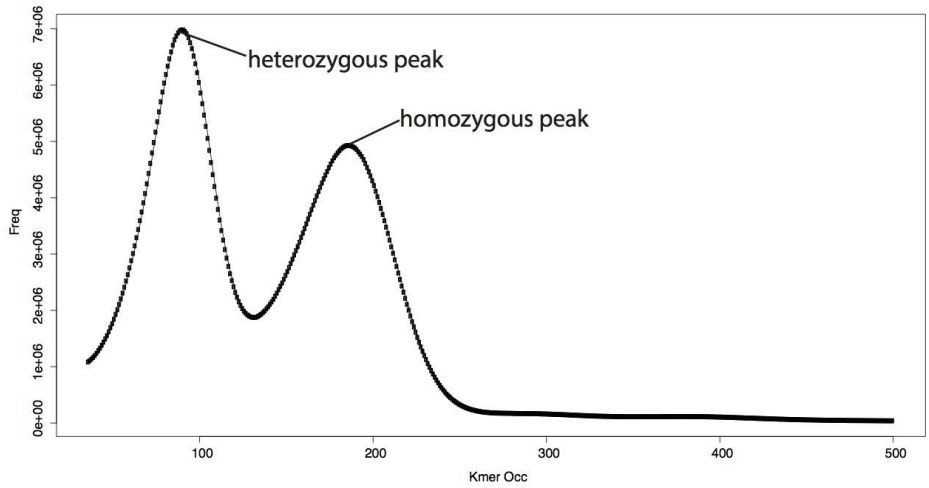
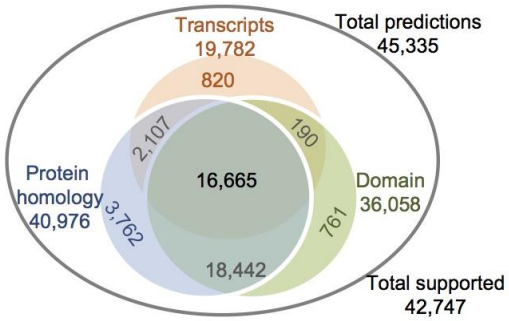


Supplementary Figure 1

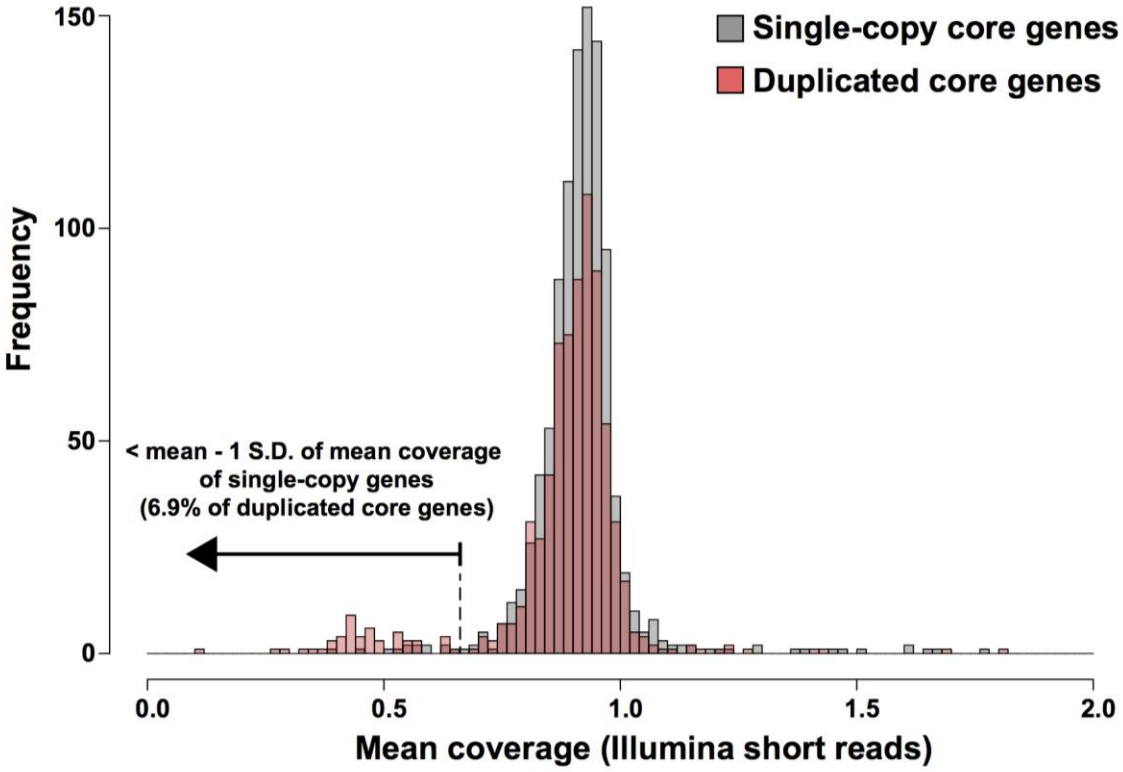
K-mer distribution analysis for genome size and heterozygosity estimation.



Supplementary Figure 2

Gene prediction support.

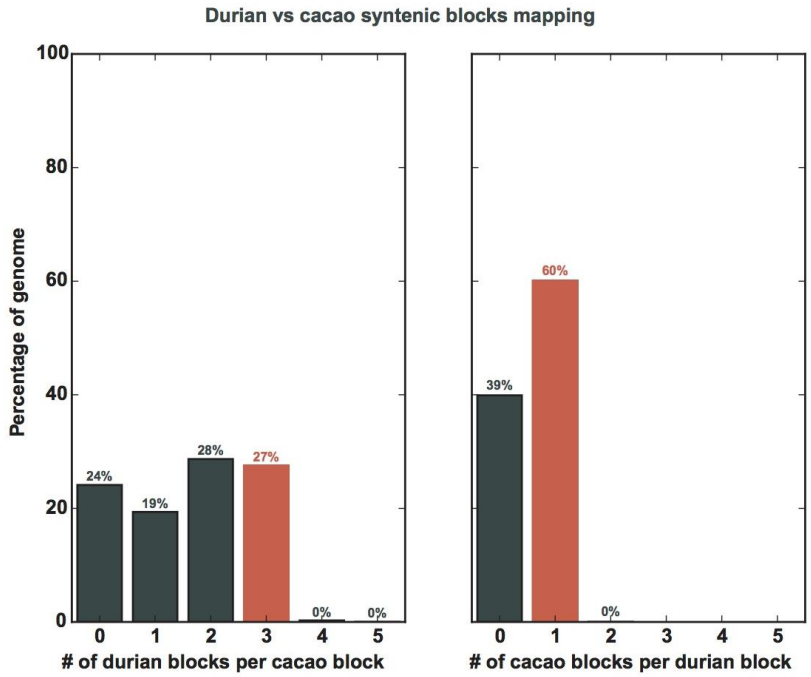
Number of predicted genes supported by RNA-seq transcripts, homology to known proteins, or functional domains.

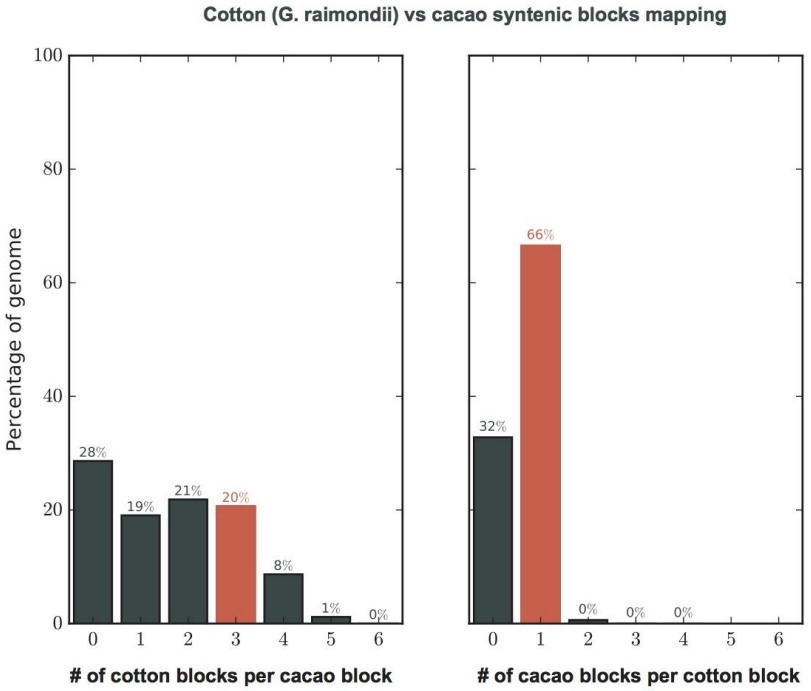


Supplementary Figure 3

Core gene coverage.

Histograms of mean coverage (Illumina short reads) of single-copy core genes (gray) and duplicated core genes (red). The vast majority of duplicated core genes (93.1%) had mean coverage within 1 standard deviation of the mean coverage of single-copy core genes.

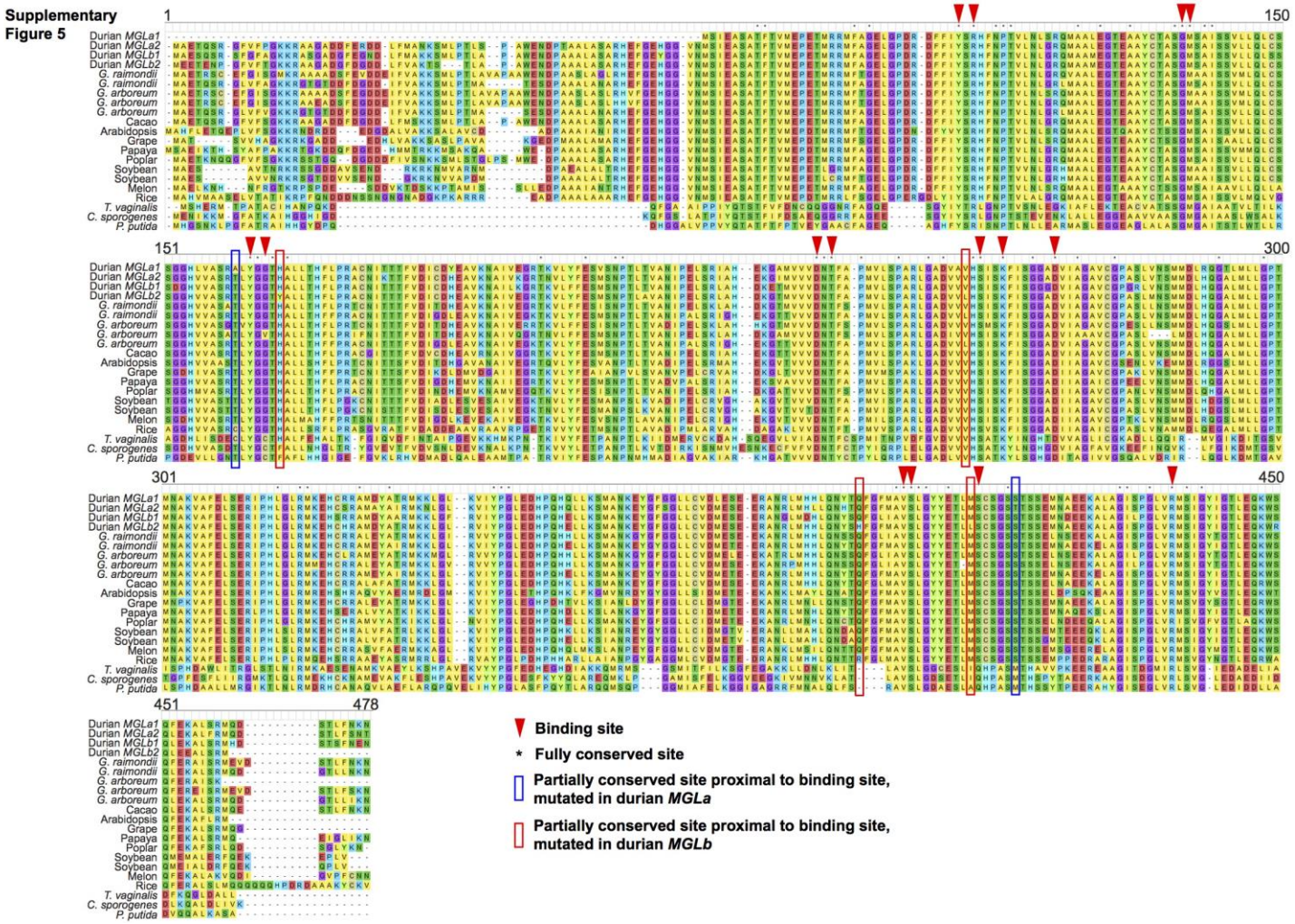




Supplementary Figure 4

Syntenic block mapping.

Distribution of duplication depths of syntenic regions between cacao-durian, and cacao-cotton.



Supplementary Figure 5

Alignment of MGL proteins.

Supplementary Note

Secondary assembly and validation of genome assembly

A second assembly was performed using a hybrid approach. Illumina short reads were preprocessed using trim_galore[1] to remove contaminating sequences from adapters as well as sequences with low base quality, and were corrected for errors using BFC[2]. The reads were assembled into contigs using SoapDenovo2[3] (k=51) and SparseAssembler[4] (k=31, g=15). Based on contig N50 results, SparseAssembler contigs were selected for hybrid assembly with long PacBio reads using the hybrid assembler DBG2OLC[5] (k=17). Consensus sequences were generated with the program PBDAGCON, run using the script split_and_run_pbdagcon.sh provided with DBG2OLC. The hybrid assembly was further scaffolded using synthetic mate-pairs derived from PacBio reads and OPERA-LG[6] (default parameters). Collapsed repeats in the assembly were identified based on 1.5X or greater Illumina read coverage compared to the genomic average.

We implemented a custom frameshift-correction tool, FSFix, which corrects many of the remaining insertions or deletions (indels) in homopolymer regions, and indels caused by reads that were derived from heterozygous haplotigs. First, FSFix aligns the Illumina short reads onto the assembly, keeping only unique alignments with mapping quality >20. Next, FSFix looks indels in the alignment supported by at least 90% of the aligned reads, and with at least 5 supporting reads. Reads for which the indel under consideration falls within the first 5 or last 5 bases are ignored in calculating this support. Finally, FSFix corrects the assembly using these supported indels: supported deletions in the aligned reads are used to delete base(s) from the assembly, while supported insertions in the aligned reads are used to insert base(s) into the assembly. FSFix also looks for a special case of indels likely derived from reads assembled from heterozygous haplotigs, where the assembly contains both alleles from the heterozygous haplotigs. This is manifested as two nearby deletions in the aligned reads. This is not detected by FSFix in the general case, because each deletion would only be supported by about 50% of reads. Thus, if FSFix finds two deletions, X1 and X2, that are are within 3bp of each other, and X1's deletion-supporting reads are distinct from X2's deletion-supporting reads, FSFix considers X1 and X2 as the same deletion event, taking its supporting reads as the union of X1 and X2's supporting reads, and calculates if the deletion is supported as before. If so, it deletes only one of either X1 or X2. FSFix corrected 1255 indels in our assembly.

We validated the 712 Mb final assembly in two ways. First, 89.8% of Illumina short reads sequenced from the same fruit stalk mapped to 98.79% of the assembly, indicating a high level of completeness and accuracy (Supplementary Table 1). Subsequently these reads were used to polish the assembly using Pilon[10] and FSFix, a custom tool to correct remaining indel errors (Supplementary Figure 2). Second, 92.2% of a secondary assembly (535 Mb), based on

a prior PacBio and Illumina hybrid-based assembly approach[11-13] from an independent fruit stalk, aligned to 67.6% of the final assembly, demonstrating concordance between assemblies and also additional regions realized in our final assembly. We also verified the completeness and proper merging of haplotigs using BUSCO[14] (see below).

Estimation of genome size and ploidy

We note that our k-mer distribution with two distinct peaks (Supplementary Figure 1) could theoretically indicate either an autotetraploid genome with low heterozygosity (such that the first peak corresponds to regions of difference between sub-genomes), or a diploid genome (such that the first peak corresponds to heterozygous regions). An autotetraploid genome with high heterozygosity, or an allotetraploid genome, would have manifested as 3 or 4 main peaks in the k-mer distribution (corresponding to 0.25X, 0.5X, and 1X coverage, and possibly 0.75X coverage), so these cases can be eliminated. We rely on two lines of evidence to show that our durian sample is likely diploid:

1. Mapping of the syntenic regions in the durian assembly to itself showed extensive shuffling of syntenic genomic regions, rather than a one-to-one correspondence expected from an autotetraploid genome with low heterozygosity. We infer that while the durian lineage underwent a whole genome duplication (WGD) event ~60 MYA (from our analysis), leading to large duplicated syntenic regions in the genome, the genome has since undergone extensive shuffling and diploidization.
2. Our assembly process used Falcon-Unzip to merge haplotigs into a single locus in the assembly. With a heterozygosity rate of 1.0%, most haplotigs regardless of ploidy should be successfully merged by Falcon-Unzip into a single locus. The assembly thus represents a merged-haplotig representation of the genome, with an assembly size of 715 Mb. This genome size is consistent with our estimates of genome size by flow cytometry assuming a diploid genome (800 Mb); in contrast, assuming a tetraploid genome would estimate the genome size at 400 Mb by flow cytometry, which would be inconsistent with our haplotig-merged assembly size. While many factors affect the completeness and correctness of haplotig merging in Falcon-Unzip, nonetheless it would require that Falcon-Unzip fail to merge ~50% of haplotigs to obtain a genome size (~350 Mb) consistent with a genome size estimate based on a tetraploid genome. Thus, we believe our sample to be diploid, and the two distinct peaks in k-mer analysis likely correspond to diploid heterozygous and homozygous regions.

Details on running Maker for genome annotation

The first iteration of Maker used these evidences to generate an initial set of gene predictions. These were filtered to keep only non-overlapping gene predictions with correct start and stop

codons and splice junctions, and used for training by the Snap[7] and Augustus[8] *ab initio* gene prediction programs. The second iteration of Maker used the trained Snap and Augustus programs in hint-based mode, with the protein and transcript evidences as hints. The predictions were again filtered and used for training by Snap and Augustus, and Maker was run for a final iteration, generating the final set of gene predictions.

Transcriptomic comparison against other plant species

To compare gene expression of the same genes across 6 different plants (durian, banana, avocado, blueberry, tomato, mango), we mapped the genes of the different plants to representative *Arabidopsis* genes. Representative *Arabidopsis* genes were obtained as follows. Among *Arabidopsis* genes that belong to the same gene family (by OrthoMCL analysis), one single *Arabidopsis* gene is chosen as the representative gene. *Arabidopsis* genes that do not belong to gene families are also considered representative genes. For each of these 6 plants under analysis, for each gene, we obtained the best BlastP hit to a representative *Arabidopsis* gene.

In our study, the plants used for transcriptomic comparison were selected using four criteria: First, the plants should be associated with fleshy fruits. Second, transcriptomic data for these plants should be publicly available. Third, while most of the fruits used (4 out of 5) are climacteric, they were also chosen on the basis of their taste profiles, as we were interested in investigating the transcriptomic rationales for their flavours and odors. Fourth, we also chose fruits that were profiled for RNA-seq using similar sequencing technology to reduce biases attributed to using different technology platforms.

Substitution models for phylogeny analysis

We performed substitution model selection using bModelTest[9]. Although no single model was consistently the most optimal across all partitions, HKY was the most frequently found among the top 5 models across all partitions. We thus used the HKY substitution model with 4 rate categories. To model varying mutation rates among different lineages, we used the uncorrelated log-normal relaxed clock model. We incorporated calibration times for the Eudicots stem node (125 MYA)[10], Brassicales stem node (89 MYA)[10], and Malvathecas stem node (60 MYA)[11]. We represented each calibration date as the minimum age of the node and allowed for uncertainty, by modeling each calibration node age as a log-normal distribution with mean = date + 10% of date, and standard deviation = 1. We performed multiple independent MCMC runs (30 runs of 20 M chain length each). Each run was analyzed with the Tracer module from the BEAST2 package, and any run with poor convergence was re-run. The combined results showed an effective sample size of at least 200 for all parameters, and the DensiTree module from the BEAST2 package showed that every clade in the final tree topology had more than 99% support.

References

1. Krueger, F. Trim Galore! http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
2. Li, H. BFC: correcting Illumina sequencing errors. *Bioinformatics* **31**, 2885-2887 (2015).
3. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 1-6 (2012).
4. Ye, C., Ma, Z.S., Cannon, C.H., Pop, M. & Yu, D.W. Exploiting sparseness in de novo genome assembly. *BMC Bioinformatics* **13 Suppl 6**, S1 (2012).
5. Ye, C., Hill, C., Ruan, J. & Zhanshan, M. DBG2OLC: Efficient Assembly of Large Genomes Using the Compressed Overlap Graph. *ArXiv*, 1410.2801 (2014).
6. Gao, S., Bertrand, D., Chia, B.K.H. & Nagarajan, N. OPERA-LG: efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees. *Genome Biology* **17**, 102 (2016).
7. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**(2004).
8. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62-62 (2006).
9. Bouckaert, R.R. & Drummond, A.J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evolutionary Biology* **17**, 42 (2017).
10. 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 Phytologist* **207**, 437-453 (2015).
11. Carvalho, M.R., Herrera, F.A., Jaramillo, C.A., Wing, S.L. & Callejas, R. Paleocene Malvaceae from northern South America and their biogeographical implications. *American Journal of Botany* **98**, 1337-1355 (2011).