**Research Article**

# Genomic insights into longan evolution from a chromosome-level genome assembly and population analysis of longan accessions

**Authors:**

Jing Wang[1,2*], Jianguang Li[1,2*†], Zaiyuan Li[3], Bo Liu[3], Lili Zhang[4], Dongliang Guo[1,2], Shilian Huang[1,2], Wanqiang Qian[3†], Li Guo[4†]

**Affiliations:**

*[1]Key Laboratory of South Subtropical Fruit Biology and Genetic Resource Utilization, Ministry of Agriculture, Key Laboratory of Tropical and Subtropical Fruit Tree Research of Guangdong Province, Guangzhou 510640, China;*

*[2]Institution of Fruit Tree Research, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China*

*[3]Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen518000, China*

*[4]MOE Key Laboratory for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049 China*

**Short title:** Longan genome assembly and population genomics

*: Equal contribution

† **Corresponding authors:** lijianguang@gdaas.cn (JL); qianwanqiang@caas.cn (WQ)**;** guo_li@xjtu.edu.cn (LG)

## ABSTRACT

Longan (*Dimocarpus longan*) is a subtropical fruit best known for its nutritious fruit and has been regarded as a precious tonic and traditional medicine since ancient times. High-quality chromosome-scale genome assembly is valuable for functional genomic study and genetic improvement of longan. Here, we report a chromosome-level reference genome sequence for longan cultivar JDB with an assembled genome of 455.5 Mb in size anchored to fifteen chromosomes, representing a significant improvement of contiguity (contig N50=12.1 Mb, scaffold N50= 29.5 Mb) over a previous draft assembly. A total of 40,420 protein-coding genes were predicted in *D. longan* genome. Synteny analysis suggests longan shares the widespread gamma event with core eudicots, but has no other whole genome duplications. Comparative genomics showed that *D. longan* genome experienced significant expansions of gene families related to phenylpropanoid biosynthesis and UDP-glucosyltransferase. Deep genome sequencing analysis of 87 longan accessions identified longan biogeography as a major contributing factor for genetic diversity, and revealed a clear population admixture and introgression among cultivars of different geographic origins, postulating a likely migration trajectory of longan overall confirmed by existing historical records. The chromosome-level reference genome assembly, annotation and population genetic resource for *D. longan* will facilitate the molecular studies and breeding of desirable longan cultivars in the future.

**Keywords:** *Dimocarpus longan*, reference genome assembly, phenylpropanoid, gene flow, population genomics

## INTRODUCTION

Longan (*Dimocarpus longan* Lour.), also known as dragon's eyeball and closely related to lychee, is a tropical/subtropical evergreen fruit tree in Sapindaceae family with a diploid genome[1] (2n = 2x = 30) It is an important economic fruit tree making great contribution to the rural economic development in tropical and subtropical areas. It is regarded as a precious tonic and traditionally used as a medicinal plant with rich pharmaceutical effects from many parts of the plant, mainly fruits. The main functional metabolites of longan include polysaccharides, polyphenols, flavonoids and alkaloids with anti-oxidative and anti-cancer activities[2]. So far the biosynthetic pathways for these metabolites in longan remain elusive due to limited genetic and genomic resources and technical difficulty of genetic transformation.

Given its high nutritional and economic values, longan was cultivated in many countries around the world, such as China, Australia, Thailand, Vietnam and other countries[3, 4]. China has the largest longan cultivation area and highest production[5], including Guangdong, Guangxi, Fujian, Hainan and other regions in China[6]. According to historical records, longan is native to South China and has been cultivated for more than 2000 years in China with rich germplasm resources[7,8] and lots of wild resources found in Yunnan and Hainan province[9, 10], from which longan was introduced to other South Asian countries such as Thailand[11, 12]. A previous study based on the differences of pollen exine patterns of fourteen longan varieties supports Yunnan as the primary center of longan origin, and Guangdong, Guangxi and Hainan as the secondary centers[13]. Thailand and Vietnam varieties have close genetic relationships indicated by ISSR (Inter-simple sequence repeat) analysis[14]. Although some molecular markers have revealed genetic differences among germplasms, the classification of longan varieties based on these markers has differed among studies, due to different markers, number of varieties and classification methods being adopted[15, 16, 17]. Additionally, the reproduction of longan can be achieved by both inbreeding and crossbreeding which both bear seeds normally, therefore making longan varieties with ambiguous genetic background. Therefore, a resolved population structure of longan varieties and understanding of its genetic diversity require a large-scale phylogenomic study of longan varieties around China and Southeast Asia based on high quality genome assembly and population resequencing data analysis. The knowledge of the longan genetic background and its migration history is also required to improve longan breeding.

Variety breeding has always been important for improving longan production, typically targeting two main traits, size and sweetness of the fruit[18, 19]. The breeding and extension of excellent varieties can enhance the stress resistance of fruit trees, improve the fruit quality and expand the planting area. At present, it is challenging and time-consumming to improve longan by biotechnological breeding due to its long juvenile period and difficulty of genetic transformation, sexual hybridization has been the main approach for longan breeding[20]. Marker-assisted selection (MAS), based on the identification of genes or genomic components related to desired new traits, is an effective biotechnological tool to promote early selection of hybrid progenies at seedling stage[21, 22]. So far, our knowledge about genetic mapping of longan is limited. Guo *et al.* (2011) constructed a low-quality male and female genetic map, consisting of 243 and 184 molecular markers separately[23]. Single nucleotide polymorphism (SNP) markers based on restriction site associated DNA sequencing (RAD-seq) was constructed for quantitative trait loci (QTL) identification by using hybrid progenies $F_1$ and two parents as materials based on a draft genome sequence of *D. longan* “HHZ”[19]. A chromosome-level reference genome sequence and knowledge of the longan genetic background would significantly facilitate the investigation of genotype-phenotype association of longan germplasms and thus expedite the longan breeding program. Although a draft genome sequence of *D. longan* “HHZ” cultivar was available[24], the assembly is essentially fragmented composed of 51,392 contigs with a contig N50 of 26kb.

Here, we produced a chromosome-level genome assembly for *D. longan* JDB cultivar combining Illumina paired-end (PE), PacBio single molecule real-time sequencing and high throughput chromatin capture sequencing (Hi-C). We annotated the genome using *ab inito* prediction, homolog evidence and multi-tissue transcriptomic data. In addition, we conducted population genome deep sequencing from a collection of 87 longan accessions, followed by an in-depth analysis of population structure using high-quality genetic variants. The analysis revealed the population genetic diversity of longan and demonstrated the population admixture and introgression among cultivars from major longan growing areas. The genome assembly, annotations and genetic variants will be valuable to functional genomic studies as well as molecular breeding of *D. longan* for improving the yield, fruit quality and exploiting its medicinal properties.

## RESULTS AND DISCUSSIONS

### Genome assembly and annotation

*D. longan* "JDB" cultivar originated from Fujian is planted in Longan Germplasm Repository of Guangdong Province (Figure 1A-1D), and the fresh young leaves were collected for genomic DNA isolation and sequencing. To generate chromosome-level genome assembly for *D. longan*, we produced 184.4 Gb PacBio single molecule sequencing reads (415x coverage), 25.3Gb (56x coverage) Illumina paired-end (PE) reads, and 57.6Gb (127x coverage) chromosome conformation capture (Hi-C) Illumina read pairs (Supplementary Table 1). We estimated the genome size of *D. longan* cultivar JDB as 474.98 Mb with a heterozygosity rate of 0.36% via k-mer frequency analysis using Illumina PE reads (Figure 1E). High-quality PacBio single-molecule sequencing reads were used to assemble the *D. longan* genome by using *Canu*[25], followed by polishing contigs using Illumina PE reads by using *Pilon*[26], which yielded a draft genome assembly of 455.5Mb (Table 1). Next, Hi-C paired-end reads were used to anchor the PacBio assembled contigs to chromosomes with *Juicer*[27] and *3D-DNA*[28]. The final *D. longan* JDB genome assembly of 455.5Mb covers 95.90% of the estimated genome size (474.98Mb) and 98.7% of sequences were anchored onto 15 chromosomes (Figure 1F) with contig and scaffold N50 of 12.1Mb and 29.6Mb, respectively (Table 1). Thus, this longan genome assembly represents a significant improvement over the highly fragmented *D. longan* HHZ genome assembly (contig N50 0.026 Mb) previously released[24]. Genome completeness was assessed using the plant dataset of the Benchmarking Universal Single Copy Orthologs (BUSCO) database v1.22[29], with e-value < 1e−5. BUSCO evaluation revealed the completeness of 98.1% for our *D. longan* genome assembly (88.4% single copy; duplicated copy 9.7%, 1.1% fragmented and 0.8% missing) (Table 1, Supplementary Table 2).

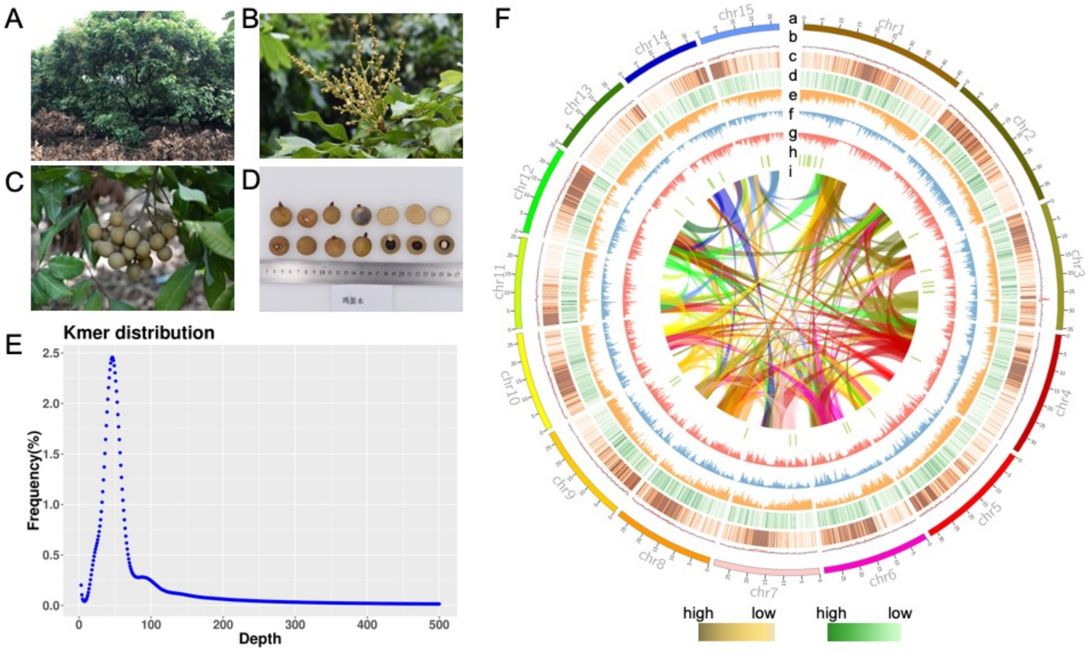

**Figure 1: Chromosome-level genomic assembly of longan (*Dimocarpus longan* Lour.).** (**A-D**): Photos of flower (**B**), fruit cluster (**C**), and fruit section (**D**) of longan cultivar JDB. (**E**) Kmer frequency distribution analysis for JDB genome based on Illumina paired-end reads. (**F**) Overview of *D. longan* genome. Track a to i: chromosomes, GC-content, density of *Gypsy* LTR (long terminal retrotransposons), density of *Copia* LTR, density of protein-coding genes, SNP density, Indel density, distribution of secondary metabolic gene cluster (predicted using *plantismash*), syntenic blocks (color ribbons). The density statistics is calculated within genomic windows of 150kb in size.

We next performed genome annotations by using the *BRAKER2* pipeline combining evidences from *ab initio* prediction, protein homologs and multi-tissue (root, shoot, leaf and fruit) transcriptome sequencing data. The genome annotation pipeline predicted a total of 40,420 protein-coding genes and 2,555 non-coding RNAs for *D. longan*, respectively (Table 1). Longan genome has an overall guanine-cytosine (GC) content of 34 % and gene density of 89 genes per Mb (Supplementary Table 2). About 89.0 % genes have been annotated with NR (non-redundant protein sequence database) and 84.6 % genes with KEGG (Kyoto encyclopedia of genes and genomes) terms

(Supplementary Table 3). Repetitive elements make up 41.7 % of *D. longan* genome, of which 54.9% and 25.4 % are long terminal repeat retrotransposons (LTRs) and DNA transposons respectively. Two major LTR subtypes, LTR-*Copia* (179.64 Mb) and LTR-*Gypsy* (66.18 Mb) represent 8.55 % and 15.53 % of the longan genome, respectively (Supplementary Table 4).

**Table 1. Statistics for *Dimocarpus longan* JDB genome assembly and annotations.**

| | **Statistics** | ***D. longan* JDB (this study)** | ***D. longan* Honghezi[24]** |
|---|---|---|---|
| **Contig** | Total number of contigs | 250 | 51,392 |
| | Assembly size (Mb) | 455.5 | 471.9 |
| | Contig N50 (Mb) | 12.1 | 0.026 |
| | Contig N90 (Mb) | 1.8 | 0.006 |
| | Largest Contig (Mb) | 31.1 | 0.17 |
| **Scaffold** | Total number of scaffolds | 90 | 17,367 |
| | Assembly size (Mb) | 455.5 | 495.3 |
| | Scaffold N50 (Mb) | 29.6 | 0.57 |
| | Scaffold N90 (Mb) | 22.3 | 0.12 |
| | Largest scaffold (Mb) | 46.6 | 6.9 |
| **Annotation** | Number of genes | 40,420 | 31,007 |
| | Repeat content (%) | 41.7 | 52.9 |
| | Number of ncRNA | 2,555 | NA |
| | BUSCO (%) | 98.1% | 94% |
| | GC content (%) | 43.9 | 33.7 |

## Comparative genomics and synteny analysis revealed longan whole genome triplication

Next, we performed intraspecies synteny analysis of *D. longan* genome to investigate its genome evolution history. Intraspecies syntenic gene pairs in *D. longan* were identified using *MCScanX,* which supported the presence of a whole genome triplication (WGT) event in longan genome (Figure 2A). Distribution of synonymous substitution rate (*Ks*) for the syntenic gene pairs also supported that the *D. longan* genome experienced the WGT (Figure 2B). The l:1 ratio of syntenic blocks between longan and grape (*Vitis vinifera*) indicated that the longan WGT was the same event as the grape WGT (γ) event, and no other whole genome duplication occurred following longan-grape divergence (Figure 2C). Furthermore, the 1:2 ratio of syntenic blocks between longan and poplar (*Populus trichocarpa*) confirmed that a species-specific WGD occurred in poplar but did not happen in longan (Figure 2C).

To reveal the genome evolution and divergence of longan, we performed phylogenomic analysis of longan and thirteen representative angiosperm species including eight Rosids (*Citrus sinensis, Carica papaya, Arabidopsis thaliana*, *Theobroma cacao, P. trichocarpa, Ricinus communis, Glycine max, V. vinifera*), two Asteroids *(Solanum tuberosum, Nicotiana attenuata),* one monocotyledon *(Oryza sativa)* and a basal angiosperm *(Amborella trichopoda).* Orthogroup (gene family) identification revealed that these plants shared 7530 orthogroups, 137 of which are single-copy ones (Figure 3A; Supplementary Table 5). Particularly, we identified 1366 orthogroups unique to *D. longan* comparing to *A. thaliana*, *C. cinensis*, *S. tuberosum* and *P. trichocarpa* (Figure 3B). The multiple sequence alignment of 137 single-copy orthologs in 14 species were concatenated and used for phylogeny construction followed by a divergence time estimation using *MCMCTREE* calibrated with fossil record time (Figure 3C). We found that among the thirteen species, longan was phylogenetically closest to *C. sinensis*, which, both belonging to Sapindales. shared a last common ancestor at around 67 million years ago (Mya) that diverged from asteroids (*N. attenuata, S. lycopersicum*) at around 125 Mya (Figure 3C).

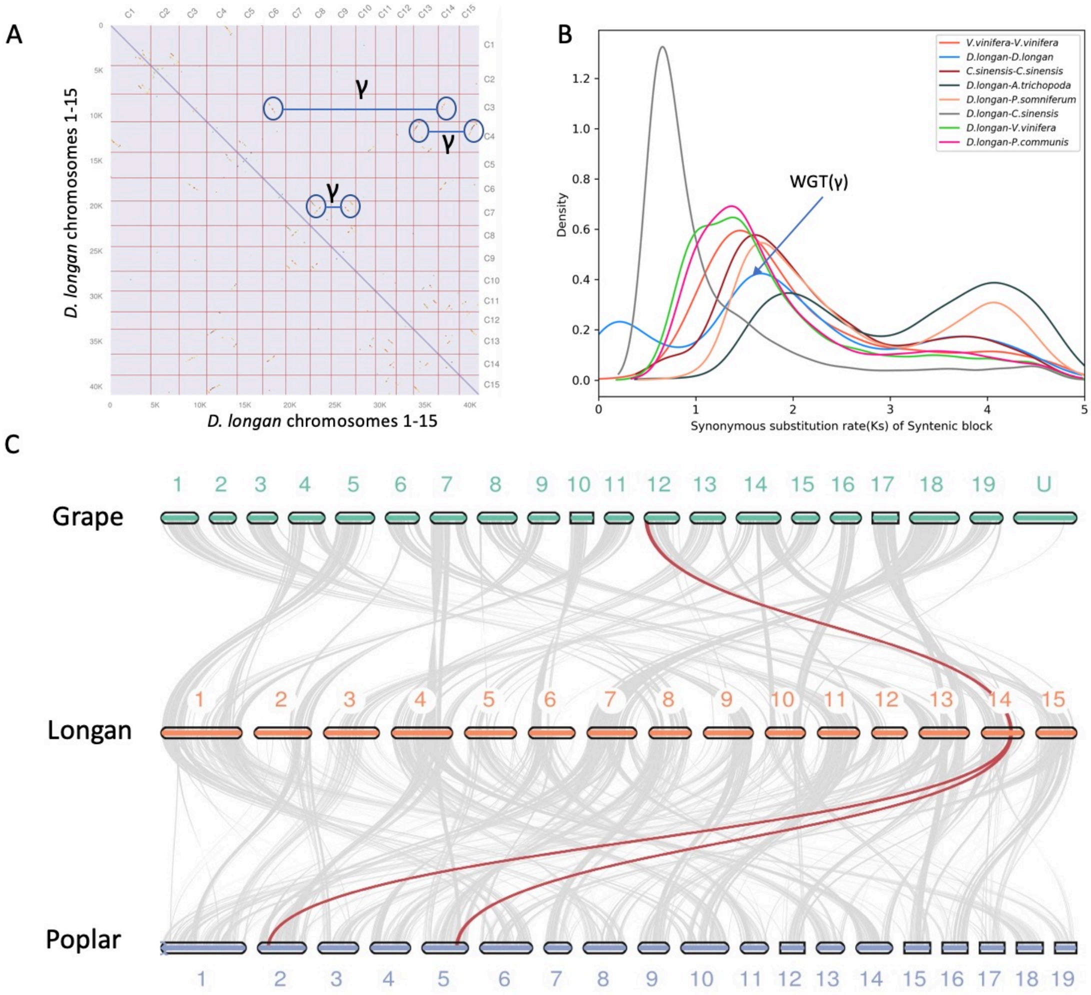

**Figure 2. Comparative genomics and synteny analysis of *Dimocarpus longan*.** (**A)** Whole genome dot plot of *D. longan* showing intraspecies genome synteny based on syntenic gene pairs. The pair of black circles connected by a straight line highlight the syntenic blocks detected in *D. longan* genome, which corresponds to the whole genome triplication (γ event). (**B**) Distribution of *Ks* (synonymous substitution rate) density for syntenic paralogs or orthologs detected in pairwise comparisons among various plant genomes. **(C)** Karyotype macrosynteny plots displaying the collinear relationships for different chromosomes among grape (*Vitis vinifera*), longan (*Dimocarpus longan*) and poplar (*Populus trichocarpa*). The colored lines highlight the syntenic blocks conserved among three species.

## Phylogenomics reveals gene family expansion for phenylpropanoid biosynthesis enzymes and UDP-glucosyltransferases

Gene family contraction/expansion are the evolutionary forces that drive the rapid speciation and result in the diversification of plants[30]. Gene family analysis suggested longan genome has experienced 1474 expanded gene families and 2424 contracted gene families (Figure 3A). KEGG (Kyoto Encyclopedia of Gene and Genomes) enrichment of expanded and contracted gene families ($P < 0.05$) showed that the 312 expanded gene families were significantly enriched with "phenylpropanoid biosynthesis", “phenylalanine metabolism”, “anthocyanin”, “sesquiterpenoid and triterpenoid biosynthesis”, “monoterpenoid biosynthesis” (Figure 3D). Longan is rich in flavonoids and polyphenols, with anti-cancer, anti-oxidant properties in leaf, flower, fruits, and seeds[24, 31, 32, 33], which are derived primarily through phenylpropanoid pathways. The branches of phenylpropanoid metabolism produce end products such as flavonoids, hydrox-ycinnamic acid esters, hydroxycinnamic acid amides (HCAAs), and the precursors of lignin, lignans, and tannins[34]. The phenylpropanoid pathway is one of the most extensively investigated specialized metabolic routes[35]. The 97 expanded longan phenylpropanoid biosynthesis genes were classified into seven gene families: phenylalanine ammonia-lyase (PAL, 5 members), peroxidase (POD, 38 members), O-methyltransferase (OMT, 3 members), glycosyl hydrolase family 1 (GH1, 26 members), aldehyde dehydrogenase family (ADH, 18 members) and AMP-binding enzyme (4 members), beta-galactosidase (BGL, 3 members) (Supplementary Table 6). They participated in the biosynthesis of P-hydroxy-phenyl lignin, quaiacyl lignin,5-hydroxyl-guaiacyl lignin and syringyl lignin, which are precursors of longan. It was speculated that lignins were involved in the longan speciation as a major component of certain plant cell walls[36]. The presence of structural lignins can provide physical barriers preventing the pathogen from entering the plant tissues[37], and required for mechanical support for plant growth and facilitate the long−distance transportation of water and nutrients[38]. In these protection processes, key enzymes of phenylpropanoid and lignin pathway were PAL, POD and PPO[39]. PALs, the first enzyme in the phenylpropanoid biosynthetic pathway, the majority in longan genome were expressed at the higher levels in the roots, leaves and stems, none PAL was highly expressed in the green fruits (Supplemental Figure 1) consistent with previous report[24]. Among the 38 PODs in longan genome, 28 showed differential expression in four major tissues (leaves, stems, roots and fruits) (Supplemental Figure 2). A previous longan genome study revealed non-expanded structural genes involved in phenylpropanoid, and flavonoid pathways[24], which dismatched with our result as expanded phenylpropanoid biosynthesis pathway. However, only PAL in phenylpropanoid pathway was studied, the other six gene families as mentioned above were not studied in the past because of their nontissue-specific expression.

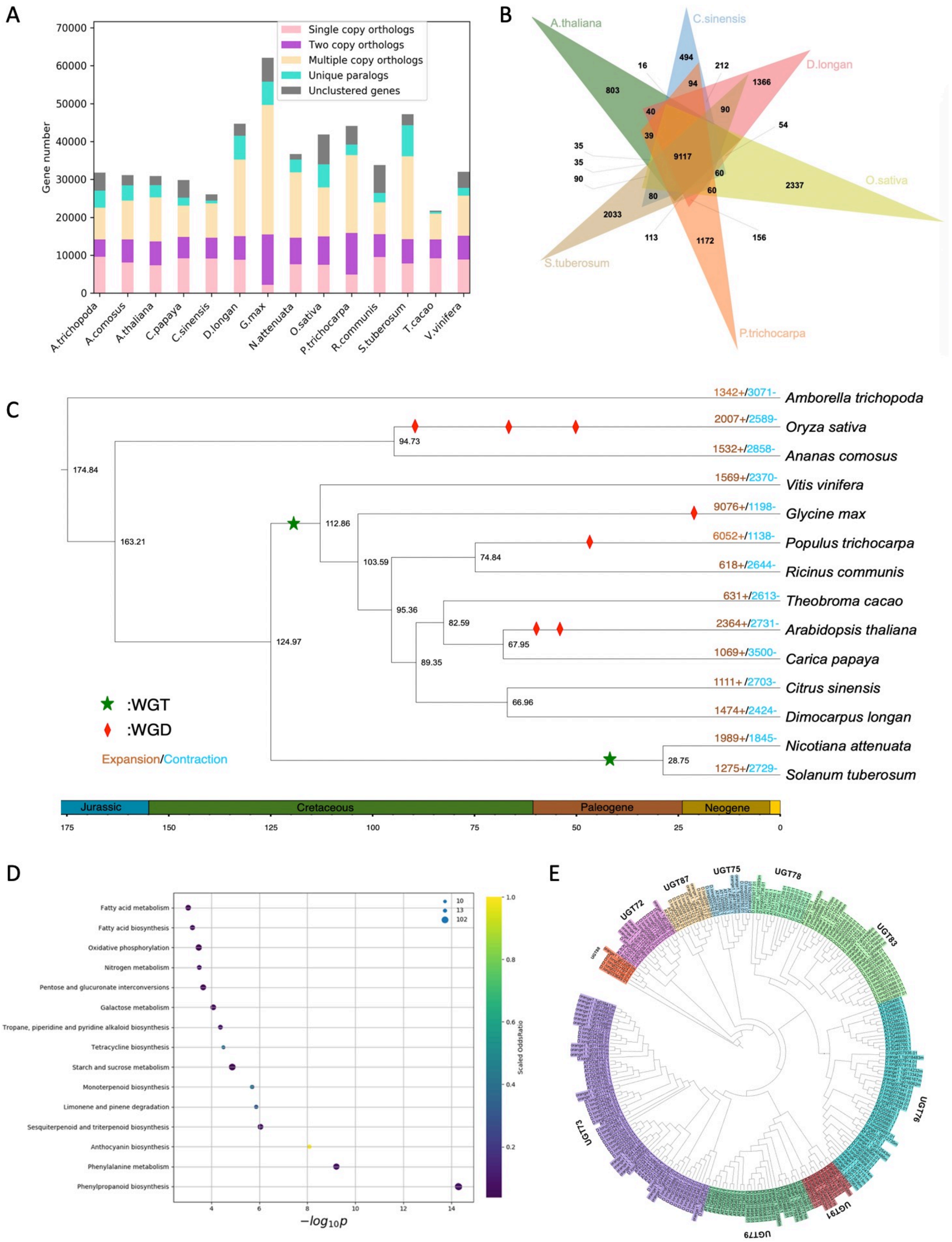

**Figure 3. Phylogenomic genomics of *Dimocarpus longan.*** **(A).** Summary of gene family clustering of *D. longan* and 13 related species. Single copy orthologs: 1-copy genes in ortholog group. Multiple copy orthologs: multiple genes in ortholog group. Unique orthologs: species-specific genes. Other orthologs: the rest of the clustered genes. Unclustered genes: number of genes out of cluster. (**B).** Comparison of orthogroups (gene families) among six angiosperm species including *D. longan* (longan), *A. thaliana* (Arabidopsis), *C. sinensis* (citrus), *S.*

*tuberosum* (potato), *P. trichocarpa* (poplar) and *O. sativa* (rice). **(C).** Phylogenetic relationship and divergence time estimation. The number of gene family expansion and contraction was indicated by red and blue number, respectively. **(D).** Bubble plot summarizing the most significantly enriched KEGG (Kyoto Encyclopedia of Genes and Genomes) terms associated with *D. longan* expanded gene families. X-axis is the log10 transformed p-value. The size of bubble is scaled to the number of genes. The color scale represents the scale of odds ratio in observed versus expected (genomic background) number of genes annotated with specific KEGG terms. **(E).** A phylogenetic tree of UGTs (UDP-glucosyltransferase) in three angiosperms including *D. longan*.

InterPro (IPR) protein domain enrichment analysis showed that the expanded gene families are significantly enriched with IPR domains such as UDP-glucosyltransferase (UGT) and Cytochrome P450s (Supplemental Figure 3). To cope with biotic and abiotic stresses and interact with ecological factors for development, plants have evolved exquisite mechanisms for the biosynthesis of secondary metabolites, through acylation, methylation, glycosylation, and hydroxylation[40, 41]. UGTs are key enzymes for glycosylation, which can stabilize and enhance the solubility of small molecular metabolites in order to maintain intracellular homeostasis[42, 43]. Most of the compounds synthesized by the phenylpropanoid pathway can be glycosylated by glycosyltransferases[44]. For example, UGTs were involved in the glycosylation of volatile benzenoids/phenylpropanoids[45], and also monoterpene linalool[46], a strawberry aroma 4-hydroxy-2,5-di-methyl-3(2H)-furanone[47] etc. A total of 215 UGTs were identified in longan genome (Supplementary Table 7), more than in *Arabidopsis* (107), *C. grandis* (145), *V. vinifera* (181), but fewer than in apple (241)[48, 49, 50]. UGTs participate in multiple plant development and growth processes, including plant defense responses[51, 52]. It has been known for a long time that phenylpropanoid metabolism plays important roles in resistance to pathogen attack[53, 54]. A new mechanism of phenylpropanoid metabolites reprogramming affecting plant immune response through UGT has been revealed[55]. In order to explore the evolutionary relationships of plant UGT families, the phylogenetic tree was constructed based on the longan and other plant UGT protein sequences, including *Arabidopsis*, *Citrus* (Figure 3e). All 115 expanded UGT members were divided into 10 phylogenetic groups. During the evolution of higher plants, the five phylogenetic groups A, D, E, G, and L appeared to expand more than others, although the number of genes found in these groups varies widely among species[50]. In longan, five phylogenetic groups A, D, H, I, and L expanded more than the other groups, whereas no expanded longan UGTs were found in group G. The number of longan UGTs in group D (31 UGTs, UGT73) and group I (19 UGTs, UGT83) was significantly increased compared to those in *Arabidopsis* and *Citrus*. A group D member UGT73C7 was reported to mediate the redirection of phenylpropanoid metabolism to hydroxycinnamic acids (HCAs) and coumarin biosynthesis under biotic stress, resulting in SNC1-dependent Arabidopsis immunity[55]. In group I, number of UGTs was highest compared to other fruits such as peach (5 UGTs), apple (11 UGTs) and grapevine (14 UGTs). UGT83A1 (GSA1) was required for metabolite reprogramming under abiotic stress through the redirection of metabolic flux from lignin biosynthesis to flavonoid biosynthesis and the accumulation of flavonoid glycosides, which coordinately confer high crop productivity and enhanced abiotic stress tolerance[56]. In addition, the number of longan UGTs in groups H (14 UGTs, UGT76) was reduced relative to *Arabidopsis* (21 UGTs). Transcript abundances of UGTs in different tissues were analyzed using RNA-seq data. Among longan *UGTs*, 96 *UGTs* were differentially expressed in longan. Additionally, four (accounting for 4.2%), fourteen (14.6%), and ten (10.4%) UGTs were uniquely expressed in leaf, root, and fruit respectively (Supplemental Figure 4, Supplementary Table 8). The functions of the significantly expanded gene families highlighted the potential roles of these secondary metabolic enzymes to the longan genome evolution and adaptations.

**Tissue-specific expression of terpene biosynthesis genes in roots**

Gene clustering is often associated with biosynthetic pathways for many plant natural products. Therefore, we sought to identify gene clusters in the assembled longan genome that may encode potential secondary metabolic pathways. Genome mining using *Plantismash* pipeline identified 29 secondary metabolic gene clusters in longan, including 21 putatively involved in biosynthesis of alkaloids, saccharide and terpenes (Supplementary Table 9). Tissue-specific transcriptome analysis showed that these gene clusters are expressed in various longan tissues (Supplementary Table 10). In the longan genome, Chr3 contains four gene clusters associated with putative terpene biosynthesis (Supplementary Table 9). It has been reported that CYP450s play critical roles in terpenoid skeleton modification and structural diversity[57, 58]. CYP450 enzymes involved in the terpenoid biosynthesis of pharmaceutical plants were mainly classified in three clans[59]: CYP71, CYP85, CYP72. We found a longan gene cluster with 13 CYP450s on Chr3, which showed root-specific co-expression of eight CYP450 genes (Supplementary Figure 5) within the Clan CYP71 (Supplementary Table 7). CYP450 superfamily was the second expanded gene family by IPR enrichment analysis. The longan CYP450s (435) accounted for 1.1% of longan genes (Supplementary Table 7), much higher than in *Arabidopsis*[60] (244), and grape[61] (236).

Root or phloem of longan has been used to treat filariasis, leucorrhea and other diseases as traditional Chinese medicine. In the past, lots of metabolomics research focused on longan leaf and fruit, whereas metabolic profiles of longan root were not clear. Terpenes were chemical compounds responsible for plant's special odor and flavor profile[62]. Puspita *et al*. (2019) reported that longan leaf ethanol extracts contained flavonoids and triterpenoids[63]. However, many of non-volatile terpenes were exuded from plant roots[64], where they serve as the first line of plant defense and mediate below-ground interactions between plants and other organisms. Therefore, the root-specific expression of a putative terpene biosynthesis gene cluster makes the terpene accumulation and exudation much more effectively in longan roots, playing a role in its biodefense against soil pathogens or herbivores.

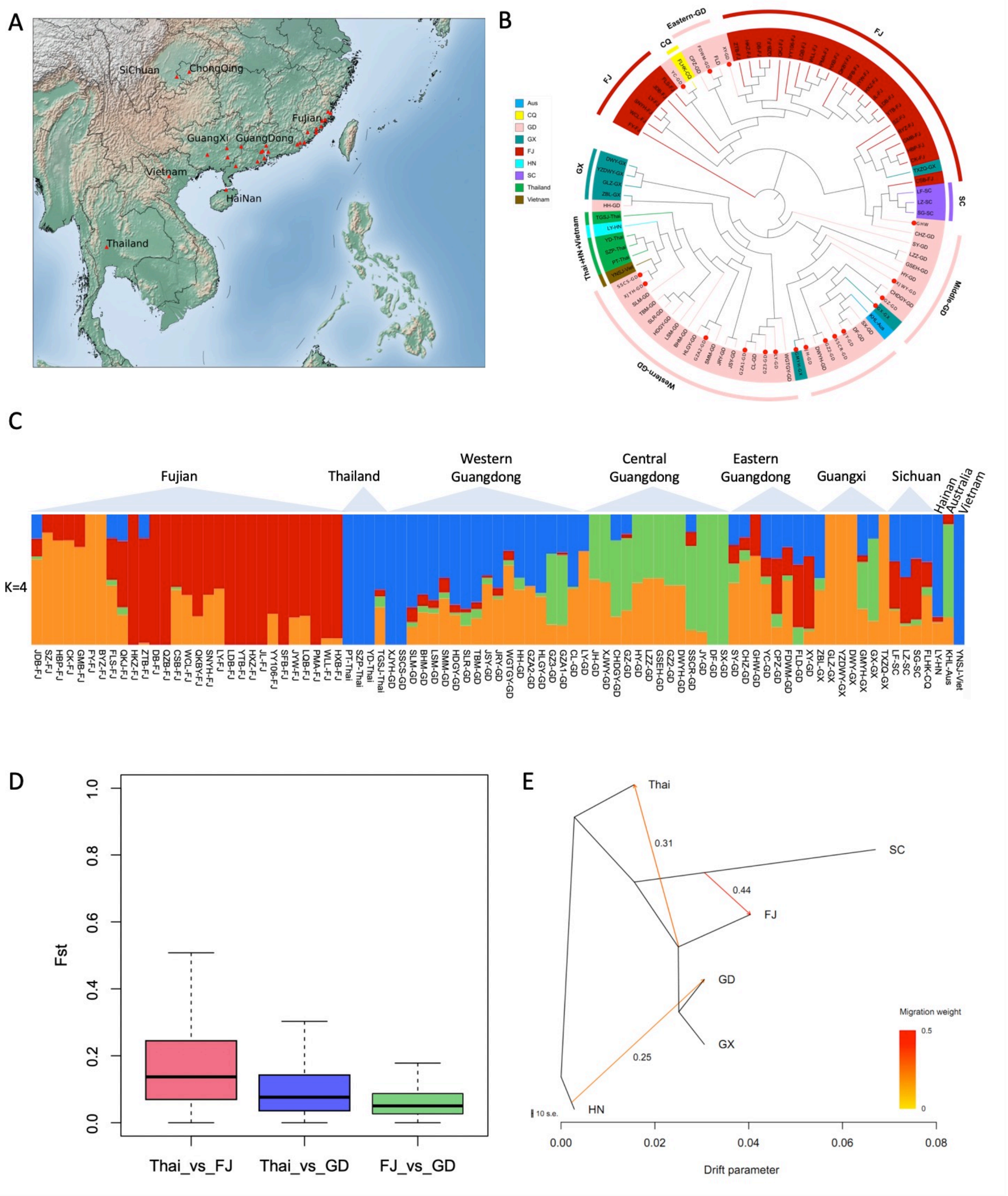

**Figure 4. Population structure and admixture analysis of *Dimocarpus longan*.** (**A**). Sampling localities of seven populations of *D. longan*, where red triangles distinguish the sampling locations. **(B)**. A neighbor-joining phylogenetic tree of all individuals of *D. longan* was constructed using SNPs. The artificial breeding individual was marked with red dots inside. Colors represent different geographic groups. **(C)**. A biogeographical ancestry (admixture) analysis of *D. longan* accessions with four ancestral clusters colored differently in the heatmap, where each column represents a longan sample. (**D**). Distribution of *Fst* values (a measure of genetic differentiation) between longan population from Thailand (Thai), Fujian (FJ) and

Guangdong (GD). **(E)**. maximum-likelihood tree and migration events among seven groups of *D. longan*. The migration events are colored according to their weight.

**Population structure, migration and genetic admixture of longan cultivars**

In the past, longan has been introduced among different populations frequently[7]. Furthermore, longan can bear fruits by both inbreeding and crossbreeding[20]. Lack of reproductive barriers between native cultivars result in ambiguous genetic background of longan germplasms until now. To understand the longan genomic dynamics across its current distribution range in southern China and southeast Asian countries, we performed genome resequencing analysis of 87 accessions (Supplementary Table 11) from five southern provinces in China: Guangdong, Fujian, Guangxi, Sichuan, Hainan, and three other countries Thailand, Vietnam and Australia, with an average sequencing depth of 50×. Read mapping to longan reference genome and variant detection yielded 1,210,426 single nucleotide polymorphisms (SNPs), 204,991 insertions (INS) and 191,681 deletions. After filtering, 7,074,864 SNP loci were polymorphic (allele frequency > 0.05), among which 2,792,700 high-quality SNPs were used for subsequent population genetic analyses.

Although Guangdong borders on Fujian (Figure 4A), the climate of the two provinces was largely different during longan growing season. After generations of planting and screening, different cultivation areas have formed their own longan variety characteristics and types. Using the genetic variant data, we analyzed the population structure within these longan cultivars using phylogenomic analysis and principal component analysis. Phylogenomic analysis clustered 87 longan samples into relatively distinct domestic Guangdong and Fujian groups after removal of artificial breeding populations (Figure 4B). Three Sichuan cultivars were next to Fujian group and distant to GuangDong group. Notebly, two GuangDong cultivars, FLD and CPZ, were clustered with Fujian group, probably because they come from eastern GuangDong adjacent to Fujian. Guangdong cultivars are divided into two subgroups as "Shixai" (SX) and "Chuliang" (CL) from central and western Guangdong, respectively (Figure 4B), also the two main cultivars widely grown in Guangdong and Guangxi. Consistent with the phylogenetic tree, the principal component analysis of the 87 accessions showed that Guangdong and Fujian cultivars were overall grouped separately, while Thailand and Vietnam populations were distant to Chinese populations, when removing artificial breeding cultivars (Supplementary Figure 6).

To investigate the genetic background of longan from various regions, we performed biogeographical ancestry (admixture) analysis based on high-quality SNPs and tested it with ancestral group value (k) ranging from 1 to 10. With a choice of four ancestral groups (k=4) giving the smallest cross-validation errors (Supplementary Figure 7), the admixture analysis discovered a distinct genetic structure within longan accessions of different geographical origins. Longan cultivars from Fujian are composed of primarily two ancestral groups, whereas Guangdong, Guangxi and Sichuan cultivars contain fractions of all four ancestral groups, indicating their more complex ancestry backgrounds than Fujian ones (Figure 4C). The more similar ancestry composition between eastern Guangdong and Fujian cultivars is accordant to the geographical closeness of the two growing regions, suggesting their common ancestral origin or a possible exchange of cultivars between the two regions. By contrast, Thailand and Vietnam cultivars overall have a simple composition with predominantly one ancestral group, most likely shared with western Guangdong and Guangxi cultivars (Figure 4C). Thailand cultivars were genetically more related to western Guangdong cultivars (Figure 4B), but distant from Fujian cultivars. Consistent with this, we have also detected a stronger genetic differentiation (measured in *Fst* value) between Thailand and Fujian than between Thailand and Guangdong (Figure 4D). Notably, the Australian cultivar has a genetic background resembling middle Guangdong cultivars, suggesting it is likely a cultivar of middle-Guangdong origin introduced into Australia lately.

With the diverse ancestry backgrounds in these longan cultivars, we are curious about the migration history of longan germplasms and therefore investigated potential gene flows among different growing areas due to such migration using Treemix analysis. Given its reported origin in China, lots of wild longan resources are present in Yunnan and Hainan province of China[9, 10]. Therefore, the Hainan cultivar was used as an outgroup in this analysis. The Treemix analysis detected a migration event directed from Hainan to Guangdong. There was the highest gene flow (migration weight 0.44) between Sichuan and Fujian (Figure 4E). Gene flows were also detected from the Fujian, Guangdong and Guangxi populations to Thailand with a high weight (migration weight 0.31) (Figure 4E). The detection of gene flows was consistent with longan migration history on record. Longan was first cultivated in ‘Ling-nan’ district of China including Guangdong, Guangxi and Hainan about 2000 years ago, recorded by painting of “San Fu Huang”. According to history records, longan was moved to northern China-Shaanxi Province unsuccessfully, but was successfully introduced to Sichuan and then Fujian with suitable climate conditions (Yang Fu, “Chronicles of the South”, 1st century A.D.). Taken together, our analysis results overall matched history records that there was gene flow from Hainan wild germplasms to Guangdong, then a strong flow from Sichuan to Fujian, and finally the gene flow from China to Thailand.

## MATERIALS AND METHODS

### Germplasm genetic resources

A 30-year-old *D. longan* tree cultivar named JDB from the Institute of Fruit Tree Research at Guangdong Academy of Agricultural Sciences in China was used for genome sequencing and *de novo* assembly in this study. Eighty-seven additional *D. longan* cultivars (Supplementary Table 11) that are widely grown in Southern China and other countries were collected for genome resequencing.

### DNA and RNA isolation

Longan cultivar JDB was planted in Longan Germplasm Repository of Guangdong Province. The fresh and healthy young leaves were collected, cleaned and used for genomic DNA isolation and sequencing. Genomic DNA was extracted from young fresh leaves of *D. longan* using the modified cetyltrimethylammonium bromide (CTAB) method[65]. The concentration and purity of the extracted DNA were assessed using a Nanodrop 2000 spectrophotometer (Thermo, MA, USA) and Qubit 3.0 (Thermo, CA, USA), and the integrity of the DNA was measured using pulsed-field electrophoresis with 0.8% agarose gel. In addition, fresh leaves and other tissues (roots, shoots, young fruits) of JDB cultivar were collected for RNA isolation and transcriptome sequencing. Total RNA was isolated with RNAprep Pure Plant Kit (Tiangen Biotech) according to the manufacturer's instructions. The integrity and quantity of extracted RNA were analyzed on an Agilent 2100 Bioanalyzer. For each tissue, three biological replicates were prepared for sequencing.

**Genome and transcriptome sequencing**

DNA sequencing libraries were constructed and sequenced on the Illumina NovaSeq 6000 platform at 50x depth according to the manufacturer's protocols (Illumina). To generate long-read sequencing reads for *D. longan*, DNA libraries for PacBio SMRT sequencing were prepared following the PacBio standard protocols and sequenced on a Sequel platform. In brief, genomic DNA was randomly sheared to an average size of 20 kb, using a g-Tube (Covaris). The sheared gDNA was end-repaired using polishing enzymes. After purification, a 20-kb insert SMRTbell library was constructed according to the PacBio standard protocol with the BluePippin size-selection system (Sage Science) and sequences were generated on a PacBio Sequel (9 cells) and PacBio RS II (1 cell) platform by Biomarker Technologies. Raw subreads was filtered based on read quality (≥0.8) and read length(≥1000bp). For chromosome-level genome scaffolding, Hi-C libraries were prepared from fresh leaves following protocol previously reported [66]and sequenced on the Illumina HiSeq X Ten platform. DNA was digested with HindIII enzyme, and the ligated DNA was sheared into size of 200-400bp. The resulting libraries was sequenced by using Illumina NovaSeq 6000. For transcriptome sequencing, RNA sequencing (RNA-seq) libraries were constructed using True-Seq kit (Illumina, CA), and sequenced using Illumina HiSeq X Ten platform. Illumina raw reads were trimmed using *Trimmomatic* (v0.39) with parameters "LEADING: 10 TRAILING:10 SLIDINGWINDOW:3:20 MINLEN:36" to remove adapter sequences and low quality reads, yielding a total of ~77.7 Gb clean RNA-seq data from four tissues.

**Genome assembly and evaluation**

To estimate the genome size and heterozygosity level of *D. longan*, cleaned Illumina PE reads were used for k-mer spectrum analysis using *kmergenie*[67] and *GenomeScope* (v2.0)[68] based on 21-mer statistics. PacBio SMRT reads were used for de novo genome assembly by using *Canu* (V1.9)[25] pipeline with parameters "correctedErrorRate=0.045 corMhapSensitivity=normal 'batOptions=-dg 3 -db 3 -dr 1 -ca 500 -cp 50". Alternative haplotig sequences was removed using *purge_dups*[69] according default settings, and only primary contigs were kept for downstream analysis. To correct the base-pair level errors in raw assembly sequences, two rounds of polishing were conducted using high-qualitied Illumina DNA reads with *Pilon* (v1.23)[26]. The longan contigs were further anchored to chromosomes using *Juicer* (v1.5.7)[27] and *3D-DNA*[28] based on Hi-C contact map, followed by manual correction using *Juicerbox* (v1.11.08)[70] to fix assembly errors. The completeness of genome assembly was assessed by BUSCO v1.22[29] using 2121 eudicotyledons_odb10 single copy genes. PacBio sequence reads and Illumina DNA reads were aligned to the genome sequences using *minimap2*[71] and BWA[72] respectively.

**Repetitive element annotation**

We used a combination of the *de novo* repeat library and homology-based strategies to identify repeat structures. *TransposonPSI*[73] was used to identify transposable elements. *GenomeTools* suite[74] (LTR harvest and LTR digest) was used to annotate LTR-RTs with protein HMMs from the Pfam database. Then, a *de novo* repeat library of longan genome was built using *RepeatModeler*[75], and each of the three repeat libraries was classified with *Repeat_Classifier*, followed by removing redundancy using *USEARCH*[76] with ≥ 90% identity threshold. Subsequently, the non-redundant repeat library was analyzed using BLASTx to search the transposase database (evalue=1e-10) and non-redundant plant protein databases (evalue=1e-10) to remove protein-coding genes. Unknown repetitive sequences were further classified using *CENSOR*[77]. Then, the *de novo* repeat library was used to discover and mask the assembled genome with

*RepeatMasker*[78] with the "-xsmall -excln" parameter.

**Prediction and annotation of protein-coding genes**

For gene structure annotations, the RNA-seq data of four different tissues were aligned to repeat-soft masked genome using *STAR*[79], which generates intron hints for gene structure annotation. The structural annotation of protein-coding genes was performed using *BRAKER2*[80], which integrates *GeneMark-ET*[81] and *AUGUSTUS*[82] by combining the aligned results from *ab initio* predictions, homologous protein mapping, and RNA-seq mapping to produce the final gene prediction. The genes with protein length < 120 amino acids and expression level < 0.5 TPM were removed. The tRNA genes were identified by *tRNAscan-SE*[83] with eukaryote parameters. For rRNA, snRNA, miRNA and other non-coding genes prediction, we used *INFERNAL*[84] software by search against Rfam database[85]. The contig-level genome sequences were used to blast against plant plastid database from NCBI (https://ftp.ncbi.nlm.nih.gov/refseq/release/plastid/). The organelle genome sequence identified was submitted to CHLOROBOX[86] website to annotate and visualize. Predicted genes were assigned function by performing BLAST against the NCBI non-redundant protein database with e-value threshold of 1e-10. In addition, a comprehensive annotation was also performed using *InterProScan* (5.36-75.0)[87], which incorporates ProDom[88], PRINTS[89], Pfam[90], SMART[91], SUPERFAMILY[92], PROSITE[93] database. Gene Ontology[94] identifiers for each gene were obtained from the corresponding InterPro entry. *KAAS*[95] and *KOBAS*[96] were used to search the KEGG GENES database for KO (KEGG Ontology) assignments and generating a KEGG pathway membership[97]. The stand-alone version of *plantiSMASH*[98] was utilized to detect plant biosynthetic gene clusters in longan genome.

**Comparative genomics analysis**

Putative orthologship was constructed from two monocots, ten eudicots and *Amborella trichopoda* and longan proteome in this study. Only longest protein sequence was selected as representative of each gene. Orthogroups were inferred by *OrthoFinder* (v2.4.1)[99], as well as a *STAG*[100] species tree rooted using *STRIDE*[101]. The species tree was used as a starting tree to estimate species divergence time using *MCMCTREE* in *paml* (v4.9)[102] package. Speciation event dates for *Ananas comosus*-*Oryza sativa* (1.02~120 MYA), *Populus trichocarpa*-*Ricinus communis* (70~86 MYA), *Arabidopsis thaliana*-*Carica Papaya* (63~82 MYA), and *Glycine max*-*Citrus sinensis* (98~117 MYA), which were obtained using *Timetree* database[103], were used to calibrate the divergence time estimation. We conducted two independent *MCMCTREE* runs using the following settings: burnin = 20000, sampfreq = 30, and nsample = 20000.

The orthologous count table and phylogenetic tree topology inferred from the *OrthoFinder* were taken into *CAFÉ* (v4.2)[104], which employed a random birth and death model to estimate the size of each family at each ancestral node and obtain a family-wise *p*-value to identify whether has a significant expansion or contraction occurred in each gene family across species. Among expanded gene families, longan genes member enriched in IPR002213 (UDP-glucuronosyl/UDP-glucosyltransferase) and IPR036396 (Cytochrome P450 superfamily) and their ortholog CDS sequences of *A.thaliana* and *C. sinensis* genome were retrieved. Genes with protein length <300 amino acids were removed. Multiple sequence alignment was conducted using *MUSCLE* (v3.8.1551)[105] software. *IQtree* was used to constructed a maximum likelihood tree with parameters “-m MF”. Tree file was loaded into the Interactive Tree of Life (iTOL) web server for tree visualization and figure generation[106].

**Transcriptomic analysis**

After removing adapters and trimming low-quality bases, RNA-seq reads were mapped to the longan reference genome using *STAR*[79] with parameters "--alignIntronMax 6000 –align IntronMin 50" and then using *RSEM* tool[107] for transcripts quantification. Outliers among the individual experimental samples were verified based on the Person correlation coefficient, $r^2 \geq 0.85$. Differential expression analysis was performed using *DEseq2*[108] package. Genes were differentially expressed between two conditions if the adjusted p-values was <0.01 and fold change > 1.

## Genetic variation detection

Genome resequencing data were mapped to chromosome-level genome assembly of longan using *BWA-mem*[72]. *Bammarkduplicates* tool in *biobambam*[109] package was used to mark and remove duplicate reads from individual sample alignments. Variant calling was performed using *Freebayes*[110] with parameter "-C 5 --min-alternate-count 5 -g 10000" and then normalized with *VT*[111], filtered using vcffilter from vcflib[112] package with parameters "QUAL / AO > 10 & SAF > 2 & SAR > 2 & RPL > 2 & RPR > 2 & AF > 0.1". Only biallelic variants occurred in more than 90% of individuals were kept and involved in further analysis.

## Population structure and history inference

The vcf-format SNP set were transformed into binary ped-format using VCFtools[113] and *PLINK*[114], and then *smartPCA*[115] was used to conduct PCA analysis based on the data generated in the last step. High-quality SNP data were used to construct individual phylogenetic relationship with SNPhylo[116] package. To estimate individual admixture assuming different numbers of clusters, the population structure and ancestry were investigated using *ADMIXTURE*[117] based on all SNPs. An LD pruning step was performed with Plink[117] with parameters "--indep-pairwise 50 10 0.1". We selected ancestry clusters number ranging from 2 to 4. The population structure result was plotted with script downloaded from https://github.com/speciationgenomics. To study the genetic relationship between longan population from different region, we computed *D*-statistics. The calculation was performed with admixr[118] package in the form of (((Population1, Population2), Population3), HN), where Population1, Population2 and Population3 represented longan from different lineage. Only combinations with absolute *Z*-score value >3 were treated as confidential results. The demographic history of longan was inferred using a hidden Markov model approach as implemented in pairwise sequentially Markovian coalescence[119]. We chose the default PSMC setting "-N25 -t15 -r5 -p 4+25*2+4+6" for all individual. To determine variance in Ne estimates, we performed 100 bootstraps. We scaled results to real time estimates of generation time and mutation rate. We used synonymous substitution rate per synonymous site and dated phylogeny tree as proxies for mutation rate estimation.

Syntenic blocks and reciprocal best hit orthologous pair were identified using *McScanX*[120] with "--full --cscore=.99" parameters. Gene CDS were used as queries to search against the genomes of the other plant genome sequences to find best matching pairs. Given both CDS and protein sequence alignment of each gene pair, *PAL2NAL*(v14)[121] was subsequently used to perform codon alignment and *KaKs_Calculator* 2.0[122] calculate *Ka* and *Ks* value under YN00 model. Gaussian mixture models were fitted to the resulting frequency distribution of *Ks* values by means of function density *Mclust* in the R *mclust* package (v5.3)[123]. The Bayesian information criterion was used to determine the best-fitting model for the data, including the optimal number of Gaussian components as one. The formula r = D/2T, where D is the median of *Ks* value, was used to estimate the neutral mutation rate. Mutation rate of 1.4 x $10^{-8}$

per site per generation, and a constant generation time were assumed in this study to convert coalescence generations into time-scale.

## AUTHOR CONTRIBUTIONS

Project design and oversight: LG, JL and WQ; Sample collection and curation: DG and SH; Conducting experiment and data analysis: JW, ZL and LG; Result interpretation: LG, JL, JW, BL and WQ; Figure and table preparation: LG, JW and LZ; Manuscript writing and revision: LG, JW, LZ, BL and WQ; Provide funding: JL and LG; All authors have read and proved the final version of this manuscript.

## ACKNOWLEDGEMENT

This project is supported by Key-Area Research and Development Program of Guangdong Province (2020B020220006) and Guangdong Provincial Crops Germplasm Nursery Construction and Resources Collection, Preservation, Identification & Evaluation Foundation. In addition, LG is supported by the National Natural Science Foundation of China (31701739 and 31970317) and National Key R&D Program of China (2018YFC0910400). The authors also would like to thank anonymous reviewers for their comments and suggestions to improve this manuscript.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## SUPPLEMENTARY MATERIALS

Supplementary Figure 1: The heatmap of phenylalanine ammonia-lyase genes (PALs) expressed in various longan tissues.

Supplementary Figure 2: The heatmap of peroxidase genes (PODs) expressed in various *Dimocarpus longan* tissues.

Supplementary Figure 3: InterPro protein domain enrichment analysis of *Dimocarpus longan* expanded gene families.

Supplementary Figure 4: The heatmap of UGTs genes expressed in various longan tissues.

Supplementary Figure 5: The heatmap of CYP450 clustered-genes expressed in various longan tissues.

Supplementary Figure 6: Principle component analysis of *Dimocarpus longan* samples based on genotypes.

Supplementary Figure 7: Biogeographical ancestry analysis with group value K.

Supplementary Table 1: Sequencing statistics.

Supplementary Table 2: Summary of Illumina data for genome survey and genome polishing.

Supplementary Table 3: Gene function annotated by different databases.

Supplementary Table 4: Statistics of repetitive elements.

Supplementary Table 5: Comparison of genes in orthogroups between *Dimocarpus longan* and 13 other species.

Supplementary Table 6: List of phenylpropanoid biosynthesis genes and their expression level in different tissues.

Supplementary Table 7: List of different expressed IPR enriched gene families.

Supplementary Table 8: List of UGTs genes ID and their expression level.

Supplementary Table 9: The gene clusters found in *Dimocarpus longan* genome.

Supplementary Table 10: Tissue-specific transcriptome analysis of gene clusters expressed in various longan tissues.

Supplementary Table 11: List of genome resequencing samples and their locations.

## REFERENCES

1. Zhang, Y. F., Lu, B. B., Wang, Y., Pan, L. J., Hu, Y. L., Zhou, J., Zhao, H.Y., Liu, C. M. The chromosomes observation of several rare germplasm in Litchi and Longan. Acta Horticulture Sinica, 2010, 37(12): 1991-1994.
2. Zhang, X. F., Guo, S., Ho, C. T., Bai, N. S. Phytochemical constituents and biological activities of longan (*Dimocarpus longan* Lour.) fruit: a review. Food Science and Human Wellness (2020). doi: https://doi.org/10.1016/j.fshw.2020.03.001.
3. Sun, J. Z., Lin, H. T., Zhang, S., Lin, Y. F., Wang, H., Lin, M. S., Hung, Y. C., Chen, Y. H. The roles of ROS production-scavenging system in Lasiodiplodia theobromae (Pat.) Griff. & Maubl.-induced pericarp browning and disease development of harvested longan fruit. Food Chem., 2018, 247: 16-22.
4. Tang, J. Y., Chen, H. B., Lin, H. T., Hung, Y. C., Xie, H. L., Chen, Y. H. Acidic electrolyzed water treatment delayed fruit disease development of harvested longans through inducing the disease resistance and maintaining the ROS metabolism systems. Postharvest Biology and Technology, 2021, 17: 111349.
5. Altendorf, S. Minor Tropical Fruits: Mainstreaming a Niche Market Food and Agriculture Organization of the United Nations (2018), pp. 67-74.
6. Chen, Y. H., Sun, J. Z., Lin, H. T., Lin, M. S., Lin, Y. F., Wang, H., Hung, Y. C. Salicylic acid reduces the incidence of *Phomopsis longanae* Chi infection in harvested longan fruit by affecting the energy status and respiratory metabolism. Postharvest Biol. Technol., 2020, 160: 111035.
7. Zheng, S. Q., Wei, X.Q., Jiang, J.M., Jiang, F., Huang, A.P. Actual state and corresponding strategy on longan breeding in China. Fujian Fruits, 2010,4 : 35-40.
8. Zeven, A. C., Zhukovsky, P. M. Dictionary of cultivated plants and their centres of diversity. Centre for Agricultural Publishing and Documentation (PUDOC), Wageningen, The Netherlands, 1975.
9. Wu, Z. Y. Flora Yunnanica, 1977, Vol. 1-21. (Science Press, 1977- 2006).
10. Zhong Y. Fruit tree resources and its geographical distribution in Hainan Island. Horticultural Plant Journal, 1983, 10(3): 145-152.
11. Anupunt, P., Sukhvibul, N. Lychee and longan production in Thailand. Acta Hort, 2005, 665: 53-59.
12. Menzel, C. M., Waite, G. K. Litchi and longan: botany, production. and uses. Trowbridge: Cromwell Press, 2005.
13. Ke, G. W., Wang, C. C., Tang, Z. F. Palynological studies on the origin of longan cultivation. Horticultural Plant Journal, 1994, 4: 323-328.

14. Zhu, J. H., Pan, L. M., Qin, X. Q., Peng, H. X., Wang, Y., Han, Z. H. Analysis on genetic relations in different ecotypes of Longan (*Dimocarpus longan* Lour.) germplasm resources by ISSR markers. Journal of Plant Genetic Resources, 2013, 14(1): 65-69.
15. Yi, G. J., Tan, W. P, Huo, H. Q., Zhang, Q. M., Li, J. G., Zhou, B. R. Studies on the genetic diversity and relationship of longan cultivars by AFLP analysis. Acta Horticulture Sinica, 2003, 30(3): 272-276.
16. Zhong, F. L., Pan, D. M., Guo, Z. X., Lin, L., Li, K.T. RAPD Analysis of longan germplasm resources. Chinese Agricultural Science Bulletin, 2007, 23(7): 558-563.
17. Hu, W. S., Huang, A. P, Jiang, F., Jiang, J. M., Chen, X. P., Zheng, S. Q. Identification and genetic diversity of reciprocal hybrids in longan (*Dimocarpus longan*) by SSR. Acta Horticulturae Sinica, 2015, 42(10): 1899-1908.
18. Zheng, S., Zeng, L., Zhang, J., Lin, H., Deng, C., Zhuang, Y. Fruit scientific research in New China in the past 70 years: longan. J. Fruit Sci, 2019, 36: 1414-1420.
19. Jue, D., Sang, X., Liu, L., Shu, B., Wang, Y., Liu, C., Wang, Y., Xie, J., Shi, S. Comprehensive analysis of the longan transcriptome reveals distinct regulatory programs during the floral transition. BMC Genomics, 2019, 20: 126.
20. Zhu, J. H, Xu, N., Qin, X. Q., Li, D. B., Huang, F. Z., Li, H. L, Lu, G. F., Peng, H. X. Longan sexual hybridization technique. Southern Horticulture, 2014, 6(25): 48-49.
21. Sun, R., Chang, Y., Yang, F., Wang, Y., Li, H., Zhao, Y., Chen, D., Wu, T., Zhang, X., Han, Z. A dense SNP genetic map constructed using restriction site-associated DNA sequencing enables detection of QTLs controlling apple fruit quality. BMC Genomics, 2015, 16: 747-747.
22. Zhang, Q., Wei, X., Liu, N., Zhang, Y., Xu, M., Zhang, Y., Ma, X., Liu, W. Construction of an SNP-based high-density genetic map for Japanese plum in a Chinese population using specific length fragment sequencing. Tree Genet. Genomes, 2020, 16: 18.
23. Guo, Y.S., Zhao, Y., Liu, C. QTLs analysis of several traits in Longan. Biotechnol & Biotechnol. Equip, 2011, 25: 2203-2209.
24. Lin, Y. L., Min, J. M., Lai, R. L., Wu, Z. Y., Chen, Y. K., Yu, L. L., Lai, Z. X. Genome-wide sequencing of longan (*Dimocarpus longan* Lour.) provides insights into molecular basis of its polyphenol-rich characteristics. Gigascience, 2017, 6(5): 1-14.

25. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res, 2017, 27(5): 722-736.

26. Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., Earl, A. M. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One, 2014, 19; 9(11): e112963.

27. Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., Aiden, E. L. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst, 2016, 3(1): 95-98.

28. Dudchenko, O., Batra, S. S., Omer, A. D, Nyquist, S. K, Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., Aiden, E. L. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science, 2017, 356(6333): 92-95.

29. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 2015, 31, 3210-3212.

30. Chen, F. C, Chen, C. J., Li, W. H., Chuang, T. J. Gene family size conservation is a good indicator of evolutionary rates. Molecular Biology and Evolution, 2010, 27(8): 1750-1758.

31. Wang, J., Guo, D. L., Han, D. M., Pan, X. W., Li, J. G. A comprehensive insight into the metabolic landscape of fruit pulp, peel, and seed in two longan (*Dimocarpus longan* Lour.) varieties. Int J Food Prop, 2020, 23(1): 1527-1539.

32. Butelli, E., Titta, L., Giorgio, M., Mock, H. P., Matros, A., Peterek, S., Schijlen, E. G, Hall, R. D., Bovy, A. G., Luo, J. Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors. Nat. Biotechnol, 2008, 26, 1301.

33. Luo, C., Zou, X., Li, Y., Sun, C., Jiang, Y., Wu, Z. Determination of flavonoids in propolis-rich functional foods by reversed phase high performance liquid chromatography with diode array detection. Food Chem, 2011, 127: 314-320.

34. Gray, J., Caparrós-Ruiz, D., Grotewold, E. Grass phenylpropanoids: regulate before using! Plant Sci, 2012, 184: 112-120.

35. Dong, N. Q., Lin, H. X. Contribution of phenylpropanoid metabolism to plant development and plant-environment interactions. J Integr Plant Biol, 2021, 63(1): 180-209.

36. Neutelings, G. Lignin variability in plant cell walls: contribution of new models. Plant Sci, 2011, 181(4): 379-386.
37. Purwar, S., Gupta, S. M., Kumar, A. Enzymes of phenylpropanoid metabolism involved in strengthening the structural barrier for providing genotype and stage dependent resistance to karnal bunt in wheat. American Journal of Plant Sciences, 2012, 3: 261-267.
38. Zhao, S., Zhao, L., Liu, F., Wu, Y., Zhu, Z., Sun, C., and Tan, L. NARROW AND ROLLED LEAF 2 regulates leaf shape, male fertility, and seed size in rice. J. Integr. Plant Biol, 2016, 58: 983-996.
39. Mohammadi, M., Kazemi, H. Changes in peroxidase and polyphenol oxidase activities in susceptible and resistant wheat heads inoculated with *Fusarum graminearum* and induced resistance. Plant Sci, 2002, 162: 491-498.
40. Yuan, L., and Grotewold, E. Plant specialized metabolism. Plant Sci, 2020, 298: 110579.
41. Le Roy, J., Huss, B., Creach, A., Hawkins, S., and Neutelings, G. Glycosylation is a major regulator of phenylpropanoid availability and biological activity in plants. Front Plant Sci, 2016, 7: 735.
42. Li, Y., Baldauf, S., Lim, E. K., Bowles, D. J. Phylogenetic analysis of the UDP-glycosyltransferase multigene family of Arabidopsis thaliana. J Biol Chem. 2001, 276(6): 4338-4343.
43. Ross, J., Li, Y., Lim, E. K., Bowles, D. J. Higher plant glycosyltransferases. Genome Biol, 2001, 2(2): 1-6.
44. Aksamit-Stachurska, A., Korobczak-Sosna, A., Kulma, A., Szopa, J. Glycosyltransferase efficiently controls phenylpropanoid pathway. BMC Biotechnol, 2008, 5, 8: 25.
45. Koeduka, T., Ueyama, Y., Kitajima, S., Ohnishi, T., Matsui, K. Molecular cloning and characterization of UDP-glucose: Volatile benzenoid/phenylpropanoid glucosyltransferase in petunia flowers. J Plant Physiol, 2020, 252:153245.
46. Wu, B., Cao, X., Liu, H., Zhu, C., Klee, H., Zhang, B., Chen, K. UDP-glucosyltransferase PpUGT85A2 controls volatile glycosylation in peach. J Exp Bot, 2019, 70: 925-936.
47. Yamada, A., Ishiuchi, K., Makino, T., Mizukami, H., Terasaka, K. A glucosyltransferase specific for 4-hydroxy-2,5-dimethyl-3(2H)-furanone in strawberry. Biosci Biotechnol Biochem, 2018, 29: 1-8.
48. Caputi, L., Malnoy, M., Goremykin, V., Nikiforova, S., Martens, S. A genome-wide phylogenetic reconstruction of family 1 UDP-glycosyltransferases revealed

the expansion of the family during the adaptation of plants to life on land. Plant J. 2012, 69(6): 1030-1042.

49. Wu, B., Gao, L., Gao, J., Xu, Y., Liu, H., Cao, X., Zhang, B., Chen, K. Genome - wide identification, expression patterns, and functional analysis of UDP Glycosyltransferase family in peach (*Prunus persica* L. Batsch). Front Plant Sci, 2017, 8: 389.

50. Wu, B., Liu, X. H., Xu, K., Zhang, B. Genome-wide characterization, evolution and expression profiling of UDP-glycosyltransferase family in pomelo (*Citrus grandis*) fruit. BMC Plant Biology, 2020, 20: 459.

51. von Saint Paul, V., Zhang, W., Kanawati, B., Geist, B., Faus-Keßler, T., Sch-mitt-Kopplin, P., Schäffner, A. R. The Arabidopsis glucosyltransferase UGT76B1 conjugates isoleucic acid and modulates plant defense and senescence. Plant Cell, 2011, 23: 4124-4145.

52. Huang, X.X., Zhu, G.Q., Liu, Q., Chen, L., Li, Y.J., Hou, B.K. Modulation of plant salicylic acid-associated immune responses via glycosylation of dihydroxybenzoic acids. Plant Physiology, 2018, 176, 3103-3119.

53. Dixon, R.A., Achnine, L., Kota, P., Liu, C.J., Reddy, M.S., Wang, L. The phenylpropanoid pathway and plant defence-a genomics perspective. Molecular Plant Pathology, 2002, 3: 371-390.

54. Vogt, T. Phenylpropanoid Biosynthesis. Molecular Plant, 2010, 3, 2-20.

55. Huang, X. X., Wang, Y., Lin, J. S., Chen, L., Li, Y. J., Liu, Q., Wang, G. F., Xu, F., Liu, L., Hou, B. K. The novel pathogen-responsive glycosyltransferase UGT73C7 mediates the redirection of phenylpropanoid metabolism and promotes SNC1-dependent Arabidopsis immunity. Plant J, 2021, 18.

56. Dong, N. Q., Sun, Y., Guo, T., Shi, C. L., Zhang, Y. M., Kan, Y., Xiang, Y. H., Zhang, H., Yang, Y. B., Li, Y. C., Zhao, H. Y., Yu, H. X., Lu, Z. Q., Wang, Y., Ye, W. W, Shan, J. X., Lin, H. X. UDP-glucosyltransferase regulates grain size and abiotic stress tolerance associated with metabolic flux redirection in rice. Nat Commun, 2020; 11(1): 2629.

57. Zheng, X., Li, P., Lu, X. Research advances in cytochrome P450-catalysed pharmaceutical terpenoid biosynthesis in plants. J Exp Bot, 2019, 70 (18): 4619-4630.

58. Morant, M., Bak, S., Møller, B. L., Werck-Reichhart, D. Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation. Curr Opin Biotechnol, 2003; 14(2): 151-162.

59. Cheng, Y., Liu, H., Tong, X. J., Liu, Z. M., Zhang, X., Li, D. L., Jiang, X. M. and Yu, X. H. Identification and analysis of CYP450 and UGT supergene family members from the transcriptome of Aralia elata (Miq.) seem reveal candidate genes for triterpenoid saponin biosynthesis. BMC Plant Biology volume, 2020, 20: 214.

60. Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S., Werck-Reichhart, D. Cytochromes p450. Arabidopsis Book, 2011, 9: e0144.

61. Jiu, S., Xu, Y., Wang, J., Wang, L., Liu, X., Sun, W., Sabir, I. A., Ma, C., Xu, W., Wang, S., Abdullah, M., Zhang, C. The Cytochrome P450 Monooxygenase Inventory of Grapevine (*Vitis vinifera* L.): Genome-Wide Identification, Evolutionary Characterization and Expression Analysis. Front Genet, 2020, 11: 44.

62. Baldwin, I. T. Plant volatiles. Current Biology, 2010, 20: 392-397.

63. Puspita, R.， Bintang, M.， Priosoeryanto, B.P. Antiproliferative activity of longan (*Dimocarpus longan* Lour.) leaf extracts. Journal of Applied Pharmaceutical Science, 2019, 9(05):102-106.

64. Xu, M., Galhano, R., Wiemann, P., Bueno, E., Tiernan, M., Wu, W., Chung, I. M., Gershenzon, J., Tudzynski, B., Sesma, A., Peters, R. J. Genetic evidence for natural product-mediated plant-plant allelopathy in rice (*Oryza sativa*). New Phytol, 2012, 193(3): 570-575.

65. Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S. & Thompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc, 2006, 1(5):2320-2325.

66. Li, Y., Liu, G. F., Ma, L. M., Liu, T. K., Zhang, C. W., Xiao, D., Zheng, H. K., Chen, F., Hou, X. L. A chromosome-level reference genome of non-heading Chinese cabbage [Brassica campestris (syn. Brassica rapa) ssp. chinensis]. Hortic Res, 2020, 7(1): 212.

67. Chikhi, R., Medvedev, P. Informed and automated k-mer size selection for genome assembly. Bioinformatics, 2014, 30(1): 31-37.

68. Ranallo-Benavidez, T. R., Jaron, K. S., Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun, 2020, 11(1): 1432.

69. Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., Durbin, R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics, 2020, 36(9): 2896-2898.

70. Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., Aiden, E. L. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst, 2016, 3(1): 99-101.
71. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 2018, 34(18): 3094-3100.
72. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009, 25(14): 1754-1760.
73. Hass, B. Transposon PSI: An application of PSI-Blast to mine (retro-) transposon ORF homologies. Broad Institute, Cambridge, MA, USA (2010).
74. Gremme, G., Steinbiss, S., Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform, 2013, 10(3): 645-656.
75. Smit, A., Hubley, R. RepeatModeler open-1.0. Available at http://www.repeatmasker.org (2015).
76. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 2010, 26(19): 2460-2461.
77. Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics, 2006, 7: 474.
78. Smit, A., Hubley, R., Green, P. RepeatMasker Open-4.0. 2013-2015. http://www.repeatmasker.org (2013).
79. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013, 29(1): 15-21.
80. Hoff, K. J., Lomsadze, A., Borodovsky, M., Stanke, M. Whole-Genome Annotation with BRAKER. Methods Mol Biol, 2019, 1962: 65-95.
81. Lomsadze, A., Burns, P. D., Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res, 2014, 42(15): e119.
82. 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-W439.
83. Lowe, T. M., Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucleic Acids Res, 1997, 25(5): 955-964.

84. Nawrocki, E. P., Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 2013, 29(22): 2933-2935.

85. Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D., Petrov, A. I. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res, 2018, 46(D1): D335-D342.

86. Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., Greiner, S. GeSeq-versatile and accurate annotation of organelle genomes. Nucleic Acids Res, 2017,45(W1): W6-W11.

87. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R. InterProScan: protein domains identifier. Nucleic Acids Res, 2005, 33(Web Server issue): W116-120.

88. Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., Kahn, D. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res, 2005, 33 (Database issue): D212-D215.

89. Attwood, T. K., Croning, M. D., Flower, D. R., Lewis, A. P., Mabey, J. E., Scordis, P., Selley, J. N., Wright, W. PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Res. 2000, 28(1): 225-227.

90. El-Gebali, S. *et al*. The Pfam protein families database in 2019. Nucleic Acids Res. 2019, 47(D1): D427-D432.

91. Letunic, I., Doerks, T., Bork, P. SMART 7: Recent updates to the protein domain annotation resource. Nucleic Acids Res. 2012, 40 (Database issue): D302-D305.

92. Gough, J., Karplus, K., Hughey, R., Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol, 2001, 313(4): 903-19.

93. Sigrist, C. J., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., Xenarios, I. New and continuing developments at PROSITE. Nucleic Acids Res. 2013, 41(Database issue): D344-347.

94. Boccacci, P. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, 25(1): 25-29.

95. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., Kanehisa, M. KAAS: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Res, 2007, 35(Web Server issue): W182-185.

96. Xie, C. *et al*. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011, 39 (Web Server issue): W316-W322.

97. Kanehisa, M., Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res, 2000, 28(1): 27-30.

98. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., Medema, M. H. PlantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res, 2017, 45(W1): W55-W63.

99. Emms, D. M., Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol, 2019, 20(1): 238.

100.Emms, D. M. STAG: Species Tree Inference from All Genes. bioRxiv (2018) doi:10.1101/267914.

101.Emms, D. M., Kelly, S. STRIDE: Species tree root inference from gene duplication events. Mol Biol Evol, 2017, 34(12):3267-3278.

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

103.Kumar, S., Stecher, G., Suleski, M., Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol, 2017, 34(7): 1812-1819.

104.Han, M. V., Thomas, G. W. C., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol, 2013, 30(8): 1987-1997.

105.Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 2004, 32(5): 1792-1797.

106.Letunic, I., Bork, P. Interactive Tree of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res, 2019, 47(W1): W256-W259.

107.Li, B., Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 2011, 12: 323.

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

109.Tischler, G.， Leonard, S. Biobambam: Tools for read pair collation based algorithms on BAM files. Source Code for Biology and Medicine (2014) doi:10.1186/1751-0473-9-13.

110.Garrison, E., Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv, 2012, 1207, 3907.

111.Tan, A., Abecasis, G. R., Kang, H. M. Unified representation of genetic variants. Bioinformatics, 2015, 31(13):2202-4.

112.Garrison, E. Vcflib: A C++ library for parsing and manipulating VCF files. GitHub (2012).

113.Danecek, P. *et al*. The variant call format and VCFtools. Bioinformatics, 2011, 27(15):2156-8.

114.Steimle, J., Weibel, N., Olberding, S., Mühlhäuser, M. & Hollan, J. D. PLink: paper-based links for cross-media information spaces. (2011) doi:10.1145/1979742.1979885.

115.Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. PLoS Genet, 2006, 2(12): e190.

116.Lee, T. H., Guo, H., Wang, X., Kim, C., Paterson, A. H. SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics, 2014, 15:162.

117.Zhou, H., Alexander, D., Lange, K. A quasi-Newton acceleration for high-dimensional optimization algorithms. Stat Comput, 2011, 21(2): 261-273.

118.Petr, M., Vernot, B., Kelso, J. Admixr-R package for reproducible analyses using ADMIXTOOLS. Bioinformatics, 2019, 35(17): 3194-3195.

119.Li, H., Durbin, R. Inference of human population history from individual whole-genome sequences. Nature, 2011, 475(7357): 493-496.

120.Wang, Y. *et al*. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res, 2012, 40(7): e49.

121.Suyama, M., Torrents, D., Bork, P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res, 2006, 34 (Web Server issue): W609-612.

122.Wang, D., Zhang, Y., Zhang, Z., Zhu, J., Yu, J. KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. Genomics Proteomics Bioinformatics, 2010, 8(1): 77-80.

123.Scrucca, L., Fop, M., Murphy, T. B., Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J. 2016, 8(1): 289-317.