

Supplementary Information

1. Supplementary notes 1 – 3
2. G+C content
3. Transposable elements
4. Introns
5. Telomeres and subtelomeric regions
6. tRNA genes
7. snRNA genes
8. Organelle genomes
9. Cytochrome P450 monooxygenase genes
10. Transcription factors
11. Photosynthesis
12. Summary of major findings

Methods

13. Differentiation of true organellar and organellar-like nuclear sequence
14. Comparative analysis of sequenced plant genomes
15. Analysis of conserved 5'-UTR regions
16. Gene tribe analysis
17. Papaya circadian clock orthologs alignments
18. Analysis of transgenic insertions

References

Supplementary Tables 1-13

Supplementary Figures 1-14

Supplementary Note 1.

Estimate of residual heterozygosity in SunUp. Arachne was used to estimate the level of heterozygosity. Single nucleotide polymorphisms and insertions/deletions were included in the analysis if at least two reads disagreed with the consensus but agreed with each other. The total number of bases in the analysis refers to regions of at least 4X sequence coverage.

Supplementary Note 2.

Genetic and physical maps. A high density genetic map was constructed using an F₂ population derived from a dioecious variety AU9 and the sequenced Hawaiian cultivar SunUp. SSR markers derived from BAC end and WGS sequences were used so the physical map and assembled WGS sequences can be easily anchored to the genetic map. A total of 707 markers were mapped into nine major and three minor linkage groups (LGs), including 706 microsatellite markers and a morphological marker fruit flesh color. Linkage groups 8 and 10 were merged because each possessed a subset of SSR markers associated with the same scaffold. Linkage groups 9 and 11 were merged

based on Fluorescence *In Situ* Hybridization (FISH) results (Supplementary Figure 1). This map spanned 1037.1 cM with an average distance of 1.45 cM between adjacent markers.

The entire set of 39,168 BAC clones of the papaya BAC library was fingerprinted using the High-Information-Content Fingerprinting system¹. One fifth of these fingerprints were excluded due to empty insert clones, incomplete restriction enzyme digestion, highly repetitive sequences, or failure to size on the capillary sequencer. A total of 30,824 fingerprints, estimated as 11X genome equivalents, were used to construct a papaya physical map. After automated overlap evaluation and manual review, 26,466 papaya BAC clones were assigned to 963 contigs. A total of 4,358 singleton clones could not be assigned to the fingerprint contigs. The three largest contigs included over 200 BACs, whereas 204 contigs contain only two BACs. The remainder of the 756 contigs contained 3 to 199 BACs.

Supplementary Note 3.

Southern blots suggested a band with a slight reaction to a fragment of the gentamycin resistance gene. However, we have not been able to confirm its presence by PCR, by analyses of BAC and fosmid clones generated from SunUp DNA, and by analysis of WGS sequences. High-throughput MUMmer analysis also revealed un-assembled singleton sequences corresponding to a fragment of the *uidA* (GUS) gene, and to a fragment of the PRSV coat protein gene. Both singletons had border sequences at least on one side, but these inserts have not been confirmed by other means such as PCR and by obtaining the border sequences of both sides.

G+C content

Papaya has an AT-rich genome, with an overall G+C content of 35.3%, nearly identical to that of *Arabidopsis* at 35%. Coding exons contain significantly higher G+C content, averaging 44.4% (Supplemental Fig. 16). The G+C excess is likely due to a combination of factors, notably the exclusion of stop codons in coding regions as well as other codon biases. If introns are included, the G+C content of transcripts excluding UTRs drops to 37.5%.

Transposable elements

More than 43.4% of the papaya genome is homologous to identifiable transposable elements (TEs) and an additional 8.5% is covered by repetitive sequences that are currently unannotated but are likely to be novel TEs (Supplementary Table 11). Given the fact that un-assembled gaps are largely composed of TEs, this could be an underestimate. Most of the major types of repeats in Repbase² are represented in papaya, with the dominant class being retrotransposons (40% of the genome) and some of the abundant types being *Ty3-gypsy* (27.8%) and *Ty1-copia* (5.5%) retrotransposons. An interesting feature of papaya is the relatively low abundance of known DNA transposons (0.20%) compared to other plant genomes. The papaya genome is dominated by papaya-specific TE families, and when these are excluded, only 14% of the genome is covered by known TEs.

Introns

We used EST-confirmed introns to estimate the intron size of papaya. The average intron length in papaya is 479 bp, the longest among plants sequenced to date (Table 1).

Comparison of 4,403 orthologous gene pairs shows that papaya introns are over twice as long as those in *Arabidopsis*, but only slightly longer than rice. These comparisons also indicate fewer introns in the papaya genes than in their orthologs in *Arabidopsis* (8,319 versus 11,718) and rice (6,874 versus 7,011).

Telomeres and subtelomeric regions

Papaya telomeres, like most other plant telomeres, are composed of tandem arrays of (TTTAGGG)_n repeats. Terminal restriction fragment analysis revealed that papaya telomere tracts are relatively long, stretching beyond 10 kb (E. Shakirov and D. Shippen, unpublished data), more than twice the size of *Arabidopsis* telomeres³. In several cases, sequenced ESTs map as close as 6-8 kb to the telomeric repeats. Additionally, sequences resembling DNA transposons and Copia- and Gypsy-type LTR elements are associated with the subtelomeric regions of 11 of the 13 chromosome arms that have been analyzed. Strikingly, these elements reside as close as 1.5 kb upstream of the telomere tracts. Furthermore, stretches of 1-1.5 kb of the subtelomere sequence immediately flanking telomeres are highly conserved, and display only a few nucleotide substitutions, deletions or insertions. Portions of these subtelomeric sequences are also found elsewhere in the genome. The papaya subtelomeric sequences are reminiscent of those in *Plasmodium falciparum* (malarial protozoan), where up to 120 kb of nearly identical subtelomeric sequence are present on all chromosome arms, consistent with frequent inter-chromosomal exchange⁴.

tRNA genes

tRNAscan-SE analysis⁵ yields 388 putative tRNA genes or pseudogenes. Predicted tRNA genes with introns include 8 Met (9 without introns), 2 Ser (27 without), and 6 Tyr (3 without). There is some clustering of tRNA genes, for example 24 predicted tRNA genes (19 Cys, 2 Tyr, and 3 pseudogenes) reside within a 9435 bp segment on scaffold 8. For comparison, *Arabidopsis* is predicted to contain 639 tRNA genes, including 8 predicted pseudogenes, and 11 Met, 2 Ser, and 70 Tyr genes with introns (<http://lowelab.ucsc.edu/GtRNAdB/Athali/>).

snRNA genes

We annotated a total of 47 snRNA genes (8, 16, 4, 11, 3, 1, 1, 1 and 2 genes for U1, U2, U4, U6, U11, U12, U4atac and U6atac, respectively). Although this is less than the 75 snRNA genes annotated for *Arabidopsis*, papaya has more snRNA pseudogenes (50 or more). Some of these are degenerate duplicate genes. Others are clearly processed pseudogenes, as is common in mammalian genomes⁶. Intact papaya snRNA genes share the highly conserved upstream element (USE; RTCCACAT), which occurs at 66-78 nt upstream of the end of the mature snRNA⁷. Several genes for the same snRNA often occur within 10 or 20 kb but there are only a few cases of different snRNAs occurring together (U2 with U6 and U1 with U5).

Organelle genomes

The WGS reads included both chloroplast (12%) and mitochondrial (7%) DNA. The papaya chloroplast genome is 160,100 bp long and encodes 78 protein, 29 tRNA and 4 rRNA unique genes. Gene content and order are the same as in *Arabidopsis*.

The mitochondrial genome is 476,890 bp long and encodes 38 proteins, 22 tRNA and 3 rRNA genes. An 11,105 bp direct repeat includes *ccmFN*. In addition there are 17

direct repeats between 53 and 156 bp in length, and 16 inverted repeats between 55 and 563 bp long. In contrast to *Arabidopsis*, papaya has intact mitochondrial genes for *rps1*, *rps10*, *rps13*, *rps14*, *rps19*, *sdh3* and *sdh4*. In addition to a 12,319 bp region of DNA transferred from the plastid, there are 17 chloroplast-like regions in the mitochondrial genome between 106 and 2495 bp long having at least 70% nucleotide identity with the chloroplast genome. Gene order is not conserved between the papaya mitochondrial genome and other plant mitochondrial genomes, but *rrn5* and *rrn18* remain linked as in all plants and the clusters *atp4/nad4L* and *cob/rps14/rpl5* have conserved synteny in non-monocot angiosperms. 255,967 bp (53%) of the genome has no detectable homology to other mitochondrial genomes. Of this, 24,808 bp match non-chloroplast nucleotide sequences in the non-redundant database.

A total of 858,190 bp (0.31%) of the nuclear genome assembly matches the mitochondrial genome in segments of 37–5,741 bp (mean: 254 bp) and 63–100% identity (mean: 79%). A total of 785,954 bp (0.28%) matches the chloroplast genome in segments of 39–5,655 bp (mean: 236 bp) and 64–100% identity (mean: 84%). The longer segments from both organelles have generally higher percentage identities, suggesting that freshly transferred segments are quickly fragmented and deleted. While only 1836 bp of the chloroplast genome is not found in the nuclear assembly, 218,702 bp (46%) of the mitochondrial genome is absent.

Twenty three different intergenic regions covering a total of 32,397 bp (6.8%) of the mitochondrial genome correspond to LINE, Copia and Gypsy retroelements. This corresponds to about 12% of the regions that have no homology to other mitochondrial genomes. If all of these regions were transferred from the nucleus with its current TE proportion we would expect about 50% TE instead of 12%. This suggests that some of the intergenic mitochondrial DNA comes from non-nuclear sources, or corresponds to old transfers that have either been shuffled beyond recognition or were derived from ancestral nuclear genomes with relatively few TEs. The G+C content of homologous mitochondrial and nuclear regions, as a function of sequence similarity, suggests that transfers from the mitochondrion to the nucleus are much more common than transfers in the other direction.

The GenBank Accession numbers for the chloroplast and mitochondrial genomes are EU431223 and EU431224 respectively.

Cytochrome P450 monooxygenase genes.

One of the most striking features of the papaya genome is that it contains only 142 CYP (cytochrome P450 monooxygenase) full-length genes and 39 pseudogenes in its compiled genome, which is significantly less than the 245 full-length genes and 27 pseudogenes present in *Arabidopsis*^{8,9} (<http://www.p450.kvl.dk/p450.shtml>; <http://arabidopsis-p450.biotec.uiuc.edu/>). The fact that both of these species exist in the Brassicales indicates that the size of the CYP gene superfamily varies widely even within a single plant order. While the majority of CYP genes in *Arabidopsis* can be accounted for by only a handful of CYP subfamilies such as CYP71B (36:7 with ratios designating *Arabidopsis* vs. papaya P450 genes), CYP705A (26:0), CYP71A (16:0), CYP96A (13:2), CYP76C (7:0) and CYP702A (6:0), no evidence indicates that a whole genome duplication in *Arabidopsis* has been responsible for its larger number of CYP genes. Instead, it is more likely that biochemical diversifications in its pathways for synthesizing defense toxins and other secondary metabolites have contributed to the expansion of these CYP subfamilies in tandem arrays. Support for this is derived from the

fact that 82 of 181 papaya P450 genes (45%, including pseudogenes) exist in clusters of 2-11 genes, suggesting that tandem duplication is an important mechanism for gene family expansion in plant genomes that do not have multicopy genes produced through extensive genome duplication.

The fact that papaya contains six P450 families (CYP80, CYP92, CYP727, CYP728, CYP733, CYP736) that are present in other plant species but missing from *Arabidopsis*, emphasizes the dramatic expansion of some P450 families and loss of others that confers species-specificity on the P450-mediated pathways in individual plants. This point is also evident in the fact that, despite its more limited number of P450 genes, papaya contains two CYP73A sequences (cinnamate 4-hydroxylase in the core of the phenylpropanoid synthetic pathway) in comparison to the single CYP73A sequence present in other plants. Suggesting that some specialization exists in the papaya phenylpropanoid pathway downstream from *p*-coumarate production, one of these CYP73A sequences is most closely related (93% identity) to CYP73A42 in *Populus tremuloides* (poplar) while the other is most closely related (82% identity) to CYP73A29 in *Citrus sinensis* (sweet orange) and only 64% identical to the first papaya CYP73A sequence. Similarly, there is some elaboration of the gibberellin synthetic pathway with three CYP701A (*ent*-kaurene oxidase mediating the first three steps in GA₁₂ synthesis) and six CYP88A (kaurenoic acid oxidase mediating the next three steps) sequences in papaya and only one CYP701A and two CYP88A sequences in *Arabidopsis*. In contrast, other subfamilies (CYP85A and CYP90A-90D in brassinosteroid synthesis, CYP734A in brassinosteroid inactivation, CYP74A in jasmonate synthesis, CYP74B in hexenal synthesis) contain the same small gene numbers as in *Arabidopsis* and other plants.

Transcription factors

Approximately 2000 transcription factors (TFs) representing over 60 families have been identified in the genomes of *Arabidopsis*¹⁰, poplar¹¹, and rice¹². While papaya has fewer members for the majority of the TF families, the number of predicted MADS-box proteins in papaya is strikingly higher (171 vs. 70-111) than in the other genomes. Eighty-one of the 171 predicted MADS-box genes (47%) were located in 28 clusters in the papaya genome that presumably result from proximal duplications.

A phylogenetic analysis was performed using all predicted MADS-box genes from papaya and selected *Arabidopsis* MADS-box genes. The predicted papaya MADS-box genes fall into the same 5 groups, M α , M β , M γ , M δ (or MIKC*), and MIKC^c, as in *Arabidopsis*, poplar, and rice¹³. However, the numbers of sequences in each group (Supplementary Table 12) were significantly different among these 3 species. There are 54-67 functional or pseudo Type II MADS-box (M δ and MIKC^c groups) genes in *Arabidopsis*, poplar, and rice, but only 25 in papaya (Supplementary Fig. 17). In contrast, 145 Type I MADS-box (M α , M β , and M γ groups) genes were predicted in papaya versus 29-94 present in the other 3 species, with the largest expansion in the M α group.

Photosynthesis

In the majority of nuclear gene families encoding diverse photosynthesis functions, papaya has fewer members of the respective gene families relative to *Arabidopsis*, apparently due to decreased duplication of loci. For example, papaya has three widely-spaced RbcS genes relative to four (three clustered on chromosome 5) in *Arabidopsis*. Similarly, papaya has six LhcB loci whereas *Arabidopsis* has 13. One exception to the pattern is Rubisco activase; papaya has three, while *Arabidopsis* has two

loci. Papaya has a single copy of the chlorophyll synthase gene as do *Arabidopsis* and rice. There were 22 Toc/Tic orthologs identified in the *C. papaya* genome compared to 28 found in *Arabidopsis*.

Summary of major findings

- Papaya has fewer genes than *Arabidopsis*, with reductions in most gene families and biosynthetic pathways, making it an excellent system in which to study the function of complex biosynthetic pathways and networks.
- The lower gene number is largely because, unlike *Arabidopsis*, the papaya genome contains no recent genome-wide duplication, with fewer opportunities for subfunctionalization, implying that papaya genes may be more representative of ancestral angiosperms than *Arabidopsis* genes. This lack of a genome-wide duplication event makes papaya a valuable outgroup for comparative genomics of the Brassicales.
- Under the assumption that a generalized angiosperm plant could potentially require only the types and minimal numbers of genes that are shared among divergent plant species, we estimate that a minimal angiosperm genome would contain just 13,311 genes.
- Papaya contains significantly fewer disease resistance gene analogs than *Arabidopsis*, suggesting that papaya may have evolved alternative defense mechanisms.
- Papaya also contains significantly fewer P450 genes than *Arabidopsis*, with some subfamilies expanded, some completely absent, and others novel to the papaya genome.
- Despite reduced gene numbers in most biosynthetic pathways, the number of predicted MADS-box family members is strikingly higher (171 vs. 78 in rice and 141 in *Arabidopsis*) in papaya than in other sequenced plant genomes.
- Papaya has fewer members of gene families involved in fruit ripening, with the exception of starch synthase, possibly reflecting a need for starch storage in the stem and during early fruit development.
- Tremendous amplification in papaya of genes related to volatile development implies strong natural selection for enhanced attractants that may be key to fruit (seed) dispersal by animals and aboriginal peoples.
- Papaya contains fewer photosynthesis, circadian clock, and light-signaling genes than either poplar or *Arabidopsis*, suggesting that papaya does not require the same level of control for daily and seasonal timing.
- Genome-wide searches for transgenic sequences revealed only three insertions that could be characterized by Southern hybridization and sequencing of the bordering host DNA; these were a functional cassette with the intact PRSV coat protein gene, a fragment of the *nptII* gene, and a fragment of the *tetA* gene. None of the insertions disrupted functional genes.

Methods

Differentiation of true organellar and organellar-like nuclear sequence.

Organellar reads were initially identified by starting with organellar seed reads and growing a contig at both ends using read overlap, quality scores, and a simple linkage

algorithm. Resulting contigs could be separated into nuclear, mitochondrial or chloroplast based on read depth since the read depths of the three genomes varied greatly. Reads corresponding to either the mitochondrion or chloroplast were assembled with cap3. The resulting hypothetical organellar genomes were iteratively screened against all reads in the assembly to identify errors until no errors remained -- that is until the high quality read depth was as expected at each base and the high quality mismatches were very low or zero. Reads that matched the final organellar genomes with few or zero high quality mismatches and that had a consistent clone end partner were considered unambiguous reads for a particular organelle's genome. The nuclear assembly was screened for regions of organelle genome similarity. Similar regions were required to be anchored to non-organellar-like sequence, by read overlap and clone end partner consistency, at one or both ends, and have clone end partner consistency throughout the organellar-like region. Regions passing this screen were considered to be organellar-like nuclear sequences.

Comparative analysis of papaya and *Arabidopsis* genomes

The longest 200 papaya scaffolds were subjected to gene colinearity analysis with *Arabidopsis* chromosomes. The top 5 blast hits with E-value $< 1e-5$ were considered in inferences of colinearity. For simplicity, the papaya genes were renamed according to their positions on the scaffolds (see attached file: cp.gene.rename.txt). We developed a multiple genome/subgenome alignment program named "MC-Scanner" to investigate colinearity both between and within *Arabidopsis* and papaya genomes. Briefly, the program first scans for pairwise colinearity by a dynamic programming procedure very similar to DAGchainer¹⁴ and ColinearScan¹⁵. The related pairwise colinear segments are then grouped into blocks and progressively re-aligned. Two colinear segments are defined as "related" if 1) they overlap with the same chromosome region, or 2) they show direct colinearity. This results in a multiple alignment of several genomic regions using genes as anchors. The detailed algorithm will be described elsewhere.

DNA synteny dotplots were produced between all papaya scaffolds and *Arabidopsis* genome sequences. Using *Arabidopsis* genes as anchor sequences, we searched for matches among the papaya scaffold sequences by using BLASTN. The matched substrings with BLASTN E-value $< 1e-5$ were used to produce the bidirectional dotplot shown in both Supplementary Figure 2 and 3. For Supplementary Figure 3.

For Supplementary Figure 2, 26,528 transcription units from *Arabidopsis* (TAIR V6) were located on the papaya assembly by TBLASTN (E-value $\leq 1e-5$). The *Arabidopsis* genes were plotted by gene order vs. the basepair position within each papaya scaffold.

For Figure 2, *Multiple alignment*: we used whole genome BLASTP (top five hits, E-value $< 1e-5$) results to predict pairwise segments by dynamic programming with an empirical scoring scheme. A pairwise segment consists of two distinct genomic regions with aligned, strictly collinear genes as anchor points. We model these regions as vertices in a graph, and define an edge between any two vertices if they are either collinear, or overlap. Then, all related regions can be found by looking for connected components within the graph. Once identified, these related regions were aligned using a heuristic that constructs the multiple alignments progressively by greedily picking one closest-related region at a time. All multiple alignments were compared against expected occurrences to ensure significance (E-value < 0.001). *Chromosomal phylogeny*: We consider "gene retention" as the ancestral state versus "gene loss" as the derived state, then the multiple

alignments can be described as binary matrices. For each multiple alignment, we searched for a hierarchical tree using “Camin-Sokal” parsimony¹⁶ criterion implemented in “mix” program in PHYLIP (version 3.66)¹⁷ based on gene contents. We applied Camin-Sokal parsimony here since genes that had been lost were highly unlikely to reemerge at original paleologous locations, *i.e.* reversal to the ancestral state is prohibited. This irreversibility property makes the tree polarized *a priori*, rendering an outgroup unnecessary.

Analysis of conserved 5'-UTR regions

C. papaya-*A. thaliana* 5'-UTR region comparisons were performed using the *UntransID* script of Windsor and coworkers¹⁸. *UntransID* expects shot-gun sequencing data from the query genome; to accommodate this requirement, the papaya assembly was decomposed to randomly generated tiles approximating 5x coverage of the available sequence prior to analysis. In comparisons where presumptive *C. papaya* 5'-UTRs were referenced to the 500 nucleotides upstream of *A. thaliana* start-codons, individual tiles were 2,000 nucleotides in length. As the papaya tiles were generated from a mature sequence source, *UntransID* was run without extra sequence quality measures.

The analysis performed by *UntransID* proceeds in two phases. In the first phase, *C. papaya* tiles were screened for significant similarity (e-value $\leq 1e-10$) to *A. thaliana* CDSs using the BLAST local alignment algorithm¹⁹ and for identity traversing the translation initiation codons of the identified *A. thaliana* CDS homologs. Tiles satisfying these requirements were filtered to isolate tiles with a minimum of 500 nucleotides 5' to presumptive initiation codons. Candidate homolog-pairs were further filtered by *UntransID* to eliminate pairs where orthology could not be assigned unambiguously owing to paralogous sequences in either or both comparison species. At the conclusion of this phase, 969 candidate 5'-UTRs from *C. papaya* were deemed effectively orthologous to *A. thaliana* upstream regions (obtained from *The Arabidopsis Information Resource* ftp-site) for the purpose of further comparison.

During the second phase of the analysis, pair-wise alignments were performed between the 969 candidate orthologous 5'-UTR-pairs using the algorithm of Needleman and Wunsch²⁰ as implemented by the *needle* program of the *EMBOSS 4.0* sequence analysis suite²¹. *Needle* alignments were performed with a gap-opening penalty of 12.0 and a gap-extension penalty of 2.0. Only alignments with scores $\geq$ a threshold value (705) were deemed significant. This threshold¹⁸ was determined by adding two standard-deviations to the mean alignment score from 10,000 random alignments of the 5' regions represented in the comparison dataset (supplemental figure 1). Alignments failing this threshold value were treated as zero-identity for subsequent portions of the analysis.

From the starting pool of 969 5'-UTR orthologs, 136 alignments (14%) had scores exceeding our threshold and were treated as sharing identity. Among these informative alignments, the mean identity of *A. thaliana* upstream regions to their *C. papaya* homologs was 75.5%. As percent identity is a one-dimensional representation of sequence similarity, the proportion of *A. thaliana* nucleotides scored as having identity to a *C. papaya* nucleotide was plotted against *A. thaliana* nucleotide position (supplemental figure 2).

Similar comparisons were performed for orthologous genomic regions 1,000 and 3,000 nucleotides 5' to *A. thaliana* start-codons (data not shown). The generated *C. papaya* tiles were 2,500 and 4,000 nucleotides in length and the threshold -alignment-

scores were determined to be 1,298 and 3,368 for the 1,000-nucleotide analysis and 3,000-nucleotide analysis, respectively. At the conclusion of these analyses, only 112 alignments from the 1,000-nucleotide comparison and 58 alignments from the 3,000-nucleotide analysis were deemed significant (data not shown). These results suggest increased sequence divergence between the species with increasing distance from conserved coding regions.

Gene Tribe analysis

The predicted protein sequences from the fully sequenced genome of *Arabidopsis thaliana* (TAIR, v7), *Oryza sativa* (rice) (TIGR, v5), *Populus trichocarpa* (JGI, v1), *Vitis vinifera* (v1), were downloaded, containing 31,921, 66,710, 45,555, and 30,434 sequences respectively. The predicted proteome of papaya was added (25,312 gene models). The combined protein set consisting of 199,932 nonredundant proteins was included in an all-against-all BLASTP search²² with an E-value cutoff of $\leq e-10$. The resulting blast reports were then used to generate a similarity matrix from the transformed E-values. Finally, the similarity matrix was used to perform MCL clustering^{23, 24} at medium stringency (inflation of 3.0). Gene ontology terms were generated for each tribe using the *Arabidopsis* GO SLIM annotations provided by TAIR, and matched to the transcription factor family annotations at the Arabidopsis Transcription Factor Database (<http://arabidopsis.med.ohio-state.edu/AtTFDB/>). The 16,362 unigenes from Papaya were searched against the predicted papaya coding sequences (using BLASTN, with median e-value of 0.0) and included 9,651 non-redundant annotated gene model hits. The reason for the lower number of gene models with unigene matches was that there were 2,958 gene models with hits from more than one unigene, indicating either insufficient overlap or sequence error.

Papaya circadian clock orthologs alignments

Protein sequences were downloaded for *Populus trichocarpa* (poplar, Pt, v1.1, http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html), *Arabidopsis thaliana* (*Arabidopsis*, At, TAIR7, <http://www.arabidopsis.org>), *Physcomitrella patens ssp patens* (moss, Pp, v1.1, http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html), chestnut²⁵, and rice²⁶. All-versus-all reciprocal BLAST²⁷ was used to identify and name putative papaya-*Arabidopsis* orthologs using the “mutual-best-blast-hit” criteria²⁸. Sequence alignments and phylogenetic tree generation were carried out using ClustalW²⁹ in MegAlign (DNASTAR, Lasergene6, Madison, WI) with default settings. The length of each pair of branches represents the distance between proteins sequences. The X-axis indicates the number of substitution events between protein sequences. The dotted line indicates a negative branch length as a result of averaging to achieve a balanced branched phenogram.

Analysis of transgenic insertions

Supplemental sequence information for the PRSV coat protein transgene insertion was obtained from a plasmid subclone (pRB6) obtained from a genomic DNA library generated from BglIII endonuclease digested, Rainbow papaya genomic DNA enriched for DNA fragments over 10 kbp and cloned into the $\square$ blueSTAR replacement vector from Novagen (Gustavo Fermín, unpublished results). Supplemental sequence information for the *tetA* fragment insertion was obtained from clone 66B4 obtained from a BAC library of SunUp papaya DNA³⁰.

References

1. Luo, M.C. *et al.* High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82**, 378-389 (2003).
2. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Res.* **110**, 462-467 (2005).
3. Richards, E. J & Ausubel, F. M. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell* **53**, 127-136 (1988).
4. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002).
5. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).
6. Denison, R. A. & Weiner, A.M. Human U1 RNA pseudogenes may be generated by both DNA- and RNA-mediated mechanisms. *Mol. Cell. Biol.* **2**, 815-828 (1982).
7. Filipowicz, W., Kiss, T., Marshall, C. & Waibel, F. U-snRNA genes, U-snRNAs and U-snRNPs of higher plants. *Mol. Biol. Rep.* **14**, 125-129 (1990).
8. Paquette, S. M., Bak, S. & Feyereisen, R. Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA Cell Biol.* **19**, 307-317 (2000).
9. Schuler M.A., Duan, H., Bilgin, M. & Ali, S. *Arabidopsis* cytochrome P450s through the looking glass: a window on plant biochemistry. *Phytochem. Reviews* **5**, 205-237 (2006).
10. Guo, A. *et al.* DATF: a database of *Arabidopsis* transcription factors. *Bioinformatics* **21**, 2568-2569 (2005).
11. Zhu, Q.-H. *et al.* DPTF: a database of poplar transcription factors. *Bioinformatics* **23**, 1307-1308 (2007).
12. Gao, G. *et al.* DRTF: a database of rice transcription factors. *Bioinformatics* **22**, 1286-1287 (2006).
13. Leseberg, C. H., Li, A., Kang, H., Duvall, M. & Mao, L. Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*. *Gene* **378**, 84-94 (2006).
14. Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643-3646 (2004).
15. Wang, X. *et al.* Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinformatics* **7**, 447 (2006).
16. Camin, J. H. & Sokal, R. R. A Method for Deducing Branching Sequences in Phylogeny. *Evolution* **19**, 311-316 (1965).
17. Retief, J. D. Phylogenetic analysis using PHYLIP. *Methods Mol. Biol.* **132**, 243-258 (2000).
18. Windsor, A. J. *et al.* Partial Shotgun Sequencing of the *Boechera stricta* Genome Reveals Extensive Microsynteny and Promoter Conservation with *Arabidopsis*. *Plant Physiology* **140**, 1169-1182 (2006).
19. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402 (1997).
20. Needleman, S. B. & Wunsch, C. D. A General Method Applicable to the Search for

- Similarities in the Amino-Acid Sequence of 2 Proteins. *Journal of Molecular Biology* **48**, 443-453 (1970).
21. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276-277 (2000).
 22. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
 23. Van Dongen, S. Graph clustering by flow simulation. PhD Thesis, University of Utrecht, The Netherlands (2000).
 24. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584 (2002).
 25. Ramos, A. *et al.* Winter disruption of the circadian clock in chestnut. *PNAS* **102**, 7037-7042 (2005).
 26. Murakami, M., Ashikari, M., Miura, K., Yamashino, T. & Mizuno, T. The evolutionarily conserved OsPRR quintet: rice pseudo-response regulators implicated in circadian rhythm. *Plant Cell Physiol.* **44**, 1229-1236 (2003).
 27. Altschul, S. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389 - 3402 (1997).
 28. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).
 29. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680 (1994).
 30. Ming, R. *et al.* Construction and characterization of a papaya BAC library as a foundation for molecular dissection of a tree-fruit genome. *Theor. Appl. Genet.* **102**, 892-899 (2001).

Supplementary Table 1. Statistics of whole genome shotgun end reads and BAC end reads.

Insert size (kb)	Vector	LUCY trimmed bases (billions)	LUCY trimmed bases after removing organellar reads (billions)	Number of reads after removing organellar reads (millions)	Sequence coverage by LUCY trimmed bases	Sequence coverage by bases with quality ≥ 20	Fraction of paired reads (%)	Fraction of assembled reads (%)
3	Plasmid	1.01	0.67	0.86	1.80 X	1.64 X	95.2	88.1
6	Plasmid	0.68	0.51	0.69	1.36 X	1.25 X	93.2	85.0
86	BAC	0.02	0.02	0.03	0.06 X	0.05 X	97.3	80.8
174	BAC	0.02	0.02	0.03	0.06 X	0.05 X	95.7	84.2
total		1.73	1.22	1.61	3.27 X	2.98 X	94.4	86.7

Supplementary Table 2. Statistics of the assembled papaya genome

Number of contigs	47,483
Total length of contigs (Mb)	271
N50 of contigs (kb)	11
Number of scaffolds	17,764
Total length of scaffolds (Mb)	370
N50 of scaffolds (Mb)	1
Number of anchored contigs	20,636
Total length of anchored and oriented contigs (Mb)	117
Total length of anchored not oriented contigs (Mb)	50
Number of anchored scaffolds	291
Total length of anchored and oriented scaffolds (Mb)	161
Total length of anchored not oriented scaffolds (Mb)	74

Supplementary Table 3. Statistics of gene models predicted by six programs and the reference gene set generated by EVM. Criteria for unigene support: identity $\geq 95\%$, score ≥ 200 , E-value $\leq 1e-10$.

	Augustus	Fgenesh	Genscan	GlimmerHMM	SNAP	TWINSKAN	EVM
Total number of models	21,008	27,878	26,747	39,783	44,550	53,356	28,629
Average gene length (bp)	2,842	2,693	10,660	5,182	2,448	882	2,312
Total number of exons	99,900	128,438	125,690	110,433	176,074	137,373	114,244
Number of exons per gene	4.8	4.6	4.7	2.8	4.0	2.6	4.0
Average exon length (bp)	224	203	199	240	163	188	220
Total number of introns	78,892	100,560	98,943	70,650	131,524	84,017	85,615
Average intron length (bp)	473	488	2,629	2,543	612	253	479
Number and proportion of models with unigene support	9,298 44.3%	10,437 37.4%	8,750 32.7%	10,724 27.0%	11,538 25.9%	12,703 23.8%	10,275 35.9%
Alignment coverage (bases and proportion) in models	5,895,698 26.3%	6,445,501 24.8%	5,588,013 22.3%	5,551,268 20.9%	5,885,673 20.6%	5,943,653 23.0%	6,523,215 25.9%
Number and proportion of unigenes matching with models	11,781 72.0%	12,462 76.2%	11,990 73.3%	11,887 72.7%	12,336 75.4%	12,301 75.1%	12,442 76.0%
Alignment coverage (bases and proportion) in unigenes	6,313,799 51.2%	6,877,605 55.8%	5,949,482 48.2%	5,913,793 47.9%	6,284,944 51.0%	6,365,671 51.6%	7,014,936 56.9%

Supplementary Table 4. EST support for gene models in *Carica* genome. Counts represent the number of gene models with corresponding ESTs or Tribes with at least one corresponding EST.

	<i>Carica</i> genome	with EST	percent
Number of gene models	25312*	9648	38.1%
Number of Tribes	12958	5858	45.2%
Tribes <i>Carica</i> shares with <i>Arabidopsis</i>	6726	4871	72.4%
Tribes unique to <i>Carica</i>	5669	741	13.1%
Singleton Tribes unique to <i>Carica</i>	5314	681	12.8%

* This is one of the four gene models used for annotating the papaya genome.

Supplementary Table 5. Global classification of protein-coding genes into Tribes (corresponding to putative gene families and subfamilies; additional details in text) and comparison of Tribe size in papaya and *Arabidopsis* based on MCL Tribe¹⁻⁴ analysis of 208,901 unique gene models from papaya, *Arabidopsis*, grape, poplar, and rice.

	# of Tribes	# <i>Arabidopsis</i> Genes	# Papaya Genes	Papaya:Arab
Papaya < <i>Arabidopsis</i>	5798			
Papaya present	2153	14253	8062	0.56
Papaya absent	3645	5590	0	
Papaya = <i>Arabidopsis</i>	3594	4685	4865	1.04
Papaya > <i>Arabidopsis</i>	7211			
<i>Arabidopsis</i> present	979	2283	4536	1.99
<i>Arabidopsis</i> absent	6232	0	7666	
Total	16603	26991	25129	0.93

1. Van Dongen, S. A cluster algorithm for graphs. *Technical Report* INS-R0010 (2000).
2. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584 (2002).
3. Enright, A.J., V. Kunin, & C.A. Ouzounis. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* **31**, 4632-4638 (2003).
4. Wall, P.K, Leebens-Mack, J., Müller, K. F. Field, D. Altman, D.S., dePamphilis, C.W. PlantTribes: A gene and gene family resource for comparative genomics in plants. *Nucleic Acids Res.* **36**, D970-6 (2008).

Supplementary Table 6 (Data supporting Figure 2). Tribe distribution for 67 Transcription factor families and superfamilies in Arabidopsis and papaya. Gene counts are for current 'best models' and are likely to be adjusted with additional genomic and cDNA sequencing.

TF Family	# Tribes	# genes		CA	C=0	C:A
		Arabidopsis	Carica					
ABI3VP1	29	113	73	14	2	13	11	0.65
Alfin-like	1	7	4	1	0	0	0	0.57
AP2-EREBP	21	195	146	8	6	7	1	0.75
ARF	1	24	17	1	0	0	0	0.71
ARID	4	8	7	2	1	1	0	0.88
AUX/IAA	2	29	20	1	1	0	0	0.69
BBR/BPC	3	8	2	3	0	0	1	0.25
BES1	6	16	16	2	0	4	0	1.00
bHLH	52	154	106	31	9	12	15	0.69
bZIP	30	80	56	17	8	5	7	0.70
C2C2-CO-like	2	6	4	1	0	1	0	0.67
C2C2-Dof	2	36	19	1	0	1	0	0.53
C2C2-GATA	12	35	34	6	1	5	3	0.97
C2C2-YABBY	2	6	8	0	0	2	0	1.33
C2H2	56	128	102	28	12	16	14	0.80
C3H	83	140	110	40	17	26	21	0.79
CAMTA	1	6	7	0	0	1	0	1.17
CCAAT_HAP3	1	1	0	1	0	0	1	0.00
CCAAT_HAP5	8	29	16	4	2	2	2	0.55
CCAAT-Dr1	2	12	10	2	0	0	0	0.83
CCAAT-HAP2	9	20	8	6	0	3	4	0.40
CPP	4	8	8	1	1	2	0	1.00
CSD	1	4	3	1	0	0	0	0.75
DBP	2	38	27	2	0	0	1	0.71
DDT	4	5	6	1	2	1	1	1.20
E2F-DP	4	8	6	3	0	1	0	0.75
EIL	2	6	4	1	0	1	0	0.67
G2-like	19	85	70	9	5	5	2	0.82
GeBP	13	27	5	12	0	1	9	0.19
GRAS	6	44	47	3	1	2	0	1.07
GRF	4	9	11	1	0	3	0	1.22
HB	31	107	93	15	6	10	3	0.87
HMG	5	11	7	3	2	0	1	0.64
HRT	2	3	2	1	0	1	0	0.67
HSF	2	24	19	1	0	1	0	0.79
Jumonji	16	18	30	1	2	13	0	1.67
LFY	1	1	1	0	1	0	0	1.00
LIM	1	6	5	1	0	0	0	0.83

LUG	1	2	4	0	0	1	0	2.00
MADS	36	127	205	18	0	18	14	1.61
MBF1	1	3	2	1	0	0	0	0.67
MYB	5	9	9	1	3	1	0	1.00
MYB-related	51	242	191	24	8	19	10	0.79
NAC	23	136	95	13	1	9	10	0.70
NOZZLE	2	1	1	1	0	1	1	1.00
Orphans	38	125	96	21	9	8	9	0.77
PBF-2-like	1	3	2	1	0	0	0	0.67
PHD	80	88	122	21	18	41	10	1.39
PLATZ	2	12	10	2	0	0	1	0.83
Pseudo ARR-B	1	6	5	1	0	0	0	0.83
RWP-RK	8	216	288	4	1	3	2	1.33
S1Fa-like	5	8	7	1	3	1	0	0.88
SAP	1	1	2	0	0	1	0	2.00
SBP	5	17	16	1	1	3	0	0.94
SET	30	49	46	5	14	11	1	0.94
Sigma70-like	3	6	6	1	1	1	0	1.00
SNF2	17	50	42	7	1	9	2	0.84
SRS	2	10	5	1	0	1	0	0.50
TAZ	4	11	7	3	0	1	1	0.64
TCP	11	28	30	4	1	6	1	1.07
Trihelix	16	29	31	3	8	5	1	1.07
TUB	2	11	6	2	0	0	1	0.55
ULT	2	2	11	0	0	2	0	5.50
VOZ	1	2	2	0	1	0	0	1.00
WRKY	9	213	66	5	1	3	3	0.31
zf-HD	1	17	11	1	0	0	0	0.65
ZIM	13	16	11	8	1	4	4	0.69
Total	815	2897	2438	375	151	289	168	61.9

Supplementary Table 7. Number of circadian clock and light signaling genes in papaya (*Cp*), poplar (*Pt*) and *Arabidopsis* (*At*)

Gene family	Domain	Function	Pt	At	Cp
PRR	PRR/CCT	clock	7	5	5
ZTL/FKF1	PAS-PAC/FBOX/KELCH	clock, flowering time	5	3	1
CCA1/LHY	sMYB-A1	clock	2	2	1
RVE/ERP1	sMYB-A2	light signaling	9	6	3
LUX	sMYB-B	clock	4	5	2
PHOT	2xPAS-PAC/TRYKc	light signaling	3	2	1
ELF3	UNKNOWN	clock	2	2	1
GI	UNKNOWN	clock, flowering time	2	1	1
SRR1	UNKNOWN	clock, light signaling	3	1	1
ELF4	UNKNOWN	clock	7	5	4
TIC/TKL	UNKNOWN	clock	4	2	2
TEJ	PARG	clock	2	2	1
PHY	PHYTOCHROME	light signaling	3	5	4
CRY	DNA photolyase, FAD binding	light signaling	4	2	2
PIF/PIL	PAS/BHLH	light signaling	5	6	2
CK	casein kinase	clock	4	4	1
COP	RING/WD40	light signaling	3	1	3
SPA	WD40	light signaling	5	4	3
HY5/HYH	BZIP	light signaling	3	2	1
DET	UNKNOWN	light signaling	1	1	1
	Total genes		78	61	40

PRR: Pseudo-response regulator

CCT: CONSTANS domain

PAS-PAC: PER-ARNT-SIM domain

sMYB: single MYB DNA binding domain

Supplementary Table 8. Circadian clock and light signaling genes in *Carica papaya*

Name	Domain (SMART)	Cp scaffolds	orf	sc	AtBestBLAST	At gene	Function
ZTL	PAS/FBOX (no Kelch)	gS_ORF_55_from_scaffold_95	55	95	At5g57360	ZTL	clock, light signaling
PHOT1	PAS/PAC/PAS/PAC/STYKc	gS_ORF_13_from_scaffold_139	13	139	At3g45780	PHOT1	light signaling
LHY/CCA1	sMYB-A	gS_ORF_119_from_scaffold_57	119	57	At1g01060	LHY	clock, light signaling
RVE1	sMYB-A	gS_ORF_31_from_scaffold_178	31	178	At5g17300	RVE1	NA
ERP1	sMYB-A	gS_ORF_157_from_scaffold_7	157	7	At1g18330	EPR1, RVE7	light signaling
RVE6	sMYB-A	gLHM_ORF_101_from_scaffold_114	101	114	At3g09600	RVE6	NA
LUX	sMYB-B	gS_ORF_90_from_scaffold_81	90	81		LUX	clock
LUX4	sMYB-B	gS_ORF_61_from_scaffold_92	61	92	At3g10760	LUX4	NA
PRR5A	PRR/CCT	gS_ORF_137_from_scaffold_3	137	3	At5g24470	PRR5	clock, light, flowering
PRR5B	PRR/CCT	gS_ORF_26_from_scaffold_193	26	193	At5g24470	PRR5	clock, light, flowering
PRR7A	PRR/CCT	gLHM_ORF_447_from_scaffold_1	447	1	At5g02810	PRR7	clock, light, flowering
PRR7B	PRR/CCT	gLHM_ORF_34_from_scaffold_139	34	139	At5g02810	PRR7	clock, light, flowering
TOC1	PRR/CCT	gS_ORF_275_from_scaffold_13	275	13	At5g61380	TOC1, PRR1	clock, light, flowering
TEJ	PARg	gS_ORF_197_from_scaffold_9	197	9	At2g31840	Unknown; PARg-like	clock
CKB3	CASEIN KINASE II	gS_ORF_91_from_scaffold_98	91	98	At4g17640	CKB2	clock

GI	UNKNOWN	gS_ORF_63_from_scaffold_26	63	26	At1g22770	gI	clock, light, flowering
ELF3	UNKNOWN	gS_ORF_12_from_scaffold_78	12	78	At2g25930	ELF3	clock, light gating
ELF4A	UNKNOWN	gS_ORF_167_from_scaffold_19	167	19	At2g40080	ELF4	clock, light gating
ELF4B	UNKNOWN	gS_ORF_3_from_scaffold_415	3	415	At2g40080	ELF4	clock, light gating
ELF4-L3	UNKNOWN	gS_ORF_?_from_scaffold_25	?	25	At2g06255	ELF4-L3	NA
ELF4-L4	UNKNOWN	gS_ORF_2_from_scaffold_554	2	554	At1g17455	ELF4-L4	NA
SRR1	UNKNOWN	GS_ORF_50_from_scaffold_30	50	30	At5g59560	SRR1	clock, light
TIC	UNKNOWN	gS_ORF_123_from_scaffold_58	123	58	At3g22380	TIC	clock
TKL	UNKNOWN	gS_ORF_33_from_scaffold_52	33	52	At3g63180	TKL	NA
PHYA	gAF/PAS/PAC/STYKc	gS_ORF_171_from_scaffold_48	171	48	At1g09570	PHYA	light sensing
PHYB	gAF/PAS/PAC/STYKc	gS_ORF_35_from_scaffold_9	35	9	At2g18790	PHYB	light sensing
PHYC	gAF/PAS/PAC/STYKc	gS_ORF_27_from_scaffold_136	27	136	At5g35840	PHYC	light sensing
PHYE	gAF/PAS/PAC/STYKc	gS_ORF_138_from_scaffold_17	138	17	At4g18130	PHYE	light sensing
CRY1	DNA photolyase, FAD-binding	gS_ORF_132_from_scaffold_36	132	36	At4g08920	CRY1	light sensing
CRY2	DNA photolyase, FAD-binding	gS_ORF_67_from_scaffold_44	67	44	At1g04400	CRY2	light sensing
PIF1,PIL5	BHLH transcription factor	gS_ORF_177_from_scaffold_8	177	8	At2g20180	PIF1	light signaling
PIL1	BHLH transcription factor	gS_ORF_45_from_scaffold_57	45	57	At2g46970	PIL1	light signaling

DET1	UNKNOWN	gS_ORF_238_from_scaffold_21	238	21	At4g10180	DET1	light signaling
COP1A	RINg/WD40	gS_ORF_2_from_scaffold_970	2	970	At2g32950	COP1	light signaling
COP1B	RINg/WD40	gS_ORF_26_from_scaffold_132	26	132	At2g32950	COP1	light signaling
COP1C	RINg/WD40	gS_ORF_51_from_scaffold_152	51	152	At2g32950	COP1	light signaling
SPA1	WD40	gS_ORF_13_from_scaffold_194	13	194	At2g46340	SPA1	light signaling
SPA2	WD40	gS_ORF_49_from_scaffold_122	49	122	At4g11110	SPA2	light signaling
SPA3	WD40	gS_ORF_49_from_scaffold_80	49	80	At3g15354	SPA3	light signaling
HY5	BZIP	GS_ORF_19_from_scaffold_115	19	115	AT5G11260	HY5	light signaling

Supplementary Table 9. The potential function of genes at the X-specific region of the X chromosome

Gene#	EST ID	Genbank hit ID	Description	Amino acid identity	Amino acid similarity	E-value
1	1436811_5_O09_033	ref NP_564443.1	Exostosin family protein [<i>Arabidopsis thaliana</i>]	100/139 (71%)	118/139 (84%)	9.00E-55
2	1436948_5_E02_011	gb ABE90444.1	Unknown	176/291 (60%)	211/291 (72%)	6.00E-89
3	1437977_5_O23_082	gb ABD33301.1	N-acetylglucosaminyl transferase component [<i>Medicago truncatula</i>]	104/184 (56%)	132/184 (71%)	8.00E-53
4	PY.4540.C1.Contig4990	gb ABE78003.1	HAD-superfamily hydrolase, subfamily IIA [<i>Medicago truncatula</i>]	254/300 (84%)	268/300 (89%)	1.00E-144
5	PY.3025.C1.Contig3398	gb AAB60921.1	F5I14.11 [<i>Arabidopsis thaliana</i>]	70/118 (59%)	89/118 (75%)	4.00E-32
6	1440261_5_O03_002	ref NP_177924.3	Trna modification gtpase, putative [<i>Arabidopsis thaliana</i>]	91/109 (83%)	99/109 (90%)	1.00E-42
7	1567097_5_O23_082	gb ABO79787.1	Rabgap/TBC	61/111 (54%)	68/111 (61%)	1.00E-26
8	PY.2093.C1.Contig2435	ref NP_198832.1	EMB506 (EMBRYO DEFECTIVE 506); protein binding [<i>Arabidopsis thaliana</i>]	115/170 (67%)	132/170 (77%)	6.00E-80
9	1607940_5_M18_067	gb AAN85202.1	Hypothetical protein	155/247 (62%)	191/247 (77%)	1.00E-82
10	1442203_5_P01_001	ref NP_174767.2	TP-dependent protease La (LON) domain-containing protein [<i>Arabidopsis thaliana</i>]	116/274 (42%)	155/274 (56%)	1.00E-45
11	1445149_5_J19_072_PY.cl.3397.single	emb CA170374.1	Alpha 1,4 fucosyltransferase [<i>Populus alba</i> x <i>Populus tremula</i>]	125/183 (68%)	145/183 (79%)	1.00E-80
12	1445612_5_N02_003	gb AAX96727.1	Expressed protein [<i>Oryza sativa</i> (japonica cultivar-group)]	32/118 (27%)	61/118 (51%)	4.00E-06
13	1608368_5_O14_049	gb AAP51059.1	Latex cyanogenic beta glucosidase [<i>Hevea brasiliensis</i>]	97/125 (77%)	110/125 (88%)	1.00E-52
14	PY.1537.C1.Contig1832	gb AAC50041.1	Poly(A) polymerase	97/138 (70%)	113/138 (81%)	4.00E-46
15	1446838_5_A04_016_PY.cl.3833.single	ref NP_177348.1	GCN5-related N-acetyltransferase (GNAT) family protein [<i>Arabidopsis thaliana</i>]	93/154 (60%)	115/154 (74%)	9.00E-48
16	1572960_5_D06_029_PY.cl.4540.single	ref NP_198495.1	Phosphoglycolate phosphatase, putative [<i>Arabidopsis thaliana</i>]	93/111 (83%)	99/111 (89%)	3.00E-48
17	PY.1818.C1.Contig2145	dbj BAF01640.1	Hypothetical	184/285 (64%)	218/285 (76%)	1.00E-101
18	1450386_5_D24_094	emb CAN81434.1	Hypothetical protein [<i>Vitis vinifera</i>]	22/29 (75%)	28/29 (96%)	4.00E-05
19	PY.1993.C1.Contig2330	ref NP_198844.1	Era1 (enhanced response to aba 1)	272/443 (61%)	334/443 (75%)	1.00E-156
20	1450057_5_G07_026	ref NP_001047387.1	Os02g0608400 [<i>Oryza sativa</i> (japonica cultivar-group)]	103/244 (42%)	137/244 (56%)	1.00E-43

21	1453673_5_M23_084	ref NP_173676.2	ATSAC1 (SUPPRESSOR OF ACTIN 1); phosphoinositide 5-phosphatase [<i>Arabidopsis thaliana</i>]	197/254 (77%)	222/254 (87%)	1.00E-108
22	PY.3273.C1.Contig3647	ref NP_565393.1	Shikimate kinase family protein	107/175 (61%)	140/175 (80%)	2.00E-55
23	1455787_5_F01_011	ref NP_564161.1	Phosphoglycerate/bisphosphoglycerate mutase family protein [<i>Arabidopsis thaliana</i>]	152/177 (85%)	167/177 (94%)	5.00E-86
24	1456599_5_G21_089	ref NP_194739.1	Trna-splicing endonuclease positive effector-related [<i>Arabidopsis thaliana</i>]	93/127 (73%)	107/127 (84%)	1.00E-47
25	1457532_5_N18_067	ref NP_176707.3	Heat shock protein binding / unfolded protein binding [<i>Arabidopsis thaliana</i>]	204/293 (69%)	247/293 (84%)	1.00E-116
26	1608308_5_M02_003	ref NP_001066320.1	Os12g0182300 [<i>Oryza sativa</i> (japonica cultivar-group)]	80/185 (43%)	119/185 (64%)	5.00E-36
27	PY.4065.C1.Contig4492	ref NP_189420.1	Monodehydroascorbate reductase 4	374/439 (85%)	402/439 (91%)	1.00E+00
28	PY.1017.C1.Contig1241	gb ABC86745.1	Pollen-specific protein []	99/159 (62%)	127/159 (79%)	3.00E-53
29	PY.1235.C1.Contig1498	gb AAG51141.1 AC069273_12	Deoxyguanosine kinase, putative [<i>Arabidopsis thaliana</i>]	190/245 (77%)	213/245 (86%)	1.00E-106
30	PY.1753.C1.Contig2066	ref NP_177929.1	Glycosyl hydrolase family 3 protein [<i>Arabidopsis thaliana</i>]	211/283 (74%)	250/283 (88%)	1.00E-123
31	PY.4322.C1.Contig4761	gb ABE84843.1	Peptidase S1 and S6	212/290 (73%)	233/290 (80%)	1.00E-103
32	PY.1789.C1.Contig2112	gb ABO14801.1	Acetyl coa carboxylase [<i>Camellia sinensis</i>]	149/167 (89%)	160/167 (95%)	1.00E-82
33	PY.4506.C1.Contig4956	emb CAN76102.1	Hypothetical protein	258/312 (82%)	286/312 (91%)	1.00E-146
34	PY.482.C1.Contig617	emb CAN61845.1	Hypothetical protein	294/315 (93%)	306/315 (97%)	1.00E-166
35	PY.2059.C1.Contig2399	sp P53393 SUT3_STYHA	Low affinity sulfate transporter 3	294/390 (75%)	343/390 (87%)	0.00E+00
36	PY.2091.C1.Contig2433	ref NP_564447.1	Permease-related [<i>Arabidopsis thaliana</i>]	208/246 (84%)	225/246 (91%)	1.00E-115
37	PY.5267.C1.Contig5713	gb AAB65822.1	Carbonic anhydrase	230/298 (77%)	258/298 (86%)	1.00E-124
38	PY.2402.C1.Contig2758	ref NP_001052831.1	Os04g0432500 [<i>Oryza sativa</i> (japonica cultivar-group)]	150/234 (64%)	189/234 (80%)	1.00E-116
39	PY.5403.C1.Contig5852	gb AAW67545.1	ADP-ribosylation factor	180/192 (93%)	188/192 (97%)	1.00E-102
40	PY.2508.C1.Contig2862	emb CAJ65923.1	Xylan 1,4-beta-xylosidase [<i>Populus alba</i> x <i>Populus tremula</i>]	116/196 (59%)	143/196 (72%)	4.00E-61
41	PY.6078.C1.Contig6483	gb ABE90442.2	Zinc finger, RING-type	77/171 (45%)	100/171 (58%)	3.00E-26

42	PY.2721.C1.Contig3077	gb AAC27138.1 AAC27138	Contains similarity to dnaj gene YM8520.10 gb 825566 from from <i>S. cerevisiae</i> cosmid gb Z49705. Ests gb Z47720 and gb Z29879 come from this gene. [<i>Arabidopsis thaliana</i>]	68/136 (50%)	91/136 (66%)	4.00E-33
43	PY.6375.C1.Contig6744	sp O49939	Peptidyl-prolyl cis-trans isomerase	130/146 (89%)	139/146 (95%)	5.00E-72
44	PY.2808.C1.Contig3169	ref NP_565037.1	Zinc finger (C3HC4-type RING finger) family protein [<i>Arabidopsis thaliana</i>]	142/185 (76%)	163/185 (88%)	1.00E-83
45	PY.2985.C1.Contig3360	gb ABE87398.1	Zinc finger, C2H2-type [<i>Medicago truncatula</i>]	57/101 (56%)	67/101 (66%)	2.00E-25
46	PY.6552.C1.Contig6901	>ref NP_568712.1	Amine oxidase-related	166/204 (81%)	185/204 (90%)	1.00E-114
47	PY.3009.C1.Contig3382	gb AAG51896.1 AC023913_4	DNA polymerase type I, putative; 54894-56354 [<i>Arabidopsis thaliana</i>]	50/125 (40%)	68/125 (54%)	2.00E-16
48	PY.6574.C1.Contig6919	ref NP_001061060.1	Os08g0162600	172/226 (76%)	190/226 (84%)	3.00E-94
49	PY.3256.C1.Contig3629	dbj BAD32780.1	Somatic embryogenesis receptor kinase 1 [<i>Citrus unshiu</i>]	287/290 (98%)	288/290 (99%)	1.00E-164
50	PY.3429.C1.Contig3805	gb AAM60857.1	Dihydroliipoamide S-acetyltransferase, putative [<i>Arabidopsis thaliana</i>]	207/222 (93%)	215/222 (96%)	1.00E-112
51	PY.3703.C1.Contig4101	ref NP_564444.1	Unknown protein [<i>Arabidopsis thaliana</i>]	115/163 (70%)	140/163 (85%)	2.00E-64
52	PY.3742.C1.Contig4140	dbj BAF62149.1	C2-H2 zinc finger protein [<i>Arabidopsis thaliana</i>]	254/353 (71%)	283/353 (80%)	1.00E-155
53	PY.4690.C1.Contig5140	gb ABA08442.1	Neutral/alkaline invertase [<i>Manihot esculenta</i>]	325/341 (95%)	335/341 (98%)	0.00E+00
54	PY.5023.C1.Contig5476	ref NP_001051835.1	Os03g0838100 [<i>Oryza sativa</i> (japonica cultivar-group)]	181/286 (63%)	216/286 (75%)	1.00E-105
55	PY.5176.C1.Contig5632	ref NP_849877.1	Unknown protein [<i>Arabidopsis thaliana</i>]	60/100 (60%)	66/100 (66%)	8.00E-23
56	PY.5283.C1.Contig5728	gb AAF86552.1 AC069252_11	F2E2.22 [<i>Arabidopsis thaliana</i>]	40/65 (61%)	51/65 (78%)	2.00E-12
57	PY.5489.C1.Contig5935	gb AAK92745.1	Putative ABC transporter protein [<i>Arabidopsis thaliana</i>]	171/279 (61%)	204/279 (73%)	1.00E-92
58	PY.5857.C1.Contig6279	gb AAM63181.1	Unknown [<i>Arabidopsis thaliana</i>]	67/89 (75%)	78/89 (87%)	7.00E-31
59	PY.5991.C1.Contig6403	ref NP_564178.2	Tetrapeptide repeat (TPR)-containing protein [<i>Arabidopsis thaliana</i>]	82/150 (54%)	96/150 (64%)	4.00E-31
60	PY.6097.C1.Contig6499	gb ABE94325.1	Dienelactone hydrolase [<i>Medicago truncatula</i>]	59/99 (59%)	79/99 (79%)	7.00E-79
61	PY.6273.C1.Contig6654	emb CAB78085.1	Putative protein [<i>Arabidopsis thaliana</i>]	87/184 (47%)	102/184 (55%)	5.00E-42
62	PY.6341.C1.Contig6713	ref NP_173675.1	C2 domain-containing protein [<i>Arabidopsis thaliana</i>]	216/272 (79%)	246/272 (90%)	1.00E-130

63	PY.6386.C1.Contig6756	ref NP_174715.2	EMB2756 (EMBRYO DEFECTIVE 2756); hydrolase, acting on carbon-nitrogen (but not peptide) bonds, in linear amides [<i>Arabidopsis thaliana</i>]	210/245 (85%)	234/245 (95%)	1.00E-126
64	PY.6733.C1.Contig7055	ref NP_177342.1	Unknown protein [<i>Arabidopsis thaliana</i>]	70/170 (41%)	88/170 (51%)	2.00E-17
65	PY.6886.C1.Contig7191	gb AAL69493.1	Putative H ⁺ -transporting ATP synthase [<i>Arabidopsis thaliana</i>]	149/252 (59%)	181/252 (71%)	5.00E-70
66	PY.699.C1.Contig874	ref NP_565035.1	Radical SAM domain-containing protein / TRAM domain-containing protein [<i>Arabidopsis thaliana</i>]	236/370 (63%)	274/370 (74%)	1.00E-122
67	PY04012X1H12.fl	ref NP_174686.1	Phosphatidylinositol-4-phosphate 5-kinase family protein [<i>Arabidopsis thaliana</i>]	119/173 (68%)	139/173 (80%)	6.00E-61
68	PY04105X1H11.fl_PY.cl.2403.single	emb CAN60958.1	Hypothetical protein [<i>Vitis vinifera</i>]	67/155 (43%)	99/155 (63%)	1.00E-27
69	PY04110X1F12.fl	gb AAL87122.1 AF479279_1	SEC6 [<i>Arabidopsis thaliana</i>]	189/205 (92%)	197/205 (96%)	1.00E-104
70-1*	PY.4171.C1.Contig4603	gb ABE93069.1	Phosphoglucosamine mutase	108/154 (70%)	128/154 (83%)	7.00E-54
70-2	1435783_5_D13_061_PY.cl.4277.single	gb ABE93069.1	Phosphoglucosamine mutase	169/214 (78%)	197/214 (92%)	6.00E-92
71-A**	PY.2187.C2.Contig2538	ref NP_565942.1	RHC1A (RING-H2 finger C1A)	205/331 (61%)	236/331 (71%)	1.00E-104
71-B	PY.2187.C1.Contig2537	ref NP_565942.1	RHC1A (RING-H2 finger C1A)	104/207 (50%)	126/207 (60%)	6.00E-40
72-A	PY.679.C2.Contig844	ref NP_001056606.1	Os06g0114700	126/201 (62%)	158/201 (78%)	1.00E-65
72-B	PY.679.C1.Contig843	ref NP_001056606.1	Os06g0114700	134/212 (63%)	167/212 (78%)	5.00E-69
73-1	PY.2499.C1.Contig2851	ref NP_177343.2	Protease-associated zinc finger (C3HC4-type RING finger) family protein [<i>Arabidopsis thaliana</i>]	140/211 (66%)	171/211 (81%)	2.00E-75
73-2	1449992_5_D14_061	ref NP_177343.2	Protease-associated zinc finger (C3HC4-type RING finger) family protein [<i>Arabidopsis thaliana</i>]	67/98 (68%)	74/98 (75%)	4.00E-29
74-1	1438843_5_D01_013	gb AAF86560.1 AC069252_19	F2E2.13 [<i>Arabidopsis thaliana</i>]	94/275 (34%)	94/275 (34%)	8.00E-32
74-2	PY.2930.C1.Contig3302	gb AAF86560.1 AC069252_19	F2E2.13 [<i>Arabidopsis thaliana</i>]	135/206 (65%)	165/206 (80%)	2.00E-66
75-1	PY04020X1E06.fl	gb AAD39604.1 AC007454_3	F23M19.3 [<i>Arabidopsis thaliana</i>]	118/143 (82%)	130/143 (90%)	5.00E-58
75-2	PY.4855.C1.Contig5311	gb AAD39604.1 AC007454_3	F23M19.3 [<i>Arabidopsis thaliana</i>]	224/308 (72%)	252/308 (81%)	6.00E-116

* - number indicates different part of one gene; **-letter indicates different isoform of the same gene

Supplementary Table 10. Composition of repetitive sequences on MSY BAC and corresponding region on X chromosome

Types of repeats	7 MSY BACs						2 X BACS and scaffolds on X					
	known		papaya		total		known		papaya		Total	
	(bp)	(%)	(bp)	(%)	(bp)	(%)	(bp)	(%)	(bp)	(%)	(bp)	(%)
Retroelements	204,550	17.7	560,376	48.4	764,926	66.0	524,456	15.9	1,128,208	34.2	1,652,664	50.1
SINEs:	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
LINEs:	0	0.0	5,862	0.5	5,862	0.5	333	0.0	47,288	1.4	47,621	1.4
LTR elements:	204,550	17.7	554,514	47.8	759,064	65.5	524,123	15.9	1,080,920	32.8	1,605,043	48.7
Ty1/Copia	10,006	0.9	11,680	1.0	21,686	1.9	121,206	3.7	87,857	2.7	209,063	6.3
Gypsy/DIRS1	185,369	16.0	446,363	38.5	631,732	54.5	348,360	10.6	802,252	24.3	1,150,612	34.9
DNA transposons	0	0.0	2,645	0.2	2,645	0.2	619	0.0	2,839	0.1	3,458	0.1
Unclassified	13,275	1.2	177,230	15.3	190,505	16.4	9,659	0.3	149,046	4.5	158,705	4.8
Total interspersed repeats	217,825	18.8	740,251	63.9	958,076	82.7	534,734	16.2	1,280,093	38.8	1,814,827	55.0
Satellites	0	0.0	928	0.1	928	0.1	0	0.0	.	0.0	0	0.0
Simple repeats	6,285	0.5	274	0.0	6,559	0.6	26,236	0.8	363	0.0	26,599	0.8
Low complexity	22,376	1.9	4,142	0.4	26,518	2.3	60,033	1.8	21,810	0.6	81,843	2.5
Bases masked	246,486	21.3	745,595	64.3	992,081	85.6	621,113	18.8	1,302,266	39.5	1,923,379	58.3
Total length used for analyses	1,158,939						3,297,440					

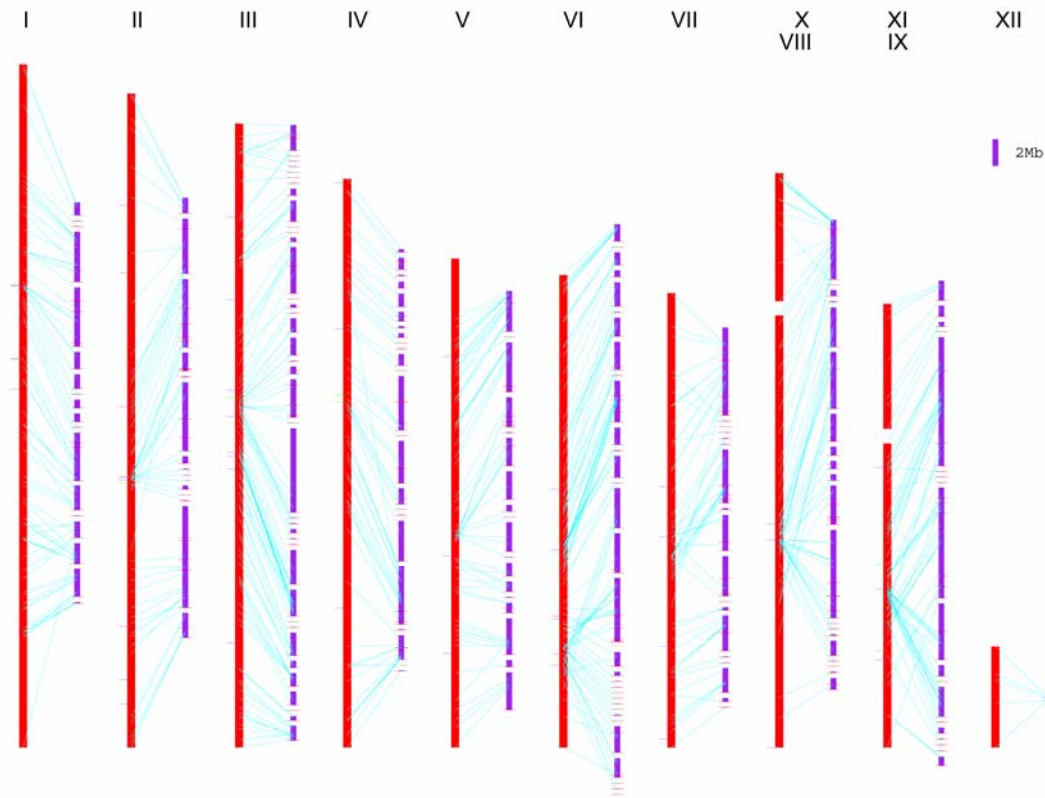
Supplementary Table 11. Summary of TE content of papaya contigs

Class	Element	Length occupied	% of total contig length
I (Retrotransposons)	<i>Ty3-gypsy</i>	77.3 Mbp (76.8 Mbp)	27.8 (27.6)
	<i>Ty1-copia</i>	15.3 Mbp (14.1 Mbp)	5.5 (5.1)
	LINE	3.0 Mbp (2.7 Mbp)	1.1 (0.96)
	SINE	1.1 Kbp	< 0.01
	Other	23.6 Mbp (22.0 Mbp)	8.4 (7.9)
II (Transposons)	CACTA, En/Spm	40.2 Kbp	0.01
	Micron	9.7 Kbp	< 0.01
	MuDR-IS905	7.6 Kbp	< 0.01
	Other	515.4 Kbp (469.1 Kbp)	0.18 (0.17)
Unclassified	Unknown	598.2 Kbp (70.7 Kbp)	0.22 (0.03)
Unannotated	Unknown	23.7 Mbp (23.7 Mbp)	8.5 (8.5)
Total	Misc.	144.1 Mbp (140.8 Mbp)	51.9 (50.4)

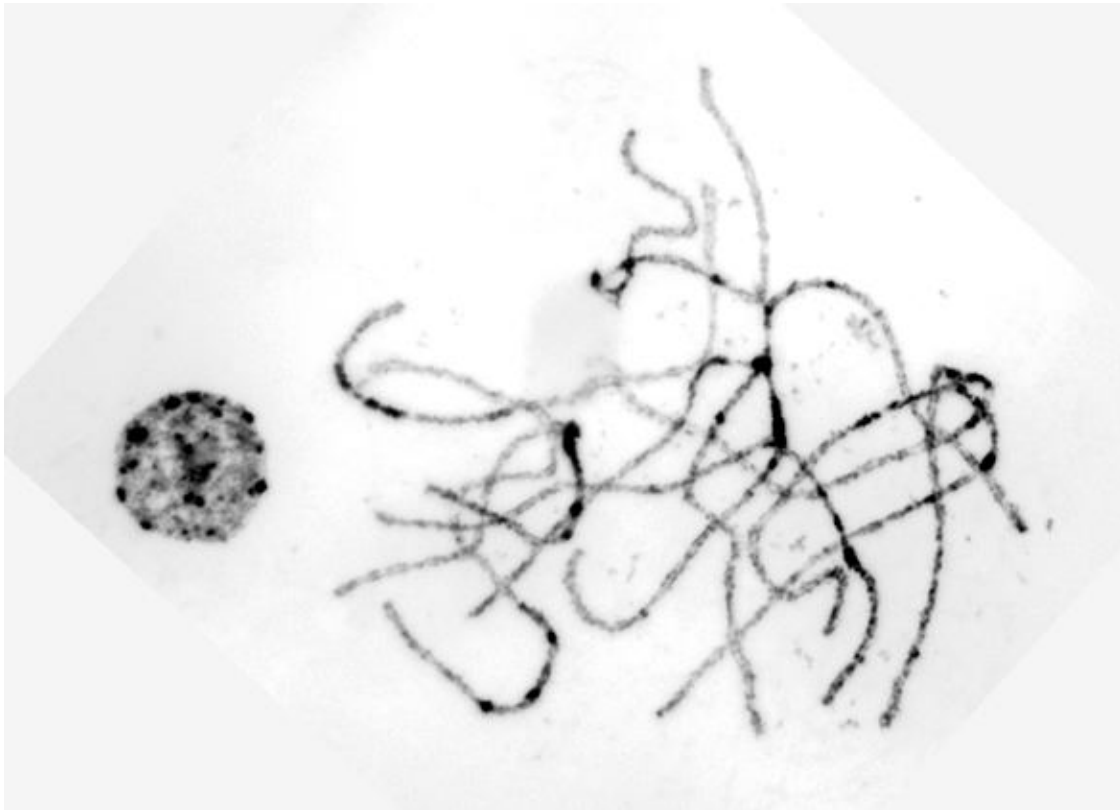
Papaya contigs were compared to known repeat elements in Repbase, the TIGR plant repeat database and a collection of papaya specific TE sequences, using RepeatMasker (with a conservative cutoff score of 350). The numbers in parentheses are those for the papaya specific repeat sequences.

Supplementary Table 12. MADS-box genes in papaya, *Arabidopsis*, poplar and rice

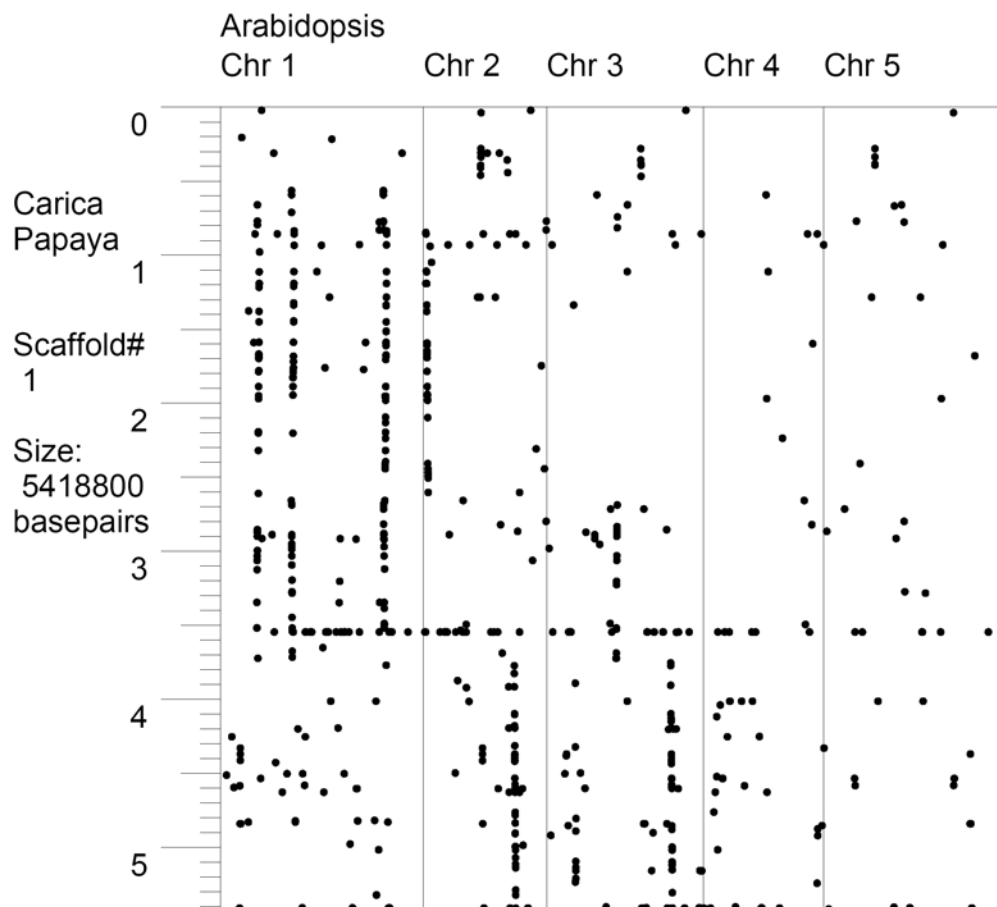
	Papaya	Arabidopsis			Poplar			Rice		
		Functional	Pseudo	Total	Functional	Pseudo	Total	Functional	Pseudo	Total
Type I	145			94			50			29
Mα	94	20	23	43	23	4	27	15	2	17
Mβ	4	17	5	22	12	5	17		1	1
Mγ	47	21	8	29	6	0	6	8	3	11
Type II	25			54			67			49
Mδ+MIKC										
*	5	6	1	7	9	1	10	1		1
MIKCa	20	43	4	47	55	2	57	47	1	48
Other	1									
Total	171			148			117			78



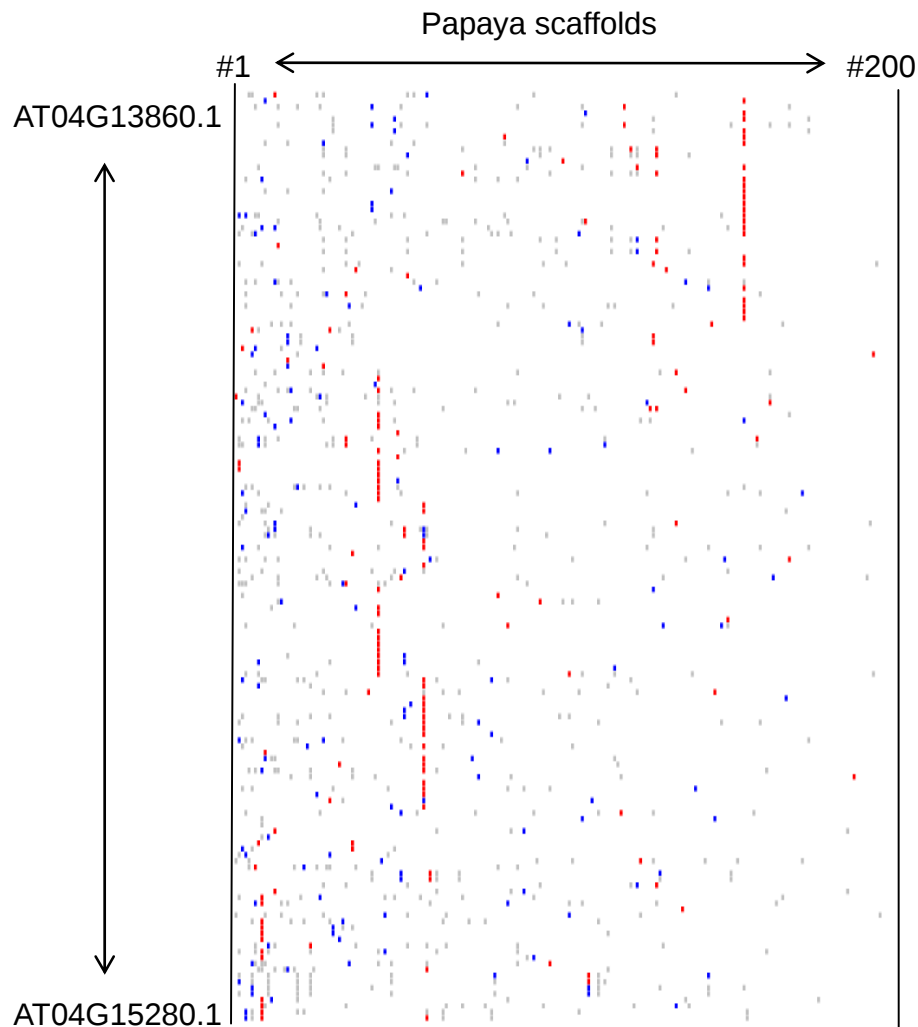
Supplementary Figure 1. Graphic representation of the assembled papaya genome (purple bars, with each scaffold marked by a red dash) anchored to the 12 papaya linkage groups (red bars) through SSR markers (blue lines). Linkage groups 8 and 10 were merged because each possessed a subset of SSR markers associated with the same scaffold. Linkage groups 9 and 11 were merged based on Fluorescence *In Situ* Hybridization (FISH) results. Un-utilized SSR markers are represented as purple dashes on the linkage groups. Conflicting SSR markers are represented as green dashes on the linkage groups.



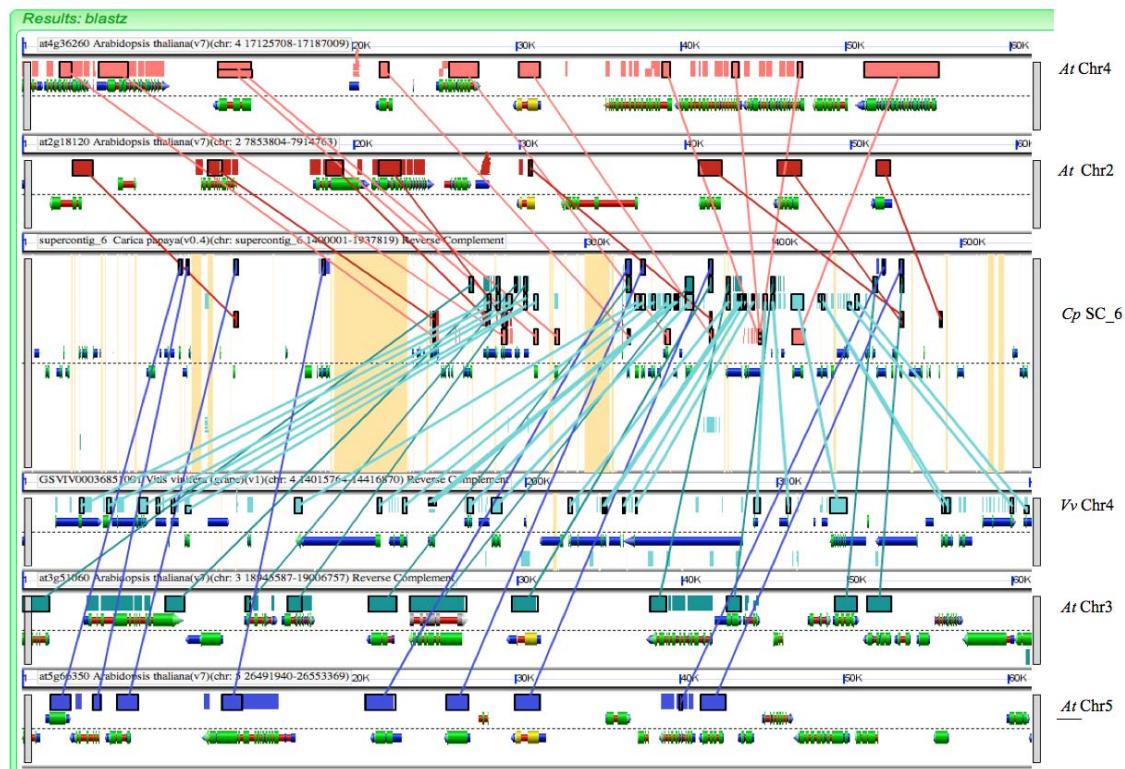
Supplementary Figure 2. Distribution of euchromatin and heterochromatin on meiotic pachytene chromosomes of papaya. Chromosomes were stained by 4', 6-diamidino-2-phenylindole (DAPI). The color image was converted into a black & white image to enhance the visualization of the darkly stained heterochromatin.



Supplementary Figure 3. Typical DNA synteny patterns between papaya scaffolds and *Arabidopsis thaliana* genome sequences. One papaya scaffold corresponds to four *Arabidopsis* chromosomal segments. The papaya scaffold is arranged vertically, with scale marks showing its length, while the 5 *Arabidopsis* chromosomes are arranged horizontally.

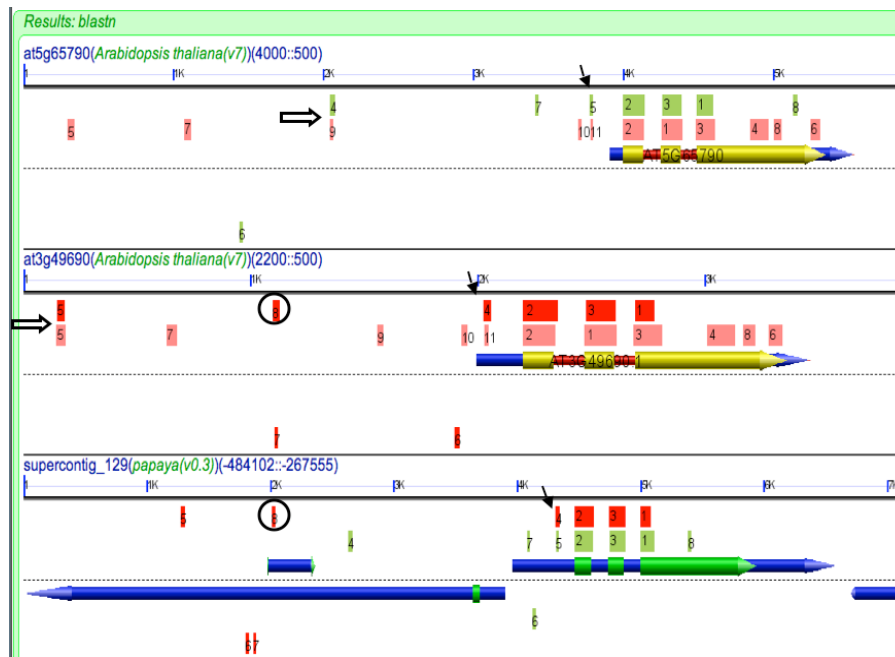


Supplementary Figure 4. Typical DNA synteny patterns between papaya scaffolds and *Arabidopsis thaliana* genome sequences. One *Arabidopsis* chromosomal segment corresponds to one best-matched papaya DNA segment. The *Arabidopsis* genes are arranged vertically, while papaya scaffolds (the longest 200 are shown here) are arranged horizontally. If an *Arabidopsis* gene has the best matched string on a papaya scaffold, a red dot will be created on the plot, the second best match produces a blue dot, others produce grey dots.



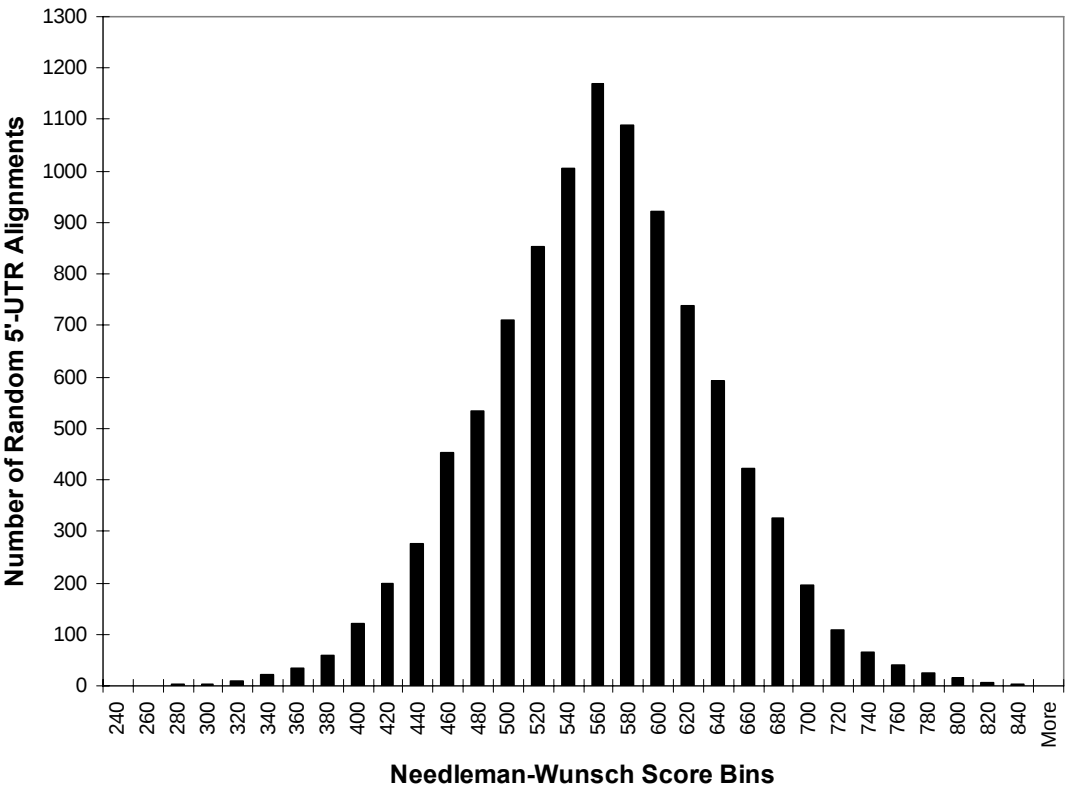
Supplementary Figure 5. BLASTZ¹ alignment of each of four syntenic chromosomal regions in *Arabidopsis* (v.7, Abbrev. “At”) to the single orthologous region of papaya (Cp) scaffold_129 (containing scaffold_6), and the orthologous region of chromosome 4 of grape². Graphic is from the GEvo alignment viewer³. Gene models have points at one end indicating transcriptional direction with putative exons in green (or yellow for the anchor gene). The gene in the center of each sequence, the anchor gene, encodes a SSR transcription factor that was retained following both of the two most recent tetraploidies, α and β . The papaya reference - the sequence third from the top and abbreviated “Cp” - is particularly informative; all other sequences are aligned with Cp. Above this sequence are stacks of high scoring pairs in one of five different colors: coral, red, light blue, green and dark blue, denoting At chromosome 4-Cp, At chromosome 2-Cp, grape chromosome 4-Cp, At chromosome 3-Cp, and At chromosome 5-Cp, respectively. These are two *Arabidopsis* a-pairs related to one another by the β tetraploidy. The result is that the approximately 27 genes in papaya for which there is syntenic support exist in *Arabidopsis* in 1, 2, 3 or 4 copies. Lines connect the HSPs that support this conclusion. Unsequenced stretches of chromosome are denoted in orange. To regenerate this GEvo graphic for research purposes use URL <http://tinyurl.com/3dpeak>, click “don’t draw HSP numbers” to reduce clutter and then “Go.”

1. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* 13, 103-107 (2003).
2. Jaillon, C. O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, doi:10.1038/nature06148 (2007).
3. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequence. *Plant J. In press* (2007).

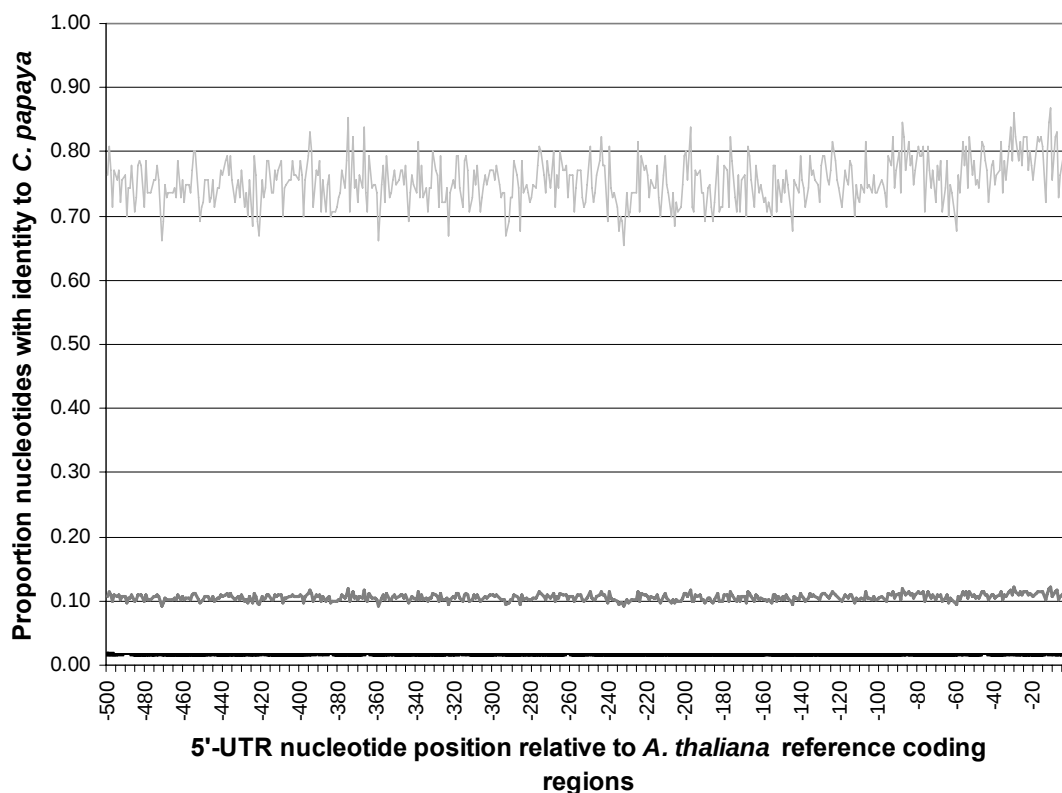


Supplementary Figure 6. Noncoding sequences are either shared or subfunctionalized when the papaya (Cp) MYB transcription factor gene is aligned with its alpha pair from *Arabidopsis* (At). Blastn¹ results were graphically represented using GEvo², with a noise cutoff set at the e-val equal to that of a 15/15 exact match. All three homologous, syntenic sequences were aligned in the three pairwise combinations using settings designed to find CNSs in plant genomes. Coral HSPs= At1 vs. At2, which generates α CNSs in noncoding sequences; light green HSPs= At1 vs. Cp; red HSPs= At2 vs. Cp. The arrows denote the single CNS that is clearly shared among all three sequences, and the hollow arrows denote sharing as well but one of the pairwise alignments did not produce a result because these CNSs are just at the limit of detection. The circle encloses a single HSP that is a CNS that was potentially subfunctionalized from at5g65790 after alpha. Cp is too diverged from At to permit efficient detection of many CNSs by sequence alone, but some genes have particularly conserved noncoding sequences like those in this example.

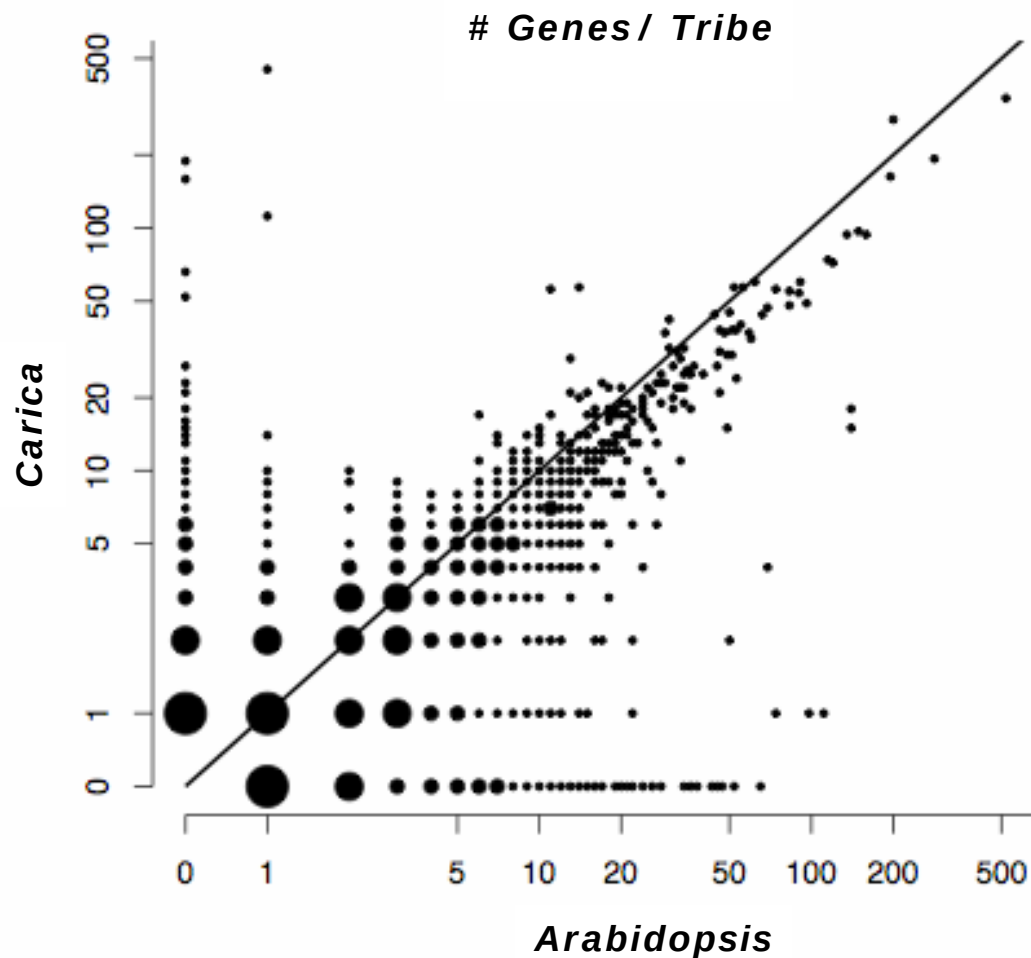
1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
2. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequence. *Plant J. In press* (2007).



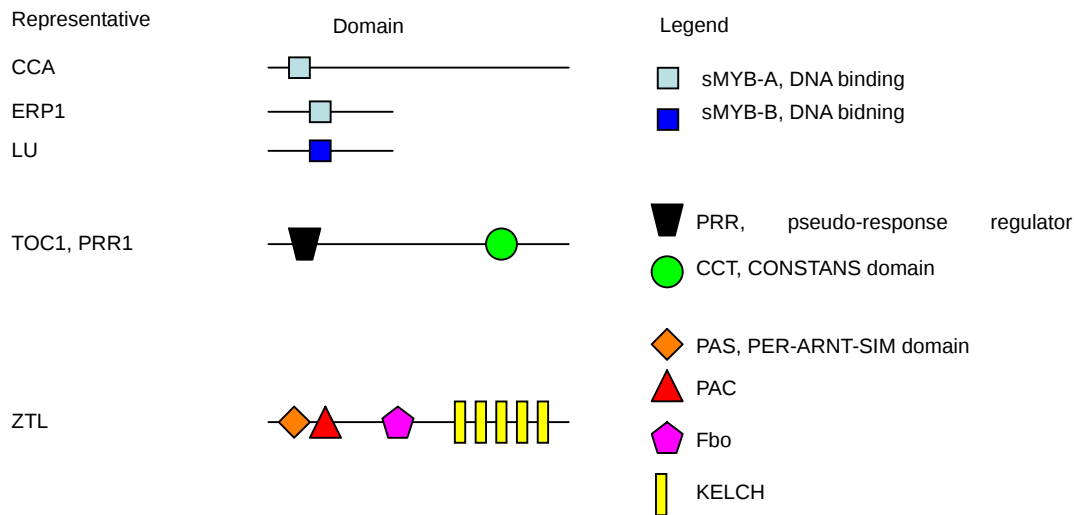
Supplemental Figure 7. Frequency distribution of Needleman-Wunsch alignment scores for 10,000 random alignments of *C. papaya* and *A. thaliana* 5'-UTR regions. The mean-score (mean) and standard deviation (stdev) are indicated. This distribution was used to determine the threshold alignment score for the analysis performed by *UntransID*.



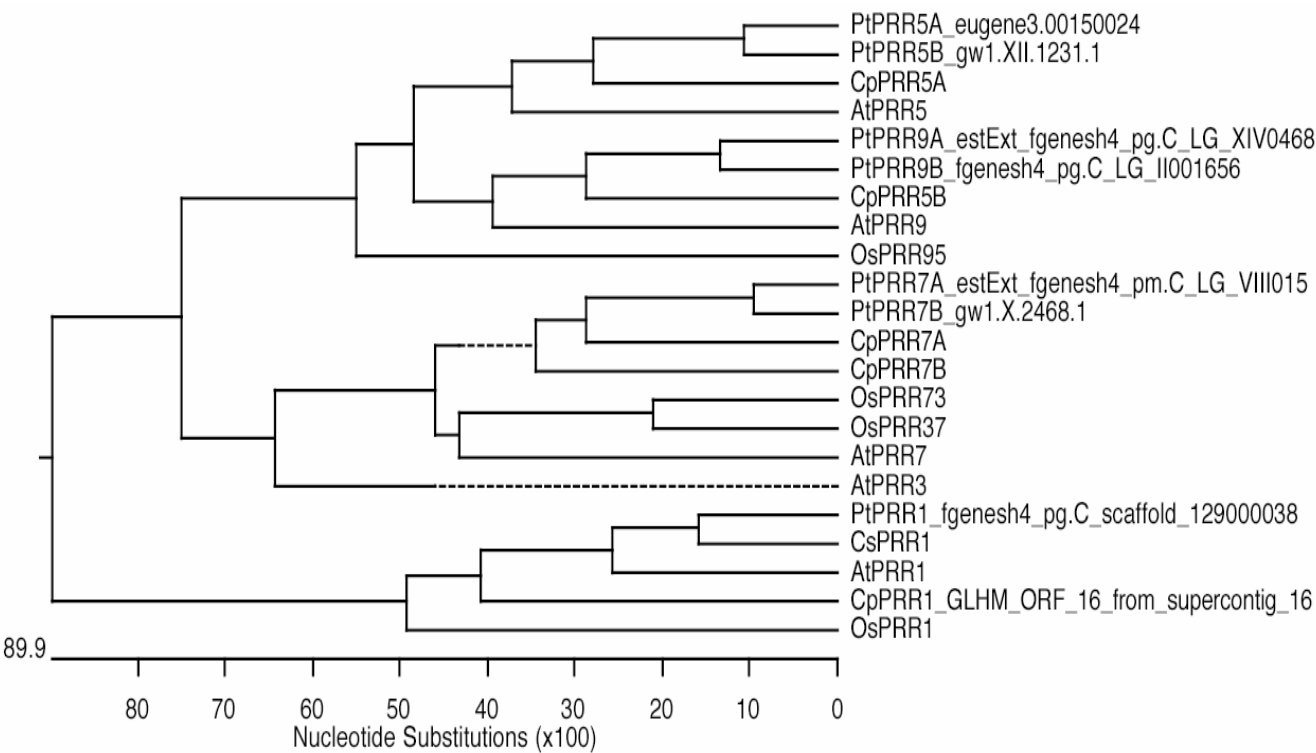
Supplemental Figure 8. Summary of the per-nucleotide identity shared between homologous 5'-UTRs in 969 *C. papaya*-*A. thaliana* comparisons. Values along the x-axis correspond to *A. thaliana* nucleotide positions 5' to coding-regions. For each position, the proportion of *A. thaliana* nucleotides scored as having identity (nucleotides from non-informative alignments are treated as "0" identity) to a *C. papaya* nucleotide is indicated by the thick dark-gray line. The thick, black line at approximately 0.01 is the mean proportion of nucleotides scored as identities with error bars (SE of each mean) for each nucleotide position as determined via a permutation test (100 iterations; see Windsor *et al.*¹, supplemental online documentation for details). The light-gray line indicates the proportion identity at each position when only alignments with scores exceeding the significance threshold (705) are included in the calculation (136 alignments).



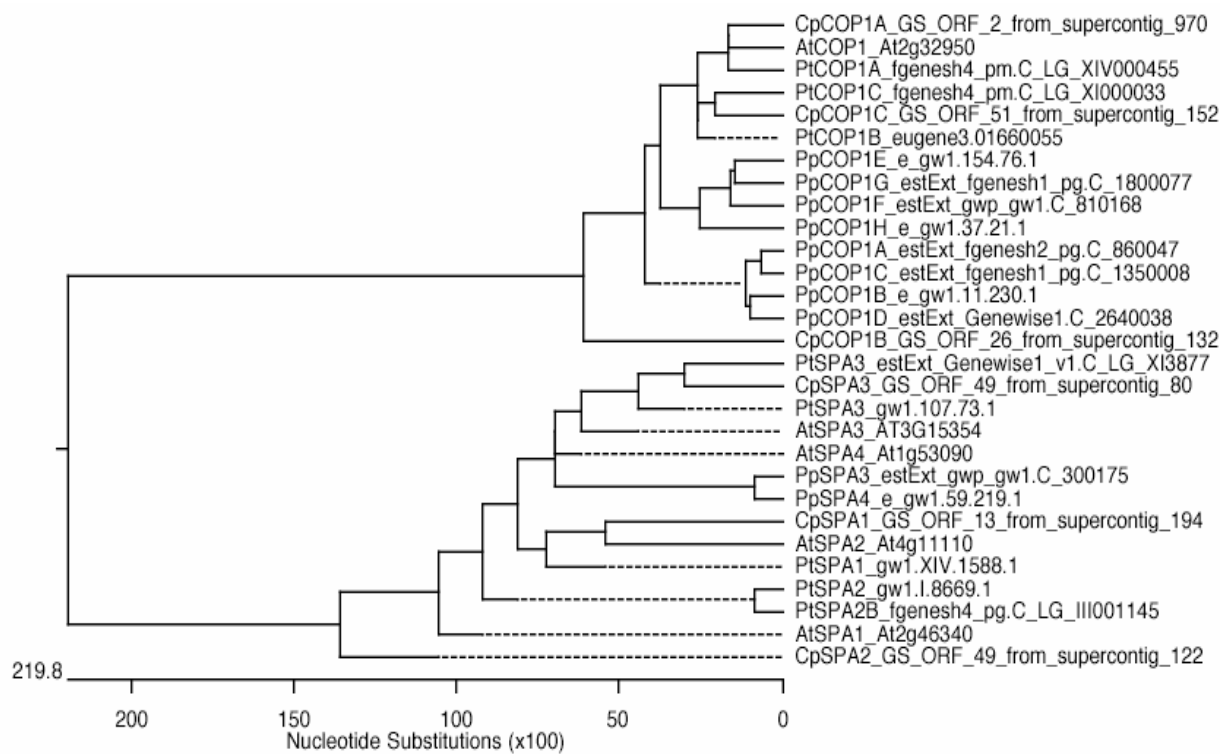
Supplemental Figure 9. Tribe size in *Carica papaya* versus *Arabidopsis thaliana* for tribes present in at least one species. Larger circles indicate larger number of tribes with those counts. The larger the circle, the more tribes (small = 1-10, medium = 11-100, large = 101-1000, super = >1000). Diagonal line shows equal numbers *Carica* and *Arabidopsis*.



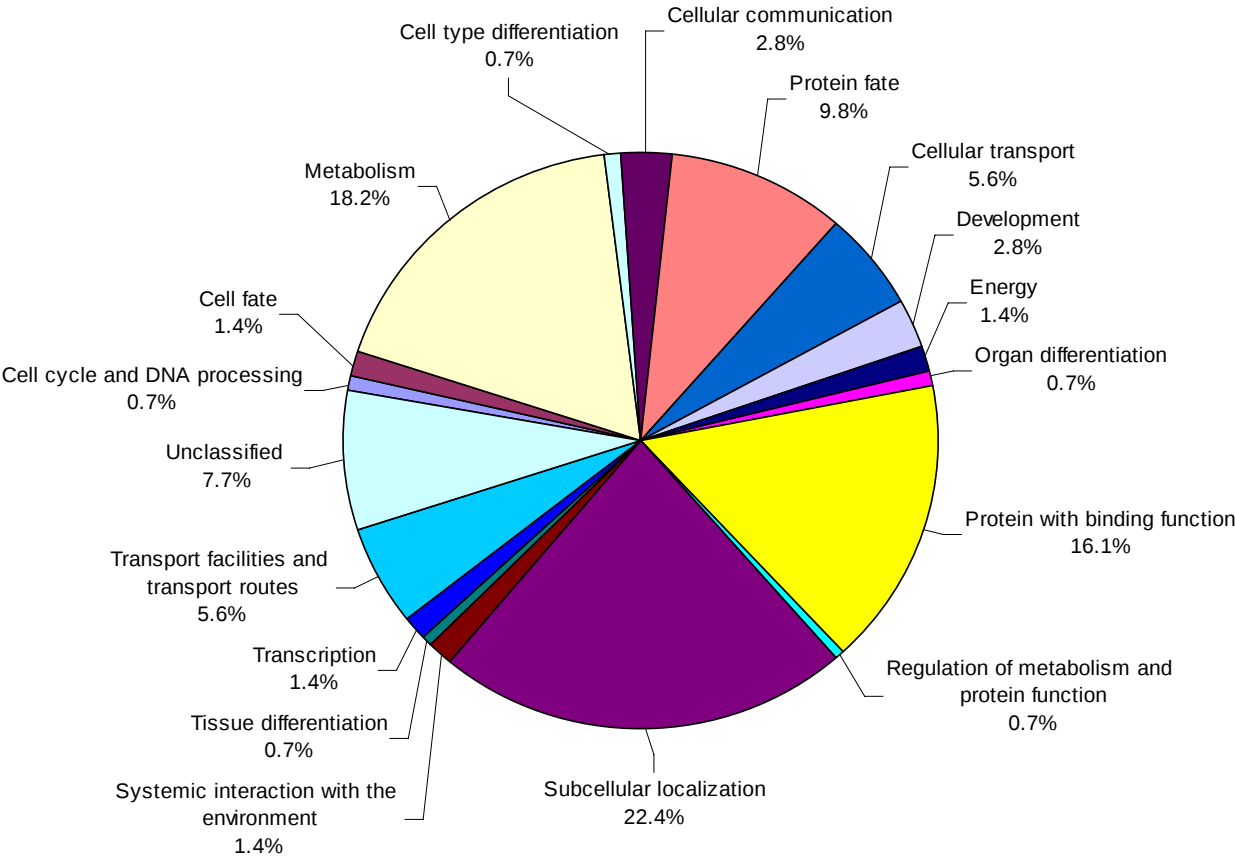
Supplementary Figure 10. Circadian clock gene models and domains. Circadian clock gene models and domains derived from *Arabidopsis thaliana*. Single MYB domain transcription factors (sMYB) CIRCADIAN CLOCK ASSOCIATED 1 (CCA1), EARLY-PHYTOCHROME-RESPONSIVE 1 (EPR1), and LUX ARRHYTHMO (LUX), contain two MYB domains of the VRSHQY subclass designated here as A (light blue) and B (dark blue). Pseudo-response regulator family represented by TIMING OF CAB 1 (TOC1)/ PSEUDORESPONSE-REGULATOR 1 (PRR1), have a pseudo-response regulator domain (black) and a CONSTANS domain (green). ZEITLUPE (ZTL) is the founding member of the PAS-PAC/FBOX/KELCH (PFK) family. The PER-ARNT-SIM (PAS) domain (orange), PAC domain (red), Fbox domain (purple) and the KELCH domain (yellow).



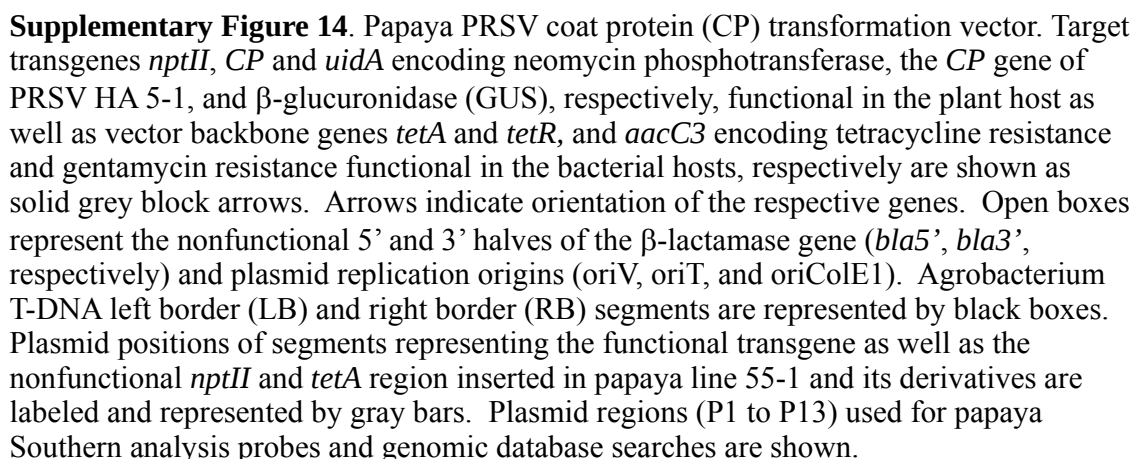
Supplementary Figure 11. Phylogenetic relationship of the PRR proteins from papaya (Cp), poplar (Pt), *Arabidopsis* (At), chestnut (Cs) and rice (Os). Protein sequences were aligned with ClustalW and the phylogenetic tree was generated in MegAlign (DNASTar, lasergene6). The length of each pair of branches represents the distance between proteins sequences. The X-axis indicates the number of substitution events between protein sequences. The dotted line indicates a negative branch length as a result of averaging to achieve a balanced branched phenogram.

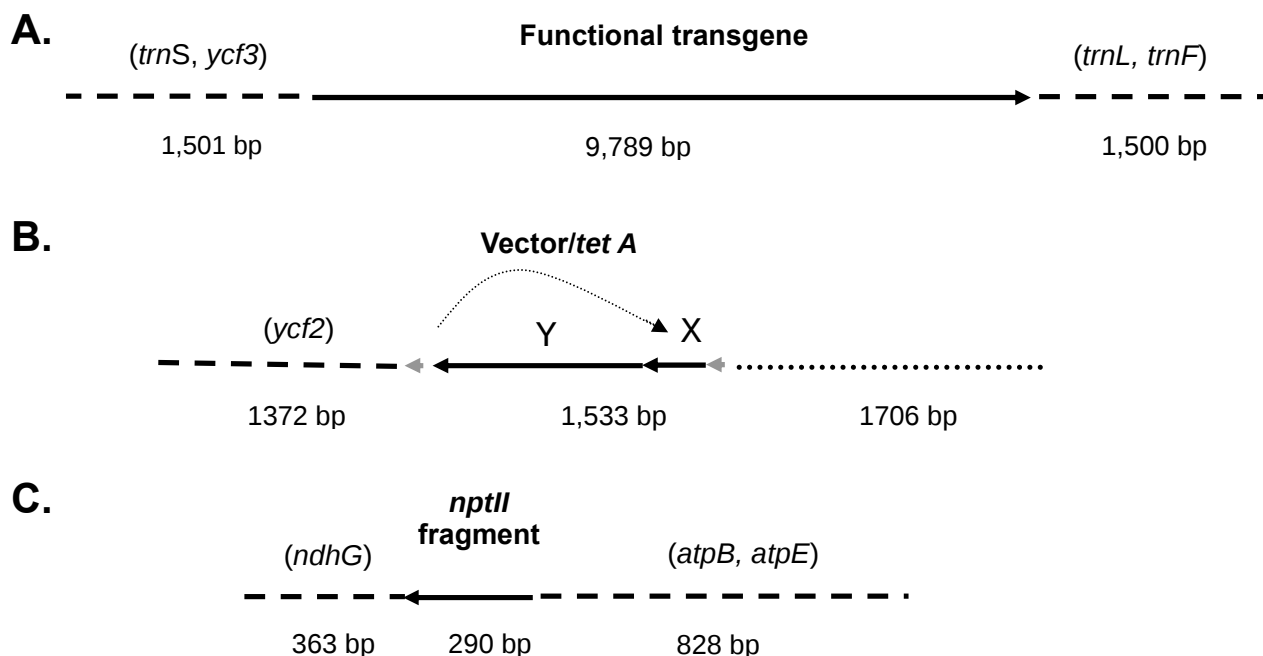


Supplementary Figure 12. Phylogenetic relationship of the COP/SPA in papaya (Cp), poplar (Pt), moss (Pp) and *Arabidopsis* (At). Protein sequences were aligned with ClustalW and the phylogenetic tree was generated in MegAlign (DNASTar, lasergene6). The length of each pair of branches represents the distance between proteins sequences. The X-axis indicates the number of substitution events between protein sequences. The dotted line indicates a negative branch length as a result of averaging to achieve a balanced branched phenogram.

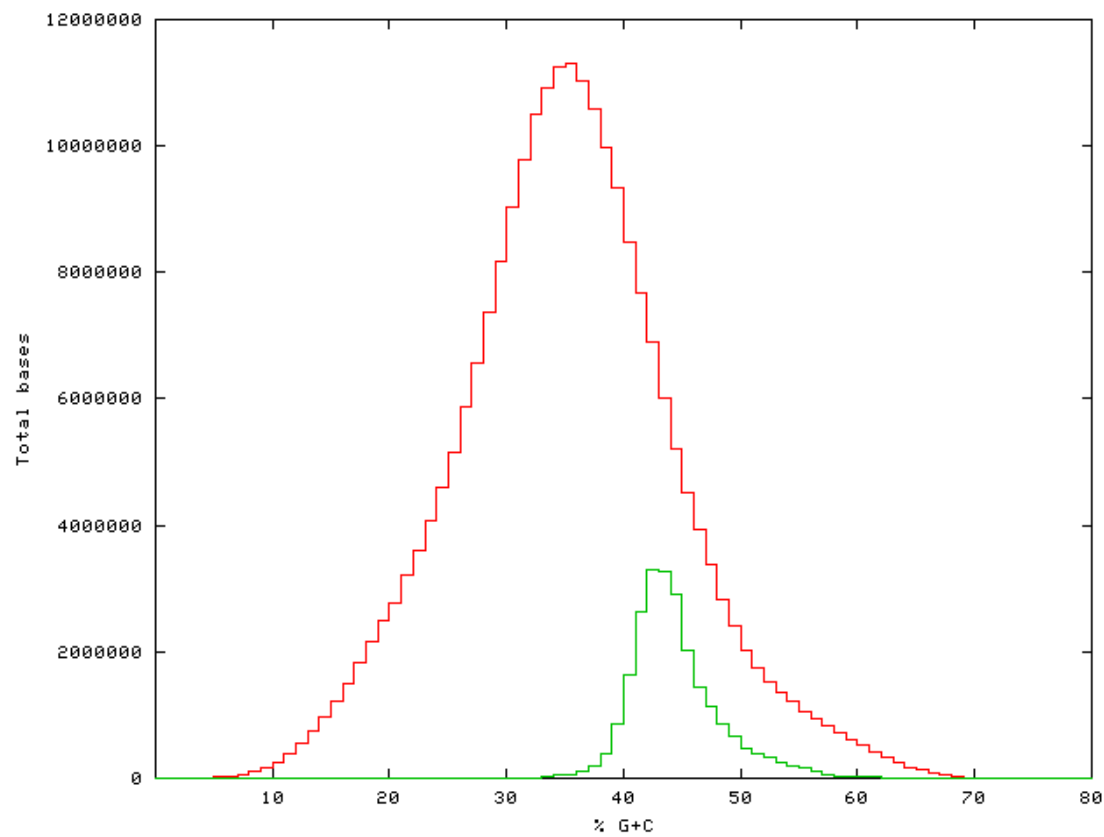


Supplementary Figure 13. The functional categories of the genes on X chromosome at papaya sex determination region. The annotated function was classified based on the hit in MIPS functional catalogue database (<http://mips.gsf.de>).

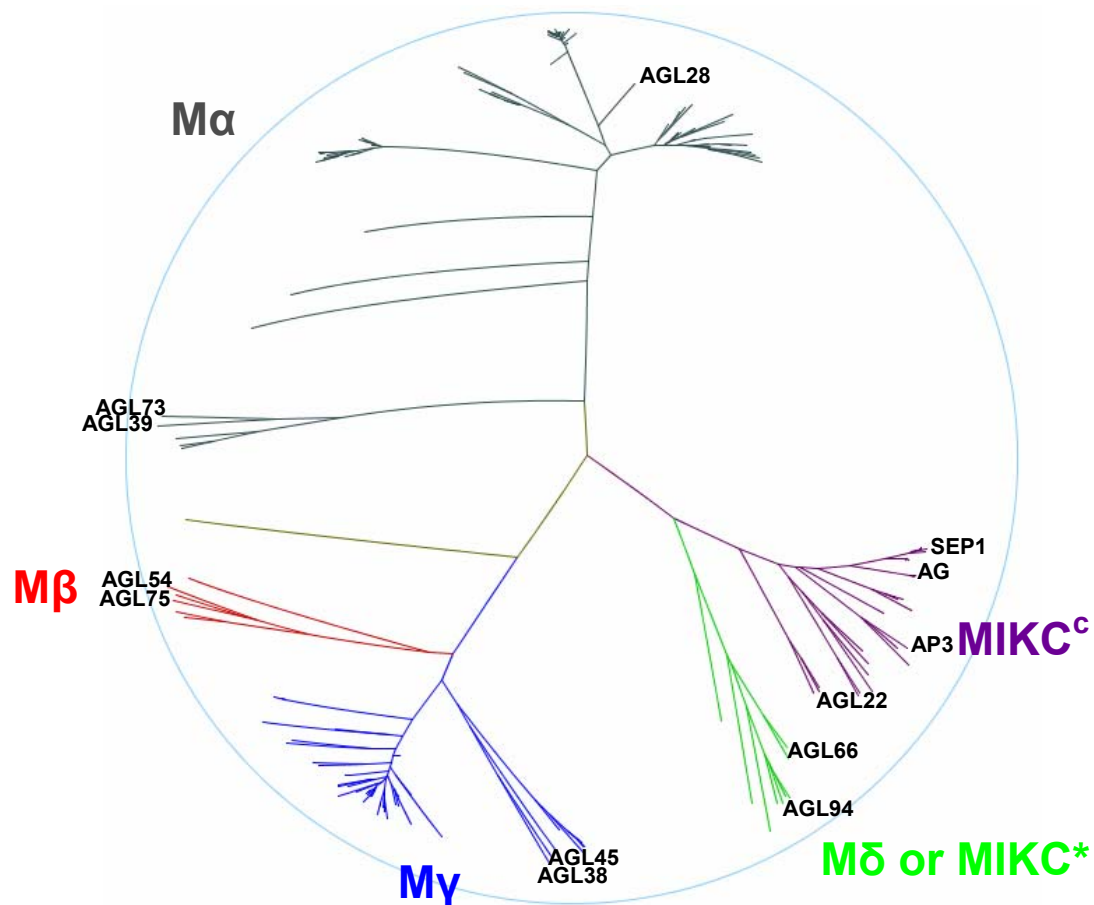




Supplementary Figure 15. Structure of transformation insertion sites in the papaya 55-1 line genome. **a.** Functional transgene insertion. **b.** Nonfunctional vector/*tetA* fragment insertion. **c.** Nonfunctional *nptII* fragment insertion. Solid arrows represent insertions and orientation with respect to the plasmid sequence (functional transgene) or gene orientation (vector/*tetA*, *nptII* fragment insertion). Gray arrowheads represent a 21 bp duplicated vector sequence and dotted arrow in the vector/*tetA* insertion indicates the transposition of a segment of the insert (X) relative to a second segment (Y) compared to their relative positions on the transformation plasmid. Dashed lines represent chloroplast DNA-like nuclear genome sequences flanking the transgene insert. Dotted line represents flanking sequences that are not chloroplast DNA-like. Lines representing insertions and flanking DNA are not drawn to scale. Gene sequences found in the flanking DNA are identified in parentheses.



Supplementary Figure 16. GC content of papaya genome. The red plot is a histogram of the total number of bases in all scaffolds, computed in 200-bp windows across the genome. The green plot shows the corresponding totals for coding exons only.



Supplementary Figure 17. Phylogenetic analysis of predicted papaya MADS-box proteins. Selected *Arabidopsis* MADS-box proteins representing the 5 subfamilies are also included (see text). The positions of the *Arabidopsis* proteins are shown (Mα, AGL28, AGL39, and AGL73; Mβ, AGL54 and AGL75; Mγ, AGL38 and AGL45; Mδ (or MIKC*), AGL66 and AGL94; MIKC^c, SEP1, AG, AP3 and AGL22).