk-gatersleben.de/misa/). A total of 6 types of SSRs were investigated, including mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats.

3. Results

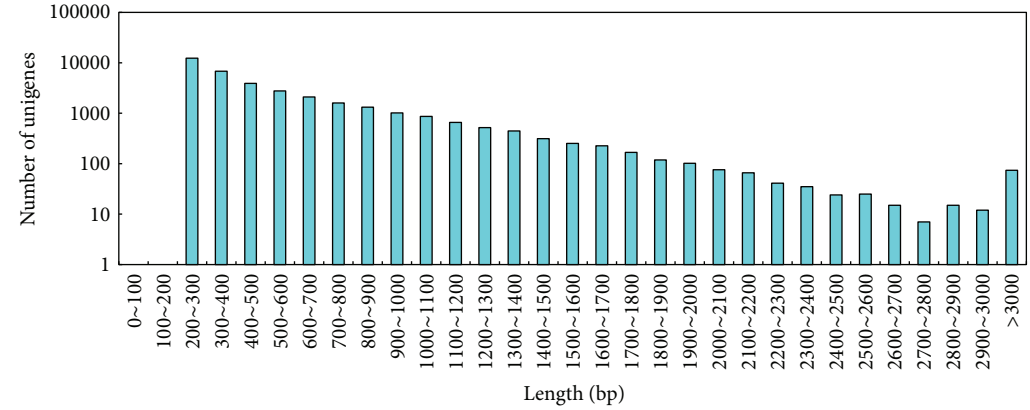
3.1. Transcriptome Sequencing and De Novo Assembly. Illumina HiSeq 2000 was used to determine the transcriptome profiles of sweet potato. After removing adaptor sequences and unknown or discarding low quality reads, 13,767,387 and 9,837,090 high-quality reads were obtained from Weiduoli and HVB-3, respectively (Table 1). With the Trinity assembly software, the high-quality reads were assembled into 1,557,001 contigs with an average length of 58 bp and N50 length of 58 bp. These contigs were assembled into 58,277 transcripts



(a)



(b)



(c)

Figure 1: Overview of the sweet potato transcriptome assembly. (a) Size distribution of contigs; (b) size distribution of transcripts; and (c) size distribution of unigenes.

with an average length of 596 bp and N50 length of 767 bp. The transcripts were further clustered into 35,909 unigenes with an average length of 533 bp and N50 length of 669 bp (Table 1). The length distributions of contigs, transcripts, and unigenes were shown in Figure 1.

3.2. Identification and Functional Annotation of DEGs. According to the BLASTX results, most of the unigenes had homologous proteins in the Nr protein database. Interestingly, 4,903 (18.71%) and 4,463 (17.03%) unigenes showed significant homology with sequences of *Nicotiana sylvestris*

Table 1: Length distribution of assembled contigs, transcripts, and unigenes from Weiduoli and HVB-3.

Length range	Contig	Transcript	Unigene
0–300	1,531,538 (98.36%)	16,317 (28.00%)	12,325 (34.32%)
300–500	12,342 (0.79%)	16,504 (28.32%)	10,710 (29.83%)
500–1000	9,349 (0.60%)	16,921 (29.04%)	8,813 (24.54%)
1000–2000	3,419 (0.22%)	7,635 (13.10%)	3,671 (10.22%)
2000+	353 (0.02%)	900 (1.54%)	390 (1.09%)
Total number	1,557,001	58,277	35,909
Total length	91,371,759	34,741,399	19,150,802
Mean length	58	596	533
N50 length	58	767	669

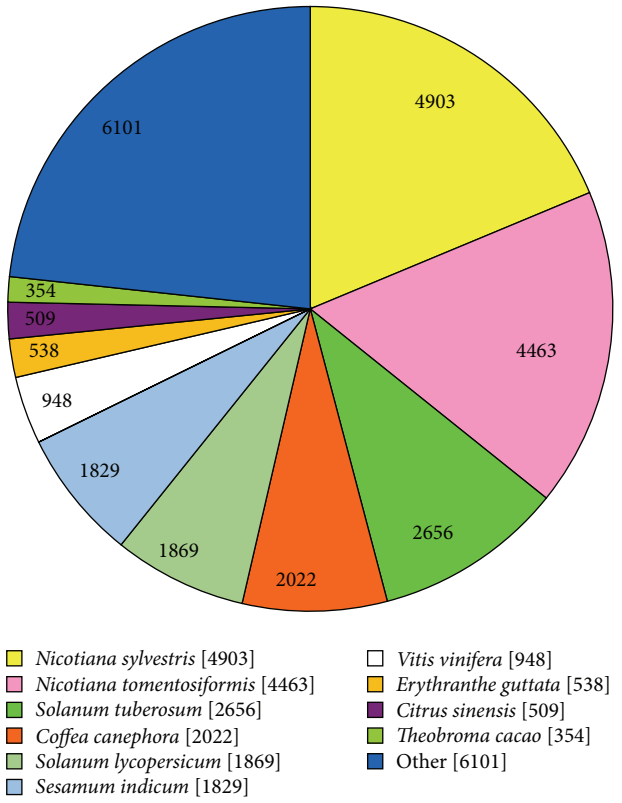


Figure 2: Species distribution of the top BlastX hits for each unigene in the Nr database.

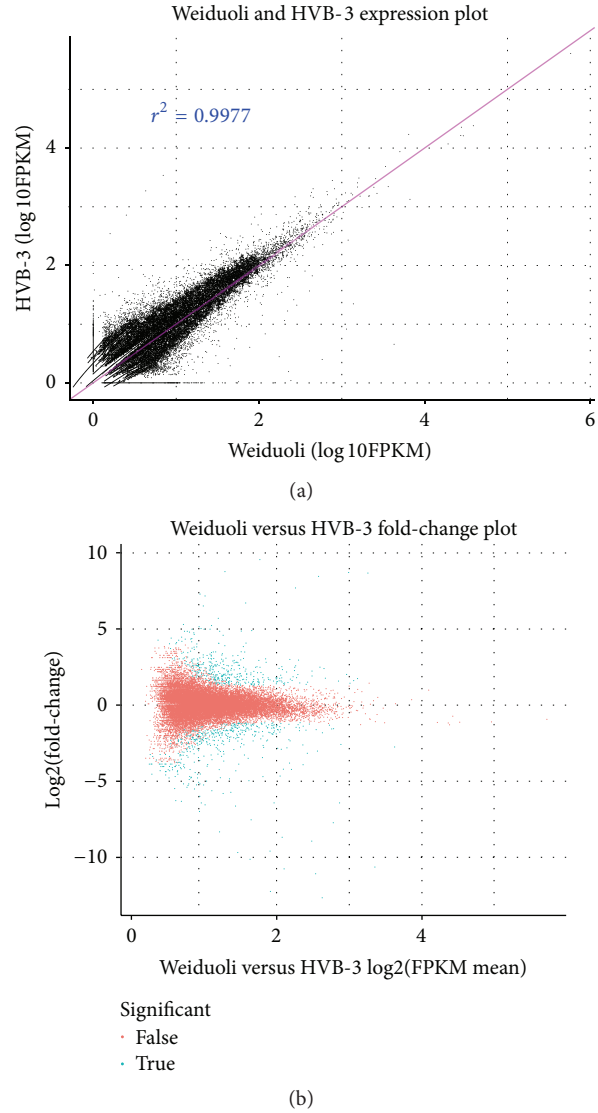


Figure 3: Comparative analysis of gene expression in Weiduoli and HVB-3. (a) A scatter plot of RPKM logarithmic values in libraries of Weiduoli and HVB-3. Each dot represents the RPKM value of a specific gene. The greater deviation from the diagonal slope shows a greater expression level of the gene in the corresponding material. (b) A scatter plot of the ratio of RPKM logarithmic numerical values of genes in Weiduoli and HVB-3. This plot graphically represents genes differentially expressed between Weiduoli and HVB-3. Blue dots represent genes that had significant difference and red dots represent genes where no significant difference was observed between Weiduoli and HVB-3.

and *Nicotiana tomentosiformis*, respectively (Figure 2). Furthermore, 14,316 (54.21%) unigenes had significant matches in the Pfam database, and 17,058 (64.59%) unigenes had similarity to proteins in the Swiss-Prot database.

The expression of unigenes was calculated according to the RPKM method. A total of 35,909 unigenes had detectable levels of expression in Weiduoli and HVB-3 (Figure 3(a)). Using the IDEG6 software, a total of 874 genes were found to be differentially expressed between HVB-3 and Weiduoli, and 401 of them were upregulated and 473 were downregulated in HVB-3 compared to Weiduoli (Figure 3(b)). A total of 697 DEGs were annotated against the public databases and 94.12% of them were mapped to the Nr library, suggesting that most of the DEGs can be translated into proteins. The mapping rate of DEGs against the Swiss-Prot protein database was 67.58%. The overall functional annotation is listed in Table 2.

A total of 14,136 unigenes were classified into three categories, cellular component, biological progress, and molecular function, through GO analysis (Figure 4). In all, 316 DEGs were classified into three categories, 149 with cellular component, 235 with biological progress, and 245 with

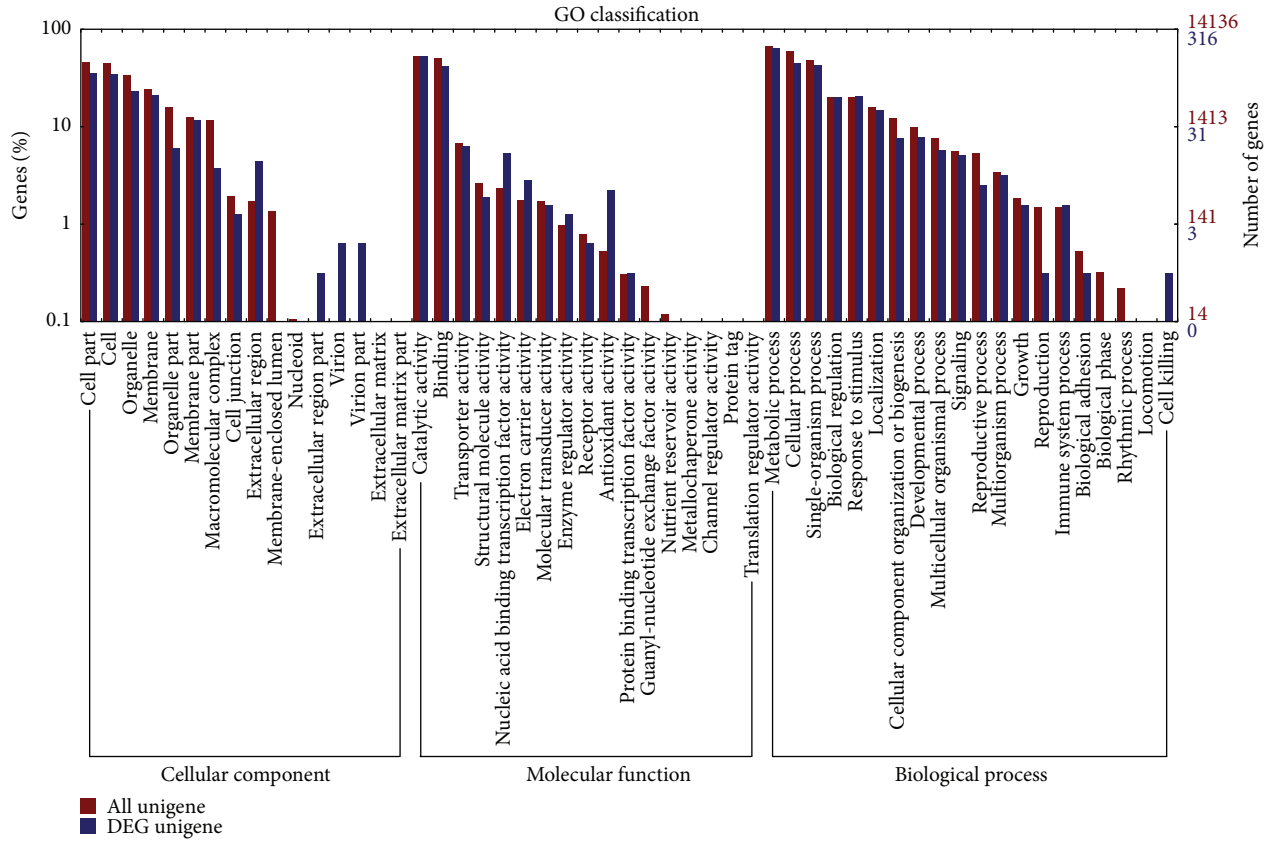


Figure 4: GO classification of unigenes in transcriptomes of Weiduoli and HVB-3. The red bars represent all the unigenes and the blue bars represent the DEGs.

Table 2: Functional annotation of DEGs between Weiduoli and HVB-3.

Annotation database	Annotation number
COG annotation	167
GO annotation	316
KEGG annotation	83
KOG annotation	290
Pfam annotation	477
Swiss-Prot annotation	471
Nr annotation	656
Total	697

molecular function, through GO analysis (Figure 4). The GO analysis revealed that most of the DEGs were involved in catalytic activity and metabolic process. Compared with the COG database, 240 DEGs were subdivided into 22 COG classifications, including secondary metabolite biosynthesis, transport, and catabolism, signal transduction mechanisms, replication, recombination, and repair, amino acid transport and metabolism, inorganic ion transport and metabolism, carbohydrate transport and metabolism, energy production and conversion, transcription, and lipid transport and metabolism (Figure 5).

Carotenoid biosynthesis which belongs to the secondary metabolisms is a dynamic and complex process catalyzed by a series of enzymes. Functional category analysis revealed that the DEGs were involved in a number of important pathways, including metabolite biosynthesis and signal transduction mechanisms (Figure 6), similar to the results of GO and COG analyses. According to the KEGG pathway enrichment results, 62 DEGs were assigned to the 50 pathways. The most noticeable pathways were terpenoid backbone biosynthesis and fatty acid metabolism. As shown in Table 3, 22 DEGs were found to be directly or indirectly involved in carotenoid biosynthesis. These 22 DEGs encoded geranylgeranyl pyrophosphate synthase (GGPS), geranylgeranyl diphosphate reductase (GGPR), dehydrodolichyl diphosphate synthase (DHDDS), alcohol dehydrogenases homologous, aldehyde dehydrogenase, alcohol dehydrogenase, long chain acyl-CoA synthetase, and 15 cytochrome P450, respectively (Table 3). Interestingly, several important transcription factors, including NAC, MYB, AP2/ERF, Zifc fingers, WRKY, bZIP, and ARF, were found to be significantly upregulated in HVB-3 compared to Weiduoli (Table 3).

3.3. SSR Markers. The MISA was used to search for SSRs. A total of 1,725 potential cDNA-derived SSRs (cSSRs) were identified from 4,061 unigenes. Most of them were mononucleotide repeats (1,005), followed by trinucleotide repeats

Table 3: Differentially expressed genes and transcription factors related to carotenoid biosynthesis between Weiduoli and HVB-3.

Gene ID	Log2 fold-change	FDR	Blast annotation
Genes			
c27179.graph_c0	2.28	0	Geranylgeranyl pyrophosphate synthase
c35889.graph_c0	−2.19	$6.71E-08$	Geranylgeranyl diphosphate reductase
c41187.graph_c4	2.29	$3.36E-11$	Dehydrodolichyl diphosphate synthase
c21698.graph_c0	−3.96	$1.68E-05$	Alcohol dehydrogenases homologous
c30364.graph_c1	2.08	0	Aldehyde dehydrogenase
c34785.graph_c0	−2.46	$9.95E-05$	Alcohol dehydrogenase
c37165.graph_c0	−2.81	$3.18E-05$	Long chain acyl-CoA synthetase
c26028.graph_c0	−1.58	$7.58E-05$	Cytochrome P450 82A2
c29728.graph_c0	2.96	0.01	Cytochrome P450 82A4
c29775.graph_c1	−2.20	0	Cytochrome P450 86B1-like
c33165.graph_c0	−2.03	0	Cytochrome p450 CYP82D47-like
c34761.graph_c0	1.77	0	Cytochrome P450 89A2
c34982.graph_c1	−2.66	$2.32E-07$	Cytochrome P450 82C4
c36936.graph_c0	−2.50	$1.63E-06$	Cytochrome p450 86B1-like
c38370.graph_c0	2.11	0	Cytochrome P450 CYP72A219
c38487.graph_c0	−1.67	0.01	Cytochrome P450 82C4
c40387.graph_c0	2.38	$7.46E-08$	Cytochrome P450 83B1
c40784.graph_c0	−1.51	0	Cytochrome P450 CYP736A12
c41229.graph_c0	−2.70	$7.30E-13$	Cytochrome P450 78A5
c41281.graph_c1	1.22	0.01	Cytochrome P450 76A2
c41491.graph_c0	1.32	0.01	Cytochrome P450 CYP72A219-like
c42321.graph_c0	3.40	0	Cytochrome P450 71A1
Transcription factors			
c39444.graph_c0	2.10	$5.87E-11$	NAC domain-containing protein
c39928.graph_c0	2.14	$1.04E-10$	NAC domain-containing protein
c41510.graph_c0	1.55	0	NAC domain-containing protein
c4748.graph_c0	3.66	$6.33E-06$	MYB-like transcription factor
c27997.graph_c0	2.18	0	Ethylene-responsive transcription factor
c32755.graph_c0	3.98	$3.64E-11$	Ethylene-responsive transcription factor
c32755.graph_c1	3.20	$5.42E-09$	Ethylene-responsive transcription factor
c34264.graph_c0	1.92	0	Ethylene-responsive transcription factor
c34616.graph_c0	1.27	0.01	Ethylene-responsive transcription factor
c35745.graph_c1	2.64	$1.63E-06$	Ethylene-responsive transcription factor
c36135.graph_c0	1.32	0	Ethylene-responsive transcription factor
c37886.graph_c1	1.41	0	Ethylene-responsive transcription factor
c39393.graph_c0	1.42	0	Ethylene-responsive transcription factor
c40060.graph_c0	1.45	0	Ethylene-responsive transcription factor
c40767.graph_c1	1.48	$8.56E-05$	Ethylene-responsive transcription factor
c41013.graph_c0	1.59	$4.20E-05$	Ethylene-responsive transcription factor
c29615.graph_c0	1.91	0	Zinc finger protein
c30636.graph_c0	1.56	0.01	Zinc finger protein
c33159.graph_c0	1.63	0.01	Zinc finger protein
c33843.graph_c0	1.91	$2.63E-05$	Zinc finger protein
c34456.graph_c0	1.43	0	Zinc finger protein
c35956.graph_c0	3.37	0	Zinc finger protein
c36278.graph_c0	1.83	0	Zinc finger protein
c36428.graph_c1	1.78	$9.34E-06$	Zinc finger protein
c33226.graph_c1	2.30	$9.80E-05$	WRKY transcription factor
c34181.graph_c0	1.59	0.01	WRKY transcription factor
c39751.graph_c0	1.36	0.01	WRKY transcription factor
c35456.graph_c0	1.71	0	bZIP transcription factor
c39175.graph_c0	1.55	0	bZIP transcription factor
c41663.graph_c1	1.95	$1.82E-08$	bZIP transcription factor
c19855.graph_c0	2.66	0	Auxin response factor

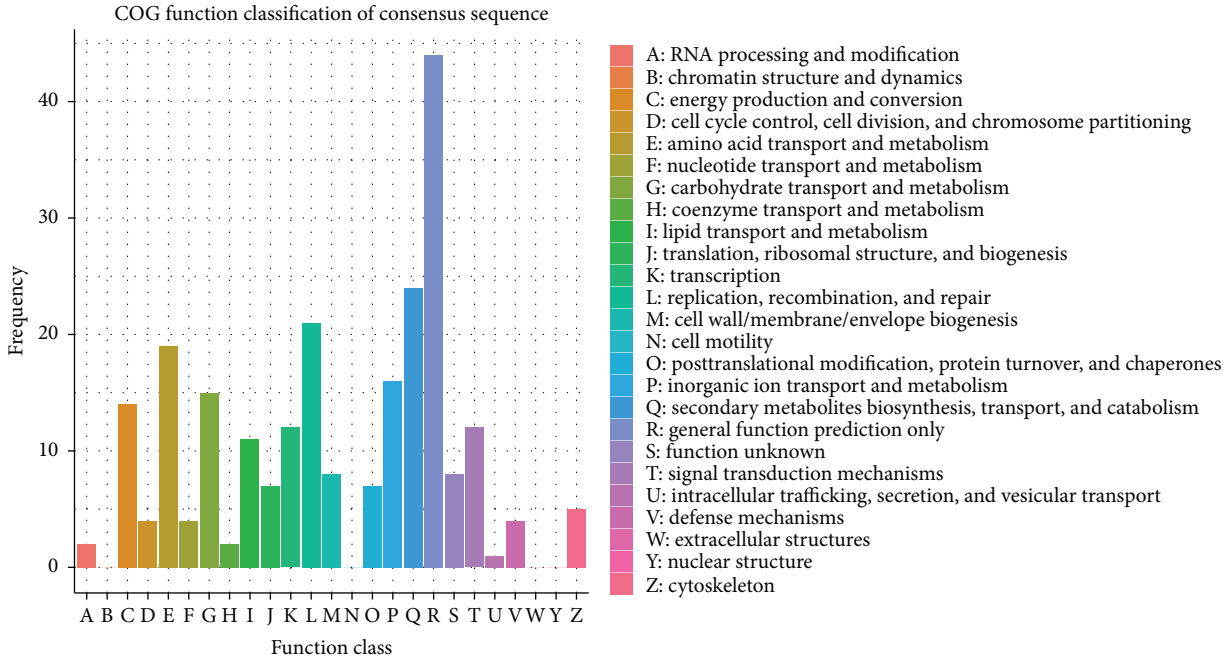


Figure 5: COG-based functional classification of DEGs between Weiduoli and HVB-3.

(388), dinucleotide repeats (298), and tetranucleotide repeats (26), with only a small portion of pentanucleotide (4) and hexanucleotide repeats (4). There were 323 sequences containing more than 1 cSSR and 125 cSSRs present in compound formation.

4. Discussion

In nonmodel plants, it is difficult to identify the candidate genes involved in complex biosynthetic pathways due to the limited availability of genomic data [28, 29]. With high-throughput transcriptome sequencing technology, this limitation has been overcome, as it can generate large amounts of data on genome wide transcription [30]. Several sweet potato transcriptomes have been sequenced, which provide an important data source for storage root formation, flower development, and anthocyanin biosynthesis of this crop [7, 18–22].

Carotenoids are widely distributed pigments in plants and play an important role as light-harvesting pigments in most photosynthetic organisms [31, 32]. In many photosynthetic and nonphotosynthetic organisms, the carotenoid biosynthesis pathway has been well studied and a series of genes involved in this pathway have been cloned and characterized. However, the molecular mechanisms regulating carotenoid synthesis have not been well understood.

The storage roots of orange-fleshed sweet potato typically have a high carotenoid content [5]. In the present study, the transcriptomes of orange-fleshed sweet potato cultivar "Weiduoli" and its high carotenoid mutant "HVB-3" were sequenced on the Illumina HiSeq 2000 sequencing platform, and 13,767,387 and 9,837,090 high-quality reads were produced from Weiduoli and HVB-3, respectively. A total of 35,909 unigenes were harvested from Weiduoli and HVB-3 (Table 1). There were 874 DEGs between HVB-3 and Weiduoli, 401 of which were upregulated and 473 were downregulated in HVB-3 compared to Weiduoli (Figure 3(b)).

The present results showed that the 22 DEGs related to carotenoid biosynthesis existed between Weiduoli and HVB-3 (Table 3). GGPS, GGPR, and DHDDS are involved in terpenoid backbone biosynthesis. GGPS is the key enzyme of carotenoid biosynthesis. Transgenic kiwifruit plants expressing GGPS exhibited the increased β-carotene content [33]. GGPR converts geranylgeranyl diphosphate (GGPP), the precursor for carotenoid biosynthesis, to phytyl diphosphate in the tocopherol and chlorophyll biosynthetic pathways. DHDDS is involved in the biosynthesis of isoprenoids, which are the precursors of carotenoid biosynthesis. Four genes encoding alcohol dehydrogenases homologous, aldehyde dehydrogenase, alcohol dehydrogenase, and long chain acyl-CoA synthetase, respectively, are involved in fatty acid metabolism. The biosynthesis of carotenoids and fatty acids requires a common precursor from pyruvate [34]. The 15 DEGs were found to encode the cytochrome P450 family (Table 3). *P450CYP707A* encoding ABA 8′-hydroxylases and *LUT1* encoding cytochrome P450-type monooxygenase (CYP97C1) have been proved to regulate carotenoid biosynthesis in *Arabidopsis* [35, 36]. Thus, it is thought that these DEGs may play important roles in carotenoid biosynthesis of sweet potato.

In the present study, several important transcription factors, including NAC, MYB, AP2/ERF, Zifc fingers, WRKY, bZIP, and ARF, were significantly upregulated in HVB-3 compared to Weiduoli (Table 3). These transcription factors may

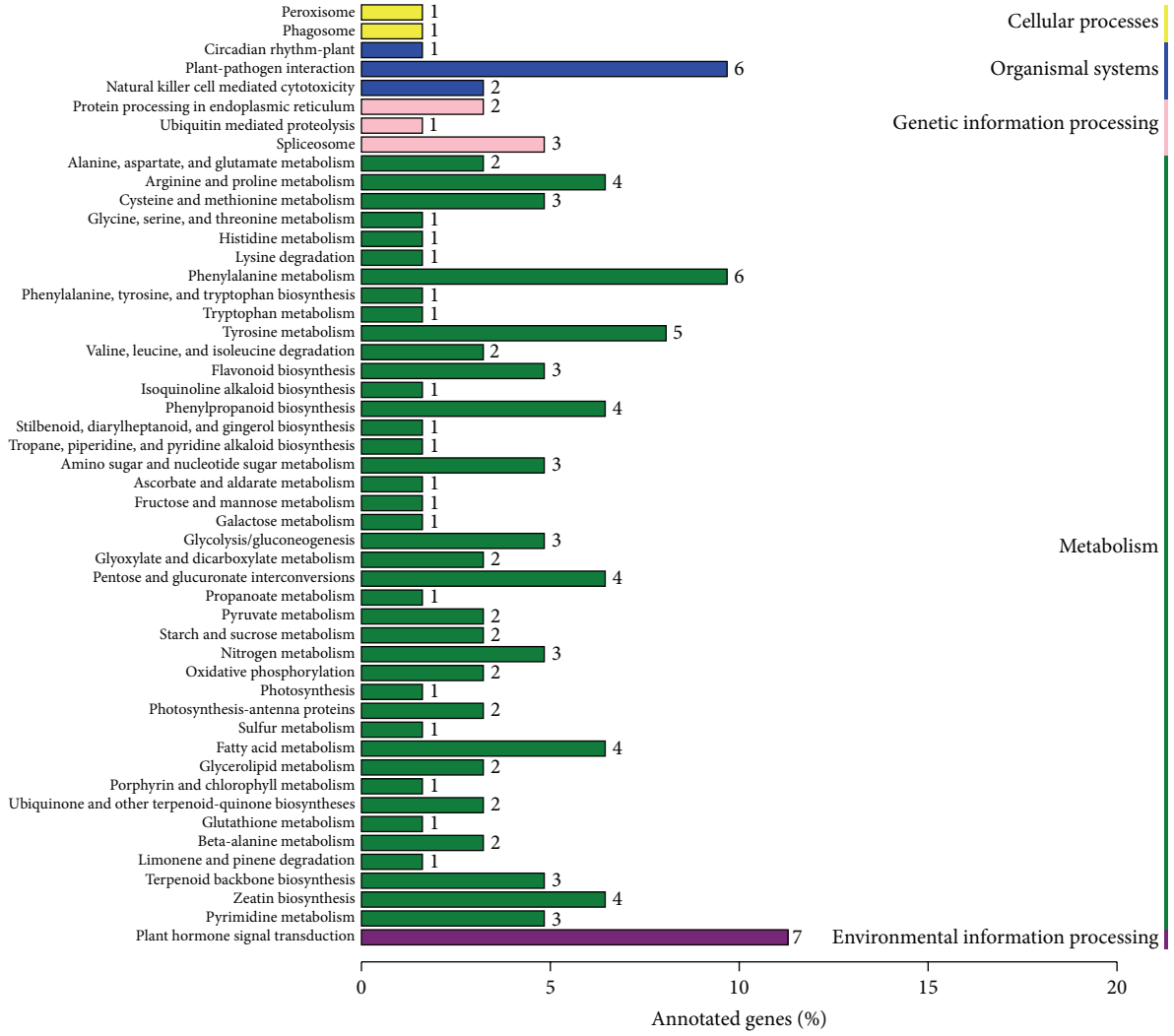


Figure 6: KEGG-based functional classification of DEGs between Weiduoli and HVB-3. Numbers beside each bar represent the actual number of DEGs classified in that descriptive term.

regulate carotenoid biosynthesis of sweet potato. In plants, transcription factors of different families have been shown to regulate secondary metabolism pathways [37]. NAC proteins are one of the largest families of plant-specific transcription factors [38]. In *Solanum lycopersicum*, a NAC transcription factor (SlNAC4) was shown to function as a positive regulator of carotenoid accumulation [39]. MYB transcription factors were found to participate in a wide range of biological processes [40, 41]. Overexpression of the *Vitis vinifera* MYB5b in tomato resulted in an increased content of β-carotene [42].

To date, the genome of sweet potato is still unavailable; therefore much of the research on sweet potato breeding and genetic linkage maps is based on molecular markers [43]. SSRs are widely distributed in both noncoding and transcribed sequences. As transferable markers, SSRs are a useful source for genome analysis due to their abundance, functionality, high polymorphism, and excellent reproducibility [44]. In the present study, a total of 1,725 potential cSSRs, including mono-, di-, tri-, tetra-, penta-, and hexa-SSR, were identified from 4,061 unigenes.

5. Conclusion

A total of 35,909 unigenes were identified from the orange-fleshed sweet potato using the high-throughput sequencing technology and most of them are protein-coding genes. There were 874 DEGs between Weiduoli and HVB-3. The 22 DEGs were found to directly or indirectly participate in carotenoid biosyntheses. The 31 important transcription factors were significantly upregulated in HVB-3 compared to Weiduoli. These DEGs and transcription factors may be involved in carotenoid biosynthesis of sweet potato.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgments

This work was supported by China Agriculture Research System (CARS-11, Sweet Potato) and the National High-Tech Research and Development Project of China (2011AA100607). The authors thank Dr. Kaitlin J. Palla, Oak Ridge National Laboratory, for her critical reading of this paper.

References

[1] Y. O. Ahn, S. H. Kim, C. Y. Kim, J.-S. Lee, S.-S. Kwak, and H.-S. Lee, "Exogenous sucrose utilization and starch biosynthesis among sweetpotato cultivars," *Carbohydrate Research*, vol. 345, no. 1, pp. 55–60, 2010.

[2] G. Padmaja, "Uses and nutritional data of sweetpotato," *The Sweetpotato*, pp. 189–234, 2009.

[3] T. Oki, M. Masuda, S. Furuta, Y. Nishida, N. Terahara, and I. Suda, "Involvement of anthocyanins and other phenolic compounds in radical-scavenging activity of purple-fleshed sweet potato cultivars," *Journal of Food Science*, vol. 67, no. 5, pp. 1752–1756, 2002.

[4] N. Zang, H. Zhai, S. Gao, W. Chen, S. Z. He, and Q. Liu, "Efficient production of transgenic plants using the bar gene for herbicide resistance in sweetpotato," *Scientia Horticulturae*, vol. 122, no. 4, pp. 649–653, 2009.

[5] M. M. Manifesto, S. M. Costa Tártara, C. M. Arizio, M. A. Alvarez, and N. R. Hompanera, "Analysis of the morphological attributes of a sweetpotato collection," *Annals of Applied Biology*, vol. 157, no. 2, pp. 273–281, 2010.

[6] L. Yu, H. Zhai, W. Chen, S.-Z. He, and Q.-C. Liu, "Cloning and functional analysis of lycopene ε-cyclase (*IbLCYe*) gene from sweetpotato, *Ipomoea batatas* (L.) Lam.," *Journal of Integrative Agriculture*, vol. 12, no. 5, pp. 773–780, 2013.

[7] X. Tao, Y.-H. Gu, H.-Y. Wang et al., "Digital gene expression analysis based on integrated *De Novo* transcriptome assembly of sweet potato [*Ipomoea batatas* (L.) Lam.]," *PLoS ONE*, vol. 7, no. 4, Article ID e36234, 2012.

[8] C. I. Cazzonelli and B. J. Pogson, "Source to sink: regulation of carotenoid biosynthesis in plants," *Trends in Plant Science*, vol. 15, no. 5, pp. 266–274, 2010.

[9] M. E. Auldridge, D. R. McCarty, and H. J. Klee, "Plant carotenoid cleavage oxygenases and their apocarotenoid products," *Current Opinion in Plant Biology*, vol. 9, no. 3, pp. 315–321, 2006.

[10] C. K. Shewmaker, J. A. Sheehy, M. Daley, S. Colburn, and D. Y. Ke, "Seed-specific overexpression of phytoene synthase: increase in carotenoids and other metabolic effects," *Plant Journal*, vol. 20, no. 4, pp. 401–412, 1999.

[11] S. Römer, P. D. Fraser, J. W. Kiano et al., "Elevation of the provitamin A content of transgenic tomato plants," *Nature Biotechnology*, vol. 18, no. 6, pp. 666–669, 2000.

[12] J. A. Paine, C. A. Shipton, S. Chaggar et al., "Improving the nutritional value of Golden Rice through increased pro-vitamin A content," *Nature Biotechnology*, vol. 23, no. 4, pp. 482–487, 2005.

[13] T. C. H. Tran, S. Al-Babili, P. Schaub, I. Potrykus, and P. Beyer, "Approaches to improve *CrtI* expression in rice endosperm for increasing the β-carotene content in golden rice," *OmonRice*, no. 13, pp. 1–11, 2005.

[14] S. H. Kim, Y. O. Ahn, M.-J. Ahn, H.-S. Lee, and S.-S. Kwak, "Down-regulation of β-carotene hydroxylase increases β-carotene and total carotenoids enhancing salt stress tolerance in transgenic cultured cells of sweetpotato," *Phytochemistry*, vol. 74, pp. 69–78, 2012.

[15] S. H. Kim, Y.-H. Kim, Y. O. Ahn et al., "Downregulation of the lycopene ε-cyclase gene increases carotenoid synthesis via the β-branch-specific pathway and enhances salt-stress tolerance in sweetpotato transgenic calli," *Physiologia Plantarum*, vol. 147, no. 4, pp. 432–442, 2013.

[16] S. H. Kim, J. C. Jeong, S. Park et al., "Down-regulation of sweetpotato lycopene β-cyclase gene enhances tolerance to abiotic stress in transgenic calli," *Molecular Biology Reports*, vol. 41, no. 12, pp. 8137–8148, 2014.

[17] S.-C. Park, S. H. Kim, S. Park et al., "Enhanced accumulation of carotenoids in sweetpotato plants overexpressing *IbOr-Ins* gene in purple-fleshed sweetpotato cultivar," *Plant Physiology and Biochemistry*, vol. 86, pp. 82–90, 2015.

[18] R. Schafleitner, L. R. Tincopa, O. Palomino et al., "A sweetpotato gene index established by de novo assembly of pyrosequencing and Sanger sequences and mining for gene-based microsatellite markers," *BMC Genomics*, vol. 11, no. 1, article 604, 2010.

[19] Z. Y. Wang, B. P. Fang, J. Y. Chen et al., "De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*)," *BMC Genomics*, vol. 11, article 726, 2010.

[20] F. L. Xie, C. E. Burklew, Y. F. Yang et al., "De novo sequencing and a comprehensive analysis of purple sweet potato (*Impomoea batatas* L.) transcriptome," *Planta*, vol. 236, no. 1, pp. 101–113, 2012.

[21] N. Firon, D. LaBonte, A. Villordon et al., "Transcriptional profiling of sweetpotato (*Ipomoea batatas*) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation," *BMC Genomics*, vol. 14, no. 1, article 460, pp. 1471–2164, 2013.

[22] X. Tao, Y.-H. Gu, Y.-S. Jiang, Y.-Z. Zhang, and H.-Y. Wang, "Transcriptome analysis to identify putative floral-specific genes and flowering regulatory-related genes of sweet potato," *Bioscience, Biotechnology and Biochemistry*, vol. 77, no. 11, pp. 2169–2174, 2013.

[23] M. G. Grabherr, B. J. Haas, M. Yassour et al., "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology*, vol. 29, no. 7, pp. U644–U652, 2011.

[24] W. J. Kent, "BLAT—the BLAST-like alignment tool," *Genome Research*, vol. 12, no. 4, pp. 656–664, 2002.

[25] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[26] C. Romualdi, S. Bortoluzzi, F. D'Alessi, and G. A. Danieli, "IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments," *Physiological Genomics*, vol. 12, no. 2, pp. 159–162, 2003.

[27] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, no. 18, pp. 3674–3676, 2005.

[28] Q. Tang, X. J. Ma, C. M. Mo et al., "An efficient approach to finding *Siraitia grosvenorii* triterpene biosynthetic genes by RNA-seq and digital gene expression analysis," *BMC Genomics*, vol. 12, article 343, 2011.

[29] K. K. Myung, B.-S. Lee, J.-G. In, H. Sun, J.-H. Yoon, and D.-C. Yang, "Comparative analysis of expressed sequence tags (ESTs) of ginseng leaf," *Plant Cell Reports*, vol. 25, no. 6, pp. 599–606, 2006.

[30] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[31] H. A. Frank and R. J. Cogdell, "Carotenoids in photosynthesis," *Photochemistry and Photobiology*, vol. 63, no. 3, pp. 257–264, 1996.

[32] G. E. Bartley and P. A. Scolnik, "Plant carotenoids: pigments for photoprotection, visual attraction, and human health," *Plant Cell*, vol. 7, no. 7, pp. 1027–1038, 1995.

[33] M. Kim, S.-C. Kim, K. J. Song et al., "Transformation of carotenoid biosynthetic genes using a micro-cross section method in kiwifruit (*Actinidia deliciosa* cv. Hayward)," *Plant Cell Reports*, vol. 29, no. 12, pp. 1339–1349, 2010.

[34] J. Schwender, M. Seemann, H. K. Lichtenthaler, and M. Rohmer, "Biosynthesis of isoprenoids (carotenoids, sterols, prenyl side-chains of chlorophylls and plastoquinone) via a novel pyruvate/glyceraldehyde 3-phosphate non-mevalonate pathway in the green alga *Scenedesmus obliquus*," *Biochemical Journal*, vol. 316, no. 1, pp. 73–80, 1996.

[35] T. Kushiro, M. Okamoto, K. Nakabayashi et al., "The *Arabidopsis* cytochrome P450 CYP707A encodes ABA 8′-hydroxylases: key enzymes in ABA catabolism," *The EMBO Journal*, vol. 23, no. 7, pp. 1647–1656, 2004.

[36] L. Tian, V. Musetti, J. Kim, M. Magallanes-Lundback, and D. DellaPenna, "The *Arabidopsis* LUT1 locus encodes a member of the cytochrome P450 family that is required for carotenoid ε-ring hydroxylation activity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 402–407, 2004.

[37] N. De Geyter, A. Gholami, S. Goormachtig, and A. Goossens, "Transcriptional machineries in jasmonate-elicited plant secondary metabolism," *Trends in Plant Science*, vol. 17, no. 6, pp. 349–359, 2012.

[38] S. Puranik, P. P. Sahu, P. S. Srivastava, and M. Prasad, "NAC proteins: regulation and role in stress tolerance," *Trends in Plant Science*, vol. 17, no. 6, pp. 369–381, 2012.

[39] M. K. Zhu, G. P. Chen, S. Zhou et al., "A new tomato NAC (NAM/ATAF1/2/CUC2) transcription factor, SlNAC4, functions as a positive regulator of fruit ripening and carotenoid accumulation," *Plant and Cell Physiology*, vol. 55, no. 1, pp. 119–135, 2014.

[40] A. M. Takos, F. W. Jaffé, S. R. Jacob, J. Bogs, S. P. Robinson, and A. R. Walker, "Light-induced expression of a *MYB* gene regulates anthocyanin biosynthesis in red apples," *Plant Physiology*, vol. 142, no. 3, pp. 1216–1232, 2006.

[41] R. V. Espley, R. P. Hellens, J. Putterill, D. E. Stevenson, S. Kutty-Amma, and A. C. Allan, "Red colouration in apple fruit is due to the activity of the MYB transcription factor, MdMYB10," *Plant Journal*, vol. 49, no. 3, pp. 414–427, 2007.

[42] A. Mahjoub, M. Hernould, J. Joubès et al., "Overexpression of a grapevine R2R3-MYB factor in tomato affects vegetative development, flower morphology and flavonoid and terpenoid metabolism," *Plant Physiology and Biochemistry*, vol. 47, no. 7, pp. 551–561, 2009.

[43] M. I. Buteler, R. L. Jarret, and D. R. LaBonte, "Sequence characterization of microsatellites in diploid and polyploid *Ipomoea*," *Theoretical and Applied Genetics*, vol. 99, no. 1-2, pp. 123–132, 1999.

[44] W. Powell, M. Morgante, C. Andre et al., "The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis," *Molecular Breeding*, vol. 2, no. 3, pp. 225–238, 1996.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 527054, 7 pages
http://dx.doi.org/10.1155/2015/527054

Research Article

Genome-Wide Identification of Genes Probably Relevant to the Uniqueness of Tea Plant (*Camellia sinensis*) and Its Cultivars

Yan Wei, Wang Jing, Zhou Youxiang, Zhao Mingming, Gong Yan, Ding Hua, Peng Lijun, and Hu Dingjin

Institute of Quality Standard and Testing Technology for Agro-Products, Hubei Academy of Agricultural Sciences, Wuhan 430064, China

Correspondence should be addressed to Hu Dingjin; djhu@hbaas.com

Received 11 June 2015; Revised 21 August 2015; Accepted 23 August 2015

Academic Editor: Yanbin Yin

Copyright © 2015 Yan Wei et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Tea (*Camellia sinensis*) is a popular beverage all over the world and a number of studies have focused on the genetic uniqueness of tea and its cultivars. However, molecular mechanisms underlying these phenomena are largely undefined. In this report, based on expression data available from public databases, we performed a series of analyses to identify genes probably relevant to the uniqueness of *C. sinensis* and two of its cultivars (LJ43 and ZH2). Evolutionary analyses showed that the evolutionary rates of genes involved in the pathways were not significantly different among *C. sinensis*, *C. oleifera*, and *C. azalea*. Interestingly, a number of gene families, including genes involved in the pathways synthesizing iconic secondary metabolites of tea plant, were significantly upregulated, expressed in *C. sinensis* (LJ43) when compared to *C. azalea*, and this may partially explain its higher content of flavonoid, theanine, and caffeine. Further investigation showed that nonsynonymous mutations may partially contribute to the differences between the two cultivars of *C. sinensis*, such as the chlorina and higher contents of amino acids in ZH2. Genes identified as candidates are probably relevant to the uniqueness of *C. sinensis* and its cultivars should be good candidates for subsequent functional analyses and marker-assisted breeding.

1. Introduction

Tea (*Camellia sinensis*) is one of the most popular beverages in the world. It belongs to the genus *Camellia* (Theaceae: Ericales) and originated from East Asia [1]. The cultivation of *C. sinensis* probably started more than 2,000 years ago [1]. The extensive secondary metabolites in tea leaves, including polyphenols, theanine, and volatile oils, are good for people's health [2]. Nowadays, many cultivars of *C. sinensis*, such as Longjing 43 (LJ43), Zhonghuang 1 (ZH1), and Zhonghuang 2 (ZH2), are cultivated extensively in China. In addition to *C. sinensis*, the genus *Camellia* includes many other species of great value, such as *C. oleifera* and *C. azalea*. *C. oleifera* has been cultivated for thousands of years in China. It is a kind of small tree with multiple trunks and branches. Its seeds can be pressed to yield edible tea oil, more than 80% of which is monounsaturated fat [3]. Unlike the aforementioned two species, *C. azalea* is very precious and was discovered in Guangdong Province decades ago. This is a very beautiful plant with great ornamental value.

A transcriptome is all the transcripts expressed in one or a population of cells at a certain time. With the advent of next-generation sequencing technology, a great number of transcriptomes, especially those from nonmodel species, have been reported. Given their important economic and ornamental values, transcriptomes of the above three *Camellia* species have been reported and some of their unique characteristics were identified. Specifically, Shi et al. described the transcriptome of *C. sinensis* and identified candidate genes related to natural product pathways that are important to tea quality, such as genes involved in flavonoid, theanine, and caffeine biosynthesis pathways [4]; Wang et al. compared the biochemical and transcriptomic differences between LJ43 and ZH2 to uncover mechanisms underlying

their phenotypic differences [5]; and Xia et al. characterized the transcriptome from tender shoots, young leaves, flower buds, and flowers of *C. oleifera* and detected many genes potentially related to lipid metabolism [6]. Despite this body of work, the molecular events underlying the uniqueness of *C. sinensis* and its cultivars remain largely undefined.

Identifying coding regions harboring mutations probably relevant to the uniqueness of a species or a cultivar has attracted the interests of many biologists for decades. Generally, nonsynonymous substitutions are harmful for their carriers and will be eliminated rapidly. However, a small number of nonsynonymous substitutions will benefit their carriers and genes harboring those mutations are termed positively selected genes. When a population extends its range or is moved by human activity to a new environment with environmental factors that are different to the original one, a series of genes may be subject to positive selection and may ultimately result in a new species. Comparing the numbers of nonsynonymous (dN) and synonymous (dS) substitutions per site is often used for diagnosing the extent and direction of selection on sequence evolution, with the ratio dN/dS >1, =1, and <1 denoting positive evolution, neutral evolution, and purifying evolution, respectively [7, 8]. Church et al. demonstrated that using likelihood-based variable selection models is feasible for comparing sequence pairs [9]. In fact, many studies have reported analyses identifying positively selected genes for several different species [10–12]. It seems highly likely that a number of genes may be subject to positive selection during the speciation of *C. sinensis*. Moreover, during the cultivation of *C. sinensis*, many important agronomical genes may be subject to artificial selection and may therefore result in a new cultivar.

In this report, based on the transcriptomes reported for the three species of the genus *Camellia*, we identified genes potentially relevant to the uniqueness of *C. sinensis*. Furthermore, candidate genes probably relevant to the divergence of LJ43 and ZH2 were also detected. Our results should be important for understanding the uniqueness of *C. sinensis* and its cultivars and provide hints for subsequent breeding.

2. Materials and Methods

2.1. Data Acquisition and Filtering. Paired-end Illumina short reads generated for the floral bud transcriptome of *C. azalea* (PRJNA257896) [13] and 454 reads for transcriptomes of tender shoots, leaves, flowers, and flower buds of *C. oleifera* [6] generated by Xia et al. (PRJNA239933), along with 454 reads of the new shoot transcriptome generated by Wang et al. for four *C. sinensis* strains (PRJNA223181) [14], were downloaded from the Sequence Read Archive at the National Center for Biotechnology Information (SRA, http://www.ncbi.nlm.nih.gov/sra/). In addition, paired-end Illumina short reads of the transcriptome of the two cultivars of *C. sinensis*, LJ43 (PRJNA261659, PRJNA240661, and PRJNA79643) and ZH2 (PRJNA261659), were downloaded from SRA.

Quality controls were implemented using NGS QC Toolkit_v2.3.3 with default settings [15]. Low quality bases that reside in short reads generated using Illumina and Roche 454 platforms were filtered by scripts included in this package. Single-end 454 reads were further processed using Seqclean (http://sourceforge.net/projects/seqclean/files/latest/download) to trim the vector sequences included in UniVec (ftp://ftp.ncbi.nih.gov/pub/UniVec/).

2.2. De Novo Assembly. A *de novo* assembly based on paired-end Illumina short reads was performed using Trinity [16] with default settings for *C. azalea*. For the assembly of 454 reads of the other two species, the iAssembler package was used, which employs MIRA (http://sourceforge.net/projects/mira-assembler) and CAP3 [17] and can assemble large-scale ESTs into consensus sequences with significantly higher accuracy [18]. For each species, contigs shorter than 200 bp were discarded in the ensuing analyses. For the *de novo* transcriptome assembly for each species, TransDecoder (http://sourceforge.net/projects/transdecoder/) was used for predicting the probable open reading frames.

2.3. Identification of Orthologous Genes and Alignment. Tentative orthologs among the three species were predicted using a transitive Reciprocal Best Hits (RBH) approach implemented in the Ortholuge pipeline [19] with default settings except the *e*-value for blastn being set to $1e-9$. For each ortholog group, we compared each nucleotide sequence with the corresponding protein sequence predicted for *C. oleifera* using Genewise [20] and used a customized Perl script to extract the matched coding regions and generate the proper alignment format for the subsequent PAML [8] analyses. We excluded alignments with premature stop codons or those shorter than 30 codons after deleting the gaps.

2.4. Evolutionary Analyses. The evolutionary rates (K_a, K_s, and K_a/K_s) for each ortholog group and each separate species in the genus *Camellia* were calculated using CODEML under the branch-free model (model = *b*). To test for selection acting on *C. sinensis*, we used CODEML's branch-site models (model = 2, NSsites = 2), which allowed dN/dS to vary among codon sites and across branches of the phylogeny. By setting the branch leading to *C. sinensis* as the foreground branch, we compared a selection model that allowed a class of codons on that branch to have $dN/dS > 1$ (Model A2, fix_omega = 0, omega = 1.5) with a neutral model that constrained this additional class of sites to have $dN/dS = 1$ (Model A1, fix_omega = 1, omega = 1). A likelihood ratio test (LRT) with χ^2 approximation was used to determine the significance of difference between the nested models. All these analyses were implemented twice with different starting omega values to check for convergence. Alignments of positive results were checked and adjusted manually and subject to analyses once again to reduce the possibility of false positives.

2.5. Mutations between LJ43 and ZH2 and Experimental Verification. All Illumina paired-end reads of LJ43 and ZH2 were used for identifying mutation sites potentially relevant to their divergence. For these two cultivars, all reads were mapped back to the nonredundant transcriptome using BWA (-n 0.005 -k 5) [21], and all duplicate reads were removed

using the MarkDuplicates program from the software package Picard (http://broadinstitute.github.io/picard/). Reads with minimum mapping quality < 20 were removed using SAMtools [22], and then a synchronized file was generated using the program mpileup2sync in the software package PoPoolation2 [23]. The synchronized file listed allele frequencies for every population at every base in the nonredundant transcriptome in a concise format. Sites with base frequencies more than five in LJ43 but absent from ZH2 were identified as unique mutations and were compared with the predicted transcript structure to decide if they were located in the coding regions and changed the encoding amino acids.

Due to the fact that RNA-Seq may produce false positives, mutation sites residing in some genes and probably related to the phenotype differentiation were selected and subject to experimental verification. Briefly, total RNAs from the leaves of LJ43 and ZH2 were extracted with Trizol Reagent Kit (Invitrogen, Madison, USA) and were reverse transcribed into cDNA using the M-MLV RTase cDNA synthesis kit (Takara, Dalian, China); sequences involved in the pathway "Porphyrin and chlorophyll metabolism" harboring nonsynonymous mutations were selected and amplified using gene-specific primers (Table 1). Sequences of the amplified productions were bidirectionally sequenced using an ABI 3730 Model DNA sequencer (Shanghai Sangon, China) and submitted to Genbank.

2.6. Pathway Analysis. KEGG pathways were assigned to all the ortholog groups using the KOBAS software [24]. After that, the K_a values for each pathway with more than fifteen genes assigned were compared between *C. sinensis* and the other two species separately using the binomial test. Contigs with mutation sites that encoded different amino acids between LJ43 and ZH2 or were identified as positively selected genes in *C. sinensis* were individually subject to pathway enrichment analysis using the KOBAS software.

2.7. Differentially Expressed Gene Families between C. sinensis and C. azalea. The identification of differentially expressed genes between species is blocked by the accurate assignment of ortholog relationships, especially for nonmodel species for which whole genome sequences are unavailable. To understand the pivotal genes probably relevant to the uniqueness of *C. sinensis* such as the high content of flavonoid, theanine, and caffeine, we employed a strategy used for the identification of differentially expressed gene families cross species [25]. Briefly, the proteome of *Arabidopsis thaliana* was downloaded from Ensembl Plants (http://plants.ensembl.org/) and clustered as gene families using CD-HIT [26]. The sequence identity threshold was set to 0.6 and a representative sequence for each gene family was selected. Then, short Illumina reads for LJ43 and *C. azalea* were mapped back to the *de novo* assembled transcriptomes of *C. sinensis* and *C. azalea* separately, using bowtie [27] with default settings. After that, each sequence of the transcriptomes of the two species was uniquely mapped to the representative of each gene family, and the number of mapped short reads was accumulated if two or more sequences were mapped to the same gene family for each species. Finally, edgeR [28] was used to identify the differentially expressed gene families between the two species, and the method of Benjamini and Hochberg was used to adjust p values for multiple comparisons [29]. Changes in expression patterns of gene families including genes involved in the pathways synthesizing "flavonoid," "theanine," and "caffeine" were scrutinized [4].

3. Results and Discussion

With *de novo* assembly methods, we obtained 246,972, 103,002, and 141,099 sequences for *C. azalea*, *C. sinensis*, and *C. oleifera*, respectively. The number of sequences obtained for *C. azalea* was similar to that reported previously [13]. However, using the Newbler software, other researchers obtained 60,479 and 120,425 sequences for *C. sinensis* and *C. oleifera*, respectively [6, 14]. The significantly greater number of sequences we obtained for these two species may be because Newbler performs best for restoring full-length transcripts [30, 31], but iAssembler can identify incorrectly assembled contigs and should be more conservative [18]. Moreover, the greater number of sequences for *C. azalea* was much more than the other two species and may result from the different sequencing platforms and assemblers, since short reads for *C. azalea* were generated by using Illumina sequencing technology and the long reads for the other two species were generated by 454 pyrosequencing.

Using the transitive RBH method, 8,787 one-one-one ortholog groups were identified and 4,617 groups were selected for subsequent evolutionary analyses. Distribution of alignment length for each ortholog group reveals that most groups are shorter than 600 bp. This may be because sequences used in the current analyses were generated by RNA-Seq and have a bias for shorter sequence reads (Figure 1).

Using a likelihood method and the binomial test, we found that the evolutionary rates of genes involved in metabolic pathways were not significantly different among the three species (data not shown). The likelihood method identified a total of 97 sequences as positively selected genes in *C. sinensis*, and pathways that these sequences may participate in are shown in Supplemental Table S1 (in the Supplementary Material available online at http://dx.doi.org/10.1155/2015/527054). The pathways are mainly related to the metabolism of carbohydrates, lipids, amino acids, and some secondary metabolites, such as "glycolysis/gluconeogenesis," "steroid biosynthesis," "fatty acid degradation," "arginine and proline metabolism," "histidine metabolism," and "butanoate metabolism." However, genes participating in pathways synthesizing important secondary metabolites of the species, such as "flavonoid," "theanine," and "caffeine," were not identified in the current analyses. These results may be because positively selected genes involved in these pathways were not included in the original dataset and the conservatism of the method in the condition of few sequences [32]. Further investigations employing the transcriptomes of more species may address this issue.

Changes of expression patterns of some pivotal genes may also contribute to the uniqueness of *C. sinensis*. To test this hypothesis, differentially expressed gene families between

Table 1: Profiles of the validation of genes harboring nonsynonymous mutations in ZH2 compared with LJ43.

Serial number	Genbank accession number (LJ43)	Genbank accession number (ZH2)	Description	Mutation site	LJ43	ZH2	Confirmed or not	Forward primer	Reverse primer
UN010458	KT427374	KT427366	Uroporphyrinogen decarboxylase	1301	A	G	No	CGCAGATTTGAGACTGGTTG	TGGCCAATCAAGTGAAGATG
UN010458	KT427374	KT427366	Uroporphyrinogen decarboxylase	1310	A	G	Yes	CGCAGATTTGAGACTGGTTG	TGGCCAATCAAGTGAAGATG
UN010458	KT427374	KT427366	Uroporphyrinogen decarboxylase	1421	A	G	Yes	CGCAGATTTGAGACTGGTTG	TGGCCAATCAAGTGAAGATG
UN014682	KT427373	KT427365	Ferrochelatase 1	1745	T	C	No	ACTTTCTCCAAGGCTGCTC	CCAACACGGGTACTAACG
UN016015	KT427372	KT427364	Glutamyl-tRNA reductase 1	674	G	C	Yes	AAGGCCAACTCAACAGAAGC	AGTTGGGCAAGGAGTCACTG
UN020635	KT427371	KT427363	Cytochrome c oxidase assembly protein COX15	446	G	C	Yes	GATCCCTCCATCGTCATCAT	CATGCAGCAGAAGCAAAAAG
UN028022	KT427367	KT427359	Chlorophyllase 1	346	T	C	Yes	TTACAGGAGCAATAGTAGGTT	CCATAGTAGAGGTGGAAAGA
UN031492	KT427368	KT427360	Chlorophyllase 2	176	A	G	Yes	GAATGTAAACCGCCAAGT	CATCCAAACAAGCCCTTA
UN043046	KT427369	KT427361	Chlorophyll(ide) b reductase NOL	183	A	T	No	TCTTCTGTTGTCTGGCGCTT	TCGATTCCTCCTAGCAACCAA
UN047188	KT427370	KT427362	Glutamyl-tRNA synthetase	128	T	A	No	CAAAAGCAAAAGCACCCAACC	GAACCCACCACCATACTCGC
UN047188	KT427370	KT427362	Glutamyl-tRNA synthetase	93	G	C	Yes	CAAAAGCAAAAGCACCCAACC	GAACCCACCACCATACTCGC

Table 2: Profiles of the expression patterns of gene families (*C. sinensis* (LJ43) versus *C. azalea*) including genes involved in the pathways synthesizing flavonoid, theanine, and caffeine.

Secondary metabolic pathway	Gene name	Number of members of the family	Fold change (log 2)	*p* value	*p* adj.
Flavonoid biosynthesis	Phenylalanine ammonia lyase	4	1.73	0.01	0.04
	Cinnamate 4-hydroxylase	1	0.82	0.21	0.40
	4-Coumarate CoA ligase	3	−0.02	0.98	0.99
	Flavanone 3-hydroxylase	1	−0.06	0.92	0.97
	Flavonol synthase	3	1.98	0.003	0.02
	Leucoanthocyanidin reductase	2	−2.35	0.0006	0.004
	Glutamine synthetase	6	1.15	0.08	0.20
Theanine biosynthesis	Gamma-glutamyl transpeptidase	1	−0.20	0.76	0.87
	Alanine aminotransferase	1	2.06	0.002	0.01
	S-Adenosylmethionine decarboxylase	3	−0.10	0.88	0.94
Caffeine biosynthesis	GMP synthase	1	0.65	0.32	0.53
	5′-Nucleotidase	1	1.56	0.02	0.07
	Adenylosuccinate synthase	1	2.29	0.0008	0.005
	AMP deaminase	1	0.002	1	1
	S-Adenosylmethionine synthase	4	0.76	0.25	0.44

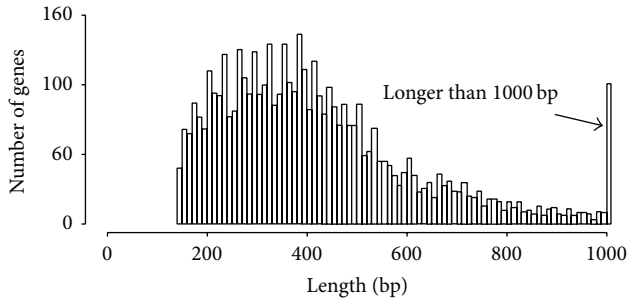


Figure 1: Length distribution of sequence alignments for the ortholog groups among *C. azalea*, *C. sinensis*, and *C. oleifera*.

C. sinensis (LJ43) and *C. azalea* were identified. Specifically, profiles of the genes that participate in the synthesis of the three important secondary metabolites in *C. sinensis* are shown in Table 2. We found that most gene families including the genes that participate in the pathways synthesizing "flavonoid," "theanine," and "caffeine" were significantly upregulated and expressed in *C. sinensis*. In contrast, the significance of gene families including those genes that were downregulated in *C. sinensis* was not so notable, except the family including the gene "leucoanthocyanidin reductase." Thus, we suggest that the uniqueness of *C. sinensis* may result from the upregulation of some pivotal genes. However, the reliability of the results needs further investigation since these results are based on the comparison of expression patterns of gene families and employing the combined expression data from different tissues.

Mutation analyses found polymorphisms present between LJ43 and ZH2 for more than 10,000 sites. Further investigations showed that 3,655 mutations were located in the coding regions and 2,021 of them were nonsynonymous mutations. Pathway analysis showed that genes harboring these nonsynonymous mutations were involved in a number of pathways. In particular, the pathways "alanine, aspartate, and glutamate metabolism," "Porphyrin and chlorophyll metabolism," "glycine, serine, and threonine metabolism," "valine, leucine, and isoleucine degradation," and "flavonoid biosynthesis" were well represented (Supplemental Table S2). Gene ontology analysis showed that genes with those nonsynonymous mutations were significantly enriched in GO terms "chloroplast part" and "chloroplast stroma" (Supplemental Table S2). To validate the mutations identified using high-throughput analyses, the sequences involved in the pathway "Porphyrin and chlorophyll metabolism" were selected and subjected to experimental verification. Bidirectional sequencing confirmed seven of the eleven nonsynonymous sites residing in these genes (Table 1). Biochemical analyses revealed that contents of free amino acids and flavonoid are different in the yellow-leaf tea cultivar ZH2 and normal cultivar LJ43 [5]. Our study suggests that nonsynonymous mutations residing in the coding regions of some genes may also take part in the formation of differences between the two cultivars, in addition to the differential expression of some other genes [5].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors express their sincere thanks to Dr. Jonathan Gardner for his excellent work on language editing for the paper. This study was supported by the Special Fund for Agroscientific Research in the Public Interest (no. 201203046) and the Natural Science Foundation of Hubei Academy of Agricultural Science (no. 2012NKYJJ18, no. 2013NKYJJ20).

References

[1] D.-W. Zhao, J.-B. Yang, S.-X. Yang, K. Kato, and J.-P. Luo, "Genetic diversity and domestication origin of tea plant *Camellia taliensis* (Theaceae) as revealed by microsatellite markers," *BMC Plant Biology*, vol. 14, article 14, 2014.

[2] P. J. Rogers, J. E. Smith, S. V. Heatherley, and C. W. Pleydell-Pearce, "Time for tea: mood, blood pressure and cognitive performance effects of caffeine and theanine administered alone and together," *Psychopharmacology*, vol. 195, no. 4, pp. 569–577, 2008.

[3] C. Dejing, F. Zili, and Z. Chenlu, "Comparison of fatty acid compositions of tea seed oil and oil—tea camellia seed oil," *China Oils and Fats*, vol. 36, no. 3, article 3, 2011.

[4] C.-Y. Shi, H. Yang, C.-L. Wei et al., "Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds," *BMC Genomics*, vol. 12, article 131, 2011.

[5] L. Wang, C. Yue, H. L. Cao et al., "Biochemical and transcriptome analyses of a novel chlorophyll-deficient chlorina tea plant cultivar," *BMC Plant Biology*, vol. 14, article 352, 2014.

[6] E. H. Xia, J. J. Jiang, H. Huang, L. P. Zhang, H. B. Zhang, and L. Z. Gao, "Transcriptome analysis of the oil-rich tea plant, *Camellia oleifera*, reveals candidate genes related to lipid metabolism," *PLoS ONE*, vol. 9, no. 8, Article ID e104150, 2014.

[7] L. D. Hurst, "The Ka/Ks ratio: diagnosing the form of sequence evolution," *Trends in Genetics*, vol. 18, no. 9, pp. 486–487, 2002.

[8] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.

[9] S. A. Church, K. Livingstone, Z. Lai et al., "Using variable rate models to identify genes under selection in sequence pairs: their validity and limitations for EST sequences," *Journal of Molecular Evolution*, vol. 64, no. 2, pp. 171–180, 2007.

[10] A. G. Clark, S. Glanowski, R. Nielsen et al., "Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios," *Science*, vol. 302, no. 5652, pp. 1960–1963, 2003.

[11] D. T. Gerrard and A. Meyer, "Positive selection and gene conversion in SPP120, a fertilization-related gene, during the East African cichlid fish radiation," *Molecular Biology and Evolution*, vol. 24, no. 10, pp. 2286–2297, 2007.

[12] C. Kosiol, T. Vinař, R. R. da Fonseca et al., "Patterns of positive selection in six mammalian genomes," *PLoS Genetics*, vol. 4, no. 8, Article ID e1000144, 2008.

[13] Z. Q. Fan, J. Y. Li, X. L. Li et al., "Genome-wide transcriptome profiling provides insights into floral bud development of

summer-flowering *Camellia azalea*," *Scientific Reports*, vol. 5, article 9729, 2015.

[14] L. Wang, X. C. Wang, C. Yue, H. L. Cao, Y. H. Zhou, and Y. J. Yang, "Development of a 44 K custom oligo microarray using 454 pyrosequencing data for large-scale gene expression analysis of *Camellia sinensis*," *Scientia Horticulturae*, vol. 174, no. 1, pp. 133–141, 2014.

[15] R. K. Patel and M. Jain, "NGS QC Toolkit: a toolkit for quality control of next generation sequencing data," *PLoS ONE*, vol. 7, no. 2, Article ID e30619, 2012.

[16] M. G. Grabherr, B. J. Haas, M. Yassour et al., "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, 2011.

[17] X. Huang and A. Madan, "CAP3: a DNA sequence assembly program," *Genome Research*, vol. 9, no. 9, pp. 868–877, 1999.

[18] Y. Zheng, L. Zhao, J. Gao, and Z. Fei, "IAssembler: a package for de novo assembly of Roche-454/Sanger transcriptome sequences," *BMC Bioinformatics*, vol. 12, article 453, 2011.

[19] D. L. Fulton, Y. Y. Li, M. R. Laird, B. G. S. Horsman, F. M. Roche, and F. S. L. Brinkman, "Improving the specificity of high-throughput ortholog prediction," *BMC Bioinformatics*, vol. 7, no. 1, article 270, 2006.

[20] E. Birney, M. Clamp, and R. Durbin, "Genewise and genomewise," *Genome Research*, vol. 14, no. 5, pp. 988–995, 2004.

[21] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[22] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[23] R. Kofler, R. V. Pandey, and C. Schlötterer, "PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq)," *Bioinformatics*, vol. 27, no. 24, Article ID btr589, pp. 3435–3436, 2011.

[24] C. Xie, X. Z. Mao, J. J. Huang et al., "KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases," *Nucleic Acids Research*, vol. 39, no. 2, pp. W316–W322, 2011.

[25] Z. Chen, C.-H. C. Cheng, J. Zhang et al., "Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 35, pp. 12944–12949, 2008.

[26] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[27] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, article R25, 2009.

[28] Z. Dai, J. M. Sheridan, L. J. Gearing et al., "edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens," *F1000Research*, vol. 3, article 95, 2014.

[29] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 57, no. 1, pp. 289–300, 1995.

[30] S.-H. Niu, Z.-X. Li, H.-W. Yuan, X.-Y. Chen, Y. Li, and W. Li, "Transcriptome characterisation of *Pinus tabuliformis* and evolution of genes in the *Pinus* phylogeny," *BMC Genomics*, vol. 14, article 263, 2013.

[31] S. Kumar and M. L. Blaxter, "Comparing de novo assemblers for 454 transcriptome data," *BMC Genomics*, vol. 11, no. 1, article 571, 2010.

[32] M. Anisimova, J. P. Bielawski, and Z. Yang, "Accuracy and power of bayes prediction of amino acid sites under positive selection," *Molecular Biology and Evolution*, vol. 19, no. 6, pp. 950–958, 2002.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 782635, 11 pages
http://dx.doi.org/10.1155/2015/782635

Research Article

Analysis of *Polygala tenuifolia* Transcriptome and Description of Secondary Metabolite Biosynthetic Pathways by Illumina Sequencing

Hongling Tian,[1] Xiaoshuang Xu,[2,3] Fusheng Zhang,[2] Yaoqin Wang,[1] Shuhong Guo,[1] Xuemei Qin,[2] and Guanhua Du[2,4]

[1]*Research Institute of Economics Crop, Shanxi Academy of Agriculture Science, Fenyang, Shanxi 032200, China*
[2]*Modern Research Center for Traditional Chinese Medicine, Shanxi University, No. 92 Wucheng Road, Taiyuan, Shanxi 030006, China*
[3]*College of Chemistry and Chemical Engineering, Shanxi University, No. 92 Wucheng Road, Taiyuan, Shanxi 030006, China*
[4]*Institute of Materia Medica, Chinese Academy of Medical Sciences, Beijing 100050, China*

Correspondence should be addressed to Fusheng Zhang; ample1007@sxu.edu.cn and Guanhua Du; dugh@imm.ac.cn

Received 22 June 2015; Accepted 3 August 2015

Academic Editor: Xiaohan Yang

Copyright © 2015 Hongling Tian et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Radix polygalae, the dried roots of *Polygala tenuifolia* and *P. sibirica*, is one of the most well-known traditional Chinese medicinal plants. Radix polygalae contains various saponins, xanthones, and oligosaccharide esters and these compounds are responsible for several pharmacological properties. To provide basic breeding information, enhance molecular biological analysis, and determine secondary metabolite biosynthetic pathways of *P. tenuifolia*, we applied Illumina sequencing technology and de novo assembly. We also applied this technique to gain an overview of *P. tenuifolia* transcriptome from samples with different years. Using Illumina sequencing, approximately 67.2% of unique sequences were annotated by basic local alignment search tool similarity searches against public sequence databases. We classified the annotated unigenes by using Nr, Nt, GO, COG, and KEGG databases compared with NCBI. We also obtained many candidates CYP450s and UGTs by the analysis of genes in the secondary metabolite biosynthetic pathways, including putative terpenoid backbone and phenylpropanoid biosynthesis pathway. With this transcriptome sequencing, future genetic and genomics studies related to the molecular mechanisms associated with the chemical composition of *P. tenuifolia* may be improved. Genes involved in the enrichment of secondary metabolite biosynthesis-related pathways could enhance the potential applications of *P. tenuifolia* in pharmaceutical industries.

1. Introduction

As a part of commonly known traditional Chinese medicinal plants, Radix polygalae (RP) can be obtained from *Polygala tenuifolia* Willd. and *P. sibirica* L., which are listed in the Chinese Pharmacopoeia (2010); RP elicits sedative, antipsychotic, expectorant, and anti-inflammatory effects. High consumption and decreased regeneration of wild resources of RP have prompted researchers to domesticate and cultivate it and a large-scale planting has been performed since early 1980s. Thus far, RP from *P. tenuifolia* has been the most commonly used variety, so *P. tenuifolia* was used in this study. *P. tenuifolia* is a perennial herb mainly distributed in northeast, north, and northwest of China. This herb contains various medicinal components, such as saponins, xanthones, and oligosaccharide esters, which exhibit tonic, sedative, antipsychotic, expectorant, and other pharmacological effects [1]. In previous studies, chemical and pharmacological properties of triterpenoid saponins, one of the most important medicinal components of *P. tenuifolia*, were investigated [2]. However, studies on the biosynthetic pathway of triterpenoid saponins in *P. tenuifolia* have been rarely conducted compared with those in *licorice*, *ginseng*, *notoginseng*, *quinquefolius*, and *salvia* [3–7]. For instance, fourteen nucleotide sequences have been deposited in the NCBI GenBank database until August 2015; among these sequences, six are related to the biosynthetic pathway of triterpenoid saponins in *P. tenuifolia*. The lack of sequence data has limited extensive and intensive

studies on *P. tenuifolia*; nevertheless, genomic research related to this medicinal plant is feasible. With the medicinal and economic importance of *P. tenuifolia*, genomic data sources of this plant species should be investigated to discover genes and develop further functional studies.

DNA sequencing technology was developed by Frederick Sanger in 1977. The "first-generation sequencing" was laborious and costly. After years of improvement, next-generation sequencing (NGS) has been developed in terms of speed and accuracy with reduced cost and manpower. Typical examples of NGS include SOLiD/Ion Torrent PGM from Life Sciences, Genome Analyzer/HiSeq 2000/MiSeq from Illumina, and GS FLX Titanium/GS Junior from Roche [8]. Among these techniques, Illumina HiSeq 2000 yields the largest output and entails the lowest reagent cost; this technique has been commonly used in deep sequencing of model and nonmodel organisms [9]. For example, Illumina HiSeq 2000 has been successfully applied in a wide range of species, including model plants (*Arabidopsis*, rice, tomato) [10–12], and crops with large genomes, including field pea, castor, and *Sorghum propinquum* [13–15]. This technique could promote studies on not only the genomics of these plants but also the biosynthetic pathways of the main active ingredients.

Triterpenoid saponins, xanthones, and oligosaccharide esters are currently considered as the main active ingredients of *P. tenuifolia* [16]. However, small amounts of secondary metabolites, such as isoquinoline alkaloids, lignins, and flavonols, are present in *P. tenuifolia*. Indeed, Illumina sequencing should be applied to determine the number of enzymes implicated in the biosynthetic pathway of secondary metabolites in *P. tenuifolia*. To date, the biosynthetic pathway of triterpenoid saponins, particularly pentacyclic triterpenoid saponins in *P. tenuifolia*, has been extensively studied [2]. The formation of a primary cucurbitane skeleton via the isoprenoid pathway is before the cyclization of 2,3-oxidosqualene of terpenoid biosynthesis in plant. Mevalonate acid (MVA) pathway located in cytoplasm and endoplasmic reticulum, and 2-C-methyl-D-erythritol-4-phosphate/1-deoxy-D-xylulose-5-phosphate (MEP/DOXP) pathway located in plasmid, which are likely to participate in the formation of a presenegenin skeleton in *P. tenuifolia*. These two pathways are present in the different cellular parts of green plants [17, 18]. In addition, phenylpropanoid biosynthetic pathways in *P. tenuifolia* are yet to be reported. Phenylalanine is an end product of the shikimate pathway, in which the aromatic amino acids tyrosine and tryptophan are also produced [19]. With phenylalanine as a precursor, lignin biosynthesis proceeds via a series of side-chain modifications, ring hydroxylations, and *O*-methylations to yield lignin monomers [20]. In this pathway, other small molecules, such as flavonoids, coumarins, hydroxycinnamic acid conjugates, and lignans, are also produced. Many of these molecules could be or have been the main focus of studies [21]. These two metabolic pathways are mainly found in *P. tenuifolia*. Therefore, whole transcripts for complete gene expression profiling should be identified by transcriptome sequencing because of limited genomic sources and information regarding the biosynthetic pathways of secondary metabolites in *P. tenuifolia*. It will support further studies on these two metabolic pathways by using high-throughput Illumina deep-sequencing techniques.

In this study, a cDNA library was constructed to obtain detailed and general data from *P. tenuifolia* by using a high-throughput Illumina deep-sequencing technique. We could determine genes encoding the enzymes involved in whole biosynthetic pathways of triterpenoid saponins, phenylpropanoids, and other secondary metabolites, and results could promote further analysis. Furthermore, our results could provide direct experimental data of *P. tenuifolia* not only for guiding genetic breeding but also for subsequent triterpenoid biosynthesis cloning and functional verification of key genes. Therefore, this transcriptome sequencing may help improve future genetics and genomics studies on molecular mechanisms associated with the chemical compositions of *P. tenuifolia*. Eventually, we can control the quality of *P. tenuifolia* from germplasm sources and thus provide important theoretical and practical significance.

2. Materials and Methods

2.1. Plant Materials. Fresh roots of one-, two-, and three-year-old *P. tenuifolia* were sampled and collected in Xinjiang County, Shanxi Province, China. Another batch of two-year-old fresh roots of *P. tenuifolia* was collected in Anguo City, Hebei Province. All of the samples were intact and healthy. The samples were snap-frozen in liquid nitrogen and stored at −80°C before use.

2.2. RNA Extraction, cDNA Library Construction, and Illumina Sequencing. According to the manufacturer's protocol, a plant RNA isolation kit (AutoLab) was used to extract the total RNA from fresh tissues and then purified by an RNeasy MiniElute cleanup kit (Qiagen). In brief, mRNA was purified with the combination of oligo-dT-containing beads and poly-A-containing mRNA and then used methods of chemicals and high temperature to fragmentate the mRNA and get fragments of 150 bp. The next step was to synthesize the double-stranded cDNA and repair the DNA fragments with protruding terminal through the function of 3′-5′ exonuclease and polymerase. In order to guarantee DNA fragments and adapters can combine with "A" and "T" complementary pair connection and prevent the inserted DNA fragments connecting with each other, single-base "A" was imported to the 3′ end of the blunt DNA and single-base "T" was imported to the 3′ end of the adapters. Adapters of the tag-containing and the DNA fragments were incubated and connected by the function of ligase. The free adapters and nonadapters DNA fragments were purified by using AMPureXP beads. DNA fragments with adapters were enriched selectively by PCR, and the PCR amplification with fewer number of cycles was operated in order to avoid errors in the library.

The methods of quantitative of library were Pico green and fluorescence spectrophotometer (Quantifluor-ST fluorometer, Promega, E6090; Quant-iT PicoGreen dsDNA Assay Kit, Invitrogen, P7589), and the quality control of the enrichment of the PCR fragments and validation of the size and distribution of DNA fragments in the library

Table 1: Functional annotation of *P. tenuifolia*.

Database	Total		*E* cutoff	Database version
	Number of annotated	Percent (%)		
Total unigenes	115,477			
Nt	69,211	59.9	1.00*e* − 05	201301
Nr	77,599	67.2	1.00*e* − 05	201301
Swissprot	54,582	47.3	1.00*e* − 10	201301
COG	30,548	26.5	1.00*e* − 10	No version
KEGG	72,636	62.9	1.00*e* − 10	Release58
Interpro	49,873	43.19	Interproscan 4.8	v36
GO	41,823	36.22		

were conducted using Agilent 2100 (Agilent 2100 Bioanalyzer, Agilent, 2100; Agilent High Sensitivity DNA Kit, Agilent, 5067-4626). The samples mixed after homogenization to 10 nM, then diluted gradually and quantitated to 4~5 pM for Illumina sequencing to construct the multiplexed DNA libraries. Illumina sequencing was conducted with Illumina HiSeqTM2000 in CapitalBio, Ltd., Beijing, China.

2.3. De Novo Assembly and Functional Annotation. Sequencing of the original data sequence through quality analysis after removal of low quality and subsequences gets the sequences available for subsequent analysis. A Perl program was written to remove reads with adapters, the reads with the average quality of less than Q20 for its 3′ end to the 5′ end, the reads with the final length of less sequence than 50 bp, and the reads with the uncertainty bases. High-quality reads were used for de novo assembly by the Trinity software to construct unique consensus sequences. Trimmed solexa transcriptome reads were mapped onto unique consensus sequences by using Bowtie [22] (Bowtie parameters: -v3 -all -best -strata). The functional annotation of unigenes was compared with National Center for Biotechnology Information (NCBI), nonredundant protein database (Nr), nonredundant nucleotide databases (Nt), UniProt/Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes (KEGG), Clusters of Orthologous Groups (COG), Gene Ontology (GO), and Interpro databases with Basic Local Alignment Search Tool (BLAST, http://www.ncbi.nlm.nih.gov/). Unigenes were identified by comparing sequence similarity against SWISS-PROT (SWISS-PROT downloaded from European Bioinformatics Institute; ftp://ftp.ebi.ac.uk/pub/databases/swissprot/), COG [23, 24], and KEGG database [25] with BLAST at *E* values ≤ 1*e*−10. The KO information retrieved from blast results by a Perl script and the pathways between databases and unigenes were established. InterProScan [26] was used to annotate the InterPro domains. Then, the functional assignments of InterPro domains were mapped onto GO and the GO classification and tree were performed by WEGO [27].

3. Results

3.1. Sequence Assembly. The RNA extracted from all the samples was mixed for Illmina sequencing in order to get the whole range of transcript diversity. In total, 58.88 million raw reads and 11.77 gigabase pairs were sequenced with an average GC content of 43.9%, $Q \geq 20$, and no ambiguous "N" (Table S1, in Supplementary Material available online at http://dx.doi.org/10.1155/2015/782635). A total of 55,432,632 high-quality reads were assembled, and 316,703 contigs and 145,857 transcripts with the N50 values were 1,636 bp. All the transcripts were subjected to cluster and assembly analyses, yielding 39,625 unigenes with a mean length of 1,378 bp and the N50 values were 1,971 bp (Figure 1). The assembly statistics of contigs, transcripts, and unigenes were shown in Table S2.

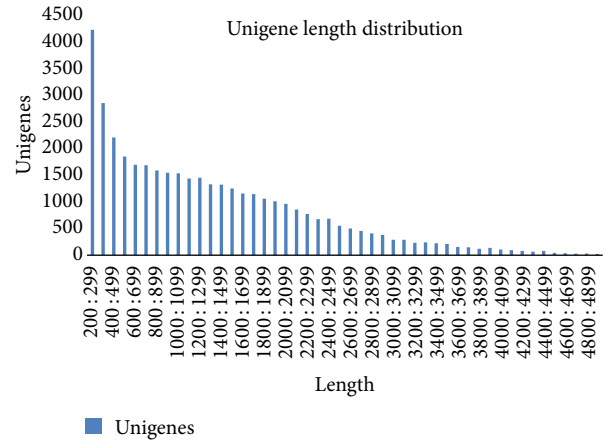


Figure 1: Length distribution of unigenes.

3.2. Functional Annotation and Classification. In the present library, approximately 67.2% of the unigenes were annotated by BLASTX and BLASTN search with a threshold of 10^{-5} against seven public databases, including Nr, Nt, UniProt/Swiss-Prot, KEGG, COG, GO, and Interpro databases (Table 1).

3.2.1. GO Annotation. A total of 41,832 unigenes for GO annotation were divided into three major categories, including cellular location, molecular function, and biological process. Among 34 functional groups with GO assignments, the dominant biochemistry, cell apoptosis, and metabolism were annotated (Figure 2).

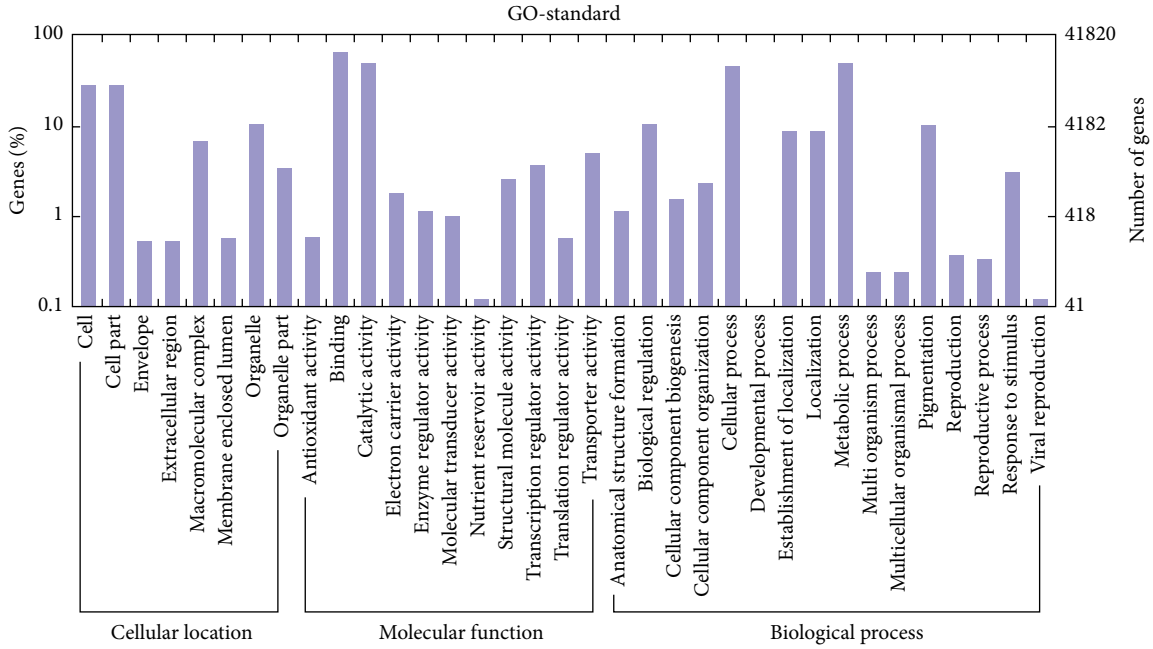


Figure 2: Gene ontology classification assigned to the unigenes.

In cellular location, eight classifications were clustered by the matched unique sequences and the larger subcategories were "cell" and "cell part." In terms of molecular function, 11 classifications were categorized, including the represented cellular components, such as "binding" and "catalytic activity." According to biological processes, 16 classifications were divided, including the represented biological processes, such as "cellular process" and "metabolic process." However, few genes were assigned to the categories "nutrient reservoir activity" and "viral reproduction," and no genes were found in the clusters of "developmental process."

3.2.2. COG Annotation. Unigenes for the COG classifications were annotated in order to further evaluate the effectiveness of the annotation process (Figure 3). Class definition, number, and percentage of this class were shown in Figure 3. The proteins in the COG categories were assumed to have the same ancestor protein, and these proteins were also assumed to be paralogs or orthologs. The largest category was general functional prediction only with 24.84%; the second categories were posttranslational modification, protein turnover, and chaperon with 8.97%; and the third category included translation, ribosomal structure, and biogenesis with 7.24%. Cell motility (24, 0.09%) and nuclear structure (10, 0.04%) represented the smallest category.

3.2.3. KEGG Annotation. The KEGG database can be used to categorize gene functions with an emphasis on biochemical pathways. The KEGG database can also be used to systematically analyze inner cell metabolic pathways and functions of gene products. Pathway-based analyses help further determine the biological functions of genes. To gain further insights into the biological pathways in *P. tenuifolia*, we performed a BLASTX search against the KEGG protein database on the assembled unigenes. A total of 72,636 (62.9%) unigenes had blast hits and 42,702 unigenes were assigned to five KEGG biochemical pathways (Table 2) including metabolism (17,217 unigenes), genetic information processing (14,593 unigenes), environmental information (2719 unigenes), cellular processes (3919 unigenes), and human diseases (4254 unigenes). The pathways with the highest unigene representation were spliceosome (ko03041, 1431 unigenes), chromosome (ko03036, 1232 unigenes), and ubiquitin system (ko04121, 1060 unigenes).

As a Chinese traditional medicinal plant, *P. tenuifolia* undergoes highly active metabolism. The higher expression of metabolism results in primary metabolite biosynthesis, such as "carbohydrate metabolism" (3702 unigenes), "amino acid metabolism" (2168 unigenes), and "energy metabolism" (1762 unigenes). However, the most active ingredients of *P. tenuifolia* were secondary metabolites. Therefore, we listed the classifications of terpenoids, polyketides, and other secondary metabolites (Figure 4) with high expression levels.

Metabolic pathways were involved in the metabolism of terpenoids and polyketides (620 unigenes; Figure 4(a)), including "terpenoid backbone biosynthesis" (176 unigenes, 28%), "carotenoid biosynthesis" (78 unigenes, 13%), "zeatin biosynthesis" (44 unigenes, 7%), "limonene and pinene degradation" (46 unigenes, 7%), "diterpenoid biosynthesis" (30 unigenes, 5%), "tetracycline biosynthesis" (16 unigenes, 3%), "brassinosteroid biosynthesis" (13 unigenes, 2%), and "polyketide sugar unit biosynthesis" (5 unigenes, 1%). The large amount of transcriptomic information may enhance the study of terpenoid biosynthesis in *P. tenuifolia*.

Metabolic pathways were also involved in the biosynthesis of other secondary metabolites (522 unigenes;

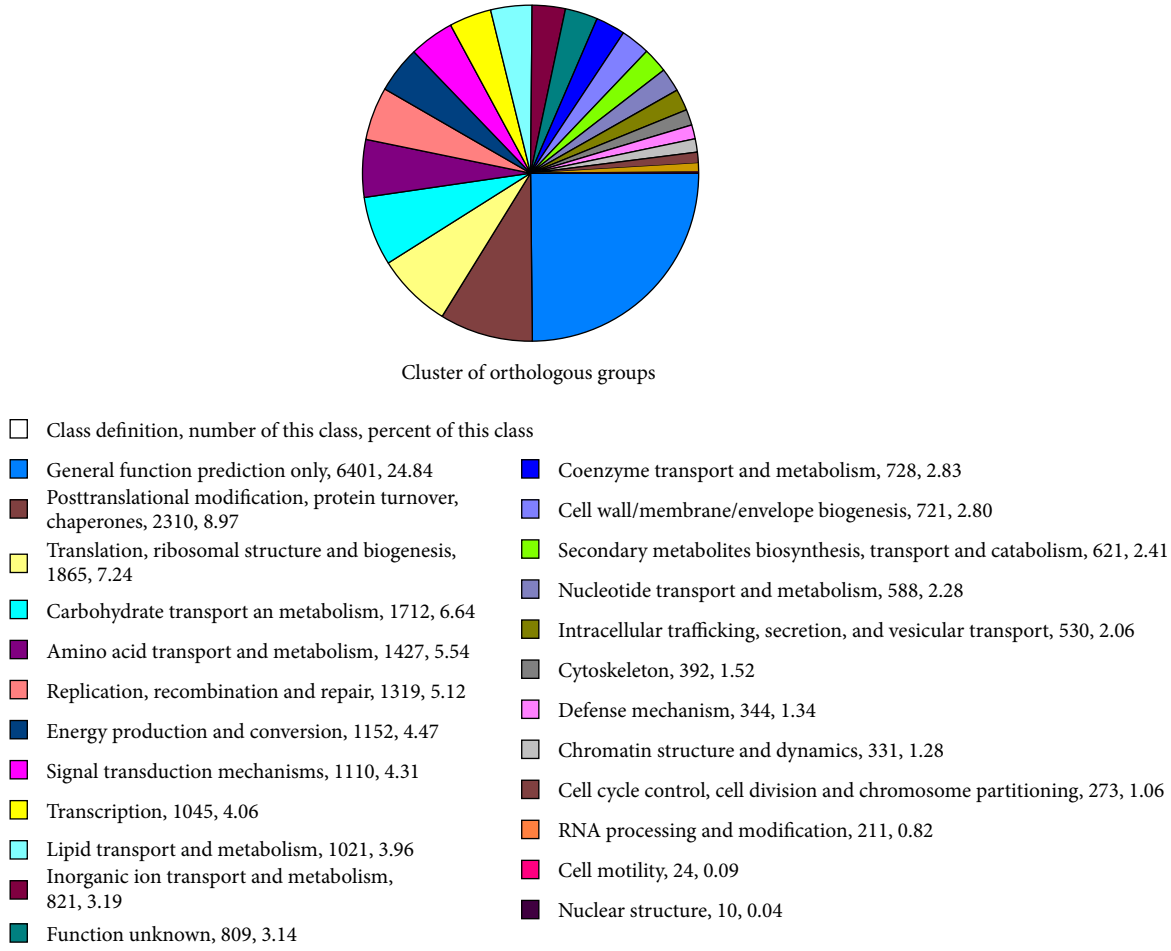


Figure 3: COG classification assigned to the unigenes.

Figure 4(b)), including "phenylpropanoid biosynthesis" (202 unigenes, 39%), "flavonoid biosynthesis" (55 unigenes, 11%), "tropane, piperidine, and pyridine alkaloid biosynthesis" (53 unigenes, 9%), "streptomycin biosynthesis" (45 unigenes, 9%), "isoquinoline alkaloid biosynthesis" (44 unigenes, 8%), "stilbenoid, diarylheptanoid, and gingerol biosynthesis" (35 unigenes, 7%), "flavone and flavonol biosynthesis" (27 unigenes, 5%), and "butirosin and neomycin biosynthesis" (25 unigenes, 5%). In our sequence dataset, a total of 176 unigenes were found to be potentially related to the biosynthesis of terpenoid biosynthesis including MEP and MVA pathways. In addition, 202 and 55 unigenes were related to the biosyntheses of phenylpropanoid and flavonoid, respectively.

(1) Characterization and Expression Analysis of the Genes Involved in the Biosynthetic Pathway of Putative Terpenoid Backbone. Dimethylallyl pyrophosphate (DMAPP) and isopentenyl pyrophosphate (IPP) are recognized as a precursor of the biosynthesis of terpenoid saponins in green plants and the compounds of "activity of isoprene" *in vivo*. DMAPP and IPP are alletic isomers, which are synthesized by MVA and MEP biosynthesis pathways mentioned above [28]. The MVA pathway starts with acetyl-CoA which is synthesized through the triterpene derivatives captured CO_2 in cytoplasm and the MEP pathway starts with pyruvate and glyceraldehydes-3-phosphate in plastid [29]. In these databases, the biosynthetic pathways of MVA and MEP were identified (Figure 5(a)). In most cases, two or more unique sequences were labeled as the same enzyme and these unique sequences could represent single gene or different fragments of gene.

The process of MVA pathway was as follows. Firstly, two molecules of acetyl-CoA were catalyzed into acetoacetyl CoA by AACT (EC 2.3.1.9, 13 unigenes) and then catalyzed into 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) by HMGS (EC 2.3.3.10, 18 unigenes). MVA was formulated by HMGR (EC 1.1.1.34, 18 unigenes) and then MVK (EC 2.7.1.36, 2 unigenes) catalyzed MVA into MVA-5-Phosphat. Next, IPP was synthesized by PMK (EC 2.7.4.2, 4 unigenes) and PMD (EC 4.1.1.33, 3 unigenes). The process of MEP pathway was as follows. Firstly, pyruvate and glyceraldehydes-3-phosphate were converted to 1-deoxy-D-xylulose-5-phosphate (DXOP) by DXS (EC 2.2.1.7, 17 unigenes) and then catalyzed into MEP by DXR (EC 1.1.1.267, 7 unigenes). Next, MCT (EC 2.7.7.60, 7 unigenes) and CMK (EC 2.7.1.148, 4 unigenes) catalyzed the MEP to 4-dicytidine triphosphate-2-methyl-D-erythritol-2-phosphate (CDP-MEP). Lastly, IPP was synthesized by MDS (EC 4.6.1.12, 1 unigene), HDS (EC 1.17.7.1, 9 unigenes), and HDR (EC 1.17.1.2, 12 unigenes). The next step was the

Table 2: Pathway classification of *P. tenuifolia*.

Category	Pathway	Count
Metabolism	Carbohydrate metabolism	3702
	Amino acid metabolism	2618
	Enzyme families	1781
	Energy metabolism	1762
	Lipid metabolism	1610
	Glycan biosynthesis and metabolism	1390
	Nucleotide metabolism	1055
	Metabolism of cofactors and vitamins	879
	Metabolism of other amino acids	716
	Metabolism of terpenoids and polyketides	620
	Xenobiotics biodegradation and metabolism	562
	Biosynthesis of other secondary metabolites	522
Genetic information processing	Folding, sorting, and degradation	4312
	Translation	4029
	Transcription	3191
	Replication and repair	3061
Environmental information processing	Signal transduction	2050
	Membrane transport	342
	Signaling molecules and interaction	327
Cellular processes	Cell growth and death	1540
	Transport and catabolism	1399
	Cell motility	507
	Cell communication	473
Human diseases	Neurodegenerative diseases	1398
	Infectious diseases	1371
	Cancers	1146
	Immune system diseases	236
	Cardiovascular diseases	103

synthesis of squalene; a series of enzymes were involved, such as IPPI (EC 5.3.3.2, 15 unigenes) and SQS (EC 2.5.1.21, 7 unigenes). However, GPPS (EC 2.5.1.29) and FPPS (EC 2.5.1.10) involved in terpenoid biosynthesis were not found from the transcriptome results maybe because some assembly cDNAs were not of full length. Squalene was catalyzed into 2,3-oxidosqualene by SE (EC 1.14.99.7, 19 unigenes). In many plants 2,3-oxisqualene cyclization was a very important step in the biosynthesis of terpenes, because it was a branch point of the synthesis of triterpenoid saponin, sterol, and other terpenes [30]. We have found that two types of OSC genes of *P. tenuifolia* are present in the dataset; these types are cycloartenol synthase (16 unigenes) and beta-amyrin synthase (54 unigenes), respectively. Moreover, except for the FPPS, another five nucleotide sequences (SQS, SE, CAS (2), cycloartenol synthase, and beta-amyrin) related to the triterpenoid saponins biosynthesis and reported in NCBI were found in our sequencing data.

(2) Characterization and Expression Analysis of the Genes Involved in the Putative Phenylpropanoid Biosynthesis Pathway. Approximately 814 unigenes, encoding 16 enzymes, were found in the phenylpropanoid biosynthesis pathway (Figure 5(b)). But flavonoids and lignins were not main compound of phenylpropanoids and researchers were rarely mentioned in *P. tenuifolia*. phenylpropanoid biosynthesis pathway starts with converting phenylalanine to Cinnamate by PAL (EC 4.3.1.24, 11 unigenes) and then Cinnamate converts to *p*-coumaryol CoA by the enzymes of C4H (EC 1.14.13.11, 10 unigenes) and 4CL (EC 6.2.1.12, 25 unigenes).

The flavonoid pathway began with the synthesis of chalcone which was catalyzed by CHS (EC 2.3.1.74, 1 unigenes) and then catalyzed to naringenin by CHI (EC 5.5.1.6, 3 unigenes). Dihydrotricetin was synthesized by F3′5′H (EC 1.14.13.88, 11 unigenes). F3H (EC 1.14.11.9, 60 unigenes) catalyzed these flavanones to dihydroflavonols. The last step was dihydroflavonols converted to flavonols by FLS (EC 1.14.11.23, 4 unigenes).

The lignin pathway began with the synthesis of *p*-coumaroyl shikimate, which was catalyzed by HCT (EC 2.3.1.133, 4 unigenes). Then, C3′H (EC 1.14.13.36) catalyzed the conversion of *p*-coumaroyl shikimate into caffeoyl shikimate. Caffeoyl shikimate was converted by hydroxycinnamoyltransferase to produce caffeoyl CoA. CCoAoMT (EC 2.1.1.104) could convert these caffeoyl CoA to feruloyl CoA. Then, CCR (EC 1.2.1.44, 6 unigenes) catalyzed the feruloyl

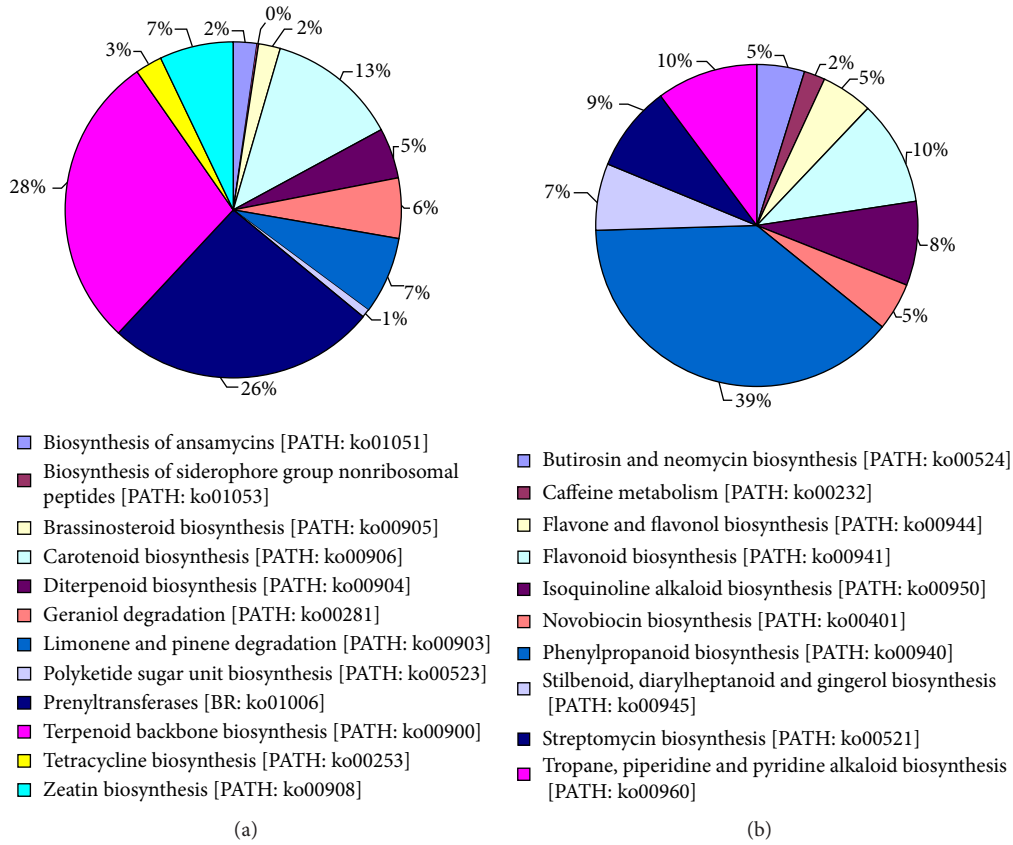


Figure 4: Classifications of unigenes involved in the metabolism of terpenoids and polyketides (a) and biosynthesis of other secondary metabolites (b).

Table 3: Frequency of identified SSR motifs in *P. tenuifolia*.

Motif length	Repeat numbers							Total	%
	5	6	7	8	9	10	>10		
Mononucleotide	—	—	—	—	—	1554	1890	3444	51.75
Dinucleotide	—	638	374	333	333	317	125	2120	31.85
Trinucleotide	646	218	117	10	1	2	2	996	14.97
Tetranucleotide	64	12	0	0	0	0	0	76	1.14
Pentanucleotide	2	1	0	0	0	0	0	3	0.045
Hexanucleotide	12	2	2	0	0	0	0	16	0.24
Total	724	871	493	343	334	1873	2017	6655	
%	10.87	13.09	7.41	5.15	5.02	4.79	1.91		

CoA into coniferaldehyde. Finally, coniferaldehyde was converted by CAD (EC 1.1.1.195, 12 unigenes) to produce coniferyl alcohol. Similar to the biosynthesis of the terpenoid backbone, C3′H was not found from the transcriptome results.

3.3. Simple Sequence Repeat (SSR) Marker Discovery. A total of 6,655 SSRs and 5784 sequences were identified from 39,625 unigenes. Among these sequences, there was more than one SSR for 755 unigene sequences. Mononucleotide accounted for 51.75% and 31.85% for dinucleotide and 14.97% for trinucleotide motify (Table 3). A/T (1539) was the most abundant repeat type, followed by AT/GA (207) and GAA/TCA (69). However, the number of tetranucleotide, pentanucleotide, and hexanucleotide motifs was <2% in the unigene sequences. In this transcriptome database, SSRs did not all exist in the unigenes and the complex SSR motifs were not detected.

4. Discussions

The advantages of high-throughput, accuracy, and reproducibility lead transcriptome sequencing to become a powerful technology [31]. Illmina sequencing 2000 has been widely

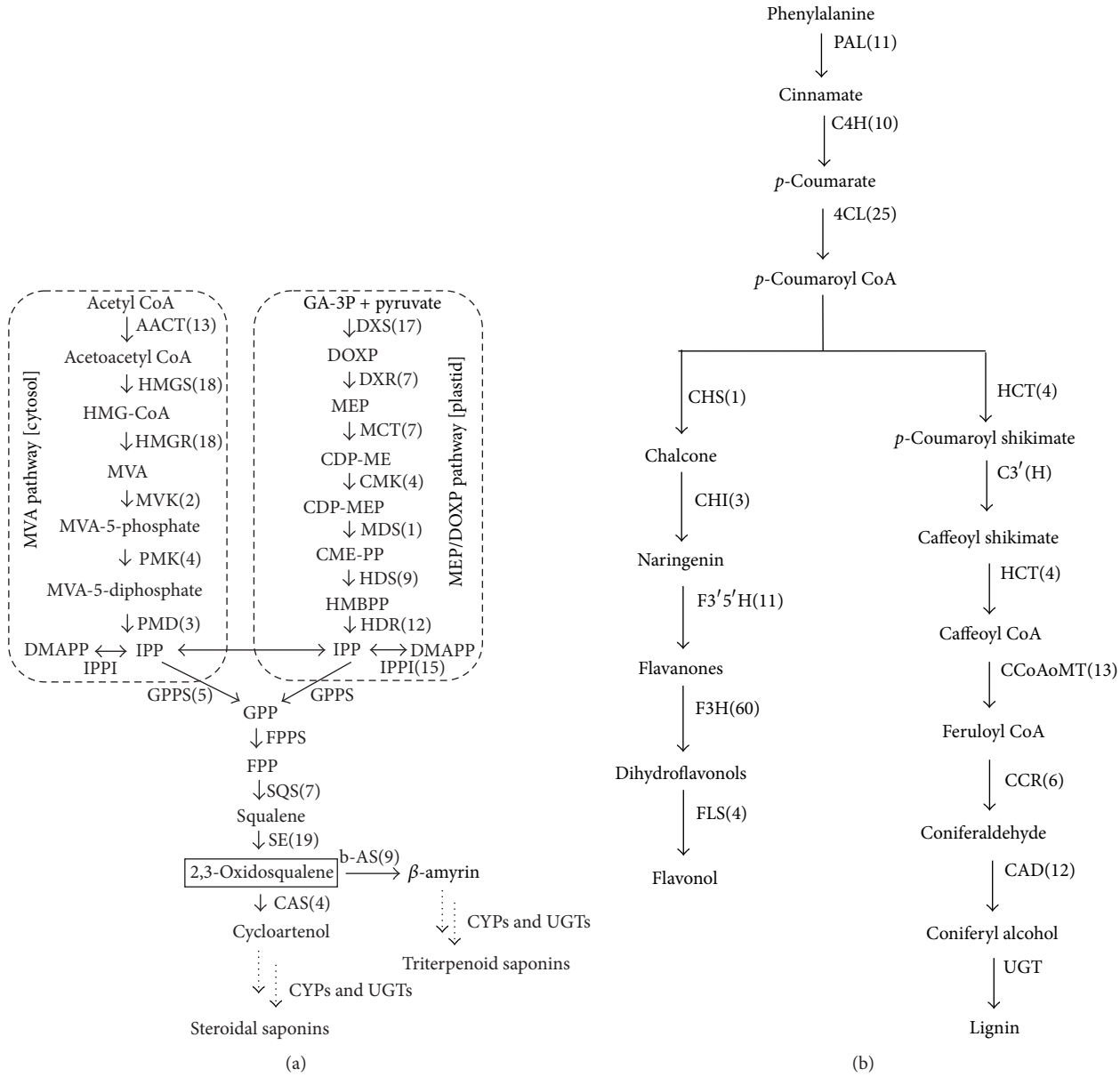


Figure 5: Schematic of the putative biosynthetic pathways of two major classes of active compounds in *P. tenuifolia*. Saponin biosynthesis pathway (a) and phenylpropanoid biosynthesis pathway (b). Enzyme abbreviations: AACT, acetyl CoA C-acetyltransferase or acetoacetyl-CoA thiolase; HMGS, 3-hydroxy-3-methylglutaryl CoA synthase; HMGR, 3-hydroxy-3-methylglutaryl CoA reductase; MVK, mevalonate kinase; PMK, phosphomevalonate kinase; PMD, mevalonate pyrophosphate decarboxylase; DXS, 1-deoxy-D-xylulose-5-phosphate synthase; DXR, 1-deoxy-D-xylulose-5-phosphate reductoisomerase; MCT, 2-C-methyl-erythritol 4-phosphate cytidylytransferase; CMK, 4-(Cytidine 5′-diphospho)-2-C-methyl-D-erythritol kinase; MDS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HDS, 1-hydroxy-2-methyl-butenyl-4-diphosphatesynthase; HDR, isopentenyl pyrophosphate (IPP)/3, 3-dimethylallyl pyrophosphate (DMAPP) synthase; IPPI, isopentenyl pyrophosphate isomerase; GPPS, geranyl diphosphate synthase; FPPS, farnesyl diphosphate synthase; SQS, squalene synthase; SE, squalene epoxidase; CAS, cycloartenol synthase; β-AS, β-amyrin synthase; DS, dammarenediol synthase; PAL, phenylalanine ammonia lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate CoA ligase; CHS, chalcone synthase; CHI, chalcone isomerase; F3′5′H, flavonoid 3′,5′-hydroxylase; F3H, flavanone 3-hydroxylase; FLS, flavonol synthase; HCT, hydroxycinnamoyl-transferase; C3′H, *p*-coumaroyl shikimate 3′ hydroxylase; CCoAoMT, caffeoyl CoA 3-*O*-methyltransferase; CCR, cinnanoyl-CoA reductase; CAD, cinnamyl alcohol dehydrogenase; CYPs, cytochrome P450s; UGTs, glycosyltransferases.

used for deep sequencing for model and nonmodel plants recently with the biggest output and lowest reagent cost. In this study, the application of Illumina HiSeq 2000 resulted in a more comprehensive understanding of genomic information and promoted the study of secondary metabolites biosynthetic pathways of *P. tenuifolia*.

Triterpenoid saponins, especially pentacyclic triterpenoid saponins, are one of the main active ingredients of *P. tenuifolia*, whose biosynthetic pathways are becoming more and more important. The upstream and midstream biosynthetic pathways of triterpenoid saponins have been studied very clearly in the past few years. Many researchers now mainly focus on the downstream biosynthetic pathways, which is also the difficult part of the studies [17]. In the downstream of biosynthesis of natural plant products, hydroxylation of CYP450s and glycosylation of UGTs play important roles in stabilizing the product and altering triterpenoid saponin bioactivity of dammarane-type and oleanane-type aglycones [32]. In this study, 466 and 157 unigenes are annotated as CYP450s and UGTs in the transcriptome of *P. tenuifolia* (data not shown), respectively. However, because of the quantity and complexity of CYP450s and UGTs, the enzymes of the downstream biosynthetic pathway of *P. tenuifolia* are still unknown, which can determine the specific steps in the accumulation of triterpenoid saponins.

CYP450s is a family of enzymes involved in the biosynthesis of lignins, terpenoids, sterols, fatty acids, hormones, pigments, and defense-related phytoalexins [33]. Only two CYP450s involved in triterpene saponin biosynthesis are functionally characterized until now, including CYP88D6 from *Glycyrrhiza uralensis* belonging to the CYP85 family [34] and CYP93E1 from *Glycine max* belonging to the CYP71 family [35]. Recently, several CYP450s and UGTs were identified as candidate genes in many studies of medicinal and nonmedicinal plants. One CYP450 (contig00248) was selected as a candidate gene, which is most likely to be involved in ginsenoside biosynthesis in *Panax quinquefolius* by MeJA-inducible and tissue-specific expression patterns [6]. Moreover, Pn02132 and Pn00158, closely related to CYP88D6 and CYP93E1, were selected as candidate CYP450s involved in triterpene saponin biosynthesis in *P. notoginseng* [5]. Furthermore, ten unigene sequences were identified corresponding to the seven different genes with a high homology to CYP79s, and four unigenes were identified corresponding to the two CYP83 genes in core structure biosynthesis of the glucosinolate metabolism of radish [36]. Our previous study once suggested that a combination of ultraperformance liquid chromatography coupled with electrospray ionization quadrupole time-of-flight mass spectrometry based on metabolomics and gene expression analysis can effectively elucidate the mechanism of biosynthesis of triterpenoid saponin, and three genes of CYP450s (CYP88D6, CYP716B1, and CYP72A1) putatively expressed in biosynthesis pathway of triterpenoid saponin in *P. tenuifolia* were identified by quantitative real-time PCR (qRT-PCR) analysis [2]. The examples mentioned above clarified that most identified CYP450s belonging to the CYP71 and CYP 88 families, and few studies were documented about the other CYP450 families until now.

UGTs are another large multigene family in plants and play an important role in the last step of biosynthesis of triterpenoid saponin. Glycosylation, the transfer of activated saccharides to an aglycone substrate, is the predominant modification in triterpenoid saponins biosynthesis. Regarding CYP450s genes, some UGT genes have also been identified. The genes UGT74M1 and UGT73K1 from *Saponaria vaccaria* [37] and UGT71G1 from *Medicago truncatula* [38] have been previously identified. Flavonoids related to UGTs (UGT78D1, UGT78D2, UGT73C6, and UGT89C1) were also identified in previous studies [39, 40]. Furthermore, four UGTs (contig01001, contig14976, contig15451, and contig16321) were selected as candidate genes in *P. quinquefolius*, which are most likely to be involved in ginsenoside biosynthesis by MeJA-inducible and tissue-specific expression patterns [6]. Pn13895, closely related to UGT71G1, was regarded as a lead candidate UGT, which is responsible for triterpene biosynthesis in *P. notoginseng* [5]. Six putative flavonol glycosides were identified in *Isatis indigotica* [41]. Thirteen unigenes were identified as UGT74s, including UGT74B1, UGT74C1, UGT74F1, and UGT74F2, in the core structure biosynthesis of the glucosinolate metabolism of radish [36]. In our previous study mentioned above [2], three genes of UGTs (UGT74B1, UGT73B2, and UGT73C6) putatively expressed in triterpenoid saponin biosynthetic pathways were also identified by qRT-PCR in *P. tenuifolia*. The examples mentioned above clarified that most identified UGTs belonged to the UGT74 and UGT73 families, and few studies have been documented about the other UGT families until now.

In other words, the genes identified in CYP450s and UGTs add little information to the knowledge of downstream triterpenoid saponin biosynthesis and thus we will continue with our in-depth studies. The large number of CYP450 and UGTs candidates could provide not only a potential gene pool for the identification of special CYP450s and UGTs involved in triterpenoid biosynthesis in *P. tenuifolia* but also a convenient method to characterize the roles in triterpenoid biosynthesis in future studies.

5. Conclusion

RP, one of the predominant Chinese medical plants, contains various medicinal ingredients with the elicits tonic, sedative, antipsychotic, expectorant, and other pharmacological effects. In this study, a large-scale unigene investigation of *P. tenuifolia* was performed by Illumina sequencing. Our results showed that many transcripts encoded by putative genes are involved in the biosynthesis of triterpene saponins and phenylpropanoid. The data we obtained presented the most abundant genomic resource and provided comprehensive information on gene discovery, transcriptome profiling, transcriptional regulation, and molecular markers of *P. tenuifolia*. This study will improve the production of active compounds through marker-assisted breeding or genetic engineering for *P. tenuifolia*, as well as other medicinal plants in the Polygalaceae family. However, many candidate CYP450s and UGTs are likely involved downstream of the biosynthetic pathway of triterpene saponins, but these candidates should be further investigated.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Hongling Tian and Xiaoshuang Xu equally contributed to this work.

Acknowledgments

This study was supported by the National Natural Sciences Foundation of China (Grant no. 31100244), the Shanxi Science and Technology Development Program (Grant no. 20140313010-1), the National Science-technology Support Plan Projects (Grant no. 2011BA107B05), the Technological Innovation Projects in Shanxi (Grant no. 2013007), and the Construction Plan for Basic Condition Platform of Shanxi (Grant no. 2014091022).

References

[1] Y. Yao, M. Jia, J.-G. Wu et al., "Anxiolytic and sedative-hypnotic activities of polygalasaponins from *Polygala tenuifolia* in mice," *Pharmaceutical Biology*, vol. 48, no. 7, pp. 801–807, 2010.

[2] F. S. Zhang, X. W. Li, Z. Y. Li et al., "UPLC/Q-TOF MS-based metabolomics and qRT-PCR in enzyme gene screening with key role in triterpenoid saponin biosynthesis of *Polygala tenuifolia*," *PLoS ONE*, vol. 9, no. 8, Article ID e105765, 2014.

[3] H. Seki, S. Sawai, K. Ohyama et al., "Triterpene functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin," *The Plant Cell*, vol. 23, no. 11, pp. 4112–4123, 2011.

[4] S. Chen, H. Luo, Y. Li et al., "454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in *Panax ginseng*," *Plant Cell Reports*, vol. 30, no. 9, pp. 1593–1601, 2011.

[5] H. Luo, C. Sun, Y. Sun et al., "Analysis of the transcriptome of *Panax notoginseng* root uncovers putative triterpene saponin-biosynthetic genes and genetic markers," *BMC Genomics*, vol. 12, no. 5, article S5, 2011.

[6] C. Sun, Y. Li, Q. Wu et al., "De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis," *BMC Genomics*, vol. 11, article 262, 2010.

[7] W. Gao, H.-X. Sun, H. Xiao et al., "Combining metabolomics and transcriptomics to characterize tanshinone biosynthesis in *Salvia miltiorrhiza*," *BMC Genomics*, vol. 15, article 73, 2014.

[8] L. Liu, Y. Li, S. Li et al., "Comparison of next-generation sequencing systems," *Journal of Biomedicine and Biotechnology*, vol. 2012, Article ID 251364, 11 pages, 2012.

[9] Y. Yang, M. Xu, Q. Luo, J. Wang, and H. Li, "De novo transcriptome analysis of *Liriodendron chinense* petals and leaves by Illumina sequencing," *Gene*, vol. 534, no. 2, pp. 155–162, 2014.

[10] S. Fowler and M. F. Thomashow, "*Arabidopsis* transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway," *Plant Cell*, vol. 14, no. 8, pp. 1675–1690, 2002.

[11] R. Zhai, Y. Feng, H. Wang et al., "Transcriptome analysis of rice root heterosis by RNA-Seq," *BMC Genomics*, vol. 14, no. 1, article 19, 2013.

[12] I. Zouari, A. Salvioli, M. Chialva et al., "From root to fruit: RNA-Seq analysis shows that arbuscular mycorrhizal symbiosis may affect tomato fruit metabolism," *BMC Genomics*, vol. 15, article 221, 2014.

[13] S. Kaur, L. W. Pembleton, N. O. I. Cogan et al., "Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers," *BMC Genomics*, vol. 13, no. 1, article 104, 2012.

[14] U. Chandrasekaran, W. Xu, and A. Liu, "Transcriptome profiling identifies ABA mediated regulatory changes towards storage filling in developing seeds of castor bean (*Ricinus communis* L.)," *Cell & Bioscience*, vol. 4, no. 1, article 33, 2014.

[15] T. Zhang, X. Zhao, W. Wang et al., "Deep transcriptome sequencing of rhizome and aerial-shoot in *Sorghum propinquum*," *Plant Molecular Biology*, vol. 84, no. 3, pp. 315–327, 2014.

[16] Y. Ikeya, S. Takeda, M. Tunakawa et al., "Cognitive improving and cerebral protective effects of acylated oligosaccharides in *Polygala tenuifolia*," *Biological and Pharmaceutical Bulletin*, vol. 27, no. 7, pp. 1081–1085, 2004.

[17] J. M. Augustin, V. Kuzina, S. B. Andersen, and S. Bak, "Molecular activities, biosynthesis and evolution of triterpenoid saponins," *Phytochemistry*, vol. 72, no. 6, pp. 435–457, 2011.

[18] Y. Wang, X. Wang, T.-H. Lee, S. Mansoor, and A. H. Paterson, "Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression in Oryza sativa (rice)," *New Phytologist*, vol. 198, no. 1, pp. 274–283, 2013.

[19] V. Tzin and G. Galili, "The biosynthetic pathways for shikimate and aromatic amino acids in *Arabidopsis thaliana*," *The Arabidopsis Book*, vol. 8, Article ID e0132, 2010.

[20] W. Boerjan, J. Ralph, and M. Baucher, "Lignin biosynthesis," *Annual Review of Plant Biology*, vol. 54, pp. 519–546, 2003.

[21] J. C. D'Auria and J. Gershenzon, "The secondary metabolism of *Arabidopsis thaliana*: growing like a weed," *Current Opinion in Plant Biology*, vol. 8, no. 3, pp. 308–316, 2005.

[22] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, article R25, 2009.

[23] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A genomic perspective on protein families," *Science*, vol. 278, no. 5338, pp. 631–637, 1997.

[24] R. L. Tatusov, N. D. Fedorova, J. D. Jackson et al., "The COG database: an updated vesion includes eukaryotes," *BMC Bioinformatics*, vol. 4, article 41, 2003.

[25] M. Kanehisa, S. Goto, M. Hattori et al., "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Research*, vol. 34, pp. D354–D357, 2006.

[26] E. M. Zdobnov and R. Apweiler, "InterProScan—an integration platform for the signature-recognition methods in InterPro," *Bioinformatics*, vol. 17, no. 9, pp. 847–848, 2001.

[27] M. L. Wise and R. Croteau, "Monoterpene biosynthesis," in *Comprehensive Natural Products Chemistry*, p. 2, Elsevier, 1998.

[28] C. A. Schuhr, T. Radykewicz, S. Sagner et al., "Quantitative assessment of crosstalk between the two isoprenoid biosynthesis pathways in plants by NMR spectroscopy," *Phytochemistry Reviews*, vol. 2, no. 1-2, pp. 3–16, 2003.

[29] W. Eisenreich, M. Schwarz, A. Cartayrade, D. Arigoni, M. H. Zenk, and A. Bacher, "The deoxyxylulose phosphate pathway of terpenoid biosynthesis in plants and microorganisms," *Chemistry and Biology*, vol. 5, no. 9, pp. R221–R233, 1998.

[30] J. D. Park, D. K. Rhee, and Y. H. Lee, "Biological activities and chemistry of saponins from *Panax ginseng* C. A. Meyer," *Phytochemistry Reviews*, vol. 4, no. 2-3, pp. 159–175, 2005.

[31] M. Rohmer, "Mevalonate-independent methylerythritol phosphate pathway for isoprenoid biosynthesis, elucidation and distribution," *Pure and Applied Chemistry*, vol. 75, no. 2-3, pp. 375–387, 2003.

[32] S. Kalra, B. L. Puniya, D. Kulshreshtha et al., "De novo transcriptome sequencing reveals important molecular networks and metabolic pathways of the plant, chlorophytum borivilianum," *PLoS ONE*, vol. 8, no. 12, Article ID e83336, 2013.

[33] A. H. Meijer, E. Souer, R. Verpoorte, and J. H. C. Hoge, "Isolation of cytochrome P450 cDNA clones from the higher plant *Catharanthus roseus* by a PCR strategy," *Plant Molecular Biology*, vol. 22, no. 2, pp. 379–383, 1993.

[34] H. Seki, K. Ohyama, S. Sawai et al., "Licorice beta-amyrin 11-oxidase, a cytochrome P450 with a key role in the biosynthesis of the triterpene sweetener glycyrrhizin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 37, pp. 14204–14209, 2008.

[35] M. Shibuya, M. Hoshino, Y. Katsube, H. Hayashi, T. Kushiro, and Y. Ebizuka, "Identification of β-amyrin and sophoradiol 24-hydroxylase by expressed sequence tag mining and functional expression assay," *FEBS Journal*, vol. 273, no. 5, pp. 948–959, 2006.

[36] Y. Wang, Y. Pan, Z. Liu et al., "De novo transcriptome sequencing of radish (*Raphanus sativus* L.) and analysis of major genes involved in glucosinolate metabolism," *BMC Genomics*, vol. 14, article 836, 2013.

[37] D. Meesapyodsuk, J. Balsevich, D. W. Reed, and P. S. Covello, "Saponin biosynthesis in *Saponaria vaccaria*. cDNAs encoding β-amyrin synthase and a triterpene carboxylic acid glucosyltransferase," *Plant Physiology*, vol. 143, no. 2, pp. 959–969, 2007.

[38] L. Achnine, D. V. Huhman, M. A. Farag, L. W. Sumner, J. W. Blount, and R. A. Dixon, "Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume *Medicago truncatula*," *Plant Journal*, vol. 41, no. 6, pp. 875–887, 2005.

[39] K. Yonekura-Sakakibara, T. Tohge, F. Matsuda et al., "Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in *Arabidopsis*," *Plant Cell*, vol. 20, no. 8, pp. 2160–2176, 2008.

[40] R. Yin, B. Messner, T. Faus-Kessler et al., "Feedback inhibition of the general phenylpropanoid and flavonol biosynthetic pathways upon a compromised flavonol-3-O-glycosylation," *Journal of Experimental Botany*, vol. 63, no. 7, pp. 2465–2478, 2012.

[41] J. Chen, X. Dong, Q. Li et al., "Biosynthesis of the active compounds of *Isatis indigotica* based on transcriptome sequencing and metabolites profiling," *BMC Genomics*, vol. 14, article 857, 2013.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 608042, 7 pages
http://dx.doi.org/10.1155/2015/608042

Research Article

PPCM: Combing Multiple Classifiers to Improve Protein-Protein Interaction Prediction

Jianzhuang Yao,[1] Hong Guo,[1] and Xiaohan Yang[2]

[1]*Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, USA*
[2]*Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*

Correspondence should be addressed to Xiaohan Yang; yangx@ornl.gov

Received 7 January 2015; Revised 22 July 2015; Accepted 26 July 2015

Academic Editor: Ian Dunham

Copyright © 2015 Jianzhuang Yao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Determining protein-protein interaction (PPI) in biological systems is of considerable importance, and prediction of PPI has become a popular research area. Although different classifiers have been developed for PPI prediction, no single classifier seems to be able to predict PPI with high confidence. We postulated that by combining individual classifiers the accuracy of PPI prediction could be improved. We developed a method called protein-protein interaction prediction classifiers merger (PPCM), and this method combines output from two PPI prediction tools, GO2PPI and Phyloprof, using Random Forests algorithm. The performance of PPCM was tested by area under the curve (AUC) using an assembled Gold Standard database that contains both positive and negative PPI pairs. Our AUC test showed that PPCM significantly improved the PPI prediction accuracy over the corresponding individual classifiers. We found that additional classifiers incorporated into PPCM could lead to further improvement in the PPI prediction accuracy. Furthermore, cross species PPCM could achieve competitive and even better prediction accuracy compared to the single species PPCM. This study established a robust pipeline for PPI prediction by integrating multiple classifiers using Random Forests algorithm. This pipeline will be useful for predicting PPI in nonmodel species.

1. Introduction

Protein-protein interaction (PPI) networks play important roles in many cellular activities, including complex formation and metabolic pathways [1], and identification of PPI pairs may provide important insights into the molecular basis of cellular processes [2]. Several high-throughput experimental approaches have been developed for PPI identification, including two-hybrid assays [3], tandem affinity purification followed by Mass Spectrometry [4], and protein microarrays [5]. These high-throughput methods have produced a large amount of PPI data, which have been accumulated in the public PPI databases, such as DIP [6] and STRING [7]. However, the results generated by these high-throughput methods may lack reliability [8] and have limited coverage of PPIs in any given organism [9]. Additional experimental information for PPI is also available, including the X-ray structures of protein complexes in the PDB databank [10]. Nevertheless, the information from protein structure complexes may be limited compared to the large volume of protein sequences available in the public databases [11].

To overcome the limitations in PPI identification using experimental methods, computational approaches have been developed to achieve large-scale PPI prediction in various organisms [12–17]. Traditional input features for PPI prediction are mainly from biological data sources, which may be divided into four categories: Gene Ontology-(GO-) based, structure-based, network topology-based, and sequence-based features [18]. Each individual computational PPI prediction method utilizes only one or few input sources for PPI prediction. For example, BIPS only takes protein sequences as input for Interolog searching [19]. Bio::Homology::InterologWalk takes protein sequences and well-known PPI networks as input [12]. Although these methods using single or several features as input can generate fairly accurate results, they are unable to take advantage of other input features that could be helpful for PPI prediction. Thus, machine learning methods (e.g., Bayesian classifiers [20], Artificial Neural Networks (ANN) [21], Support Vector Machines (SVM) [22], and Random Forests [23]) have been developed to integrate multiple features as inputs. Machine learning approaches have shown better

performances compared to some other methods; among them, Random Forests method seems to show the best performance [24]. In addition, PPI prediction is associated with imbalanced data problem. Zhang et al. [25] proved that the imbalanced data problem could be solved by ensemble methods. Augusty and Izudheen [26] further showed that Random Forests method could improve Zhang's methods in dealing with the imbalanced data problem.

In addition to the progress in identification of informative features for PPI prediction, a variety of algorithms have been developed to improve the PPI prediction accuracy [18]. For instance, Phylogenetic Profiling (PP) uses genome-scale and network-based features as inputs for PPI prediction founded on the assumption that the cooccurrence of two proteins across taxa indicates a good chance for them to function together [27, 28]. Although PPI prediction by PP has shown good performance in prokaryotes, it has poor performance in PPI prediction in eukaryotes, probably due to modularity of eukaryotic proteins, biased diversity of available genomes, and large evolutionary distances [29, 30]. Several studies indicate that the accuracy of PPI prediction by PP can be improved by selecting the appropriate reference taxa and matching the reference taxa to the known PPI network [30–32]. Recently, Simonsen et al. developed a PPI prediction software Phyloprof [33] that integrates four PPI prediction methods including the original PP method [27], mutual information (MI) method [34], hypergeometric distribution based method [35], and the extension of the hypergeometric distribution (RUN) method [36]. Also, Phyloprof provides six reference taxa optimization methods including Tree Level Filtering, Iterative Taxon Selection, Genetic Algorithm, and Tree based search [33, 37]. Furthermore, there are four PPI networks available in Phyloprof, including the networks from *Escherichia coli* (EC), *Saccharomyces cerevisiae* (hereafter referred to as SC), *Drosophila melanogaster* (DM), and *Arabidopsis thaliana* (AT). In short, Phyloprof provides a series of PPI prediction classifiers as a result of various combinations of PPI prediction methods, reference taxa optimization methods, and networks from different species.

Another sophisticated PPI prediction software called GO2PPI has been developed to use Gene Ontology and PPI networks as input [38]. By introducing a concept called inducer to combine machine learning and semantic similarity techniques, GO2PPI can provide a series of PPI prediction classifiers that are combinations of machine learning methods (i.e., Naïve Bayes (NB) and Random Forests), GO categories (i.e., biological process (BP), cellular component (CC), and molecular function (MF)), and networks from seven species (*Homo sapiens* (HS), *Mus musculus* (MM), *S. pombe* (SP), SC, AT, EC, and DM).

A variety of ensemble classifiers have been proposed in different bioinformatics studies and showed generally better performance than individual classifiers [39–41]. To build on this research, we developed a pipeline PPCM (i.e., PPI prediction classifiers merger) to enhance the PPI prediction accuracy by merging multiple PPI prediction classifiers using Random Forests algorithm. To the best of our knowledge, this study is the first effort to merge multiple classifiers (Phyloprof and GO2PPI) by machine learning for PPI prediction.

2. Methods

2.1. Construction of a Gold Standard Dataset. We created training and test dataset containing direct interacted protein pairs of yeast for protein-protein interaction (PPI) prediction using a method described by Qi et al. [24]. Briefly, 2865 positive PPI pairs were obtained from the DIP database [6]. These direct interaction protein pairs were tested to be highly confident PPI pairs by small-scale experiments. Since there was insufficient high-confidence negative data [42], negative PPI pairs were generated by randomly pairing proteins followed by removing the positive PPI pairs [43]. Finally, the positive PPI pairs and the negative PPI pairs were combined by a ratio of 1 to 100 into a "Gold Standard" dataset. It has been proved that the AUC value is not sensitive to the different positive-to-negative ratios (e.g., from 1 : 2 to 1 : 100) by both GO2PPI and Phyloprof.

2.2. Selection of Features for PPI Prediction. The results of PPI prediction classifiers were used as features of PPCM. Specifically, Phyloprof has three kinds of input parameters, including four PPI prediction methods, eight Reference Taxa Optimization methods, and four PPI networks. Without the time-consuming PPI prediction method "RUN," there were 96 different classifiers based on different combinations of parameters provided by Phyloprof (Table S2 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/608042). As mentioned above, GO2PPI has three kinds of input parameters as well, including two machine learning methods, seven GO terms or terms combinations (BP, CC, MF, BPCC, BPMF, CCMF, and BPCCMF), and seven PPI networks. In the same way, there were 98 different combinations of classifiers provided by GO2PPI (Table S1). We used combined GO terms in this study, because the best accuracy was achieved by the integration of three GO terms in the GO2PPI paper [38].

2.3. PPI Prediction Using PPCM Pipeline. The PPCM pipeline, as illustrated in Figure 1, was developed to combine multiple classifiers for enhancing PPI prediction accuracy. Specifically, a protein pair is first evaluated by classifiers provided by PPI prediction software, such as GO2PPI [38] and Phyloprof [33]. Then, the classification scores from individual classifiers are used as input features to generate the final PPI prediction score using Random Forests algorithm, implemented in the Berkeley Random Forests package [44]. GO2PPI has 98 PPI prediction classifiers, among which 14 are SC-related and 84 are not SC-related (cross species) classifiers (Table S1). Phyloprof has 96 PPI prediction classifiers, among which 24 are SC-related and 72 are not SC-related (cross species) classifiers (Table S2).

2.4. Evaluation of PPI Prediction Accuracy. The aforementioned Gold Standard database that contains about 30,000 PPI pairs with a positive-to-negative PPI ratio of 1 : 100 was used to evaluate the PPI prediction accuracy. The following measures were used to evaluate PPI prediction results: the true positive rate (TPR, also called sensitivity), defined as the ratio of correctly predicted positive PPI pairs among

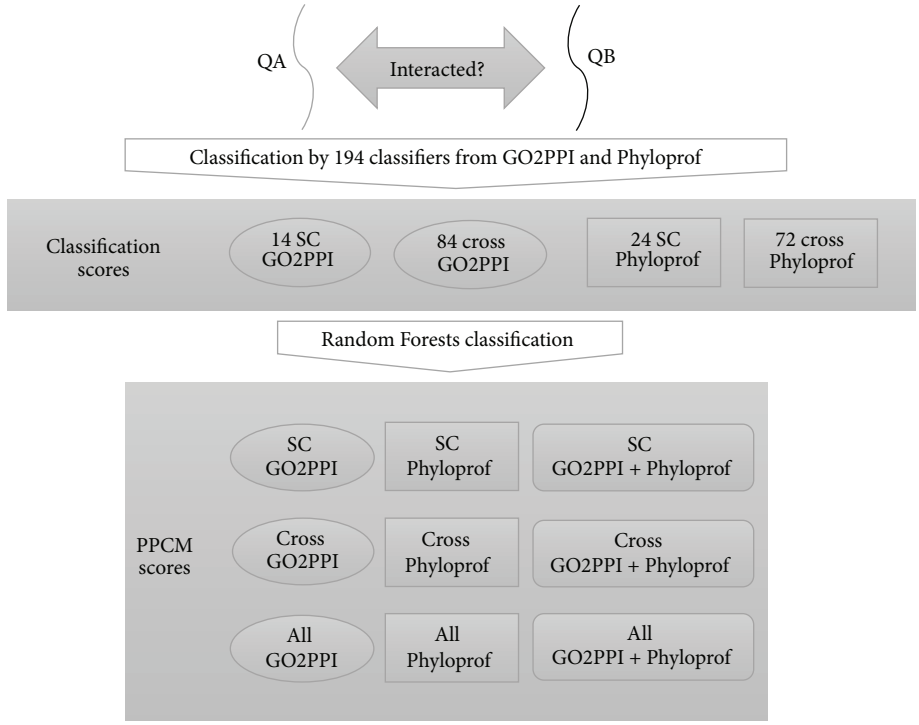


Figure 1: The PPCM pipeline for protein-protein interaction prediction. Given a pair of query proteins QA and QB, their interaction possibility was first predicted by each of the 194 classifiers from GO2PPI and Phyloprof. Then, the classification scores were merged using Random Forests algorithm to generate the final PPI prediction score. Nine PPI classification scores were provided by PPCM. "SC" represents PPI networks in *Saccharomyces cerevisiae*. "Cross" represents all PPI networks except SC. "All" represents all PPI networks in both SC and cross species.

all positive PPI pairs, the true negative rate (TNR, also called specificity), defined as the ratio of correctly predicted negative PPI pairs among all negative PPI pairs, and the false positive rate (FPR, also called Type I error), defined as the ratio of incorrectly predicted PPI pairs among all negative PPI pairs. FPR is one minus TNR. The receiver operating characteristic (ROC) curves were created by plotting TPR versus FPR. The area under the curve (AUC) was used as a measure of the prediction accuracy. The AUC value was calculated using the following equation:

$$\mathrm{AUC} = \frac{1}{2}\sum_{k=1}^{n}\left(\left(X_k - X_{k-1}\right)\left(Y_k + Y_{k-1}\right)\right), \quad (1)$$

where X_k is the FPR at k pair and Y_k is the TPR at k pair in the ranked PPI pair list. The prediction process was repeated 25 times, and the average AUC value was reported.

We evaluated the PPI prediction accuracy of PPCMs and the classifiers in GO2PPI and Phyloprof using AUC. We introduced three categories of PPCM, including GO2PPI, Phyloprof, and GO2PPI + Phyloprof, with each further divided to three subcategories: SC, cross species, and all species (i.e., SC plus cross species) (Figure 1).

3. Results and Discussion

3.1. Performance of PPCM in GO2PPI Category. Using our Gold Standard dataset, the average AUC of the 14 SC-related classifiers in GO2PPI (Table S1) was 0.63 and rf|bpcc|SC was the most accurate classifier, with an AUC of 0.64, among these 14 classifiers (Figure 2(a)). The average AUC of the 84 cross species related classifiers in GO2PPI (Table S1) was 0.57 and rf|bpcc|HS was the most accurate classifier, with an AUC of 0.61, among these 84 classifiers (Figure 2(b)). The average AUC of all the 96 (all species) classifiers in GO2PPI (Table S1) was 0.58 and rf|bpcc|SC was the most accurate classifier, with an AUC of 0.64, among these 98 classifiers (Figure 2(c)). The AUCs of PPCMs are 0.70, 0.68, and 0.70 for SC, cross species, and all species PPCM, respectively (Figure 2). These results indicate that PPCMs significantly improved PPI prediction accuracy compared with their corresponding classifiers in GO2PPI category.

Compared with the most accurate classifier in GO2PPI category, the cross species PPCM improves AUC by 11%. The improvement of PPCM in SC PPCM was only 9% (Figure 2), indicating that the cross species PPCM had better performance than the SC classifier. The better performance of cross species PPCM (containing 84 features) than SC PPCM (containing 14 features) suggests that the larger number of features incorporated into PPCM enhanced PPI prediction accuracy in GO2PPI category.

3.2. Performance of PPCM in the Phyloprof Category. Again, using our Gold Standard dataset, the average AUC of the 24 SC-related classifiers in Phyloprof (Table S2) was 0.64 and

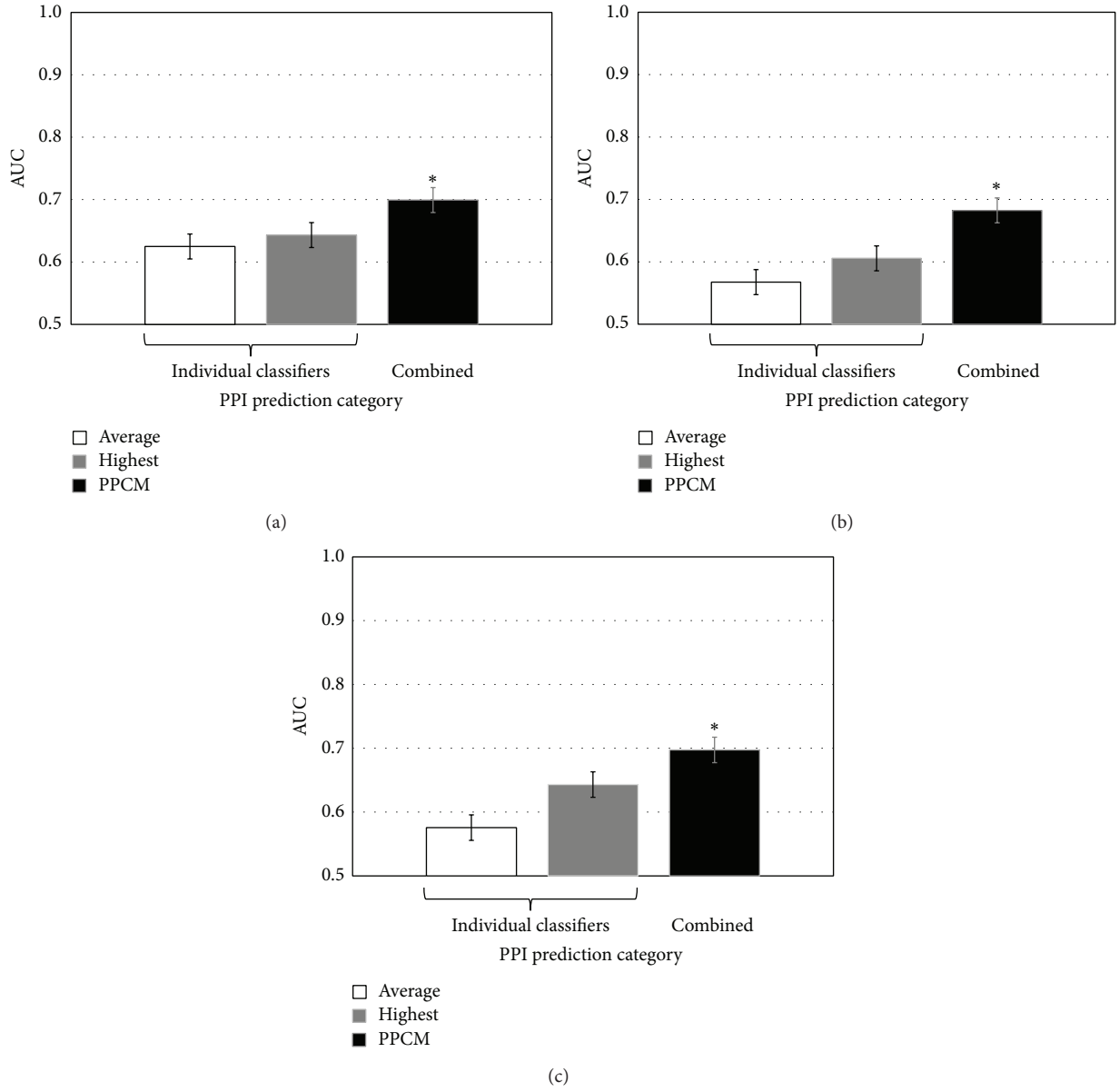


Figure 2: Comparison of PPI prediction accuracy in the GO2PPI category. (a) PPI prediction based on classifiers related to SC. (b) PPI prediction based on classifiers related to cross species. (c) PPI prediction based on classifiers related to all species. "Average" represents the mean AUC of all the classifiers in each category. "Highest" represents the classifier with highest AUC among all the classifiers in each category. Error bars show standard deviation. "∗" indicates that AUC of PPCM was significantly (P value < 0.05; t-test) higher than that of the most accurate classifier in each category.

SC|mi|et was the most accurate classifier, with an AUC of 0.71, among these 24 classifiers (Figure 3(a)). The average AUC of the 72 cross species related classifiers in Phyloprof (Table S2) was 0.61 and EC|mi|et was the most accurate classifier, with an AUC of 0.72, among these 84 classifiers (Figure 3(b)). The average AUC of all the 96 (all species) classifiers in Phyloprof (Table S2) was 0.62 and mi|et|EC was the most accurate classifier, with an AUC of 0.72, among these 96 classifiers (Figure 3(c)). The AUCs of PPCMs are 0.72, 0.76, and 0.77 for SC, cross species, and all species PPCM, respectively (Figure 3). These results indicate that PPCMs significantly improved PPI prediction accuracy compared with their corresponding classifiers in the Phyloprof category. Compared with the most accurate classifier in the Phyloprof category, the cross species PPCM improves AUC by 6%, while the improvement by SC PPCM is only 1% (Figure 3), indicating that the cross species PPCM had better performance in AUC improvement. The better performance of cross species PPCM (containing 72 features) than SC PPCM (containing 24 features) suggests that more features incorporated into PPCM could enhance PPI prediction accuracy in the Phyloprof category.

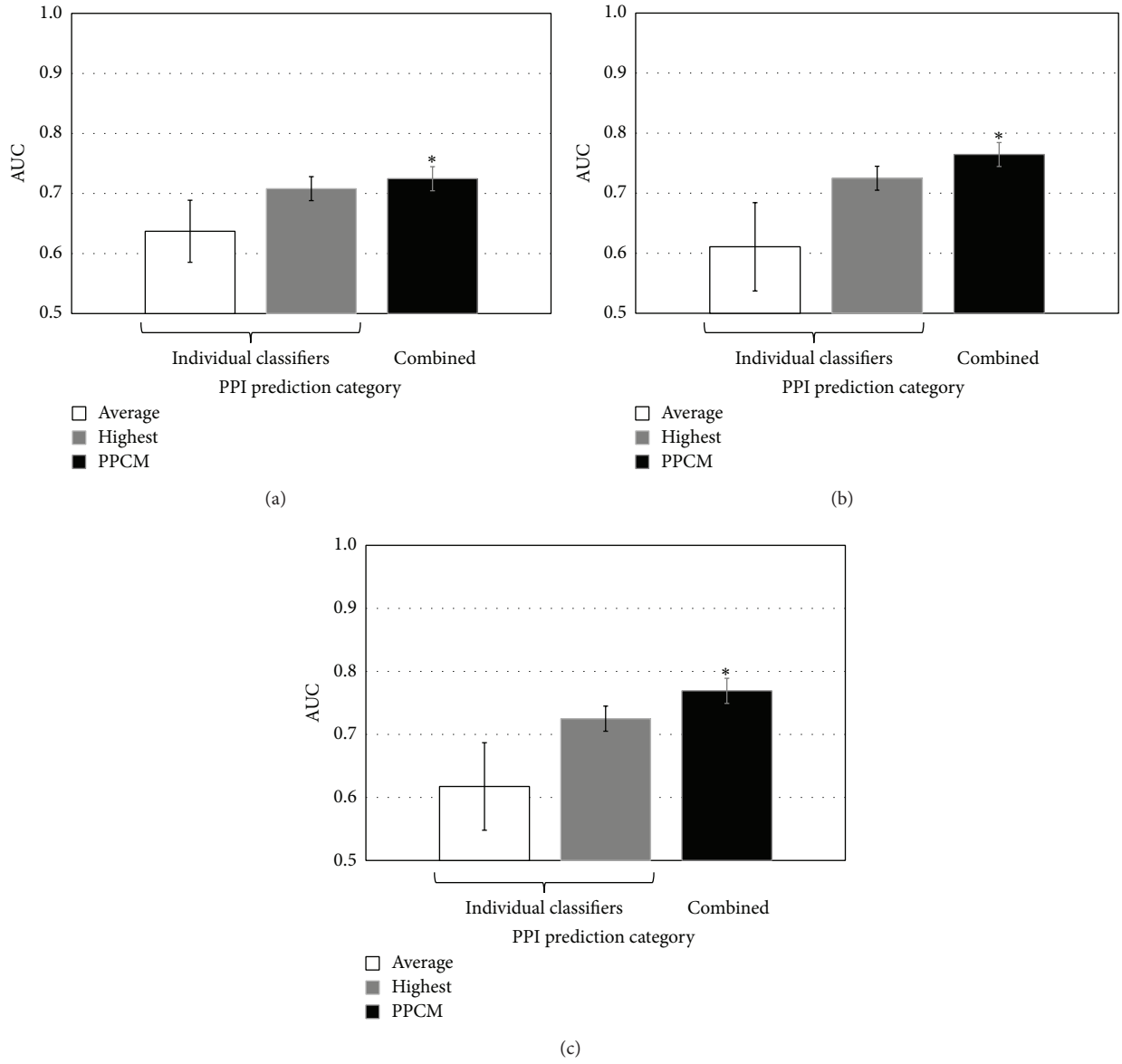


Figure 3: Comparison of PPI prediction accuracy in the Phyloprof category. (a) PPI prediction based on classifiers related to SC. (b) PPI prediction based on classifiers related to cross species. (c) PPI prediction based on classifiers related to all species. "Average" represents the mean AUC of all the classifiers in each category. "Highest" represents the classifier with highest AUC among all the classifiers in each category. Error bars show standard deviation. "∗" indicates that AUC of PPCM was significantly (P value < 0.05; t-test) higher than that of the most accurate classifier in each category.

3.3. Performance of PPCM in GO2PPI + Phyloprof Category. After separate evaluation of PPCM in the GO2PPI and Phyloprof categories, we assessed the performance of PPCM in the GO2PPI + Phyloprof category which combined all the classifiers in both GO2PPI and Phyloprof. The AUCs of PPCMs in the GO2PPI + Phyloprof category were 0.83, 0.85, and 0.86 for SC, cross species, and all species PPCM, respectively (Figure 4), which are significantly higher than those of PPCMs in either GO2PPI or Phyloprof category separately (Figures 2 and 3). Compared with the highest AUCs of individual classifiers in GO2PPI and Phyloprof category, the cross species PPCM improves AUC by 18% and the improvement by SC PPCM was 17% (Figures 2, 3, and 4). These results indicate that PPCM based on all the 194 classifiers from both GO2PPI and Phyloprof could generate more accurate PPI prediction than PPCM based on a fewer number of classifiers in GO2PPI or Phyloprof individually, further supporting the aforementioned premise that more features incorporated into PPCM would enhance PPI prediction accuracy. In summation, based on our combinatorial approach, our cross species PPCM results yield informative predictions that will help build high-quality PPI networks

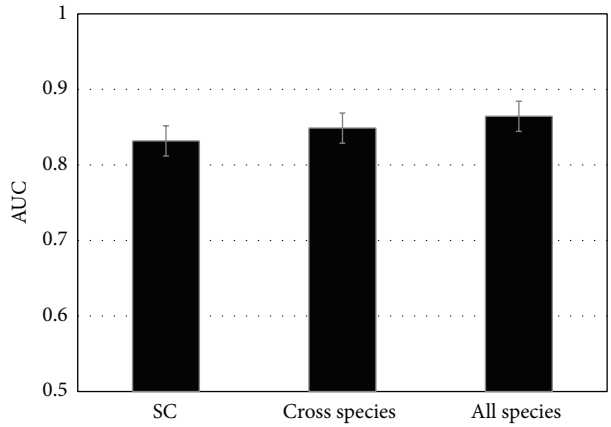


Figure 4: Comparison of PPI prediction accuracy in the GO2PPI + Phyloprof category. Error bars show standard deviation.

for nonmodel organisms. Such prediction will be valuable for nonmodel organisms that lack biological data and PPI prediction software for nonmodel organisms [18].

Recently, ensemble classifiers, for example, LibD3C, were developed based on a clustering and dynamic selection strategy [39]. In order to compare the performance of Random Forests method applied by our PPCM with the latest ensemble classifiers, we performed ensemble classifiers calculation on our all species training and testing datasets of the GO2PPI + Phyloprof category (see Figure 4) by LibD3C in Weka-3.7.12 with default setting. The average AUC by LibD3C was 0.86 ± 0.03 which is in an excellent agreement with our Random Forests result (0.86 ± 0.02). Therefore, Random Forests method applied by our PPCM shows very similar performance with the latest ensemble classifiers (LibD3C).

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors wish to thank G. A. Tuskan and T. J. Tschaplinski for providing edits and constructive comments. This research was supported by the Department of Energy, Office of Science, Genomic Science Program (under Award no. DESC0008834). Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy (under Contract no. DE–AC05–00OR22725).

References

[1] A.-C. Gavin, M. Bösche, R. Krause et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[2] B. Alberts, "The cell as a collection of protein machines: preparing the next generation of molecular biologists," *Cell*, vol. 92, no. 3, pp. 291–294, 1998.

[3] D. Devos and R. B. Russell, "A more complete, complexed and structured interactome," *Current Opinion in Structural Biology*, vol. 17, no. 3, pp. 370–377, 2007.

[4] A.-C. Gavin, P. Aloy, P. Grandi et al., "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–636, 2006.

[5] A. Kumar and M. Snyder, "Proteomics: protein complexes take the bait," *Nature*, vol. 415, no. 6868, pp. 123–124, 2002.

[6] I. Xenarios, "DIP: the database of interacting proteins," *Nucleic Acids Research*, vol. 28, no. 1, pp. 289–291, 2000.

[7] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. D1, pp. D808–D815, 2013.

[8] C. von Mering, R. Krause, B. Snel et al., "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[9] J. Planas-Iglesias, J. Bonet, J. García-García, M. A. Marín-López, E. Feliu, and B. Oliva, "Understanding protein–protein interactions using local structural features," *Journal of Molecular Biology*, vol. 425, no. 7, pp. 1210–1224, 2013.

[10] J. L. Sussman, D. Lin, J. Jiang et al., "Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules," *Acta Crystallographica Section D Biological Crystallography*, vol. 54, no. 6, part 1, pp. 1078–1084, 1998.

[11] G. T. Hart, A. Ramani, and E. Marcotte, "How complete are current yeast and human protein-interaction networks?" *Genome Biology*, vol. 7, no. 11, p. 120, 2006.

[12] G. Gallone, T. I. Simpson, J. D. Armstrong, and A. P. Jarman, "Bio::Homology::InterologWalk—a Perl module to build putative protein-protein interaction networks through interolog mapping," *BMC Bioinformatics*, vol. 12, article 289, 2011.

[13] C. Y. Yu, L. C. Chou, and D. T. H. Chang, "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins," *BMC Bioinformatics*, vol. 11, article 167, 2010.

[14] J. Garcia-Garcia, E. Guney, R. Aragues, J. Planas-Iglesias, and B. Oliva, "Biana: a software framework for compiling biological interactions and analyzing networks," *BMC Bioinformatics*, vol. 11, no. 1, article 56, 2010.

[15] Y. Liu, I. Kim, and H. Zhao, "Protein interaction predictions from diverse sources," *Drug Discovery Today*, vol. 13, no. 9-10, pp. 409–416, 2008.

[16] Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph, "Random forest similarity for protein-protein interaction prediction from multiple sources," in *Proceedings of the Pacific Symposium on Biocomputing*, pp. 531–542, January 2005.

[17] X. W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, 2005.

[18] K. A. Theofilatos, C. M. Dimitrakopoulos, A. K. Tsakalidis, S. D. Likothanassis, S. T. Papadimitriou, and S. P. Mavroudi, "Computational approaches for the prediction of protein-protein interactions: a survey," *Current Bioinformatics*, vol. 6, no. 4, pp. 398–414, 2011.

[19] J. Garcia-Garcia, S. Schleker, J. Klein-Seetharaman, and B. Oliva, "BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference," *Nucleic Acids Research*, vol. 40, no. W1, pp. W147–W151, 2012.

[20] R. Jansen, H. Yu, D. Greenbaum, and et al, "A bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003.

[21] X.-W. Chen, M. Liu, and Y. Hu, "Integrative neural network approach for protein interaction prediction from heterogeneous data," in *Advanced Data Mining and Applications*, C. Tang, C. X. Ling, X. Zhou, N. J. Cercone, and X. Li, Eds., vol. 5139 of *Lecture Notes in Computer Science*, pp. 532–539, Springer, Berlin, Germany, 2008.

[22] S. M. Gomez, W. S. Noble, and A. Rzhetsky, "Learning to predict protein–protein interactions from protein sequences," *Bioinformatics*, vol. 19, no. 15, pp. 1875–1881, 2003.

[23] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, article 25, 2007.

[24] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 63, no. 3, pp. 490–500, 2006.

[25] Y. Zhang, D. Zhang, G. Mi et al., "Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions," *Computational Biology and Chemistry*, vol. 36, pp. 36–41, 2012.

[26] S. M. Augusty and S. Izudheen, "A survey: evaluation of ensemble classifiers and data level methods to deal with imbalanced data problem in protein-protein interactions," *Review of Bioinformatics and Biometrics*, vol. 2, no. 1, 2013.

[27] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of the National Academy of Sciences*, vol. 96, no. 8, pp. 4285–4288, 1999.

[28] T. Gaasterland and M. A. Ragan, "Constructing multigenome views of whole microbial genomes," *Microbial & Comparative Genomics*, vol. 3, no. 3, pp. 177–192, 1998.

[29] E. S. Snitkin, A. M. Gustafson, J. Mellor, J. Wu, and C. DeLisi, "Comparative assessment of performance and genome dependence among phylogenetic profiling methods," *BMC Bioinformatics*, vol. 7, no. 1, article 420, 2006.

[30] R. Jothi, T. M. Przytycka, and L. Aravind, "Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment," *BMC Bioinformatics*, vol. 8, no. 1, article 173, 17 pages, 2007.

[31] J. Sun, Y. Li, and Z. Zhao, "Phylogenetic profiles for the prediction of protein–protein interactions: how to select reference organisms?" *Biochemical and Biophysical Research Communications*, vol. 353, no. 4, pp. 985–991, 2007.

[32] D. Herman, D. Ochoa, D. Juan, D. Lopez, A. Valencia, and F. Pazos, "Selection of organisms for the co-evolution-based study of protein interactions," *BMC Bioinformatics*, vol. 12, no. 1, article 363, 2011.

[33] M. Simonsen, S. R. Maetschke, and M. A. Ragan, "Automatic selection of reference taxa for protein-protein interaction prediction with phylogenetic profiling," *Bioinformatics*, vol. 28, no. 6, pp. 851–857, 2012.

[34] S. V. Date and E. M. Marcotte, "Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages," *Nature Biotechnology*, vol. 21, no. 9, pp. 1055–1062, 2003.

[35] J. Wu, S. Kasif, and C. DeLisi, "Identification of functional links between genes using phylogenetic profiles," *Bioinformatics*, vol. 19, no. 12, pp. 1524–1530, 2003.

[36] S. Cokus, S. Mizutani, and M. Pellegrini, "An improved method for identifying functionally linked proteins using phylogenetic profiles," *BMC Bioinformatics*, vol. 8, supplement 4, article S7, 2007.

[37] S. Singh and D. P. Wall, "Testing the accuracy of eukaryotic phylogenetic profiles for prediction of biological function," *Evolutionary Bioinformatics*, vol. 4, pp. 217–223, 2008.

[38] S. R. Maetschke, M. Simonsen, M. J. Davis, and M. A. Ragan, "Gene Ontology-driven inference of protein-protein interactions using inducers," *Bioinformatics*, vol. 28, no. 1, pp. 69–75, 2011.

[39] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, and Q. Zou, "LibD3C: ensemble classifiers with a clustering and dynamic selection strategy," *Neurocomputing*, vol. 123, pp. 424–435, 2014.

[40] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, and Q. Zou, "nDNA-prot: identification of DNA-binding proteins based on unbalanced classification," *BMC Bioinformatics*, vol. 15, no. 1, article 298, 2014.

[41] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, Article ID e56499, 2013.

[42] P. Smialowski, P. Pagel, P. Wong et al., "The Negatome database: a reference set of non-interacting protein pairs," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D540–D544, 2010.

[43] L. Zhang, S. Wong, O. King, and F. P. Roth, "Predicting co-complexed protein pairs using genomic and proteomic data integration," *BMC Bioinformatics*, vol. 5, no. 1, article 38, 2004.

[44] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 824287, 12 pages
http://dx.doi.org/10.1155/2015/824287

Research Article

Significant Microsynteny with New Evolutionary Highlights Is Detected through Comparative Genomic Sequence Analysis of Maize CCCH IX Gene Subfamily

Wei-Jun Chen, Yang Zhao, Xiao-Jian Peng, Qing Dong, Jing Jin, Wei Zhou, Bei-Jiu Cheng, and Qing Ma

Key Laboratory of Crop Biology of Anhui Province, School of Life Sciences, Anhui Agricultural University, Hefei, Anhui 230036, China

Correspondence should be addressed to Qing Ma; mqqmmq@ahau.edu.cn

Received 8 May 2015; Revised 16 August 2015; Accepted 17 August 2015

Academic Editor: Yanbin Yin

Copyright © 2015 Wei-Jun Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

CCCH zinc finger proteins, which are characterized by the presence of three cysteine residues and one histidine residue, play important roles in RNA processing in plants. Subfamily IX CCCH proteins were recently shown to function in stress tolerances. In this study, we analyzed CCCH IX genes in *Zea mays, Oryza sativa*, and *Sorghum bicolor*. These genes, which are almost intronless, were divided into four groups based on phylogenetic analysis. Microsynteny analysis revealed microsynteny in regions of some gene pairs, indicating that segmental duplication has played an important role in the expansion of this gene family. In addition, we calculated the dates of duplication by Ks analysis, finding that all microsynteny blocks were formed after the monocot-eudicot divergence. We found that deletions, multiplications, and inversions were shown to have occurred over the course of evolution. Moreover, the Ka/Ks ratios indicated that the genes in these three grass species are under strong purifying selection. Finally, we investigated the evolutionary patterns of some gene pairs conferring tolerance to abiotic stress, laying the foundation for future functional studies of these transcription factors.

1. Introduction

Transcription factors (TFs) are critical regulators of gene expression that control many important biological processes, such as cellular morphogenesis, signal transduction, and environmental stress responses [1]. Zinc finger TFs belong to one of the largest TF families in plants and can be categorized into at least 14 families, such as RING finger, WRKY, DOF, and LIM families [2–4]. These TFs have been proven to regulate gene expression with the aid of DNA-binding or protein-binding proteins. However, previous reports discovered a new type of *Arabidopsis* zinc finger proteins, named CCCH zinc finger family that is involved in mRNA binding and processing [5]. CCCH-type proteins are TFs with a typical motif consisting of three cysteine residues and one histidine residue. CCCH proteins, containing one to six copies of CCCH-type zinc finger motifs, were originally defined as C-X6-14-C-X4-5-C-X3-H, but a recent study has redefined them as C-X4-15-C-X4-6-C-X3-H, following genome-wide analysis of rice and *Arabidopsis thaliana* CCCH proteins [5].

Recent studies have revealed that CCCH proteins participate in the regulation of plant growth, developmental processes, and environmental responses. In rice, a novel nuclear-localized CCCH-type zinc finger protein, OsDOS, is involved in delaying leaf senescence by integrating developmental cues to the jasmonic pathway [6]. In pepper and rice, the CCCH-domain proteins CaKR1 and OsC3H12 were shown to protect plants from bacterial blight [7, 8]. During senescence in *Arabidopsis*, HUA1, a CCCH-type zinc finger protein with six tandem CCCH motifs, likely participates in regulating flower development [9]. In addition, some CCCH zinc finger proteins are also involved in the abiotic stress response. Two closely related proteins in *Arabidopsis*, AtSZF1 and AtSZF2, are both involved in modulating salt stress tolerance in plants [10]. Recently, *ZmC3H4* and *ZmC3H28*, which are indirectly regulated by ABA or drought, and 10 other maize CCCH IX genes were found to be responsive to abiotic stress [11].

The CCCH-type zinc finger protein family had been studied in some model organisms on a phylogenetic scale, but its particular evolutionary pathway is still poorly understood. Gramineae, which evolved approximately 60–70 mya (million years ago) from a common ancestor, includes a number of agronomically important crops, such as *Oryza sativa*, *Zea mays*, and *Sorghum bicolor* [12]. The origin of these crops dates back to approximately 50 to 65 mya, and the family has now expanded to over 10,000 species [13]. Whole-genome analyses have revealed high levels of genetic conservation in grasses over the course of evolution, yet these studies have revealed no trace of microsynteny (preservation of a specific local gene order) across grasses [14, 15]. The goal of this study was to identify stress-responsive genes in *Z. mays*, *O. sativa*, and *S. bicolor* and to analyze the evolutionary relationships of CCCH IX subfamily at the molecular level using microsynteny analysis. Specifically, we searched for CCCH IX subfamily genes and predicted their structures. To determine the expression patterns of CCCH IX genes in maize tissue, we utilized publically available microarray data from Sekhon et al. [16]. The expression map was shown in Figure S1 (in Supplementary Material available online at http://dx.doi.org/10.1155/2015/824287). In addition, we identified duplication events and calculated Ka/Ks values. Finally, we designed microsynteny maps to identified conservative CCCH IX genes during evolution. The results of this study lay the foundation for future functional studies of CCCH IX subfamily genes.

2. Materials and Methods

2.1. Identification of Genes Encoding CCCH IX Proteins. The recent versions of genome, protein, and cDNA sequences for the following three grass species were downloaded from the respective genome sequence sites: *Oryza sativa* (version 7.0) from the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/), *Zea mays* (version 2.0) from the B73 Maize Genome Project (http://www.maizesequence.org/index.html), and *Sorghum bicolor* (version 1.0) from the DOE-JGI Community Sequencing Program (CSP) (http://www.phytozome.net/sorghum.php). These nucleotide and protein sequences were used to build local databases using DNATOOLS software [17]. The conserved CCCH domain (PF00642) based on the Hidden Markov Model (HMM) was obtained from http://pfam.sanger.ac.uk/ (Pfam database) [18]. This HMM profile was used as a query to search against the protein database with the BLASTp program (version blast-2.2.9-ia32-win32) (p value = 0.001). This step was crucial for finding as many similar sequences as possible. All predicted protein sequences of genes were analyzed in the Pfam HMM database and the SMART tool (http://smart.embl-heidelberg.de/) to identify CCCH domains, and proteins without these regions were excluded from the dataset [19]. All potential sequences were aligned using Clustal W, and all identical sequences were checked manually to remove redundant genes prior to subsequent analysis [20].

The molecular weight (kDa) and isoelectric point (pI) of each gene were calculated using the online ExPASy tools (http://www.expasy.org/tools/) [21]. The intron distribution patterns and intron/exon boundaries of the CCCH IX genes were deduced by using Gene Structure Display Server (http://gsds.cbi.pku.edu.cn/) to compare the predicted full-length cDNA or coding sequences with the corresponding genomic sequences [22].

2.2. Phylogenetic Analysis of CCCH IX Genes. The phylogeny of CCCH IX genes was performed by clustering and aligning the protein sequences using Clustal W. The phylogenetic tree was constructed using MEGA 6.0 by the neighbor-joining (NJ) method with the following parameters: Poisson correction, pairwise deletion, and bootstrapping (1,000 replicates) [23]. To confirm the robustness of the NJ tree, an ML tree was constructed using the maximum likelihood method (MEGA 6.0; bootstrap = 1,000 replicates, amino acid substitution model, and Jones-Taylor-Thornton matrix), and an MP tree was constructed using the maximum parsimony method (MEGA 6.0; bootstrap = 1,000 replicates).

2.3. Detection of CCCH IX Gene Expansion. The following analysis was performed to obtain in-depth knowledge of the evolutionary relationships among CCCH IX genes and to determine whether these genes were derived from segmental duplication or tandem duplication events. Tandem duplication is characterized as the presence of multiple gene family members within the same or neighboring intergenic regions. To be defined as a segmental duplication event, each pair of protein-coding genes (excluding noncoding RNA genes, pseudogenes, and so on) in each genome must reside within a duplicated block; moreover, there must be a high similarity between their neighboring protein-coding genes at the amino acid level [24]. First, all identified CCCH IX genes were used as the original anchor points. Next, 100 kb sequences upstream and downstream of each anchor point were compared by pairwise BLASTp (E-value $\leq 10^{-10}$) analysis to identify duplicated genes between two independent regions. The number of protein-coding genes flanking any anchor point was then counted [25]. When three or more such genes pairs with syntenic relationships were identified in two regions, the regions were considered to have been derived from a large-scale duplication event [26, 27].

2.4. Microsynteny Analysis. Microsynteny analysis across the three species was carried out based on comparisons of the specific regions containing CCCH IX genes. Levels of similarity between the flanking genes of each CCCH IX gene in one species and those in the other species were determined by pairwise comparisons using the BLASTp program. A syntenic block was defined as a region where three or more conserved homologs (BLASTp E-value $\leq 10^{-20}$) were located within a 100 kb region between genomes [28, 29]. The relative syntenic quality in a region was calculated based on the sum of the total number of genes in both conserved sequence regions, excluding tandem duplication. A circular microsynteny map was also constructed using the program Circos-0.54, which utilizes the Perl language [30].

Table 1: The 27 CCCH IX genes identified in three species and their sequence characteristics (gene ID, ORF, MW, PI, and chromosome locations).

Gene name	Gene identifier	ORF (aa)	MW (Da)	pI	Chromosome	Chromosomal localization	
						Star	End
OsC3H2	LOC_Os01g09620.1	386	41422.51	6.41	Os1	4949047	4951126
OsC3H10	LOC_Os01g53650.1	225	24982.17	6.06	Os1	30824689	30825670
OsC3H24	LOC_Os03g49170.1	764	81043.54	6.63	Os3	28008600	28012204
OsC3H33	LOC_Os05g03760.1	601	63235.89	8.65	Os5	1662021	1664438
OsC3H35	LOC_Os05g10670.1	464	49684.7	9.02	Os5	5846045	5848291
OsC3H37	LOC_Os05g45020.1	255	28257.88	5.39	Os5	26171092	26172349
OsC3H50	LOC_Os07g38090.1	657	69391.93	6.61	Os7	22840986	22843954
OsC3H52	LOC_Os07g47240.1	280	31590.77	8.04	Os7	28233256	28234642
OsC3H67	LOC_Os12g33090.1	619	64725.96	6.04	Os12	20018555	20021302
ZmC3H4	GRMZM2G180979_P01	746	79771.77	6.4	Zm1	263731420	263734946
ZmC3H10	GRMZM2G099622_P03	360	37921.57	6.37	Zm2	205904041	205906953
ZmC3H12	GRMZM5G853245_P03	370	39786.59	6.46	Zm3	6681798	6683631
ZmC3H28	GRMZM5G845366_P01	482	52092.46	7.97	Zm5	12382780	12384893
ZmC3H34	AC233871.1_FGP008	416	45437.96	8.78	Zm6	1823029	1825102
ZmC3H38	GRMZM5G801627_P01	394	42104.18	8.02	Zm6	132663265	132665063
ZmC3H39	GRMZM2G004795_P01	270	29691.44	5.67	Zm6	160013893	160016282
ZmC3H43	GRMZM5G842019_P01	656	69821.35	6.88	Zm7	158677416	158680576
ZmC3H51	GRMZM2G173124_P03	378	40178.89	6.46	Zm8	20669510	20671733
ZmC3H53	GRMZM2G093404_P01	262	28025.75	6.4	Zm8	124841672	124843159
ZmC3H54	GRMZM2G117007_P01	372	40034.27	8.65	Zm8	131041707	131043172
ZmC3H63	GRMZM2G027298_P01	594	62657.48	5.71	Zm10	27949011	27951381
SbC3H2	SB01G011150	745	79472.39	6.51	Sb1	10009063	10012269
SbC3H10	SB02G036710	680	72441.2	6.88	Sb2	71102658	71104964
SbC3H12	SB03G003110	350	37662.05	8.2	Sb3	3207828	3209693
SbC3H44	SB08G016640	533	56473.43	5.89	Sb8	44663680	44665480
SbC3H45	SB09G002390	611	64185.57	8.21	Sb9	2607622	2610024
SbC3H47	SB09G006050	399	42664.56	6.21	Sb9	8731871	8733975

Open reading frame (ORF), molecular weight (MW), and isoelectric point (IP).

2.5. Ks Analysis of Homologous Segments. The time of divergence of duplicated gene pairs within each duplicated block or the divergence of homologous segments was estimated by calculating Ks values between homologous genes using DnaSP (version 5.10) [31–33]. Sliding window analysis of nonsynonymous substitutions per nonsynonymous site (Ka/Ks) ratios was conducted with the following parameters: window size, 150 bp; step size, 9 bp [34].

Ks values can also be used to calculate the timing of large-scale replications. For each pair of duplicated regions, the mean Ks value for individual homologs in flanking conserved genes was calculated and used to determine the approximate time of divergence. Hence, Ks could be converted to the divergence time beyond the Gramineae evolutionary rate of each locus. The divergence time (T) was calculated as $T = \mathrm{Ks}/(2 \times 6.5 \times 10^{-9}) \times 10^{-6}$ mya [35].

3. Results

3.1. Phylogenetic and Sequence Structure Analysis of CCCH IX. In the previous study, 67 CCCH genes were identified in *Z. mays*. Then, we identified 55 genes encoding CCCH zinc finger proteins in *S. bicolor* (Supplementary Table 1) using the BLASTp program. For convenience, we assigned names to these genes (*SbC3H1–SbC3H55*) according to their chromosomal positions. Based on previous studies, we identified six *S. bicolor* genes in the CCCH IX subfamily [5, 11]. A total of 27 CCCH IX genes, which are listed in Table 1 (nine genes from *O. sativa,* six genes from *S. bicolor*, and twelve genes from *Z. mays*), were subjected to further analysis. The lengths of the 27 encoded CCCH IX proteins vary from 225 to 764 aa, with an average of 476 aa. Other pieces of information, including the clone number, chromosomal location, molecular weight (Mw), and isoelectric point (pI) of each CCCH IX gene/protein, are listed in Table 1. To determine the organization and distribution of CCCH IX genes on different chromosomes, we constructed a chromosome map. The 27 CCCH IX genes are randomly distributed on chromosomes, as shown in Figure 1.

To explore the evolutionary relationships between members of the CCCH IX zinc finger subfamily, we constructed a phylogenetic tree using the neighbor-joining (NJ) method

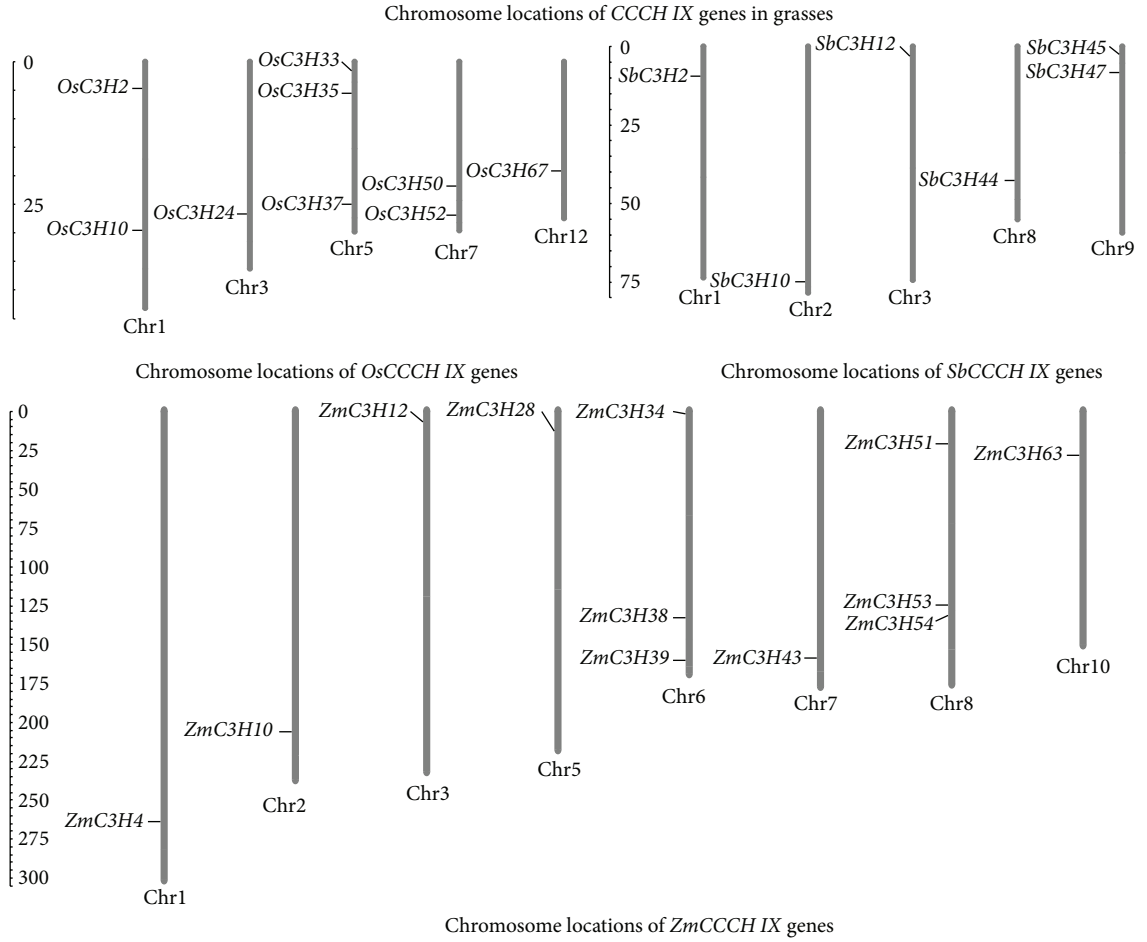


Figure 1: Chromosomal locations of CCCH IX genes in three Gramineae species (*O. sativa*, *S. bicolor,* and *Z. mays*). The 27 CCCH IX genes are randomly distributed on chromosomes, including nine rice genes, six sorghum genes, and twelve maize genes.

based on protein sequence alignment. The phylogenetic tree is divided into four clades (Figure 2(a)). Although different clades have different numbers of members, clades 1–3 consist of proteins from all three grass species, whereas *S. bicolor* proteins are absented from clade 4. These differences may have been derived from partial gene loss that may have occurred after large-scale duplication events after the formation of new species, which drove species separation. The phylogenetic relationships depicted in the ML and MP trees are largely consistent with these results (Figure S2). We investigated the sequence structure by exon-intron structure analysis (Figure 2(b)) (http://gsds.cbi.pku.edu.cn/) [36]. The most closely related CCCH IX members in the same clades share similar gene lengths and exon lengths. The only exception was observed in the sequence (*SbC3H44*) from *S. bicolor* that contained one intron and one CCCH domain (Figure 3). Interestingly, the remaining 26 genes are entirely intronless, and they contain two CCCH domains without exception. These findings imply that this subfamily of genes has retained sequences, which have been conserved at the structural level, including exon-intron structure and the number of CCCH domains, throughout millions of years of evolution.

3.2. Complicated Duplication Events Have Contributed to CCCH IX Expansion. We estimated the chromosomal locations of CCCH IX genes in these grasses and examined the evolutionary relationships between these genes. However, in the relevant chromosomes, there was no universal tandem duplication, due to the irregular distribution of CCCH IX genes. To determine whether the regions flanking CCCH IX genes have undergone large-scale duplication events during the evolution, we compared the flanking genes of any two CCCH IX genes. If three or more flanking genes had a best nonself-match according to BLASTp (E-value $\leq 10^{-10}$ within species and E-value $\leq 10^{-20}$ between species), we considered that these members belonged to a duplicated block. Based on this information above, we investigated the evolutionary origins and evolutionary relationships within and between grasses species using a Perl script (Figure 4). Initially, we found 11 duplicated gene segments constituting a network within species, including five maize genes and four rice genes. However, we subsequently identified five groups containing 23 genes (the five groups are shown in Figure 5, with red arrows representing the 23 genes) from species exhibiting tight microsynteny relationships.

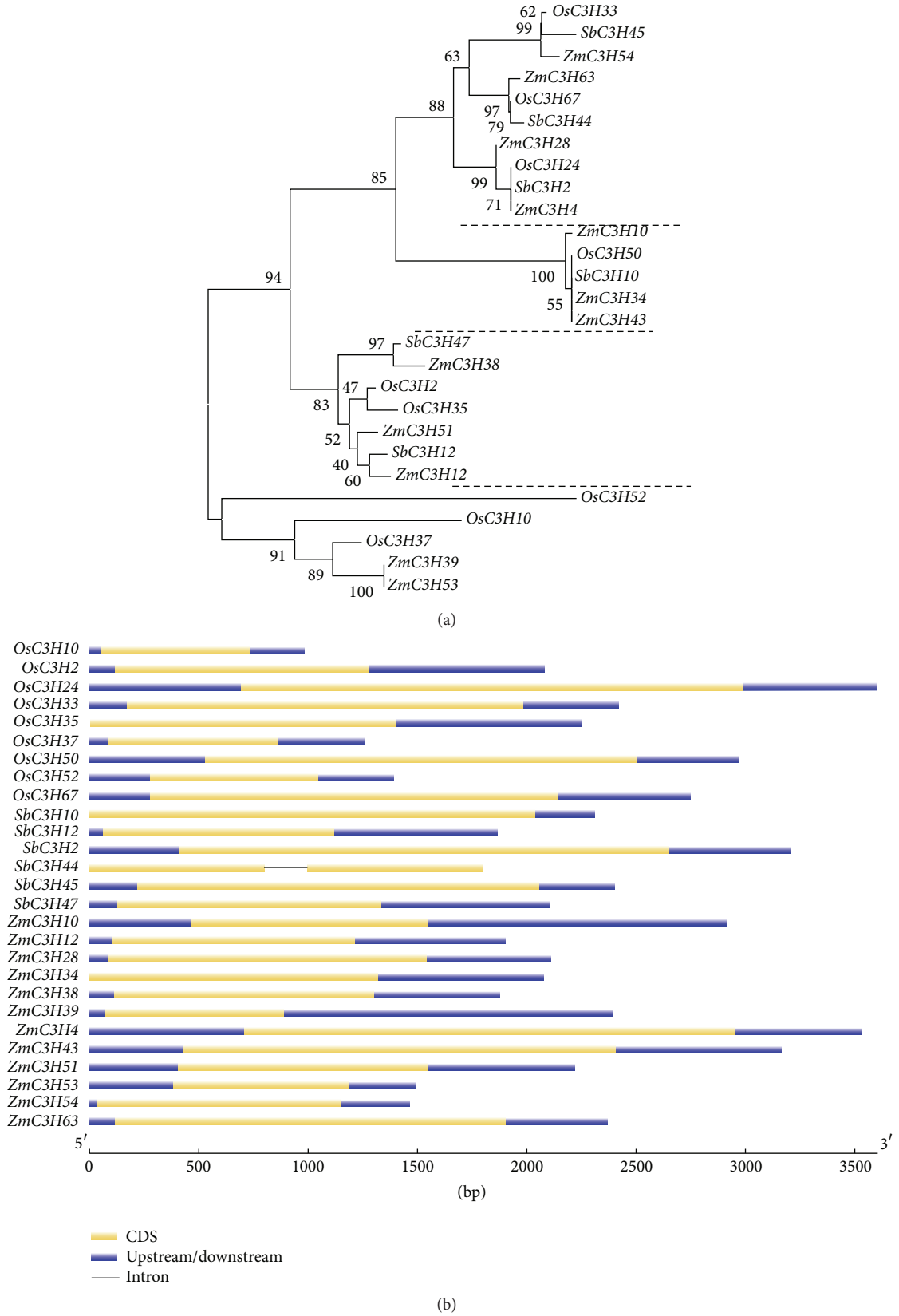


Figure 2: (a) Phylogenetic tree of CCCH IX proteins from *O. sativa*, *S. bicolor*, and *Z. mays*. This tree was constructed using the MEGA 6.0 program by the N-J method with 1,000 bootstrap replicates based on amino acid sequence. The tree is divided into four clades (clades I–IV). (b) Exon-intron structures of 27 CCCH IX genes in three species. Exons and introns are indicated by green thick lines and thin gray lines, respectively. The untranslated regions (UTRs) are indicated by blue lines.

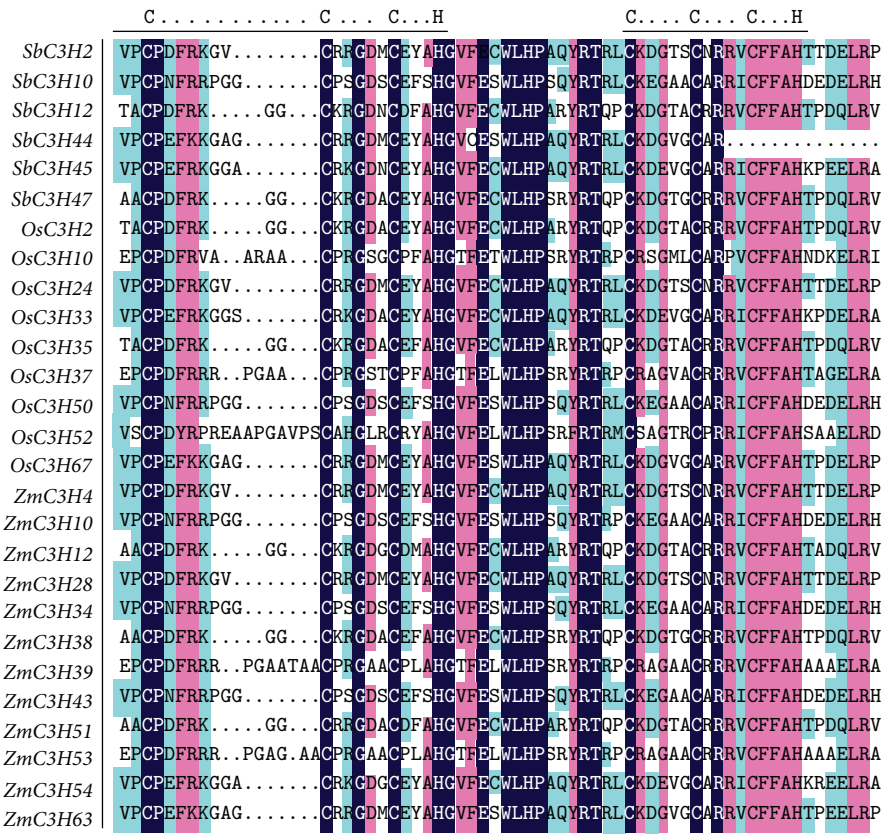


FIGURE 3: Alignment of the amino acid sequences of CCCH IX proteins in three grass species. Identical (100%), conservative (75–99%), and blocks (50–74%) of similar amino acid residues are shaded in deep blue, dark pink, and light blue, respectively. The conserved CCCH zinc finger motifs are indicated by straight lines.

In rice, we found four highly similar genes among the sequences flanking both sides of *OsC3H2/OsC3H35* (Figure 5(a)). A duplication event may have occurred during the evolutionary history of this pair of genes. Based on this notion, we reasoned that this pair of genes from one group shares a closer phylogenetic relationship than the others, which helps confirm the results of phylogenetic analysis (Figure 2). In addition, the gene pair *OsC3H10/OsC3H37* is surrounded by six conserved genes compared with the gene pair *OsC3H2/OsC3H35*, and their microsyntenic relationship is closer than that of the latter pair (Figure 5). In addition, the collinear gene pair *OsC3H2/OsC3H35* is located on OsChr1/OsChr5, and the gene pair *OsC3H10/OsC3H37* is present on the same chromosome (Figure 1). These results suggest that, during the evolution of the rice genome, these two chromosome segments were generated by whole-genome duplication; a large-scale duplication has affected the evolution of CCCH IX genes in rice. In sorghum, the gene pair *SbC3H12/SbC3H47* was identified as microsyntenic relationship (Figure 5(a)), because only six CCCH IX genes belong to this subfamily and they exhibit relatively sparse microsynteny. Based on the phylogenetic tree (Figure 2), each branch of *S. bicolor* genes has a high degree of similarity with that of other grasses.

In maize, five CCCH IX genes are located in the duplicated section of the genome, and they share a microsyntenic relationship (Figures 5(c) and 5(d)). *ZmC3H10/ZmC3H34/ZmC3H43* exhibit tight microsynteny. These results suggest that a large-scale duplication event has occurred during the process of evolution of the maize genome. This duplication event can also be deduced by examining the gene pair *ZmC3H4/ZmC3H28*. We compared 100 kb sequences upstream and downstream of each anchor point by pairwise alignment, finding that many duplicated regions occurred as mentioned above in the three species. These results indicate that large-scale duplication events have resulted in the production of paralogous genes throughout evolutionary history.

Microsynteny analysis can be used to predict the locations of homologous genes in different species [37]. Regions with 80% of close homologs in the same order and transcriptional orientation are characterized as exhibiting conserved microsynteny [38]. We used this analysis to deduce the molecular evolutionary origins and orthologous relationships within the chromosome regions containing CCCH IX genes in the three grass species. We performed a stepwise gene-by-gene reciprocal comparison to gauge the linkages between CCCH IX regions. Based on these detailed comparisons and rigorous analysis, the results of mapping of conserved

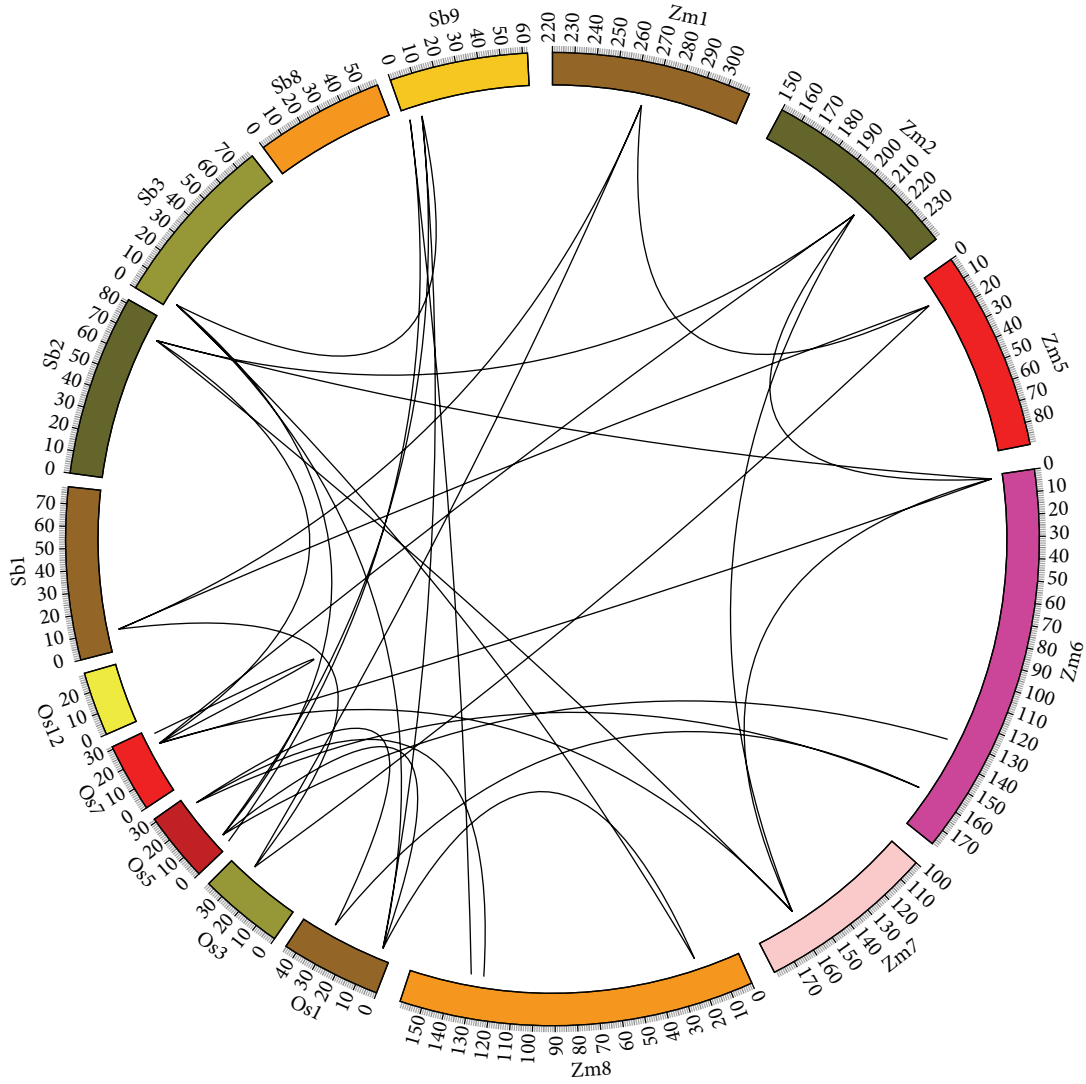


FIGURE 4: Extensive microsynteny of CCCH regions across *O. sativa*, *S. bicolor*, and *Z. mays*. *O. sativa* chromosomes (labeled Os) are indicated by orange boxes. *S. bicolor* and *Z. mays* chromosomes (labeled Sb and Zm, resp.) are shown in blue and green, respectively. CCCH IX regions in the three grasses are shown in the circle. Black lines show syntenic relationships.

microsynteny regions are similar to those shown in the phylogenetic tree, but analysis of the flanking fragments represents a more thorough approach than phylogenetic analysis.

3.3. Conserved Microsynteny of CCCH IX Genes between Species. The genes from the three grasses exhibiting tight microsyntenic relationships were divided into five groups (Figure 5). Genome segments in the same group were likely derived from a single sequence during evolution, which resulted in species differentiation. Sequence segments from the same group are considered to be homologous genes whereby genetic evolution led to species separation. In group (a), we observed a marked opposite-direction microsynteny relationship among *OsC3H2*/*OsC3H35*, *SbC3H12*/*SbC3H47*, and *ZmC3H38*/*ZmC3H51* segments (Figure 5). In this group, these gene fragments were derived from a duplication event and are orthologous to each other. The same situation exists in group (b), where we observed microsynteny in a series of genes: *OsC3H10*, *OsC3H37*, *ZmC3H39*, and *ZmC3H53* (Figure 5). Chromosome inversions result in reversals in gene order. Group (c) exhibited a higher level of compact microsynteny, especially *OsC3H50*, *SbC3H10*, *ZmC3H10*, and *ZmC3H43* (Figure 5). On the contrary, *SbC3H10*, *ZmC3H34*, and *ZmC3H43* exhibited opposite-orientation microsynteny. In group (d), *OsC3H24*, *SbC3H2*, *ZmC3H4*, and *ZmC3H28* appeared microsyntenic, especially the pair *OsC3H24* and *SbC3H2*, which has nine best matching genes according to BLASTp alignment (Figure 5). Group (e) genes *OsC3H33* and *SbC3H45* were identified as having successive same-direction microsynteny (Figure 5). In CCCH IX regions in these species, the chromosomal composition of one species is often assembled from two or more successive segments. For example, fragment SbChr3 appears to match OsChr1

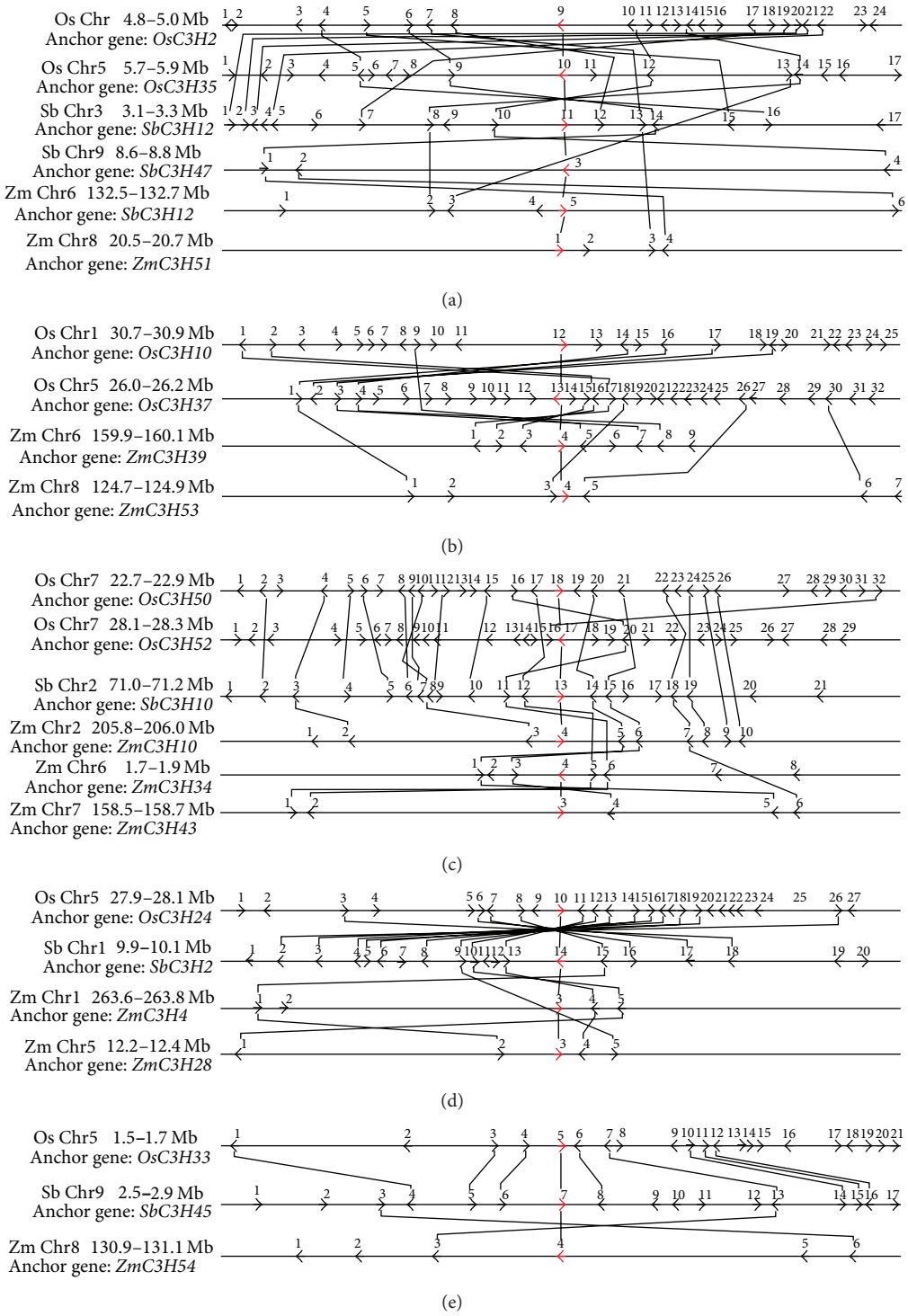


Figure 5: Microsynteny maps of CCCH IX genes in grasses. Red arrows represent anchor (CCCH IX) genes, and upstream and downstream genes are represented by black arrows. All genes are numbered from left to right for each segment. Black lines connect conserved gene pairs.

and OsChr5, and fragment SbChr1 appears to match OsChr3 and ZmChr1. In the CCCH IX subfamily, chromosomal translocation is a common phenomenon that occurs during the process of differentiation. Based on previous study (in which 12 stress-responsive CCCH IX genes were identified in maize [11]), *OsC3H24*, *SbC3H2*, *SbC3H45*, and so on (as mentioned above) may be involved in regulating abiotic stress through microsynteny mapping, as shown in Figure 5.

3.4. Estimating the Dates of Duplication Events. Based on the assumption that the synonymous mutation rate at each site is stable over time [39], we calculated the duplication event

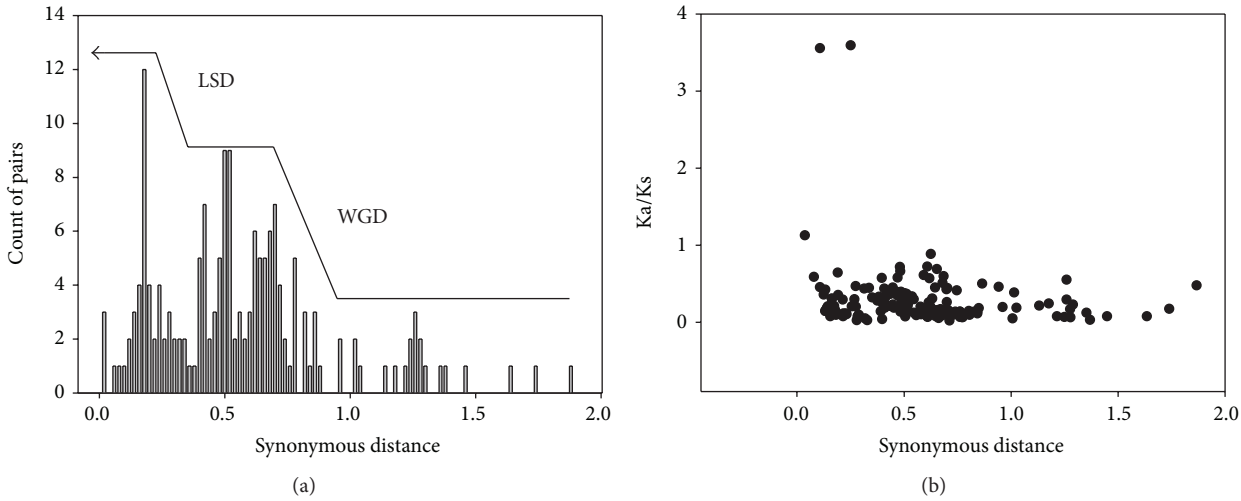


Figure 6: (a) Distribution of synonymous distance (Ks) between paralogous genes flanking duplicated CCCH IX genes in the three grass species. The histogram depicts the number of duplicate gene pairs (y-axis) versus synonymous distance between pairs (x-axis). CCCH IX blocks experienced whole-genome duplication (WGD) in the first stage of their evolution and large-scale duplication (LSD) in the second stage. (b) Ka/Ks ratios of duplicated CCCH IX genes and their flanking paralogs in the three grasses. The y- and x-axes denote the Ka/Ks ratio and synonymous distance for each pair, respectively.

date based on the conserved flanking protein-coding genes. Each pair of proteins in a microsynteny block was aligned at the amino acid level, and codons from gapless aligned regions were used to calculate Ks values using CodeML [24]. We removed any Ks values > 2.0 due to the risk of saturation [40]. The approximate date of the duplication event was then calculated using the mean Ks and an estimated rate of silent site substitutions of 6.5×10^{-9} substitutions/synonymous site/year. The divergence time (T) was calculated as $T = \text{Ks}/(2 \times 6.5 \times 10^{-9}) \times 10^{-6}$ mya [35]. The mean Ks values for each duplication event and the estimated dates are listed in Table 2. The results reveal that, for these duplication events, few genes and their flanking fragments expanded before Gramineae speciation (60–70 mya) [40]. The subsequent whole-genome duplication played an important role in the expansion of genes containing CCCH IX regions, leading to complete genome diploidization along with gene rearrangement and loss. The related gene duplication events occurred frequently, leading to further integration of these genes in maize and sorghum approximately 15 mya (Figure 6(a)). We identified the same pattern for synchronous replication during evolutionary history, which helps confirm that section is synchronous in function.

3.5. Selection Pressure on CCCH IX Genes. Since large-scale duplication has contributed to genome evolution, we also calculated the selection pressure among CCCH IX duplicated genes [41]. We calculated the Ka/Ks ratios for 31 pairs of conservative homogenous genes, along with their flanking segments. We found that the Ka/Ks ratios of homologous replication groups were less than 1, except for those of two flanking genes (Figure 6(b)), suggesting that these genes were subjected to purifying selection over the course of evolution [42]. To investigate the selection pressure on these genes in distinct regions, we performed sliding window analysis of Ka/Ks ratios using the following parameters: window size, 150 bp; step size, 9 bp. The Ka/Ks values reveal that the selection pressure differed among sites with sequence differences. We detected the stronger purifying selection in the CCCH domain, except for *ZmC3H10/ZmC3H34* and *SbC3H45/ZmC3H54* (Figure S3). The Ka/Ks ratios of most sequences were < 1 suggesting that these gene pairs evolved under purifying selection. Purifying selection can remove detrimental mutations and has probably made the CCCH IX sequences consistent across evolutionary history. Hence, the CCCH IX genes are important for plant growth and development.

4. Discussion

CCCH IX genes are thought to play a variety of roles in plant growth, development, and stress resistance [6, 9, 10]. In this study, we selected 27 abiotic stress-responsive CCCH IX genes in *Z. mays*, *O. sativa,* and *S. bicolor* based on phylogenetic tree analysis and their genetic structures, as described in previous reports [5, 11]. The CCCH IX subfamily was characterized into four classes based on our interspecific phylogenetic tree (Figure 2). We determined that these specific genes are almost intronless and that they respond to various adverse environmental factors throughout the plant's life cycle [43].

By calculating the dates of duplication of homologous segments and examining the phylogenetic tree, we determined that the most recent (15 mya) duplication events likely occurred in maize and sorghum. Thus, maize and sorghum have probably undergone a series of evolutionary events and experienced a higher rate of evolution than rice. We observed strong microsynteny among rice, maize, and sorghum genes,

Table 2: Estimation of the dates of large-scale duplication events in three grasses.

Synteny blocks of CCCH IX genes	Number of conserved flanking protein-coding genes	Synonymous sites	Ks (mean ± s.d.)	Date (mya)
OsC3H2 & *OsC3H35*	5	299.25	1.0927 ± 0.4254	84.0538
OsC3H2 & *SbC3H47*	3	292.50	1.0915 ± 0.4868	83.9615
OsC3H2 & *SbC3H12*	13	268.75	0.5470 ± 0.2981	42.0769
OsC3H2 & *ZmC3H51*	3	293.00	0.7827 ± 0.4078	60.2077
OsC3H35 & *SbC3H12*	5	274.42	0.9811 ± 0.3190	75.4692
OsC3H35 & *SbC3H47*	3	296.67	0.4635 ± 0.1314	35.6538
SbC3H12 & *SbC3H47*	3	271.58	0.8632 ± 0.5594	66.4000
SbC3H12 & *ZmC3H51*	3	275.17	0.2552 ± 0.0552	19.6308
OsC3H35 & *ZmC3H38*	3	302.58	0.9811 ± 0.6874	75.4692
OsC3H10 & *OsC3H37*	7	173.67	0.7084 ± 0.2090	54.4923
OsC3H10 & *ZmC3H39*	6	169.08	0.6818 ± 0.1138	52.4462
OsC3H37 & *ZmC3H39*	6	193.00	0.5051 ± 0.0373	38.8538
OsC3H37 & *ZmC3H53*	5	197.83	0.5696 ± 0.1238	43.8154
OsC3H50 & *ZmC3H10*	9	289.33	0.5867 ± 0.1815	45.1285
OsC3H50 & *SbC3H10*	16	515.00	0.5893 ± 0.1738	45.3308
OsC3H50 & *ZmC3H34*	5	298.42	0.6653 ± 0.1309	51.1769
OsC3H50 & *ZmC3H43*	6	501.00	0.6014 ± 0.1371	46.2615
SbC3H10 & *ZmC3H10*	7	292.67	0.1816 ± 0.0469	13.9692
SbC3H10 & *ZmC3H34*	5	307.92	0.1692 ± 0.0579	13.0154
SbC3H10 & *ZmC3H43*	6	514.33	0.1498 ± 0.0200	11.5231
ZmC3H10 & *ZmC3H34*	3	275.50	0.2019 ± 0.0367	15.5308
ZmC3H10 & *ZmC3H43*	4	291.83	0.1990 ± 0.0326	15.3077
ZmC3H34 & *ZmC3H43*	5	310.50	0.0157 ± 0.0352	1.2077
OsC3H24 & *SbC3H2*	14	572.67	0.8338 ± 0.4394	64.1385
OsC3H24 & *ZmC3H4*	4	571.75	0.6299 ± 0.2791	48.4539
OsC3H24 & *ZmC3H28*	5	359.50	0.5979 ± 0.1667	45.9923
SbC3H2 & *ZmC3H4*	4	570.75	0.1853 ± 0.1206	14.2538
SbC3H2 & *ZmC3H28*	5	359.92	0.1921 ± 0.0503	14.7769
ZmC3H4 & *ZmC3H28*	4	361.00	0.2170 ± 0.0754	16.6923
OsC3H33 & *SbC3H45*	9	458.25	0.5963 ± 0.1419	45.8692
SbC3H45 & *ZmC3H54*	3	269.67	0.2567 ± 0.0433	19.7462

but this process is not simple, as transsituation, inversion, loss, and segmental duplication have occurred to varying degrees, which act as the driving force in evolution. Such a process is necessary for the expansion of gene families over the course of evolution. For example, through the analysis of microsynteny, we identified segmental duplication in OsChr1/OsChr5 and ZmChr6/ZmChr8; such duplication frequently occurs among genes.

The CCCH IX microsynteny maps suggested that these genes have been conserved over the course of evolution. Chromosomes contain many syntenic segments that have undergone transsituation, inversion, deletion, and duplication. The gene order has been retained in syntenic segments. In such segments, key genes can be identified from other closely related species on homologous chromosomes at the same relative locations. In the CCCH IX subfamily, *ZmC3H4*/*ZmC3H28*/*ZmC3H54* are stress-response genes, and the expression level of *ZmC3H54* is the highest (Figure S1). Therefore, we can deduce that *OsC3H33*, *SbC3H45*, *OsC3H24*, and *SbC3H2* might be responsive to abiotic stress according to the detailed microsynteny analysis (Figure 5) [11]. This conclusion is supported by the close relationship between ATSZF1, ATSZF2, and CaKR1, which were previously identified as stress-response genes [7, 10]. Our analysis of CCCH IX genes demonstrated that the CCCH IX genes and their flanking protein-coding genes are subjected to purifying selection. Subsequently, we conducted sliding window analysis to detect gene sequences with unusual selection pressure, which provided more insights into the effects of the environment and abiotic stress on this subfamily.

In the current study, comparisons among the genes across the three Gramineae genomic sequences demonstrated that extensive large-scale genome duplication has occurred in the CCCH IX subfamily before the species separated 60–70 mya

[39]. CCCH IX genes have undergone dramatic expansion followed by whole-genome duplication, which led to speciation approximately 60 mya (Figure 6(a)). In general, we found that CCCH IX genes evolved through multiple large-scale duplication events, which are similar to the events that have driven the evolution of protein-coding genes, but the structure, order, and transcriptional orientation of the CCCH IX genes have stayed the same. We then analyzed the evolutionary history of CCCH IX genes in the subsequent tens of thousands of years at the molecular level and performed detailed microsynteny analysis of the abiotic-responsive gene pairs. The results of this study provide a foundation for further investigating the molecular evolution and functions of CCCH IX genes, particularly for members with potentially important roles in regulating abiotic stress responses in plants. However, further experiments should be conducted to directly explore the functions of CCCH IX genes.

5. Conclusion

In this study, we identified and analyzed stress-responsive members of the conserved CCCH IX subfamily through comparative genomic analysis. Some pairs of regions exhibited microsyntenic relationships during evolution according to microsynteny maps. In addition, we calculated the date of duplication by performing Ks analysis and examining Ka/Ks ratios. Through microsynteny analysis, we investigated the evolutionary patterns of *OsC3H24/OsC3H33* and *SbC3H2/SbC3H45*, which function in the response to abiotic stress. The results of this study lay foundation for future functional analyses of these TFs.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Wei-Jun Chen and Yang Zhao contributed equally to this work.

Acknowledgments

The authors thank the members of the Key Laboratory of Crop Biology of Anhui province for their helpful suggestions regarding the experimental design and data processing. This research was supported by grants from the Natural Science Foundation Project of Anhui Province (1508085QC64) and the Biology Key Subject Construction of Anhui.

References

[1] H. Wang, X. Yin, X. Li et al., "Genome-wide identification, evolution and expression analysis of the grape (*Vitis vinifera* L.) zinc finger-homeodomain gene family," *International Journal of Molecular Sciences*, vol. 15, no. 4, pp. 5730–5748, 2014.

[2] D. Arnaud, A. Déjardin, J.-C. Leplé, M.-C. Lesage-Descauses, and G. Pilate, "Genome-wide analysis of LIM gene family in *Populus trichocarpa*, A*rabidopsis thaliana*, and *Oryza sativa*," *DNA Research*, vol. 14, no. 3, pp. 103–116, 2007.

[3] S. Yanagisawa, "Dof domain proteins: plant-specific transcription factors associated with diverse phenomena unique to plants," *Plant and Cell Physiology*, vol. 45, no. 4, pp. 386–391, 2004.

[4] P. Kosarev, K. F. Mayer, and C. S. Hardtke, "Evaluation and classification of RING-finger domains encoded by the *Arabidopsis* genome," *Genome Biology*, vol. 3, no. 4, pp. 16.1–16.12, 2002.

[5] D. Wang, Y. Guo, C. Wu, G. Yang, Y. Li, and C. Zheng, "Genome-wide analysis of CCCH zinc finger family in *Arabidopsis* and rice," *BMC Genomics*, vol. 9, article 44, 2008.

[6] Z. Kong, M. Li, W. Yang, W. Xu, and Y. Xue, "A novel nuclear-localized CCCH-type zinc finger protein, OsDOS, is involved in delaying leaf senescence in rice," *Plant Physiology*, vol. 141, no. 4, pp. 1376–1388, 2006.

[7] E. S. Seong, D. Choi, S. C. Hye, K. L. Chun, J. C. Hye, and M.-H. Wang, "Characterization of a stress-responsive ankyrin repeat-containing zinc finger protein of *Capsicum annuum* (CaKR1)," *Journal of Biochemistry and Molecular Biology*, vol. 40, no. 6, pp. 952–958, 2007.

[8] H. Deng, H. Liu, X. Li, J. Xiao, and S. Wang, "A CCCH-type zinc finger nucleic acid-binding protein quantitatively confers resistance against rice bacterial blight disease," *Plant Physiology*, vol. 158, no. 2, pp. 876–889, 2012.

[9] J. Li, D. Jia, and X. Chen, "HUA1, a regulator of stamen and carpel identities in *Arabidopsis*, codes for a nuclear RNA binding protein," *Plant Cell*, vol. 13, no. 10, pp. 2269–2281, 2001.

[10] J. Sun, H. Jiang, Y. Xu et al., "The CCCH-type zinc finger proteins AtSZF1 and AtSZF2 regulate salt stress responses in *Arabidopsis*," *Plant and Cell Physiology*, vol. 48, no. 8, pp. 1148–1158, 2007.

[11] X. Peng, Y. Zhao, J. Cao et al., "CCCH-type zinc finger family in maize: genome-wide identification, classification and expression profiling under abscisic acid and drought treatments," *PLoS ONE*, vol. 7, no. 7, Article ID e40120, 2012.

[12] E.-C. Oerke and H.-W. Dehne, "Safeguarding production—losses in major crops and the role of crop protection," *Crop Protection*, vol. 23, no. 4, pp. 275–285, 2004.

[13] A. Lawton-Rauh, "Evolutionary dynamics of duplicated genes in plants," *Molecular Phylogenetics and Evolution*, vol. 29, no. 3, pp. 396–409, 2003.

[14] F. Choulet, T. Wicker, C. Rustenholz et al., "Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces," *The Plant Cell*, vol. 22, no. 6, pp. 1686–1701, 2010.

[15] K. Vandepoele, C. Simillion, and Y. Van de Peer, "Evidence that rice and other cereals are ancient aneuploids," *The Plant Cell*, vol. 15, no. 9, pp. 2192–2202, 2003.

[16] R. S. Sekhon, H. Lin, K. L. Childs et al., "Genome-wide atlas of transcription during maize development," *The Plant Journal*, vol. 66, no. 4, pp. 553–563, 2011.

[17] T. Tvedebrink and J. Curran, "DNAtools: statistical functions for analysing forensic DNA databases," R package version 0.1-2, 2011.

[18] R. D. Finn, J. Mistry, B. Schuster-Böckler et al., "Pfam: clans, web tools and services," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D247–D251, 2006.

[19] I. Letunic, T. Doerks, and P. Bork, "SMART 6: recent updates and new developments," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D229–D232, 2009.

[20] M. A. Larkin, G. Blackshields, N. P. Brown et al., "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[21] M. R. Wilkins, E. Gasteiger, A. Bairoch et al., *Protein Identification and Analysis Tools in the ExPASy Server*, Humana Press, 1999.

[22] A. Y. Guo, Q. H. Zhu, X. Chen et al., "GSDS: a gene structure display server," *Yi Chuan*, vol. 29, no. 8, pp. 1023–1026, 2007.

[23] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: molecular evolutionary genetics analysis version 6.0," *Molecular Biology & Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[24] C. Maher, L. Stein, and D. Ware, "Evolution of *Arabidopsis* microRNA families through duplication events," *Genome Research*, vol. 16, no. 4, pp. 510–519, 2006.

[25] Z. Yang, Y. Zhou, X. Wang et al., "Genomewide comparative phylogenetic and molecular evolutionary analysis of tubby-like protein family in *Arabidopsis*, rice, and poplar," *Genomics*, vol. 92, no. 4, pp. 246–253, 2008.

[26] C. Kidgell, S. K. Volkman, J. Daily et al., "A systematic map of genetic variation in *Plasmodium falciparum*," *PLoS Pathogens*, vol. 2, no. 6, article e57, 2006.

[27] R. W. Lusk and M. B. Eisen, "Evolutionary mirages: Selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers," *PLoS Genetics*, vol. 6, no. 1, Article ID e1000829, 2010.

[28] S. Sato, Y. Nakamura, T. Kaneko et al., "Genome structure of the legume, *Lotus japonicus*," *DNA Research*, vol. 15, no. 4, pp. 227–239, 2008.

[29] Z. Li, H. Jiang, L. Zhou et al., "Molecular evolution of the HD-ZIP I gene family in legume genomes," *Gene*, vol. 533, no. 1, pp. 218–228, 2014.

[30] M. Krzywinski, J. Schein, I. Birol et al., "Circos: an information aesthetic for comparative genomics," *Genome Research*, vol. 19, no. 9, pp. 1639–1645, 2009.

[31] M. Nei and T. Gojobori, "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions," *Molecular Biology and Evolution*, vol. 3, no. 5, pp. 418–426, 1986.

[32] P. Librado and J. Rozas, "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data," *Bioinformatics*, vol. 25, no. 11, pp. 1451–1452, 2009.

[33] A. Nekrutenko, K. D. Makova, and W.-H. Li, "The KA/KS ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study," *Genome Research*, vol. 12, no. 1, pp. 198–202, 2002.

[34] B. S. Gaut, B. R. Morton, B. C. Mccaig, and M. T. Clegg, "Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 19, pp. 10274–10279, 1996.

[35] M. Lynch and J. S. Conery, "The evolutionary fate and consequences of duplicate genes," *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000.

[36] D. Michael and L. Manyuan, "Intron—exon structures of eukaryotic model organisms," *Nucleic Acids Research*, vol. 27, no. 15, pp. 3219–3228, 1999.

[37] H.-K. Choi, J.-H. Mun, D.-J. Kim et al., "Estimating genome conservation between crop and model legume species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 43, pp. 15289–15294, 2004.

[38] J. L. Bennetzen, "Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions," *The Plant Cell*, vol. 12, no. 7, pp. 1021–1029, 2000.

[39] E. P. Rocha, J. M. Smith, L. D. Hurst et al., "Comparisons of dN/dS are time dependent for closely related bacterial genomes," *Journal of Theoretical Biology*, vol. 239, no. 2, pp. 226–235, 2006.

[40] G. Blanc and K. H. Wolfe, "Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes," *The Plant Cell*, vol. 16, no. 7, pp. 1667–1678, 2004.

[41] G. Blanc, K. Hokamp, and K. H. Wolfe, "A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome," *Genome Research*, vol. 13, no. 2, pp. 137–144, 2003.

[42] F. Burki and H. Kaessmann, "Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux," *Nature Genetics*, vol. 36, no. 10, pp. 1061–1063, 2004.

[43] H. Tsukaya, "Leaf shape: genetic controls and environmental factors," *International Journal of Developmental Biology*, vol. 49, no. 5-6, pp. 547–555, 2005.

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 473028, 11 pages
http://dx.doi.org/10.1155/2015/473028

Research Article

Characterization and Development of EST-SSRs by Deep Transcriptome Sequencing in Chinese Cabbage (*Brassica rapa* L. ssp. *pekinensis*)

Qian Ding,[1] Jingjuan Li,[2] Fengde Wang,[2] Yihui Zhang,[2] Huayin Li,[2] Jiannong Zhang,[1] and Jianwei Gao[2]

[1]*College of Horticulture, Gansu Agricultural University, Lanzhou 730070, China*
[2]*Institute of Vegetables and Flowers, Shandong Academy of Agricultural Sciences and Shandong Key Laboratory of Greenhouse Vegetable Biology and Shandong Branch of National Vegetable Improvement Center, Jinan 250100, China*

Correspondence should be addressed to Jiannong Zhang; zjn0101@sina.com and Jianwei Gao; jianweigao3@yahoo.com

Received 15 February 2015; Accepted 26 March 2015

Academic Editor: Yanbin Yin

Copyright © 2015 Qian Ding et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Simple sequence repeats (SSRs) are among the most important markers for population analysis and have been widely used in plant genetic mapping and molecular breeding. Expressed sequence tag-SSR (EST-SSR) markers, located in the coding regions, are potentially more efficient for QTL mapping, gene targeting, and marker-assisted breeding. In this study, we investigated 51,694 nonredundant unigenes, assembled from clean reads from deep transcriptome sequencing with a Solexa/Illumina platform, for identification and development of EST-SSRs in Chinese cabbage. In total, 10,420 EST-SSRs with over 12 bp were identified and characterized, among which 2744 EST-SSRs are new and 2317 are known ones showing polymorphism with previously reported SSRs. A total of 7877 PCR primer pairs for 1561 EST-SSR loci were designed, and primer pairs for twenty-four EST-SSRs were selected for primer evaluation. In nineteen EST-SSR loci (79.2%), amplicons were successfully generated with high quality. Seventeen (89.5%) showed polymorphism in twenty-four cultivars of Chinese cabbage. The polymorphic alleles of each polymorphic locus were sequenced, and the results showed that most polymorphisms were due to variations of SSR repeat motifs. The EST-SSRs identified and characterized in this study have important implications for developing new tools for genetics and molecular breeding in Chinese cabbage.

1. Introduction

Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*) is a diploid ($2n = 2x = 20$) dicot with a genomic size of 550 Mb (http://www.brassica.info/resource/). It is a subspecies of *B. rapa* with the A genome [1]. The species originated in China and now has become one of the most important and widely cultivated leaf vegetables in Asia. Chinese cabbage has rosette leaves (RLs) and folding leaves (FLs). The tight leafy head is the main edible part. After a long history of domestication, Chinese cabbage evolves into different cultivars with a variety of characteristics, such as rosette leaf morphology, heading leaf morphology, leafy head shape, size, and structure, flowering time, nutrient composition, and resistance to biotic and abiotic. A better understanding of the molecular mechanism of evolution of Chinese cabbage and further development of marker-assisted selection (MAS) will accelerate the selection process of improved cultivars to meet the growing consumers and environmental needs. Although progress has been made in underlining the molecular mechanism [2–5], many aspects are still unclear.

Molecular markers have been widely used to study the genetic basis of important traits and map regulatory genes in plants. Markers tightly linked with important agronomic traits can potentially be used for molecular breeding to develop improved cultivars. Many molecular markers and genetic maps of Chinese cabbage have been reported previously [6–25]. However, there is still a great need to develop novel molecular markers for construction of high-density linkage maps for genetics and molecular studies of important traits in Chinese cabbage.

Simple sequence repeat (SSR) markers or microsatellite markers are among the most important markers in plants. SSRs have been widely used in genetic mapping and molecular breeding in plants because they are highly abundant and have significant polymorphism. Other factors, like accessibility for detection, reliability, and codominance, also make them perfect markers for such purposes [26]. SSRs found in transcribed sequences are called expressed sequence-simple sequence repeats (EST-SSRs). Compared with genomic-SSRs detected in noncoding sequences, EST-SSRs are more efficient for QTL mapping, gene targeting, and MAS [27]. As transcribed sequences are more conserved than noncoding sequences, the transferability of EST-SSRs is better than genomic-SSRs [28–30], which can be utilized for cross genome comparison and evolutionary analysis [27, 31]. Additionally, abundant ESTs were generated in recent years with the development of next-generation sequencing approaches, making identification of EST-SSRs more practical and cost-efficient [32]. Many EST-SSRs have been identified in Chinese cabbage [16, 20, 25, 33–36]. Because whole genome sequencing of Chinese cabbage is still underway, new EST-SSRs could also be identified for further studies such as high-density genetic linkage map construction, gene/QTL mapping, and cultivar identification.

In our previous study, the whole transcriptomes were analyzed for the rosette leaves and folding leaves of a typical heading Chinese cabbage, namely, FuShanBaoTou, using a Solexa/Illumina RNA-Seq platform, and a large-scale EST database was generated [37]. In this study, we further assembled those ESTs from the RL and FL libraries into nonredundant unigenes. A total of 10,420 EST-SSRs were identified, among which 2744 EST-SSRs are detected for the first time, according to the SSR marker database for *Brassica* (http://oilcrops.info/SSRdb). We characterized these identified EST-SSRs and designed 7877 PCR primer pairs for 1561 EST-SSRs. Furthermore, serving as a validation purpose, we tested polymorphisms of 24 EST-SSRs. We expect this study can pave the road for further investigation of new EST-SSR markers and for construction of high-density genetic maps.

2. Materials and Methods

2.1. Plant Materials. For EST-SSR identification and primer design, a typical heading Chinese cabbage, namely, FuShanBaoTou, was used in this study. For primer assessment and SSR polymorphism analysis, a panel of twenty-four cultivars of Chinese cabbage was used, including nineteen morphologically diverse cultivars of *Brassica rapa* L. ssp. *pekinensis* (*B. pekinensis* L.) and five *Brassica rapa* L. *chinensis* (*B. chinensis* L.). All plants were grown in a greenhouse with 16/8 photoperiod at $22 \pm 2^{\circ}$C. Leaves were collected after they were grown for two weeks from ten seedlings of each cultivar and were pooled together for DNA extraction.

2.2. De Novo Assembly. We assembled the clean read dataset presented by Wang et al. [37] from the RL and FL libraries according to the methods described by Wang et al. [38] using the Trinity software (http://trinityrnaseq.sourceforge.net/). Contigs and unigenes were obtained from these two libraries, respectively. Redundant sequences were removed and overlapping unigenes were assembled into continuous sequences by the TIGR Gene Indices Clustering (TGICL) tools [39]. Similarity was set at 94% and an overlap length was set at 100 bp.

2.3. Identification of EST-Derived SSRs and Primer Design. SSRs were detected with the MicroSAtellite software (MISA; http://pgrc.ipk-gatersleben.de/misa/). Parameters were set with a minimum number of 12, 6, 5, 5, 4, and 4 repeat units for identification of mono-, di-, tri-, tetra-, penta-, and hexanucleotide motifs, respectively. Primers were designed using primer 3 with no SSR allowed in primers. Primer length ranged from 18 to 28 bp (with an optimality at 23). Annealing temperature was set at 55–65°C (with an optimality at 60°C). The size of a PCR product ranged from 80 to 300 bp.

2.4. Mapping EST-SSRs. The physical positions of the EST-SSRs identified in the study were determined by aligning the SSRs and flanking sequences (50 bp at each side) to the *Brassica rapa* (Chiifu-401) reference genome (http://brassicadb.org/brad/) using BLASTN. New EST-SSRs were identified by comparing with previously reported SSRs in the SSR marker database for *Brassica* (http://oilcrops.info/SSRdb) [25].

2.5. SSR Amplification and SSR Polymorphism Analysis. DNA was extracted following a CTAB DNA extraction protocol [40]. The DNA sample of the Chinese cabbage FuShanBaoTou was used as template to detect the availability of SSR primers designed above. The DNA samples of those aforementioned twenty-four cultivars of Chinese cabbage were used as templates for SSR polymorphism analysis. The polymorphisms of EST-SSRs were validated by 6% denaturing polyacrylamide gel electrophoresis, 12% nondenaturing polyacrylamide gel electrophoresis, and sequencing.

3. Results

3.1. De Novo Assembly. High quality clean read data from the RL and FL libraries by Wang et al. [37] were assembled using the Trinity software package [41]. A total of 99,684 and 95,411 contigs were obtained, with an average length of 333 and 342 bp and a median length (N50) of 531 and 536 bp, from the RL and FL libraries, respectively (Table 1).

Contigs from the same transcript were detected with paired-end reads, as well as the distances between these contigs. Using the Trinity software package, we assembled these contigs into unigenes, in which Ns were removed. These unigenes were set to be not extendable on either end of the sequences. A total of 46,294 and 48,473 unigenes from the RL and FL libraries were obtained with an average length of 707 and 680 bp and a median length (N50) of 1000 and 980 bp, respectively (Table 1). Size distribution of the contigs and unigenes is consistent with the RL and FL libraries as shown in Figure 1, indicating that our Illumina sequencing solution is reliable and reproducible. Unigenes from the two samples were combined; redundant unigenes were removed;

Table 1: Overview of the sequencing and assembly.

	Sample	Total number	Total length (nt)	Average length (nt)	N50	Total consensus sequences	Distinct clusters	Distinct singletons
Contig	RL	99,684	33,205,708	333	531	—	—	—
	FL	95,441	32,596,297	342	536	—	—	—
Unigene	RL	46,294	32,729,586	707	1000	46,294	19,512	26,782
	FL	48,473	32,971,187	680	980	48,473	19,749	28,724
	All	51,694	40,724,256	788	1154	51,694	23,850	27,844

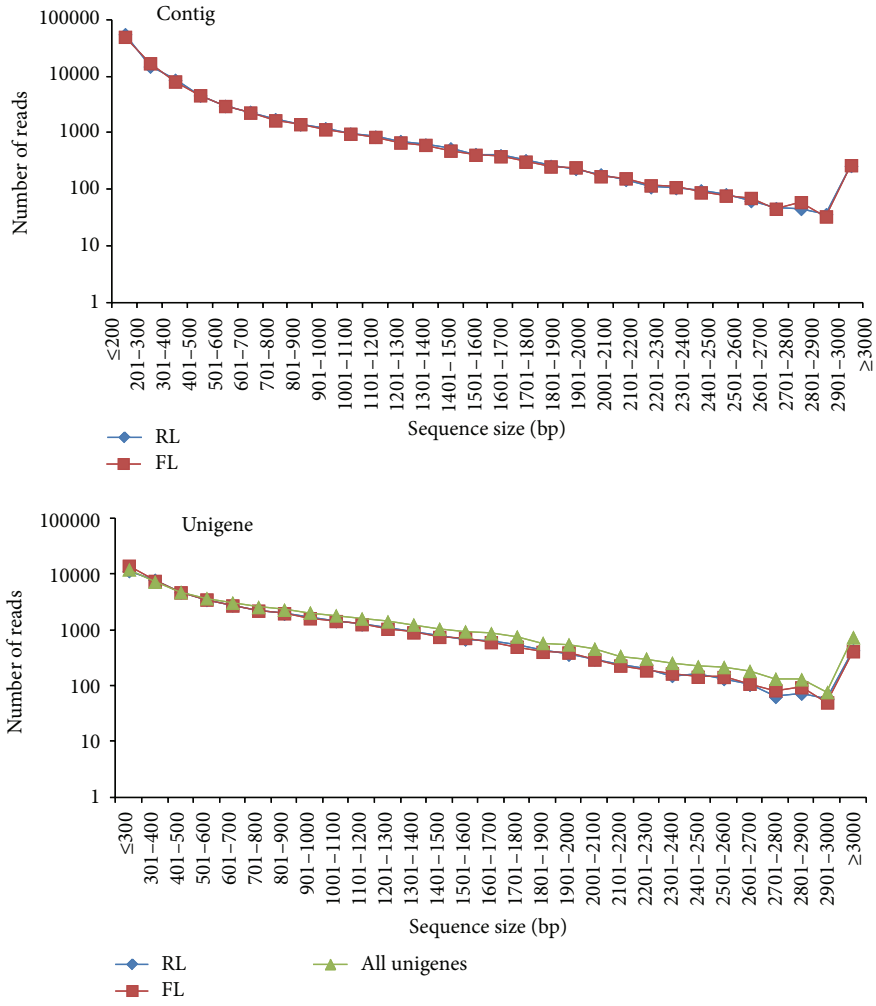


Figure 1: Size distribution of the assembled contigs and unigenes in RL and FL libraries.

and the rest was assembled with TGICL [39] to form a single dataset, which represents 40.7 Mb of sequence and contains a total of 51,694 nonredundant unigenes, with an average read length of 788 bp, and a median read length (N50) of 1154 bp (Table 1). The sequences of the unigenes are listed in Table s1 (see Supplementary Material available online at http://dx.doi.org/10.1155/2015/473028).

The length of 24,271 nonredundant unigenes (46.95%) is between 200 and 500 bp; the length of 13,613 (26.33%) is between 501 and 1,000 bp, and the length of 13,810 (26.72%) is longer than 1,000 bp (Figure 1).

3.2. Characterization of EST-SSRs in Chinese Cabbage. A total of 10420 EST-SSRs were detected with the MicroSAtellite software (MISA; http://pgrc.ipk-gatersleben.de/misa/) in 8571 unigenes, accounting for 16.6% of total nonredundant unigenes (Tables 2 and s2). The mean SSR density is one per 3.9 Kb, corresponding to one for every 5.0 nonredundant unigenes. 1502 unigenes (17.5%) harbored more than one SSR and 666 SSRs (6.4%) were present in compound formation that had more than one repeat type (Table 2).

The size of SSR repeat units ranged from one to six. The number of SSRs with each repeat unit was found to be

Table 2: Summary of EST-SSR searching results.

Searching items	Numbers
Total number of sequences examined	51694
Total size of examined sequences (bp)	40724256
Total number of identified SSRs	10420
Number of SSR-containing sequences	8571
Number of sequences containing more than one SSR	1502
Number of SSRs present in compound formation	666
% EST-SSRs	16.6%

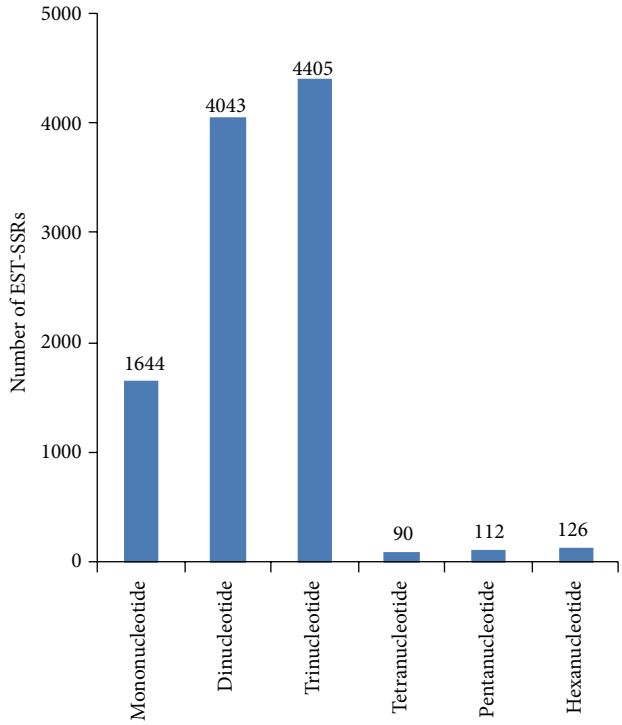


Figure 2: EST-SSR statistics.

quite different. The SSRs with tri- and dinucleotide repeat motifs were the most common (4,405, 42.27%; 4,043, 38.80%, resp.), followed by mono- (1,644, 15.78%), hexa- (126, 1.21%), penta- (112, 1.07%) and tetra- (90, 0.86%) nucleotide repeat motifs (Figure 2). The most common two repeat motif types accounted for 81.07% of the total SSRs detected, and the rest repeat motifs types only accounted for 18.93%.

The iterate number of repeat units in an EST-SSR ranged from 4 to 25. The occurrence frequency of EST-SSTs with different iterate numbers was found to be unequal either. EST-SSRs with iterate number of 5 (2832, 27.18%) were the most common ones, followed by 6 (2739, 26.29%), 7 (1368, 13.13%), 8 (703, 6.75%), 12 (542, 5.20%), and 9 (480, 4.61%) (Table s3). A dinucleotide containing EST-SSRs with a maximum of 25 repeat units was identified. For EST-SSRs with more than 10 repeat units, the mononucleotide repeat motifs were the most abundant, accounting for 93.46% of these EST-SSRs. The lengths of EST-SSR sequences ranged from 12 to 65 bp (Table s4). The longest one is a pentanucleotide containing EST-SSR with 65 bp in length. The lengths of most EST-SSRs are from 12 to 20 bp, accounting for 91.47% of the total EST-SSRs, followed by EST-SSRs with 21–30 bp in length (874 SSRs, 8.39%). Only 13 EST-SSRs were identified with over 30 bp, accounting for 0.12% of the total EST-SSRs.

A total of 124 EST-SSR motifs were identified, including 2 mono-, 3 di-, 10 tri-, 13 tetra-, 33 penta-, and 63 hexanucleotide repeat units containing EST-SSRs. The dominant motif identified in our EST-SSRs was AG/CT (3,519, 33.8%), followed by A/T (1,562, 15.0%), AAG/CTT (1,445, 13.9%), AGG/CCT (776, 7.4%), ATC/ATG (627, 6.0%), AAC/GTT (392, 4.4%), ACC/GGT (392, 3.8%), AC/GT (349, 3.3%), and AGC/CTG (317, 3.0%) (Figure 3). The other 115 motifs have low frequency, accounting only for 9.3% of total EST-SSRs.

Physical locations of the EST-SSRs were assigned by searching against the nonredundant (nr) protein database of NCBI (http://www.ncbi.nlm.nih.gov/) and the *Brassica* database (http://brassicadb.org/brad/) using BLASTX. Our results showed that 4329 EST-SSRs (44.4%) were located in coding regions (CDSs), 3456 (35.5%) in 5′-UTRs, and 1297 (13.3%) in 3′-UTRs (Figure 4, Table s4). Locations of the remaining 672 EST-SSRs (6.9%) were not successfully assigned (Figure 4, Table s4). For the EST-SSRs localized in the CDS region, trinucleotide repeats were the most common ones, accounting for 62.72% of the total EST-SSRs localized in this region, followed by dinucleotide repeats (897, 20.72%), mononucleotide repeats (325, 7.51), and compound formation (287, 6.63%) (Table s4). Dinucleotide repeats (1909, 55.24%) were the dominant types in 5′-UTRs, followed by trinucleotide repeats (730, 21.12%), mononucleotide repeats (483, 13.98%), and compound formation ones (214, 6.19%) (Table s4). Mono-, di-, and trinucleotide repeat EST-SSRs were the top three types found in 3′-UTRs, accounting for 35.08%, 30.07%, and 28.60% of the total EST-SSRs localized in these regions, respectively.

3.3. New EST-SSRs Identification. The EST-SSRs and the flanking sequences (50 bp on each side) were aligned to the *Brassica rapa* (Chiifu-401) reference genome (http://brassicadb.org/brad/) using BLASTN to determine their physical positions. New EST-SSRs were identified by comparing with the earlier reported SSRs in the SSR marker database for *Brassica* (http://oilcrops.info/SSRdb). A total of 2744 new EST-SSRs (26.3%) were identified in the study. Of the 7676 known SSRs (73.6%), 2317 EST-SSRs (22.2%) show polymorphism with different repeat numbers, and 5359 (51.4%) were exactly the same with the earlier reported SSRs based on the *Brassica rapa* (Chiifu-401) genomic sequence [25] (Table s2).

3.4. Primer Design and Evaluation of EST-SSRs in Chinese Cabbage. A total of 7877 PCR primer pairs from the unique sequences flanking 1561 EST-SSR loci were designed according to the criteria described in Section 2 using primer 3 (Table s5). For each EST-SSR locus, a maximum of 5 alternative primer pairs was designed. The other 8859 EST-SSRs, which had no appropriate PCR primer pairs designed as their flanking sequences, did not fulfill the primer design criteria

Table 3: Details of 19 EST-SSRs that successfully yielded PCR amplicons in FuShanBaoTou.

Code	EST-SSR name	Motif	Product size expected (bp)	Product size validated (bp)	SSR location	Number of alleles
BR-es1	CL3455.Contig1_All-2	(TC)11	97	97	CDS	2
BR-es2	CL4114.Contig2_All-2	(TCA)7	157	157	3-UTR	3
BR-es3	CL2525.Contig4_All-1	(TAG)9	160	160	CDS	4
BR-es4	Unigene10387_All-1	(CTC)9	127	127	CDS	2
BR-es5	CL7077.Contig2_All-1	(AATC)5	153	153	3-UTR	3
BR-es6	Unigene16359_All-1	(AACC)5	**134**	**130**	5-UTR	1
BR-es7	CL4685.Contig1_All-1	(CCTT)6	**160**	**148**	3-UTR	2
BR-es8	CL5247.Contig3_All-1	(TTTC)6	**133**	**141**	CDS	3
BR-es9	CL3462.Contig4_All-1	(AATCG)4	155	155	CDS	2
BR-es10	CL5726.Contig2_All-2	(TCTCT)4	146	146	5-UTR	3
BR-es11	Unigene6713_All-1	(AAAAC)4	118	118	CDS	1
BR-es12	CL7282.Contig2_All-1	(GAGGA)5	**140**	**117**	CDS	4
BR-es13	Unigene2970_All-1	(GAACT)5	106	106	3-UTR	3
BR-es14	Unigene8739_All-1	(GATTT)5	130	130	5-UTR	4
BR-es15	CL5873.Contig4_All-2	(CCCTAA)4	146	146	3-UTR	4
BR-es16	Unigene14449_All-1	(CTCAAG)5	**99**	**185**	CDS	6
BR-es17	Unigene5096_All-1	(ACTCCC)5	141	141	CDS	3
BR-es18	CL4691.Contig2_All-1	(GATGGT)7	**155**	**117**	CDS	6
BR-es19	Unigene13507_All-1	(ATTTG)4	152	152	CDS	2

EST-SSRs shown in bold have sizes different from the expected sizes.

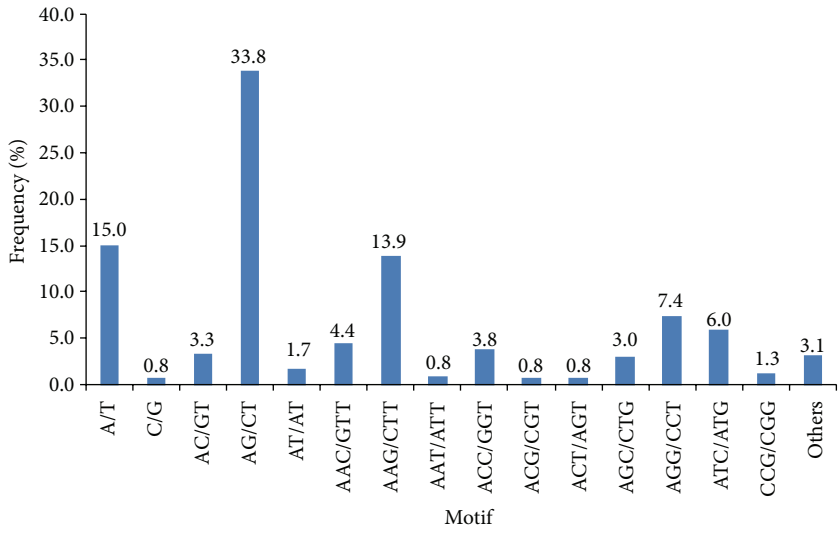


Figure 3: Frequency distribution of EST-SSRs according to motif sequence types.

mentioned above. For the 1561 EST-SSRs with PCR primers designed, PCR primers of those aforementioned 24 loci with $n \geq 20$ bp were selected for primer synthesis and amplification evaluation in Chinese cabbage FuShanBaoTou. Nineteen (79.2%) of these 24 EST-SSR loci successfully yielded PCR amplicons in FuShanBaoTou. We sequenced these nineteen PCR amplicons and found that the amplicons in thirteen loci were exactly the same as expected; two were longer than the expected size, and four were shorter (Table 3). Size deviation of five EST-SSRs loci with the expected sizes (BR-es6, BR-es7, BR-es8, BR-es12, and BR-es18) was due to the variations of SSR repeat motifs (Table s6). One amplicon (BR-es16) deviated from the expected sizes and had an additional 86 bp containing a $(TC)_9$ motif near the SSR repeat motif region (Table s6).

3.5. Validation of Polymorphism of EST-SSRs. Nineteen effective primer pairs were used for polymorphism validation for these aforementioned 24 Chinese cabbage cultivars. The results showed that 17 loci (89.5%) were polymorphic (Figure 5). A total of 56 alleles at the 17 polymorphic loci were identified and the average number of alleles per SSR locus was 3.29 with a range between 2 and 6. A maximum of 6 alleles was detected for BR-es16 and BR-es18 loci. BR-es6 and

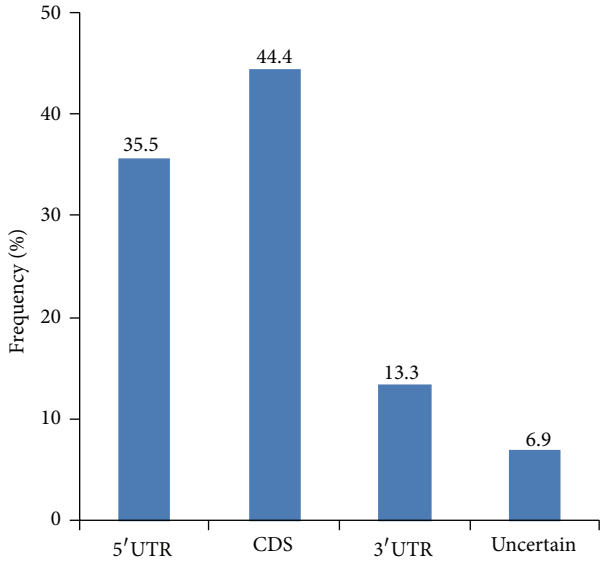


Figure 4: Frequency distribution of EST-SSRs based on locations.

BR-es11 had no polymorphic allele in all 24 cultivars in this study (Figure 5, Tables 3 and s4). Of the 17 polymorphic loci, twelve loci were polymorphic in all cultivars of *B. pekinensis* L. and *B. chinensis* L. Three loci (BR-es2, BR-es9, and BR-es19) had no polymorphism in the cultivars of *B. pekinensis* L. but had polymorphism in the cultivars of *B. chinensis* L., while two loci (BR-es4 and BR-es7) were polymorphic in the cultivars of *B. pekinensis* L. but were not polymorphic in the cultivars of *B. chinensis* L. (Figure 5, Table s8).

We sequenced the polymorphic alleles of the 17 polymorphic loci and found that polymorphisms of 9 loci (BR-es1, BR-es4, BR-es7, BR-es8, BR-es10, BR-es14, BR-es17, BR-es18, and BR-es19) were because of different iterate numbers of SSR repeat motifs. In another 6 polymorphic loci (BR-es2, BR-es3, BR-es12, BR-es13, BR-es15, and BR-es16), the most polymorphic alleles were found in the repeat motifs with additional changes in other regions (Table s7). For example, compared with the allele BR-es3-160 bp in FuShanBaoTou, the polymorphic alleles BR-es3-163 bp and 145 bp had different iterate numbers of the TAG/ATC repeat motif, while the polymorphic allele 99 bp had not only a different number of the repeat motif, but also a deletion in another region (Table s7). The other two polymorphic loci, BR-es5 and BR-es9, had polymorphisms that are not related with the repeat numbers of SSR motifs (Table s7).

4. Discussion

4.1. High-Throughput RNA Sequencing Provides Substantial Knowledge for EST-SSRs. Illumina paired-end RNA sequencing is one of the fast immerging next-generation sequencing (NGS) technologies. Because of its advantages in high-throughput, high accuracy, and low cost, Illumina paired-end sequencing has been widely used for *de novo* transcriptome sequencing and assembly and transcriptome quality and quantity analysis in many plants [37, 38, 42, 43]. In our previous study, the transcriptome of rosette and folding leaves in Chinese cabbage was analyzed using the Illumina paired-end RNA sequencing technology, and abundant clean reads and ESTs with high quality were obtained [37]. The large quantity of clean reads would increase coverage depth of transcriptome nucleotide, enhance sequencing accuracy, and provide useful information for developing new tools for genetic mapping and molecular breeding of Chinese cabbage. In this study, we further assembled the clean reads into contigs and unigenes from the RL and FL libraries, respectively. The parameters for both contigs and unigenes between the two libraries had no significant differences (Table 1), indicating our Illumina sequencing solutions have high reliability and reproducibility. The unigenes of the two libraries were further assembled and a total of 51,694 nonredundant unigenes were obtained from the 40.7 Mb sequence data. We discovered more nonredundant unigenes than those in previous studies [35, 36], which represent a large portion of the Chinese cabbage transcriptome and are important for a comprehensive understanding of EST-SSRs.

4.2. Frequency and Distribution of EST-SSRs in Chinese Cabbage. A total of 10,420 SSRs with over 12 bp were identified from the deep transcriptome sequence dataset of Chinese cabbage. About 16.6% of the unigenes have SSRs. The frequency of occurrence of SSRs is slightly higher than those reported in previous studies on Chinese cabbage (about 8.4–15.6%) [20, 34–36] and also higher than those of other dicotyledonous species such as peanut (6.8%) [44], sweetpotato (8.2%) [21], sesame (8.9%) [43], pigeonpea (7.6%) [45], grapes (2.5%) [46], pepper (4.9%) [47], and flax (3.5%) [48], but it is lower than those of coffee (18.5%) [49], radish (23.8%) [38], and caster bean (28.4%) [50]. Detection of EST-SSRs depends on a number of factors such as genome structure [51], tools and parameters for EST-SSRs detection and exploration [43], and size of dataset for unigene assembly [27].

The frequency of SSRs with different sizes of repeat units is not evenly distributed in plants. Previous studies showed dinucleotide SSR loci are the most abundant class in safflower [52], pigeonpea [45], and sesame [43], whereas trinucleotide repeats are the most frequent ones in barley [53], sweetpotato [21], *Jatropha curcas* [54], iris [55], pepper [47], caster bean [50], flax [48], *Cucurbita pepo* [56], and radish [38]. In ramie [57] and wheat [58], dinucleotide and trinucleotide repeat motifs are the most two abundant types. In the present study, trinucleotide (4405, 42.3%) was found to be the most common repeat motif class in Chinese cabbage, followed by dinucleotide (4043, 38.8%) (Figure 2). It is consistent with previous reports for SSRs identification from unigenes of Chinese cabbage [20]. However, on the genomic level, of Chinese cabbage, dinucleotide is the most common repeat motif, followed by trinucleotide [25].

We found the most dominant mononucleotide repeat motif in Chinese cabbage was A/T (1,562, accounting for 15.0% of the total EST-SSRs), which is consistent with previous reports for Chinese cabbage [25] and for other plants such as *Arabidopsis* [59], rice [59], wheat [60], radish [38], castor bean [50], *Gossypium raimondii* [61], oil palm [62],

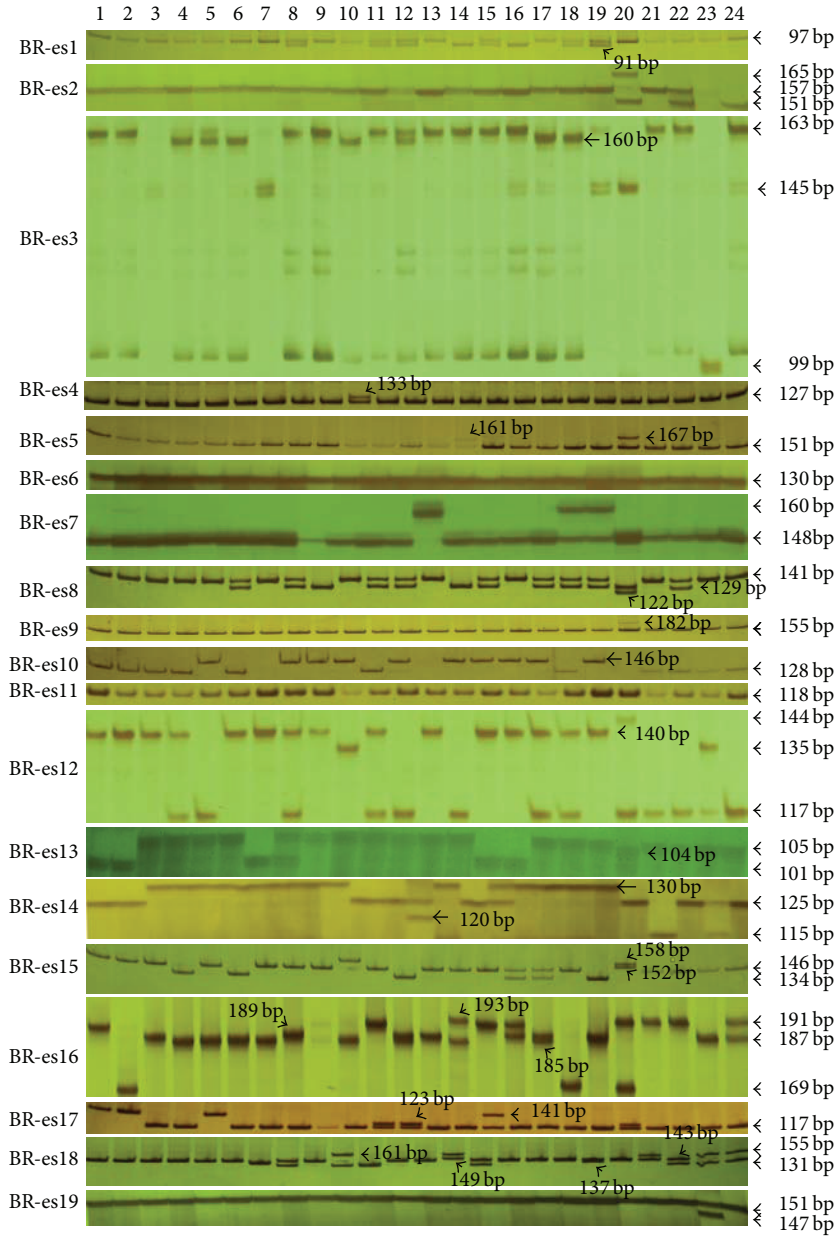


Figure 5: PCR products amplified by nineteen effective EST-SSR primer pairs in twenty-four cultivars of Chinese cabbage. The order of DNA samples from lane 1 to lane 24 within each primer pair image panel is 682, GuangDongZao, ZaoHuangBai, Z61-8, FuShanBaoTou, Li-3, 212-7, TianJinQingMaYe, KuaiCai number 6-5, JinHuangXiaoBaiCai, SiJiXiaoBaiCai, SiJiHuangYangXiaoBaiCai, PinZao number 1, HanYuTeXuanHuangXin, QuanNengSiJiKuaiCai, JingYouXiaoBaiCaiKuaiCai, GaoLiWaWaCai, KeYiXiaWaWa, JinNuoChunQiuWaWaCai, SiJiLvGanXiaoKuCai, YouLv157, ShuYaoYouCai, DeGaoYouLiangQingGengCai, and QingXiuF1QingGengCai. PCR products amplified by BR-es2, BR-es3, BR-es6, BR-es7, BR-es12, BR-es13, BR-es14, BR-es16, and BR-es19 primer pairs were separated on 6% denaturing polyacrylamide gels, while those amplified by BR-es1, BR-es4, BR-es5, BR-es8, BR-es9, BR-es10, BR-es11, BR-es15, BR-es17, and BR-es18 primer pairs were separated on 12% nondenaturing polyacrylamide gels.

and eggplant [63]. For dinucleotide motif, AG/CT was the most common repeat motif, accounting for 87.0% of the total dinucleotide EST-SSRs. It is in close agreement with the results in previous studies for genic SSRs in Chinese cabbage [20, 36] and those in most other plants such as sweetpotato [21], iris [55], sesame [43], and radish [38]. The AG/CT repeat motif was also the most dominant repeat among all the EST-SSRs identified in this study, accounting for 33.8% of the total EST-SSRs. However, for genomic-SSRs in Chinese cabbage, AT/TA is the most common dinucleotide motif [25]. The AAG/CTT (1,445, 13.9%) motif was the most frequent motif among trinucleotide EST-SSRs in the study, which is consistent with the results in previous studies in Chinese cabbage [25, 36] and many dicot species, for example, *Arabidopsis* [64], soybean [65], peanut [44], sweetpotato [21], radish [38], and sesame [43]. In many monocot species such

as maize, barley, and sorghum [66, 67], CCG/GGC is the most dominant trinucleotide repeat motif. It is considered a specific feature of monocot genomes due to the high GC content in monocot genomes [68].

4.3. New EST-SSRs Identification. Of all 10420 EST-SSRs identified in this study, more than 70% have been identified and presented in the SSR marker database (http://oilcrops.info/SSRdb), among which over half were exactly the same with the earlier reported SSRs based on the *Brassica rapa* (Chiifu-401) genomic sequence (Table s2) [25]. It demonstrates that our method is highly reliable for EST-SSR identification. 2317 EST-SSRs (22.2%) with polymorphism in different repeat numbers could further be used for identification of Chiifu-401 and FuShanBaoTou and for genetic linkage map constructions using these two cultivars as parents. A total of 2744 new EST-SSRs (26.3%) were identified in the study, which, in combination with previously discovered EST-SSRs, could be used for high-density genetic linkage map construction, gene/QTL mapping, cultivar identification, and so forth.

4.4. High Polymorphism of Chinese Cabbage EST-SSRs. In the present study, 79.2% of the EST-SSRs primer pairs selected for primer evaluation successfully generated high quality amplicons, indicating that the ESTs from the high-throughput RNA sequencing of Chinese cabbage transcriptome are suitable for specific primer design. The unsuccessfully designed primer pairs may be due to splice sites, large introns, chimeric primer(s), or poor quality sequences [27]. We sequenced all PCR amplicons in Chinese cabbage FuShanBaoTou yielding 19 successful primer pairs. We found that all amplicons contained the expected SSRs and the SSRs in 13 amplicons were exactly the same as predicted (Table s6). The deviation of EST-SSR PCR amplicons from the expected size is likely due to the presence of introns, large insertions or repeat number variations, a lack of specificity, or assembly errors [43]. In the present study, we found five of six amplicons with unexpected sizes had different iterate number of SSR repeat units, while the other one had a 86 bp insertion near the expected SSR repeat motif region (Table s6). These results suggested that the unigenes assembled from the high-throughput RNA sequencing of Chinese cabbage transcriptome are reliable, and the EST-SSRs identified in our dataset could be used for further studies, such as genetic mapping and cultivar identification.

Most of the EST-SSR loci (accounting for 89.5% of the tested loci) were found to be polymorphic among the 24 tested cabbage cultivars. The mean number of alleles per SSR locus was 3.29 with a range between 2 and 6 (Table 3), indicating that polymorphism of EST-SSRs in Chinese cabbage is relatively high. Most of the polymorphisms of the tested EST-SSR loci are due to the variations of SSR repeat motifs in this study. There were only two loci where the polymorphisms were not related to the SSR repeat motif variations (Table s6). The results indicate that the EST-SSRs identified and the PCR primers designed in this study could further be used for constructing high-density genetic linkage maps, mapping quantitative trait loci, assessing germplasm polymorphism and evolution, marker-assisted selection, and cloning functional gene in Chinese cabbage.

In summary, we assembled a large set of clean reads with high quality derived from the Chinese cabbage transcriptome using high-throughput RNA sequencing technology with a Solexa/Illumina platform. A total of 51,694 nonredundant unigenes were obtained from 40.7 Mb sequence data, providing substantial knowledge for EST-SSR identification and characterization. 10,420 EST-SSRs were identified and characterized, and PCR primer pairs for 1561 EST-SSRs were designed. By comparing with previously reported SSRs in the SSR marker database for *Brassica* (http://oilcrops.info/SSRdb), we identified a total of 2744 new EST-SSRs. Primer pairs for 24 EST-SSRs were selected for primer evaluation, and 79.2% of the 24 EST-SSR loci successfully generated high quality amplicons. Among the effective primers, 89.5% of them showed polymorphism in 24 cultivars of Chinese cabbage. The EST-SSRs developed in this study, in combination with previously reported EST-SSRs, will provide valuable resources for constructing high-density genetic linkage maps, mapping quantitative trait loci, assessing germplasm polymorphism and evolution, marker-assisted selection, and cloning functional gene in Chinese cabbage. To our knowledge, this is the first successful attempt to develop large quantity of EST-SSRs with high quality based on the transcriptome of Chinese cabbage using high-throughput RNA sequencing technology.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Qian Ding and Jingjuan Li contributed equally to this work.

Acknowledgments

This research was supported by the China Postdoctoral Science Foundation Funded Project (2013M541948), the Shandong Postdoctoral Science Foundation Funded Project, China (201303030), the National Natural Science Foundation of China (31401869), the National High-tech R&D Program of China (863 Program) (Grant 2012AA100103), the Modern Agricultural Industrial Technology System Funding of Shandong Province, China (SDAIT-02-022-04), and the Project for Cultivation of Major Achievements in Science and Technology in SAAS (2015CGPY09).

References

[1] X. Wang, H. Wang, J. Wang et al., "The genome of the mesopolyploid crop species *Brassicarapa*," *Nature Genetics*, vol. 43, pp. 1035–1039, 2011.

[2] X. Yu, J. Peng, X. Feng et al., "Cloning and structural and expressional characterization of BcpLH gene preferentially expressed in folding leaf of Chinese cabbage," *Science in China, Series C: Life Sciences*, vol. 43, no. 3, pp. 328–329, 2000.

[3] H. Wu, L. Yu, X.-R. Tang, R.-J. Shen, and Y.-K. He, "Leaf downward curvature and delayed flowering caused by AtLH overexpression in *Arabidopsis thaliana*," *Acta Botanica Sinica*, vol. 46, no. 9, pp. 1106–1113, 2004.

[4] J. Lee, C. T. Han, and Y. Hur, "Overexpression of *BrMORN*, a novel 'membrane occupation and recognition nexus' motif protein gene from Chinese cabbage, promotes vegetative growth and seed production in *Arabidopsis*," *Molecules and Cells*, vol. 29, no. 2, pp. 113–122, 2010.

[5] B. Wang, X. Zhou, F. Xu, and J. Gao, "Ectopic expression of a Chinese cabbage BrARGOS gene in Arabidopsis increases organ size," *Transgenic Research*, vol. 19, no. 3, pp. 461–472, 2010.

[6] K. M. Song, J. Y. Suzuki, M. K. Slocum, P. M. Williams, and T. C. Osborn, "A linkage map of *Brassica rapa* (syn. *campestris*) based on restriction fragment length polymorphism loci," *Theoretical and Applied Genetics*, vol. 82, no. 3, pp. 296–304, 1991.

[7] Y.-S. Chyi, M. E. Hoenecke, and J. L. Sernyk, "A genetic linkage map of restriction fragment length polymorphism loci for *Brassica rapa* (syn. campestris)," *Genome*, vol. 35, no. 5, pp. 746–757, 1992.

[8] K. Suwabe, H. Iketani, T. Nunome, T. Kage, and M. Hirai, "Isolation and characterization of microsatellites in *Brassica rapa* L," *Theoretical and Applied Genetics*, vol. 104, no. 6-7, pp. 1092–1098, 2002.

[9] A. J. Lowe, C. Moule, M. Trick, and K. J. Edwards, "Efficient large-scale development of microsatellites for marker and mapping applications in *Brassica* crop species," *Theoretical and Applied Genetics*, vol. 108, no. 6, pp. 1103–1112, 2004.

[10] K. Suwabe, H. Iketani, T. Nunome, A. Ohyama, M. Hirai, and H. Fukuoka, "Characteristics of microsatellites in *Brassica rapa* genome and their potential utilization for comparative genomics in cruciferae," *Breeding Science*, vol. 54, no. 2, pp. 85–90, 2004.

[11] S. R. Choi, G. R. Teakle, P. Plaha et al., "The reference genetic linkage map for the multinational *Brassica rapa* genome sequencing project," *Theoretical and Applied Genetics*, vol. 115, no. 6, pp. 777–792, 2007.

[12] S. K. Jung, Y. C. Tae, G. J. King et al., "A sequence-tagged linkage map of *Brassica rapa*," *Genetics*, vol. 174, no. 1, pp. 29–39, 2006.

[13] K. Suwabe, H. Tsukazaki, H. Iketani et al., "Simple sequence repeat-based comparative genomics between *Brassica rapa* and *Arabidopsis thaliana*: the genetic origin of clubroot resistance," *Genetics*, vol. 173, no. 1, pp. 309–319, 2006.

[14] J. Wu, Y.-X. Yuan, X.-W. Zhang et al., "Mapping QTLs for mineral accumulation and shoot dry biomass under different Zn nutritional conditions in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*)," *Plant and Soil*, vol. 310, no. 1-2, pp. 25–40, 2008.

[15] F. Li, H. Kitashiba, K. Inaba, and T. Nishio, "A brassica rapa linkage map of EST-based SNP markers for identification of candidate genes controlling flowering time and leaf morphological traits," *DNA Research*, vol. 16, no. 6, pp. 311–323, 2009.

[16] L. Li and X. Y. Zheng, "The development of multiplex EST-SSR markers to identification Chinese cabbage [*Brassica campestris* L. chinensis (L.) Makino and *Brassica campestris* L. pekinensis (Lour.) Olsson] cultivars," *Acta Horticulturae Sinica*, vol. 37, no. 11, pp. 1627–1634, 2009.

[17] F. L. Iniguez-Luy, L. Lukens, M. W. Farnham, R. M. Amasino, and T. C. Osborn, "Development of public immortal mapping populations, molecular markers and linkage maps for rapid cycling *Brassica rapa* and *B. oleracea*," *Theoretical and Applied Genetics*, vol. 120, no. 1, pp. 31–43, 2009.

[18] H. Feng, P. Wei, Z.-Y. Piao et al., "SSR and SCAR mapping of a multiple-allele male-sterile gene in Chinese cabbage (*Brassica rapa* L.)," *Theoretical and Applied Genetics*, vol. 119, no. 2, pp. 333–339, 2009.

[19] X. Li, N. Ramchiary, S. R. Choi et al., "Development of a high density integrated reference genetic linkage map for the multinational *Brassica rapa* Genome Sequencing Project," *Genome*, vol. 53, no. 11, pp. 939–947, 2010.

[20] S. K. Parida, D. K. Yadava, and T. Mohapatra, "Microsatellites in *Brassica* unigenes: relative abundance, marker design, and use in comparative physical mapping and genome analysis," *Genome*, vol. 53, no. 1, pp. 55–67, 2010.

[21] Z. Wang, J. Li, Z. Luo et al., "Characterization and development of EST-derived SSR markers in cultivated sweetpotato (*Ipomoea batatas*)," *BMC Plant Biology*, vol. 11, article 139, 2011.

[22] W. Li, J. Zhang, Y. Mou et al., "Integration of Solexa sequences on an ultradense genetic map in *Brassica rapa* L.," *BMC Genomics*, vol. 12, article 249, 2011.

[23] J. Zou, D. Fu, H. Gong et al., "De novo genetic variation associated with retrotransposon activation, genomic rearrangements and trait variation in a recombinant inbred line population of *Brassica napus* derived from interspecific hybridization with *Brassica rapa*," *Plant Journal*, vol. 68, no. 2, pp. 212–214, 2011.

[24] H. Bagheri, M. El-Soda, I. van Oorschot et al., "Genetic analysis of morphological traits in a new, versatile, rapid-cycling *Brassica rapa* recombinant inbred line population," *Frontiers in Plant Science*, vol. 3, article 183, 2012.

[25] J. Shi, S. Huang, J. Zhan et al., "Genome-wide microsatellite characterization and marker development in the sequenced *Brassica* crop species," *DNA Research*, vol. 21, no. 1, pp. 53–68, 2014.

[26] W. Powell, M. Morgante, C. Andre et al., "The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis," *Molecular Breeding*, vol. 2, no. 3, pp. 225–238, 1996.

[27] R. K. Varshney, A. Graner, and M. E. Sorrells, "Genic microsatellite markers in plants: features and applications," *Trends in Biotechnology*, vol. 23, no. 1, pp. 48–55, 2005.

[28] I. Eujayl, M. K. Sledge, L. Wang et al., "*Medicago truncatula* EST-SSRs reveal cross-species genetic markers for *Medicago* spp," *Theoretical and Applied Genetics*, vol. 108, no. 3, pp. 414–422, 2004.

[29] L. Y. Zhang, M. Bernard, P. Leroy, C. Feuillet, and P. Sourdille, "High transferability of bread wheat EST-derived SSRs to other cereals," *Theoretical and Applied Genetics*, vol. 111, no. 4, pp. 677–687, 2005.

[30] M. C. Saha, J. D. Cooper, M. A. R. Mian, K. Chekhovskiy, and G. D. May, "Tall fescue genomic SSR markers: development and transferability across multiple grass species," *Theoretical and Applied Genetics*, vol. 113, no. 8, pp. 1449–1458, 2006.

[31] R. K. Varshney, R. Sigmund, A. Börner et al., "Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice," *Plant Science*, vol. 168, no. 1, pp. 195–202, 2005.

[32] J. E. Zalapa, H. Cuevas, H. Zhu et al., "Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences," *The American Journal of Botany*, vol. 99, no. 2, pp. 193–208, 2012.

[33] L. Li, W. M. He, L. P. Ma et al., "Construction Chinese cabbage (*Brassica rapa* L.) core collection and its EST-SSR fingerprint database by EST-SSR molecular markers," *Genomics and Applied Biology*, vol. 28, pp. 76–88, 2009.

[34] Y. Xin, H. Cui, M. Lu et al., "Data mining for SSRs in ESTs and EST-SSR marker development in Chinese cabbage," *Acta Horticulturae Sinica*, vol. 33, no. 3, pp. 549–554, 2006.

[35] Y. Ge, N. Ramchiary, T. Wang et al., "Development and linkage mapping of unigene-derived microsatellite markers in *Brassica rapa* L," *Breeding Science*, vol. 61, no. 2, pp. 160–167, 2011.

[36] N. Ramchiary, V. D. Nguyen, X. Li et al., "Genic microsatellite markers in brassica rapa: development, characterization, mapping, and their utility in other cultivated and wild brassica relatives," *DNA Research*, vol. 18, no. 5, pp. 305–320, 2011.

[37] F. Wang, L. Li, H. Li et al., "Transcriptome analysis of rosette and folding leaves in Chinese cabbage using high-throughput RNA sequencing," *Genomics*, vol. 99, no. 5, pp. 299–307, 2012.

[38] S. Wang, X. Wang, Q. He et al., "Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish," *Plant Cell Reports*, vol. 31, no. 8, pp. 1437–1447, 2012.

[39] G. Pertea, X. Huang, F. Liang et al., "TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets," *Bioinformatics*, vol. 19, no. 5, pp. 651–652, 2003.

[40] B. Winnepenninckx, T. Backeljau, and R. de Wachter, "Extraction of high molecular weight DNA from molluscs," *Trends in Genetics*, vol. 9, no. 12, p. 407, 1993.

[41] M. G. Grabherr, B. J. Haas, M. Yassour et al., "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, 2011.

[42] Z. Wang, B. P. Fang, J. Chen et al., "De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*)," *BMC Genomics*, vol. 11, article 726, 2010.

[43] W. Wei, X. Qi, L. Wang et al., "Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers," *BMC Genomics*, vol. 12, article 451, 2011.

[44] X. Liang, X. Chen, Y. Hong et al., "Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and *Arachis* wild species," *BMC Plant Biology*, vol. 9, article 35, 2009.

[45] S. Dutta, G. Kumawat, B. P. Singh et al., "Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]," *BMC Plant Biology*, vol. 11, article 17, 2011.

[46] K. D. Scott, P. Eggler, G. Seaton et al., "Analysis of SSRs derived from grape ESTs," *Theoretical and Applied Genetics*, vol. 100, no. 5, pp. 723–726, 2000.

[47] K. Shirasawa, K. Ishii, C. Kim et al., "Development of *Capsicum* EST-SSR markers for species identification and *in silico* mapping onto the tomato genome sequence," *Molecular Breeding*, vol. 31, no. 1, pp. 101–110, 2013.

[48] S. Cloutier, Z. Niu, R. Datla, and S. Duguid, "Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.)," *Theoretical and Applied Genetics*, vol. 119, no. 1, pp. 53–63, 2009.

[49] R. K. Aggarwal, P. S. Hendre, R. K. Varshney, P. R. Bhat, V. Krishnakumar, and L. Singh, "Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species," *Theoretical and Applied Genetics*, vol. 114, no. 2, pp. 359–372, 2007.

[50] L. Qiu, C. Yang, B. Tian, J.-B. Yang, and A. Liu, "Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.)," *BMC Plant Biology*, vol. 10, article 278, 2010.

[51] S. P. Kumpatla and S. Mukhopadhyay, "Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species," *Genome*, vol. 48, no. 6, pp. 985–998, 2005.

[52] K. N. Yamini, K. Ramesh, V. Naresh, P. Rajendrakumar, K. Anjani, and V. Dinesh Kumar, "Development of EST-SSR markers and their utility in revealing cryptic diversity in safflower (*Carthamus tinctorius* L.)," *Journal of Plant Biochemistry and Biotechnology*, vol. 22, no. 1, pp. 90–102, 2013.

[53] T. Thiel, W. Michalek, R. K. Varshney, and A. Graner, "Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.)," *Theoretical and Applied Genetics*, vol. 106, no. 3, pp. 411–422, 2003.

[54] M. Wen, H. Wang, Z. Xia, M. Zou, C. Lu, and W. Wang, "Developmenrt of EST-SSR and genomic-SSR markers to assess genetic diversity in *Jatropha Curcas* L.," *BMC Research Notes*, vol. 3, article 42, 2010.

[55] S. Tang, R. A. Okashah, M.-M. Cordonnier-Pratt et al., "EST and EST-SSR marker resources for Iris," *BMC Plant Biology*, vol. 9, article 72, 2009.

[56] J. Blanca, J. Cañizares, C. Roig, P. Ziarsolo, F. Nuez, and B. Picó, "Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae)," *BMC Genomics*, vol. 12, article 104, 2011.

[57] T. Liu, S. Zhu, L. Fu et al., "Development and characterization of 1,827 expressed sequence tag-derived simple sequence repeat markers for ramie (*Boehmeria nivea* L. Gaud)," *PLoS ONE*, vol. 8, no. 4, Article ID e60346, pp. 1091–1104, 2013.

[58] H. Pan, J. Wang, Y. Wang, Z. Qi, and S. Li, "Development and mapping of EST-SSR markers in wheat," *Scientia Agricultura Sinica*, vol. 43, pp. 452–461, 2010.

[59] H. Sonah, R. K. Deshmukh, A. Sharma et al., "Genome-wide distribution and organization of Microsatellites in plants: an insight into marker development in *Brachypodium*," *PLoS ONE*, vol. 6, no. 6, Article ID e21298, 2011.

[60] J. H. Peng and N. L. V. Lapitan, "Characterization of EST-derived microsatellites in the wheat genome and development of eSSR markers," *Functional and Integrative Genomics*, vol. 5, no. 2, pp. 80–96, 2005.

[61] C. Wang, W. Guo, C. Cai, and T. Zhang, "Characterization, development and exploitation of EST-derived microsatellites in *Gossypium raimondii* Ulbrich," *Chinese Science Bulletin*, vol. 51, no. 5, pp. 557–561, 2006.

[62] R. Singh, N. M. Zaki, N.-C. Ting et al., "Exploiting an oil palm EST database for the development of gene-derived SSR markers and their exploitation for assessment of genetic diversity," *Biologia*, vol. 63, no. 2, pp. 227–235, 2008.

[63] A. Stàgel, E. Portis, L. Toppino, G. L. Rotino, and S. Lanteri, "Gene-based microsatellite development for mapping and phylogeny studies in eggplant," *BMC Genomics*, vol. 9, article 357, 2008.

[64] L. Cardle, L. Ramsay, D. Milbourne, M. Macaulay, D. Marshall, and R. Waugh, "Computational and experimental characterization of physically clustered simple sequence repeats in plants," *Genetics*, vol. 156, no. 2, pp. 847–854, 2000.

[65] L. Gao, J. Tang, H. Li, and J. Jia, "Analysis of microsatellites in major crops assessed by computational and experimental approaches," *Molecular Breeding*, vol. 12, no. 3, pp. 245–261, 2003.

[66] R. K. Varshney, T. Thiel, N. S. P. Langridge, and A. Graner, "In silico analysis on frequency and distribution of microsatellites

in ESTs of some cereal species," *Cellular and Molecular Biology Letters*, vol. 7, no. 2, pp. 537–546, 2002.

[67] M. La Rota, R. V. Kantety, J.-K. Yu, and M. E. Sorrells, "Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley," *BMC Genomics*, vol. 6, article 23, 2005.

[68] M. Morgante, M. Hanafey, and W. Powell, "Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes," *Nature Genetics*, vol. 30, no. 2, pp. 194–200, 2002.