






## OPEN

# The draft genome of tropical fruit durian (*Durio zibethinus*)

Bin Tean Teh<sup>1–6,18</sup> , Kevin Lim<sup>2,7,18</sup> , Chern Han Yong<sup>2,7,18</sup>, Cedric Chuan Young Ng<sup>3,18</sup>, Sushma Ramesh Rao<sup>8–11</sup>, Vikneswari Rajasegaran<sup>3</sup>, Weng Khong Lim<sup>2,4,7</sup> , Choon Kiat Ong<sup>12</sup> , Ki Chan<sup>13</sup>, Vincent Kin Yuen Cheng<sup>14</sup>, Poh Sheng Soh<sup>15</sup>, Sanjay Swarup<sup>8–11,16</sup>, Steven G Rozen<sup>2,4,7</sup> , Niranjana Nagarajan<sup>17</sup> & Patrick Tan<sup>1,2,4,5,17</sup>

**Durian (*Durio zibethinus*) is a Southeast Asian tropical plant known for its hefty, spine-covered fruit and sulfury and onion-like odor. Here we present a draft genome assembly of *D. zibethinus*, representing the third plant genus in the Malvales order and first in the Helicteroideae subfamily to be sequenced. Single-molecule sequencing and chromosome contact maps enabled assembly of the highly heterozygous durian genome at chromosome-scale resolution. Transcriptomic analysis showed upregulation of sulfur-, ethylene-, and lipid-related pathways in durian fruits. We observed paleopolyploidization events shared by durian and cotton and durian-specific gene expansions in *MGL* (methionine  $\gamma$ -lyase), associated with production of volatile sulfur compounds (VSCs). *MGL* and the ethylene-related gene *ACS* (aminocyclopropane-1-carboxylic acid synthase) were upregulated in fruits concomitantly with their downstream metabolites (VSCs and ethylene), suggesting a potential association between ethylene biosynthesis and methionine regeneration via the Yang cycle. The durian genome provides a resource for tropical fruit biology and agronomy.**

Durian (*D. zibethinus*) is an edible tropical fruit endemic to Southeast Asia. In the region, durian is known as the ‘king of fruits’ for its formidable spiny husk, overpowering flavor, and unique odor, described as an onion-like, sulfury aroma with notes of sweet fruitiness and savory soup seasoning<sup>1</sup>. Durian can elicit opposing opinions, from devotion to revulsion; the famed naturalist Alfred Russel Wallace once remarked on durian: “the more you eat of it, the less you feel inclined to stop” (ref. 2). In contrast, durian is banned from public transportation and many hotels because of its characteristic and pungent smell, described by some detractors as “turpentine and onions, garnished with a gym sock” (ref. 3).

Among the 30 known species in the *Durio* genus, *D. zibethinus* is the most prized as a major Southeast Asian food crop. The three leading durian-producing countries are Thailand, Malaysia, and Indonesia, with more than 250,000 ha cultivated in 2008 (ref. 4). Durian is also of major economic value, as it has recently gained market penetrance in China: in 2016 alone, durian imports into China accounted for about \$600 million, as compared to \$200 million for oranges, one of China’s other main fruit imports (UN Trade Statistics; see URLs). More than 200 different cultivars of durian exist, encompassing a range of fruit textures, flavors, and aromas. Distinct regional demands for different

cultivars reflect local idiosyncrasies in consumer tastes: pungent and bitter varieties are prized in Malaysia and Singapore (for example, Musang King), whereas sweeter cultivars with a mild odor are popular in Thailand (for example, Monthong). The distinctive odors of different durian cultivars have also been biochemically studied and characterized as a complex suite of odor-active compounds including sulfur volatiles, esters, alcohols, and acids<sup>4–6</sup>.

Despite the importance of durian as a tropical fruit crop, durian-related genetic research is almost nonexistent. Important resources such as genetic maps are not publicly available for durian, and no plant species in the Helicteroideae subfamily have had their genomes sequenced and assembled. Relatives whose genomes are available include only the more distantly related cash crops within the larger Malvaceae family: *Theobroma cacao* (cacao; used in chocolate) and members of the *Gossypium* genus (cotton). Besides *D. zibethinus*, several other *Durio* species can produce edible fruits (for example, *Durio graveolens* and *Durio testudinarum*), occupying specific ecological niches where they have relationships with specific pollinators, primarily fruit bats and birds<sup>1,7</sup>. However, many of the other *Durio* species are currently listed as vulnerable or endangered.

<sup>1</sup>Thorn Biosystems Pte Ltd, Singapore. <sup>2</sup>Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore. <sup>3</sup>Laboratory of Cancer Epigenome, Division of Medical Science, National Cancer Centre Singapore, Singapore. <sup>4</sup>SingHealth/Duke-NUS Institute of Precision Medicine, National Heart Centre, Singapore. <sup>5</sup>Cancer Science Institute of Singapore, National University of Singapore, Singapore. <sup>6</sup>Institute of Molecular and Cellular Biology, Singapore. <sup>7</sup>Centre for Computational Biology, Duke-NUS Medical School, Singapore. <sup>8</sup>Department of Biological Sciences, National University of Singapore, Singapore. <sup>9</sup>Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore. <sup>10</sup>Metabolites Biology Laboratory, National University of Singapore, Singapore. <sup>11</sup>NUS Synthetic Biology for Clinical and Technological Innovation, Life Sciences Institute, National University of Singapore, Singapore. <sup>12</sup>Lymphoma Genomic Translational Research Laboratory, National Cancer Centre, Singapore. <sup>13</sup>Global Databank, Singapore. <sup>14</sup>Verdant Foundation, Hong Kong. <sup>15</sup>Samsoney Group, Johor Bahru, Malaysia. <sup>16</sup>NUS Environmental Research Institute, National University of Singapore, Singapore. <sup>17</sup>Genome Institute of Singapore, Singapore. <sup>18</sup>These authors contributed equally to this work. Correspondence should be addressed to B.T.T. (te.hin.tean@singhealth.com.sg) or P.T. (gmstanp@duke-nus.edu.sg).

Here we report a draft whole-genome assembly of the *D. zibethinus* Musang King cultivar, employing a combination of sequencing technologies. We used the durian genome to assess phylogenetic relationships with other Malvaceae members and to compare transcriptome data between different plant organs (fruit aril, leaf, stem, and root) and between fruit arils from different cultivars. Genomic and transcriptomic analyses provided insights into the evolution and regulation of fruit-related processes in durian, including ripening, flavonoid production, and sulfur metabolism. The durian genome provides a valuable resource for biological research and agronomy of this tropical fruit.

RESULTS

Genome assembly

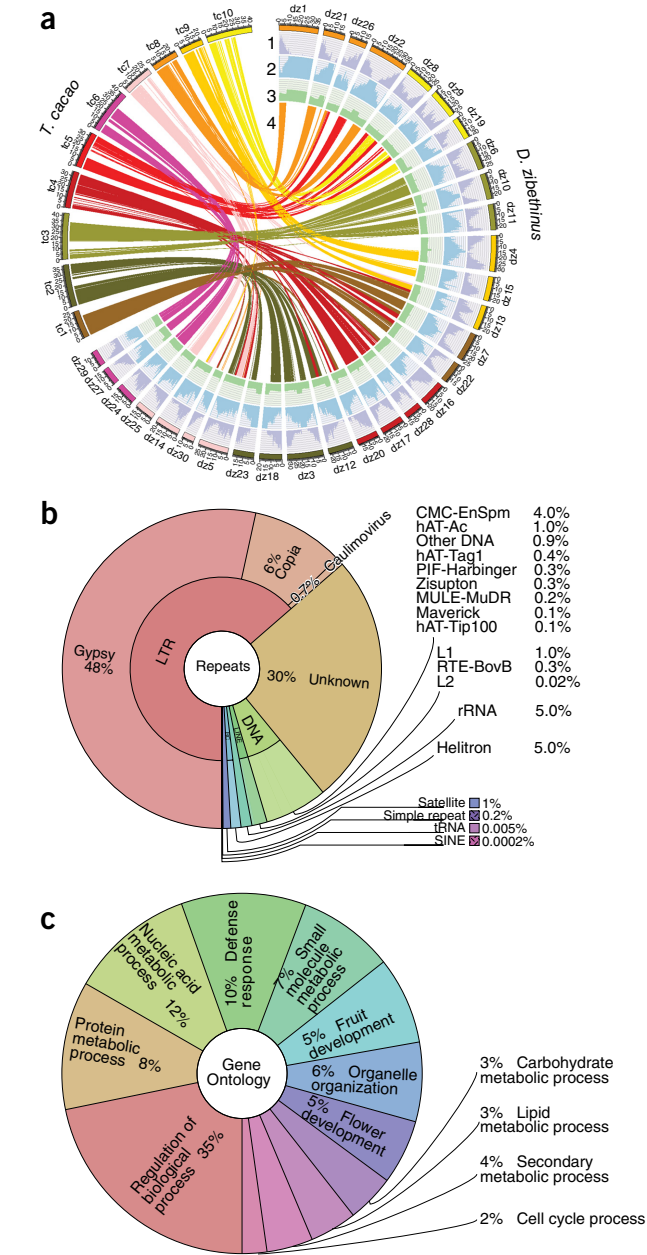
We extracted DNA from a *D. zibethinus* Musang King durian fruit stalk, generating 18.3 million PacBio single-molecule long reads (average read length of 6.2 kb) corresponding to ~153× coverage of the ~738-Mb durian genome whose size was estimated by *k*-mer distribution analysis (Supplementary Fig. 1 and Supplementary Table 1). We performed a PacBio-only assembly using an overlap-layout-consensus method implemented in FALCON<sup>8</sup>; as the durian genome is highly heterozygous, possibly owing to frequent outcrossing during cultivation, we additionally phased the contigs using diploid-aware FALCON-Unzip<sup>9</sup>. The subsequent haplotig-merged assembly was polished using Arrow over three iterations. For details on validation of the genome assembly, see the Supplementary Note.

The durian assembly was further refined in two stages using chromosome contact maps. CHiCAGO (*in vitro* chromatin reconstitution of high-molecular-weight DNA) and Hi-C (*in vivo* fixation of chromosomes) libraries<sup>10,11</sup> (~280 million read pairs each) were used to improve scaffold N50 to 22.7 Mb, with the longest scaffold being 36.3 Mb (Fig. 1a; GC content of 32.5%). The final reference assembly comprised chromosome-scale pseudomolecules, with 30 pseudomolecules greater than 10 Mb in length and covering 95% of the 712-Mb assembly (the pseudomolecules are hereafter referred to as chromosomes, numbered according to size; Table 1). These figures correspond closely to previous estimates of the haploid chromosome number of durian ( $1n = 28$ ,  $2n = 56$ )<sup>12</sup>.

We experimentally estimated the size of the durian genome to be about 800 Mb by flow cytometry, close to the size from our *k*-mer analysis (738 Mb; Supplementary Table 2). *k*-mer distributions showed two distinct peaks, indicative of either a diploid genome or autotetraploid genome with low heterozygosity (Supplementary Fig. 1). Mapping of syntenic regions within the assembly showed extensive shuffling of syntenic regions rather than the one-to-one correspondence expected for an autotetraploid genome. SNP calling on the final assembly yielded a heterozygosity rate of 1.14%, supporting the estimate from *k*-mer analysis (1%) and consistent with durian being a highly heterozygous organism.

We identified and masked 54.8% of the assembly as repeat regions (Fig. 1a). LTR/Gypsy repeats were the most abundant, making up 26.2% of the genome, followed by LTR/Copia elements (3.2%; Fig. 1b and Supplementary Table 3). In comparison to other genomes in the Malvaceae family, LTR/Gypsy families of repeats appear to have expanded in *D. zibethinus* (26.2% of the genome), *Gossypium raimondii* (33.1%), and *Gossypium arboreum* (55.8%) as compared to *T. cacao* (9%). Conversely, LTR/Copia repeats appear to have contracted in *D. zibethinus* (3.2% of the genome) and *G. arboreum* (5.5%) as compared to *T. cacao* (7%) and *G. raimondii* (11.1%).

We annotated the remaining unmasked *D. zibethinus* genome using a comprehensive strategy combining evidence-based and *ab initio*



**Figure 1** Characterization of the *D. zibethinus* genome. (a) Circos plot of the multidimensional topography of the *D. zibethinus* genome (right), comprising 30 pseudomolecules that cover ~95% of the assembly. Concentric circles, from outermost to innermost, show (1) gene density, (2) repeat element density, (3) GC content, and (4) syntenic regions with *T. cacao* (left), the closest sequenced relative in the Malvaceae family that did not undergo a recent WGD event. (b) Distribution of repeat classes in the durian genome. (c) Distribution of predicted genes among different high-level Gene Ontology (GO) biological process terms.

gene prediction (Fig. 1a). Using the Maker pipeline<sup>13</sup>, we incorporated 151,593 protein sequences from four plant species and 43,129 transcripts assembled from *D. zibethinus* Musang King RNA-seq data. 45,335 gene models were identified, with an average coding-sequence length of 1.7 kb and an average of 5.8 exons per gene. The vast majority of gene predictions (42,747) were supported by homology to known proteins, existence of known functional domains, or the presence of expressed transcripts (Supplementary Fig. 2). Using Blast2GO<sup>14</sup>, we

**Table 1 Statistics for the *D. zibethinus* draft genome**

Assembly feature	Statistic
Estimated genome size (by <i>k</i> -mer analysis)	738 Mb
Number of scaffolds	677
Scaffold N50	22.7 Mb
Longest scaffold	36.3 Mb
Assembly length	715 Mb
Assembly % of genome	96.88
Repeat region % of assembly	54.8
Predicted gene models	45,335
Average coding sequence length	1,700.4 bp
Average exons per gene	5.8

annotated 35,975 predicted gene models with Gene Ontology (GO) terms. The annotated gene models were involved in processes such as defense response, fruit development, and carbohydrate and lipid metabolism (Fig. 1c), which may be of interest in the study of genetic variation between domesticated cultivars.

To verify the sensitivity of our gene predictions and the completeness and proper haplotig merging of our assembly, we checked core gene statistics using BUSCO<sup>15</sup>. Our gene predictions recovered 1,300 of the 1,440 (90.3%) highly conserved core proteins in the Embryophyta lineage, of which 68.1% were single-copy genes and 22.2% were duplicated. We checked whether this high duplication rate (22.2%) indicated unmerged haplotigs in our assembly, using coverage statistics from the Illumina short reads. Among these duplicated genes, 93.1% had mean read coverage within 1 s.d. of the mean read coverage for single-copy core genes, showing that these duplicated genes likely exist as independent and distinct copies in the genome (Supplementary Fig. 3). The high number of independent duplicate genes is suggestive of a recent whole-genome duplication (WGD) event in the durian lineage, similar to the scenarios in *G. raimondii* (11.5% duplicate genes) and *G. arboreum* (12.2%), both of which had recent WGDs, and distinct from the scenario in *T. cacao* (1.2%), which has not undergone a recent WGD. Visualization of syntenic blocks between durian and cacao confirmed that chromosome-level blocks in cacao are duplicated across multiple durian chromosomes (Fig. 1a).

### Comparative phylogenomics shows durian paleopolyploidy

To investigate the evolution of distinct durian traits, we compared the durian genome to ten other sequenced plant species. These included three plants in the same Malvales order (*G. raimondii*, *G. arboreum*, and *T. cacao*), six plants in the same Eudicots clade (*Arabidopsis thaliana*, *Populus trichocarpa* (poplar), *Glycine max* (soybean), *Vitis vinifera* (grape), *Coffea canephora* (coffee), and *Carica papaya* (papaya)), and *Oryza sativa japonica* (rice) as an outgroup.

Gene family clustering with OrthoMCL<sup>16</sup> identified 32,159 gene families consisting of 327,136 genes (Fig. 2a; for clarity, only durian, *T. cacao*, *G. arboreum*, *G. raimondii*, and *A. thaliana* gene families are shown). 607 gene families, consisting of 1,764 genes, were unique to durian (Supplementary Table 4). Durian shared the most gene families with the other Malvales plants (cotton and cacao), consistent with the placement of these three species in the same taxonomic order. Durian, cotton, and cacao also had similar numbers of gene families as *Arabidopsis* (within 99% of each other), further validating the accuracy and completeness of our gene predictions at the gene family level.

We derived 47 single-copy orthologous genes among the 11 plant species for phylogenetic analysis, taking the intersection of 344 single-copy gene families found by OrthoMCL and 395 single-copy gene families present across diverse plant species<sup>17</sup> (Supplementary Table 5).

We used BEAST2 (ref. 18) to generate and date phylogenetic trees on the basis of Bayesian analysis. Consistent with the phylogenetic ordering of durian, cotton, and cacao first proposed by Alverson *et al.*<sup>19</sup>, our results suggest that cacao first diverged from the shared durian–cotton lineage 62–85 million years ago (95% highest posterior density (HPD) interval), followed by the divergence of durian and cotton 60–77 million years ago (Fig. 2b). For the other plant species, phylogenetic placement and estimated speciation dates were mostly in broad consensus with other studies<sup>20,21</sup>.

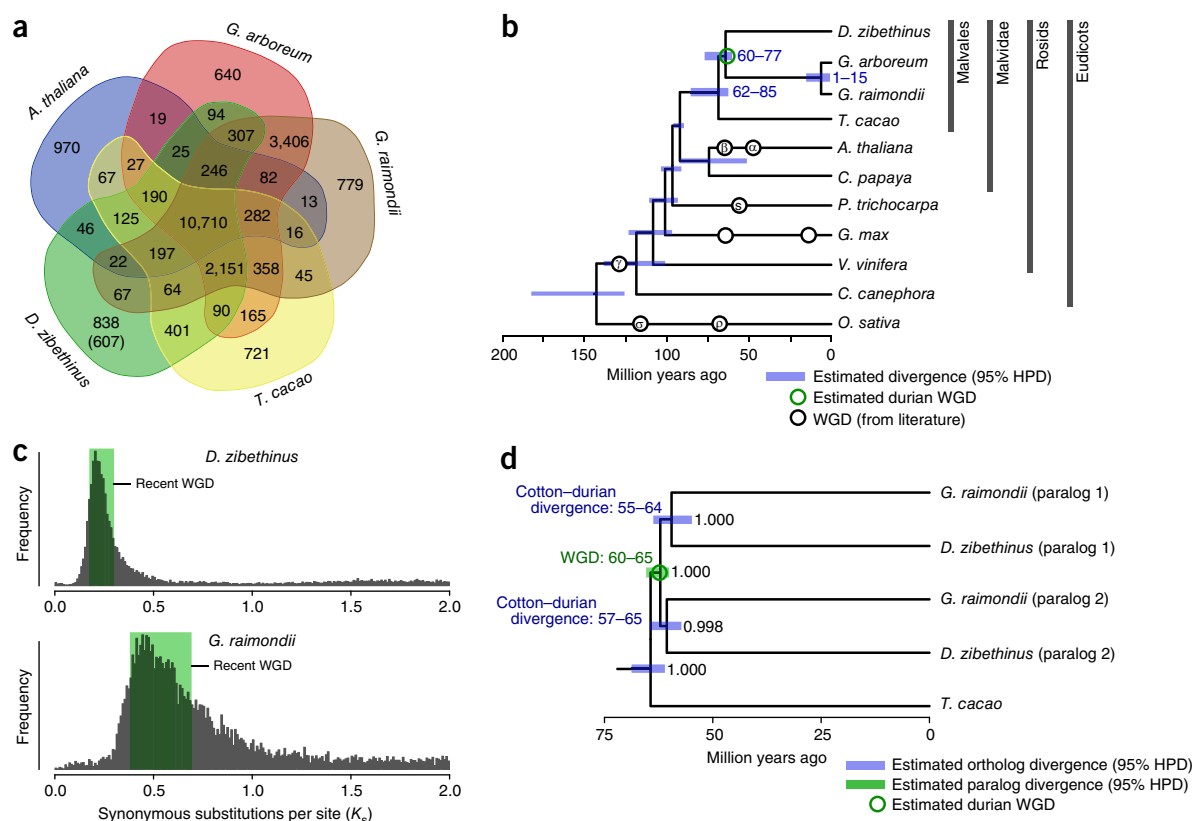
Ancient WGDs (also known as paleopolyploidization events) are widespread in plant lineages and represent a powerful evolutionary force for the development of novel gene functions and the emergence of new species<sup>22</sup>. All core eudicots share an ancient WGD termed the  $\gamma$  event<sup>23</sup>, and, among the 11 plant species analyzed, at least eight additional ancient WGDs have previously been identified<sup>24</sup>. To investigate WGDs in the durian lineage, we identified syntenic regions across the durian, cotton, and cacao genomes, with each region consisting of at least five collinear homologous genes (Supplementary Table 6). We found 60% of the cacao genome represented by syntenic regions in the durian genome. Of these syntenic regions in durian, 25% were present in one copy, 37% in two copies, 36% in three copies, and the remainder (2%) in four or more copies (Supplementary Fig. 4). The observation that 75% of the genomic regions syntenic between durian and cacao are present in multiple copies in durian strongly suggests that the durian lineage underwent a WGD after speciation from cacao. The  $K_s$  (synonymous substitution rate) distribution between syntenic durian genes also exhibited a peak characteristic of a recent WGD ( $K_s = 0.24$ ; Fig. 2c).

A WGD has also previously been shown for the cotton lineage, after its speciation with cacao<sup>25,26</sup>. We investigated whether the WGDs associated with cotton and durian reflect the same event or represent two distinct evolutionary events, by resolving the ordering and dates of the WGDs alongside the durian–cotton divergence event. We extracted paralogous pairs of durian and cotton genes arising from their respective WGDs, established orthologous relationships between them, and estimated their phylogenetic relationships and divergence times (Supplementary Table 7). Our analysis showed that paralogs within cotton and durian (originating from the WGD) each predated their orthologs (originating from the durian–cotton divergence; posterior probability for the phylogenetic tree branches  $\geq 0.9981$ ; Fig. 2d). This provides evidence that durian and cotton share a WGD that likely occurred before their lineages diverged and places the known cotton-specific WGD also within the durian lineage.

### Upregulated gene expression in sulfur and ripening

To investigate biological processes associated with durian fruit traits, we sequenced expressed RNAs (RNA-seq) from different plant organs of the Musang King cultivar, including ripe fruit arils and stem, leaf, and roots. In addition, we sequenced RNA from ripe fruit arils of two other durian cultivars (Monthong and Puang Manee). We performed three independent comparative transcriptomic analyses: (i) comparison of durian transcriptomic data from fruit arils (Musang King, Monthong, and Puang Manee combined, three arils each) against root, stem, and leaf transcriptomic profiles (combined); (ii) comparison of durian fruit arils (Musang King, Monthong, and Puang Manee combined, three arils each) against fruits of other plant species combined, specifically *Musa acuminata* (banana; SRA259656), *Mangifera indica* (mango; SRA289054), *Persea americana* (avocado; SRA172282), *Solanum lycopersicum* (tomato; SRA049915), and *Vaccinium corymbosum* (blueberry; SRA145423); and (iii) comparison of Musang King fruit arils (three arils) against Monthong and Puang Manee fruit arils combined (three arils each).



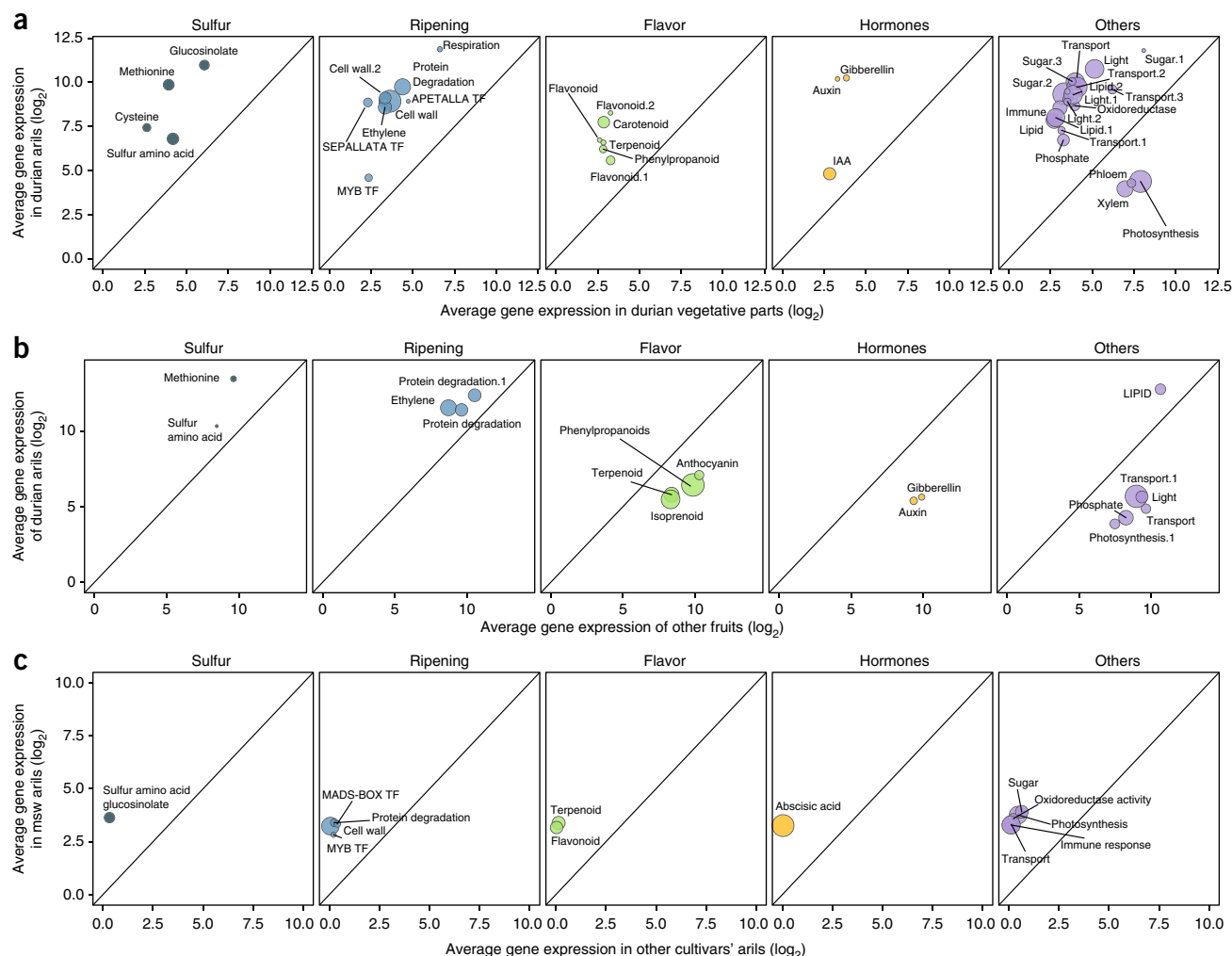


**Figure 2** Comparative genomic analysis of durian with other plants. **(a)** Sharing of gene families by durian and three other Malvales plants, with *A. thaliana* as an outgroup. The number in parentheses indicates durian-specific gene families among all 11 plants considered. **(b)** Inferred phylogenetic tree across 11 plant species. Established WGD events are placed according to Vanneste *et al.*<sup>24</sup>. Posterior probabilities for all branches are >0.9999. **(c)** Frequency distributions of synonymous substitution rates ( $K_s$ ) between collinear genes in syntenic blocks in durian and *G. raimondii*. Peak boundaries corresponding to recent WGDs (green shading) were derived from Gaussian mixture modeling. **(d)** Inferred phylogenetic tree of paralogous durian and *G. raimondii* genes originating from their respective WGDs. Numbers represent branching posterior probabilities. The tree topology shows that the divergence of durian and cotton paralogs (derived from a WGD) predates the divergence of durian–cotton orthologs (derived from speciation) with high branching posterior probability, suggesting a shared WGD occurring before durian–cotton speciation (60–65 million years ago; green circle).

Gene set enrichment analysis (GSEA)<sup>27</sup> showed that genes upregulated in durian fruits relative to non-fruit organs were associated with distinct cellular pathways, such as sulfur-, ripening-, and flavor-related processes (Fig. 3a and Supplementary Table 8). Enrichment in gene sets related to sulfur comprised five clusters, of which acid-thiol ligase enzymes (containing a number of key enzymes in carbon–sulfur reactions and flavonoid production) and methionine metabolic pathways (of which *MGL* is a key member)<sup>28</sup> contained the most significantly upregulated leading-edge genes. Ripening-related gene sets included genes regulated by the MADS-BOX transcription factor family<sup>29,30</sup>, SEPALLATA transcription factor family, and ethylene-related genes such as ACS (aminocyclopropane-1-carboxylic acid synthase), a key ethylene-production enzyme involved in ripening<sup>31</sup>. Other upregulated processes related to taste and odor included triterpenoid metabolism, which has been associated with the intensely bitter taste of *Ganoderma lucidum* (Lingzhi)<sup>32</sup>, and lipid-related genes involved in the production of  $C_6$  volatile compounds (hexanal and hexanol), which have been associated with the green odor profiles of apples, tomatoes, and bananas, as well as undesirable rancid odors when present at high levels<sup>33,34</sup>. Conversely, processes enriched in non-fruit organs included photosynthetic pathways and the regulation of nitrogen metabolism, likely required by roots and leaves to manufacture amino acids.

To determine whether the fruit-enriched processes are specific to durian, we next compared the transcriptomes of durian fruits to those from the fruits of five other species. GSEA showed that sulfur-related pathways, lipid-oxidation pathways, and ripening pathways related to ethylene as well as protein degradation were upregulated in durian as compared to other fruits, whereas some flavor-related pathways were downregulated (Fig. 3b). These results suggest that both sulfur-related pathways and specific ripening processes related to ethylene could be highly regulated in durian fruits, even in comparison to other climacteric fruits. A prominent role for these pathways in durian is supported in the literature, as VSCs have been identified as important components in durian odor<sup>5,6,35</sup>. Ethylene production and cellular respiration are also increased in durian fruits during the climacteric stage of ripening and are associated with important physiological changes, including odor production, pulp softening, sugar release, and water loss followed by dehiscence<sup>1,36</sup>.

We also investigated whether differences in sulfur-, ripening-, and flavor-related pathways might be associated with the distinct and multifarious odor descriptors ascribed to different durian cultivars. GSEA results showed that all three pathways were significantly upregulated (sulfur metabolic process,  $P = 0.0019$  by Kolmogorov–Smirnov test; response to ethylene,  $P = 0.0009$ ; flavonoid metabolic process,  $P = 0.0009$ ) in Musang King as compared to Monthong and Puang Manee



**Figure 3** Transcriptome analysis identifies differentially expressed pathways in durian fruit. (a–c) Differentially expressed pathways in durian fruit arils versus stem, root, and leaf (a), durian fruit arils versus fruits from five other plant species (banana, mango, avocado, tomato, and blueberry) (b), and fruit arils from the Musang King cultivar versus fruit arils from other cultivars (Monthong and Puang Manee) (c). Circle sizes correspond to the number of differentially expressed genes in enriched pathways. The diagonal in each plot corresponds to a region where gene expression was not different between the two groups. Genes in the regions above this line correspond to gene sets that are prominently upregulated in durian or Musang King fruit.

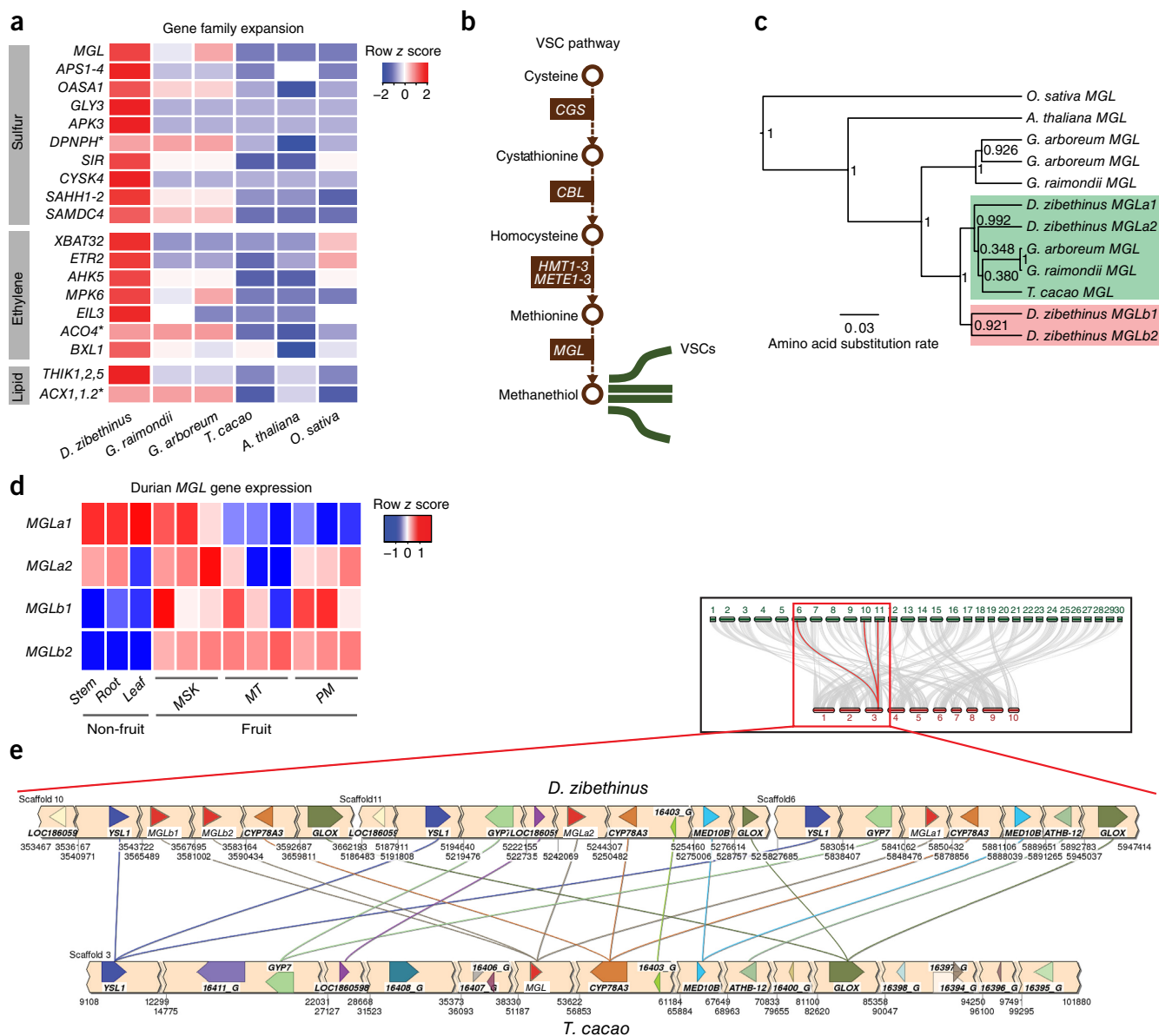
(Fig. 3c). These findings correlate well with the stronger perceived taste and smell of the Musang King cultivar.

### Genomic expansions of volatile sulfur compound genes

Durian aroma comprises at least two major groups of volatile compounds: (i) sulfur-containing volatiles, such as thiols, disulfides, and trisulfides, which may contribute to the distinct roasty, onion-like odor of durian, and (ii) esters, which may contribute to a sweeter, fruity odor<sup>5,6,37</sup>. The former group (VSCs) is of particular interest, as its odor descriptors most closely reflect the distinctive nature of the durian aroma. Interestingly, analysis of the durian genome identified evolutionary gene family expansions of the *MGL* gene (four copies in durian, three copies in *G. arboreum*, two copies in *G. raimondii*, and one copy in *T. cacao*, *A. thaliana*, and *O. sativa*; Fig. 4a). Studies in microbes and plants have shown that *MGL* is a major functional contributor to VSC biosynthesis, degrading the sulfur-containing amino acids cysteine and methionine into methanethiol, which is further broken down into di- and trisulfides<sup>1,38,39</sup> (Fig. 4b). Phylogenetic analysis of *MGL* proteins from durian and other species identified two evolutionary *MGL* subgroups (posterior probability

> 0.9999; Fig. 4c). The first group, referred to as *MGLa*, included singular *MGL* genes from *T. cacao* and both cotton species, as well as two durian *MGL* genes, *MGLa1* (Duzib1.0C006G000732) and *MGLa2* (Duzib1.0C011G000528). This group likely represents the ancestral, conserved *MGL* gene family. The second *MGL* group included only the two remaining duplicated copies of *MGL* from durian, suggesting that it comprises *MGL* genes that have diverged in sequence after genome duplication. We denoted the *MGL* genes in this second group *MGLb1* (Duzib1.0C010G000484) and *MGLb2* (Duzib1.0C010G000488).

To investigate the functions of the *MGLa* and *MGLb* genes in durian, we examined differences in their expression between durian fruit arils and non-fruit organs (stem, root, and leaf). *MGLb1* and *MGLb2* were consistently upregulated in fruits, with *MGLb2* in particular showing dramatically higher expression (DESeq2; 12.73- and 2,241-fold increase relative to non-fruit organs,  $P = 2.07 \times 10^{-4}$  and  $9.26 \times 10^{-100}$  for *MGLb1* and *MGLb2*, respectively; Fig. 4d and Supplementary Table 9). In contrast, among the *MGLa* genes, *MGLa1* exhibited decreased expression in fruits (7.14-fold decrease,  $P = 4.44 \times 10^{-3}$ ). These results suggest a novel function for durian *MGLb* in durian fruits.

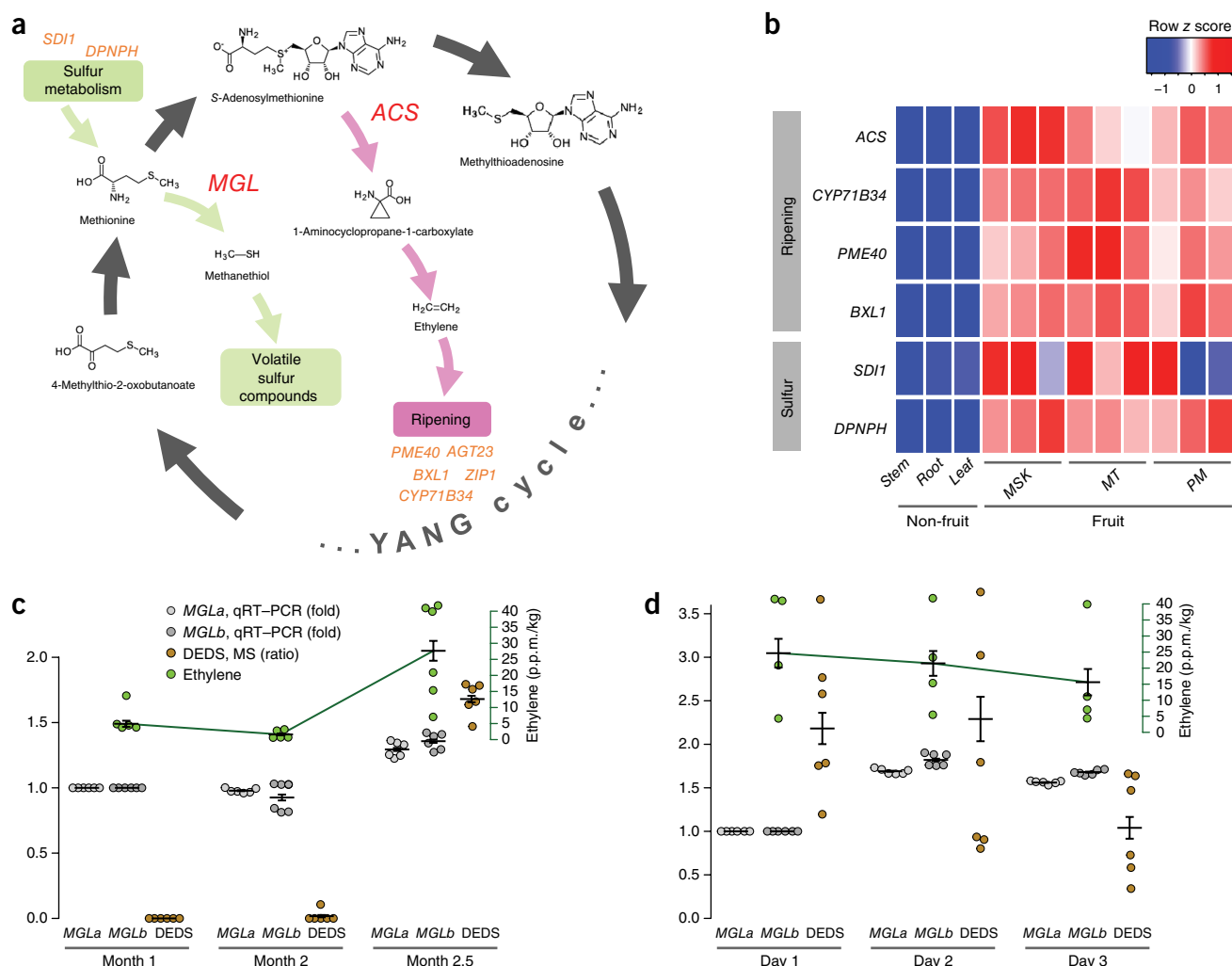


**Figure 4** Gene family analysis showing expansion in sulfur-related pathways associated with volatile sulfur compounds. **(a)** Gene families in sulfur, ethylene, and lipid metabolism pathways expanded in durian, including the *MGL* family. An asterisk indicates gene families expanded in both *Gossypium* and durian. **(b)** The VSC pathway, which breaks down cysteine and methionine into methanethiol and other VSCs, via the *MGL* enzyme. **(c)** Phylogeny of the *MGL* genes in durian, cotton, cacao, *Arabidopsis*, and rice showing two groups of durian *MGL* genes, denoted *MGLa* and *MGLb*. Numbers correspond to branching posterior probabilities. **(d)** *MGLb* durian genes are upregulated in durian fruit. MSK, Musang King; MT, Monthong; PM, Puang Manee. **(e)** The three genomic regions containing the durian *MGLa* and *MGLb* genes show clear synteny with the cacao genome, while the two *MGLb* genes occur in tandem. This suggests the involvement of both WGD and tandem duplication events in *MGL* family expansion.

To investigate potential mechanisms by which *MGL* might have expanded from a single copy in the durian–cacao ancestor to four copies in durian, we compared the *MGL*-containing genomic regions in cacao and durian. The two durian *MGLa* genes were present on two separate scaffolds, both of which showed regional synteny to the cacao *MGL* region. In contrast, the two durian *MGLb* genes were present in close proximity on the same scaffold, which also displayed synteny with the cacao *MGL* region (Fig. 4e). This suggests that *MGL* initially expanded in the durian lineage via a large-scale genomic duplication event (such as a WGD), followed by a subsequent tandem duplication event that doubled the durian *MGLb* genes.

Besides *MGL*, we also observed significant genomic expansions at the pathway level in gene families related to sulfur metabolism

(false discovery rate (FDR) = 0.032), fatty acid metabolism (FDR = 0.018), and ethylene processes (FDR =  $1.76 \times 10^{-7}$ ) (Fig. 4a and Supplementary Table 10). These included the *APS* (ATP sulfurylase) family, involved in hydrogen sulfide biosynthesis; *OASA1* (cysteine synthase), involved in maintaining sulfur levels; the *THIK* (3-ketoacyl-CoA thiolase) family, involved in the production of lipid volatiles; *XBAT32* (an E3 ubiquitin–protein ligase), a regulator of ethylene biosynthesis; and *ETR2* (ethylene receptor) and *AHK5* (histidine kinase), involved in ethylene signaling. Interestingly, some genes in these pathways were expanded in both durian and cotton, supporting findings of the importance of ethylene processing in cotton<sup>21</sup>. Taken together, these results suggest that a WGD event in durian's lineage led to the expansion and diversification of pathways related to durian



**Figure 5** The Yang cycle may link ethylene production and volatile sulfur compound production. **(a)** The Yang cycle (black arrows) salvages methionine consumed for ethylene production (pink), while sulfur-related processes (green) replenish and consume the sulfur moiety from methionine. **(b)** Upregulation of ripening- and sulfur-related genes in durian fruit. **(c)** MS analysis of the VSC DEDS, qRT-PCR analysis of *MGL*, and measurement of ethylene across durian fruits at three ripening stages (1 month, 2 months, and 2.5 months post-anthesis) show a coordinated spike toward later ripening stages. Dots correspond to individual samples, and bars represent means  $\pm$  s.e.m. **(d)** Increased levels of DEDS, *MGL*, and ethylene are maintained in post-abscission fruits (fully ripened) over 3 d at room temperature. Dots correspond to individual samples, and bars represent means  $\pm$  s.e.m.

volatiles, such as those involved in sulfur processing (including *MGL*), lipid volatiles, and ethylene. Upregulation of these genes in durian may be linked to increased VSC production, thereby contributing to durian odor.

### Potential association between VSC production and ripening

In climacteric plant species, ethylene sensing has been shown to act as a major trigger of fruit ripening. Ethylene production is a three-step reaction, beginning with transfer of the adenosyl group in ATP to L-methionine by SAM synthase, followed by conversion of L-methionine to 1-aminocyclopropane-1-carboxylic acid and ethylene by the enzymes ACS and ACO, respectively<sup>40</sup>. Our genomic and transcriptomic findings suggest a potential association between durian odor and fruit ripening, based on the central role of methionine in both processes. In this model (Fig. 5a), methionine acts as a precursor amino acid for both VSCs and ethylene via *MGL* and ACS activity, respectively, and both *MGL* and ACS are upregulated in durian fruits (Figs. 4d and 5b). We also observed upregulation of other genes plausibly related to durian ripening, including *PME40*, which

encodes a pectinesterase that has an important role in fruit softening in strawberries<sup>41</sup>; *C71BV*, which encodes a cytochrome P450 upregulated in banana fruits<sup>42</sup>; and *BXL1*, which encodes a  $\beta$ -D-xylosidase involved in cell wall modification of ripening tomatoes<sup>43</sup>. To study associations between *MGL*, ACS, VSCs, and ethylene, we profiled six durian fruits at different ripening stages (six arils for each stage): 1 month post-anthesis, 2 months post-anthesis, 2.5 months post-anthesis (ripe), and over 3 d post-abscission (fully ripe). qRT-PCR analysis of *MGLa* and *MGLb* in durian arils over the post-anthesis period demonstrated a coordinated spike in expression over the ripening period, in concert with increasing VSC levels as detected by profiling with headspace gas chromatography coupled with mass spectrometry (GC-MS) and increasing ethylene levels (Fig. 5c and Supplementary Tables 11–13). The availability of global mass spectrometry profiles also allowed us to identify different types of VSCs such as sulfide complexes and disulfide analogs (details of different VSCs are provided in Supplementary Table 11), the most abundant of which was diethyl disulfide (DEDS), a compound produced through redox reactions from organic thiols. The levels of *MGL* expression, DEDS,



and ethylene remained high for post-abscission fruits maintained over 3 d at room temperature (Fig. 5d).

The usage of methionine as a precursor for both odor (VSCs) and ripening (ethylene) in durian suggests the presence of a rapidly sulfur-depleted environment in durian fruits, as (i) sulfur is a scarce nutrient, (ii) VSC production via MGL is a sulfur sink, and (iii) 5-methylthioadenosine (MTA), the byproduct of ACS, contains a reduced sulfur group. Consistent with a low-sulfur state, we observed upregulation in fruits of several genes related to sulfur sensing and scavenging, including *SDI1*, a sulfur-sensing gene, and upregulation of *BGL*, which encodes a  $\beta$ -glucosidase that cleaves sulfur from glucosinolate<sup>44</sup>. An additional pathway for salvaging sulfur from MTA is the Yang cycle, which allows sulfur moieties from MTA to be recycled back to methionine<sup>45,46</sup>. In summary, our results suggest that the complex aroma of durian is possibly linked to durian fruit ripening, as both ethylene and VSC production are connected to the same regulatory processes and metabolites. However, we emphasize that, in the current study, the potential associations between MGL expression, VSC generation, and ethylene production are correlative and that further functional experiments will be required to establish a direct causal link between these molecular entities.

## DISCUSSION

Although durian is well known as a delicacy in tropical food, research focused on it has been hampered by an absence of genetic resources. The *D. zibethinus* Musang King draft genome thus provides a useful resource for durian agronomy and is, to our knowledge, the first and most complete genome assembly of durian. The *D. zibethinus* assembly also represents the first from the Helicteroideae subfamily and is thus also valuable for evolutionary phylogenomic studies, similar to cotton and cacao, whose genomes are available and for which genetic research is actively pursued.

Ancient polyploidization events are important evolutionary drivers for the emergence of new phenotypes and new species<sup>22</sup>. Once duplication occurs, novel or specialized functions can arise in redundant alleles, and new regulatory mechanisms can develop through genomic rearrangements. Our observation that durian likely shares the previously described cotton WGD has two implications. First, other subfamilies within Malvaceae, after divergence from the Byttnerioideae subfamily (including cacao), are also likely to have the cotton-specific WGD. Second, this WGD event, which has been proposed to have driven the evolution of unique *Gossypium* traits<sup>26</sup>, may also have been involved in driving the evolution of unique durian traits, as well as the radiation of different Malvaceae subfamilies around the same time<sup>47,48</sup>.

Our current study focused on analyzing biological processes related to durian odor. VSCs have been identified as major contributors to durian smell, and we observed upregulation of sulfur-related pathways in durian fruit arils in comparison to other durian plant organs and fruits from other species. At the genomic level, we also identified expansions in gene families in sulfur-related pathways, most notably in the MGL gene, whose product has been shown to catalyze the breakdown of methionine into VSCs in microbes and plants<sup>38,39,49</sup>. More specifically, our analysis suggests a distinct role for *MGLb* genes in the durian fruit, where these genes were found to be significantly upregulated. MGL functions as a homotetramer in which each subunit requires a pyridoxal 5'-phosphate (PLP) cofactor, with the active sites located in the vicinity of PLP and the substrate-binding site<sup>39,49</sup>. While these active sites are conserved in both durian MGLa and MGLb protein sequences as compared to other plant and bacterial MGL sequences, regions proximal to the active sites, which are conserved in other plant species, exhibited more mutations in MGLb

than in MGLa (four versus two, respectively; Supplementary Fig. 5). The impact of these mutations on the function of durian MGL gene products requires further investigation.

Our analysis also suggests a potential association between MGL and ACS in the durian fruit, whose coordination may involve the Yang cycle. Previous durian studies have also noted correlations between VSC production and increased ethylene production<sup>36</sup>; however, a genetic link behind these observations has not been demonstrated. It is possible that linking odor and ripening may provide an evolutionary advantage for durian in facilitating fruit dispersal. Certain plants whose primary dispersal vectors are primates with more advanced olfactory systems show a shift in odor at ripening<sup>50–52</sup>. Similarly, durian—by emanating an extremely pungent odor at ripening—appears to have the characteristic of a plant whose main dispersal vectors are odor-enticed primates rather than visually enticed animals.

The durian genome assembly creates a large scope for further studies. As an example, rapid commercialization of durian has led to the proliferation of cultivars with a wide discrepancy in prices and little way to verify the authenticity of the fruit products at scale. A high-quality genome assembly may aid in identifying cultivar-specific sequences, including SNPs related to important cultivar-specific traits (such as taste, texture, and odor), and allow molecular barcoding of different durian cultivars for rapid quality control. At a more basic level, such genetic information is vital to better understanding of durian biodiversity. The *Durio* genus comprises more than 30 known species, some of which do not produce edible fruit (for example, *Durio singaporensis*) and others that are outcompeted in nature and face extinction. Further studies will help to elucidate the ecological roles of these important and fascinating tropical plants.

**URLs.** Arrow, <https://github.com/PacificBiosciences/GenomicConsensus>; Picardtools, <http://broadinstitute.github.io/picard/>; UN Trade Statistics, <http://unstats.un.org/>; RepeatMasker and RepeatModeler, <http://www.repeatmasker.org/>.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

This work was supported by Thorn Biosystems Pte Ltd through funding from an anonymous donor and by Dovetail Genomics through their End-of-Year Matching Funds Award. We thank all co-authors and colleagues for donating their after-work hours to support this study. We thank C.N. Lee for his support. We thank A. Tan from 101 Fruits for education and guidance in durian cultivars.

## AUTHOR CONTRIBUTIONS

B.T.T. and P.T. conceived the project. B.T.T., P.T., N.N., and S.S. directed the study and supervised the research. K.L., C.H.Y., C.C.Y.N., N.N., B.T.T., and P.T. designed and interpreted the data and wrote the manuscript. C.C.Y.N., S.R.R., and V.R. performed the experiments. K.C. and V.K.Y.C. provided support for the experiments. K.L., C.H.Y., N.N., and W.K.L. performed the data analysis. S.G.R. contributed to the data analysis. C.K.O. and P.S.S. contributed to sample collection. All authors have read and approved the final manuscript.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.





This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

1. Li, J.X., Schieberle, P. & Steinhaus, M. Characterization of the major odor-active compounds in Thai durian (*Durio zibethinus* L. 'Monthong') by aroma extract dilution analysis and headspace gas chromatography-olfactometry. *J. Agric. Food Chem.* **60**, 11253–11262 (2012).
2. Wallace, A.R. On the bamboo and durian of Borneo. *Hooker's. J. Bot.* **8**, 225–230 (1856).
3. Winokur, J. *The Traveling Curmudgeon: Irreverent Notes, Quotes, and Anecdotes on Dismal Destinations, Excess Baggage, the Full Upright Position, and Other Reasons Not to Go There* (Sasquatch Books, 2003).
4. Siriphanich, J. *Postharvest Biology and Technology of Tropical and Subtropical Fruits* (Woodhead Publishing Limited, 2011).
5. Jaswir, I., Che Man, Y.B., Selamat, J., Ahmad, F. & Sugisawa, H. Retention of volatile components of durian fruit leather during processing and storage. *J. Food Process. Preserv.* **32**, 740–750 (2008).
6. Chin, S.T. et al. Analysis of volatile compounds from Malaysian durians (*Durio zibethinus*) using headspace SPME coupled to fast GC–MS. *J. Food Compos. Anal.* **20**, 31–44 (2007).
7. Bumrungsri, S., Sripaoraya, E., Chongsiri, T., Sridith, K. & Racey, P.A. The pollination ecology of durian (*Durio zibethinus*, Bombacaceae) in southern Thailand. *J. Trop. Ecol.* **25**, 85–92 (2009).
8. Yumoto, T. Bird-pollination of three *Durio* species (Bombacaceae) in a tropical rainforest in Sarawak, Malaysia. *Am. J. Bot.* **87**, 1181–1188 (2000).
9. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
10. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
11. Putnam, N.H. et al. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
12. Manganot, S. & Manganot, G. Enquête sur les nombres chromosomiques dans une collection d'espèces tropicales. *Bulletin de la Société Botanique de France* **109**, 411–447 (1962).
13. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
14. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
15. 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* **31**, 3210–3212 (2015).
16. Li, L., Stoeckert, C.J. Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
17. Duarte, J.M. et al. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* **10**, 61 (2010).
18. Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
19. Alverson, W.S., Whitlock, B.A., Nyffeler, R., Bayer, C. & Baum, D.A. Phylogeny of the core Malvales: evidence from *ndhF* sequence data. *Am. J. Bot.* **86**, 1474–1486 (1999).
20. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L.L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* **207**, 437–453 (2015).
21. Li, F. et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat. Genet.* **46**, 567–572 (2014).
22. Otto, S.P. The evolutionary consequences of polyploidy. *Cell* **131**, 452–462 (2007).
23. Jiao, Y. et al. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, R3 (2012).
24. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).
25. Wang, K. et al. The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* **44**, 1098–1103 (2012).
26. Paterson, A.H. et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423–427 (2012).
27. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
28. Joshi, V. & Jander, G. *Arabidopsis* methionine  $\gamma$ -lyase is regulated according to isoleucine biosynthesis needs but plays a subordinate role to threonine deaminase. *Plant Physiol.* **151**, 367–378 (2009).
29. Giovannoni, J.J. Fruit ripening mutants yield insights into ripening control. *Curr. Opin. Plant Biol.* **10**, 283–289 (2007).
30. Seymour, G., Poole, M., Manning, K. & King, G.J. Genetics and epigenetics of fruit development and ripening. *Curr. Opin. Plant Biol.* **11**, 58–63 (2008).
31. Burg, S.P. & Burg, E.A. Role of ethylene in fruit ripening. *Plant Physiol.* **37**, 179–189 (1962).
32. Nishitoba, T., Sato, H., Kasai, T., Kawagishi, H. & Sakamura, S. New bitter C<sub>27</sub> and C<sub>30</sub> terpenoids from the fungus *Ganoderma lucidum* (Reishi). *Agric. Biol. Chem.* **49**, 5 (1984).
33. Eskin, N.A., Grossman, S. & Pinsky, A. Biochemistry of lipoxygenase in relation to food quality. *CRC Crit. Rev. Food Sci. Nutr.* **9**, 1–40 (1977).
34. Rowan, D.D., Allen, J.M., Fielder, S. & Hunt, M.B. Biosynthesis of straight-chain ester volatiles in Red Delicious and Granny Smith apples using deuterium-labeled precursors. *J. Agric. Food Chem.* **47**, 2553–2562 (1999).
35. Ketsa, S. & Daengkanit, T. Firmness and activities of polygalacturonase, pectinesterase,  $\beta$ -galactosidase and cellulase in ripening durian harvested at different stages of maturity. *Sci. Hortic. (Amsterdam)* **80**, 181–188 (1999).
36. Maninang, J.S., Wongs-Aree, C., Kanlayanarat, S., Sugaya, S. & Gemma, H. Influence of maturity and postharvest treatment on the volatile profile and physiological properties of the durian (*Durio zibethinus* Murray) fruit. *Int. Food Res. J.* **18**, 1067–1075 (2011).
37. Landaud, S., Helinck, S. & Bonnarne, P. Formation of volatile sulfur compounds and metabolism of methionine and other sulfur compounds in fermented food. *Appl. Microbiol. Biotechnol.* **77**, 1191–1205 (2008).
38. Rébeillé, F. et al. Methionine catabolism in *Arabidopsis* cells is initiated by a  $\gamma$ -cleavage process and leads to S-methylcysteine and isoleucine syntheses. *Proc. Natl. Acad. Sci. USA* **103**, 15687–15692 (2006).
39. Gonda, I. et al. Catabolism of l-methionine in the formation of sulfur and other volatiles in melon (*Cucumis melo* L.) fruit. *Plant J.* **74**, 458–472 (2013).
40. Rodrigues, M.A., Bianchetti, R.E. & Freschi, L. Shedding light on ethylene metabolism in higher plants. *Front. Plant Sci.* **5**, 665 (2014).
41. Castillejo, C., de la Fuente, J.I., Iannetta, P., Botella, M.A. & Valpuesta, V. Pectin esterase gene family in strawberry fruit: study of *FaPE1*, a ripening-specific isoform. *J. Exp. Bot.* **55**, 909–918 (2004).
42. Asif, M.H. et al. Transcriptome analysis of ripe and unripe fruit tissue of banana identifies major metabolic networks involved in fruit ripening process. *BMC Plant Biol.* **14**, 316 (2014).
43. Itai, A., Ishihara, K. & Bewley, J.D. Characterization of expression, and cloning, of  $\beta$ -D-xylosidase and  $\alpha$ -L-arabinofuranosidase in developing and ripening tomato (*Lycopersicon esculentum* Mill.) fruit. *J. Exp. Bot.* **54**, 2615–2622 (2003).
44. Zheng, Z.-L., Zhang, B. & Leustek, T. Transceptors at the boundary of nutrient transporters and receptors: a new role for *Arabidopsis* SULTR1;2 in sulfur sensing. *Front. Plant Sci.* **5**, 710 (2014).
45. Baur, A.H. & Yang, S.F. Methionine metabolism in apple tissue in relation to ethylene biosynthesis. *Phytochemistry* **11**, 3207–3214 (1972).
46. Miyazaki, J.H. & Yang, S.F. Metabolism of 5-methylthioribose to methionine. *Plant Physiol.* **84**, 277–281 (1987).
47. Salman-Minkov, A., Sabath, N. & Mayrose, I. Whole-genome duplication as a key factor in crop domestication. *Nature Plants* **2**, 16115 (2016).
48. Soltis, P.S. & Soltis, D.E. Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* **30**, 159–165 (2016).
49. Kudou, D. et al. Structure of the antitumor enzyme l-methionine  $\gamma$ -lyase from *Pseudomonas putida* at 1.8 Å resolution. *J. Biochem.* **141**, 535–544 (2007).
50. Hodgkinson, R. et al. Fruit bats and bat fruits: the evolution of fruit scent in relation to the foraging behaviour of bats in the New and Old World tropics. *Funct. Ecol.* **27**, 1075–1084 (2013).
51. Borges, R.M., Bessière, J.M. & Hossaert-McKey, M. The chemical ecology of seed dispersal in monoecious and dioecious figs. *Funct. Ecol.* **22**, 484–493 (2008).
52. Lomáscolo, S.B., Levey, D.J., Kimball, R.T., Bolker, B.M. & Alborn, H.T. Dispersers shape fruit diversity in *Ficus* (Moraceae). *Proc. Natl. Acad. Sci. USA* **107**, 14668–14672 (2010).

## ONLINE METHODS

**Sample collection.** Durian samples used for genomic analysis (Musang King fruit stalk) and transcriptomic analysis (three Musang King fruit arils, three Monthong fruit arils, and three Puang Manee fruit arils, as well as three vegetative parts (stem, root, and leaf)) were from a plantation in the Bentong region of Pahang, West Malaysia, while durian fruits used for validation studies (MS profiling of VSCs, qRT-PCR of *MGL*, and measurement of ethylene) were from a separate plantation in Sendenak, Johor, West Malaysia, at  $-1^{\circ} 41' 55.8''$  N,  $103^{\circ} 28' 57.7''$  E. For the different ripening stages, we sourced fruiting trees in this plantation in July 2016 and isolated three trees of ages 28, 45, and 45 years. Two durian fruits were collected per tree in each ripening stage: 1 month post-anthesis, 2 months post-anthesis, 2.5 months post-anthesis, and post-abscission. In each ripening stage (1 month post-anthesis, 2 months post-anthesis, 2.5 months post-anthesis, and post-abscission), we profiled six arils from six durian fruits for qRT-PCR and MS and profiled ethylene for six durian whole fruits. For ethylene measurements, samples that did not obtain stable profiles were discarded.

**DNA extraction and library preparation.** Genomic DNA was extracted from a fruit stalk using Plant DNAzol reagent (Life Technologies) following the manufacturer's recommendations.

SMRTbell DNA library preparation and sequencing with P6-C4 chemistry were performed in accordance with the manufacturer's protocols (Pacific Biosciences). 50  $\mu$ g of high-quality genomic DNA was used to generate a 20-kb SMRTbell library using a lower end of 10 kb in size selection on the BluePippin (Sage Science). Initial titration runs were performed to optimize SMRT cell loading and yield. The genome was sequenced employing 52 SMRT cells on the PacBio RSII platform (Pacific Biosciences) by sequencing provider Icahn Mount Sinai School of Medicine.

A short-read genomic library was prepared using the TruSeq Nano DNA Library Preparation kit (Illumina) in accordance with the manufacturer's recommendations. Briefly, 1  $\mu$ g of extracted intact genomic DNA was sheared (Covaris), with an insert size of 200–300 bp, and ligated to adaptors. DNA fragments were subjected to PCR, and the DNA library was used in paired-end sequencing ( $2 \times 100$  bp,  $\sim 750$  million paired-end reads,  $\sim 202\times$  coverage) on the Illumina HiSeq 2500 sequencer.

**RNA extraction and library preparation.** RNA-seq experiments were conducted for Musang King fruit arils (three arils, 51.6 million paired-end reads each on average) and a combined non-fruit group comprising stem (55.6 million reads), leaf (56.9 million reads), and root (53.2 million reads) profiles. We also sequenced fruit arils of Monthong and Puang Manee cultivars (three arils each, 90 million reads each on average). Our sample sizes and read depths enabled significance testing of differentially expressed genes and genes sets<sup>53</sup> (by DESeq2 and GSEA).

RNA was extracted using the PureLink RNA mini kit (Life Technologies) following the manufacturer's recommended protocol.

For the construction of RNA-seq libraries, 2  $\mu$ g of total RNA was processed using the TruSeq RNA Sample Preparation kit (Illumina) followed by sequencing on the Illumina HiSeq 2500 platform.

**Genome assembly.** The main assembly was performed on full PacBio long reads using two diploid assembly approaches. CANU<sup>54</sup> was run with both default and high-sensitivity settings, and FALCON was run with length\_cutoff = 5000 for initial mapping of seed reads for the error-correction phase. The CANU assembly had a N50 value indicating a less contiguous assembly, and the assembly did not substantially improve even after applying Redundans, a tool that detects redundant heterozygous contigs and attempts to reduce assembly fragmentation. We subsequently used FALCON because it was better able to phase the diploid genome as well as indicating better contiguity from N50 evaluations. Errors in the PacBio reads were corrected within the FALCON pipeline. Screening for contamination was handled with PacBio's whitelisting pipeline. We subsequently compared a number of length\_cutoff values in FALCON for the mapping of seed reads for preassembly ranging from 2,000 to 11,000. On the basis of the contig N50 results, we chose length\_cutoff = 9,500 for the preassembly step. We additionally configured DBsplit ( $-x500$ ,  $-s400$ ), daligner ( $-B128$ ,  $-L4800$ ,  $-s100$ ,  $-k18$ ,  $-h480$ ), and overlap\_filtering

( $--max\_diff$  100,  $--max\_cov$  100) with parameters optimized for our plant assembly. The results from this step were used to perform phased diploid genome assembly using FALCON-Unzip (default parameters). Subsequently, the draft assembly was polished using Arrow (see URLs) over three iterations and finally corrected using Illumina short reads with Pilon.

For additional details about genome assembly validation, see the **Supplementary Note**.

**Estimation of genome size and ploidy.** We estimated genome size by flow cytometry<sup>55</sup> and obtained an estimate of 800 Mb (**Supplementary Table 2**). We additionally estimated genome size by *k*-mer distribution analysis with the program Jellyfish<sup>56</sup> ( $k = 31$ ), using Illumina short reads, and obtained an estimate of 738 Mb. For additional details about ploidy estimation, see the **Supplementary Note**.

**ChICAGO and Hi-C library preparation and sequencing.**  $\sim 500$  ng of high-molecular-weight genomic DNA (mean fragment length = 150 kb) was reconstituted into chromatin *in vitro* for the ChICAGO library. Chromatin was fixed with formaldehyde for the ChICAGO library. For the Hi-C library, chromatin was fixed in place with formaldehyde in the nucleus. Fixed chromatin was digested with DpnII, 5' overhangs filled in with biotinylated nucleotides, and free blunt ends were ligated. After ligation, cross-links were reversed and the DNA was purified from protein. Purified DNA was treated to remove biotin that was not internal to the ligated fragments. The DNA was then sheared to a mean fragment size of  $\sim 350$  bp, and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adaptors. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq X to produce 233 million (ChICAGO) and 280 million (Hi-C)  $2 \times 150$ -bp paired-end reads, which provided  $379.52\times$  (1- to 100-kb pairs) and  $4,370.55\times$  (100- to 1,000-kb pairs) physical coverage of the genome, respectively.

**Scaffolding with HiRise.** The input *de novo* assembly, shotgun reads, ChICAGO library reads, and Dovetail Hi-C library reads were used as input data for HiRise, a software pipeline designed specifically to use proximity-ligation data to scaffold genome assemblies. An iterative analysis was conducted. First, shotgun and ChICAGO library sequences were aligned to the draft input assembly using a modified SNAP read mapper<sup>57</sup>. The separation between ChICAGO read pairs mapping within draft scaffolds was analyzed by HiRise to produce a likelihood model for the genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and to make joins above a selected threshold. After aligning and scaffolding ChICAGO data, Dovetail Hi-C library sequences were aligned and scaffolded following the same method. After scaffolding, shotgun sequences were used to close gaps between contigs. Scaffolding with ChICAGO libraries made 167 breaks and 2,359 joins to the contig assembly. We additionally used Hi-C libraries to detect 7 breaks and 706 joins.

**Transcriptome assembly.** RNA-seq reads were preprocessed by trim\_galore to remove contaminating sequences from adaptors as well as sequences with low base quality. Reads were then aligned to the reference genome using HiSat<sup>58</sup>. Reference-guided assembly was performed on the resulting alignment to produce a plant-organ-specific set of transcripts using StringTie<sup>59</sup>. A non-redundant set of transcripts observed in all samples assembled was merged using the merge functionality in StringTie.

**Genome annotation.** We generated a *de novo* repeat library from our assembly using RepeatModeler. We removed *de novo* repeats that matched known *Arabidopsis* genes (BLASTN  $E < 10^{-5}$ ). We annotated our *de novo* repeats using the RepeatClassifier module of RepeatModeler. We combined our *de novo* library with the RepBase plant repeat database<sup>60</sup> and ran RepeatMasker on the assembly (default parameters).

We used the Maker genome annotation pipeline for gene prediction. Gene evidence was provided in the form of 151,594 protein sequences from four plant species (*Arabidopsis*, grape, rice, and soybean) and 178,840 transcripts assembled from our RNA-seq data. For details on how Maker was run, see the **Supplementary Note**.

We transferred GO annotations to our predicted genes using Blast2GO, incorporating sequence homology with *Arabidopsis* genes using BLASTP (best hit with  $E < 10^{-5}$ ) and functional domain search with InterProScan<sup>61</sup>.

**Short-read alignment and SNP calling.** Previously preprocessed Illumina short reads were used for paired-end remapping to the reference assembly using BWA-MEM<sup>62</sup>. The resulting alignments were sorted using SAMtools<sup>63</sup> and PCR, and optical duplicates were marked and removed using Picardtools (see URLs). The processed alignments were then used to infer SNPs with FreeBayes<sup>64</sup> using parameters -C 5, -m 20 --F 0.01. Resulting SNP calls were further filtered to remove those with quality less than 20.

**Molecular phylogenetic analysis.** We performed gene family clustering for 11 plant species using OrthoMCL (default parameters) on protein sequences. Molecular phylogenetic analysis was performed using a strict set of 47 single-copy orthologous genes (Supplementary Table 5), derived by intersecting 344 single-copy gene families found by OrthoMCL with 395 single-copy gene families published. We codon aligned each gene family with MUSCLE<sup>65</sup> and curated the alignments with Gblocks<sup>66</sup>. Phylogeny analysis was performed with BEAST2. For details about model selection for phylogeny, see the Supplementary Note.

**Whole-genome duplication analysis.** We used MCSan<sup>67</sup> and the CoGe Comparative Genomics Platform<sup>68</sup> to detect syntenic blocks (regions with at least five collinear genes) and calculate  $K_s$  rates for syntenic genes. To analyze the durian and *G. raimondii* WGDs, we first identified paralogous durian and *G. raimondii* gene pairs originating from their respective WGDs. In durian and *G. raimondii* separately, we modeled the distribution of  $K_s$  rates as a mixture model and identified syntenic gene pairs falling within the first peak  $\pm 1$  s.d. as paralogs likely derived from the most recent WGD event. Next, we identified OrthoMCL gene families consisting of one such paralogous durian pair, one such paralogous *G. raimondii* pair, and one gene each from cacao and *Arabidopsis*. In each of these 62 families, we matched each durian paralog with its cotton ortholog (among the two cotton paralogs), by setting the durian–cotton pair with the lowest synonymous distance as orthologs (Supplementary Table 7). We then used BEAST2 to estimate the phylogenetic tree and divergence times of the paralogous and orthologous genes in durian and *G. raimondii*, along with cacao and *Arabidopsis*. BEAST2 was run as described above.

**Transcriptome analysis.** Previously aligned RNA-seq data were counted against the predicted gene models using HTSeq-count<sup>69</sup>. Raw counts were fitted on a negative binomial distribution, and differential expression was tested for using DESeq2. Hidden variables were removed using the svu package and added to the study design before running DESeq2. Genes were sorted according to their log<sub>2</sub>-transformed fold-change values after shrinkage in DESeq2 and used for gene set analysis with fgsea's implementation of GSEA. Significant gene sets were used to perform leading-edge analysis. We detected clusters of pathways that shared many leading-edge genes using a community detection algorithm and manually curated these clusters to elucidate the important phenotype-associated pathway groups visualized on the bubble plots. Leading edges that also had a fold change of 2 were used to generate bubble plots. DESeq normalized gene expression was used in all bubble plots and heat maps. For additional details about transcriptomic comparisons against other plant species, see the Supplementary Note.

**Gene family expansion analysis.** For gene family expansion analysis, we considered only six plants from the OrthoMCL gene family results: durian, two cotton species (*G. raimondii* and *G. arboreum*), cacao, *Arabidopsis*, and rice. We considered a gene family expanded in durian if (i) the number of durian gene members was greater than or equal to the number of gene members for every other plant considered, (ii) the number of durian gene members was greater than the number of gene members for cacao, *Arabidopsis*, and rice, and (iii) the number of gene members for all plants considered was greater than 0. These criteria were chosen to detect gene family expansions in durian, as well as expansions in both durian and the closely related cotton species. At the pathway level, we considered whether each pathway was expanded in

durian by checking if its durian gene members were enriched in belonging to expanded gene families (Fisher's exact test with Benjamini–Hochberg multiple-hypothesis correction).

**Quantitative real-time PCR.** cDNA synthesis was performed on 200 ng of total RNA with the SensiFAST cDNA Synthesis kit (Bioline) followed by preamplification with SsoAdvanced PreAmp Supermix (Bio-Rad) according to the manufacturer's recommendations. The cDNA was diluted tenfold, and a 1- $\mu$ l aliquot of cDNA was used for each qRT-PCR reaction using SsoAdvanced SYBR Green Supermix (Bio-Rad). All samples were run in triplicate for each primer combination. Thermal cycling was performed on a CFX C1000 System (Bio-Rad) in 96-well plates with an initiation step at 95 °C for 30 s, 40 cycles of denaturation at 95 °C for 5 s and extension at 60 °C for 15 s, and a final melting curve analysis on each reaction to confirm the specificity of amplification. Relative quantification of gene expression was performed using three durian housekeeping genes, *CAC*, *DNAJ*, and *APT*, as reference. Ct values were determined using CFXTM manager software (Bio-Rad) and exported into MS Excel (Microsoft) for statistical analysis. Real-time efficiencies ( $E$ ) were calculated from the slopes of standard curves for each gene ( $E = 10(1/\text{slope})$ ).

**Mass spectrometry.** Headspace solid-phase microextraction (HS-SPME) was used to identify sulfur-containing volatiles in durian fruits. Harvested fruits were kept in a room maintained at 25 °C with a relative humidity of 80–90%. 500 mg of each sample aril was ground under liquid nitrogen and transferred into a 20-ml brown glass headspace vial (Restek). 50 nmol of the internal standard, thiophene (Sigma-Aldrich), was added to each sample vial, and vials were immediately locked with air-tight headspace screw caps. An 85- $\mu$ m carboxen on polymethylsiloxane (CAR/PDMS) (Supelco, Bellefonte) was used as the SPME. The SPME fiber was conditioned according to the manufacturer's instructions. Online headspace sampling, extraction, and GC–MS analysis were performed on the Agilent 7200 GC-QTOF mass spectrometer equipped with a PAL CTC Headspace autosampler. Headspace extraction and desorption were carried out accordingly. Each sample was incubated for 30 min at 40 °C, followed by extraction of the volatiles for 30 min at 40 °C and 250 r.p.m. Conditioning was performed for 5 min at 250 °C followed by injection into the GC–MS instrument for 1 min at 230 °C. The injector was operated in a 'split' mode at a ratio of 50:1, and the flow rate was maintained at 1 ml/min. An Agilent 19091S-433HP-5MS 5% Phenyl Methyl Silox (30 m  $\times$  0.25 mm  $\times$  0.25  $\mu$ m) column was used to separate the volatiles. The oven temperature was increased from 40 °C to 200 °C at a rate of 5 °C/min followed by further increase to 280 °C at a rate of 40 °C/min for 5 min. The total run time was 39 min. Other parameters are as follows: ion source, EI; source temperature, 230 °C; electron energy, 70 eV; acquisition rate, 2 spectra/s; acquisition time, 500 ms/spectrum; mass range, 40–250 a.m.u. Mass spectra were analyzed using Agilent MassHunter Qualitative Analysis software version B.05.00. Identification of metabolites was performed using the NIST version 11 library based on spectral matching. Quantifications of sulfur-containing metabolites are expressed as follows: peak area of metabolite/peak area of internal standard/0.5 g = abundance/gram.

**Ethylene gas measurement.** Harvested fruits were all kept in a room maintained at 25 °C with a relative humidity of 80–90%. All daily samplings for the post-abscission fruits were conducted at the same starting time to ensure consistency. Each durian fruit was placed in an airtight 8-L container fitted with gas-sampling ports connected to the ethylene gas analyzer (model F-900, based on electrochemical sensors; Felix Instruments). After 60 min, the headspace air was sampled and measured under polarcept settings following the manufacturer's recommendations. Ethylene measurements were obtained as a mean of three readings sampled at 10-min intervals.

**Statistics.** To determine differentially expressed genes, we used DESeq2 to model raw gene counts for gene expression on a negative binomial model.  $P$  values were obtained from a two-tailed test with Benjamini–Hochberg multiple-hypothesis correction. To determine dysregulated pathways based on gene expression, we used GSEA, which applies a Kolmogorov–Smirnov-like statistic.  $P$  values were derived from a permutation test to generate empirical null distributions. To determine significant pathways with gene family expansion, we used a two-tailed Fisher's exact test with Benjamini–Hochberg

multiple-hypothesis correction. All *P* values were considered significant if below a threshold of 0.05.

**Data availability.** All custom scripts have been made available at <https://github.com/chernycherny/FSFix>. Raw reads and genome assembly have been deposited as a BioProject under accession [PRJNA400310](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA400310). A **Life Sciences Reporting Summary** is available.

53. Ching, T., Huang, S. & Garmire, L.X. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* **20**, 1684–1696 (2014).
54. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
55. Marie, D. & Brown, S.C. A cytometric exercise in plant DNA histograms, with 2*C* values for 70 species. *Biol. Cell* **78**, 41–51 (1993).
56. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
57. Zaharia, M. *et al.* Faster and more accurate sequence alignment with SNAP. Preprint at <https://arxiv.org/abs/1111.5572> (2011).
58. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
59. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
60. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
61. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
62. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
63. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
64. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at <https://arxiv.org/abs/1207.3907> (2012).
65. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
66. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
67. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
68. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53**, 661–673 (2008).
69. Anders, S., Pyl, P.T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).



## Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### ► Experimental design

#### 1. Sample size

Describe how sample size was determined.

Our sample sizes and read depths enabled significance testing of differentially-expressed genes and gene sets (by DESeq2 and GSEA).

#### 2. Data exclusions

Describe any data exclusions.

As per a Reviewer's request, cultivars that did not have at least 3 biological replicates were removed from this study. This did not change the conclusions of our study.

#### 3. Replication

Describe whether the experimental findings were reliably reproduced.

All replications were successful.

#### 4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

For transcriptomic analysis 1, samples were grouped based on their tissue type (fruit aril vs non-fruit [stem/root/leaf]).  
For transcriptomic analysis 2, samples were grouped based on their tissue type (fruit aril vs other species' fruits).  
For transcriptomic analysis 3, samples were grouped based on their cultivar (Musang King vs non-Musang King [Monthong and Puang Manee]).

#### 5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No blinding was used, as no randomized controlled trials were conducted, and measurements were not prone to observer bias.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

#### 6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

- 1/a Confirmed
- ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
  - ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
  - ☒ A statement indicating how many times each experiment was replicated
  - ☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
  - ☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
  - ☒ The test results (e.g.  $p$  values) given as exact values whenever possible and with confidence intervals noted
  - ☒ A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
  - ☒ Clearly defined error bars

See the web collection on [statistics for biologists](#) for further resources and guidance.

## ► Software

Policy information about [availability of computer code](#)

### 7. Software

Describe the software used to analyze the data in this study.

Custom code: FSFix, deposited in Github.

Other software (all cited in manuscript): DESeq2, GSEA v3.0, CANU v1.3, FALCON v0.3.0, trim\_galore, BFC, SoapDenovo2, SparseAssembler, DBG2OLC, OPERA-LG, Jellyfish v2, HiRise, SNAP (read mapper), HiSat, StringTe v1.2.3, RepeatModeler v1.0.10, RepeatMasker v4.0.6, Maker v2.31.9, Snap v11-29-2013 (gene prediction), Augustus v2.5.5, Blast2GO v4.1.9, InterproScan v5.21, BWA-MEM v0.7.5, SAMtools v0.1.19, PicardTools v2, FreeBayes v0.9.21, OrthoMCL v2, Muscle, Gblocks v0.91b, BEAST2, MCScan v0.8, CoGe, HTSeq-Count v0.9.1.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* [guidance for providing algorithms and software for publication](#) may be useful for any submission.

## ► Materials and reagents

Policy information about [availability of materials](#)

### 8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used.

### 9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

### 10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used.

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

## Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

### 11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about [studies involving human research participants](#)

### 12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Study did not involve human participants.