DR WENPING ZHANG (Orcid ID : 0000-0002-0746-0873)

PROFESSOR JISEN ZHANG (Orcid ID : 0000-0003-1041-2757)

DR RAY MING (Orcid ID : 0000-0002-9417-5789)

Article type : Original Article

Rambutan genome revealed gene networks for spine formation and aril development

Wenping Zhang[1†], Jishan Lin[1†], Jianguo Li[2†], Shaoquan Zheng[3†], Xingtan Zhang[1], Shuai Chen[1], Xiaokai Ma[1], Fei Dong[1], Haifeng Jia[1], Xiuming Xu[1], Ziqin Yang[4], Panpan Ma[1], Fang Deng[1], Ban Deng[1], Yongji Huang[1], Zhanjie Li[1], Xiaozhou Lv[5], Yaying Ma[1], Zhenyang Liao[1], Zhicong Lin[1], Jing Lin[1], Shengcheng Zhang[1], Tracie Matsumoto[6], Rui Xia[5], Jisen Zhang[1], Ray Ming [7*]

[1.] Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Key Laboratory of Genetics, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China

[2.] State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, Guangdong Litchi Engineering Research Center, South China Agricultural University, Guangzhou, 510642, China.

[3.] Fujian Fruit Breeding Engineering Technology Research Center for Longan and Loquat, Fuzhou, Fujian 350013, China

[4.] Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou, Hainan 570100, China

[5.] Tropical Crops Institute, Baoting, Hainan 572311, China

[6.] USDA-ARS, Pacific Basin Agricultural Research Center, Hilo, HI, USA.

[7] Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 6180, USA

Running title: The rambutan genome and genetics analysis

Keywords: Aril development, Chromosome fusion, Domestication, Sapindaceae, Selective sweeps.

*Correspondence, Email: rayming@illinois.edu

†These authors contributed equally to this work

Summary

Rambutan is a popular tropical fruit with exotic appearance with long flexible spines on shells, extraordinary aril growth, desirable nutrition and favorable taste. The genome of an elite rambutan cultivar Baoyan 7 was assembled into 328 Mb in 16 pseudo-chromosomes. Comparative genomics analysis between rambutan and lychee revealed that rambutan chromosomes 8 and 12 are collinear with lychee chromosome 1, resulted a chromosome fission event in rambutan (n = 16) or a fusion event in lychee (n = 15) after their divergence from a common ancestor 15.7 million years ago. Root development genes played a crucial role in spine development, such as endoplasmic reticulum pathway genes, jasmonic acid response genes, vascular bundle development genes, potassium ion transport genes. Aril development was regulated by D-class genes (*STK* and *SHP1*), plant hormone and phenylpropanoid biosynthesis genes, and sugar metabolism genes. The lower rate male sterility of hermaphroditic flowers appears to be regulated by *MYB24*. Population genomic analyses revealed genes in selective sweeps during domestication that are related to fruit morphology and environment stress response. These findings enhance our understanding of spine and aril development and provide genomic resources for rambutan improvement.

Keywords

Aril development, Chromosome fusion, Domestication, *Nephelium lappaceum* L., Selective sweeps.

Significance Statement

Rambutan genome revealed a chromosomal fission or fusion event causing karyotype variation between rambutan and lychee. Root development genes played a crucial role in spine development, including endoplasmic reticulum pathway, jasmonic acid response, vascular bundle development, and potassium ion transport genes.

Introduction

Rambutan (*Nephelium lappaceum* L.) is a popular tropical fruit for its refreshing flavor and exotic appearance (Almeyada *et al.*, 1979). It is native to tropical low land rain forest of Indonesia and Malaysia, and widely cultivated throughout archipelago and Southeast Asia, Hawaii, Central America, Africa and China (Huang *et al.*, 2005; Nakasone and Paull, 1998; Zee *et al.*, 1998). Between 2015 and 2017, average Asian production of rambutan was 1.4 million metric tons, and the largest producer was Indonesia (692.0 thousand metric tons (TMT), 49.9% of total production), followed by Thailand (344.8 TMT; 24.9%), Vietnam (261.4 TMT; 18.8%), Malaysia (62.7 TMT; 4.5%), and the Philippines (7.70 TMT;0.6%) (Altendorf, 2018).

Rambutan belongs to the family Sapindaceae, which includes two other tropical fruits: longan (*Dimocarpus longan* Lour., 2n = 30) and lychee (*Litchi chinensis* Sonn., 2n = 30). A notable difference separating rambutan from longan and lychee is their sex types. Rambutan is trioecious, and can be classified into three groups: male trees producing only staminate flowers, hermaphroditic trees functioning as female (HF), and hermaphroditic trees producing both hermaphrodite and female flowers (HF) or hermaphrodite and male flowers (HM) (the most common type in cultivar selections) (Muhamed *et al.*, 2019; Valmayor *et al.*, 1970). However, others thought male trees belong to dioecious, hermaphrodite flower trees belong to andro- or gyno-monoecious, and hermaphroditic flowers have viable pollen with low rate of anthers dehiscence (5%) (Manuel *et al.*, 2014; Windarsih and Efendi, 2019). The fruit morphological characteristics of rambutan is similar with lychee, having oval to spherical drupe, white and fleshy aril, leathery skin, except one distinguishable trait: long and flexible hairy spines on rambutan fruit shell (VanBuren *et al.*, 2011). The length of spines, thickness of fruit rind, thickness of aril, taste of aril, and easy or hard to peel off aril from seed are a serials of major indicators for evaluating rambutan fruit characteristics (Windarsih and Efendi, 2019).

The spines are derived from trichomes on the external epidermis of the ovary (Moncur, 1988). Anatomical structure showed the spines with anomocytic stomata, and contains well-developed vascular tissues, which extended along the spines (Arévalo-Galarza *et al.*, 2018). In *Arabidopsis*, more than 17 genes were involved in trichomes initiation, and plant hormones gibberellins (GA), jasmonates (JA), and cytokinins (CK) regulated trichomes development (An *et al.*, 2011; Chopra *et al.*, 2019).

Rambutan fruit flesh is aril, which is delicious, nutritious, and rich in carbohydrates, vitamin C, calcium, and other minerals. The fruit is usually consumed fresh or preserved in syrupy canned. Comparing with lychee and longan, the seed coat of rambutan was rough and not easily peeled off from aril. Aril has an important role in the formation of reproductive structures in both gymnosperm and angiosperm, but the origin and genesis are largely unknown (Silveira *et al.*, 2016). Research of aril development regulate genes were scarce in Sapindoideae. In lychee and longan, the fleshy arils develop from ovules stalk (Huang *et al.*, 1983; Ke *et al.*, 1992), and rambutan has similar development pattern. A *LcHMGR1* (*3-Hydroxy-3-methylglutaryl coenzyme A reductase*) gene regulated lychee fruit size through the regulation of early cell division (Xia *et al.*, 2012). In Ginkgo and Taxus, transcription factors of the ABCE model of flower development genes involved in regulate aril development, including B class genes (*APETALA3 and PISTILLATA*), C class gene (*AGAMOUS*), D class genes (*SEEDSTICK* and *SHATTERPROOF*) and E class genes (*AGL6* and *SEPALLATA*) involved in regulating the formation of fleshy structures (Lovisetto *et al.*, 2015; Lovisetto *et al.*, 2012).

Beside the sweet and refreshingly flavorful fruits, most organs of rambutan trees have economic value. Polyphenols are abundantly contained in leaves, barks, peels and seeds and have potential applications in food and health-related products owing to their wide range of biological activities (Hajimehdipoor *et al.*, 2014; Rakariyatham *et al.*, 2020; Thitilertdecha *et al.*, 2010). The abundant phenolics compounds are secondary metabolites and affect various parts of a plant, such as synthesis of photosynthetic pigments (Tanase *et al.*, 2019), response to stress (Cheynier *et al.*, 2013), and improve the tolerance and adaptability of plants under suboptimal conditions (Dixon and Paiva, 1995; Lattanzio *et al.*, 2009). Plant phenolics are synthesized through a shikimate/phenylpropanoid pathway. In peels and seeds of rambutan fruits, the dominant phenolic is geraniin, which exhibited significant therapeutic activity to metabolic dysfunction induced by safely mitigating obesity (Thitilertdecha *et al.*, 2010), and antiviral activity effective against dengue virus type-2 (DENV-2) (Sukmandari *et al.*, 2017). Rambutan has been used as traditional medicine for centuries (Burkill and I.H, 1935; Marry and Josephine , 2016). The seeds also contain a high amount of fat, which is similar to cocoa butter and can be used as edible oil or making soap and candles (Issara *et al.*, 2014).

Despite being widely cultivated in suitable climates in tropics for its edible fruit, as tradition medicine plants, and also as an ornamental crop, the suitable planting area and yield of rambutan are still smaller than those of lychee and longan. We sequenced and assembled the genome of Baoyan 7, which is an important cultivar from Hainan in China, and generated transcriptome data from leaves, flowers, and different fruit tissues in 4 developmental stages. The genome and transcriptome analyses led to the discovery of regulatory genes affecting species-specific spines, aril development in fruits, geraniin biosynthesis pathway, sex determined of male and hermaphroditic flowers. Population genomic analyses revealed polymorphic regions among different cultivars and selective sweeps from domestication.

Results

Genome sequencing, assembly, and annotation

The rambutan genome was assembled into 3,884 contigs with a total 467 Mb and an N50 at 326 kb (**Table S1**). However, the genomic size of rambutan was estimated as a range of 324-343Mb by flow cytometry (VanBuren *et al.*, 2011). We removed redundant sequences and retained 330 Mb non-redundant contigs. We have counted the chromosome number of rambutan and it has 2n = 32 chromosomes, different from 2n = 30 in lychee and longan (**Figure S1**). Hi-C-based physical map was used to anchor 16 pseudo-chromosomes with a total length of 328Mb **(Figure 1)**. The range of single chromosome length was from 16.8 to 27.7 Mb **(Table S2)**. The genome completeness was estimated to be 96.2% (**Table S3**).

Genome annotation resulted in 26,500 genes with the BUSCO completeness 90.9% and duplication 4.4%, and these genes were evenly distributed across 16 chromosomes **(Figure 1).** Repetitive sequences were analyzed and accounted for 41.2% of the genome assembly, totally 136 Mb, and retrotransposons accounted for 69.0% of repetitive sequences and 28.3% of the assembled genome (**Table S4)**. Intact LTR retrotransposons were identified and their insert time ranged from 0.01 to 3.77 MYA (**Figure S2 (a) and (b)**). The genome contains 149,997 miniature inverted-repeat transposable elements (MITEs), and they were classified into 42 families, totaling 23.8 Mb (7.3%) of the genome.

Genome evolution

Rambutan had no additional whole genome duplication event after the triplication event shared among eudicots (**Figure S3**). To investigate the evolutionary history of rambutan genome, we performed a gene family clustering using 8 representative angiosperm species. From the results, rambutan was at the same branch with lychee. We estimated that rambutan diverged from the last common ancestor with lychee about 15.7 million years ago (MYA) **(Figure 2)**. The numbers of gene families showing expansions and contractions were 131 and 684, respectively **(Figure 2)**.

The rambutan genome contained 726 species-specific genes comparing with the other 8 species, and these genes randomly distributed into 16 chromosomes (**Figure S4 (a)**). Among them, 167 genes had expression at 9 tissues, 46 genes with expression in vegetative growth tissue (young shoot, young leaf, and mature leaf), 24 in flower tissue (hermaphrodite flower and male flower), 23 in fruit tissues (**Figure S4 (b)**). In addition, 12 genes displayed tissue-specific expression (**Figure S4 (c)**).

Comparative analysis of karyotype evolution between rambutan and lychee

Rambutan and lychee belong to different genera in the family Sapindaceae, and they have similar morphology in tree architecture and fruit morphology. But they have different chromosome numbers, 16 for rambutan and 15 for lychee. Whole genome collinearity analysis revealed a chromosome fusion event, and chromosomes 8 and 12 in rambutan shared collinearity with chromosome 1 (Chr1) of lychee (**Figure 3**). 0.41 Mb of chromosome 12 shared collinear with lychee chromosome 7. Synteny analysis between rambutan and lychee revealed chromosomal rearrangements in other 13 collinear chromosomes (**Figure S5**). In details, 8 chromosomes of rambutan had translocations, 10 chromosomes had inversions, 9 chromosomes had deletions (more than 1 Mb) (**Table S5**). Among them, chromosome 1 of rambutan had two large inversions and one small inversion and translocation each at the two ends of chromosome, their length was 3.27, 2.74, 0.6, and 0.08 Mb, respectively; chromosome 2 had one 4.7Mb large inversion and two 0.74 and 0.28Mb translocations at one terminal of chromosome; chromosome 3 had two 1.51 and 0.41 Mb large inversions and one 0.48 Mb translocations at one end; chromosome 5 had three 5.07 Mb, 4.28 Mb, 4.49 Mb large inversions and 3.46 Mb large translocation in the middle of chromosomes. Chromosome 4 had a 5.04 Mb deletion, which was the largest deletion one.

Sequence rearrangements occurred around the fusion site of chromosome 1 in lychee compared with rambutan. Two regions of rambutan chromosomes 8 and 12, which were about 0.37 Mb and 2.27 Mb, respectively, did not have collinear sequence in lychee close to the fusion site. These two regions of rambutan contained 137 genes. However, only 6 of 137 were rambutan specific genes, and orthologs of the other genes were distribute in different chromosomes in the lychee genome. No expression of these genes was detected in leaves, flowers, and fruit tissues.

The rinds of rambutan participate in photosynthesis and adaptation to environments

To investigate the difference of rinds between rambutan and lychee, differentially expression genes (DEGs) were analyzed among five developmental stages of fruit rinds in rambutan (including spines and pericarps) and lychee. There were 2,897 and 2,823 DEGs in rambutan and lychee, respectively. Among these DEGs, 1130 DEGs of rambutan found homologous with 994 DEGs of lychee. The top 263 of 1130 genes were highly expressed with the FPKM more than 200 in rinds, and these genes were enriched in cell periphery (39), extracellular region (21), cell wall or external encapsulating structure (14), thylakoid membrane or photosynthetic membrane (7). They can be categorized into transcription factor activity or nucleic acid binding (25), and response to stimulus (69), such as, wounding, water, cold and chemical (**Figure S6**).

We further analyzed 1767 rambutan specific DEGs, 169 of them showed higher expression in spines than that in other fruit tissues (**Figure S7 (a)**). They were mainly enriched in plastid, such as chloroplast thylakoid membrane, photosynthetic membrane, and involved in photosynthesis, light response, sucrose catabolic and metabolic process (**Figure S7 (b) – (d)**). There were 29 chloroplast genes, and 3 sucrose catabolic and metabolic genes, including sugar transporter *SWEET17*, *GT* and *UGT74E2* encoding UDP-glycosyltransferase*,* which showed high expression in spines.

Candidate genes for spine development

Among the 14 homologous genes identified for trichomes initiation, 13 of them showed down-regulated with spine development, but *JAZ1* was up-regulated during spine development (**Figure S8**). To further investigate spine development genes, 1,774 spine-specific DEGs were identified. They were classified into 6 clusters according to the expression trend at four spine developmental stages, 283 of them in the up-regulated cluster (including *JAZ1*) (subcluster3) (**Figure S9, Table S6**). GO enrichment analysis revealed the 283 up-regulated genes mainly involved in developmental process (38), cellular potassium ion transport (4) cellular carbohydrate biosynthetic process (5), response to endoplasmic reticulum stress (3), and cell wall pectin biosynthetic process (2) (**Figure S10**). Developmental processes included anatomical structure development, post-embryonic morphogenesis, gametophyte development, lateral root development, tissue development, and trichome morphogenesis. The cluster of up-regulated genes of the highest expression included 25 genes, containing four development genes two cellular potassium ion transporters. The four development genes, were *HSC70-1* (heat shock cognate 70 kDa, *Nl02g24320*), *BIP1* (luminal-binding protein, *Nl07g01780*), *BIP2* (*Nl01g15350*), *LOX3* (*linoleate 13 S-lipoxygenase 3-1*, *Nl03g11310*). *HSC70-1*, showed the highest expression level than other upregulated genes in spines. BIP1 and BIP2 are molecular chaperone of the heat shock protein 70, and they are involved in embryo sac central cell differentiation. HSC70-1, BIP1, and BIP2 are also involved in the process of endoplasmic reticulum. TIFY family member *JAZ1* (*Nl03g13930*) and *JAZ8* (*Nl16g09270*) showed high expression along with spine development. *JAZ1* responds to jasmonic acid and inhibits trichome development. Another jasmonic acid response gene *AOS* (*allene oxide synthase*, *Nl05g05670*) was upregulated during spine development. A WRKY gene—*WRKY23* in rambutan showed up-regulated along with spines development, and different from trichome initiation gene *TTG2*. The spines contain abundant vascular bundle, the genes involved in vascular bundle development and lignin biosynthesis likely regulate spine development, such as *SCL1* and *MOT3*, which showed a significant upregulation along with spine development. *SCL1* is a member of SCARECROW-like transcription factors, responding to auxin. Other members likes *SHORT-ROOT* (*SHR*), *SCARECROW* (*SCR*) and *SCARECROW-LIKE 23* (*SCL23*) affect bundle sheath cell in leaf, and regulated the endodermis development in both roots and shoots of *Arabidopsis thaliana* (Cui *et al.*, 2014; Yoon *et al.*, 2016). MOT3, a S-adenosylmethionine synthase, was a methyl donor for several reactions in lignin biosynthesis (Shen *et al.*, 2010). The two genes of cellular potassium ion transport were tandem in chromosome 7, and are homologous with *AtHAK5*, which was the main K^+ transport proteins of root in the *Arabidopsis* (Nieves-Cordones *et al.*, 2010). ARF related gene *GNOM* also exhibited an increase trend along with spine development, and it is required for the establishment of the auxin response maximum for lateral root initiation in *Arabidopsis* (Okumura *et al.*, 2013) (**Figure 4**).

We analyzed the top 30 hub genes (top 10%) interaction with other genes, and the top 3 genes were *FT1*, *HoP3* and *MNS1*. Ten genes involved in eight KEGG pathways, including protein processing in endoplasmic reticulum (*MNS3*, *BIP1* and *Nl02g09980*), carbon metabolism (*G6PD2*), fatty acid metabolism (*FAD2*), glycosphingolipid biosynthesis - lacto and neolacto series (*Nl16g13700*), glycan biosynthesis and metabolism (*FUT13*), transcription (*RAP74*), amino acid metabolism (*Nl11g01810*), carbohydrate metabolism (*Mik*) (**Figure S11**).

POR genes of the ellagitannin biosynthesis pathway expanded in rambutan

The biosynthetic pathways of ellagitannin and shikimate acid share the first several steps, and we identified homologous genes in rambutan, including 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase (DHS) (2), 3-dehydroquinate synthase (DHQS) (1), bifunctional 3-dehydroquinate dehydratase/shikimate dehydrogenase (SDH) (3), UDP-glucose:gallate glucosyltransferase (UGGT) (1), and pentagalloylglucose oxygen oxidoreductase (POR) (48) (**Figure 5**). The *POR* genes are substantially expanded to 48 comparing with the 11 genes in papaya (Ming *et al.*, 2008), 20 in sweet orange (Xu *et al.*, 2013) and 34 in pomegranate (Qin *et al.*,

2017) (**Figure 5 (b), Figure S12**).

The expression level of all 56 genes in ellagitannin biosynthesis pathway were analyzed in four fruit tissues (spines, rinds, arils, and seeds) at four developmental stages (**Figure 5 (a)**). The genes in the first step (*DHS1*, *DHS2*), the second step (*DHQS1*), and the third step (*SDH1, SDH2*, and *SDH3*) of ellagitannin biosynthetic pathway showed significantly higher expression in spines or pericarps (**Figure 5 (c)**). POR works in the finally step, is a laccase-type phenol oxidase catalyzed pentagalloylglucose to yield to ellimagrandin II (Niemetz and Gross, 2003). They showed diverse expression in different fruit tissues, 8 of them in spines, 7 of them in pericarps and 7 in seeds（The purple box in the Figure 5 (c)）.

Transcription factors and plant hormone related genes regulating arils development

Weighted correlation network analysis (WGCNA) found four modules associated with aril development at different stages (**Figure 6 (a)**). The MEdarkgreen module contained 476 genes, including 75 transcription factors which mainly related to the development at both A1+S1 (Aril1 and Seed1) and A2+S2 (Aril2 and Seed2) stages. The MEdarkturquoise module contains 257 genes, including 24 transcription factors, mainly related with Stage2 and also related with Stage 1 development. The MEpaleturquosie module contains 686 genes, including 76 genes transcription factors, mainly related with Aril 3, also related with Aril1 and Aril2 development, and the correlation and expression taken an increased trend along with aril development; the MEyellow module contains 632 genes including 53 transcription factors, mainly related with Aril4. We found the genes of MEpaleturquosie module regulate aril development, which showed an increase during aril development and declined at Aril4 (fleshy stage). The number of *B3, Dof, MIKC_MADS, NAC, bZIP, LBD, ARR-B, FAR1, NF-YA, Nin-like, SRS, WOX* gene family in the up-regulated module was higher than other modules, and *ARR-B, FAR1, NF-YA, Nin-like, SRS, WOX* was special expression at up-regulated module (**Table S7**).

Four hundred and one genes in the up-regulated module (MEpaleturquosie) showed a higher expression in aril than other fruit tissues (more than two times), and 22 of them were aril specific expression genes. These 401 genes mainly located in cell periphery, cell wall, external region, external encapsulating structure, had nucleic acid binding transcription factor activity, regulated nucleic acid-template transcription by GO enrichment analysis. They mainly involved in metabolic pathway (such as phenylpropanoid biosynthesis, glycosaminoglycan degradation) and signal transduction (plant hormone transduction and MAPK signaling pathway – plant) by KEGG enrichment.

There were 52 transcriptions factors highly expressed in aril. Two MIKC_MADS family member *STK* and *SHP1* were detected, and they belong to the floral D class genes, which regulate ovule development. *STK* showed the highest expression level in aril than other transcriptions factors. *BZIP44* and *ERF-1* also showed higher expression than others. Seven genes involved in phenylpropanoid biosynthesis pathway, including three genes (*Nl10g09570*, *Nl15g09330*, and *Nl08g10880*) encoding peroxidase, *BGLU40* (*Nl14g07850*), one gene (*Nl15g08780*) encoding oxoglutarate/iron-dependent dioxygenase, *SAM* (*Nl08g01160*), and *Nl15g08790*. Beta-glucosidase (BGLU) hydrolyses cellulose-based oligosaccharides into monomeric sugars. *BGLU40* also involved in starch and sucrose metabolism pathway. The aril stores abundant glucose and sucrose, and the genes of monomeric sugar metabolism were also detected, including two homologous genes of *GUS3* (*Nl08g04910* and *Nl08g07570*) involved in glycosaminoglycan degradation and sugar transporters *SWEET11* (*Nl02g14710*) and *SWEET12* (*Nl09g06780*), which showed specific high expression in aril.

Auxin and GA involved in fruit initiation and development, and auxin promotes cell division and GA functions in later cell expansion (Gorguet *et al.*, 2005; Serrani *et al.*, 2007). Auxin and GA related genes also found in regulating rambutan aril development, such as *GH3.1* (*Nl13g08960*), *GH3.9* (*Nl02g14290*), *ARF1* (*Nl04g02640*, *Nl04g02670*, and *Nl04g04550*), and *GA20ox2*. We also found other types of plant hormone related genes involved in aril development, including Abscisic acid (*PYL4* and *PYR1*), Ethylene (*EIN4*, *ERF-1*, *Nl08g16100*), and Cytokinine (*ARR11*). All above genes showed co-expression in aril development (**Figure 6 (b)**). The top 15 hub genes of the co-expression network including seven transcriptions factors, Auxin related genes *GH3.1* and *ARF1*, glycosaminoglycan degradation pathway gene *GUS3-1* (*Nl08g04910*), Starch and sucrose metabolism gene *Nl16g13620*. We selected nine genes from aril candidate genes to verify the expression trend by RT-qPCR, and they showed lower or not expression in stage 4 (aril4), which were consistent with RNAseq data (Figure S13a and b).

Lipid biosynthetic and metabolic related genes showed high expression in seed

We analyzed two developmental stages of seeds and found 797 seed unique DEGs (**Figure S14**). Both top 20 GO and KEGG enrichment including lipid metabolic (27, $p<0.003$) and biosynthetic process (17, $p<0.006$), and fatty acid biosynthetic process related genes (7, $p<0.009$). There were totally 27 genes, 17 are involved in lipid metabolic and biosynthetic process, and 7 are involved in three processes. And 19 of them expressed higher in seeds than other fruit tissues (**Figure S15**). *AGL15*, *GA3ox*, *GA5*, *GA20ox*, and *BAT1* showed special high expression in rambutan seeds. *AGL15* and GA related genes *GA3ox* and *GA20ox* have been reported regulating seed development in *Arabidopsis* and soybean (Heck *et al.*, 1995; Kang *et al.*, 1999; Perry *et al.*, 1996; Thakare *et al.*, 2008).

Candidate genes for sex differentiation

To investigate sex determination mechanism in rambutan, the RNAseq data of male and hermaphrodite flowers were analyzed. There are 173 differentially expression genes between male and hermaphrodite flowers. Among them 65 genes showed high expression in male flower, and 108 genes showed high expression in hermaphrodite flowers. Among them, there are nine fertility genes, including two female (*ANT* and *CUC2*) and seven male (*LAP6*, *FAR2*, *MYB24*, *MYB103*, *ABCG15*, *QRT1*, and *GPAT1*) fertility genes (**Figure S16**). Expression analysis showed five male fertility genes (*LAP6*, *FAR2*, *ABCG15*, *QRT1*, and *GPAT1*) had high expression in hermaphroditic flowers. Male fertility related gene *MYB103* and *MYB24* showed higher expression level in male flowers than in hermaphrodite flowers. Female fertility gene *ANT* showed higher expression in hermaphrodite flowers. The expression profiles of these genes were verified by RT-qPCR (**Figure S17 a and b**). In addition, seven transcription factors from five family showed higher expression in male flowers than in hermaphroditic flowers, including two of *NACs*, two of *ERFs*, one of *bHLH*, one of *WRKY*, and one of *HD-zip*.

Population and genetic diversity of rambutan germplasm

Twenty-seven accessions of rambutan were selected for whole-genome resequencing, including 18 accessions from USDA germplasm resource, 8 cultivars and 1 male tree collected from Hainan, China. The 27 accessions were splits into three distinct groups in phylogenetic tree, group1 with three accessions, group 2 with seventeen accessions, and group3 with seven accessions (Admixture value with K=2) (**Figure 7**), and the principal component analysis (PCA) generated the same three groups.

Pairwise population differentiation of population (F_{ST}) was calculated in pairs of the three groups, and coefficient of differentiation of group 1 and group 3 ($F_{\mathrm{ST}} = 0.40$) was significantly higher than the other two pairs ($F_{\mathrm{ST}} = 0.11$ and $F_{\mathrm{ST}} = 0.10$) (**Figure S18**). To reveal diversity of the population and find the positive selective sweeps, standard population genetic summary statistics (π and Tajima's D) were calculated based on non-overlapping with 50 kb sliding windows across the whole genome.

The lowest 1% of nucleotide diversity (π) regions were about 3.19M and contained 171 genes (**Table S8, S9**). KEGG enrichment analysis revealed these genes mainly involved in plant hormone signal transduction, plant-pathogen interaction, fatty acid biosynthesis and metabolism, protein processing in endoplasmic reticulum, photosynthesis. Based on the analysis above, these genes were involved in spine, aril, and seed development, indicating these selected genes related to fruit morphology and quality (**Figure S19 (a)**). We found that 66 of 171 genes had expression in 9 organs (**Figure S19 (b)**). Nine genes in the highest expression level cluster, including 2 members of prolyl oligopeptidase family (*Nl05g02300* and *Nl05g02320*), 2 homologous genes of *MLN51* (*Nl03g09270* and *Nl03g09300*), *RGLG2* (*Nl07g07940*), chloroplast gene *sufE* (*Nl01g17160*). The prolyl oligopeptidase family members and *RGLG2* respond to abiotic stress (Cheng *et al.*, 2012; Tan *et al.*, 2013). And a gene (*Nl09g01410*) encoding maintenance of PSII under high light 1 (MPH1) showed significantly higher expression in mature leaves (ML) (about 10 times than young leaf and shoot, less or no express in other tissues).

Tajima's D ranged from -1.62 to – 4.67. The lowest 1% value of Tajima's D contained

3.29Mb sequences and 141 genes (**Table S10, S11**). These genes were mainly enriched in plant-pathogen, endocytosis, pentose and glucoronate interconversion, phosphatidylinositol signal system, inositol phosphate metabolism pathway (**Figure S20 (a)**). Transcriptome sequencing analysis identified 12 genes from the highest expression cluster, such as 3 rambutan species-specific genes, 2 chloroplast genes (*sufE-like 1*: *Nl01g17160*, *reticulate-related 1*: *Nl04g00840*), *EMBRYO DEFECTIVE 3006* (*EMB3006*), *UMAMIT2*, *nep-1* (*Nl04g00690*), a gene of auxin-responsive family (*Nl09g16950*). (**Figure S20 (b)**). KEGG and annotation analysis revealed these genes were involved in plant photosynthesis and development, such as, *EMB3006* was related with embryo development in *Arabidopsis* (*AT4G19350*), WAT1-related protein related with bast fiber development in ramie (Liu *et al.*, 2018). Auxin participate in adventitious roots development (Gutierrez *et al.*, 2009).

Discussion

Although rambutan has one more chromosome than lychee, the genome size of rambutan is smaller than lychee (average 554Mb) due to greater content of repetitive sequences in lychee (41.2% vs. 56.5%) (VanBuren et al., 2011). Comparative genomic analysis revealed chromosome rearrangements between these two genomes. There was either a chromosome fusion event involving chromosomes 8 and 12 in rambutan or a chromosome fission event involving chromosome 1 in lychee, resulting 16 chromosomes in rambutan and 15 in lychee.

Spines and pericarp of rambutan were different from those in lychee and longan. We analyzed the rinds (including spines and pericarps in rambutan) development genes and detected species-specific DEGs in rambutan. The types of genes showed high expression in spine and pericarp mainly located in plastid, such as chloroplast genes (29) and sucrose catabolic and metabolic genes (3). These results indicate that the spines and pericarps of rambutan participate in photosynthesis beside to protecting aril and provide nutrients for growth of spines and aril. Spines derived from trichomes on the external epidermis of the ovary (Moncur, 1988). Our results found that only *JAZ1*, a gene inhibit trichomes initiation, showed upregulation during spine development, likely to limit the number of trichomes to ensure the development of existing spines, because the diameters of spines in rambutan are so large, at 1 – 3 mm, about 30 – 100 times of regular cell size, indicating that one spine spans 30 – 100 epidermal cells, whereas the length ranged from 1.0 cm to 2.0 cm. *JAZ1* is a signal transduction gene of jasmonic acid (JA), and JA is always related to stress response from environment, indicating that the spine development is related to response to external environment. The genes controlling trichome development might work in spine's initiation but not in further development. The candidate genes of spine growth include 38 development genes, such as, post-embryonic morphogenesis and lateral root development.

Anatomical picture displays the vascular bundle fills the spines (Arévalo-Galarza *et al.*, 2018). *SCL1* showed higher expression along with spines development and it is from a gene family that participate in bundle sheath cell development in leaf, roots and shoots of *Arabidopsis thaliana* (Cui *et al*., 2014; Yoon *et al*., 2016). In addition, lignin biosynthesis gene *MOT3*, the main K^+ transport proteins of root *AtHAK5*, and ARF related gene were also highly expressed along with spine development.

The rinds and seeds of rambutan contains 50 compounds of phenolics, and the most important one is geraniin. We analyzed the ellagitannin biosynthesis pathway genes, including nine main enzymes, and found five genes (*DHS*, *DHQS*, *SDH*, *UGGT*, and *POR*) through homologous comparison, while found no orthologs of the other four genes (*GLUG*, *1,6GALT*, *1,2,6GALT*, *1,2,3,6GALT*) in rambutan. The first three steps include genes *DHS*, *DHQS*, and *SDH* of the biosynthesis pathway that showed significantly higher expression in spine or pericarp. They were also key enzymes of shikimic acid biosynthesis pathway that produces anthocyanins, which were abundant in spine and pericarp. The POR genes family showed expansion than in other fruit crops, and the genes were differential expression in different tissues, indicating the crucial role of them in geraniin contents in fruit tissues of rambutan.

The ABCE model genes of flower development are involved in regulating formation of fleshy structures of ginkgo and taxus (Lovisetto *et al.*, 2015; Lovisetto *et al.*, 2012). The D-class genes *STK* and *SHP1* showed crucial role in aril development, likely because of the aril development from ovule stalk (Huang *et al.*, 1983; Ke *et al.*, 1992). *SHP1* as a hub gene interacted with many genes and transcription factors. Beside MIKC-MADS genes, there were many transcription factors involved in aril development, and different transcription factors worked in different aril development stages. Five types of plant hormones interact with transcription factors regulating aril development, including Auxin, Abscisic acid, Ethylene, Cytokinine, and Gibberellin.

We analyzed the expression profile of male and hermaphroditic flowers and anther dehisced gene *NlMYB24* showed significantly higher expression in male than in hermaphroditic flowers, while other genes regulating male development also expressed in hermaphroditic flowers. In *Arabidopsis*, *MYB24* regulates pollen maturation and dehiscence (Mandaokar and Browse, 2009). Female fertility gene *ANT* showed higher expression in rambutan hermaphrodite flowers, and it promotes integument formation and plays a critical role in regulating ovule and female gametophyte development in *Arabidopsis* (Klucher *et al.*, 1996). These results proved that hermaphroditic flowers with normal pollen development and the lower rate of set fruit may be related with anthers dehiscence. Although previous reports documented hermaphroditic flowers with only 5% anthers dehiscence (Manuel *et al.*, 2014; Windarsih and Efendi, 2019), modern cultivars always only plant one or no male tree in an orchard, indicating modern cultivars suffer chosen in male fertility genes. Population genomic analysis found two homologous genes of *GEX1* in selected regions, which regulated male and female gametophytes development (Alandete-Saez *et al.*, 2011), suggesting improved fertility of cultivars through breeding.

The F_{ST} range of rambutan population was 0.1~0.4, and it was widely ranged in other fruits as well, such as pomegranate (0.26) (Qin et al., 2017), mango (0.1358 ~ 0.1856) (Wang *et al.*, 2020), and apple (0.14 ~ 0.21) (Velasco *et al.*, 2010). It may relate to differences in population size and composition. The average nucleotide diversity of rambutan was 0.005, and it was similar with mango and apple (Table S12). Analysis of these genes in the lowest 1% nucleotide diversity genes of cultivars revealed the types of genes related to fruit tissues development, photosynthesis, response to stress and the pathway genes of plant hormone signal transduction and plant-pathogen interaction through unconscious selected. Especially, the homologous gene of maintenance of PSII under high light 1 (MPH1) showed significantly higher expression in mature leaves. These results were in line with high temperature growth environment of rambutan and improved fruit quality during domestication.

EXPERIMENTAL PROCEDURES

Genome assembly

A total of 50Gb long reads (~150X) were generated from Pacific Biosciences (PacBio) Sequel II Platform and de novo assembled using CANU (v.1.8) (Koren *et al.*, 2017), and corrected using the pair-end short reads from HiSeq2500 by Pilon (v.1.24) (Walker *et al.*, 2014). Redundancy contigs were purged using Redundans (Pryszcz and Gabaldon, 2016) with a series of parameters (identify 0.9 and coverage 0.8, identify 0.9 and coverage 0.6, identify 0.9 and coverage 0.56, identify 0.9 and coverage 0.5), and each result was submitted to Universal Single-Copy Orthology (BUSCU) (v.3.0.2) to evaluated the complete, and the verison with identify 0.9 and coverage 0.56 was optimal (Simao *et al.*, 2015). Softwares, scripts, and some pipelines are listed in https://github.com/lvvn/genome-analysis. After that, ALLHiC was used to anchored purged contigs into super-scaffolds with the Hi-C library (Zhang *et al.*, 2019). Finally, the genome assembly contains 16 pseudo chromosomal molecules, and 8 unplaced scaffolds were further evaluated the complete using BUSCU.

Genome annotation

The *de novo* repeat library of the genome was customized using RepeatModeler (v.1.0.8) (http://www.repeatmasker.org/RepeatModeler/), then retrotransposons were identified by LTR_FINDER (v.1.0.6) (Xu and Wang, 2007), LTRharvest (Ellinghaus *et al.*, 2008), and LTR_retriever (Ou and Jiang, 2018), then they were masked by Repeatmask (v.4.0.5) (http://www.repeatmasker.org/). TEclass (v.2.1.3) and Tandem Repeat Finder (TRF) package (v.4.09) were used to further classified and identify tandem repeats within the genome (Abrusán *et al.*, 2009) (Benson, 1999). In order to calculate LTR inserted time, the intact LTR retrotransposons were further identified by LTRdigest according to fourth domains (Steinbiss *et al.*, 2009). The two ends of intact LTR retrotransposons were aligned with MUSCLE (v.3.8.31) (Edgar 2004) and the distance was calculated using the distmat program in the EMBOSS package (Rice *et al.*, 2000) (v.6.5.7) with parameter of '-nucmethod 2', and the insertion time was estimated as T=K/2*r* (K is the divergence rate and *r* is the neutral mutation rate, default is 1.38 × 10^{-8}). Miniature inverted-repeat transposable elements (MITEs) were estimated by MITE Hunter (Han and Wessler, 2010).

All RNA-seq data were imported into Trinity (v.2.8.6) *de novo* assembly and genome-guided assembly pipelines with default parameters (Haas *et al.*, 2013) . Meanwhile, retro-transposon (RT) genes were identified by HMMER (v.3.1b2) (Eddy, 1998) with the Pfam data-base and removed from the final annotation. RSEM was used to calculate transcript abundance (Li and Dewey, 2011). Then *ab initio* gene predictors by GENEMARK (Lomsadze *et al.*, 2005), and a series of homolog proteins were used to comparison. After that, MAKER pipeline was used to integrate multiple tiers of coding evidence, including *ab initio* gene prediction, transcript evidence and protein evidence, and generate a comprehensive set of protein-coding genes. The transcripts were aligned to genome by GMAP (v.2016-07-11) (Wu and Watanabe, 2005). Proteins annotations were carried out by searching NCBI non-redundant protein database, UniProt plant protein database, SwissProt database, and KEGG databases (Du *et al.*, 2014). Transcription factors identification was conducted in the plant TF database (http://planttfdb.cbi.pku.edu.cn/prediction.php) (Jin *et al.*, 2014).

WGD event analysis

Lychee (Zhang et al., 2021 not published) and sweet orange (Xu *et al.*, 2013) genome were chosen to predicate the whole genome duplication event of rambutan. They were the most closely related species of rambutan that have been published. The whole genome duplication event analysis by Wgd - simple command line tools (Zwaenepoel and Van de Peer, 2019). The synonymous substitution rates (Ks) of sweet orange, rambutan, lychee, sweet orange and rambutan, lychee and rambutan were calculated.

Comparative genome analysis

The phylogenetic tree of 8 species (*O. sativa*, *P. alba*, *A. thaliana*, *C. papaya*, *L. chinesis*, *C. sinensis*, *C. clementina*, and *N. lappaceum*) were construct by OrthoFinder (v.2.2.7) (Emms and Kelly, 2018) with single-copy orthologous genes, which were identified using OrthoMCL (v.2.0) (Li *et al.*, 2003). The amino acid and cds sequence of the 7 species were download from public database NCBI (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/plant/). Ultrametric tree was drawn by the CAFÉ (v.4.2) software (De Bie *et al.*, 2006). The *A. thaliana* and *C. papaya* divergence time (68–72 million years ago) (Ming *et al.*, 2008), and dicotyledons and monocotyledons divergence time (130–240 million years ago)(Jaillon *et al.*, 2007) was applied as calibrators. and CAFÉ was used to identify the gene family that had undergone expansions or contractions among 8 plant genomes.

The synteny of rambutan with lychee using MCScanX (Wang *et al.*, 2012). The chromosomes structure rearrangement analysis was used Mauve (v.2015-02-13) (http://www.darlinglab.org/mauve/).

The ellagitannin biosynthetic pathway genes identification

The genes of ellagitannin biosynthetic pathway in rambutan were blastp by the amnio acid sequence of the pomegranate (*Punica granatum* L.) genome (Qin *et al.*, 2017). The protein domain were predicated in Conserved Domain Database (CDD）(https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi).

Transcriptome data collected and analysis

We select young leaves (YL), mature leaves (ML), young shoot (YS), hermaphroditic flowers (H_Flower), male flowers (M_Flower), and significantly different 4 stages of fruits development from fertilization to mature (The different fruit tissues were sampled at one time). Fruit tissues collect spines, pericarps, arils, and seeds separately, each sample of fruit tissues have three replicates. The four stages of fruit including three stages of fruit rapid development (stage1-stage3) and a ripe stage (stage4, arils turn fleshy and spines turn senescent). The first two stages of arils and seeds were small and hard to separate, so they were mixed in one sample at stage1 (A1+S1) and stage2 (A2+S2). The Stage1 about 10 days after fertilization. Stage 2, Stage 3, and Stage 4 with significantly different in aril development. In detail, Stage 2, aril still invisible;

Stage 3, aril has completely covered the seed; Stage 4, aril turn to fleshy. The other tissues were mix with more than three replicates in one sample.

RNA-Seq libraries were constructed according to the protocol of NEBNext Ultra DNA Library Prep Kit for Illumina sequencing. Each library got 5G database from Illumina Hiseq2500, and they were trimmed to remove the adaptors and low-quality bases using trimmomatic (v.0.36) (Bolger *et al.*, 2014). And the RNAseq data were analysis by Trinity (v.2.8.6) and the different expression genes were normalized by TMM and calculated by edgeR ($p < 0.001$) (Robinson *et al.*, 2010). The GO and KEGG ID of genes were obtained by eggNOG-mapper (Huerta-Cepas *et al.*, 2017) (http://eggnogdb.embl.de/#/app/emapper) and KOBAS (http://kobas.cbi.pku.edu.cn/kobas3/annotate/).The differentially expression genes enrichment by GO and KEGG using OmicShare tools, an online platform for data analysis (http://www.omicshare.com/tools). The genes of the pathways with Pvalue < 0.05 were used to further analysis. All heatmaps were drawn by R script (pheatmap(log2(Exp+1), scale="row", cluster_cols=F, cellwidth=30, cellheight=15, fontsize=12, color = colorRampPalette (c("green", "white","red")) (100))), the parameters of cellwidth, cellheight, fontsize were changed according to the number of genes and samples.

Weighted correlation network analysis (WGCNA) were used to find the candidate genes of aril development. There were total 17,345 genes (FPKM > 3) were used in the WGCNA (Langfelder and Horvath, 2008). Module detection was performed using the TOM-based similarity measure and the dynamic tree cutting algorithm to cut the hierarchal clustering tree and defined modules as branches from the tree cutting. The minimum number of genes per module was set as 30 genes by default, and the threshold of module merging correlation for eigengene similarity was 0.8. For the module-tissue association analysis, the eigengene value was calculated for each module and used to test the association with each tissue type. The total connectivity and intramodular connectivity, kME, and kME-p-value were calculated and represent the Pearson correlation between the expression level of that gene and the ME. Additionally, genes with high degree of intramodular connectivity within a module are referred to as hub genes, and top 10% to 20% genes were chosen in the thesis. The co-expression network (weight value of edge is more than 0.1) was drawn by cytoscape (v.3.8.0)(Shannon *et al.*, 2003).

Expression Analysis

We selected 18 genes from aril and flower development related genes to verify expression profiles by RT-qPCR. The primers were design by Primer 5.0 (Table S13). Total mRNA was extracted with an RNA extraction kit (Tiangen Biotech, Beijing) and reverse transcribed with Hifair® III 1st strand cDNA Synthesis SuperMix (11141ES60, YEASEN Biotech Co., Ltd, China). Real-time qPCR (RT-qPCR) was performed on Roche LightCycler 480 with SYBR Premix ExTaqII (Takara), and data were quantified by the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001). The expression data were normalized by ACTIN. Heatmaps were drawn by R v.3.5.3 (pheatmap (log2(Exp+1), scale="row", cluster_cols=F, cellwidth=30, cellheight=15, fontsize=12, color = colorRampPalette (c("green", "white","red"))(100))).

Re-sequencing and population analysis

The re-sequence data of 29 accessions (20 from USDA germplasm resource, 8 cultivars and 1 male tree from Hainan province of China) (~ 20× genomic coverage for each genome) were trimmed to remove the adaptors and low-quality bases using Trimmomatic (v.0.36) (Bolger *et al.*, 2014), and then they were mapped to the rambutan genome with Bowtie2 (v.2.4.1) with default parameters (Langmead and Salzberg, 2012). The Realigner Target Creator and Indel Realigner programs from the Genome Analysis Toolkit (GATK) package (v.4.0.4.0) (McKenna *et al.*, 2010) were used for global realignment of reads around indels from the sorted BAM files. Variant calling for each genome was carried out by GATK HaplotypeCaller to produce VCF files. Low depths and repetitive variants were removed from the raw VCF file with parameter (min-alleles 2, max-alleles 2, minDP 4, maxDP 60, minQ 30, max missing 0.9, maf 0.05) by VCFtools (Danecek *et al.*, 2011). After 2 accessions with low depth were remove, 27 accessions with high confidence SNPs were used for genetic distance calculation, and the population structure analysis. All SNPs and indels of each accession were combined to one vcf file, and 2,155,374 high-confidence SNPs were left after filter out indels by vcftools (Danecek *et al.*, 2011). A phylogenetic tree was constructed using the neighbor-joining method implemented in PHYLIP (v.3.6) (Felsenstein, 2005) and displayed in figtree (v.0.9.3) (Price *et al.*, 2010). Population structure analysis was performed with STRUCTURE using the admixture (Pritchard *et al.*, 2000), and the best k (k = 2) was selected by Structure Harvester (v.2.3) (Earl and vonHoldt, 2012). Population genetic statistics of pairwise population differentiation (F_{ST}), nucleotide diversity (π) and Tajima's D were calculated directly from the high confidence SNPs variant set by VCFtools (Danecek *et al.*, 2011). VCFtools was further used to analysis F_{ST} (with 200 kb window and 20 kb step), π (with 50 kb window no overlap), and Tajima's D (with 50kb window no overlap). The lowest 1% genes of π and Tajima's D were chosen to further analysis.

Data availability statement

The raw data of whole genome sequence data, resequencing data, and RNA-seq data of *Nephelium lappaceum* L. were deposited at NCBI GenBank under the accession PRJNA728838. The assembled genome sequence and gff3 file were uploadedhave been deposited in the Genome Warehouse in National Genomics Data Center (CNCB-NGDC Members and Partners, 2021; Chen et al., 2021), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number GWHBECQ00000000 that are publicly accessible at https://ngdc.cncb.ac.cn/gsa. The variation data reported in this paper has been deposited in the Genome Variation Map (Li et al., 2021) in National Genomics Data Center (CNCB-NGDC Members and Partners, 2021), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number GVM000267 that can be publicly accessible at http://bigd.big.ac.cn/gvm/getProjectDetail?project=GVM000267.

Acknowledgments

This work was supported by a startup fund from Fujian Agriculture and Forestry University. Natural Science Foundation of Fujian Province, China (Grant No. 2019J05055). The Outstanding Talent Program of the Ministry of Agriculture and the China Agricultural Research System (CARS-32).

Author Contributions

R.M. conceived this genome project and coordinated research activities; R,M. J.L., S.Z., and W. Z. designed experiments; W.Z., X.Z, and S.C. assembled and annotated the genome; R.M., T.M., W.Z., Jishan L, and X.M generated and analyzed re-sequenced genome; R.M., J.L., S.Z., R.X. W.Z., and X.L designed RNA-seq experiments and generated RNA-seq data; W.Z., F.D., H.J., X.X., Z.Y., P.M., Fang D., B.D., Y.H., Z.L., Y.M., Z.L., Zhicong L., J.L., S.Z. analyzed genome, gene expression, and gene network; W.Z., R.M., and Jishan L. wrote the manuscript.

Conflict of Interest Statement

The authors declare no conflict of interest.

Supporting information

Additional Supporting Information may be found in the online version of this article.

Supporting tables

Table S1. Statistics of contigs and scaffolds level assembly

Table S2. Statistics of chromosome length

Table S3. The completeness of genome assembly and gene annotation using BUSCO pipeline

Table S4. Statistics of repeat sequence in rambutan (Nephelium lappaceum L.)

Table S5. Statistics of chromosomal rearrangement of rambutan genome compared with lychee genome

Table S6. The genes of 6 subclusters for spine development

Table S7. The distribution of transcription factors in aril developmental stages

Table S8. Statistics of the regions and genes of nucleotide diversity (π) (The lowest 1%)

Table S9. Annotation information of genes in the lowest 1% of nucleotide diversity (π)

Table S10. Statistics of the regions and genes of Tajima's D (The lowest 1%)

Table S11. Annotation information of genes in the lowest 1% of Tajima's D

Table S12. Statistics of the genetic diversity of perennial tree fruits crop

Table S13. The RT-qPCR primers sequence

Supporting figures

Figure S1. The chromosome number of rambutan (Baoyan7) was 2n=2x=32.

Figure S2. The phylogenetic tree and insert time analysis of LTR retrotransposons.

(a). Phylogenetic tree of LTR retrotransposons with RT domains. The maximum likehood tree including 342 LTRs from copia and gypsy. (b). Insert time of intact LTR retroelements. The insert time range from 0.01 to 3.77MYA.

Figure S3. Rambutan whole genome duplication event analysis.

The synonymous substitution rates (Ks) distribution of syntenic blocks for three species paralogs (sweet orange, rambutan and lychee), and rambutan orthologs with sweet orange and lychee.

Figure S4. The expression profile of species-specific genes during different tissues in rambutan.

(a). Distribute of 762 species-specific genes in 16 chromosomes. (b). The expression level of 167 species-specific genes of rambutan during 9 tissues. (c). The expression of 12 tissue-specific genes. A1+S1 indicated the sample mixed with Aril1 and Seed1, A2+S2 indicated the sample mixed with Aril2 and Seed2.

Figure S5. The chromosomal rearrangements of rambutan compare with lychee chromosomes.

(a). The inter-genomic comparison of rambutan (16 chromosomes) and lychee (15 chromosomes). (b)-(l). The inter-genomic comparison of single chromosome pair of rambutan and lychee. Red arrows indicated inversion. Green ellipses represent translocations.

Figure S6. The GO enrichment analysis of 263 high expression genes of lychee homologous genes in rambutan.

(a). GO enrichment (Pvalue <0.05) in Cellular Component.

(b). Top 20 of GO enrichment (Pvalue <0.05) in Molecular Function.

(c). Top 20 of GO enrichment (Pvalue <0.05) in Biological Process.

Figure S7. The expression and GO enrichment analysis of 169 spines higher expression genes of rambutan species-specific DEGs compared with lychee.

(a). The expression profile of 169 genes in four fruit tissues. (b). Top 20 of GO enrichment (Pvalue <0.05) in Cellular Component. (c). Top 20 of GO enrichment (Pvalue <0.05) in Molecular

Function. (d). Top 20 of GO enrichment (Pvalue <0.05) in Biological Process.

Figure S8. The expression profile of trichome initiation genes in four developmental stages of spines. *JAZ1* showed upregulation along with spines development.

Figure S9. The 6 expression trend clusters of 1774 spine-specific DEGs.

Figure S10. GO and KEGG enrichment of the spines development candidate genes.

GO enrichment analysis of the 283 up-regulated genes of spines development in Biological Process (Pvalue < 0.05) (a), and the top 20 of KEGG pathway (the top four pathways Pvalue < 0.05) (b).

Figure S11. The co-expression network of top 30 hub genes of spine up-regulated genes.

Figure S12. The phylogenetic tree of pentagalloylglucose oxygen oxidoreductase (POR) gene family in rambutan (48), sweet orange (18), and papaya (11).

Figure S13. The expression profile of aril developmental candidate genes.

The expression profile of aril developmental candidate genes with RNA-seq data (a) and RT-qPCR (b).

Figure S14. The seed tissue-specific DEGs analysis.

(a). The Venn fruit tissue-specific DEGs.

(b). The top 20 of GO enrichment (Pvalue < 0.05) of seed-specific 797 DEGs in biological process.

(c). KEGG enrichment pathways (Pvalue < 0.05) in seed-specific 797 DEGs.

Figure S15. The expression profile of lipid and fatty biosynthetic and metabolic related genes.

Figure S16. The expression profile of fertility genes (a) and transcription factors (b).

Figure S17. The expression profile of sex differentiation genes.

The expression trend of sex differentiation genes with RNA-seq data (a) and RT-qPCR (b).

Figure S18. The pairwise population differentiation (F_{ST}) of 3 groups.

The coefficient of differentiation of group 1 and group 3 was significantly higher than the other two pairs.

Figure S19. The KEGG enrichment and expression analysis of genes in the lowest 1% of nucleotide diversity (π) regions.

(a). The KEGG enrichment analysis of 171 genes in the lowest 1% of nucleotide diversity (π) regions in rambutan. The pathways in the red box with Pvalue < 0.05. (b). The expression analysis of 66 genes (FPKM value more than 3) of 171 in 9 tissues (log2(FPKM+1)).

Figure S20. The KEGG enrichment and expression analysis of genes in the lowest 1% of Tajima's D regions.

(a). The GO enrichment analysis of 141 genes in the lowest 1% of Tajima's D regions in rambutan. The pathways in the red box with Pvalue < 0.05. (b). The expression analysis of 58 genes (FPKM value more than 3) of 141 in 9 tissues (log2(FPKM+1)).

Reference

Abrusán, G., Grundmann, N., DeMester, L. and Makalowski, W. (2009) TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics*, **25**, 1329–1330.

Alandete-Saez, M., Ron, M., Leiboff, S. and McCormick, S. (2011) *Arabidopsis thaliana GEX1* has dual functions in gametophyte development and early embryogenesis. *Plant J*, **68**, 620–632.

Almeyada, N., Mab, S.E. and Martin, F.W. (1979) The rambutan. *Citrus Sub-Trop Fruit J*, **544**, 10–12.

Altendorf, S. (2018) Minor tropical fruits: Mainstreaming a niche market, 69–76.

An, L., Zhou, Z., Yan, A. and Gan, Y. (2011) Progress on trichome development regulated by phytohormone signaling. *Plant Signal Behav*, **6**, 1959–1962.

Arévalo-Galarza, M.L., Caballero-Pérez, J.F., Valdovinos-Ponce, G., Cadena-Iñiguez, J. and Avendaño-Arrazate, C.H. (2018) Growth and histological development of the fruit pericarp in rambutan (*Nephelium lappaceum* Linn.). *Acta Horticulturae*, 165–172.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–20.

Burkill and I.H (1935) A dictionary of the economic products of the Malay Peninsula. *A Dictionary of the Economic Products of the Malay Peninsula*, **137**, 255.

Chen, M., Ma, Y., Wu, S., et al. (2021) Genome Warehouse: A Public Repository Housing Genome-scale Data. *Genomics, Proteomics & Bioinformatics*.

Cheng, M.-C., Hsieh, E.-J., Chen, J.-H., Chen, H.-Y. and Lin, T.-P. (2012) *Arabidopsis* RGLG2, functioning as a RING E3 ligase, interacts with AtERF53 and negatively regulates the plant drought stress response. *Plant Physiol*, **158**, 363–375.

Cheynier, V., Comte, G., Davies, K.M., Lattanzio, V. and Martens, S. (2013) Plant phenolics: Recent advances on their biosynthesis, genetics, and ecophysiology. *Plant Physiology and Biochemistry*, **72**, 1–20.

Chopra, D., Mapar, M., Stephan, L., et al. (2019) Genetic and molecular analysis of trichome development in *Arabis alpina*. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 12078–12083.

CNCB-NGDC Members and Partners (2021) Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. *Nucleic Acids Res*, **49**, D18–D28.

Cui, H., Kong, D., Liu, X. and Hao, Y. (2014) *SCARECROW*, *SCR-LIKE 23* and *SHORT-ROOT* control bundle sheath cell fate and function in *Arabidopsis thaliana*. *Plant J*, **78**, 319–327.

Danecek, P., Auton, A., Abecasis, G., et al. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, p.2156-2158.

De Bie, T., Cristianini, N., Demuth, J.P. and Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–71.

Dixon, R.A. and Paiva, N.L. (1995) Stress-induced phenylpropanoid metabolism. *The Plant Cell*, **7**, 1085–

1097.

Du, J., Yuan, Z., Ma, Z., Song, J., Xie, X. and Chen, Y. (2014) KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Mol Biosyst*, **10**, 2441–7.

Earl, D.A. and vonHoldt, B.M. (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genet Resour*, **4**, 359–361.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–63.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.

Emms, D. and Kelly, S. (2018) OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences, *bioRxiv*.

Felsenstein, J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Gorguet, B., Heusden, A.W. van and Lindhout, P. (2005) Parthenocarpic fruit development in tomato. *Plant Biol (Stuttg)*, **7**, 131–139.

Gutierrez, L., Bussell, J.D., Pacurar, D.I., Schwambach, J., Pacurar, M. and Bellini, C. (2009) Phenotypic plasticity of adventitious rooting in *Arabidopsis* is controlled by complex regulation of AUXIN RESPONSE FACTOR transcripts and microRNA abundance. *Plant Cell*, **21**, 3119–3132.

Haas, B.J., Papanicolaou, A., Yassour, M., et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*, **8**, 1494–1512.

Hajimehdipoor, H., Shahrestani, R. and Shekarchi, M. (2014) Investigating the synergistic antioxidant effects of some flavonoid and phenolic compounds. *Research Journal of Pharmacognosy*, **1**, 35–40.

Han, Y. and Wessler, S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res*, **38**, e199–e199.

Heck, G.R., Perry, S.E., Nichols, K.W. and Fernandez, D.E. (1995) AGL15, a MADS domain protein expressed in developing embryos. *Plant Cell*, **7**, 1271–1282.

Huang, H.B., Jiang Shi Yao and Xie, C. (1983) The initiation of aril and ontogeny of fruit in *Litchi Chinensis* Sonn. *Journal of South China Agricultural Collega*, **4**, 78–83.

Huang, X., Huang, H.B., Gao, A. and Xiao, Z. (2005) Production of rambutan in China. *Acta Horticulturae*, 73–80.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Mering, C. von and Bork, P. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.*, **34**, 2115–2122.

Issara, U., Zzaman, W. and Yang, T.A. (2014) Rambutan seed fat as a potential source of cocoa butter substitute in confectionary product. *International Food Research Journal*, **21**, 25–31.

Jaillon, O., Aury, J.M., Noel, B., et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–7.

Jin, J., Zhang, H., Kong, L., Gao, G. and Luo, J. (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res*, **42**, D1182-7.

Kang, H.G., Jun, S.H., Kim, J., Kawaide, H., Kamiya, Y. and An, G. (1999) Cloning and molecular analyses of a *gibberellin 20-oxidase* gene expressed specifically in developing seeds of watermelon. *Plant Physiol.*, **121**, 373–382.

Ke, G.W., Wang, C.C. and Huang, J.H. (1992) The aril initiation and on to genes of longan fruit. *Journal of Fujian Academy of Agricultural Sciences*, **7**, 22–26.

Klucher, K.M., Chow, H., Reiser, L. and Fischer, R.L. (1996) The *AINTEGUMENTA* gene of *Arabidopsis* required for ovule and female gametophyte development is related to the floral homeotic gene

APETALA2. *Plant Cell*, **8**, 137–153.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*, **27**, 722–736.

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *Bmc Bioinformatics*, **9**, 559.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357–9.

Lattanzio, V., Cardinali, A., Ruta, C., Fortunato, I.M., Lattanzio, V.M.T., Linsalata, V. and Cicco, N. (2009) Relationship of secondary metabolism to growth in oregano (*Origanum vulgare* L.) shoot cultures under nutritional stress. *Environmental and Experimental Botany*, **65**, 54–62.

Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li, C., Tian, D., Tang, B., Liu, X., Teng, X., Zhao, W., Zhang, Z. and Song, S. (2021) Genome Variation Map: a worldwide collection of genome variations across multiple species. *Nucleic Acids Res*, **49**, D1186–D1191.

Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.

Liu, C., Zeng, L., Zhu, S., et al. (2018) Draft genome analysis provides insights into the fiber yield, crude protein biosynthesis, and vegetative growth of domesticated ramie (*Boehmeria nivea* L. Gaud). *DNA Res*, **25**, 173–181.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. and Borodovsky, M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*, **33**, 6494–506.

Lovisetto, A., Baldan, B., Pavanello, A. and Casadoro, G. (2015) Characterization of an *AGAMOUS* gene expressed throughout development of the fleshy fruit-like structure produced by *Ginkgo biloba* around its seeds. *BMC Evol. Biol.*, **15**, 139.

Lovisetto, A., Guzzo, F., Tadiello, A., Toffali, K., Favretto, A. and Casadoro, G. (2012) Molecular analyses of MADS-box genes trace back to Gymnosperms the invention of fleshy fruits. *Mol. Biol. Evol.*, **29**, 409–419.

Mandaokar, A. and Browse, J. (2009) MYB108 acts together with MYB24 to regulate jasmonate-mediated stamen maturation in *Arabidopsis*. *Plant Physiol.*, **149**, 851–862.

Manuel, R.-R., David, W.R., Miguel, A.G., Miguel, S.-F., Lourdes, A.-A. and Isidro, O. (2014) High yields and bee pollination of hermaphroditic rambutan *Nephelium lappaceum* L. in Chiapas, Mexico. *Fruits*, **70**, 23–27.

Marry, S.A. and Josephine R. (2016) *Nephelium Lappaceum* (L.): An overview. *International Journal of Pharmaceutical Science and Research*, 5(**1**), 36–39.

McKenna, A., Hanna, M., Banks, E., et al. (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, **20**, 1297–303.

Ming, R., Hou, S., Feng, Y., et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, **452**, 991–6.

Moncur, M.W. Commonwealth S. and I.R.O. (1988) Floral development of tropical and subtropical fruit and nut species - an atlas of scanning electron micrographs. *Natural Resources Series - CSIRO Division of Water and Land Resources (Australia)*.

Muhamed, S., Kurien, S., Iyer, K.S., Remzeena, A. and Thomas, S. (2019) Natural diversity of rambutan (*Nephelium lappaceum* L.) in Kerala, India. *Genet Resour Crop Evol*, **66**, 1073–1090.

Nakasone, H.Y. and Paull, R.E. (1998) Tropical Fruits. CAB International, Wallingford, UK, 464 pp.

National Genomics Data Center Members and Partners (2020) Database Resources of the National

Genomics Data Center in 2020. *Nucleic Acids Research*, **48**, D24–D33.

Niemetz, R. and Gross, G.G. (2003) Oxidation of pentagalloylglucose to the ellagitannin, tellimagrandin II, by a phenol oxidase from Tellima grandiflora leaves. *Phytochemistry*, **62**, 301–306.

Nieves-Cordones, M., Alemán, F., Martínez, V. and Rubio, F. (2010) The *Arabidopsis thaliana* HAK5 K^+ transporter is required for plant growth and K^+ acquisition from low K^+ solutions under saline conditions. *Mol Plant*, **3**, 326–333.

Okumura, K., Goh, T., Toyokura, K., Kasahara, H., Takebayashi, Y., Mimura, T., Kamiya, Y. and Fukaki, H. (2013) *GNOM/FEWER ROOTS* is required for the establishment of an auxin response maximum for *Arabidopsis* lateral root initiation. *Plant Cell Physiol.*, **54**, 406–417.

Ou, S. and Jiang, N. (2018) LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.*, **176**, 1410–1422.

Perry, S.E., Nichols, K.W. and Fernandez, D.E. (1996) The MADS domain protein AGL15 localizes to the nucleus during early stages of seed development. *Plant Cell*, **8**, 1977–1989.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2 - approximately maximum-likelihood trees for large alignments. *Plos One*, **5**, e9490.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–59.

Pryszcz, L.P. and Gabaldon, T. (2016) Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res*, **44**, e113.

Qin, G., Xu, C., Ming, R., et al. (2017) The pomegranate (*Punica granatum* L.) genome and the genomics of punicalagin biosynthesis. *Plant J.*, **91**, 1108–1128.

Rakariyatham, K., Zhou, D., Rakariyatham, N. and Shahidi, F. (2020) Sapindaceae (*Dimocarpus longan* and *Nephelium lappaceum*) seed and peel by-products: Potential sources for phenolic compounds and use as functional ingredients in food and health applications. *Journal of Functional Foods*, **67**, 103846.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the european molecular biology open software suite. *Trends Genet*, **16**, 276–7.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Serrani, J.C., Fos, M., Atarés, A. and García-Martínez, J.L. (2007) Effect of gibberellin and Auxin on parthenocarpic fruit growth induction in the cv Micro-Tom of tomato. *J Plant Growth Regul*, **26**, 211–221.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498–2504.

Shen, B., Li, C. and Tarczynski, M.C. (2010) High free-methionine and decreased lignin content result from a mutation in the *Arabidopsis* S-adenosyl-L-methionine synthetase 3 gene. *Plant Journal*, **29**, 371–380.

Silveira, S.R., Dornelas, M.C. and Martinelli, A.P. (2016) Perspectives for a framework to understand aril initiation and development. *Front Plant Sci*, **7**, 1919.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–2.

Song, S., Tian, D., Li, C., et al. (2018) Genome Variation Map: a data repository of genome variations in BIG Data Center. *Nucleic Acids Research*, **46**, D944–D949.

Steinbiss, S., Willhoeft, U., Gremme, G. and Kurtz, S. (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res*, **37**, 7002–13.

Sukmandari, N.S., Dash, G.K., Jusof, W.H.W. and Hanafi, M. (2017) A review on *Nephelium lappaceum* L.

Research Journal of Pharmacy and Technology, **10**, 9.

Tan, C.-M., Chen, R.-J., Zhang, J.-H., Gao, X.-L., Li, L.-H., Wang, P.-R., Deng, X.-J. and Xu, Z.-J. (2013) *OsPOP5*, a prolyl oligopeptidase family gene from rice confers abiotic stress tolerance in *Escherichia coli*. *Int J Mol Sci*, **14**, 20204–20219.

Tanase, C., Bujor, O.-C. and Popa, V.I. (2019) Chapter 3 - Phenolic natural compounds and their influence on physiological processes in plants. In R. R. Watson, ed. *Polyphenols in Plants* (*Second Edition*). Academic Press, pp. 45–58.

Thakare, D., Tang, W., Hill, K. and Perry, S.E. (2008) The MADS-domain transcriptional regulator *AGAMOUS-LIKE15* promotes somatic embryo development in *Arabidopsis* and soybean. *Plant Physiol.*, **146**, 1663–1672.

Thitilertdecha, N., Teerawutgulrag, A., Kilburn, J.D. and Rakariyatham, N. (2010) Identification of major phenolic compounds from *Nephelium lappaceum* L. and their antioxidant activities. *Molecules*, **15**, 1453–65.

Valmayor, Aycardo, and Palencia (1970) Growth and flowering habits, floral biology and yield of rambutan (*Nephelium lappaceum* Linn.). *Philippine Agr*.

VanBuren, R., Li, J., Zee, F., Zhu, J., Liu, C., Arumuganathan, A.K. and Ming, R. (2011) Longli is not a hybrid of longan and lychee as revealed by genome size analysis and trichome morphology. *Tropical Plant Biol.*, **4**, 228–236.

Velasco, R., Zharkikh, A., Affourtit, J., et al. (2010) The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nature Genetics*, 42, 833–839.

Walker, B.J., Abeel, T., Shea, T., et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.

Wang, P., Luo, Y., Huang, J., et al. (2020) The genome evolution and domestication of tropical fruit mango.

Genome Biol, **21**, 60.

Wang, Y., Tang, H., Debarry, J.D., et al. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, **40**, e49.

Windarsih and Efendi, M. (2019) Short communication: Morphological characteristics of flower and fruit in several rambutan (*Nephelium lappaceum*) cultivars in Serang City, Banten, Indonesia. *Biodiversitas Journal of Biological Diversity*, **20**.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Xia, R., Li, C., Lu, W., Du, J., Wang, Z. and Li, Z. (2012) 3-Hydroxy-3-methylglutaryl coenzyme A reductase 1 (HMG1) is highly associated with the cell division during the early stage of fruit development which determines the final fruit size in *Litchi chinensis*. *Gene*, **498**, 28–35.

Xu, Q., Chen, L.-L., Ruan, X., et al. (2013) The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.*, **45**, 59–66.

Xu, Z. and Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.*, **35**, W265-268.

Yoon, E.K., Dhar, S., Lee, M.-H., et al. (2016) Conservation and diversification of the SHR-SCR-SCL23 regulatory network in the development of the functional endodermis in *Arabidopsis* shoots. *Mol Plant*, **9**, 1197–1209.

Zee, F.T.P., Chan, H.T.J. and Yen, C.R. (1998) Lychee, longan, rambutan and pulasan.

Zhang, X., Zhang, S., Zhao, Q., Ming, R. and Tang, H. (2019) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants*, **5**.

Zwaenepoel, A. and Van de Peer, Y. (2019) wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics*, **35**, 2153–2155.

Figure legends

Figure 1. High-quality genome of *Nephelium lappaceum* L. (a). Distribution of genomic feature along the rambutan genome. (b). 328Mb sequence was anchored to 16 chromosomes by ALLHiC. The map was drawn by ALLHiC_Plot.

Figure 2. The time tree of 8 angiosperm species.

The divergence time of *Carica papaya* and *Arabidopsis* (68~72MYA), dicotyledons and monocotyledons divergence time (130~240 MYA) was used as refer to estimate the species diversity time.

Figure 3. Collinearity analysis of rambutan and lychee.

(a). Intergenomic synteny of rambutan and lychee from the family of Sapindaceae. (b). The Chr8 and Chr12 of rambutan fusion into chromosome 1 of lychee, and part segment of Chr12 collinear with Chr7 of lychee.

Figure 4. The expression profile of spines development candidate genes.

(a). The four developmental stages of fruit. The Stage1 about 10 days after fertilization; Stage 2, Stage 3, and Stage 4 with significantly different in aril development. Stage 2, aril still invisible; Stage 3, aril has completely covered the seed; Stage 4, aril turn to fleshy. (b). The expression of 282 genes in the up-regulated cluster of four spines development stages. Spines in the fourth stage (Spine4) was senescent. c. The expression profile of top 26 high expression genes of spine up-regulated cluster.

Figure 5. The genes of ellagitannin biosynthesis pathway in rambutan.

(a). The abridged general view of ellagitannin biosynthesis pathway. (b). The number of POR gene family in rambutan, pomegranate, apple, grape, sweet orange, and papaya. (c). The expression level of genes in ellagitannin biosynthesis pathway during four developmental stages of spines, pericarps, arils and seeds. Heatmap was drawn by R with $\log_2$(FPKM+1). The purple box indicated the genes specific expression in one tissue.

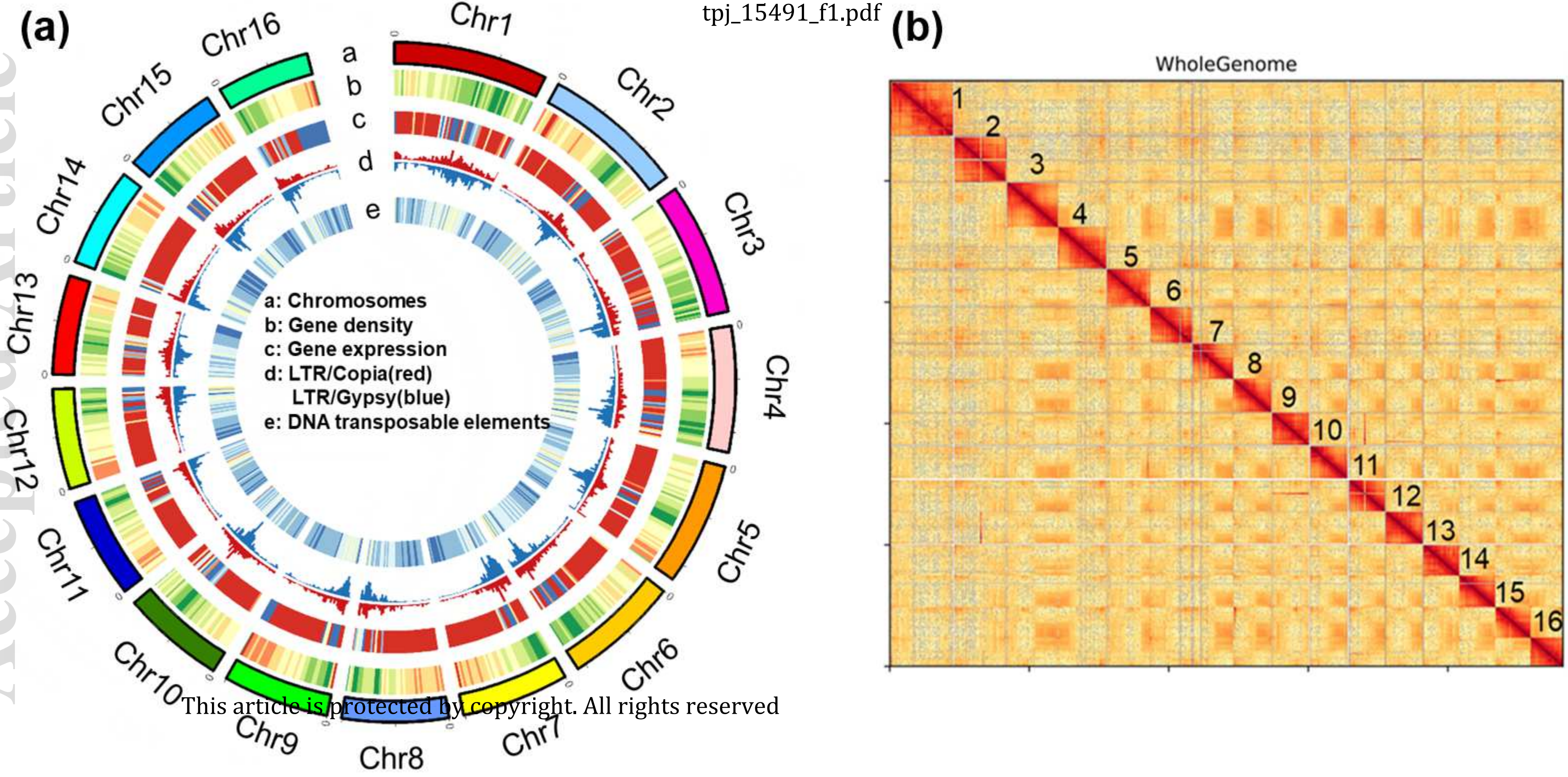
Figure 6. The diagram of WGCNA module (a) and co-expression network (b) of aril development candidate genes.

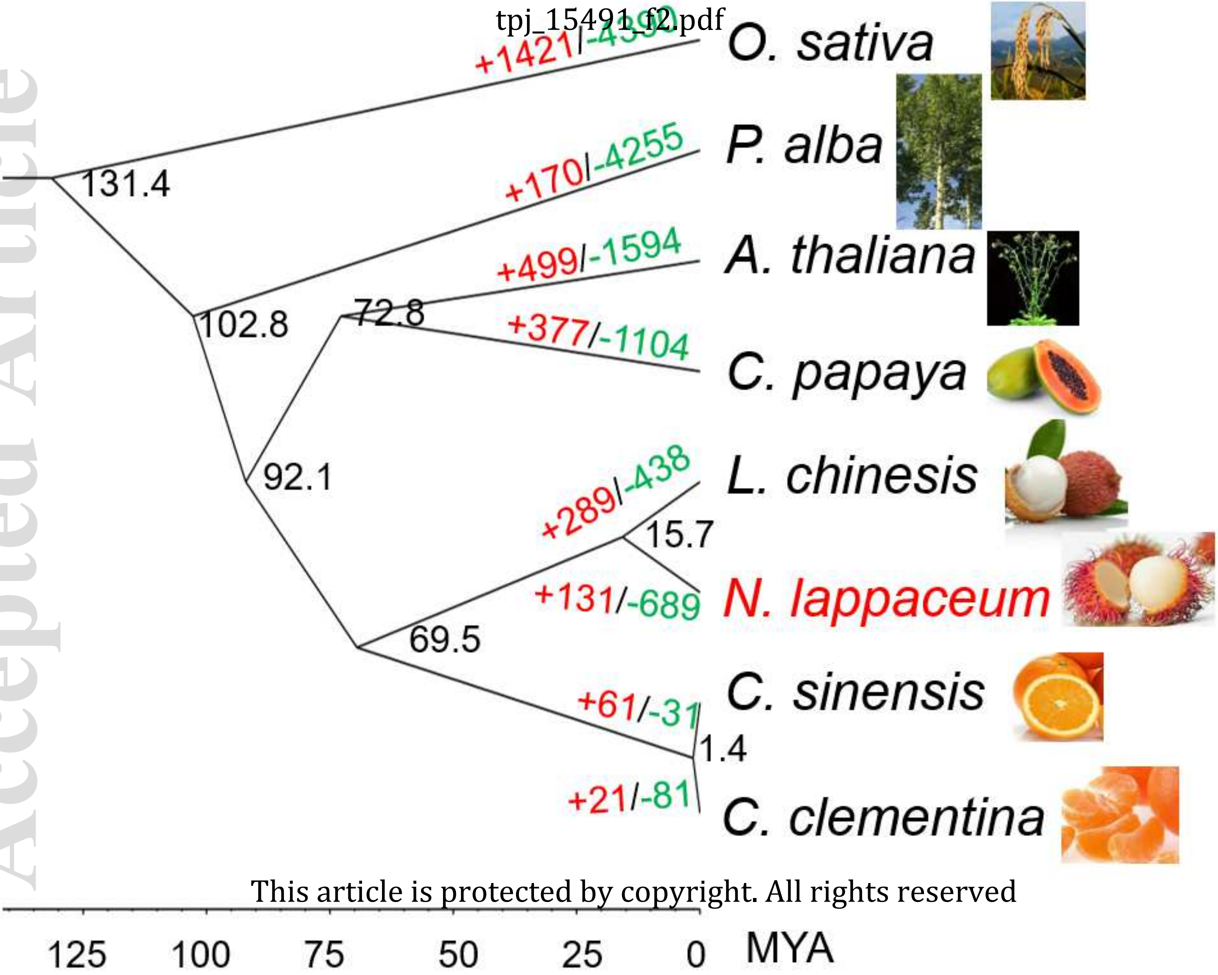
Each column corresponds to a module indicated by different colors. Each row corresponds to a tissue. The color from blue to red indicated the correlation of the genes with tissue in the module from low to high. The genes with red triangle indicated hub genes in the co-expression network. A1+S1 indicated the sample mixed with Aril1 and Seed1, and A2+S2 indicated the sample mixed with Aril2 and Seed2.

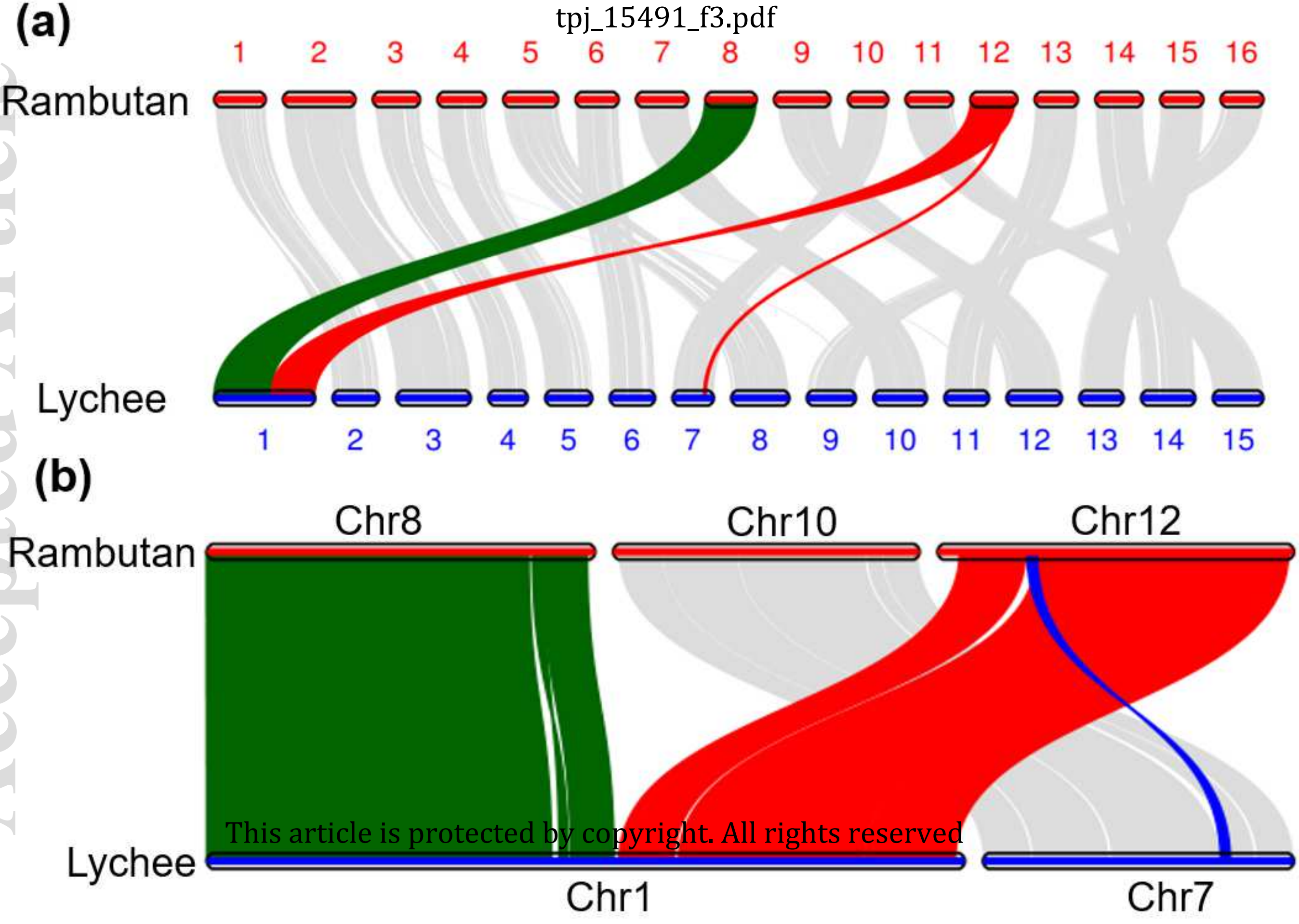
Figure 7. The population structure of 27 rambutan accessions.

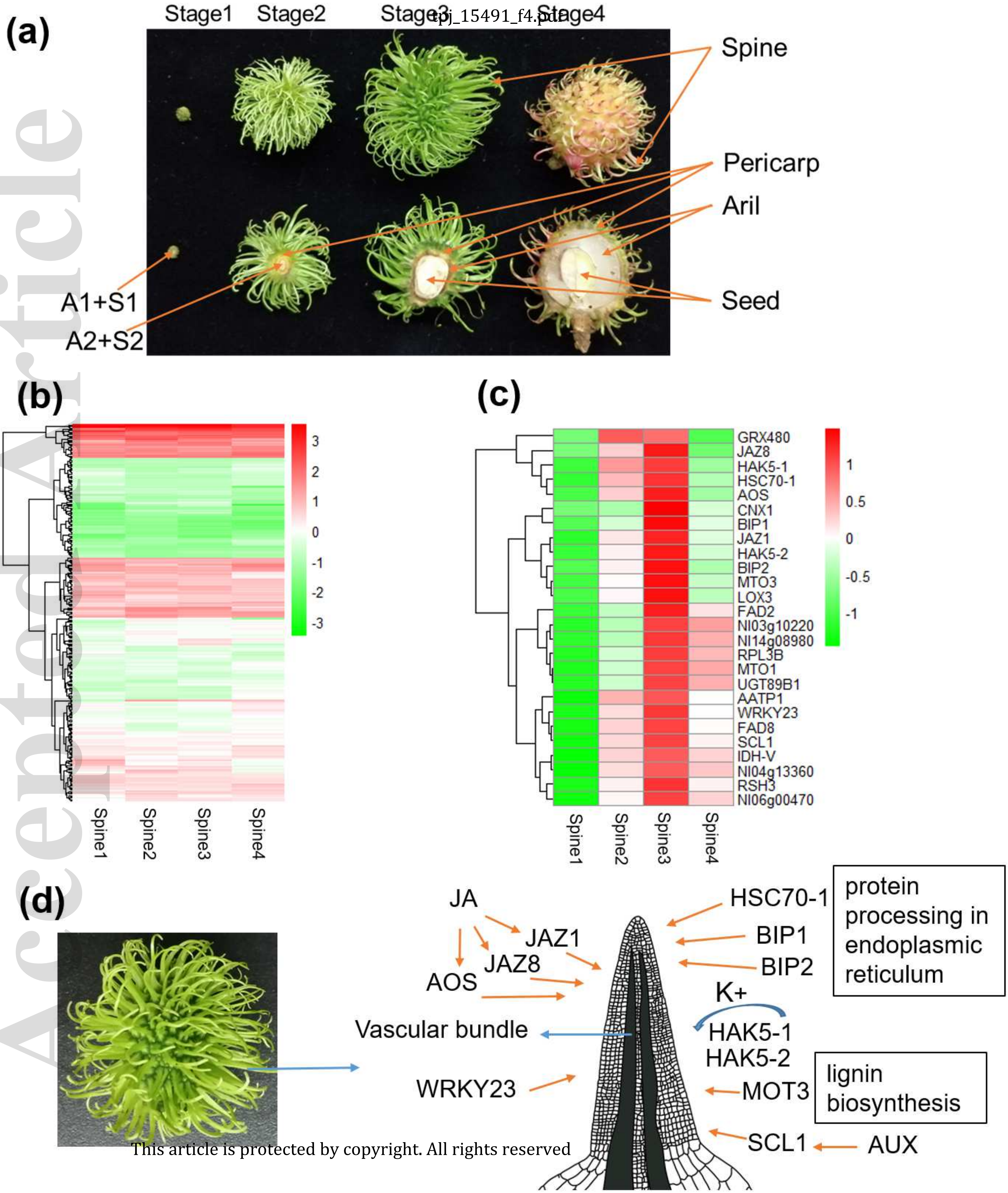
(a). The phylogentic tree of 27 accessions based on the high-confidence 2,155,374 SNPs variant from whole genome re-sequencing. (b). PCA of the 27 accessions using high-confidence SNPs as markers. (c). Population structure of rambutan accessions (the best was K=2).

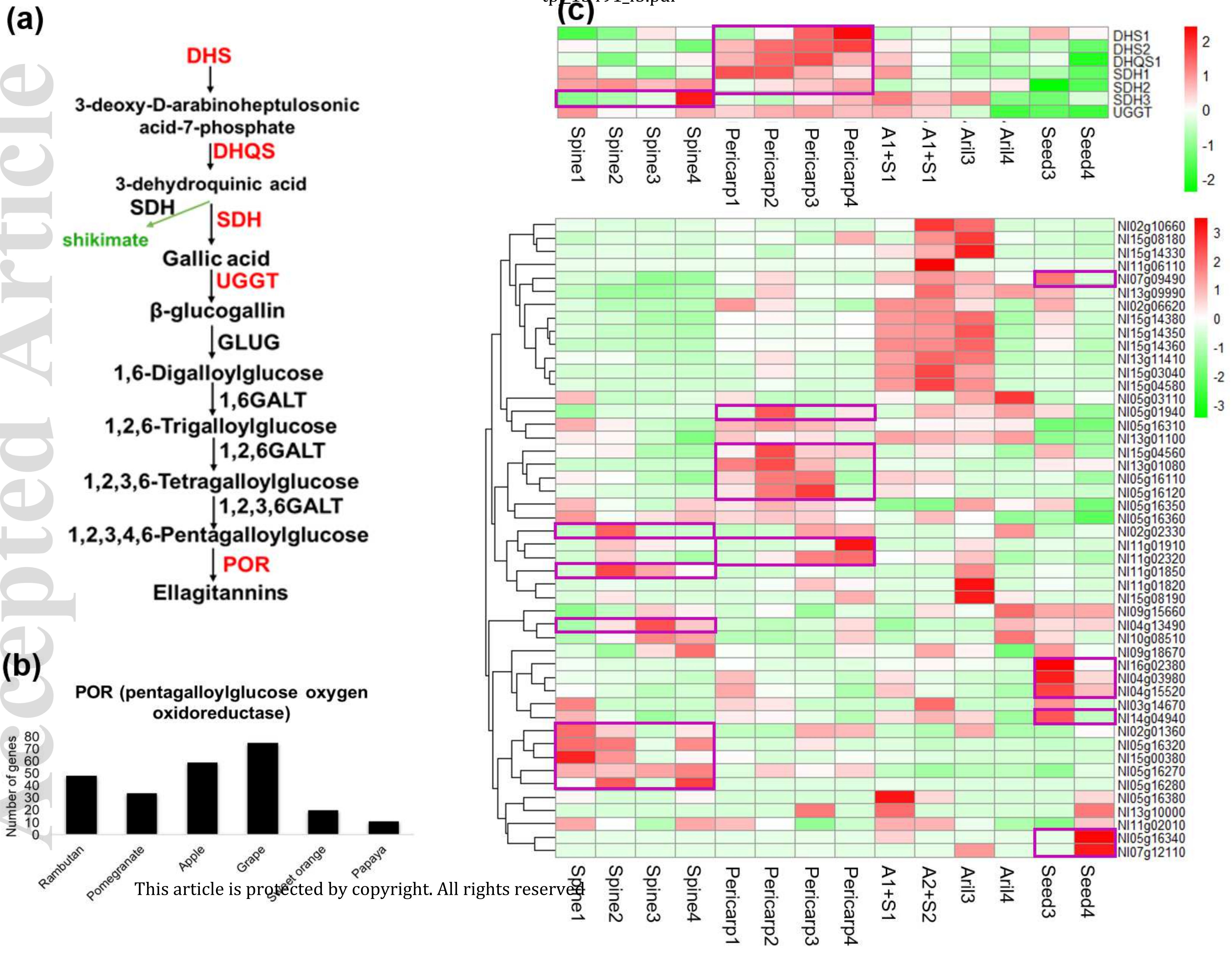
tpj_15491_f1.pdf

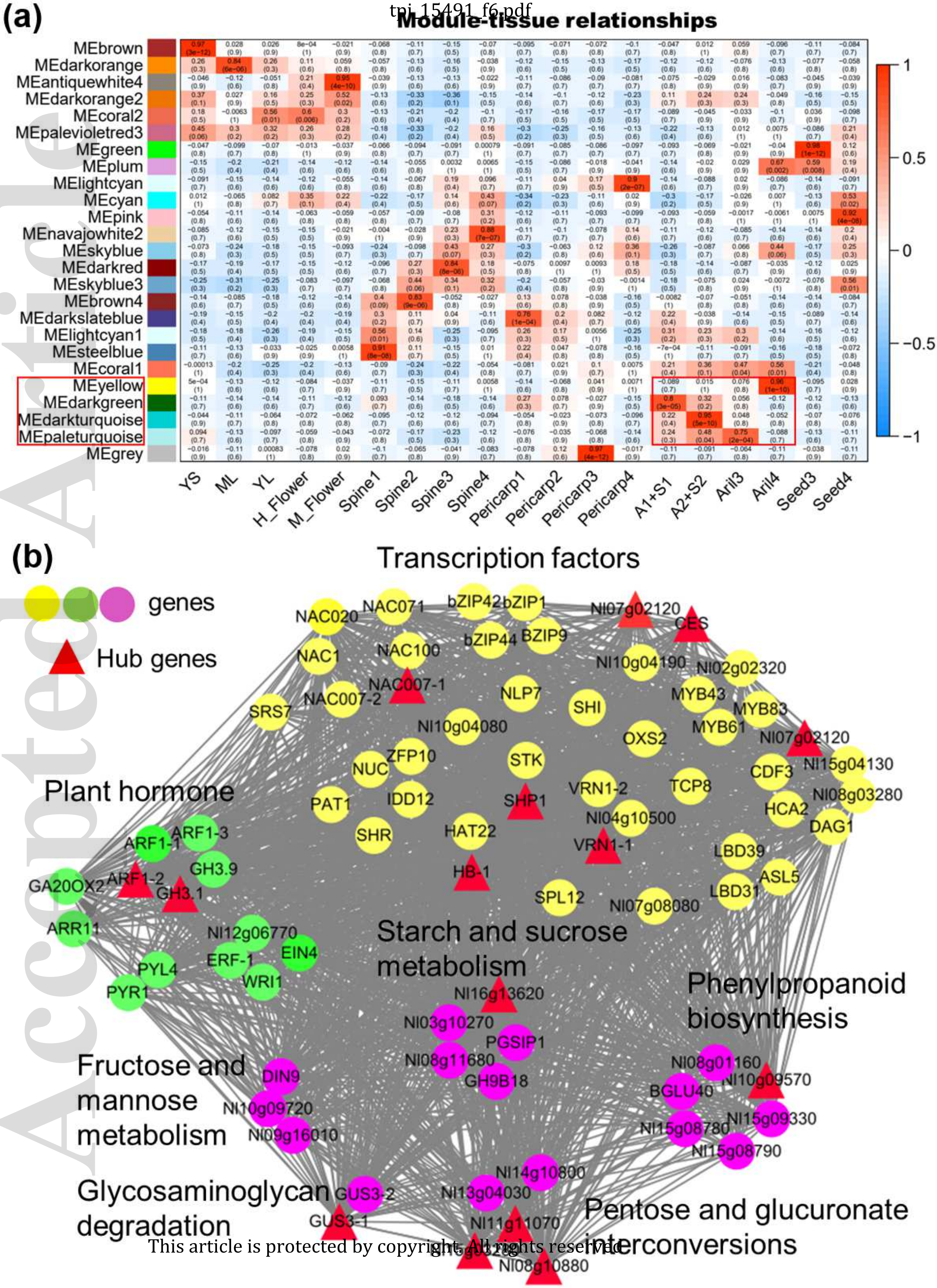


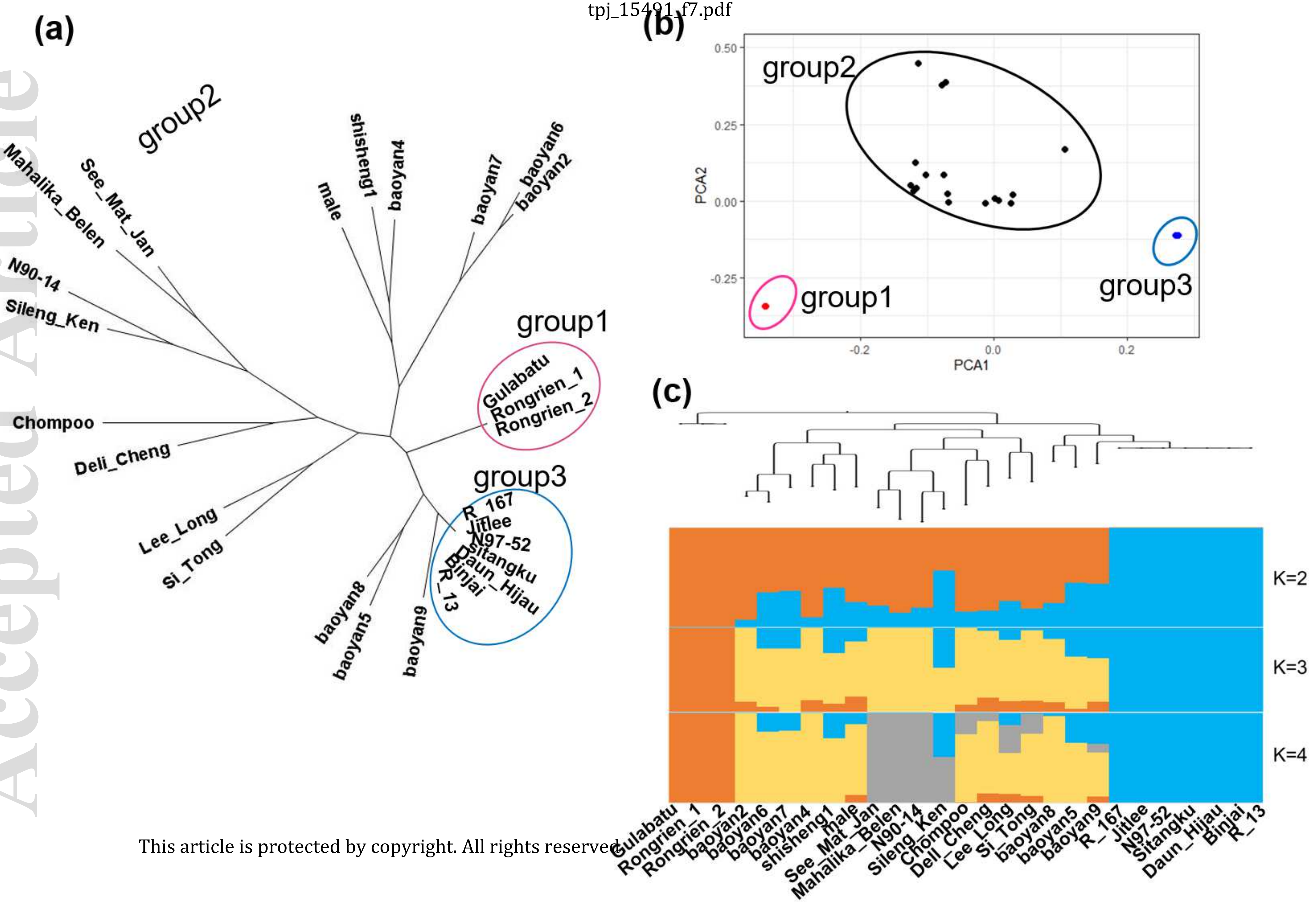

+1421/-4390
O. sativa
+170/-4255
P. alba
131.4
+499/-1594
A. thaliana
102.8
72.8
+377/-1104
C. papaya
+289/-438
L. chinesis
92.1
15.7
+131/-689
N. lappaceum
69.5
+61/-31
C. sinensis
1.4
+21/-81
C. clementina
125
100
75
50
25
0
MYA


(a)
Rambutan
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Lychee
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
(b)
Rambutan
Chr8
Chr10
Chr12
Lychee
Chr1
Chr7


(a)
Stage1
Stage2
Stage3
Stage4
Spine
Pericarp
Aril
Seed
A1+S1
A2+S2
(b)
3
2
1
0
-1
-2
-3
Spine1
Spine2
Spine3
Spine4
(c)
GRX480
JAZ8
HAK5-1
HSC70-1
AOS
CNX1
BIP1
JAZ1
HAK5-2
BIP2
MTO3
LOX3
FAD2
NI03g10220
NI14g08980
RPL3B
MTO1
UGT89B1
AATP1
WRKY23
FAD8
SCL1
IDH-V
NI04g13360
RSH3
NI06g00470
1
0.5
0
-0.5
-1
Spine1
Spine2
Spine3
Spine4
(d)
JA
JAZ1
JAZ8
AOS
Vascular bundle
WRKY23
HSC70-1
BIP1
BIP2
protein processing in endoplasmic reticulum
K+
HAK5-1
HAK5-2
MOT3
lignin biosynthesis
SCL1
AUX


(a)
DHS
3-deoxy-D-arabinoheptulosonic acid-7-phosphate
DHQS
3-dehydroquinic acid
SDH
SDH
shikimate
Gallic acid
UGGT
β-glucogallin
GLUG
1,6-Digalloylglucose
1,6GALT
1,2,6-Trigalloylglucose
1,2,6GALT
1,2,3,6-Tetragalloylglucose
1,2,3,6GALT
1,2,3,4,6-Pentagalloylglucose
POR
Ellagitannins
(b)
POR (pentagalloylglucose oxygen oxidoreductase)
Number of genes
80
70
60
50
40
30
20
10
0
Rambutan
Pomegranate
Apple
Grape
Sweet orange
Papaya
(c)
DHS1
DHS2
DHQS1
SDH1
SDH2
SDH3
UGGT
2
1
0
-1
-2
Spine1
Spine2
Spine3
Spine4
Pericarp1
Pericarp2
Pericarp3
Pericarp4
A1+S1
A1+S1
Aril3
Aril4
Seed3
Seed4
Nl02g10660
Nl15g08180
Nl15g14330
Nl11g06110
Nl07g09490
Nl13g09990
Nl02g06620
Nl15g14380
Nl15g14350
Nl15g14360
Nl13g11410
Nl15g03040
Nl15g04580
Nl05g03110
Nl05g01940
Nl05g16310
Nl13g01100
Nl15g04560
Nl13g01080
Nl05g16110
Nl05g16120
Nl05g16350
Nl05g16360
Nl02g02330
Nl11g01910
Nl11g02320
Nl11g01850
Nl11g01820
Nl15g08190
Nl09g15660
Nl04g13490
Nl10g08510
Nl09g18670
Nl16g02380
Nl04g03980
Nl04g15520
Nl03g14670
Nl14g04940
Nl02g01360
Nl05g16320
Nl15g00380
Nl05g16270
Nl05g16280
Nl05g16380
Nl13g10000
Nl11g02010
Nl05g16340
Nl07g12110
3
2
1
0
-1
-2
-3
Spine1
Spine2
Spine3
Spine4
Pericarp1
Pericarp2
Pericarp3
Pericarp4
A1+S1
A2+S2
Aril3
Aril4
Seed3
Seed4


(a)
Module-tissue relationships
MEbrown
MEdarkorange
MEantiquewhite4
MEdarkorange2
MEcoral2
MEpalevioletred3
MEgreen
MEplum
MElightcyan
MEcyan
MEpink
MEnavajowhite2
MEskyblue
MEdarkred
MEskyblue3
MEbrown4
MEdarkslateblue
MElightcyan1
MEsteelblue
MEcoral1
MEyellow
MEdarkgreen
MEdarkturquoise
MEpaleturquoise
MEgrey
YS
ML
YL
H_Flower
M_Flower
Spine1
Spine2
Spine3
Spine4
Pericarp1
Pericarp2
Pericarp3
Pericarp4
A1+S1
A2+S2
Aril3
Aril4
Seed3
Seed4
1
0.5
0
−0.5
−1
(b)
genes
Hub genes
Transcription factors
NAC020
NAC071
bZIP42
bZIP1
NI07g02120
CES
NAC1
NAC100
bZIP44
BZIP9
NI10g04190
NI02g02320
NAC007-1
NAC007-2
SRS7
NLP7
SHI
MYB43
MYB83
MYB61
NI10g04080
OXS2
NI07g02120
ZFP10
STK
NUC
NI15g04130
CDF3
TCP8
VRN1-2
NI08g03280
PAT1
IDD12
SHP1
HCA2
NI04g10500
DAG1
SHR
HAT22
VRN1-1
LBD39
HB-1
ASL5
SPL12
LBD31
NI07g08080
Plant hormone
ARF1-1
ARF1-3
GH3.9
GA20OX2
ARF1-2
GH3.1
ARR11
NI12g06770
ERF-1
EIN4
PYL4
WRI1
PYR1
Starch and sucrose metabolism
NI16g13620
NI03g10270
PGSIP1
NI08g11680
GH9B18
Phenylpropanoid biosynthesis
NI08g01160
NI10g09570
BGLU40
NI15g08780
NI15g09330
NI15g08790
Fructose and mannose metabolism
DIN9
NI10g09720
NI09g16010
Glycosaminoglycan degradation
GUS3-2
GUS3-1
NI14g10800
NI13g04030
NI11g11070
NI08g10880
Pentose and glucuronate interconversions


(a)
group2
Mahalika_Belen
See_Mat_Jan
male
shisheng1
baoyan4
baoyan7
baoyan6
baoyan2
N90-14
Sileng_Ken
group1
Gulabatu
Rongrien_1
Rongrien_2
Chompoo
Deli_Cheng
group3
R_167
Jitlee
N97-52
Sitangku
Daun_Hijau
Binjai
R_13
Lee_Long
Si_Tong
baoyan8
baoyan5
baoyan9
(b)
0.50
0.25
0.00
-0.25
PCA2
-0.2
0.0
0.2
PCA1
group2
group1
group3
(c)
K=2
K=3
K=4
Gulabatu
Rongrien_1
Rongrien_2
baoyan2
baoyan6
baoyan7
baoyan4
shisheng1
male
See_Mat_Jan
Mahalika_Belen
N90-14
Sileng_Ken
Chompoo
Deli_Cheng
Lee_Long
Si_Tong
baoyan8
baoyan5
baoyan9
R_167
Jitlee
N97-52
Sitangku
Daun_Hijau
Binjai
R_13