

1  
2 **Genome-wide sequencing of longan (*Dimocarpus longan* Lour.)**  
3  
4 **provides insights into molecular basis of its polyphenol-rich**  
5  
6 **characteristics**  
7  
8  
9

10  
11 YuLing Lin<sup>1†</sup>, JiuMeng Min<sup>2†</sup>, RuiLian Lai<sup>1</sup>, ZhangYan Wu<sup>2</sup>, YuKun Chen<sup>1</sup>, LiLi Yu<sup>2</sup>,  
12  
13 ChunZhen Cheng<sup>1</sup>, YuanChun Jin<sup>2</sup>, QiLin Tian<sup>1</sup>, QingFeng Liu<sup>2</sup>, WeiHua Liu<sup>1</sup>,  
14  
15 ChengGuang Zhang<sup>2</sup>, LiXia Lin<sup>1</sup>, YanHu<sup>2</sup>, DongMin Zhang<sup>1</sup>, MinKyaw Thu<sup>1</sup>, ZiHao  
16  
17 Zhang<sup>1</sup>, ShengCai Liu<sup>1</sup>, ChunShui Zhong<sup>1</sup>, XiaoDong Fang<sup>2</sup>, Jian Wang<sup>2, 3</sup>,  
18  
19 Huanming Yang<sup>2, 3</sup>, Rajeev K Varshney<sup>4,5\*</sup>, YeYin<sup>2\*</sup>, ZhongXiong Lai<sup>1\*</sup>  
20  
21  
22  
23  
24

25  
26 <sup>1</sup>Institute of Horticultural Biotechnology, Fujian Agriculture and Forestry University,  
27 Fuzhou, Fujian 350002, China.  
28

29  
30 <sup>2</sup>BGI-Shenzhen, Shenzhen 518083, China.  
31

32  
33 <sup>3</sup>James D. Watson Institute of Genome Sciences, Hangzhou 310058, China  
34

35  
36 <sup>4</sup>International Crops Research Institute for the Semi-Arid Tropics (ICRISAT),  
37 Hyderabad, India.  
38

39  
40 <sup>5</sup>School of Plant Biology, The University of Western Australia, Crawley, Perth,  
41 Australia.  
42

43 <sup>†</sup>These authors contributed equally to this work.  
44

45  
46 \*Corresponding authors  
47

48 Email: R.K.Varshney@CGIAR.ORG, yinye@genomics.cn, laizx01@163.com.  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## Abstract

### Background

Longan (*Dimocarpus longan* Lour.), an important subtropical fruit, is grown in more than 10 countries of the world. China's longan acreage and production rank first in the world. Longan is an edible drupe fruit and source of traditional medicine with polyphenol-rich traits, while tree size, alternate bearing, and witches' broom disease still pose serious problems. To gain insights into the genomic basis of longan traits, a draft genome sequence was assembled and genetic diversity and polyphenol-rich traits were identified and are presented here for the first time.

### Results

The draft genome (about 471.88 Mb) of a China longan cultivar 'Honghezi' was estimated to contain 31,007 high-quality genes and 261.88 Mb of repetitive sequences. No recent whole-genome wide duplication event was detected in the genome. Whole-genome resequencing and analysis of 13 cultivated *D. longan* accessions revealed the extent of genetic diversity and breeding-associated balancing selection. RNA sequencing revealed single nucleotide polymorphisms, insertions/deletions, differentially expressed genes, and alternative splicing events in different tissues of 'Sijimi' longan. Comparative transcriptome studies combined with genome-wide analysis revealed polyphenol-rich and pathogen-resistance characteristics. Genes involved in secondary metabolism, especially those from significantly expanded (*DHS*, *SDH*, *F3'H*, *ANR*, and *UFGT*) and contracted (*PAL*,

1 *CHS*, and *F3'5'H*) gene families with tissue-specific expression, may be important  
2  
3 contributors to the high accumulation levels of polyphenolic compounds observed in  
4  
5 longan fruit. The high number of genes encoding nucleotide-binding site leucine- rich  
6  
7 repeat (NBS-LRR) and leucine-rich repeat receptor-like kinase proteins, and the  
8  
9 recent expansion and contraction of the NBS-LRR family in the longan genome  
10  
11 suggested a genomic basis for resistance to insects, fungus, and bacteria in this fruit  
12  
13  
14  
15  
16  
17 tree.

## 22 **Conclusions**

23 These data provide insights into the evolution and diversity of the longan genome.  
24  
25 The comparative genomic and transcriptome analyses provided information about  
26  
27 longan-specific traits, particularly genes involved in its polyphenol-rich and pathogen-  
28  
29 resistance characteristics.  
30  
31  
32  
33  
34  
35  
36

## 37 **Keywords**

38 longan genome; genetic diversity; polyphenols biosynthesis; pathogen resistance  
39  
40  
41  
42  
43

## 44 **Background**

45 *Dimocarpus longan* Lour. (*D. longan*) originated from South China or Southeast Asia  
46  
47 and is commonly called longan or ‘dragon eye’ in Asia. It is an important  
48  
49 tropical/subtropical evergreen fruit tree that has a diploid genome ( $2n=2x=30$ ) and  
50  
51 belongs to the family Sapindaceae. Longan is widely cultivated in Southeast Asia,  
52  
53 South Asia, Australia, and Hawaii [1]. China's longan acreage and production rank  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 first, accounting for 70% and more than 50% of the world's acreage and production,  
2  
3 respectively [2]. As an edible drupe fruit and source of traditional medicine, longan is  
4  
5 grown in most areas of Southern China, including Guangdong, Guangxi, Fujian,  
6  
7 Sichuan, Yunnan, and Hainan [3]. Traditionally, longan leaves, flowers, fruit, and  
8  
9 seeds all have been widely used as traditional Chinese medicines for several diseases,  
10  
11 including leucorrhea, kidney disorders, allergies, cancer, diabetes, and cardiovascular  
12  
13 disease, because they contain bioactive compounds such as phenolic acids, flavonoids,  
14  
15 and polysaccharides [4], which exhibit antimicrobial, antioxidant, anticancer,  
16  
17 antityrosinase, and inflammatory properties [5, 6]. However, tree size, alternate  
18  
19 bearing, and witches' broom disease still pose serious problems in longan production  
20  
21 [1]. Cultivar identification and characterization are the first steps for fruit introduction  
22  
23 and breeding improvement [7]. In China, there are more than 300 longan varieties,  
24  
25 most are landraces and farm varieties, although a few wild populations exist in Hainan,  
26  
27 Guangdong, Guangxi, and Yunnan provinces [7, 8]. However, only 30–40 varieties  
28  
29 are grown commercially worldwide. Longan breeding improvement via conventional  
30  
31 breeding strategies has been hindered by its long juvenility, genetic heterozygosity,  
32  
33 and plant size [1]. To identify cultivars and improve longan breeding, knowledge of  
34  
35 the longan genetic background is required.

36  
37 Recently, many draft genome sequences for fruit trees have become available,  
38  
39 including papaya (*Carica papaya*) [9], grape (*Vitis vinifera*) [10], apple (*Malus*  
40  
41 *domestica*) [11], plum (*Prunus mume*) [12], orange (*Citrus sinensis*) [13], peach  
42  
43 (*Prunus persica*) [14], pear (*Pyrus bretschneideri*) [15], kiwifruit (*Actinidia chinensis*)  
44  
45

1 [16], pineapple (*Ananas comosus*) [17], banana (*Musa acuminata*) [18], jujube  
2  
3 (*Ziziphus jujuba*) [19], and strawberry (*Fragaria vesca*) [20]. However, draft genome  
4  
5  
6 sequences are still lacking for the subtropical and tropical fruits of the Sapindaceae  
7  
8  
9 family, which include longan, lychee (*Litchi chinensis*), and rambutan (*Nephelium*  
10  
11  
12 *lappaceum*). To accelerate improved breeding and utilization of the secondary  
13  
14  
15 metabolic products of longan, a fundamental understanding of its complete genome  
16  
17  
18 sequence is crucial. Here, we report the draft genome sequence of the longan cultivar  
19  
20 ‘Honghezi’ (HHZ) ( $2n=2x=30$ ) and the extent of genetic diversity in this species  
21  
22  
23 based on whole genome re-sequencing of 13 cultivated *D. longan* accessions.  
24  
25  
26 Comparative transcriptome studies combined with genome-wide analysis provided  
27  
28  
29 insights into the structure and evolution of the longan genome, the molecular  
30  
31  
32 mechanisms of the biosynthesis of polyphenol, and the pathogen resistance of longan.  
33  
34  
35 Together, these results provided insights into the evolution and diversity of the longan  
36  
37  
38 genome, and will help to improve the efficiency of longan conventional breeding by  
39  
40  
41 integrating biotechnological tools.

## 42 **Results and Discussion**

### 43 **Genome sequencing and assembly**

44  
45  
46 We selected the *D. longan* ‘HHZ’ cultivar for genome sequencing. In brief, a total of  
47  
48  
49 316.84 Gb of raw data was generated by Illumina sequencing of 12 genome shotgun  
50  
51  
52 libraries with different fragment lengths ranging from 170 bp to 40 kb (Additional file  
53  
54  
55 1: Table S1). After stringent filtering and correction steps, a total of 121.68 Gb of  
56  
57  
58 high-quality sequence data, representing 273.44-fold coverage of the entire genome,  
59  
60  
61

1 were obtained (Additional file 1: Table S2). Based on K-mer frequency methods [21],  
2  
3 the *D. longan* genome was estimated to be 445 Mb with a 0.88% heterozygosity rate  
4  
5 (Additional file 2: Fig. S1, Additional file 1: Table S3). Compared with other  
6  
7 sequenced fruit trees genomes, the *D. longan* genome was bigger than papaya [9],  
8  
9 orange [13], peach [14], and plum [12], and smaller than grape [10], apple [11], pear  
10  
11 [15], pineapple [17], and kiwifruit [16]. Longan trees are generally thought to have  
12  
13 highly heterozygous traits. The estimated 0.88% heterozygosity rate in the whole  
14  
15 genome of the longan ‘HHZ’ cultivar is reported here for the first time. This  
16  
17 heterozygosity rate is higher than the rates reported for kiwifruit (0.536%) [16], plum  
18  
19 (0.03%) [12, 22], and poplar (about 0.5%) [23], and lower than the rates for pear  
20  
21 (1–2% sequence divergence) [15] and pineapple (1.89% in F153, 1.98% in MD2,  
22  
23 2.93% in CB5) [17]. These results imply that the idea that fruit trees always have high  
24  
25 heterozygosity may be due to artificial grafting and/or asexual reproduction.  
26  
27

28  
29 Using the SOAPdenovo program [24], all the high-quality reads were assembled into  
30  
31 51,392 contigs and 17,367 scaffolds ( $\geq 200$  bp) totaling 471.88 Mb without N  
32  
33 sequences (Table 1). These assembled sequences accounted for approximately  
34  
35 106.04% of the estimated longan genome, which conflicts with previously reported  
36  
37 genome assemblies where the sequences accounted for less than 100% of the  
38  
39 estimated genome [13-15]. The higher percentage might be due to the high  
40  
41 heterozygosity of the longan genome, suggesting that, in the future, a single- molecule  
42  
43 sequencing technology should be used to correct the longan genome assembly.  
44  
45

46  
47 Generally, the N50 number is used to measure the contig or scaffold lengths for  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57

1 estimating the quality of a genome assembly. Here, the N50s of contigs and scaffolds  
2  
3 were 26.04 kb (longest, 173.29 kb) and 566.63 kb (longest, 6942.32 kb), respectively  
4  
5  
6 (Table 1), suggesting the high quality of the assembly. The GC content of the *D.*  
7  
8  
9 *longan* genome was 33.7%, which is comparable with the GC content of the genomes  
10  
11 of pineapple (33%) [17], jujube (33.41%) [19], and orange (34.06 %) [13], but lower  
12  
13 than the GC content of the genomes of kiwifruit (35.2 %) [16], papaya (35.3%) [9],  
14  
15 and grape (36.2%) [10] (Table 2, Additional file 2: Fig. S2). Analysis of the percent  
16  
17 GC content among different fruit trees can provide important clues about gene density,  
18  
19  
20  
21  
22  
23 gene expression, replication timing, recombination, and evolutionary relationships  
24  
25 [25]. The GC-depth graph and distribution indicated no contamination of any bacterial  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

estimating the quality of a genome assembly. Here, the N50s of contigs and scaffolds were 26.04 kb (longest, 173.29 kb) and 566.63 kb (longest, 6942.32 kb), respectively (Table 1), suggesting the high quality of the assembly. The GC content of the *D. longan* genome was 33.7%, which is comparable with the GC content of the genomes of pineapple (33%) [17], jujube (33.41%) [19], and orange (34.06 %) [13], but lower than the GC content of the genomes of kiwifruit (35.2 %) [16], papaya (35.3%) [9], and grape (36.2%) [10] (Table 2, Additional file 2: Fig. S2). Analysis of the percent GC content among different fruit trees can provide important clues about gene density, gene expression, replication timing, recombination, and evolutionary relationships [25]. The GC-depth graph and distribution indicated no contamination of any bacterial sequence in the longan genome assembly, and 99.2% of the assembly was sequenced with more than 20× coverage (Additional file 2: Fig. S3). The statistics and comparison of the *D. longan* assembly with 12 other twelve fruit tree genomes are shown in detail in Table 2. The quality of the assembly was assessed by aligning the scaffolds to a longan transcriptome assembly from the NCBI Sequence Read Archive (SRA) [SRA050205]. Of the 96,251 longan transcriptome sequences ( $\geq 100$ ) reported previously [26], 97.55% were identified in the genome assembly (Additional file 1: Table S4), confirming the high quality of the assembly.

### **BUSCO analysis**

We further evaluated the quality and completeness of the draft longan genome assembly using the BUSCO (Benchmarking Universal Single-Copy Orthologs) datasets [27]. The BUSCO analysis indicated that 94% of the longan draft assembly

1 was complete. Of the total of 956 BUSCO ortholog groups searched in the longan  
2  
3 assembly, 900 (94%) BUSCO genes were “complete single-copy”, 288 (30%) were  
4  
5 “complete duplicated”, 16 (1.6%) were “fragmented”, and 40 (4.1%) were “missing”  
6  
7  
8 (Additional file 1: Tables S5). The percentage of missing BUSCO genes was  
9  
10 comparable to the percentages missing in the assemblies of banana (3%), *Brassica*  
11  
12 *napus* (3%), and *Arabidopsis* (2%), which have served as well-assembled standards at  
13  
14 the chromosomal level [28], further suggesting the high quality of our assembly.  
15  
16  
17  
18  
19

### 20 **Repetitive elements and gene annotation**

21  
22 Repetitive elements are major components of eukaryotic genomes, and they have been  
23  
24 used extensively to analyze genome structure, karyotype, ploidy, and evolution. In the  
25  
26 longan assembly, we found a total of 261.88 Mb (52.87%, 445 Mb) was repetitive  
27  
28 sequences (Additional file 1: Table S6), which is higher than the amount observed in  
29  
30 orange (20%, 367 Mb) [13], peach (29.6%, 265 Mb) [14], kiwifruit (36%, 758 Mb)  
31  
32 [16], pineapple (38.3%, 526 Mb) [17], grape (41.4%, 475 Mb) [10], jujuba (49.49%,  
33  
34 444 Mb) [10], and papaya (51.9%, 372 Mb) [9], and lower than the amount reported  
35  
36 in pear (53.1%, 527 Mb) [15] and apple (67.4%, 742.3 Mb) [11] (Table 2), indicating  
37  
38 that the size of fruit tree genomes differed as a result of the variable amounts of  
39  
40 repetitive elements that they contained. Accordingly, the bigger plant genomes often  
41  
42 possessed higher percentages of repetitive elements than the smaller plant genomes.  
43  
44 Most plant genomes appear to contain abundant long-terminal repeat (LTR)  
45  
46 retrotransposons and a small number of short interspersed elements (SINEs) and long  
47  
48 interspersed elements (LINEs) [29]. We found that the repetitive fraction of the  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



1 longan genome comprised LTR retrotransposons, which were the most abundant  
2  
3 (36.54%), and SINEs (2.43%) and LINEs (0.04%), which were the least abundant;  
4  
5  
6 other repeats, including tandem repeats and unknown repeats, made up 7.59% and  
7  
8  
9 7.71% of the repetitive fraction, respectively (Additional file 1: Table S7). A large  
10  
11 number of the unknown repetitive sequences may be longan-specific. The  
12  
13 characterization of repetitive sequences is of primary importance for understanding  
14  
15  
16 the structure and evolution of the longan genome.  
17

18  
19  
20 Using a combination of *de novo* prediction, homology-based searches, and a  
21  
22 transcriptome assembly, we predicted a total of 39,282 genes yielding a set of 31,007  
23  
24 high-quality proteins in the longan genome. The average gene size was 3,266.02 bp,  
25  
26 the average length of the coding sequence was 1,232.18 bp, and the average number  
27  
28 of exons per gene was 4.68 (Additional file 1: Table S8). The number of genes  
29  
30 predicted in the longan genome was close to the number of genes predicted in jujube  
31  
32 (32,808) [10], higher than in papaya (24,746) [9], pineapple (27,024) [17], peach  
33  
34 (27,852) [14], orange (29,445) [13], and grape (30,434) [10], and lower than in  
35  
36 kiwifruit (39,040) [16], pear (42,812) [15], and apple (57,386) [11]. This analysis  
37  
38 showed that the number of genes in the longan genome was similar to the numbers  
39  
40 found in other sequenced fruit tree genomes of equivalent size, and also indicated that  
41  
42 the bigger plant genomes usually contained higher numbers of genes. Of 31,007  
43  
44 protein-coding genes, 27,862 (89.86%) had TrEMBL homologs, 22,986 (74.13 %)  
45  
46 had SwissProt homologs, and 23,398 (75.46%) had InterPro homologs (Additional  
47  
48 file 1: Table S9). A total of 1,611 putative transcription factors (TFs) distributed in 64  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 families were identified, which represented 4.1% of the genes in the longan genome  
2  
3 (39,282). The percentage of TFs in longan genome was close to the percentages  
4  
5 reported in strawberry (4.6%) [20] and rice (4.8 %), but lower than the percentages in  
6  
7 Arabidopsis (6%), kiwifruit (6.2%) [16], grape (6.7%) [30], poplar (6.7%), and  
8  
9 banana (11.75%) [18]. In the longan genome, the largest numbers of genes encoded  
10  
11 TFs in the following TF families: MYB (186 genes), ERF (115), MADS (109), NAC  
12  
13 (107), bHLH (107), C2H2 (98), B3 superfamily (86), HB (71), WRKY(58), bZIP (55),  
14  
15 GRAS (52), and C3H (49) (Supplemental EXCEL File 1). The identification of these  
16  
17 TFs will help to lay a solid foundation for functional verification of longan traits in  
18  
19 the future. Among the non-coding genes detected in the longan genome assembly, we  
20  
21 identified 359 microRNAs, 212 rRNA, 506 tRNAs, and 399 small nuclear RNAs  
22  
23 (Additional file 1: Table S10).  
24  
25  
26  
27  
28  
29  
30  
31

### 32 **Gene family evolution and comparison**

33  
34 Orthologous clustering analysis was conducted with the longan genome and eight  
35  
36 other selected plant genomes, Arabidopsis, orange, papaya, grapevine, banana, peach,  
37  
38 kiwifruit, and apple. Of the 31,007 protein-coding genes in the genome, 26,261 were  
39  
40 grouped into 14,961 gene families (763 of which were longan-unique families) giving  
41  
42 an average of 1.76 genes per family (Additional file 1: Table S11). The remaining  
43  
44 5,834 genes were classed as un-clustered genes. Among the 31,007 genes, 4,653 were  
45  
46 longan-unique paralogs, 5,184 were multiple-copy orthologs, 3,606 were single-copy  
47  
48 orthologs, and 12,818 were other orthologs (Fig. 1b). Comparative analysis of the  
49  
50 longan genome with eight other selected plant genomes indicated that the number of  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 gene families in the longan genome was similar to the numbers in the genomes of  
2  
3 orange (15,000) [13] and peach (15,326) [14], higher than in banana (12,519) [18],  
4  
5 Arabidopsis (13,406), grape (13,570) [10], kiwifruit (13,702) [16], and papaya  
6  
7 (13,763) [9], and lower than in apple (17,740) [11] (Fig. 1b, Additional file 1: Table  
8  
9 S11). These comparisons indicated that differences in gene families in plant genomes  
10  
11 may be important sources of genetic traits and adaptation in different species.  
12  
13 Comparative analysis of the longan genome with the genomes of citrus, banana, peach,  
14  
15 and Arabidopsis showed that these five species contained a core set of 9,215 genes in  
16  
17 common, whereas 1,207 genes were specific to longan, which is more than the  
18  
19 numbers of genes specific to citrus and Arabidopsis, and lower than the numbers  
20  
21 specific to *M. acuminata* and peach (Fig. 1d).  
22  
23  
24  
25  
26  
27  
28  
29  
30

31 Expansion or contraction of gene families may provide clues to the evolutionary  
32  
33 forces that have shaped plant genomes and have an important role in the  
34  
35 diversification of plants. In this study, we used CAFÉ [31] to identify gene families  
36  
37 that had potentially undergone expansion or contraction in the longan genome. We  
38  
39 found a total of 2,849 expanded gene families and 2,842 contracted families; however,  
40  
41 only 386 expanded families (7,839 genes) and 12 contracted families (53 genes),  
42  
43 accounting for 19.96% and 0.13% of the total coding-genes (39,282), respectively,  
44  
45 were found to be statistical significant at  $P < 0.05$  (Supplemental EXCEL Files 2 and  
46  
47 3). The genes in the significantly expanded and contracted families ( $P < 0.05$ ) were  
48  
49 annotated with gene ontology (GO) terms. Genes in a total of 32 (expanded) and 11  
50  
51 (contracted) families were assigned GO terms under the three GO categories,  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 biological process, cellular component, and molecular function. Almost all the  
2  
3 expanded or contracted families contained genes that were assigned terms under  
4  
5 biological process, and a few genes in the contracted families were assigned terms  
6  
7 under the cellular component and molecular function categories (Additional file 2: Fig.  
8  
9 S4a, b). The dominant terms in the expanded or contracted gene families were  
10  
11 ‘cellular component organization’, ‘locomotion’, ‘auxiliary transport protein’, and  
12  
13 ‘binding’, revealing important clues to the evolutionary forces that may have shaped  
14  
15 the longan genomes.  
16  
17  
18  
19  
20  
21

## 22 **Genome evolution**

23  
24  
25 Whole-genome duplication is common in most plant species and it represents an  
26  
27 important molecular mechanism that has shaped modern plant karyotypes [32].  
28  
29 Characterization and annotation of the longan genome provided comprehensive  
30  
31 information for us to further investigate the evolutionary history of longan.  
32  
33  
34 Single-copy nuclear genes from orange, Arabidopsis, cacao (*Theobroma cacao*),  
35  
36 poplar (*Populus trichocarpa*), grape, apple, papaya, soybean, peach, kiwifruit, and  
37  
38 banana [18] were used in a genome-scale phylogenetic analysis using the maximum  
39  
40 likelihood method. The phylogenetic analysis showed that longan was  
41  
42 phylogenetically closest to orange, close to papaya, Arabidopsis, and cacao, and most  
43  
44 distant from monocotyledon fruits (banana). From the phylogenetic tree, we estimated  
45  
46 that longan diverged about 69.3 million years ago (Fig. 1a). To determine the nature  
47  
48 of the evolutionary events that led to the modern longan genome structure, we  
49  
50 analyzed the syntenic relationships between longan and poplar. We detected a total of  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 2,106 and 883 syntenic blocks containing 17,901 and 17,447 colinear genes for  
2  
3 longan and poplar, respectively (Additional file 1: Table S12), which supported the  
4  
5 reported conserved colinearity and close evolutionary relationship in these two plant  
6  
7 species. To further analyze the evolutionary divergence and the relative age of  
8  
9 duplication events in longan and other related species, we calculated the  
10  
11 distance–transversion rates at fourfold degenerate sites (4DTv) (Fig. 1c). The 4DTv  
12  
13 value peaked at 0.5 for paralog pairs in grape, highlighting the recent whole-genome  
14  
15 duplication in this species. Two 4DTv values that peaked at 0.72 and 0.6 for orthologs  
16  
17 between longan and banana, and between longan and Arabidopsis, respectively,  
18  
19 supported species divergence. These results are consistent with the more ancient  
20  
21 divergence between monocotyledons and dicotyledons. The orthologs between longan  
22  
23 and grape, longan and peach, and longan and orange showed 4DTv distances peaks at  
24  
25 0.36, 0.36, and 0.26, respectively, which is consistent with the 4DTv peaks reported  
26  
27 previously for Vitaceae and Rosaceae species, and more ancient than the 4DTv values  
28  
29 for Rutaceae or Sapindaceae. In longan, the analysis showed ancient duplication  
30  
31 events (the 4DTv peak at about 0.55) but did not reveal a recent whole-genome  
32  
33 duplication. These results complement the results for the longan genome and will  
34  
35 contribute to studies into ancestral forms and arrangements of plant genes [33].  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49

### 50 **Assessment of genetic diversity in longan germplasm**

51  
52 A representative characteristic of longan cultivars is their high heterozygosity, which  
53  
54 has resulted in the low efficiency of longan germplasm management and utilization.  
55  
56

57 Traditionally, molecular markers (RAPD, AFLP, SCAR, SCTP, and SRAP) and  
58  
59  
60  
61  
62  
63  
64  
65

1 single nucleotide polymorphisms (SNPs) based on transcriptome data [34] have been  
2  
3 used for accurate identification of longan varieties. However, the extent of  
4  
5 heterozygosity in the whole genome is not well understood [7]. The availability of the  
6  
7 longan draft genome provided the foundation for a comprehensive assessment of  
8  
9 heterozygosity in the longan genome.  
10  
11  
12

13 We selected 13 representative commercially cultivated accessions with early-maturing,  
14  
15 middle-maturing, late-maturing, multiple-flowering, aborted-seeded, and disease-  
16  
17 resistant characteristics for whole-genome resequencing (Additional file 1: Table S13).  
18  
19  
20

21 A total of 45.77 Gb of raw data were generated by Illumina sequencing. After  
22  
23 alignment of the clean reads corresponding to 5.02- to 7.31-fold depths and >78%  
24  
25 coverage to the reference genome (Additional file 1: Table S14), we identified  
26  
27 357,737 SNPs (Additional file 1: Table S15), and 23,225 small insertions/deletions  
28  
29 (indels) (Additional file 1: Table S16). The overall polymorphism density was  
30  
31 0.05–0.12 SNPs and 0.004–0.007 indels per 10 kb of the genome sequence, which is  
32  
33 much lower than the diversity reported in orange [13]. Notably, the major variations  
34  
35 existed among the ‘FY’, ‘MQ’, and ‘SJM’ accessions, whereas variations within the  
36  
37 cultivated longan accessions, particularly the ‘LDB’ accessions, were relatively low  
38  
39 (Additional file 1: Tables S15 and S16).  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49

50 To further investigate the population structure and relationships among the longan  
51  
52 accessions, we constructed a neighbor-joining tree (Fig. 2a) and carried out a principal  
53  
54 component analysis (PCA) (Fig. 2b). The neighbor-joining tree, constructed based on  
55  
56 all the identified SNPs, indicated that the 13 longan accessions clustered into two  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 subfamilies. The first subfamily consisted only of ‘FY’, which showed the highest  
2  
3 variations and clear separation from other cultivars. This result is quite different from  
4  
5 results reported previously [35, 36]. In previous studies using molecular markers,  
6  
7 ‘FY’, which originated from Quanzhou, China, was found to cluster together with  
8  
9 other Chinese longan accessions. In our study, which was conducted at an overall  
10  
11 genomic level, ‘FY’ was found to possess more genetic differences compared with the  
12  
13 other longan accessions tested. This result might be due to the special traits of ‘FY’,  
14  
15 such as witches' broom disease-resistant, middle-maturity, and canned processing  
16  
17 products. This result also supports the observed diversity of ‘FY’ at the overall  
18  
19 genomic level. The second subfamily neighbor-joining tree consisted of three clades  
20  
21 (Fig. 2a). The first clade included ‘JHLY’, ‘WLL’, ‘JYW’, and ‘SN1H’; the second  
22  
23 contained ‘MQ’, ‘SX’, ‘SJM’, and ‘SEY’; and the third consisted of ‘DB’, ‘HHZ’,  
24  
25 ‘LDB’ and ‘YTB’. Moreover, the PCA showed that the samples that originated from  
26  
27 China tended to cluster together (‘HHZ’, ‘DB’, ‘JYW’, ‘LDB’, ‘WLL’, ‘SN1H’,  
28  
29 ‘YTB’, ‘SEY’, ‘JHLY’, and ‘SX’). The PCA also showed the clear separation of ‘FY’,  
30  
31 ‘SJM’, and ‘MQ’. The ‘SJM’ and ‘MQ’ accessions, which originated from Southeast  
32  
33 Asia and Thailand, respectively, possessed apparent differences compared with the  
34  
35 Chinese longan accessions tested in this study. Together these results indicated  
36  
37 geographic patterns of genetic differentiation, which agree with findings reported  
38  
39 previously [34]. The relatively low levels of genetic variation among the Chinese  
40  
41 cultivars also suggested that they might have suffered a bottleneck during  
42  
43 domestication [7, 34]. These results suggested the relationship among the 13 selected  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 longan accessions was, at least partly, determined by their geographical distributions.

2  
3 An additional analysis of the population structure was conducted using the FRAPPE  
4 program [37] with K (the number of populations) set from 2 to 7 (Fig. 2c). For K=7, a  
5  
6 new subgroup was detected among the 13 longan accessions. This subgroup had  
7  
8 characteristics, such as various maturity levels, high yielding, aborted-seeding,  
9  
10 disease-resistant, and multiple flowering. The cultivars ‘SX’ and ‘YTB’, which are  
11  
12 susceptible to disease, contained more variations in resistance genes, such as  
13  
14 NBS-LRR and LRR-RLK, than the disease resistant cultivars (‘FY’, ‘SN1H’, ‘MQ’,  
15  
16 ‘LDB’, and ‘JYW’) (Supplemental EXCEL Files 4 and 5). These results provided a  
17  
18 measure of the changes in genetic diversity and a theoretical estimate of the genetic  
19  
20 relationships among the selected longan cultivars.  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30

31 **RNA sequencing revealed SNPs, indels, differentially expressed genes, and**  
32  
33 **alternative-splicing events in different tissues of ‘SJM’ longan**  
34  
35

36 To improve the gene annotation of the longan genome sequence and get more  
37  
38 information about longan traits, we constructed nine cDNA libraries corresponding to  
39  
40 nine different organs (root, stem, mature leaf, flower bud, flower, young fruit,  
41  
42 pericarp, pulp, and seed) from a representative ‘SJM’ cultivar. ‘SJM’, which  
43  
44 originated in Southeast Asia, blossoms and bears fruit throughout the year, with no  
45  
46 requirement of environmental control [38]. Here, a total of 490,502,822 clean reads  
47  
48 from nine RNA sequencing (RNA-seq) data sets were obtained after removing  
49  
50 low-quality reads and adaptor sequences, and about 53.55–79.40% of the clean reads  
51  
52 mapped to the longan draft genome (Additional file 1: Table S17). This percentage of  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



1 mapped reads is lower than the 90% previously reported in peach [39], suggesting that  
2  
3 the ‘SJM’ cultivar contained high variations compared with the sequenced ‘HHZ’  
4  
5 genome, probably because of their different origins. Moreover, the BUSCO analysis  
6  
7 [27] showed that 483 (87%) of BUSCO genes were “complete single-copy”, 352  
8  
9 (36%) were “complete duplicated”, 53 (5.5%) were “fragmented”, and 68 (7.1%)  
10  
11 were “missing” (Additional file 1: Table S18), indicating the high quality of our  
12  
13 assembled transcriptome.  
14  
15  
16  
17  
18  
19

20 The transcribed regions/units were constructed independently for individual tissues.  
21  
22 We found that transcripts/genes ranged from 19,322 (pulp) to 23,118 (flower bud),  
23  
24 completely or partially (49.18–58.85%) overlapped with 39,282 annotated genes in  
25  
26 the longan genome. The numbers of expressed transcripts in each longan tissue were  
27  
28 much lower than the numbers previously reported in *Brassica rapa* (32,335 genes  
29  
30 expressed in at least one tissue, equivalent to 78.8% of the 41,020 annotated genes)  
31  
32 [40]. The lower numbers of transcripts detected in each tissue, may be due to the high  
33  
34 variations and genetic heterozygosity in the ‘SJM’ cultivar. The coverage of the  
35  
36 longan gene set by our transcripts indicated the broad representation of our unigenes,  
37  
38 and provided the opportunity to identify alternative splicing (AS) events. In addition  
39  
40 to the predicted genes, novel transcripts, ranged from 1,621 (stem) to 1,999 (young  
41  
42 fruit), were detected across all nine samples. Among the novel transcripts, 798 (flower)  
43  
44 – 988 (young fruit) contained open reading frames, while 820 (stem) – 1,011 (young  
45  
46 fruit) were identified as non-coding RNAs in the longan genome (Additional file 1:  
47  
48 Table S17). Most of these non-coding RNAs were longer than 200 nt and had no  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 ORFs encoding sequences longer than 300 amino acids, suggesting they may be long  
2  
3 intergenic non-coding RNAs [41] or *cis*-natural antisense transcripts [42], which will  
4  
5  
6 need further analysis. The numbers of novel encoding and non-coding transcripts in  
7  
8  
9 young fruit were the highest among the nine samples, suggesting the development of  
10  
11  
12 young fruit required more complicate gene regulatory networks than the other stages.  
13  
14 To further optimize of the structure of the transcripts, we compared the assembled  
15  
16 transcripts and annotated genes from the reference longan genome and extended the 5'  
17  
18 or 3' ends of the transcripts according to the annotated gene information. In total, the  
19  
20 extending 5' or 3' end of annotated genes ranged from 8,126 (pulp) to 9,995 (flower  
21  
22 bud) across nine tissues, and about almost half the number of total genes extended by  
23  
24  
25 5' end in each sample. We identified a total of 1,255,816 SNPs and 34,390 indels  
26  
27  
28 across the nine longan tissues, and found that the highest number of SNPs and indels  
29  
30  
31 were detected in young fruit (161,897) and leaf (4,673), respectively, suggesting the  
32  
33  
34 expressed transcripts may be more diverse in these two tissues. Notably, the lowest  
35  
36  
37 frequencies of SNPs and indels were detected in pulp (105,007 and 2,587  
38  
39  
40 respectively). The SNPs and indels detected in the transcript sequences will be a  
41  
42  
43 valuable resource from which to identify candidate genes, analyze population  
44  
45  
46 structures and evolution, and accelerate plant breeding [39]. The identification of  
47  
48  
49 novel genes, extended annotated genes, SNPs, and indels from different  
50  
51  
52 developmental stages, imply our gene set can serve as a valuable complementary  
53  
54  
55 resource for longan genomics.  
56

57  
58 To identify significantly differentially expressed genes (DEGs), we used 12 pair-wise  
59  
60  
61  
62  
63  
64  
65

1 comparisons among the nine samples as follows: root VS stem, root VS leaf, leaf VS  
2  
3 stem, flower bud VS flower, flower bud VS young fruit, flower VS young fruit, young  
4  
5 fruit VS pulp, young fruit VS seed, pericarp VS pulp, pericarp VS seed, and pulp VS  
6  
7 seed. Among the detected DEGs (Additional file 2: Fig. S5), an average of  
8  
9 3,922±2,391 were up-regulated and an average of 4,859±2,666 were down-regulated  
10  
11 in the 12 comparisons. The highest number of DEGs was detected in young fruit VS  
12  
13 seed (9,737), followed by root VS leaf (9,702) and flower VS young fruit (9,101), and  
14  
15 the lowest number of DEGs was detected in flower bud VS flower (3,722). The  
16  
17 numbers of organ-specific genes ranged from 87 in young fruit to 530 in root, and the  
18  
19 significantly differentially expressed transcription factors in each comparison ranged  
20  
21 from 272 (flower bud VS flower) to 732 (young fruit VS pulp). To evaluate the  
22  
23 potential functions of the DEGs, we annotated them by assigning GO terms under the  
24  
25 three main categories, biological process, cellular component, and molecular function.  
26  
27 DEGs in each pair were categorized into 43 (flower bud VS flower) - 47 (young fruit  
28  
29 VS pulp). Details of the GO annotations are provided in Additional file 2: Fig. S6. The  
30  
31 dominant terms in all 12 comparisons were ‘Metabolic process’, ‘Cellular process’,  
32  
33 ‘Cell’, ‘Cell part’, ‘Catalytic activity’, and ‘Binding’, which is similar to results  
34  
35 previously reported in the ‘SJM’ and ‘LDB’ cultivars [43]. To further understand the  
36  
37 biological functions of the DEGs, we carried out a KEGG (Kyoto Encyclopedia of  
38  
39 Genes and Genomes) pathway-based analysis. In nine of the 12 comparisons, the  
40  
41 highest numbers of DEGs were involved in ‘metabolic pathway’, followed by the  
42  
43 ‘biosynthesis of secondary metabolites’ and ‘plant–pathogen interaction’ pathways. In  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 pericarp VS seed, root VS leaf, and pericarp VS pulp, ‘biosynthesis of secondary  
2 metabolites’, ‘pyrimidine metabolism’, and ‘stilbenoid, diarylheptanoid and gingerol  
3 biosynthesis’ were the most represented pathways, respectively (Additional file 2: Fig.  
4 S7). These results are fully consistent with the view that *D. longan* contains high  
5 levels of polyphenolic compounds, and a large number of pathogen resistance genes  
6 [44, 45].  
7

8  
9 To determine the types of AS events represented in our assembled transcripts data set,  
10 we used the TopHat software [46]. First, the nine longan tissues were analyzed at the  
11 exon level, which can provide important information about the types of gene isoforms  
12 that are expressed and variable [47]. Expressed exons were detected in the range of  
13 96,105 (pulp) to 111,476 (flower bud) across the nine tissues (Additional file 1: Table  
14 S17). A total of 298,914 AS events were detected across all the tissues, representing  
15 the four known types of AS, namely intron retention, exon skipping, alternative 5’  
16 splice site donor, and alternative 3’ splice site acceptor. Alternative transcripts have  
17 been shown to be tissue- or condition-specific [47, 48]. We also found that the largest  
18 numbers of AS events were detected in leaf (37,216), followed by young fruit  
19 (35,998), and pericarp (35,384), and the smallest numbers were found in pulp  
20 (28,058), corresponding to the least expressed exons. The predominant and rare types  
21 of AS events in all nine tissues were intron retention and exon skipping, respectively.  
22 This result is consistent with prior findings in rice [49], Arabidopsis [50], grape [48,  
23 51], and *B. rapa* [40], but contradicts a previous finding that exon-skipping was  
24 predominant in peach [39] and metazoans [52], indicating the complexity of the AS  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 landscape in plants and the important consequences this may have on plant/crop  
2  
3 phenotypes.  
4

### 5 **Biosynthesis of polyphenols and MYB transcription factors in longan**

6  
7  
8  
9 Polyphenols, potential antioxidative compounds, are the major category of secondary  
10  
11 metabolites in longan leaf, flower, fruit, and seed [4]. Phenolic compounds are derived  
12  
13 primarily through the shikimic acid, phenylpropanoid, and flavonoid pathways. Our  
14  
15 transcriptome data showed that the significant DEGs in the nine longan tissues were  
16  
17 involved mainly in ‘biosynthesis of secondary metabolites’. To further assess changes  
18  
19 between the primary and secondary metabolism of polyphenols during the longan  
20  
21 vegetative and reproductive growth stages, the copy numbers of 26 selected structural  
22  
23 genes within the shikimate acid, phenylpropanoid, and flavonoid biosynthesis  
24  
25 pathways were compared with those in corresponding pathways of Arabidopsis,  
26  
27 orange, peach, grape, poplar, and eucalyptus (Fig. 3a, Supplemental EXCEL File 6).  
28  
29  
30  
31  
32  
33  
34  
35

36 Comparison analysis showed that the 26 structural genes showed up and down  
37  
38 variations in copy numbers among the seven plants tested (Supplemental EXCEL File  
39  
40 6). The significant expanded gene families in longan, orange, peach, poplar, and  
41  
42 eucalyptus were *DHS*, *SDH*, *F3'H*, *ANR*, and *UFGT*, when compared with the  
43  
44 corresponding families in grape, which is considered to be the oldest among the seven  
45  
46 selected plants in evolutionary history [53]. *SDH*, catalyzes the NADPH-dependent  
47  
48 reduction of 3-dehydroshikimate to shikimate in the fourth step of the shikimate  
49  
50 pathway, which is the metabolic route required for the biosynthesis of the aromatic  
51  
52 amino acids. *SDH* had six copy numbers in longan, which is the same as in Populus,  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 but much higher than in Arabidopsis (1 copy), peach and grape (2 copies each), and  
2  
3 orange and eucalyptus (3 copies each). *F3'H* is involved in flavonoid biosynthesis and  
4  
5 is important for flower color and fruit skin. We found 65 copies of *F3'H* in the  
6  
7 eucalyptus genome, 35 in longan, 28 in peach, 25 in orange, 26 in *Populus*, and only  
8  
9 12 in grape and 10 in Arabidopsis, suggesting that the *F3'H* family was significantly  
10  
11 expanded in woody plants and a little contracted in herbs. These findings may provide  
12  
13 important clues for the mechanism of flavonoid biosynthesis in plants. The gene  
14  
15 encoding ANR, which is involved in the biosynthesis of proanthocyanidins (also  
16  
17 called condensed tannins), had higher copy numbers (6) in longan than in Arabidopsis  
18  
19 (2), orange (1), peach (1), grape (4), and *Populus* (5), implying that the expanded *ANR*  
20  
21 numbers may play a role in proanthocyanidin biosynthesis. Significantly smaller  
22  
23 numbers of the structural genes *PAL*, *CHS*, and *F3'5'H* were detected in longan (6, 14,  
24  
25 3), Arabidopsis (4, 1, 1), orange (4, 15, 4), peach (3, 7, 4), eucalyptus (9, 16, 8), and  
26  
27 *Populus* (5, 12, 2), compared with the higher numbers detected in grape (13, 34, 12).  
28  
29 *PAL* and *CHS* are involved in the key regulatory step in the branch pathway of  
30  
31 phenylpropanoid biosynthesis specific for synthesis of ubiquitous flavonoid pigments  
32  
33 [54], and *F3'5'H* is important for determining flower color [55], which may  
34  
35 suggesting that the *PAL*, *CHS*, and *F3'5'H* encoding genes that were discarded in the  
36  
37 evolution history of longan, Arabidopsis, orange, peach, eucalyptus, and *Populus*  
38  
39 compared with grape were functionally redundant. Besides the expanded and  
40  
41 contracted numbers of structural genes, other structural genes, namely *DHS*, *DHQS*,  
42  
43 *SK*, *EPSP*, *CS*, *CM*, *ADT*, *C4H*, *4CL*, *CHI*, *F3H*, *DFR*, and *ANS*, showed little  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 variations in copy numbers among longan, Arabidopsis, orange, peach, grape, poplar,  
2  
3 and eucalyptus, which indicated their evolutionary conservation in different plant  
4  
5 species. Overall, the expanded, contracted, and conserved copy numbers of the 26  
6  
7 selected structural genes among the seven selected plants, defined the different  
8  
9 characteristics of polyphenol biosynthesis in the different species.  
10  
11  
12

13  
14 To further understand the functions of the 26 structural genes, we measured their  
15  
16 expression levels between primary and secondary metabolism during longan  
17  
18 vegetative and reproductive growth (Fig. 3b, Supplemental EXCEL File 7). The PCA  
19  
20 showed that all the genes related to the biosynthesis of polyphenols were similarly  
21  
22 expressed in leaf, pulp, and pericarp, but their expression levels differed among root,  
23  
24 stem, flower bud, flower, young fruit, and seed (Fig. 3b), suggesting these genes may  
25  
26 have tissue-specific roles in longan. Thirteen of the 26 structural genes were found to  
27  
28 be expressed in specific tissues, such as root, flower, flower bud, and/or seed  
29  
30 (Supplemental EXCFL File 7). For example, two members of the *SDH* family,  
31  
32 Cs9g05070.1-D1 and Cs9g05070.1-D5, showed high expression levels during the  
33  
34 vegetative and reproductive stages, especially in pulp and pericarp, while the other  
35  
36 members of the family were barely detectable, suggesting that Cs9g05070.1-D1 and  
37  
38 Cs9g05070.1-D5 may play major roles in the shikimate acid pathway. The six  
39  
40 members of the *PAL* family all exhibited low or undetectable expression levels in pulp,  
41  
42 two had the highest expression levels in stem, and the other four were strongly  
43  
44 expressed in stem, root, leaf, flower, and pericarp. The tissue-specific expression  
45  
46 pattern of *PAL* further confirmed that *PAL* was related to lignin, the structural  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 component of the cell wall in longan [56]. Five of the 14 members of the *CHS* family  
2  
3 were barely detectable among the nine samples; among the other members, the  
4  
5 highest expression levels were observed for four in seed, three in flower bud, and two  
6  
7 in root, suggesting that CHS played important roles in the synthesis of flavonoid  
8  
9 pigments in flower bud and seed. The 35 members of the *F3'H* family (Fig. 3c),  
10  
11 exhibited different temporal and spatial expression levels (Fig. 3d). Among them, the  
12  
13 highest expression levels were observed for one of the members in root, two in stem,  
14  
15 five in leaf, eleven in flower bud, three in flower, six in young fruit, three in pericarp,  
16  
17 and three in seed; while 11 *F3'H* family members were barely detectable in pericarp,  
18  
19 pulp, and seed. For the three members of the *F3'5'H* family, one was detected only in  
20  
21 root and one only in flower bud, implying *F3'H* and *F3'5'H* both played major roles in  
22  
23 determining longan flower colors. Proanthocyanidin synthesis involves both LAR and  
24  
25 ANR (Fig. 3c). The six *ANR* family members and two of the four *LAR* members were  
26  
27 barely detectable in pulp, and all the *ANR* and *LAR* genes were highly expressed in  
28  
29 pericarp, and relatively less expressed in seed (Fig. 3d). Previous studies of 12  
30  
31 varieties of Chinese longan fruit have shown that total polyphenols, tannins, and  
32  
33 proanthocyanidins were most abundant in pericarp, followed by seed and pulp [57].  
34  
35 The high expression levels of *ANR* and *LAR* in pericarp and seed, and their lowest  
36  
37 expression levels in pulp indicated they may determine the tannin composition of  
38  
39 longan fruit, further indicating why whole longan fruit is dried for use in sweet  
40  
41 desserts and soups for human health [58].  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56

57  
58 The MYB family of TFs are involved in the regulation of flavonoid biosynthesis [59].  
59  
60  
61  
62  
63  
64  
65



1 To further investigate the biosynthesis of polyphenols in longan, we compared the  
2  
3 numbers of MYB-encoding genes in longan with their numbers in Arabidopsis,  
4  
5 orange, peach, and grape. We also investigated their expression levels in longan using  
6  
7 the genome and transcriptome data. We detected 94 *R2R3-MYB* genes in longan,  
8  
9 which was more than in orange (74) and peach (88), but less than in grape (116), and  
10  
11 Arabidopsis (141) (Fig. 4a). A neighbor-joining tree of the *MYB* gene family was  
12  
13 constructed (Fig. 4b). The expression profiles of the *MYB* gene family in each tissue  
14  
15 were clustered by PCA. The plots showed that the expression profiles in three of the  
16  
17 tissues (stem, pericarp, and seed) formed one cluster, while the expression profiles of  
18  
19 the other tissues were independently separated, implying that each had a distinct *MYB*  
20  
21 expression profile (Fig. 4c). All members of the *MYB* gene family were expressed at  
22  
23 varying levels among the nine vegetative growth and reproductive growth tissues,  
24  
25 with some preferentially expressed in specific tissues (Fig. 4d, Supplemental EXCEL  
26  
27 File 8). In Arabidopsis, specific *R2R3-MYB* family members, namely *MYB3* -5, -7, -11,  
28  
29 -12, -32, -75, -90, -111, -113, -114, and -123, are known to be involved in regulating  
30  
31 the flavonoid pathway [59]. In longan, only four *R2R3-MYB* genes, which are  
32  
33 homologs of *AtMYB4*, -12, and -123, were found. In Arabidopsis, *AtMYB4*  
34  
35 down-regulated *C4H* and controlled sinapate ester biosynthesis in a UV-dependent  
36  
37 manner; *AtMYB12* up-regulated *CHS*, *CHI*, *F3H*, and *F3'H*, and controlled flavonol  
38  
39 biosynthesis in all the tissues tested; and *AtMYB123* up-regulated *DNS* and controlled  
40  
41 the biosynthesis of proanthocyanidins in the seed coat [59]. In longan, three of the  
42  
43 four homologous *R2R3-MYB* genes reached peaks in root, but were undetected or  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 lowly expressed in pericarp, pulp, and seed (Fig. 4d). The tissue-specific expression  
2  
3 of these genes indicated they may be required for flavonoid biosynthesis.  
4  
5

### 6 **Identification and classification of genes encoding NBS-LRR and LRR-RLK**

7

8  
9 Transcriptome data analysis showed that longan contained a large number of  
10  
11 significantly differentially expressed plant pathogen resistance genes. To further  
12  
13 investigate the molecular basis for longan pathogen susceptibility, we searched for  
14  
15 two classes of resistance genes in the longan genome, those encoding nucleotide  
16  
17 binding site-leucine rich repeat (NBS-LRR) proteins and those encoding leucine rich  
18  
19 repeat-receptor-like kinases (LRR-RLK). We identified 594 NBS-LRR and 338  
20  
21 LRR-RLK encoding genes, which accounted for approximately 1.51% and 0.86% of  
22  
23 the annotated protein-coding genes in longan, respectively. These numbers of  
24  
25 NBS-LRR and LRR-RLK coding genes in the longan genome were more than those  
26  
27 in orange (509, 325) [13], grape (341, 234) [10], kiwifruit (110, 259) [16], peach (425,  
28  
29 268) [14], mei (411, 253) [12], and papaya (60, 134) [9], but nearly half that in apple  
30  
31 (1035, 477) [11] (Additional file 1: Table S19). *NBS* and *LRR* existed before the  
32  
33 divergence of prokaryotes and eukaryotes, but their fusion has been detected only in  
34  
35 land plant lineages [60], which are assumed to have originated from a common  
36  
37 ancestor. A previous study showed that grape was the oldest among the fruits tested  
38  
39 [53]. In this study, the numbers of *NBS-LRR* and *LRR-RLK* genes were either more or  
40  
41 less in longan, orange, kiwifruit, peach, papaya, mei, and apple compared with grape.  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 the selected genomes, which is similar to what was found in grass species [60].  
2  
3 Moreover, the NBS- and LRR-encoding genes were significantly more in apple than  
4  
5 in the other selected fruits, possibly as a result of a whole-genome wide duplication  
6  
7 event in apple [53]. The uneven distribution of NBS-, and LRR-encoding genes on  
8  
9 chromosomes was reported previously in Arabidopsis, rice, grapevine, and poplar [61].  
10  
11 These results suggest that changes in the numbers of genes encoding NBS-LRR and  
12  
13 LRR-RLK in different species may alter the resistance of these species to different  
14  
15 diseases.  
16  
17  
18  
19  
20  
21

22 The 594 encoded NBS-LRRs in longan were classified into six subgroups based on  
23  
24 their protein domains: NBS-LRR (258, 43.43%), coiled-coil-NBS-LRR (150,  
25  
26 25.25%), NBS (122, 20.54%), coiled-coil-NBS (37, 6.23 %), Toll interleukin receptor  
27  
28 (TIR)-NBS-LRR (23, 3.87%), and TIR-NBS (4, 0.67%) (Additional file 1: Table S19).  
29  
30  
31  
32

33 Previous studies have shown that the deduced NBS-LRR proteins can be divided into  
34  
35 two subfamilies, TIR and non-TIR proteins based on their N-terminal features [62].  
36  
37  
38

39 The TIR family of *NBS-LRR* genes probably originated earlier than the non-TIR  
40  
41 family [60]. Here, the number of genes encoding the TIR proteins (TIR-NBS-LRR  
42  
43 and TIR-NBS) varied from one (kiwifruit) to 288 (apple), and the number of genes  
44  
45 encoding the non-TIR proteins was 567 in longan, 415 in orange, 320 in grape, 109 in  
46  
47 kiwifruit, 282 in peach, 53 in papaya, and 753 in apple. The ratio of TIR to non-TIR  
48  
49 genes was found to differ markedly in different species [62], suggesting ancient  
50  
51 origins and subsequent divergence between the two NBS gene types. The distribution  
52  
53 of resistance genes in the longan genome and the encoded domains are similar to  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 those of the resistance proteins in other sequenced genomes, as shown in Additional  
2  
3 file 1: Table S19. In addition, we noted that allelic variations due to the presence of  
4  
5 SNPs in NBS-encoding genes were associated with the phenotypic divergence  
6  
7 between resistant ('FY', 'SN1H', 'MQ', 'LDB', and 'JYW') and susceptible ('SX', and  
8  
9 'YTB') longan accessions. Such detailed knowledge of the longan genome will help  
10  
11 to accelerate the development of genetic strategies to counter fruit loss caused by  
12  
13 diverse pathogens [30].  
14  
15  
16  
17  
18  
19  
20

## 21 **Conclusions**

22  
23 Here, a draft genome of *D. longan* is presented for the first time. The assembled  
24  
25 genome sequence is 471.88 Mb with 273.44-fold coverage obtained by paired-end  
26  
27 sequencing. Whole-genome resequencing and analysis of 13 representative cultivated  
28  
29 *D. longan* accessions revealed the extent of genetic diversity and contributed to trait  
30  
31 discovery. Annotation of the protein-coding genes, comparative genomic analysis,  
32  
33 and transcriptome analyses provided insights into longan-specific traits, particularly  
34  
35 those involved in the biosynthesis of secondary metabolites and pathogen resistance.  
36  
37  
38  
39  
40  
41  
42

## 43 **Methods**

### 44 **Germplasm genetic resources**

45  
46 An 80-year old *D. longan* 'HHZ' cultivar from the Fujian Agriculture and Forestry University,  
47  
48 China, was used for genomic DNA isolation and sequencing. RNA samples from root, leaf,  
49  
50 floral bud, flower, young fruit, mature fruit, pericarp, pulp, and seed tissues of the *D. longan*  
51  
52 'SJM' cultivar from the experimental fields of Fujian Academy of Agricultural Science in  
53  
54 Putian, Fujian Province, were collected for transcriptome sequencing. Fourteen *D. longan*  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 cultivars, ‘HHZ’, ‘SJM’, ‘SN1H’, ‘JYW’, ‘SX’, ‘WLL’, ‘MQ’, ‘YTB’, ‘SEY’, ‘LDB’,  
2  
3 ‘JHLY’, ‘FY’, ‘DB’, and ‘SFB’, that originated or are widely grown in Asia and other regions  
4  
5  
6 of the world, were collected for resequencing.  
7

### 8 **DNA extraction, library construction, whole-genome shotgun sequencing and assembly**

9  
10 Whole-genome shotgun sequencing was performed using the Illumina HiSeq 2000 system.  
11  
12 Genomic DNA was extracted from fresh mature leaves of the *D. longan* ‘HHZ’ cultivar using  
13  
14 the modified SDS method. DNA sequencing libraries were constructed according to the  
15  
16 standard Illumina library preparation protocols. A total of 12 paired-end sequencing libraries,  
17  
18 spanning 170, 250, 500, 800, 2,000, 5,000, 10,000, 20,000, and 40,000 bp, were constructed  
19  
20 and sequenced on an Illumina HiSeq 2000 system. After stringent filtering and correction  
21  
22 steps using K-mer frequency-based methods [21], a total of 121.68 Gb of data were obtained,  
23  
24 and then assembled using SOAPdenovo and SSPACE software [63]. To check the  
25  
26 completeness of the assembly, a longan transcriptome assembly comprising 68,925 unigenes  
27  
28 [SRA050205] was mapped to the genome assembly using BLAT32 with various sequence  
29  
30 homology and coverage parameters. The BUSCO pipeline [27] was also used to check the  
31  
32 genome completeness.  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43

### 44 **Repetitive elements identification**

45  
46 Tandem repeats and interspersed repeats are two main types of repeats found in genomes.  
47  
48 Tandem repeats were identified using LTR\_FINDER[64] with the default parameters.  
49  
50 Interspersed repeats were identified by Repeat Masker (<http://www.repeatmasker.org/>) and  
51  
52 RepeatProteinMask using the Repbase library [65] and the *de novo* transposable element  
53  
54 library. Identified repeats were then classified into different known classes, as previously  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 described [33].  
2

### 3 **Gene prediction and annotation** 4

5  
6 For gene prediction, the scaffolds were first repeat-masked [65]. Then, three *de novo*  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

For gene prediction, the scaffolds were first repeat-masked [65]. Then, three *de novo* homology-based and RNA-seq unigenes-based prediction methods, Augustus[66], GENSCAN [67], and GlimmerHMM [68], were used with parameters trained on *Arabidopsis thaliana* and *Carica papaya*. The *de novo* predictions were then merged into a unigene set. For the homology search, translated protein sequences from three sequenced plant genomes (*Glycine max*, *Populus trichocarpa*, and *Vitis vinifera*) were mapped to the longan genome assembly using TBLASTN (E-value cutoff  $1 \times 10^{-5}$ ). To extract accurate exon–intron information, the homologous genome sequences were aligned against the matching proteins using GeneWise [69]. Subsequently, the Illumina RNA-seq unigenes sequences [26] were aligned to the longan genome assembly using BLAT [70] to detect spliced alignments.

Finally, to generate the consensus gene set, the results obtained using the three methods described above were integrated using the GLEAN program [71]. The final gene set contained 39,282 genes. TFs were identified and classified using the TAK program [72]. Non-coding RNAs were predicted and classified, as previously described [73]. Functions of the predicted protein genes were obtained by BLAST searches (E-value cutoff  $1 \times 10^{-5}$ ) against the InterproScan [74], GO [75], KEGG [76], SwissProt [77], and TrEMBL databases.

### 50 **Gene families and phylogenetic analysis** 51

52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

To identify gene families, the translated proteins sequences from *T. cacao*, *C. sinensis*, *A. thaliana*, *C. papaya*, *Populus trichocarpa*, *Glycine max*, *V. vinifera*, *M. acuminata*, *P. persica*, *A.chinensis*, and *M.domestica* genomes were scanned using BLASTP (E-value cutoff

1e-5), and gene family clusters among the different plant species were identified by OrthoMCL [78]. Single-copy families that were represented in all the selected species were alignment using MUSCLE [79]. 4DTv in the 12 species, including longan, were used to construct a phylogenetic tree by MRBAYES [80]. The divergence time was estimated using the MultiDivtime software [79]. Colinearity between *D. longan* and *P. trichocarpa* was computed by SyMAP v3.4 [81]. Subsequently, TF families were identified using the IPR2genomes tool in GreenphylDB v2.0 [82] based on InterPro domains, and gene family expansion and contraction within phylogenetically-related organisms was detected by CAFÉ, a tool for computational analysis of gene family evolution [31].

#### **Resequencing, SNPs, indels, and sequence variations analysis**

Paired-end Illumina libraries for 13 *D. longan* cultivars were prepared following the manufacturer's instructions and sequenced on an Illumina HiSeq 2000 system. After stringent filtering and correction steps, the resulting sequence data were uniquely aligned to the reference longan genome. SNPs, indels, and sequence variations were identified using SOAPsnp (<http://soap.genomics.org.cn/soapsnp.html>), SOAPindel [83], and SOAPsv [84].

We used all and high quality SNPs to infer the phylogeography and population structure for *D. longan*. A phylogenetic tree was subsequently generated using the neighbor-joining method implemented in TreeBeST. The bootstrap was set as 1000 replicates.

Population structure was examined primarily via PCA using our own program and model-based clustering algorithms implemented in FRAPPE v1.1 (<http://smstaging.stanford.edu/tanglab/software/frappe.html>), We increased the pre-defined genetic clusters from K2 to K7 and ran the analysis with 10,000 maximum iterations.

## Transcriptome sequencing

Transcriptome sequencing was performed on the Illumina HiSeq 2000 system. Total RNAs from the samples described above were isolated using a TRIzol Reagent kit (Invitrogen, Carlsbad, CA). cDNA libraries were constructed and sequenced using the Illumina protocols. All the raw reads were first processed to remove the adaptor sequences, low quality reads, and possible contaminations from chloroplast, mitochondrion, and ribosomal DNA. The clean reads were then aligned to the longan genome sequence using TopHat [46] to identify exons and splice junctions *ab initio*. The expression levels of matched genes in each cDNA library were derived and normalized to fragments per kilobase of exon per million fragments mapped. Cluster 3.0 [85] was used to analyze hierarchical clustering of genes. DEGs among different samples were identified using the EBSeq packages [86]. Subsequently, GATK (<http://www.broadinstitute.org/gatk/>) with default parameters was used to call SNPs based on the transcript sequence data.

## Identification of genes associated with secondary metabolites

We downloaded all the proteins from Arabidopsis, orange, peach, and grape, and identified the genes encoding them using the following methods. First, we collected previously published related genome sequences as the query sequences. We then used TBLASTN (Legacy Blast v2.2.23) [70] to align the query sequences against each genome sequence (E-value cutoff  $<1e-10$ . Because many query sequences aligned to the same genomic region, we extracted only the high quality alignments (Query\_align\_ratio  $\geq 70\%$  and Identity  $\geq 40\%$ ). Functional intact genes were confirmed as follows. First, we collected the blast-hits as described above. Then, we extended each of the blast-hits sequences in both the 3' and 5'



1 directions along the genome sequences and predicted the gene structure by Genewise (v2.2.0)  
2  
3 [69]. Using this approach, we obtained all the pathway genes in longan and the other fruit  
4  
5  
6 plants.

### 7 8 9 **Identification of *MYB* genes**

10  
11 We download the annotated *MYB* genes from Arabidopsis, orange, peach, and grape,  
12  
13 and applied identification methods that were similar to those described in the  
14  
15 ‘Identification of genes associated with secondary metabolites’ section.  
16  
17

### 18 19 20 **Disease resistance genes analysis**

21  
22 Identification of longan resistance-related genes was based on the most conserved  
23  
24 motif structures of plant resistance proteins. Details of the methods used were as  
25  
26 described in [30].  
27  
28

### 29 30 31 **Availability of data and material**

32  
33 The draft genome sequencing project of *D. longan* is registered at NCBI under  
34  
35 BioProject [PRJNA305337]. The NCBI SRA database with accession numbers  
36  
37 [SRA315202], and the sample Accession were [SRS1272137], [SRS1272138],  
38  
39 [SRS1272139], and [SRS1272140]. Sequencing data, annotations. The *D. longan*  
40  
41 ‘SJM’ transcriptome data is deposited at NCBI under BioProject [PRJNA326792].  
42  
43  
44  
45  
46  
47  
48  
49

### 50 51 **Abbreviations**

52  
53 **Mb:** million base; **SNPs:** single nucleotide polymorphisms; **indels:** insertions/  
54  
55 deletions; **PCA:** principal component analysis; **DHS:** 3-deoxy-D-arabino-  
56  
57 heptulosonate 7-phosphate synthase; **DHQS:** 3-dehydroquininate synthase; **SDH:**  
58  
59  
60  
61

1 bifunctional 3- dehydroquinate dehydratase/ shikimate dehydrogenase; **SK:** shikimate  
2  
3 kinase; **EPSPS:** 3-phosphoshikimate 1-carboxyvinyltransferase/  
4  
5  
6 5-enolpyruvylshikimate- 3- phosphate/ EPSP synthase; **CS:** chorismate synthase; **CM:**  
7  
8 chorismate mutase; **ADT:** arogenate dehydratase/ prephenate dehydratase; **PAL:**  
9  
10 phenylalanine ammonia lyase; **C4H:** cinnamate 4-hydroxylase; **4CL:** 4-coumaroyl-  
11  
12 coenzyme A ligase; **CHS:** chalcone synthase; **CHI:** chalcone-flavanone isomerase;  
13  
14 **F3H:** flavanone 3-hydroxylase; **F3'H:** flavonoid 3'-hydroxylase; **F3'5'H:** flavonoid  
15  
16 3',5'-hydroxylase; **ANS:** anthocyanidin synthase; **LDOX:** leucoanthocyanidin  
17  
18 dioxygenase; **DFR:** dihydroflavonol 4-reductase; **LAR:** leucoanthocyanidin  
19  
20 reductase.  
21  
22  
23  
24  
25  
26

## 27 **Declarations**

28  
29 The authors declare no competing financial interests.  
30  
31

## 32 **Additional files**

33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Supplementary EXCEL File 1 Identification of transcription factors in the  
*Dimocarpus longan* genome

Supplementary EXCEL File 2 Significantly expanded gene families detected in the  
*Dimocarpus longan* genome (Viterbi  $p \leq 0.05$ )

Supplementary EXCEL File 3 Significantly contracted gene families detected in the  
*Dimocarpus longan* genome (Viterbi  $p \leq 0.05$ )

Supplementary EXCEL File 4 SNP analysis of FY, SN1H, MQ, LDB, and JYW  
cultivars

Supplementary EXCEL File 5 SNP analysis of SX and YTB cultivars

1 Supplementary EXCEL File 6 Statistics of copy numbers of genes involved in the  
2 biosynthesis of polyphenols in different plants

3  
4 Supplementary EXCEL File 7 Expression levels of genes involved in the biosynthesis  
5 of polyphenols in *Dimocarpus longan*

6  
7  
8 Supplementary EXCEL File 8 MYB genes expressed in nine different tissues of  
9 *Dimocarpus longan*

### 10 11 12 **Consent for publication**

13  
14  
15  
16 Not applicable

### 17 18 19 **COMPETING FINANCIAL INTERESTS**

20  
21  
22 The authors declare no competing financial interests.

### 23 24 **Funding**

25  
26  
27 This work was funded by the Research Funds for the National Natural Science Foundation of  
28 China (31672127, 31572088, 31272149, 31201614, and 31078717), the Science and  
29 Technology Plan Major Projects of Fujian Province (2015NZ0002-1), the Natural Science  
30 Funds for Distinguished Young Scholar in Fujian Province (2015J06004), the program for  
31 New Century Excellent Talents in Fujian Province University (20151104), the Doctoral  
32 Program of Higher Education of the Chinese Ministry of Education (20093515110005 and  
33 20123515120008), the Education Department of Fujian Province Science and Technology  
34 Project (JA14099), the Program for High-level University Construction of the Fujian  
35 Agriculture and Forestry University (612014028), and the Natural Science Funds for  
36 Distinguished Young Scholar of the Fujian Agriculture and Forestry University (xjq201405).

### 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 **Authors' contributions**

54  
55  
56  
57 ZXL, YLL, YY, and RKV designed the research; YLL, ZXL, RLL, YKC, CZC, QLT,  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000

1 the DNA and RNA. LLY, ZYW, QFL, and YH did the sequencing, processed the raw  
2  
3 data, and assembled the sequences. XDF, ZYW, CGZ, JW, and MHY coordinated the  
4  
5 project. JMM, LLY, ZYW, QFL, YH, and YLL analyzed the data. YLL, ZXL, YY,  
6  
7  
8  
9 JMM, and RKV wrote and revised the paper.

## 11 Acknowledgments

12  
13  
14  
15 We thank the following colleagues from the experimental fields of the Fujian  
16  
17  
18 Academy of Agricultural Science in Putian for samples.

## 21 References:

- 22 1. Lai Z, Chen C, Zeng L, Chen Z: **Somatic embryogenesis in longan**  
23 **[*Dimocarpus longan* Lour.].** In: *Somatic Embryogenesis in Woody Plants.*  
24 Edited by Jain SM, Gupta P, Newton R, vol. 67: Springer Netherlands; 2000:  
25 415-431.
- 26 2. Luo J, Zhou C-f, Wan Z: **Analysis on the Development Status of**  
27 **Lychee Industry in Guangdong Province in 2010.** *Guangdong*  
28 *Agricultural Sciences* 2011, **4**:16-18.
- 29 3. Mei ZQ, Fu SY, Yu HQ, Yang LQ, Duan CG, Liu XY, Gong S, Fu JJ: **Genetic**  
30 **characterization and authentication of *Dimocarpus longan* Lour. using an**  
31 **improved RAPD technique.** *Genet Mol Res* 2014, **13**(1):1447-1455.
- 32 4. Jiang G, Jiang Y, Yang B, Yu C, Tsao R, Zhang H, Chen F: **Structural**  
33 **characteristics and antioxidant activities of oligosaccharides from longan**  
34 **fruit pericarp.** *Journal of agricultural and food chemistry* 2009,  
35 **57**(19):9293-9298.
- 36 5. Chung YC, Lin CC, Chou CC, Hsu CP: **The effect of Longan seed**  
37 **polyphenols on colorectal carcinoma cells.** *European journal of clinical*  
38 *investigation* 2010, **40**(8):713-721.
- 39 6. Prasad KN, Yang B, Shi J, Yu C, Zhao M, Xue S, Jiang Y: **Enhanced**  
40 **antioxidant and antityrosinase activities of longan fruit pericarp by**  
41 **ultra-high-pressure-assisted extraction.** *Journal of pharmaceutical and*  
42 *biomedical analysis* 2010, **51**(2):471-477.
- 43 7. Lin T, Lin Y, Ishiki K: **Genetic diversity of *Dimocarpus longan* in China**  
44 **revealed by AFLP markers and partial rbcL gene sequences.** *Scientia*  
45 *Horticulturae* 2005, **103**(4):489-498.
- 46 8. Yonemoto Y, Chowdhury AK, Kato H, Macha MM: **Cultivars identification**  
47 **and their genetic relationships in *Dimocarpus longan* subspecies based on**  
48 **RAPD markers.** *Scientia Horticulturae* 2006, **109**(2):147-152.

- 1  
2  
3 9. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W,  
4 Ly BV, Lewis KL *et al*: **The draft genome of the transgenic tropical fruit**  
5 **tree papaya (*Carica papaya* Linnaeus)**. *Nature* 2008, **452**(7190):991-996.
- 6  
7 10. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N,  
8 Aubourg S, Vitulo N, Jubin C *et al*: **The grapevine genome sequence**  
9 **suggests ancestral hexaploidization in major angiosperm phyla**. *Nature*  
10 2007, **449**(7161):463-467.
- 11  
12 11. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A,  
13 Fontana P, Bhatnagar SK, Troggio M, Pruss D *et al*: **The genome of the**  
14 **domesticated apple (*Malus x domestica* Borkh.)**. *Nature genetics* 2010,  
15 **42**(10):833-839.
- 16  
17 12. Zhang Q, Chen W, Sun L, Zhao F, Huang B, Yang W, Tao Y, Wang J, Yuan Z,  
18 Fan G *et al*: **The genome of *Prunus mume***. *Nature communications* 2012,  
19 **3**:1318.
- 20  
21 13. Xu Q, Chen LL, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao WB, Hao  
22 BH, Lyon MP *et al*: **The draft genome of sweet orange (*Citrus sinensis*)**.  
23 *Nature genetics* 2013, **45**(1):59-66.
- 24  
25 14. Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T,  
26 Dettori MT, Grimwood J, Cattonaro F *et al*: **The high-quality draft genome**  
27 **of peach (*Prunus persica*) identifies unique patterns of genetic diversity,**  
28 **domestication and genome evolution**. *Nature genetics* 2013, **45**(5):487-494.
- 29  
30 15. Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, Khan MA, Tao S, Korban SS,  
31 Wang H *et al*: **The genome of the pear (*Pyrus bretschneideri* Rehd.)**.  
32 *Genome Res* 2013, **23**(2):396-408.
- 33  
34 16. Huang S, Ding J, Deng D, Tang W, Sun H, Liu D, Zhang L, Niu X, Zhang X,  
35 Meng M *et al*: **Draft genome of the kiwifruit *Actinidia chinensis***. *Nature*  
36 *communications* 2013, **4**:2640.
- 37  
38 17. Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, Lyons E,  
39 Wang M-L, Chen J, Biggers E *et al*: **The pineapple genome and the**  
40 **evolution of CAM photosynthesis**. *Nature genetics* 2015, **advance online**  
41 **publication**.
- 42  
43 18. D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B,  
44 Bocs S, Droc G, Rouard M *et al*: **The banana (*Musa acuminata*) genome and**  
45 **the evolution of monocotyledonous plants**. *Nature* 2012,  
46 **488**(7410):213-217.
- 47  
48 19. Ma Q, Feng K, Yang W, Chen Y, Yu F, Yin T: **Identification and**  
49 **characterization of nucleotide variations in the genome of *Ziziphus jujuba***  
50 **(Rhamnaceae) by next generation sequencing**. *Mol Biol Rep* 2014,  
51 **41**(5):3219-3223.
- 52  
53 20. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL,  
54 Jaiswal P, Mockaitis K, Liston A, Mane SP *et al*: **The genome of woodland**  
55 **strawberry (*Fragaria vesca*)**. *Nature genetics* 2011, **43**(2):109-116.
- 56  
57 21. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y *et al*:  
58 **The sequence and de novo assembly of the giant panda genome**. *Nature*  
59  
60  
61  
62  
63  
64  
65

- 2010, **463**(7279):311-317.
22. Sun L, Zhang Q, Xu Z, Yang W, Guo Y, Lu J, Pan H, Cheng T, Cai M: **Genome-wide DNA polymorphisms in two cultivars of mei (*Prunus mume sieb. et zucc.*)**. *BMC Genet* 2013, **14**:98.
  23. Brunner AM, Busov VB, Strauss SH: **Poplar genome sequence: functional genomics in an ecologically dominant plant species**. *Trends in plant science* 2004, **9**(1):49-56.
  24. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K *et al*: **De novo assembly of human genomes with massively parallel short read sequencing**. *Genome Res* 2010, **20**(2):265-272.
  25. Du H, Hu H, Meng Y, Zheng W, Ling F, Wang J, Zhang X, Nie Q, Wang X: **The correlation coefficient of GC content of the genome-wide genes is positively correlated with animal evolutionary relationships**. *FEBS Lett* 2010, **584**(18):3990-3994.
  26. Lai Z, Lin Y: **Analysis of the global transcriptome of longan (*Dimocarpus longan* Lour.) embryogenic callus using Illumina paired-end sequencing**. *BMC Genomics* 2013, **14**:561.
  27. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.
  28. Lee H, Golicz AA, Bayer PE, Jiao Y, Tang H, Paterson AH, Sablok G, Krishnaraj RR, Chan CK, Batley J *et al*: **The Genome of a Southern Hemisphere Seagrass Species (*Zostera muelleri*)**. *Plant Physiol* 2016, **172**(1):272-283.
  29. Meyers BC, Tingey SV, Morgante M: **Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome**. *Genome Res* 2001, **11**(10):1660-1676.
  30. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J *et al*: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety**. *PLoS One* 2007, **2**(12):e1326.
  31. De Bie T, Cristianini N, Demuth JP, Hahn MW: **CAFE: a computational tool for the study of gene family evolution**. *Bioinformatics* 2006, **22**(10):1269-1271.
  32. Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z *et al*: **The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions**. *Nature genetics* 2013, **45**(1):51-58.
  33. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P *et al*: **The genome of the cucumber, *Cucumis sativus* L.** *Nature genetics* 2009, **41**(12):1275-1281.
  34. Wang B, Tan HW, Fang W, Meinhardt LW, Mischke S, Matsumoto T, Zhang D: **Developing single nucleotide polymorphism (SNP) markers from transcriptome sequences for identification of longan (*Dimocarpus longan*) germplasm**. *Horticulture research* 2015, **2**:14065.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
35. Zhu J, Pan L, Qin X, Peng H, Wang Y, Hang Z: **Analysis on genetic relations in different ecotypes of longan (*Dimocarpus longan* Lour.) germplasm resources by ISSR markers.** *Journal of Plant Genetic Resources* 2013(01):65-69.
  36. Zhong F, Pan D, Guo Z, Lin L, Li K: **RAPD Analysis of Longan Germplasm Resources.** *Chinese agricultural science bulletin* 2007(07):558-563.
  37. Tang H, Peng J, Wang P, Risch NJ: **Estimation of individual admixture: analytical and study design considerations.** *Genetic epidemiology* 2005, **28**(4):289-301.
  38. Peng J, Xie L, Xu B, Dang J, Li Y, Lu Z, Zhang S, Yu Z, Bai X, Cai Z: **Study on Biological Characters of 'Sijihua'Longan.** In: *III International Symposium on Longan, Lychee, and other Fruit Trees in Sapindaceae Family 863: 2008.* 249-258.
  39. Wang L, Zhao S, Gu C, Zhou Y, Zhou H, Ma J, Cheng J, Han Y: **Deep RNA-Seq uncovers the peach transcriptome landscape.** *Plant molecular biology* 2013, **83**(4-5):365-377.
  40. Tong C, Wang X, Yu J, Wu J, Li W, Huang J, Dong C, Hua W, Liu S: **Comprehensive analysis of RNA-seq data reveals the complexity of the transcriptome in *Brassica rapa*.** *BMC Genomics* 2013, **14**:689.
  41. Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH: **Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*.** *Plant Cell* 2012, **24**(11):4333-4345.
  42. Wang XJ, Gaasterland T, Chua NH: **Genome-wide prediction and identification of cis-natural antisense transcripts in *Arabidopsis thaliana*.** *Genome Biol* 2005, **6**(4):R30.
  43. Jia T, Wei D, Meng S, Allan AC, Zeng L: **Identification of regulatory genes implicated in continuous flowering of longan (*Dimocarpus longan* L.).** *PLoS One* 2014, **9**(12):e114568.
  44. Lin Y, Lai Z: **Comparative analysis reveals dynamic changes in miRNAs and their targets and expression during somatic embryogenesis in longan (*Dimocarpus longan* Lour.).** *PLoS One* 2013, **8**(4):e60337.
  45. Lin CC, Chung YC, Hsu CP: **Potential roles of longan flower and seed extracts for anti-cancer.** *World journal of experimental medicine* 2012, **2**(4):78-85.
  46. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**(9):1105-1111.
  47. Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, Wang H, Williams A, Blume JE: **Discovery of tissue-specific exons using comprehensive human exon microarrays.** *Genome Biol* 2007, **8**(4):R64.
  48. Vitulo N, Forcato C, Carpinelli EC, Telatin A, Campagna D, D'Angelo M, Zimbello R, Corso M, Vannozzi A, Bonghi C *et al*: **A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype.** *BMC Plant Biol* 2014, **14**:99.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
49. Wang BB, Brendel V: **Genomewide comparative analysis of alternative splicing in plants**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(18):7175-7180.
  50. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC: **Genome-wide mapping of alternative splicing in *Arabidopsis thaliana***. *Genome Res* 2010, **20**(1):45-58.
  51. Potenza E, Racchi ML, Sterck L, Collier E, Asquini E, Tosatto SC, Velasco R, Van de Peer Y, Cestaro A: **Exploration of alternative splicing events in ten different grapevine cultivars**. *BMC Genomics* 2015, **16**:706.
  52. Reddy AS, Marquez Y, Kalyna M, Barta A: **Complexity of the alternative splicing landscape in plants**. *Plant Cell* 2013, **25**(10):3657-3683.
  53. Michael TP, VanBuren R: **Progress, challenges and the future of crop genomes**. *Curr Opin Plant Biol* 2015, **24**:71-81.
  54. Assis JS, Maldonado R, Muñoz T, Escribano MaI, Merodio C: **Effect of high carbon dioxide concentration on PAL activity and phenolic contents in ripening cherimoya fruit**. *Postharvest Biology and Technology* 2001, **23**(1):33-39.
  55. Togami J, Tamura M, Ishiguro K, Hirose C, Okuhara H, Ueyama Y, Nakamura N, Yonekura-Sakakibara K, Fukuchi-Mizutani M, Suzuki K-i *et al*: **Molecular characterization of the flavonoid biosynthesis of *Verbena hybrida* and the functional analysis of verbena and *Clitoria ternatea* F3'5'H genes in transgenic verbena**. *Plant Biotechnology* 2006, **23**(1):5-11.
  56. Zhang X, Gou M, Liu CJ: **Arabidopsis Kelch repeat F-box proteins regulate phenylpropanoid biosynthesis via controlling the turnover of phenylalanine ammonia-lyase**. *Plant Cell* 2013, **25**(12):4994-5010.
  57. He N, Wang Z, Yang C, Lu Y, Sun D, Wang Y, Shao W, Li Q: **Isolation and identification of polyphenolic compounds in longan pericarp**. *Separation and Purification Technology* 2009, **70**(2):219-224.
  58. Tseng HC, Wu WT, Huang HS, Wu MC: **Antimicrobial activities of various fractions of longan (*Dimocarpus longan* Lour. Fen Ke) seed extract**. *International journal of food sciences and nutrition* 2014.
  59. Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L: **MYB transcription factors in Arabidopsis**. *Trends in plant science* 2010, **15**(10):573-581.
  60. Yue JX, Meyers BC, Chen JQ, Tian D, Yang S: **Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes**. *New Phytol* 2012, **193**(4):1049-1063.
  61. Li J, Ding J, Zhang W, Zhang Y, Tang P, Chen JQ, Tian D, Yang S: **Unique evolutionary pattern of numbers of gramineous NBS-LRR genes**. *Mol Genet Genomics* 2010, **283**(5):427-438.
  62. Yang S, Zhang X, Yue JX, Tian D, Chen JQ: **Recent duplications dominate NBS-encoding gene expansion in two woody species**. *Mol Genet Genomics* 2008, **280**(3):187-198.
  63. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding**



- pre-assembled contigs using SSPACE. *Bioinformatics* 2011, **27**(4):578-579.
64. Xu Z, Wang H: **LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.** *Nucleic acids research* 2007, **35**(Web Server issue):W265-268.
  65. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenetic and genome research* 2005, **110**(1-4):462-467.
  66. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts.** *Nucleic acids research* 2006, **34**(Web Server issue):W435-439.
  67. Salamov AA, Solovyev VV: **Ab initio gene finding in Drosophila genomic DNA.** *Genome Res* 2000, **10**(4):516-522.
  68. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.** *Bioinformatics* 2004, **20**(16):2878-2879.
  69. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res* 2004, **14**(5):988-995.
  70. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656-664.
  71. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM: **Creating a honey bee consensus gene set.** *Genome Biol* 2007, **8**(1):R13.
  72. Zheng Y, Jiao C, Sun H, Rosli HG, Pombo MA, Zhang P, Banf M, Dai X, Martin GB, Giovannoni JJ *et al*: **iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases.** *Mol Plant* 2016, **9**(12):1667-1670.
  73. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B *et al*: **Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement.** *Nature biotechnology* 2013, **31**(3):240-246.
  74. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**(9):847-848.
  75. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25**(1):25-29.
  76. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic acids research* 2000, **28**(1):27-30.
  77. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic acids research* 2000, **28**(1):45-48.
  78. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**(9):2178-2189.
  79. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy**

**and high throughput.** *Nucleic acids research* 2004, **32**(5):1792-1797.

80. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**(8):754-755.
81. Soderlund C, Bomhoff M, Nelson WM: **SyMAP v3.4: a turnkey synteny system with application to plant genomes.** *Nucleic acids research* 2011, **39**(10):e68.
82. Rouard M, Guignon V, Aluome C, Laporte MA, Droc G, Walde C, Zmasek CM, Perin C, Conte MG: **GreenPhylDB v2.0: comparative and functional genomics in plants.** *Nucleic acids research* 2011, **39**(Database issue):D1095-1102.
83. Li S, Li R, Li H, Lu J, Li Y, Bolund L, Schierup MH, Wang J: **SOAPindel: efficient identification of indels from short paired reads.** *Genome Res* 2013, **23**(1):195-200.
84. Li Y, Zheng H, Luo R, Wu H, Zhu H, Li R, Cao H, Wu B, Huang S, Shao H *et al*: **Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly.** *Nature biotechnology* 2011, **29**(8):723-730.
85. de Hoon MJ, Imoto S, Nolan J, Miyano S: **Open source clustering software.** *Bioinformatics* 2004, **20**(9):1453-1454.
86. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C: **EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments.** *Bioinformatics* 2013, **29**(8):1035-1043.

**Tables****Table 1 D. *longan* genome assembly**

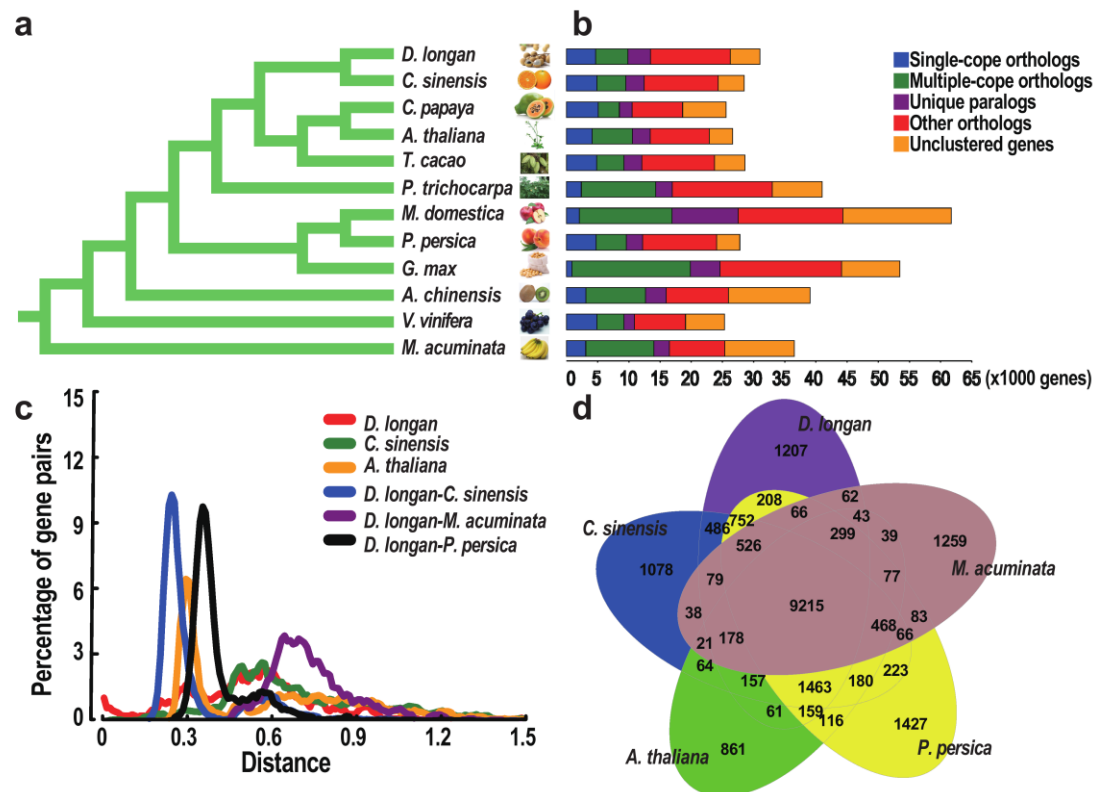
	Contig		Scaffold	
	Size(bp)	Number	Size(bp)	Number
N90	6,457	18,861	122,626	983
N80	11,286	13,434	197,247	668
N70	15,938	9,933	283,489	459
N60	20,685	7,339	396,999	309
N50	26,035	5,306	566,629	204
Longest	173,288		6,942,318	
Total size	471,874,380		495,332,425	
Total number(>=200bp)		51,392		17,367
Total number(>=2Kb)		27,296		2,282

**Table 2 Statistics and comparison of the *D. longan* assembly to other twelve genomes.** Dl, *Dimocarpus longan*; Cs, *Citrus sinensis*; Cc, *Citrus Clementina*; Cp, *Carica papaya*; Ac, *Actinidia chinensis*; Md, *Malus domestica*; Pp, *Prunus persica*; Pb, *Pyrus bretschneideri*; Vv, *Vitis vinifera*; An, *Ananas comosus (L.) Merr.*; Zj, *Ziziphus jujuba* Mill.; Mn, *Morus notabilis*; Tc, *Theobroma cacao*.

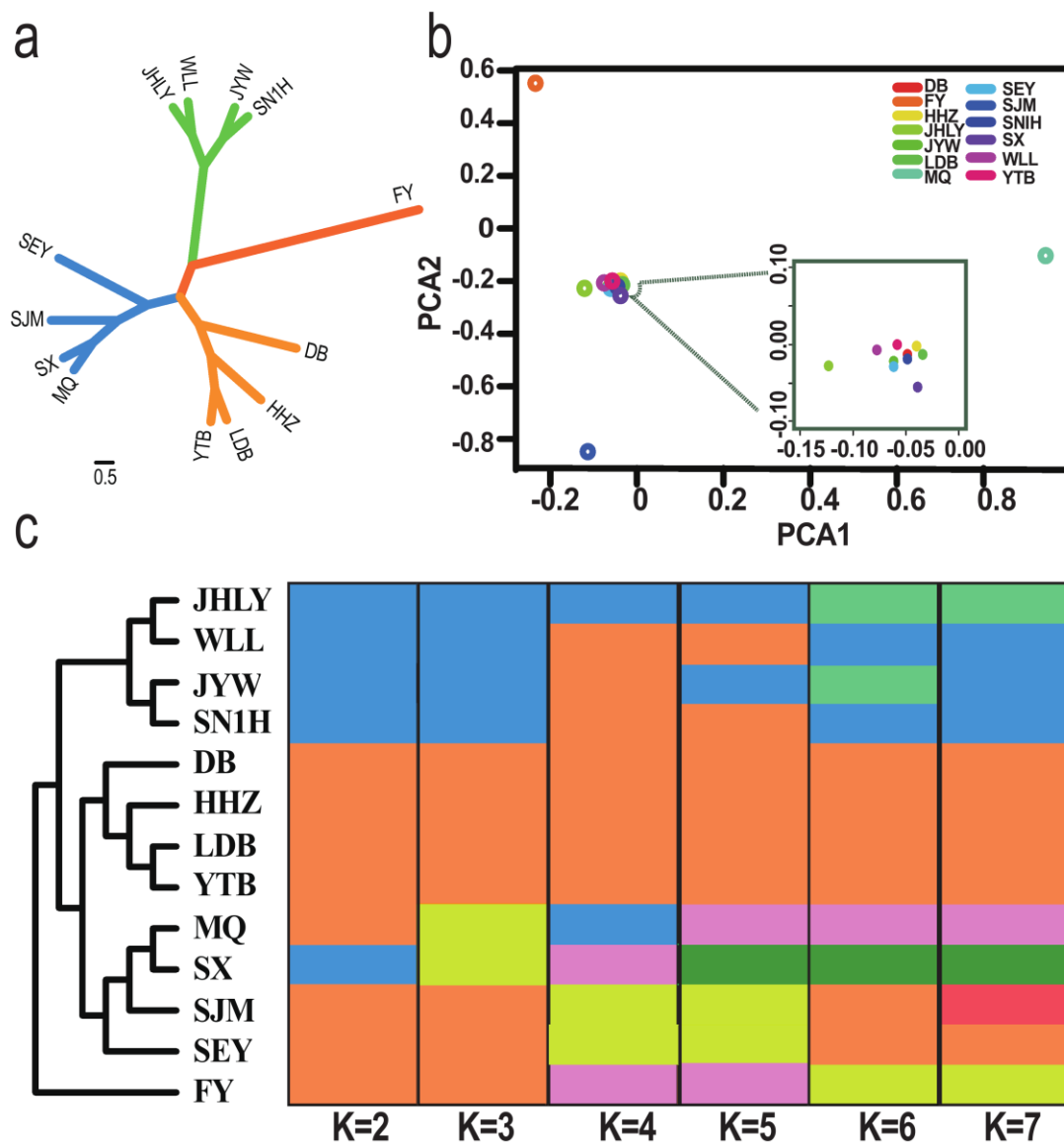
	Dl	Cs	Cc	Cp	Ac	Md	Pp	Pb	Vv	An	Zj	Mn	Tc
Chromosome number ( $2n$ )	30	18	18	18	58	34	16	34	38	50	24	14	20
Estimate of genome size (Mb)	445	367	370	372	758	742.3	265	527	475	526	444	357	430
Sequence Coverage	273.43	214	7	NA	140	16.9	8.47	194	8.4	400	390	236	16.7
Assembled (Mb)	471.88	320	301	271	616.1	603.9	226.6	512	487	382	437.65	330	326.9
Assembling represent percentage of genome (%)	106.4	87.30	81.4	75	81	81.3	85.50	97.10	102.5	73	98.60	92.4	76
N50 length of contig (Kb)	26.03	49.89	NA	NA	58.9	16.17	294	35.7	65.9	126.5	33.9	34.4	19.8
N50 length of scaffolds (Mb)	0.56662	1.69	NA	NA	0.646	NA	4	0.54	2	11.8	0.3	0.39	0.4738
GC content (%)	33.7	34.06	NA	35.3	35.20	NA	NA	NA	35	33	33.41	35	NA
Repeat content (%)	52.87	20	NA	51.90	36	67.4	29.60	53.10	41.40	38.30	49.49	38.8	25.70
Number of gene models	31,007	29,445	24,533	24,746	39,040	57,386	27,852	42,812	30,434	27,024	32,808	27,085	28,798

NA, no available.

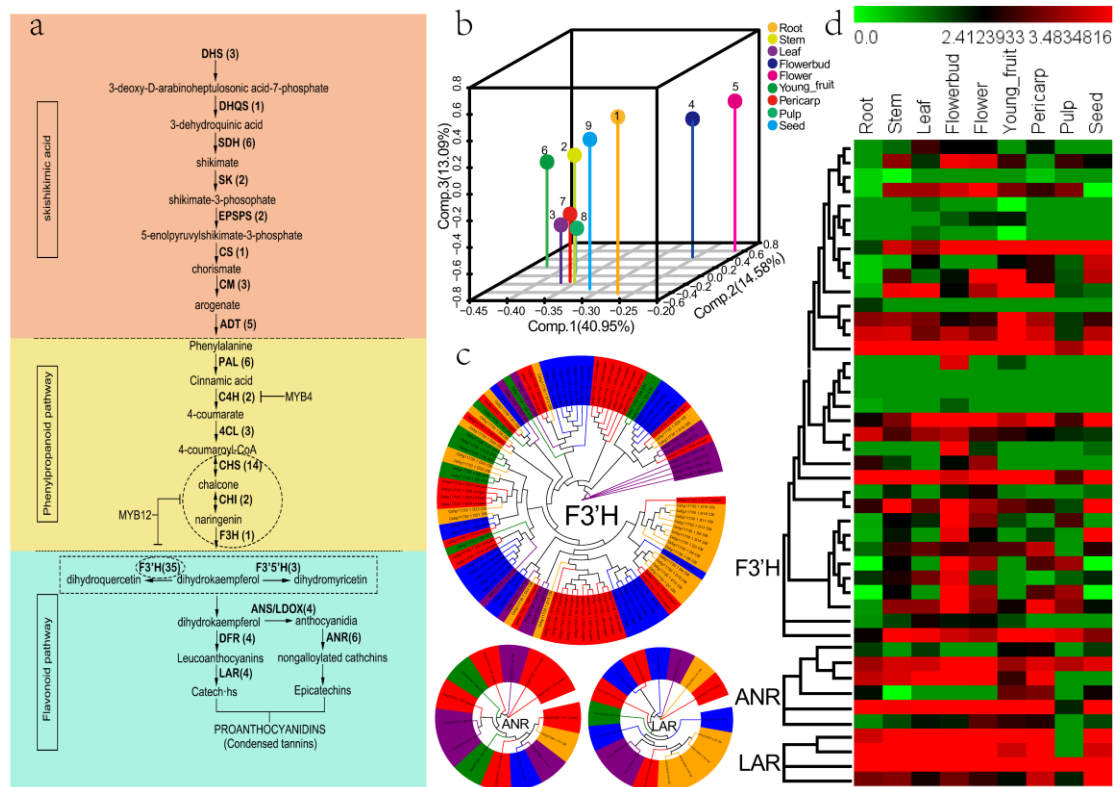
**Figure 1 Phylogenetic and evolutionary analysis of the longan genome.** (a) Molecular phylogenetic analysis based on single-copy genes shared among orange, papaya, Arabidopsis, cacao, poplar, banana, grape, soybean, apple, peach, kiwifruit, and banana from genome data. (b) Comparison of the number of gene families in eleven plant species, such as *T. cacao*, *A. thaliana*, *C. sinensis*, *C. papaya*, *P. trichocarpa*, *G. max*, *V. vinifera*, *M. acuminata*, *D. longan*, *P. persica*, *A. chinensis*, and *M. domestica*. (c) Distribution of 4DTv distance between syntenic gene pairs among banana, peach, orange, Arabidopsis and grape. (d) Distribution of gene families among *D. longan*, *C. sinensis*, *C. papaya*, *V. vinifera*, and *P. persica*. Homologous genes in longan, orange, papaya, grape, and peach were clustered to gene families. The numbers of gene families are indicated for each species and species intersection.



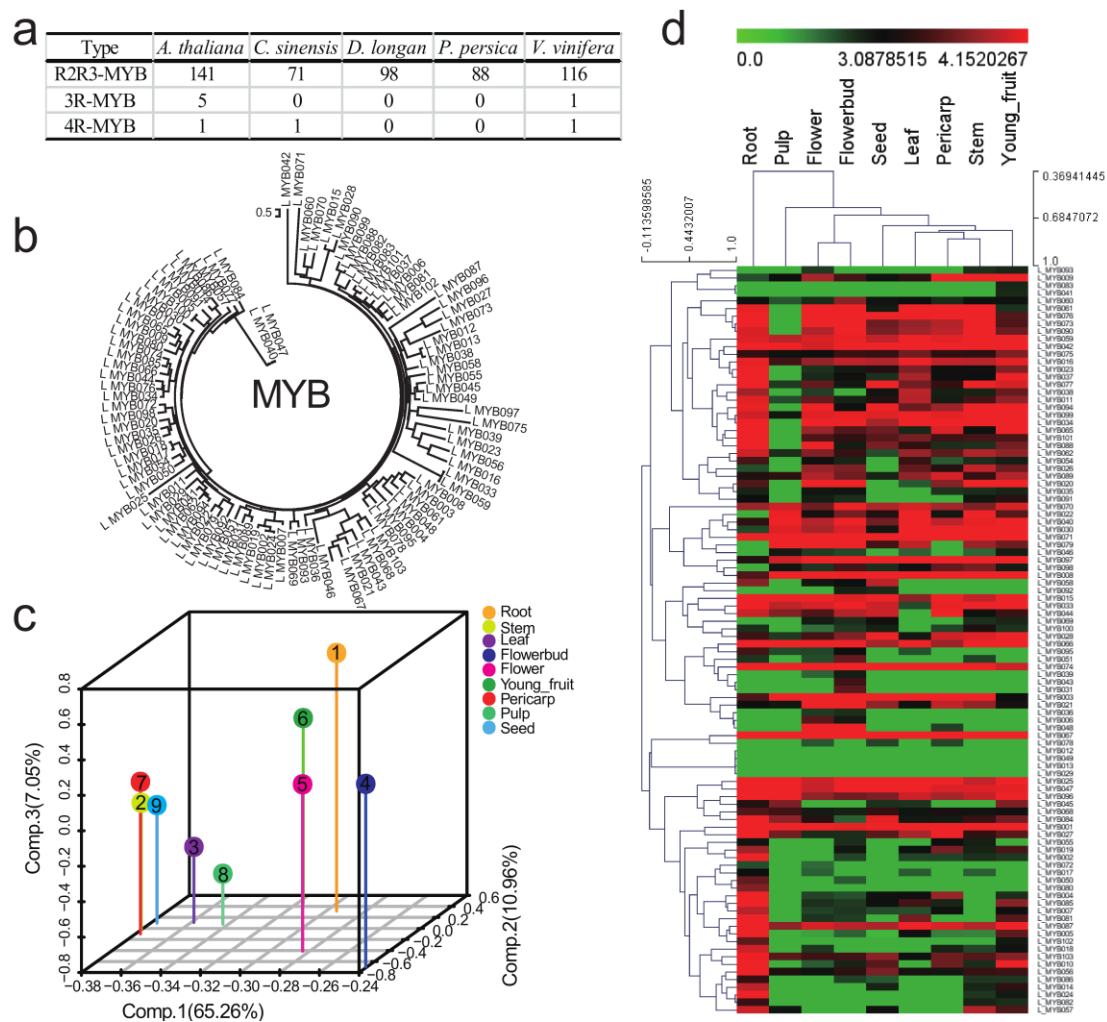
**Figure 2 Genetic diversity and population structure of longan accessions.** (a) Neighbor-joining tree of the 13 longan accessions on the basis of all SNPs. (b) PCA of the 13 longan accessions using SNPs as markers. Different colors represent for different longan accession. HHZ, DB, JYW, LDB, WLL, SN1H, YTB, SEY, JHLY, and SX, are clustered together, FY (Quanzhou, China), SJM (South-East Asia), and MQ (Thailand) showed a clear separation. (c) Population structure of longan accessions. The distribution of the accessions to different populations is indicated by different color. Each accession is represented by a vertical bar. Numbers on the x-axis show represents the K number, and the y-axis shows the different accession.



**Figure 3 Simplified diagram of polyphenols biosynthetic pathway.** (a) Simplified diagram of polyphenols biosynthetic pathway. Numbers in brackets represent genes' copy number. (b) PCA scatter plot of 9 samples using genes related to polyphenols biosynthetic pathway. (c) Neighbor-joining tree of the F3'H, ANR, and LAR from longan, peach, orange, Arabidopsis and grape. (d) Cluster analysis of expression profiles of *F3'H*, *ANR*, and *LAR*. The bar represents the scale of relative expression levels of genes, and colors indicate relative signal intensities of genes. Each column represents a sample, and each row represents a single gene.



**Figure 4 The MYB transcription factor in longan genome.** (a) Numbers of the members in the three different MYB classes in Arabidopsis, orange, longan, peach, and grape. (b) Neighbor-joining tree of the MYB gene family. (c) PCA scatter plot of 9 samples using 94 R2R3-MYB genes. (d) Cluster analysis of expression profiles of MYB transcription factor. The bar represents the scale of relative expression levels of genes, and colors indicate relative signal intensities of genes. Each column represents a sample, and each row represents a single gene.







Click here to access/download  
**Supplementary Material**  
Additional file 1-12.9.doc



Click here to access/download  
**Supplementary Material**  
Additional file 2-12.9.doc



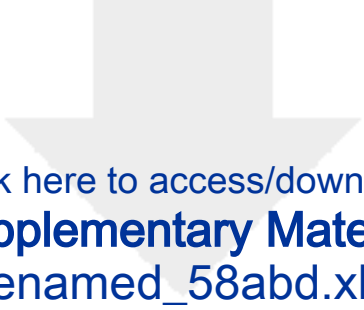


Click here to access/download

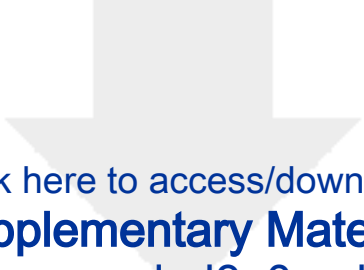
**Supplementary Material**

Supplementary EXCEL file 1- Identification of  
transcription factors in the Dimocarpus longan  
genome.xls

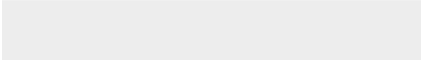





Click here to access/download  
**Supplementary Material**  
renamed\_58abd.xls



Click here to access/download  
**Supplementary Material**  
renamed\_d3c0a.xls





Click here to access/download

**Supplementary Material**

Supplementary EXCEL file 4 SNP analysis of FY, SN1H,  
MQ, LDB, and JYW cultivars.xls



Click here to access/download

**Supplementary Material**

Supplementary EXCEL file 5 SNP analysis of SX and  
YTB cultivars-12.9.xls



[Click here to access/download](#)

**Supplementary Material**

Supplementary EXCEL file 6 Statistics of copy numbers of genes involved in the biosynthesis of polyphenols in different plants.xls





[Click here to access/download](#)

**Supplementary Material**

Supplementary EXCEL file 7 Expression levels of genes  
involved in the biosynthesis of polyphenols in  
*Dimocarpus longan*.xls



Click here to access/download

**Supplementary Material**

Supplementary EXCEL file 8 MYB genes expressed in  
nine different tissues of Dimocarpus longan.xls