**[Supplementary file]**

**Diploid wax apple (*Syzygium samarangense*) genome identified NAC genes regulating its fruit development**

Junyu Zhang1,2#, 21210901000055@hainanu.edu.cn

Zhidong Li1,2#, m15132506079\_1@163.com

Yufan Liang1,2, 22210901000041@hainanu.edu.cn

Shuang He1,2, 21220951310023@hainanu.edu.cn

Guilian Guo1,2, 22220951310071@hainanu.edu.cn

Shuting Yang1,2, 22220951310027@hainanu.edu.cn

Fei Chen1,2**\***, feichen@hainanu.edu.cn

Yinhua Chen1,2**\***, yhchen@hainanu.edu.cn

Wenquan Wang1,2\*, wangwenquan@itbb.org.cn

**Affiliations**

1, College of breeding and multiplication (Sanya Institute of Breeding and Multiplication), Hainan University, Sanya 572025, China

2, College of tropical agriculture and forestry, Hainan University, Danzhou 571737, China

**Plant materials and sequencing**

Wax Apple materials were sampled at the South Lu Base of *Syzygium samarangense* in Sanya, Hainan, China. Leaves, stems, buds, and peduncles of *Syzygium samarangense* collected from this tree were stored at −80 °C until nucleic acid extraction. For the transcriptome analysis, we harvested fruits at different stages of development from Wax Apple: small fruits at 20 days, small fruits at 40 days, medium-sized fruits at 60 days, medium-sized fruits at 75 days, large fruits at 95 days, large fruits at 110 days, and mature fruits at 120 days. Additionally, we collected functional leaves, flowers, and stems. Each sample was replicated twice for analysis. High-molecular-weight DNA was extracted from tender leaves for ultra-long ONT sequencing. We employed the SQK-LSK114 kit to create a standard library after performing quality checks using a NanoDrop One spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). The OXFORD\_NANOPORE PromethION was utilized for sequencing (Oxford Nanopore Technologies, Oxford, UK). For Hi-C and transcriptome sequencing, we utilized the BGI DNBSEQ platform.

**Estimation of genome size and percentage heterozygosity**

Flow cytometry technology was employed, and Cleaned Illumina PE reads from DNAseq libraries were analyzed using GenomeScope 2.0 [1] to estimate genome size and percentage of heterozygosity, utilizing a k-mer size of 21 bp.

**Genome assembly and assessment**

To reduce the raw error rate, the cleaned ONT reads were initially corrected using NanoFilt (v2.8.0) [2]. Illumina reads were corrected using the fastp (v0.20.0) [3] software. The ONT data were put together using the NextDenovo (v2.5.2) [4] program (https://github.com/Nextomics/NextDenovo) with the following settings: genome size =680 Mb, read cutoff = 50 000) . Combining second-generation and third-generation genome sequencing data, we performed five rounds of polishing on the initial assembly contigs using the NextPolish (v1.4.1) [5] software. Hi-C data were utilized to anchor the contigs onto chromosomes. After integrating the eleven contigs generated from ONT data with ALLHiC (v0.9.8) [6], eleven scaffolds were produced, representing eleven pseudochromosomes. The completeness of the final chromosome-level assembly was assessed using the genome evaluation mode of BUSCO (v5.2.2) [7] and the eudicots\_odb10 lineage dataset.

**Transcriptome expression quantification**

We aligned transcriptome data to the reference genome separately using HISAT2 (v2.2.1) [8] and then quantified the expression of each gene in every sample based on the FPKM and TPM calculation formulas using the featureCounts [9] software , and normalize the expression matrix for all genes by log2(FMM+1). Subsequently, we identified key genes and generated heatmaps using tbtools (v2.001) [10] .

**Genome annotation**

For the identification and classification of repetitive sequences, we utilized RepeatModeler (v2.0.3) [11] for de novo prediction and compiled its output into a repeat library. The de novo and known repeat libraries were consolidated and employed to predict repetitive sequences throughout the entire genome using RepeatMasker (v4.1.2) [12] with the following parameters: -nolow, -no\_is, -norna, -parallel 2. Subsequently, RepeatMasker was used to predict TE types with the parameters: RepeatProteinMask, -noLowSimple, -pvalue 0.0001. Finally, we integrated all the predicted repetitive sequences.

We employed *de novo* prediction, homology-based prediction, and RNA-seq-based methods to predict the protein-coding gene structure domains in the genome of Black Sugar Baby Wax Apple. Before conducting ab initio prediction with Augustus (v3.4.0) [13] and GlimmerHMM (v3.0.4) [14], we utilized BUSCO to acquire the training datasets. We performed homology prediction by integrating *Syzygium grande*, *Syzygium aromaticum*, and *Eucalyptus grandis*. Transcriptome data prediction was carried out using third-generation full-length transcripts and RNA-seq in conjunction with Augustus. Finally, all prediction results were integrated using EVidenceModeler (EVM) (v2.1.0) [15] to generate the final gene models. Protein-coding genes were predicted using the online version of Eggnog-Mapper (v2.1.12) [16].

**Karyotype inference**

Download protein sequences of *Syzygium aromaticum*, *Eucalyptus grandis*, and *Vitis vinifera* from the literature. We constructed the ancestral angiosperm karyotype (AAK) using the -km subroutine of WGDI (v0.6.4) [17], and subsequently utilized the proteins of the ancestral core eudicot karyotype (AEK) to infer the karyotypes of the three related species and *Syzygium samarangense*. To calculate fission and fusion events, we initially tallied all collinear color blocks to determine all splitting times. Then, fusion and fission times were computed based on the total count.

**Phylogenomic and phylogenetic inference**

We employed OrthoFinder (v2.5.5) [18] to identify and align orthogroups in the eight species. The resulting alignment was utilized as input for RAxML (v1.2.0) [19] to construct a phylogenetic tree, and the MCMCTree (v4.9) [20] pipeline of PAML (v4.10.6) [21] was utilized to estimate species divergence times. And perform gene family contraction and expansion analysis using CAFE (v5.1.0) [22] software, and generate graphical representations using the iTol (v6) [23] website.

For the gene tree construction, we performed multiple sequence alignment of protein sequences from Arabidopsis thaliana, Syzygium aromaticum, and Syzygium samarangense using the MAFFT (v7.520) [24] software, we constructed a phylogenetic tree using the FastTree (v2.1.11) [25] software, and visualized the evolutionary tree using MEGA (v11) [26] software.

**Venn diagram**

Using the results from OrthoFinder, we created a Venn diagram using the online software (<http://jvenn.toulouse.inra.fr/app/example.html>) [27].

**Co-expression network in wax apple**

Using the R package WGCNA[28], filter out genes that exhibit low expression or minimal variation across all samples (with an expression of 0) to construct a weighted gene co-expression network. Establish modules with the following parameters: maxBlockSize = 100000, power = 7, networkType = "unsigned," mergeCutHeight = 0.25, deepSplit = 3, minModuleSize = 30, corType = "pearson" (indicating Pearson correlation), and "biweight midcorrelation," which serves as a correlation measure for co-expression network analysis.

**Construction of differentially expressed genes**

Divide all samples into 10 distinct groups, grouping replicates by calculating the mean. Conduct pairwise comparisons between different groups using DESeq2[29] and edgeR[30] software. Filter the resulting differentially expressed genes based on the following criteria: |log2FoldChange| > 2, padj < 0.05. Combine the filtered differentially expressed genes.

**Gene expression profile at different time points and in different organs in the transcriptome**

Using the sample expression data correlated with the selected differentially expressed genes as the input table, employ the online Hiplot[31] platform to cluster all differential genes into 12 clusters and generate a trend plot illustrating gene clustering.

**Results and discussions**

**Gene mining related to fruit aroma on terpene metabolism**

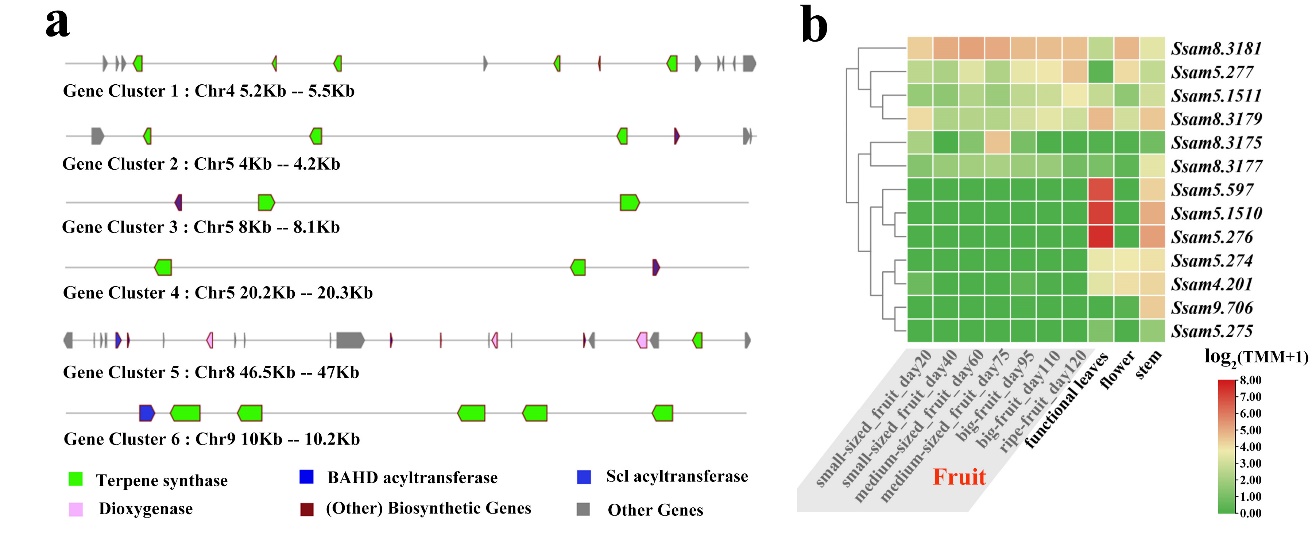


Figure S1 (a) The six gene clusters were predicted to involve in the terpene metabolism. (b) Expressional profile of all genes on six gene clusters. The genes do not express in any of the organ were not displayed.

Wax apple plants are renowned for their rich aroma composed of Volatile Organic Compounds (VOCs), with the ripe fruit emanating the most intense fragrance. The primary components of this fragrance comprise terpenoids, alcohols, ketones, and esters [32]. We identified 61 secondary metabolism-related gene clusters in the wax apple genome by submitting it to the specialized plantiSMASH database (Fig. S1a). Of these, six gene clusters are involved in terpenoid biosynthesis, located on chromosomes 4, 5 (3 clusters), 8, and 9. Notably, we observed that 12 genes are expressed in different plant organs, yet none of them exhibited fruit-specific expression, aligning with the characteristic presence of VOCs throughout the entire wax apple plant. Among these, the gene Ssam8.3181 displayed increased expression as the fruit matured (Fig. S1b), suggesting its potential role as a key contributor to fruit flavor. Conversely, genes such as Ssam5.597, Ssam5.1510, and Ssam5.276 were exclusively expressed in vegetative organs, possibly participating in VOC metabolism within these plant parts.

**Gene mining on fruit maturation**

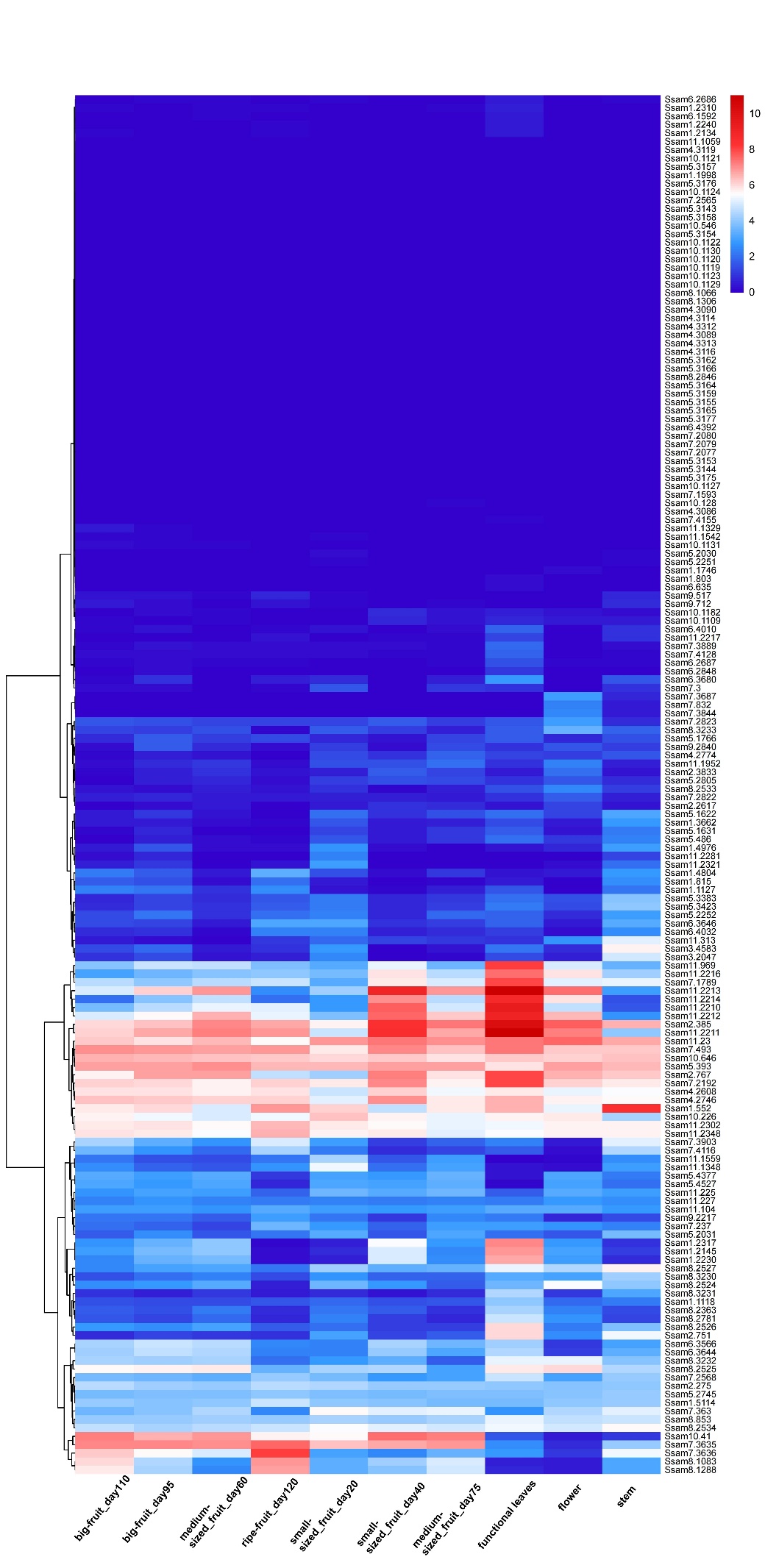


Figure S2 Heatmap of all genes in the wax apple NAC gene family at different time points and in various organs

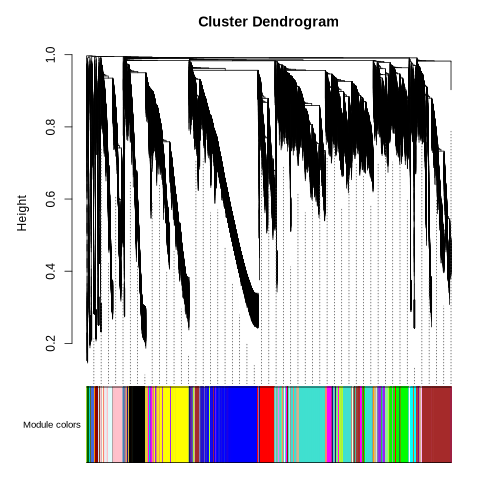


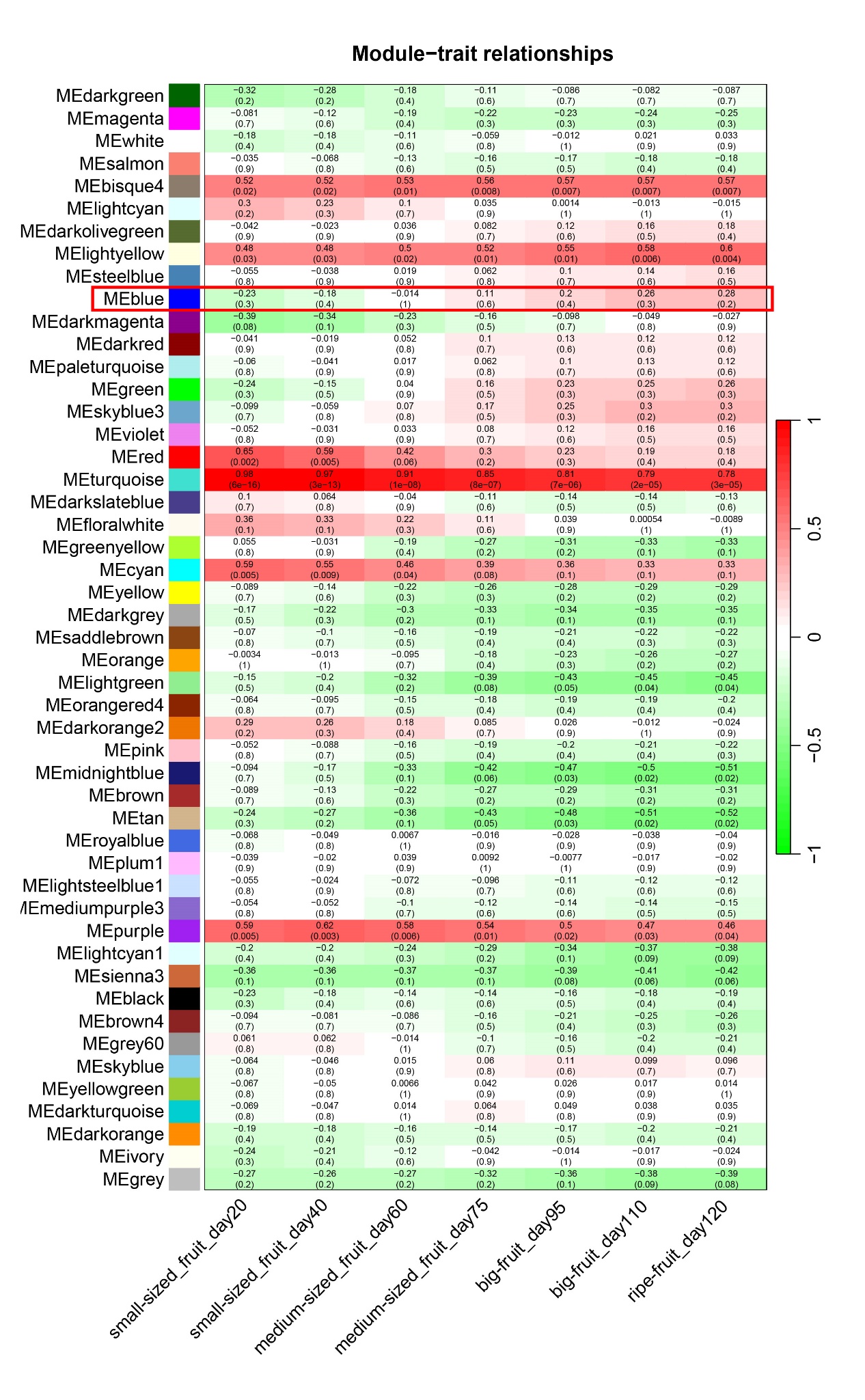
Figure S3 Hierarchical clustering tree depicting co-expression modules identified by WGCNA. Each branch represents a gene, and the tree branches cluster into 49 groups, distinguished by different colors

Figure S4: Relationship between consistent module feature genes and different stages. Each row in the table corresponds to a co-expression module with the grouping in Figure S3. The seven columns in the table represent seven different developmental stages of wax apple.

**References**

1. Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes**. *Nat Commun* 2020, **11**(1):1432.

2. De Coster W, Rademakers R: **NanoPack2: population-scale evaluation of long-read sequencing data**. *Bioinformatics* 2023, **39**(5).

3. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics* 2018, **34**(17):i884-i890.

4. Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K *et al*: **An efficient error correction and accurate assembly tool for noisy long reads**. *bioRxiv* 2023:2023.2003.2009.531669.

5. Hu J, Fan J, Sun Z, Liu S: **NextPolish: a fast and efficient genome polishing tool for long-read assembly**. *Bioinformatics* 2020, **36**(7):2253-2255.

6. Zhang X, Zhang S, Zhao Q, Ming R, Tang H: **Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data**. *Nat Plants* 2019, **5**(8):833-845.

7. Manni M, Berkeley MR, Seppey M, Zdobnov EM: **BUSCO: Assessing Genomic Data Quality and Beyond**. *Curr Protoc* 2021, **1**(12):e323.

8. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL: **Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype**. *Nat Biotechnol* 2019, **37**(8):907-915.

9. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features**. *Bioinformatics* 2014, **30**(7):923-930.

10. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R: **TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data**. *Mol Plant* 2020, **13**(8):1194-1202.

11. Bao Z, Eddy SR: **Automated de novo identification of repeat sequence families in sequenced genomes**. *Genome Res* 2002, **12**(8):1269-1276.

12. Tarailo-Graovac M, Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences**. *Curr Protoc Bioinformatics* 2009, **Chapter 4**:4 10 11-14 10 14.

13. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W435-439.

14. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders**. *Bioinformatics* 2004, **20**(16):2878-2879.

15. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments**. *Genome Biol* 2008, **9**(1):R7.

16. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P: **Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper**. *Mol Biol Evol* 2017, **34**(8):2115-2122.

17. Sun P, Jiao B, Yang Y, Shan L, Li T, Li X, Xi Z, Wang X, Liu J: **WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes**. *Mol Plant* 2022, **15**(12):1841-1851.

18. Emms DM, Kelly S: **OrthoFinder: phylogenetic orthology inference for comparative genomics**. *Genome Biol* 2019, **20**(1):238.

19. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014, **30**(9):1312-1313.

20. Puttick MN: **MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees**. *Bioinformatics* 2019, **35**(24):5321-5322.

21. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. *Mol Biol Evol* 2007, **24**(8):1586-1591.

22. 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.

23. Letunic I, Bork P: **Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation**. *Nucleic Acids Res* 2021, **49**(W1):W293-W296.

24. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment**. *Nucleic Acids Res* 2005, **33**(2):511-518.

25. Price MN, Dehal PS, Arkin AP: **FastTree: computing large minimum evolution trees with profiles instead of a distance matrix**. *Mol Biol Evol* 2009, **26**(7):1641-1650.

26. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: **MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms**. *Mol Biol Evol* 2018, **35**(6):1547-1549.

27. Bardou P, Mariette J, Escudie F, Djemiel C, Klopp C: **jvenn: an interactive Venn diagram viewer**. *BMC Bioinformatics* 2014, **15**(1):293.

28. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics* 2008, **9**:559.

29. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.

30. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.

31. Li J, Miao B, Wang S, Dong W, Xu H, Si C, Wang W, Duan S, Lou J, Bao Z *et al*: **Hiplot: a comprehensive and easy-to-use web service for boosting publication-ready biomedical data visualization**. *Brief Bioinform* 2022, **23**(4).

32. Ruan Z, Jin J, Fan X: **Volatile constituent of black peral wax apple fruit at different developmental stages**. *Journal of South China Agricultural University* 2013, **34**:182-186.