www.sciencemag.org/content/346/6209/646/suppl/DC1

Supplementary Materials for

A Y-chromosome–encoded small RNA acts as a sex determinant in persimmons

Takashi Akagi, Isabelle M. Henry, Ryutaro Tao,* Luca Comai*

*Corresponding author. E-mail: rtao@kais.kyoto-u.ac.jp (R.T.); lcomai@ucdavis.edu (L.C.)

Published 31 October 2014, *Science* **346**, 646 (2014)
DOI: 10.1126/science.1257225

This PDF file includes:

Supplementary Materials:

Materials and Methods:

Materials 1. Plant materials

Trees from the KK population (*21*) (*D. lotus*) and their parents, planted in Kyoto University (Kyoto, Japan, 35º03', 135º78' L/L) (Table S5), were used for the sequencing analyses. A total of 46 trees (22 female and 24 male) were used for the initial male k-mer identification (Fig. 1A) and genomic contig construction (Fig. 1B) while data from one male and 10 female trees were added for the recombination analysis (Fig. 1C). For expression analyses, mixed buds were sampled on June 17th, and July 4th, 2013, corresponding to the early differentiation stages of male/female primordia and male and female developing flowers were sampled on 29th April, 2014,. Nine individuals of each gender and the parents were sampled. For information about the other *Diospyros* species, see Table S5.

Method 1. Illumina library construction and sequencing

1. Genomic libraries

Genomic DNA was extracted from young leaves using the CTAB method and purified by phenol/chloroform extraction. Approximately 1.0 µg of genomic DNA was fragmented using NEBNext dsDNA Fragmentase (New England BioLabs; NEB) for 40-60 min at 37°C and cleaned using Agencourt AMPure XP (Beckman Coulter Genomics) for size-selection. To select fragments ranging between 200 and 600 bp, 25 µl AMPure were added to the initial 50 µl reaction. After a brief incubation, 72 µl of the supernatant was transferred to a new tube, and an additional 12 µl water and 36 µl AMPure were added. After a second brief incubation, the supernatant was discarded and the DNA was eluted from the beads in 20 µl of EB, as recommended. Next, DNA fragments were subjected to end repair using NEB's End Repair Module Enzyme Mix, and A-base overhangs were added with Klenow (NEB), as recommended by the manufacturer. End repair and A-base addition were both followed by AMPure cleanup using 1.8 : 1 (v/v) AMPure / reaction. Barcoded NEXTflex adaptors (Bioo Scientific) were ligated at room temperature using NEB Quick Ligase (NEB), following the manufacturer's recommendations. To remove contamination of self-ligated adapter dimers, libraries were size-selected using AMPure in 0.8 : 1 (v/v) AMPure : reaction volume, in order to select for adapter-ligated DNA fragments at least 300-bp long. Half of the eluted DNA was enriched by PCR reaction using Phusion 2X HF master mix (NEB), with the following PCR conditions: 30 s at 95°C; 8 cycles of 10 s at 95°C, 30 s at 65°C, and 30 s at 72°C and a final extension step of 1 min at 72°C. Enriched libraries were purified with AMPure (0.8 : 1 v/v AMPure to reaction), and quality and quantity were assessed using the Agilent BioAnalyzer (Agilent Technologies) and Qubit fluorometer (Invitrogen). Libraries were sequenced using Illumina's HiSeq 2000 (100-bp paired-end reads), according to the manufacturer's instructions.

2. mRNA libraries

Mixed buds samples corresponding to the early differentiation stages of male or female primordia were harvested on July 4th, 2013. Total RNA was extracted using the CTAB method and purified by phenol/chloroform extraction. Fifteen to twenty micrograms of total RNA was processed

in preparation for Illumina Sequencing, according to the previous report (*26*). Briefly, mRNA was purified using the Dynabeads mRNA purification kit (Life Technologies). Next, cDNA was synthesized via random priming using Superscript III (Life Technologies) followed by heat inactivation for 5 min at 65°C. Second-strand cDNA was synthesized using the second-strand buffer (200 mM Tris-HCl, pH 7.0, 22 mM $MgCl_2$, and 425 mM KCl), DNA polymerase I (NEB) and RNaseH (NEB) with incubation at 16°C for 2.5 h. Double-stranded cDNA was purified using AMPure with a 1.8 : 1 (v/v) AMPure to reaction volume ratio. The resulting double-stranded cDNAs were subjected to fragmentation and following library construction, as described above for the genomic library preparation. Ten cycles of PCR enrichment were performed using the same temperature and time conditions as described above. The constructed libraries were sequenced on Illumina's HiSeq 2500 sequencer (150-bp paired-end reads).

3. small RNA library

Total RNA was extracted from pooled buds sampled on July 17th, 2013, using the CTAB method and purified by phenol/chloroform extraction. The small RNA fraction was concentrated from total RNA using the mirVana miRNA Isolation kit (Life Technologies). Approximately 150 ng of concentrated small RNA was processed to library construction using the NEBNext Small RNA Library Prep Set (NEB), according to the manufacture's instruction. PCR enrichment reactions used 12 cycles of amplifications, followed by DNA cleanup using AMPure (AMPure : reaction = 1.1 : 1 v/v) to remove self-ligated adapter dimers. Library quality and quantity were assessed using the Agilent BioAnalyzer (Agilent Technologies) and Qubit fluorometer (Invitrogen). The constructed libraries were sequenced using Illumina's HiSeq 2000 sequencer (50-bp single-end reads).

4. Sequence processing

Illumina sequencing reads were processed with custom Python scripts developed in the Comai laboratory and available online (http://comailab.genomecenter.ucdavis.edu/index.php/Barcoded_data_preparation_tools). In short, sequences were split according to barcode information (Table S6), trimmed for quality (average Phred sequence quality > 20 over a 5 bp sliding window) and adaptor sequence contamination after which reads shorter than 35-bp were discarded (except for the small RNA reads, for which read length cut-off was reduced to 19 bps). For the smRNA data, a maximum read length of 25 bp was applied as well. After application of those size thresholds a total of approximately 12.0 M and 9.6 M reads were obtained from the female and male samples, respectively.

Method 2. Sex-specific k-mer extraction

To select sex-biased reads, the quality trimmed read files from both male and female samples were processed to identify gender-specific subsequences using custom python scripts. For the genomic reads, all 35 bp kmers (words) starting with the "AG" dinucleotide were selected from all reads, while keeping track of the number of times each specific subsequence was collected. The use of a dinucleotide trigger sequence allowed us to restrict file size while retaining the ability to compare k-mers between reads by effectively phasing them. For the RNA-Seq data, no trigger sequence was used, resulting in the selection of all possible k-mers. Next, the set of subsequences that met a minimum total (male + female) count threshold of 10 for genomic kmers and 20 for RNA-Seq kmers, and a maximum total count threshold of 200 for genomic kmers and 2000 for RNA-Seq kmers were retained. The k-mer counts were then compared between male and female reads (Fig. 1B for genomic data and Fig. S4 for mRNA data). Finally, fully male specific k-mers (count of 0 in the female set) were identified and used to extract the sex-biased reads from the original quality trimmed read set as follows: all pair-ended reads containing at least one of the selected fully male-specific kmers were retained.

Method 3. Alignment and construction of the MSY-linked genomic contigs

De novo assembly for the genomic contigs was performed by the CLC assembler using all 100-bp paired-end (PE) reads that included male-specific k-mers (MSK) (see Fig. 1 and Method 2 for the MSK). Alignment of the Illumina reads to the 5,100 contigs, which were generated by seeded assembly including the male specific k-mers (MSK), were conducted using Burrows-Wheeler Aligner (BWA) version 0.7.7 using default parameters (*27*) (http://bio-bwa.sourceforge.net/). The number of reads mapping to each reference sequences was recorded from the alignment file produced by Sequence Alignment/Map (SAM) tool (*28*) (http://samtools.sourceforge.net/) using custom R scripts. For each contig, informative SNPs and short indels were identified using the mapped reads and custom Python script (http://comailab.genomecenter.ucdavis.edu/index.php/Mpileup) following BWA/SAM alignment. Only SNPs and/or indels observed in multiple individuals (N > 5) and gender-specific were defined as informative polymorphisms.

Recombination mapping was performed in two steps (see Fig. S2): first reads from all male and female samples were aligned to the 5,100 seed contigs with high stringency (no mismatches allowed), to classify the contigs into three categories: male-specific contigs, pending contigs and repetitive contigs. Male-specific contigs were defined by a complete lack of mapped reads from the female individuals. Contigs were labeled "repetitive" when the coverage of mapped reads from both male and female pools was 20-fold higher than expected coverage for Y (= ca 23). The other contigs were defined as "pending". Here, we identified approximately 300, 4600, and 200 "male-specific", "pending", and "repetitive" contigs, respectively. The pending contigs were used as reference for a new round of read mapping using lower stringency (number of mismatches allowed between 2 and 12) in order to further examine their linkage to the MSY (approximately 0-4 cM).

To identify contigs that are Y-linked but not fully gender-specific in our sample set, we assessed the percentage of recombination around the MSY region. Information from all informative polymorphism present in each contig were pooled to derive the X/Y genotype of each individual (XY or XX). The haplotype of each contig was manually confirmed by visualization of polymorphisms using the Integrative Genomics Viewer (IGV) version 2.2 (*29*). For each contig and each individual, the number of reads containing X or Y-specific polymorphisms were recorded. Individuals with at least eight independent reads containing an X-specific polymorphism, but no read containing Y-specific polymorphisms, were labeled as homozygous XX for that particular contig. Individuals exhibiting at least one Y-specific polymorphism were labeled as heterozygotic XY for that contig. If six or more reads included an X-Y polymorphism, at least 2 Y-specific polymorphism were required to assign the XY genotype to that individual. Otherwise, it was labeled as "unknown". Contigs with ambiguous polymorphic reads or insufficient data were labeled as "unknown". Next, genotype information was required from at least 40 individuals for a contig to be classified as Y-linked or Y-specific. Following these thresholds and based on data from 57 individuals, 1076 contigs were labeled as Y-specific, where all individuals genotypes matched their gender, *ca* 300 contigs were found to be gender-linked and *ca* 150 contigs remained unknown.

The Y allelic contigs, which remained after testing for recombination, were further expanded and integrated by Paired-Read Iterative Contig Extension (PRICE) (*30*), and CAP3 (*31*). Y allelic PE reads mapping at least partially onto approximately 300 male-specific contigs, to which no female reads mapped, were isolated from the SAM file using a custom Python script, and subjected to PRICE assembly using the following parameters: -fs "PE reads file" 500 (insertion size),, –icf "seed contig file" 20 (nos of addition steps) 3 (nos cycles per step) 1 (constant by which to multiply quality scores) –nc 60-100 (nos of cycles). The –icf parameter was set such that 1/20 of the seeds would be added at the first cycle, and the next 1/20 is added 3 cycles later (at cycle 4). The resulting contigs were integrated with the Y-specific, the *SD*-linked polymorphic, and the unknown (*N* = *ca* 1,500) contigs using CAP3 and default parameters. These expansion/integration steps were repeated seven times, and consequently submitted to the recombination test described above. The resulting 796

contigs exhibiting no evidence of recombination from the SD locus (Fig. 1C).

To construct X-specific contigs homologous to the Y-specific contigs described above, genomic reads from female individuals of the KK population were mapped to these 796 contigs using BWA, allowing at maximum 12-bp mismatches per read. Assembling the mapped PE reads using Trinity (*32*) and CAP3 using default parameters generated candidate X allelic contigs. Of those, 777 contigs showing significant homology to the 796 Y allelic contigs (> 90% nucleotide homology at least in 100-bp word size) and were defined as putative X allelic contigs. Note that at most three different X alleles could be expected in the KK sibling population, one from the XY male parent and the other two from the XX female parent.

Method 4. Identification of expressed genes involved in sex determination

RNA-Seq reads from 9 male and 9 female individuals from the KK population and their parents (N = 2 x 10) were subjected to three independent analyses in order to identify candidate genes involved in sex determination in *D. lotus*. (i) RNA-Seq 150-bp paired-end reads were fragmented to 3 x 50-bp (x 2), and mapped to 796 genomic contigs on the MSY and the 777 corresponding X allelic sequences assembled from the female genomic reads, which were obtained as described above (Method 3). The original 150-bp PE RNA-Seq reads containing one or more 50-bp kmers mapping to these genomic contigs were used to assemble cDNA contigs (described below, Method 5). (ii) In an approach analogous to the one applied to the genomic contigs (Fig. 1, Method 2-3), male-specific k-mers (MSKs) were isolated directly from the RNA-Seq read sequences (k = 35, ≥ 20X coverage). Next, all reads containing one or more of these MSKs were assembled into cDNA contigs using the CLC assembler. Using the genomic sequence reads obtained from 57 individuals (Fig. 1), recombination mapping was performed to define a subset of MSY-linked contigs. The cDNA contigs constructed using approaches (i) and (ii) were used for differential expression analysis by determining reads per kilobase per millions (RPKM) values (> 1.0), and annotated using the TAIR/uniprot databases. The approach followed to integrate all data concerning the putative SD loci is described below. For the third approach, (iii), all male and female full-length RNA-Seq reads were aligned using the CLC assembler to produce approximately 400,000 contigs, including allelic polymorphisms (see Method 6). Selection for contigs that exhibited RPKM values of at least 1.0 reduced that number to approximately 80,000 contigs. The number of RNA-Seq reads from each individual mapping to these contigs, were used for DESeq analysis described below.

Method 5. Construction of cDNA contigs located in the MSY

To map the RNA-Seq reads to the 796 and 777, respectively, Y- and X genomic contigs, each 150-bp read sequence was first converted to three 50-bp fragments using a custom Python script, to decrease the negative effect of intron sequences on mapping efficiency. Here, each fragmented read from the male and female individuals of the KK population was mapped, respectively, to the 796 putative Y and 777 putative X MSY allelic contigs using BWA and allowing no nucleotide mismatches. The original PE reads for which at least one of the fragmented read was mapped to a genomic contigs, were extracted using custom Python scripts. The mapped PE reads were assembled into contigs using Trinity and CAP3 and default parameters. Next, cDNA contigs exhibiting > 99% nucleotide homology to the 796 Y or 777 X allelic genomic contigs over at least 100-bp, were defined as cDNA contigs located in the sex-determining region. From the Y and X allelic genomic contigs, respectively, 99 and 81 cDNA contigs were retained. Alignment of the genomic and cDNA contigs to each other indicated that these cDNA contigs were derived from 60 of the 796 Y allelic genomic contigs and / or their corresponding X allelic contigs.

Method 6. Expression profiling

cDNA contigs corresponding to genes expressed in male or female developing bud primordia were assembled using the full-length cDNA reads from all male and female individuals ("RNA-Seq"

in the third set in Table S6; 340,457,970 reads for the male samples, and 264,929,840 reads for the female samples) using the CLC assembler and a minimum contig length of 200-bp. Next, the resulting 400,000 cDNA contigs, including some alternative splicing and isoforms were used as reference sequences for alignment of the reads using BWA with default parameters. The read counts per contig were generated from the aligned SAM files using a custom R script. Differential expression between male and female individuals was analyzed in R (version 3.0.1) using the R package DESeq (version 1.14; http://bioconductor.org/packages/release/bioc/html/DESeq.html) (*33*). We conducted DESeq analysis using 10 biological replicates from male and female individuals (see Table S6), with the following parameters: method="per-condition" and sharingMode="gene-est-only". A False discovery rate (FDR) threshold of 0.01 was used to identify differentially expressed genes.

Method 7. Identification of SD candidates based on genetic diversity and phylogenetic information

The X and Y allelic sequences of the SD candidate genes were isolated from the cDNA contigs located on the MSY (described in Method 4), and expanded/integrated using flanking PE reads using PRICE and CAP3. Next, the allelic states of each polymorphism was reassessed by mapping cDNA and genomic reads to the expanded contigs and assessing possible recombination events within the MSY-linked contigs, as described above (Method 3). Nine ORF fragments for which polymorphic data were consistent with incomplete linkage to the SD were excluded. Furthermore, potential transposable elements identified by blast using the TAIR/nr databases were removed, and alternative splicing and isoforms were integrated, based on putative functional orthologous sequences (i.e. all spliced cDNA sequences) in model plants (Arabidopsis, grape, and tomato) and on the sequences of RNA-Seq reads bridging two separated putative exons. Finally, 22 SD candidate genes remained (Table 1), including 18 genes exhibiting distinct X and Y alleles, 2 genes with X and Y alleles, but containing mostly repetitive sequences, 1 repetitive gene specific to the Y chromosome, and 1 completely Y-specific gene (*OGI*). The X and Y allelic nucleotide sequences from the first 18 candidate genes were aligned using MAFFT ver. 7 (*34*) and the L-INS-i model and by SeaView ver. 4 (*35*) for manual pruning. The resulting alignments were subjected to DnaSP 5.1 (*36*) to calculate the number of segregation sites (*S*), Jukes and Cantor corrected values of synonymous (Ks) and non-synonymous (Ka) substitutions, and the index of evolutionary speed (Ka/Ks ratio) (Table S3).

To investigate the timing of divergence between the X vs Y alleles of the candidate genes (Fig. S9), we compared their Ks value and all site divergence with those of phytochrome A (phyA) in *Diospyros lotus* compared to phyA in various more or less distant *Diospyros* species. The sequences of phyA in *D. mespiliformis*, *D. virginiana*, *D. kaki*, and *D. lotus* were obtained from KF291691, KF291789, KF291677, and KF291684, respectively. Those from *D. digyna*, *D. montana*, and *D. oleifera* were amplified using specific primers sets (*37*), and directly sequenced. These sequences were aligned using ClustalX2 and SeaView ver. 4 and subjected to DnaSP 5.1 to analyze informative SNPs.

To investigate the pattern of sequence divergence in the closely-related species, primers for all 22 SD candidates were designed, and used for amplification of orthologous sequences in the following other *Diospyros* species: *D. kaki* (cv. Fuyu for female, and cv. Tohachi for male), *D. virginiana* (cv. Weber for female, and DDIO 69 0003A for male), *D. digyna* (cv. Reineke for female), and *D. mespiliformis* (MIA3483 for male) as an outgroup species. Candidate genes ENOD, TLC, and COX were excluded from this analysis because there was no polymorphism between the X and Y amino acid sequences (Table S3). Furthermore, FE-Contig392, contig42779, contig60102, contig40011, contig60218, and contig77762 (Table S3) were also removed from further analysis because they either produced a fragment size different from that obtained from *D. lotus*, suggesting that they originated from a recent lineage-specific event, or because the PCR product was either multiple bands or a smear suggesting that they originated from repetitive sequences. For the

remaining 12 genes, the sequences from female individuals (putatively homozygous X) were aligned by MAFFT ver. 7 with L-INS-i model and by SeaView ver. 4 for manual pruning. A maximum likelihood (ML) approach was applied to the resulting alignment file for phylogenetic analysis using Mega v.5.05 with 1,000 replications for bootstraps (1/10 values of the calculated bootstraps were shown on the branch). For the ML method, the general time reversible (GTR) model was used as the substitution model, with consideration of invariable site and gamma distribution (nos. of discrete gamma categorization = 4). All sites including missing and gap data were used for the construction of phylogenic trees, and the Nearest-Neighbor-Interchange (NNI) was used as the ML heuristic method.

Method 8. Evolution of *OGI*/*MeGI* family

The genomic and cDNA full length of *OGI* and *MeGI* in *D. lotus* were expanded from seeded contigs using PRICE and CAP3 as described. Multiple primers were designed in or surrounding the *OGI*/*MeGI* ORFs to obtain sequences from other *Diospyros* species. The amplified fragments were sequenced and subjected to phylogenetic analysis. Here, the two inverted-repeats present in *OGI* [forwarded (FR) and inverted repeats (IR)], which are homologous to the third exon of *MeGI*, were analyzed independently to infer the order of *OGI*/*MeGI* divergence, and investigate the possibility of coevolution among the three types of sequences. The nucleotide sequences were aligned by MAFFT ver. 7 and SeaView, followed by a ML approach using Mega v. 5.05 to construct a phylogenic tree, as described (Method 7). To estimate the divergence time between *OGI* and *MeGI*, synonymous substitution ratio (Ks) between *MeGI* and *OGI* were calculated using DnaSP v 5.1, as previously described, and as previously reported from Arabidopsis relatives (*38*) and papaya (*14*), and using an estimated rate of 4×10^{-9} substitutions per synonymous site per year. Note that *OGI* putatively acts as a non-coding RNA with the potential to form hairpin loops resulting in small RNA production. Synonymous sites in *MeGI* could therefore be under significant selective pressure, suggesting higher substitution rates under neutral selection than the Ks we used here.

The putative full-length sequences of the *OGI* and *MeGI* from *D. lotus*, Vrs1 from barley and 80 genes with significant homology to the *MeGI* gene were identified from the genomes of 32 angiosperms, using BLASTp in Phytozome (JGI release version 9.1, http://www.phytozome.net/). *AtHB5* from Arabidopsis was considered as an outgroup gene, according to a previous phylogenetic analysis of this class I HD-ZIP family (*39*). In this study, we selected genes that exhibited significant homology to *MeGI* in *D. lotus* (e^{-19} cut off in BLASTp). Alignment analyses on amino acid sequences were conducted using MAFFT ver. 7 with L-INS-i model. The raw alignments were subjected to manual revision using SeaView ver. 4, and subsequently, to examination of the evolutionary topology by Mega v. 5.05 with WAG+I+G model with 1,000 replications for bootstraps (nos. of discrete gamma categorization = 3). All sites including missing and gap data were used for the construction of phylogenic trees, and the Nearest-Neighbor-Interchange (NNI) was used. Bootstraps were shown on the branches as 1/10 values of the calculated values.

Method 9. Sequencing genomic regions surrounding *OGI*

A BAC library of genomic DNA from *D. lotus* cv. Kunsenshi Male, the male parent of the KK F_1 population, was constructed using the CopyControl pCC1BAC system, as recommended (Epicentre). Clones were screened by colony-hybridization using digoxigenin-labeled nucleotide probes corresponding to *OGI* or flanking sequences. Two clones that tested positive for the presence of the target sequences were selected and sequenced using PacBio technology, as recommended, using one SMRT cell per BAC clone. Data obtained from each SMRT cell was used for assembly using HGAP2.0 and default parameters.

For each BAC clone, the longest assembled contig was retained (~110 and 80 kb). Vector sequence was removed by sequence comparison (BLASTN). The remaining sequences were compared to each other and ~20kb of overlap was found between the two sequences (including a

single mismatched bp). Last, the two sequences were combined to create a single contig, containing 155,135 bps.

Method 10. Transformation

The full lengths of the *MeGI* and *OGI* genomic sequences were amplified by PCR using 2X Phusion High Fidelity PCR Master Mix (NEB) and genomic DNA from *D. lotus* cv. Kunsenshi Male, which is the male parent of the KK population. We used the following primers for PCR amplification: MeGI-LIC26-stF-Gib (5'-CGAGCTAGTTGGAATAGGTTATGACAGCCAACTTTAATCCTCCG-3') and MeGI-LIC26-spR-Gib (5'-TGCAGTATGGAGTTGGGTTTCATATAAGGTTAACCCATTCCATGCC-3') for *MeGI*, and OGI-LIC26-stF-Gib (5'-CGAGCTAGTTGGAATAGGTTCACATATATAAATCATATAAGGTTAACACATTC-3') and OGI-LIC26-spR-Gib (5'-TGCAGTATGGAGTTGGGTTTCCTGGCACACAAAATATTTTCAACCCT-3') for *OGI*, to connect the place the gene under the control of CaMV35S promoter in the pPLV26 vector (*40*). The full length of the *MeGI* and the surrounding genomic sequences were also amplified by PCR using the same conditions as described, and the following primers: NatMeGI-LIC2-stF-Gib (5'-GAATTCTAGTTGGAATGGGTTTTGTAATTTCGACCTGCACTCTCTAC-3') and NatMeGI-LIC2-spR-Gib (5' -TCCTTATGGAGTTGGGTTTGTGCGAGAGAAGCCTAATGTAATT- 3'), to connect to the pPLV2 vector (*40*). We constructed pPLV26-*MeGI*, pPLV26-*OGI*, and pPLV2-native-promoters-*MeGI* using the Gibson Assembly Master Mix (NEB), using the underlined sequences of the primers listed above as overlap with pPLV26 or pPLV2 digested by *Hpa*I, according the manufacturer's recommendations.

Arabidopsis (*Arabidopsis thaliana*) ecotype Columbia-0 was grown under white light with 16-h-light and 8-h-dark cycles at 22°C until transformation. The binary construct was introduced into *Agrobacterium tumefaciens* strain GV3101 (pMP90) using the helper vector pSOUP by electroporation (*40*). Next, Arabidopsis wild-types plants were transformed using the flower-dipping method, according to the previous protocol (*41*). Screening of transgenic plants was conducted on Murashige and Skoog media containing 50 μg/mL kanamycin. Pollen viability was assessed by Alexander staining (*42*). Pollen tube germination was assessed two hours after placing the pollen grains on 10 % sucrose / 1.5% agarose media at room temperature.

Tobacco plants (*Nicotiana tabacum*) cv. Petit Havana SR1 was grown *in vitro* under white light with 16-h-light and 8-h-dark cycles at 23°C until transformation. The binary construct was introduced into *Agrobacterium tumefaciens* strain GV3101 (pMP90) as described above. Young petioles and leaves of tobacco plants were transformed by the leaf disk method (*43*). Transgenic plants were selected on Murashige and Skoog medium supplemented with 100 μg/mL kanamycin. Pollen tube germination was assessed three hours after placing the pollen grains on 15 % sucrose / 0.005 % borate / 1.0% agarose media at 25°C. The pollen germination ratio was counted as average percentages, in batched of 200 pollen grain from the first three flowers. Pollen grains exhibiting pollen tubes longer than the length of the grain were counted as "germinated pollen".

To assess the effect of *OGI* on *MeGI* expression level, we conducted co-expression assays in *Nicotiana benthamiana*. Specifically, pPLV26-*MeGI*, pPLV26-*OGI*, and the external control pPLV26-ppa003808m from the peach (*Prunus persica*) genome (see Table S7) were introduced into *Agrobacterium tumefaciens* strain GV3101 (pMP90) using the helper vector pSOUP, and transiently introduced to *Nicotiana benthamiana* plants carrying 10-12 leaves, using agrobacterium infiltration. The transformed agrobacterium was cultured at 28 ºC for 32 hours, and then suspended to Murashige and Skoog (MS) medium (pH 5.7) including 20 μg/mL acetosyringone. The concentration was adjusted to OD600 = 1.0. For the co-expression experiments, 500 μl of pPLV26-*MeGI*, pPLV26-*OGI*, and pPLV26-ppa003808m agrobacterium suspensions were combined together prior to infection. For

the control experiments, 500 µl of pPLV-26-MeGI, pPLV26-empty, and pPLV26-ppa003808m were agrobacterium suspensions were combined. We infected the third to fifth leaves (N = 3) per plant, and 4 plants per treatment. Three days after the infection, total RNA was extracted from all 3 leaves for each plant, and used for cDNA synthesis using superscript III (Life Technology). *MeGI* expression levels were analyzed by iQ SYBR Green Supermix and qPCR DNA Engine Opticon 2 Continuous Fluorescence Detector (Bio-Rad). Infection efficiency, expression induction by the CaMV35S promoter were normalized to the expression level of ppa003808m.

Supplementary Text:

Text S1. *OGI* is the single best SD candidate amongst the genes located on the MSY

Of the 21 SD candidates except *OGI* (Table 1), 18 genes, while present in distinct X and Y allelic forms, displayed apparently lower rate of substitutions than expected for the SD region, whose origin corresponds to the establishment of dioecy (Fig. S9, Table S3), and suggesting more recent divergence potentially due to historical recombination. Phylogenetic analysis of these genes indicated that their current allelic states postdated the divergence of *D. virginiana* and *D. lotus* at the earliest, later than the establishment of the Y chromosome (Fig. S12). The remaining 3 candidate loci (contig40011, contig60218, and contig77762, in Table S3) were highly repetitive and/or Y-specific and thus no obvious homologous sequence from the X chromosome could be identified to assess the timing of their origin. PCR analyses using various primer sets under low annealing temperatures suggested that these sequences are not conserved in other *Diospyros* species, suggesting that they are not associated with the SD. Furthermore, regions with coding potential, such as open reading frames could not be identified in these sequences (see Table S3). These results suggested that they are lineage specific transposable elements. Together, our results suggest that *OGI* is the best SD candidate amongst the genes located on the MSY identified in this study.

Text S2. Establishment of dioecy in *Diospyros*

Most species within the *Diospyros* genus are dioecious, although a few species/lines are hermaphroditic, monoecious, or polygamous (*18*). It is thus possible that dioecy evolved multiple times within the *Diospyros* genus. On the other hand, fossil records dating from the Eocene suggest that the ancestral status of the Ebenaceae bear unisexual flowers, with features similar to those in the current Ebenaceae species (*20*). One of the important features of unisexuality in the fossil flowers (*20*), the presence of a rudimentary carpel, is also observed in *D. lotus* (Fig. S1C), as well as in the other dioecious *Diospyros* species. This information supports the hypothesis that the establishment of flower unisexuality predates the divergence of the *Diospyros* genus.

In this study, we demonstrated that one of the best SD candidates is *OGI*, a gene that is well conserved within the *Diospyros* genus and associated with maleness (Fig. 2B-C). This remains true in hexaploid *D. kaki* which nests within monoecious cultivars and occasionally bears hermaphroditic flowers (*21*) (Fig. 2B). This indicates that sexuality in these varieties might be controlled by the same factor, and further supports the idea that the establishment of dioecy or at least the control of maleness by *OGI* predates the divergence of the species used in this study, which covers a wide range within the *Diospyros* genus, which is thought to have diverged 25-35 Mya (*19*). Taken together, these data are consistent with the establishment of dioecy, genetically controlled by *OGI* predating at least the radiation of the *Diospyros* species used in this study.

Supplementary Figures:

Figure S1. Male and female flower characteristics in D. lotus

Development of male (top) and female (bottom) lowers in *D. lotus*. **A-F.** Early flower developmental stages. **A and B.** Cross sections of flowers. In male flowers, (residual) carpels cannot be observed, while defective stamens (DS) develop early in female flowers. **C and D.** Scanning electron microscope (SEM) image of dissected flowers. Sepals, petals and stamens have been removed from the male flowers (C), and sepals and petals have been removed from the female flowers (D). Residual carpel (RC) on the receptacle is indicated in (C). **E and F.** Cross sections of (defective) stamens stained by safranin and fast green FCF. Pollen sacs (PS) including putative meiocytes were developing in male (E), while residual stamens (DS) in female (F) generally showed no pollen sacs. **G.** In male pollen sacs, meiocytes (Mi) and tapetum cells (Tp) were differentiated. **H-K.** Flower maturing stages. Male (H and J) and female (I and K) dissected flowers from which the sepals and petals have been removed. Flowers shown in J and K were also cut in half vertically for ease of visualization. **L and M.** Cross sections of (defective) stamens stained by safranin and fast green FCF. Distinct pollen sacs were observed in male (L) but not in female (M) flowers. In pollen sacs of male (**N**), mature microspores (Ms) and tapetum cells (Tp) were observed. Pollen tube (PT) growth was observed from pollen grains collected from male flowers (**O**). Female flowers generally cannot produce pollen grains, although occasionally produce a few pollen-like grains. We could ~~get~~ obtain four pollen-like grains from 10 female flowers, none of which exhibited pollen tube growth (**P**). Mi: meiocytes, Ms: microspores, Ov: ovary, Pe: petal, PS: pollen sac, RC: residual carpel, DS: defective stamens, Se: sepal, Sg: stigma, St: stamens, Tp: tapetum cells. **Q**. Primordium (left) and almost mature (right) stages of male (top) and female (bottom) flowers. Central flower (CF) and lateral flowers (LF) in the simple cyme-like male inflorescence are indicated by black and white arrows, respectively. In contrast, female flowers generally do not fully develop as a three-flower cyme structure. br: bract. **R**. Stunted growth of lateral flower in female *D. lotus*. Stunted lateral flowers can occasionally develop on the pedicel of the single, well developed female flowers, but they are mostly infertile. This difference between male and female in *D. lotus* is similar to what was observed between WT and *vrs1* mutants in barley. Bars indicate 1mm for A, B, H-K and R, 0.5mm for Q, 0.1mm for C-F, L and M, 0.05mm for G and N, and 0.025mm for O and P.

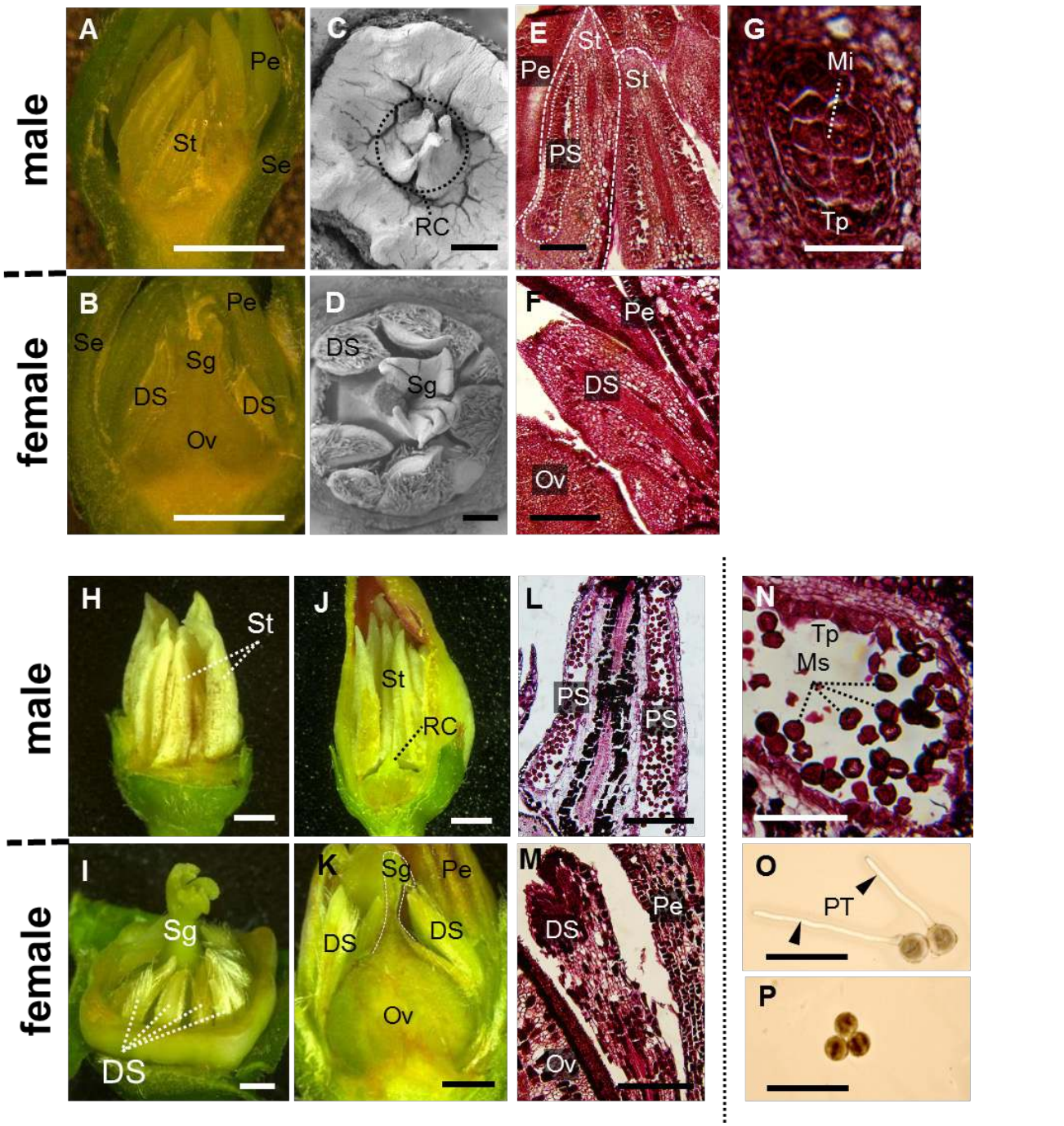

male
female
A
Pe
St
Se
C
RC
E
St
Pe
St
PS
G
Mi
Tp
B
Pe
Se
Sg
DS
DS
Ov
D
DS
Sg
F
Pe
DS
Ov
H
St
J
St
RC
L
PS
PS
N
Tp
Ms
I
Sg
DS
K
Sg
Pe
DS
DS
Ov
M
Pe
DS
Ov
O
PT
P
Q
Primordium
Almost matured
Male
Female
br
LF
CF
LF
br
CF
LF
LF
br
CF
br
R
RF
Fr
Se
Pd
Sm

Figure S2. Pipeline for the identification of the MSY-linked contigs

Multiple iterations of recombination mapping and expansion/integration of the contigs were performed from the 5,100 seed contigs initially constructed by the CLC assembler using all 100-bp paired-end (PE) reads that included male-specific k-mers (MSK) (see Materials and Methods section, Method 3). Finally, 720 contigs were identified that are putatively located on the MSY and for which no recombination could be observed in the 57 sibling F_1 individuals sequenced (KK population).

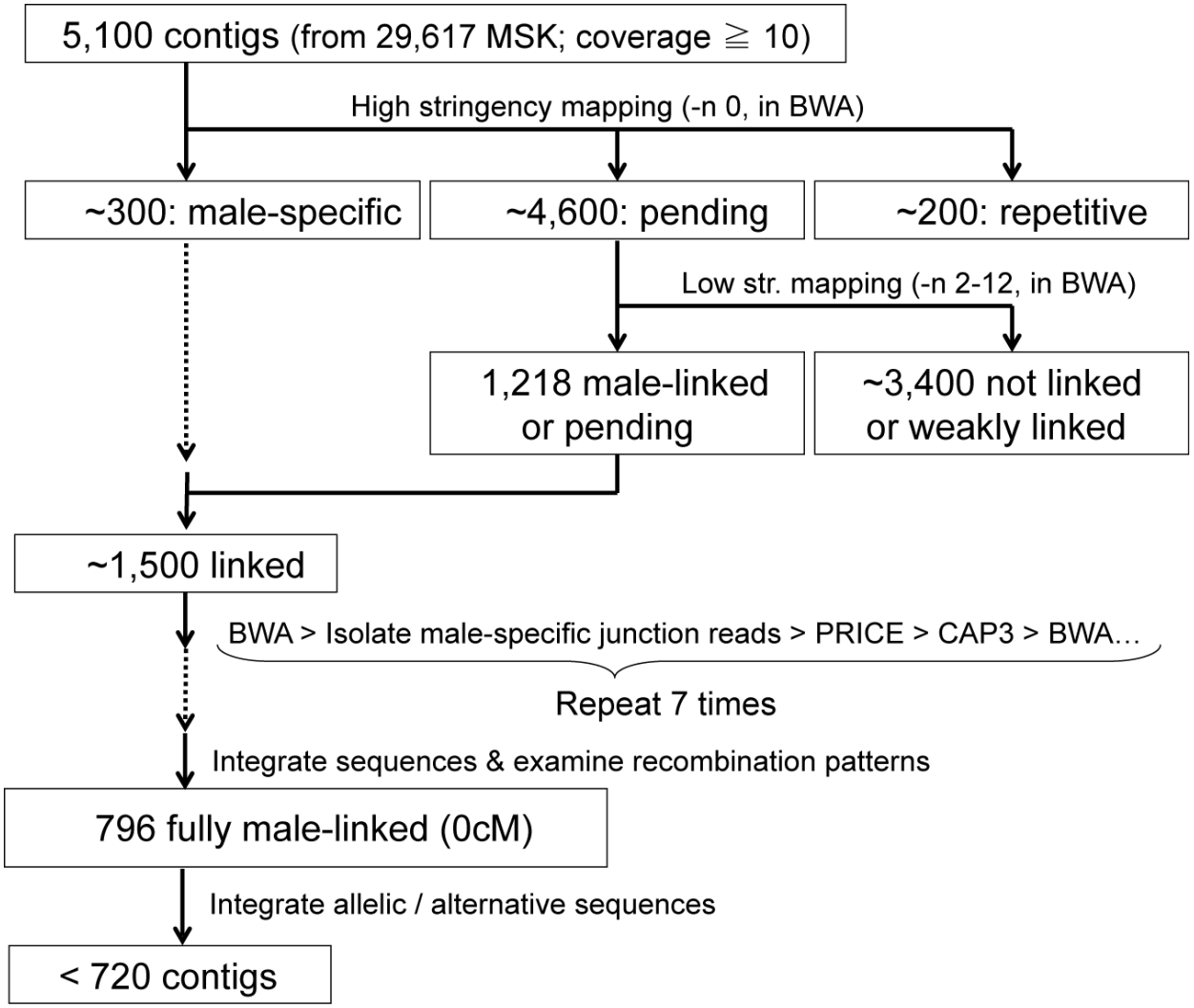


Figure S3. Model of the relationship between male-specific k-mers and the male specific region of the Y chromosome (MSY).

The male specific region of the Y-chromosome is expected to contain a central region, which never undergoes recombination with the X chromosome (red), flanking regions (orange), which undergo low rates of recombination, and peripheral pseudo-autosomal regions (yellow) which undergo rates of recombination indistinguishable from autosomal regions. Our definition of MSY thus covers a wider region than the MSY presented in previous reviews of plant sex chromosomes architecture (*5*-*8*). Male-specific k-mers were identified as such if they were only present in male reads and absent in female reads. They were assemble into male-specific contigs. Such contigs and k-mers can originate either from regions present both in male and female genomes but containing male-specific polymorphisms (small indels and/or nucleic acid substitutions, white contigs) or from regions specific to the male genome (blue contigs).

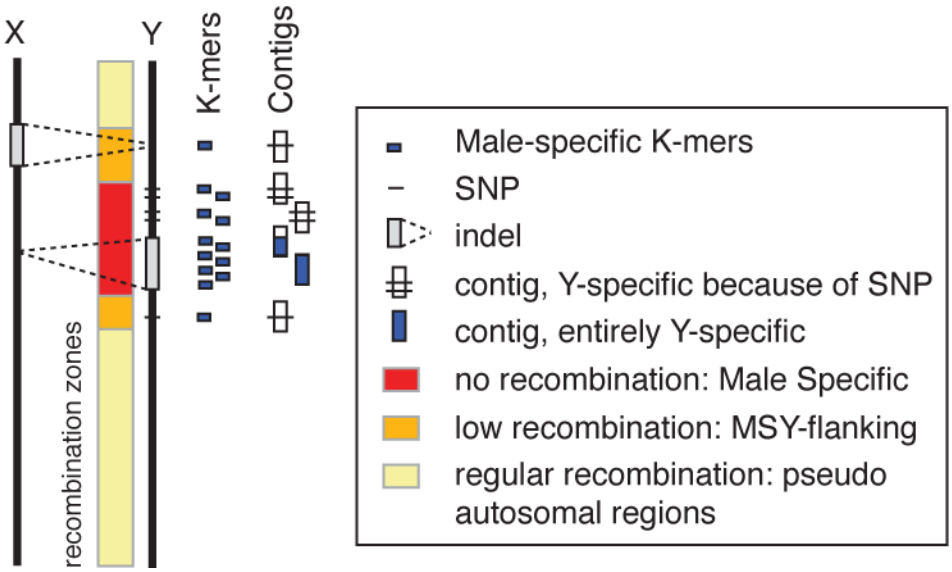


Figure S4. Identification of male-specific k-mers

Numbers of sex specific k-mers isolated from RNA-Seq reads obtained from 9 male and 9 female individuals from the KK population and their parents (N = 2 x 10). In an approach analogous to the one applied to the genomic contigs (Fig. 1), male-specific k-mers (MSKs) were isolated directly from the RNA-Seq read sequences (k = 35 bp, ≥ 20X coverage). Next, all reads containing one or more of these MSKs were assembled into cDNA contigs using the CLC assembler. All kmers are displayed in the inset while only the biased and sex-specific k-mers are displayed in the main graph.

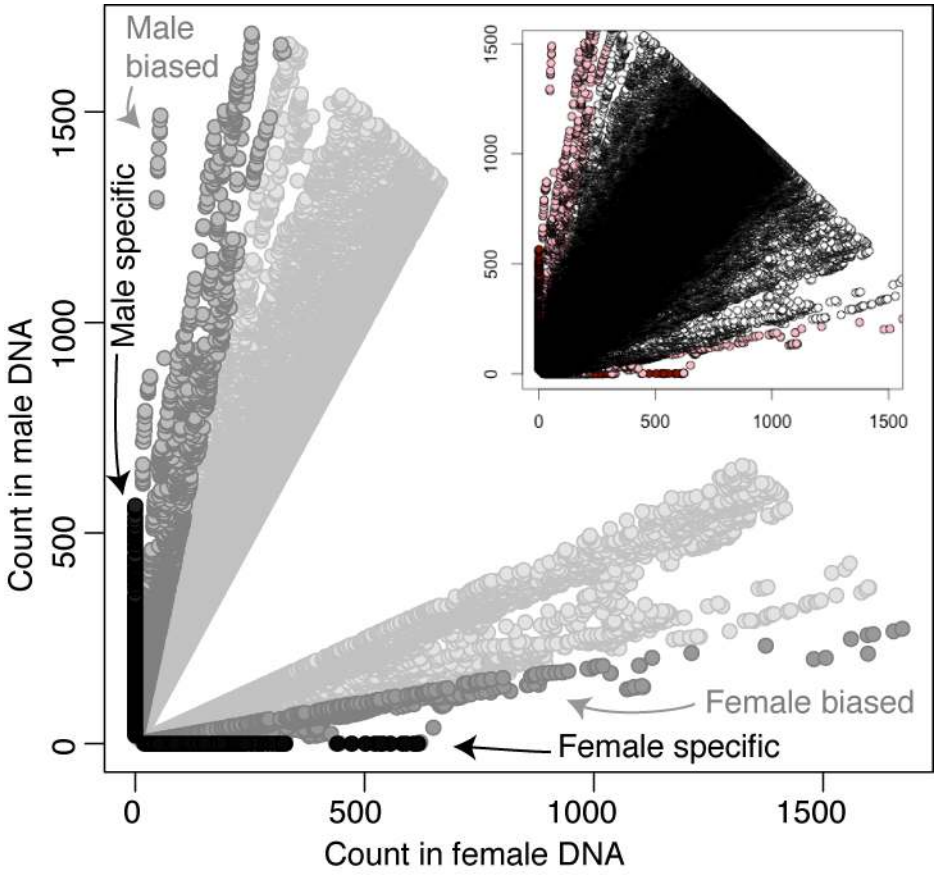


Figure S5. Integrated analysis of expressed genes underlying the SD locus

Overlap between the MSY-linked cDNA contigs derived from the genomic and the RNA-Seq reads (i vs ii, see Text and Materials and Methods, "Identification of expressed genes involved in sex determination") was very high; 90% of the contigs identified using approach (ii) (27/30) shared the sequences of the contigs from approach (i). On the other hand, > 75% of the contigs derived from approach (i) (46/60) shared the sequences of the contigs identified using approach (ii). Generally, the cDNA contigs identified from approach (i) were supposed to be fragmented ORFs, and were shorter than the contigs identified from approach (ii). The 3 cDNA contigs identified specific to (ii) contained highly repetitive sequences, which could have prevented their identification from whole genome sequencing data. Twelve of the 14 contigs identified only from approach (i) were non-polymorphic short fragments i.e. these genes only exhibited X-Y polymorphisms in introns and upstream / downstream regions, preventing their identification as polymorphic using RNA-Seq data only. These results confirmed the validity of our approach and suggested that we had identified the majority of the polymorphic regions in the MSY. After integration of the genes identified in the different approaches, 22 genes were selected as best candidates underlying the SD locus (Tables 1 and S3).

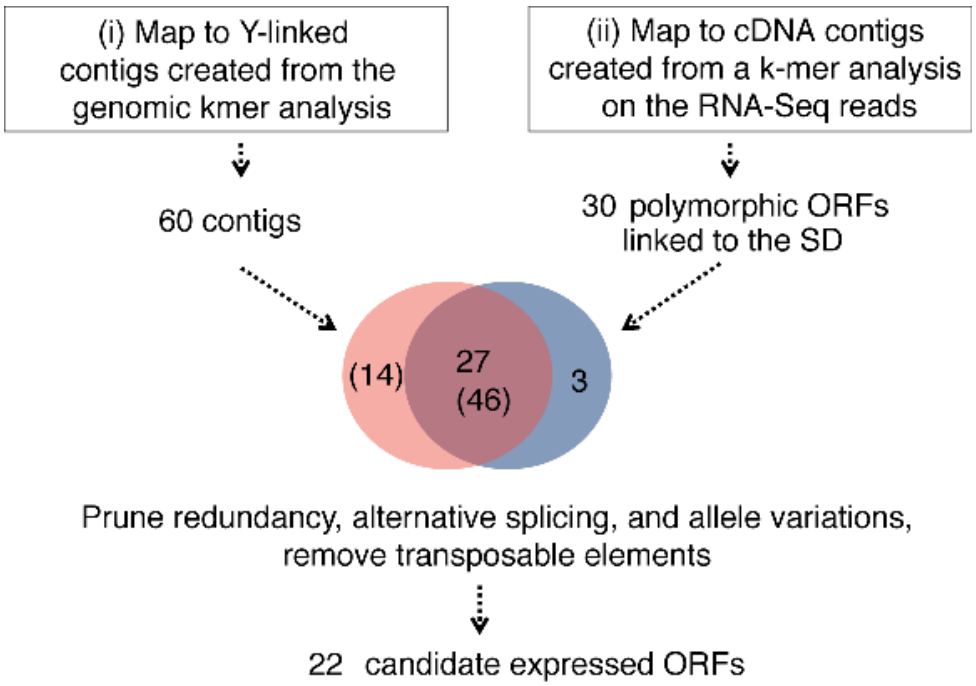


Figure S6. RNA and smRNA Expression patterns of MeGI and OGI in organs of D. lotus

To assess the abundance levels of *MeGI* mRNA and *OGI* primary RNA transcript in various *D. lotus* organs, cDNA synthesized from total RNA from each organ was subjected to PCR analysis using *MeGI*- and *OGI*-specific primers. **A.** *MeGI* exhibited bud- and flower-specific expression and higher expression in female than in male buds, both in dormant/developing buds and developing flowers. **B**. *OGI* exhibited male-specific and bud- and flower-specific expression. The expression level of *OGI* in male flowers was, however, much lower than in the buds. Samples were collected from adult trees from the KK population. The developmental stages investigated are: whole flowers developing gyno/androecia (corresponding to Fig. S1A and B for male and female, respectively, sampled on April 17th 2014), buds during primordium development (Dev. buds, sampled on July 4th 2013) and dormant buds (Dor. buds, sampled on Feb. 1st, 2014), from both male (M) and female (F) trees. Additionally, for *MeGI* (A), young leaves (YL) and mature leaves (ML), stem (St), developing seeds (Se), calyx (Cx), and young (YF) and mature (MF) fruit flesh for *MeGI* (A) were sampled from the female parent of the KK population. For *OGI* (B), young and mature leaves, and stem were sampled from the male parent of the KK population. The young leaves and stems were sampled on May 11th 2011. The mature leaves, developing seeds, calyx, and young fruit flesh were sampled on July 28th 2011. The mature fruit flesh was sampled on October 11th 2011. All plant materials were planted and sampled at the experimental orchard of Kyoto University, Kyoto, Japan. [a] The *actin* gene was used as an expression reference gene, and exhibited generally high expression in all organs tested (see Table S7). **C.** Expression levels of *OGI* and *MeGI* transcripts and their smRNA forms in developing bud and developing flowers. *OGI* and *MeGI* transcript levels were measured by qPCR using the *actin* gene as an expression reference gene. Mean relative expression levels and standard deviation values are shown. Small RNA levels were measured after mapping to the *OGI* and *MeGI* reference sequence reads derived from small RNA libraries, which were made from developing buds (Dev. Buds, sampled on June 17th 2013) and developing flowers (Dev. Flower, sampled on April 17th 2014).

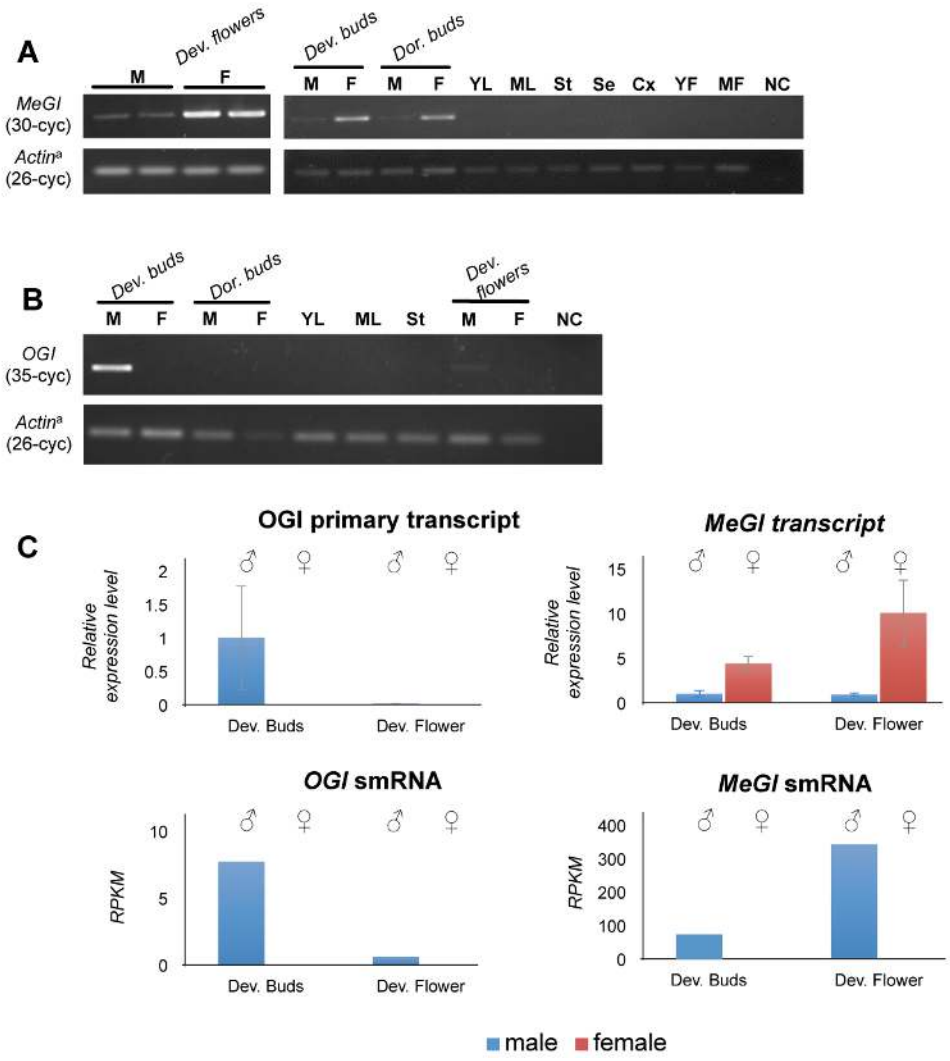

A
Dev. flowers
Dev. buds
Dor. buds
M
F
M
F
M
F
YL
ML
St
Se
Cx
YF
MF
NC
MeGI
(30-cyc)
Actin[a]
(26-cyc)
B
Dev. buds
Dor. buds
Dev. flowers
M
F
M
F
YL
ML
St
M
F
NC
OGI
(35-cyc)
Actin[a]
(26-cyc)
C
OGI primary transcript
Relative expression level
2
1.5
1
0.5
0
♂
♀
♂
♀
Dev. Buds
Dev. Flower
MeGI transcript
Relative expression level
15
10
5
0
♂
♀
♂
♀
Dev. Buds
Dev. Flower
OGI smRNA
RPKM
10
5
0
♂
♀
♂
♀
Dev. Buds
Dev. Flower
MeGI smRNA
RPKM
400
300
200
100
0
♂
♀
♂
♀
Dev. Buds
Dev. Flower
male
female

Figure S7. Evolutionary tree of MeGI/OGI HD-Zip orthologs from angiosperm genomes

A. Maximum-likelihood (ML)-based phylogenetic analysis of *MeGI*/*OGI* HD-Zip orthologs identified in the whole genome sequences of 32 angiosperms, including *MeGI*/*OGI* sequences from *Diospyros* and *Vrs1* from *Hordeum*, and using *AtHB5* as an outgroup gene (class I HD-Zip, see Method 8) (*38*). **B.** ML-based phylogenetic analysis focusing on 20 species nested within the eurosids, with *Vrs1* as an outgroup. Different colors indicate different families within the eurosids. Bootstrap values were calculated from 1000 replications and indicated as 1/10 values on the branches. The results of the phylogenetic analysis suggest the existence of a single gene as the most recent common ancestor (MRCA) in angiosperm, eudicot, eurosid, and asterid, respectively, consistent with an overall monophyletic evolution of this gene family until the establishment of the different families, after which, it exhibits lineage-specific divergences and duplications.

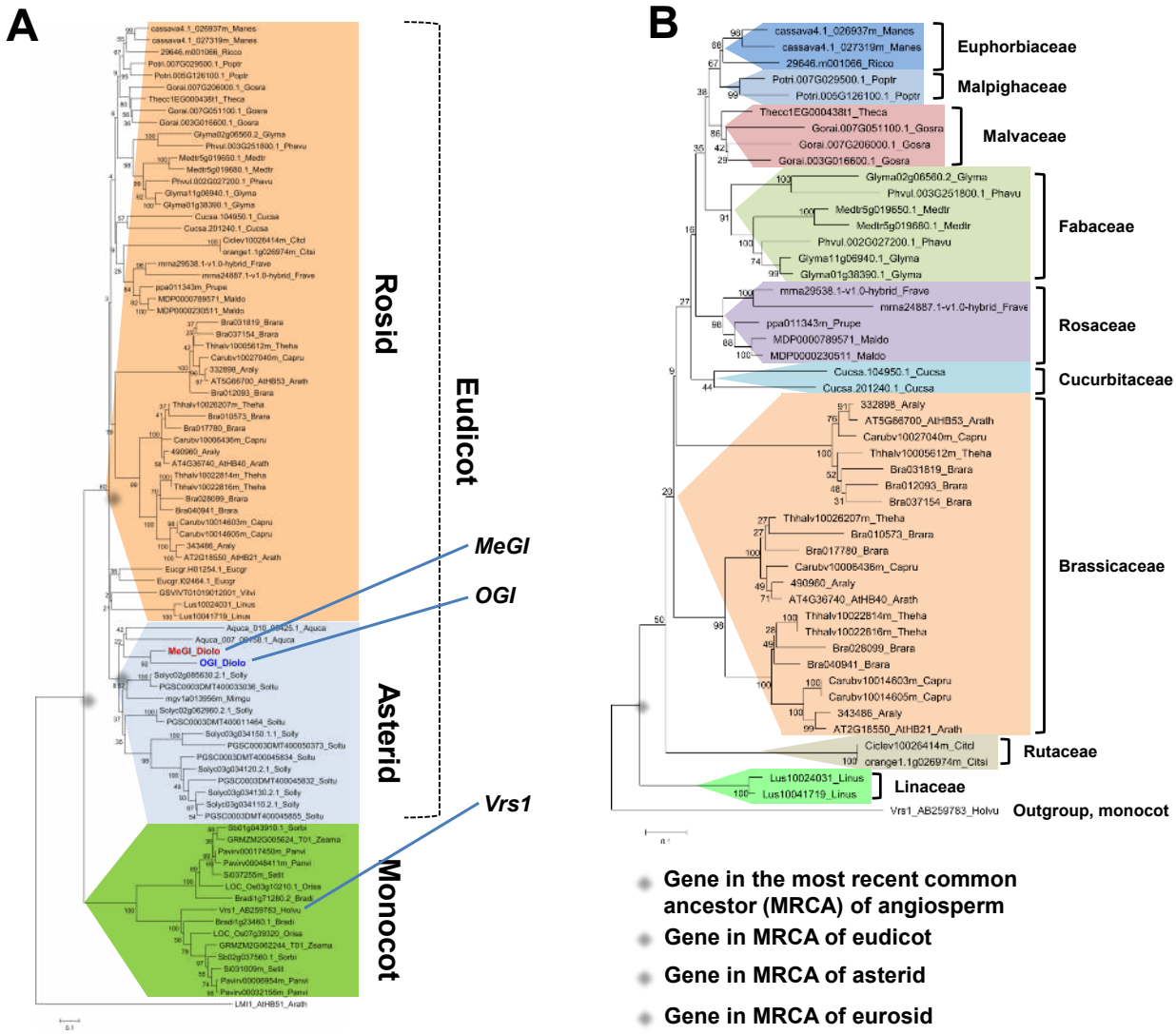


Figure S8. Prediction of the secondary structure of the OGI mRNA and homology to the MeGI gene

A: Reads from the RNA-Seq analysis were assembled to derive a sequence, putatively corresponding to a primary micro (small) RNA (pri-miRNA) for the non-coding *OGI* gene. This sequence was subjected to secondary structure and binding strength prediction using CentroidFold (http://www.ncrna.org/centroidfold/). The *OGI* sequences were predicted to form a double stranded RNA.

B: Sequence homology between the duplicated sequences of *OGI,* predicted to form dsRNA, and the corresponding *MeGI* sequence.

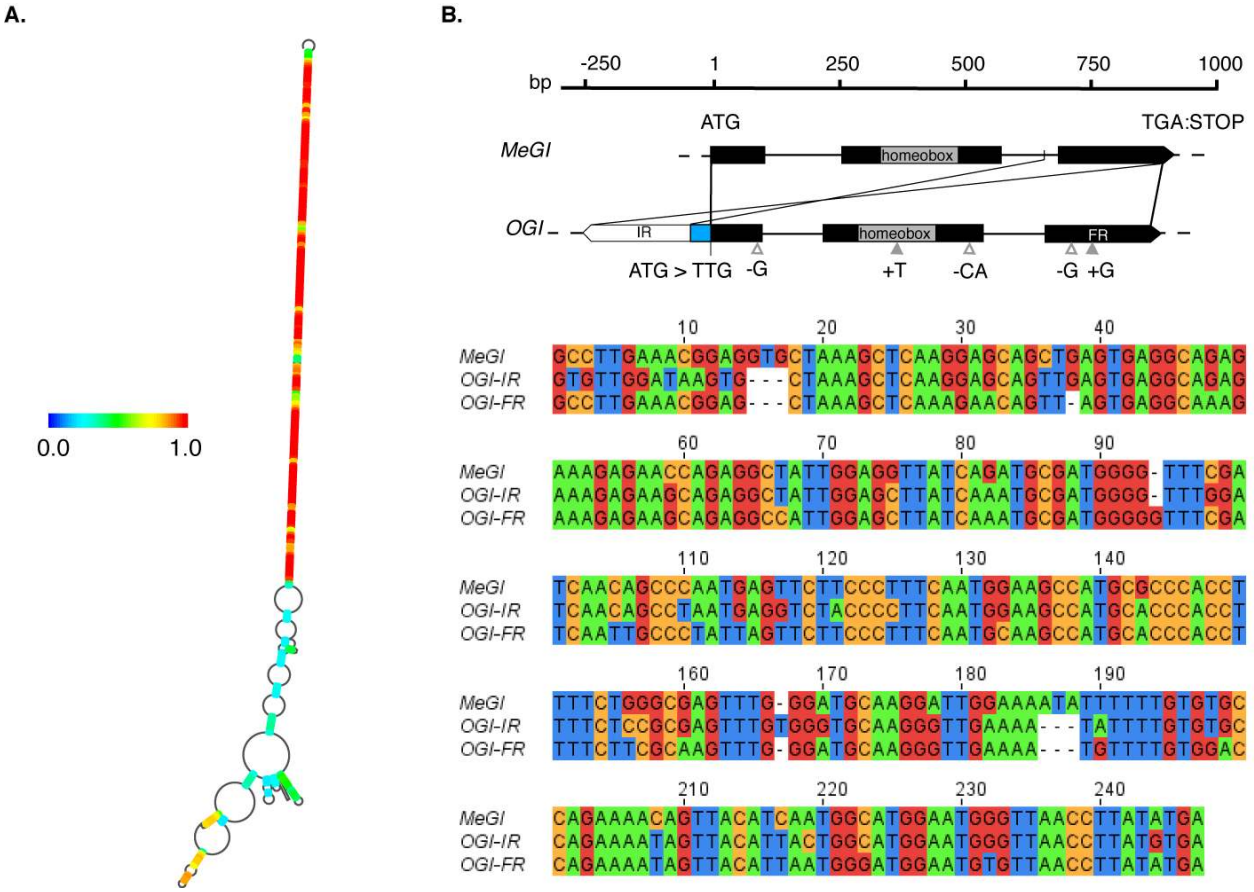


Figure S9. Silent and net divergences in X vs Y alleles of the SD candidate genes
All-sites divergence (**A**) and Jukes and Cantor corrected values of synonymous substitution ratios (Ks) (**B**) between the X and Y allelic sequences of SD candidate genes in *D. lotus* were calculated. The pattern of divergence observed in the sequences of the autosomal gene phytochrome A (blue circles) from a variety of *Diospyros* species (*18-19, 37*), was used to estimate the timing of divergence between *OGI* and *MeGI* (green circle), and between X and Y alleles from other SD candidates (black circles). These results suggest that the current X and Y alleles of the SD candidates tested, diverged relatively recently, at the earliest after divergence between *D. lotus* and *D. virginiana*, while *OGI* and *MeGI* diverged much earlier (red circles indicate potential divergence times; see Text S2). Only the 14 SD candidate genes for which the open reading frames could be annotated are included in panel B (Table S3).

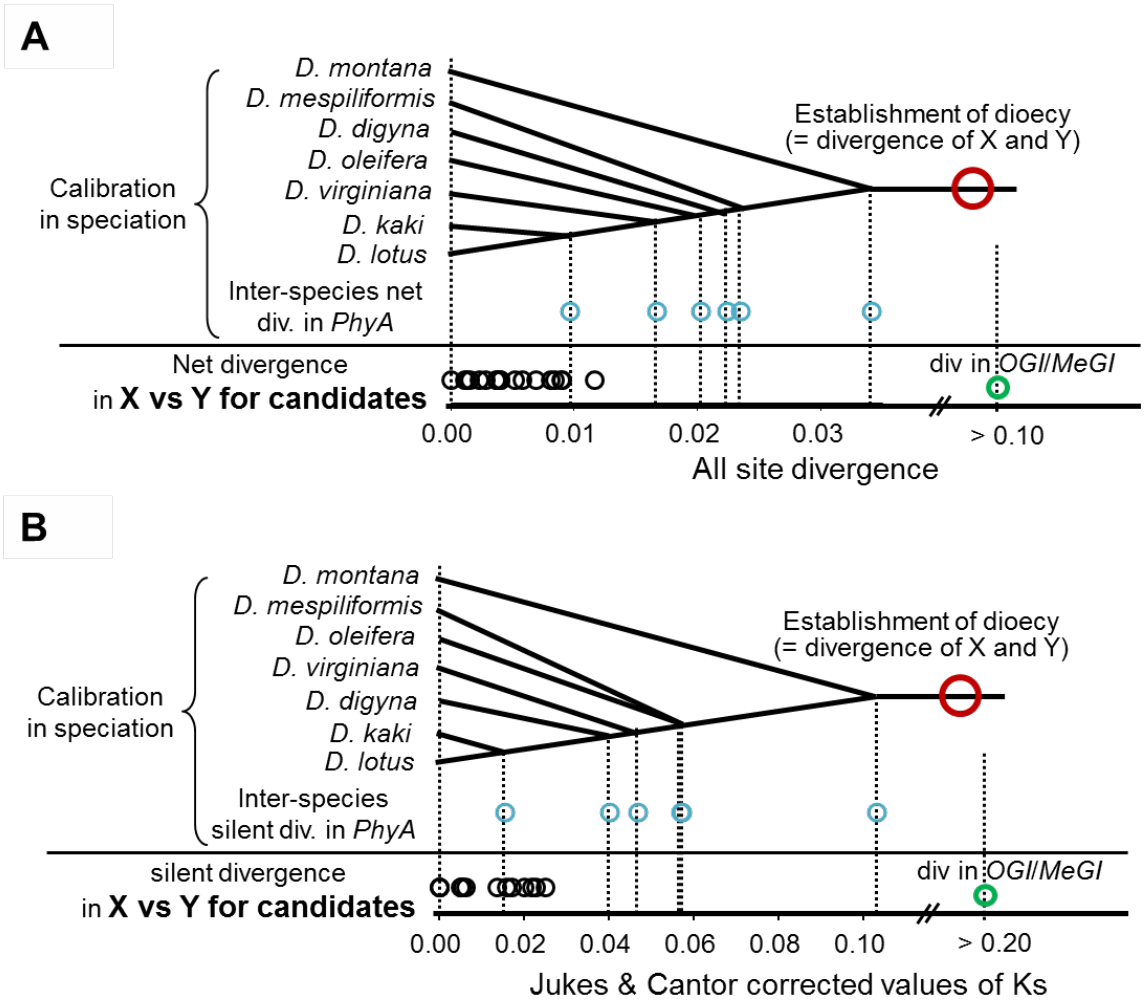


Figure S10. Co-evolution between forward and inverted repeats in OGI

The sequences of the forward (FR) and inverted (IR) repeats of *OGI*, and the corresponding *MeGI* sequences from *D. lotus, D. mespiliformis and D. virgiana* were aligned. Our results suggest that the divergence between the FR and IR of *OGI* predates the origin of the *Diospyros* species (Fig. 2C). Nevertheless, lineage-specific nucleotide substitutions common to the FR and IR (pink boxes) of *D. mespiliformis* or *D. virginiana* (indicated by "M" and "V", respectively) could be identified, suggesting constrained evolution of the IR and FR sequences within a species. This is consistent with the idea that substitutions that could affect the formation of doubled stranded RNA and the production of smRNA targeting *MeGI* have been under strong selective pressures independently in each species.

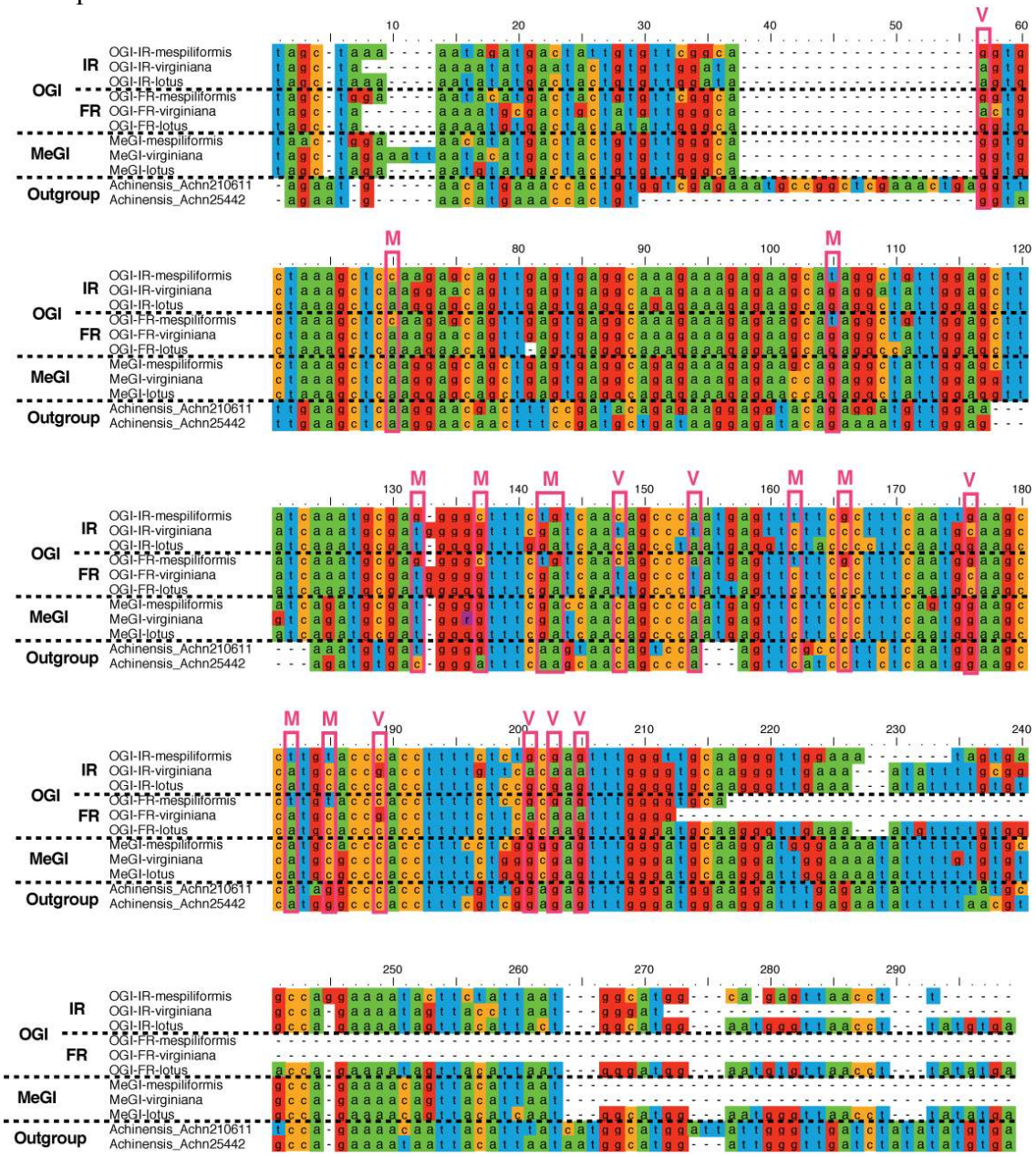


Figure S11. Genomic sequences surrounding the OGI locus.
Two BAC clones were selected based on positive hybridization signal after probing with *OGI* or linked sequences and sequenced using PacBio technology. After removal of vector sequences, the two largest contigs overlapped by ~ 20 kb. These two sequences were merged to produce a single contig of 155,135 bps. Sequencing reads obtained from genomic and RNA libraries were pooled according to gender and mapped to this contig. In the case of the mRNA sequences, reads were trimmed to 50 bps each prior to mapping to minimize the risks of mapping over intron-exon junctions. In order to identify true MSY-specific reads, only reads that mapped perfectly (no mismatches) to the contig were retained. Next, reads were pooled in consecutive non-overlapping 500 bp bins and the number of reads falling into each bin was recorded. Top panel: The number of male genomic read per bin is depicted (log2 scale), demonstrating the existence of many regions of high coverage, a feature diagnostic of repeated regions. Indeed, using this subset of the genomic reads, an average of 10X coverage is expected for male-specific single copy sequences. For both read types (RNA and genomic), gender bias was measured by calculating the percentage of male reads in each bin (see color legend on the right). Bins for which less than 10 reads total (male + female reads) were obtained were labeled undetermined (colored white). Several fully male-specific regions can be identified, including one located close to the middle of the contig and including the *OGI* sequence (as indicated on top).

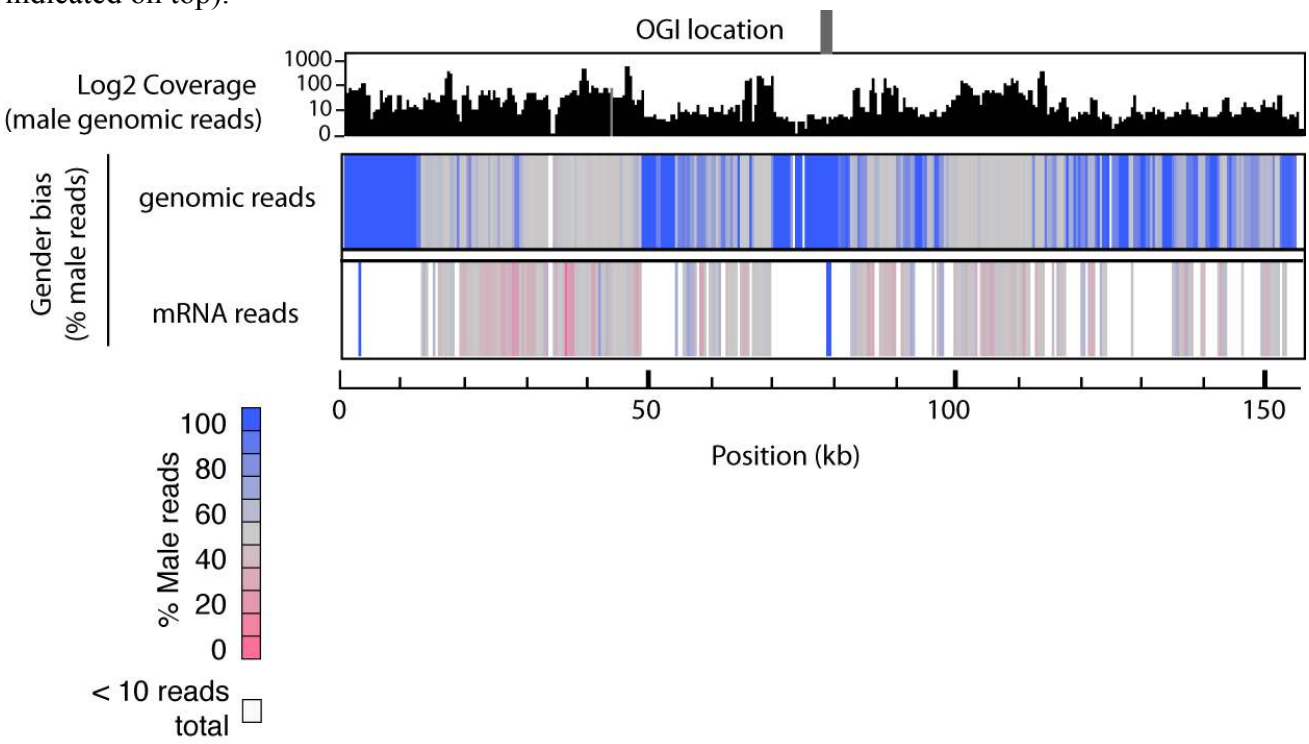


Figure S12. Phylogenetic analysis of the divergence between the X and Y alleles of SD candidate genes in Diospyros

A: Theoretical tree depicting the expected topology of divergence between the X and Y alleles of the SD gene located in the male-specific region of the Y chromosome (MSY) since before the divergence of the *Diospyros* species. In this case, the X and Y alleles are expected to form independent clades.

B: Phylogenetic trees of 12 SD candidate genes exhibiting significant synonymous substitutions between the X and Y alleles in *D. lotus* (see Table S3). The X and Y allelic sequences from *D. lotus* were aligned to putative X alleles from other *Diospyros* species, which were all isolated from female accessions (see Method 8). The resulting topologies indicated that the divergence of the current X and Y alleles in all of these *D. lotus* SD candidates postdates at least the divergence of *D. lotus* and *D. virginiana*.

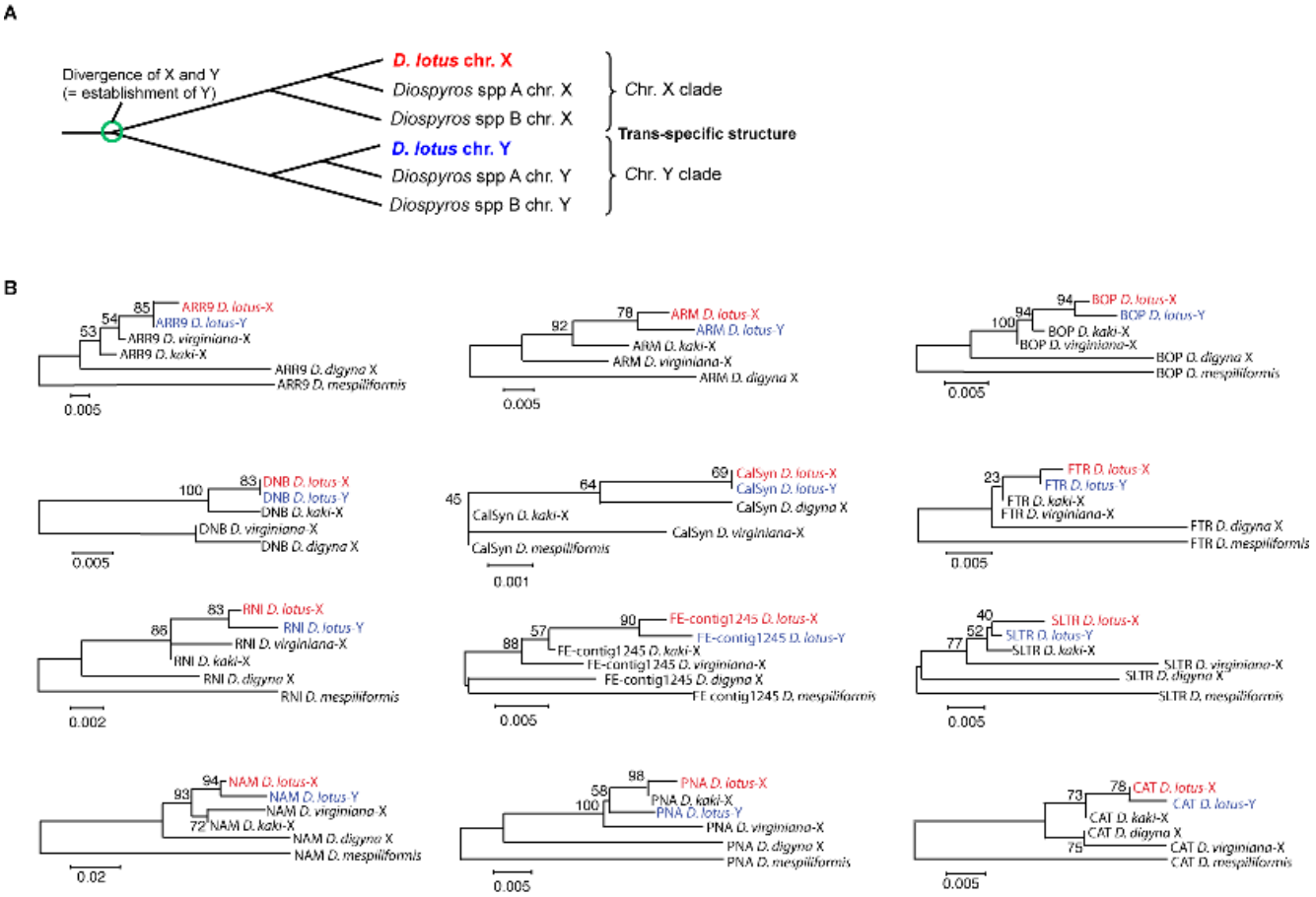


Figure S13. Repression of MeGI by OGI and evidence of alternative splicing of the MeGI transcript from smRNA mapping

A. Standardized *MeGI* expression level after transient co-expression with *OGI* in *N. benthamiana* leaves (Method 10). Overexpression of *OGI* suppressed the expression of *MeGI* (P=0.00082, Student's T-test). Bars indicated standard errors (biological replicates N = 4). **B**. smRNA produced from developing buds of male *D. lotus* trees were sequenced and mapped simultaneously to both the cDNA and genomic sequence of the *MeGI*. Mapped reads are shown in different color depending on their mapping quality, with unambiguously mapped reads in pink (forward mapped reads) or blue (reversely mapped reads) and ambiguously mapped reads in grey. By mapping simultaneously to both the genomic and cDNA reference of the *MeGI* gene, all reads that map equally well to both sequences are colored in grey. The presence of reads specific to either the intron/exon junctions of the cDNA or the region corresponding to the second intron of the genomic sequence suggests alternative splicing of the second intron but full splicing of the first intron.

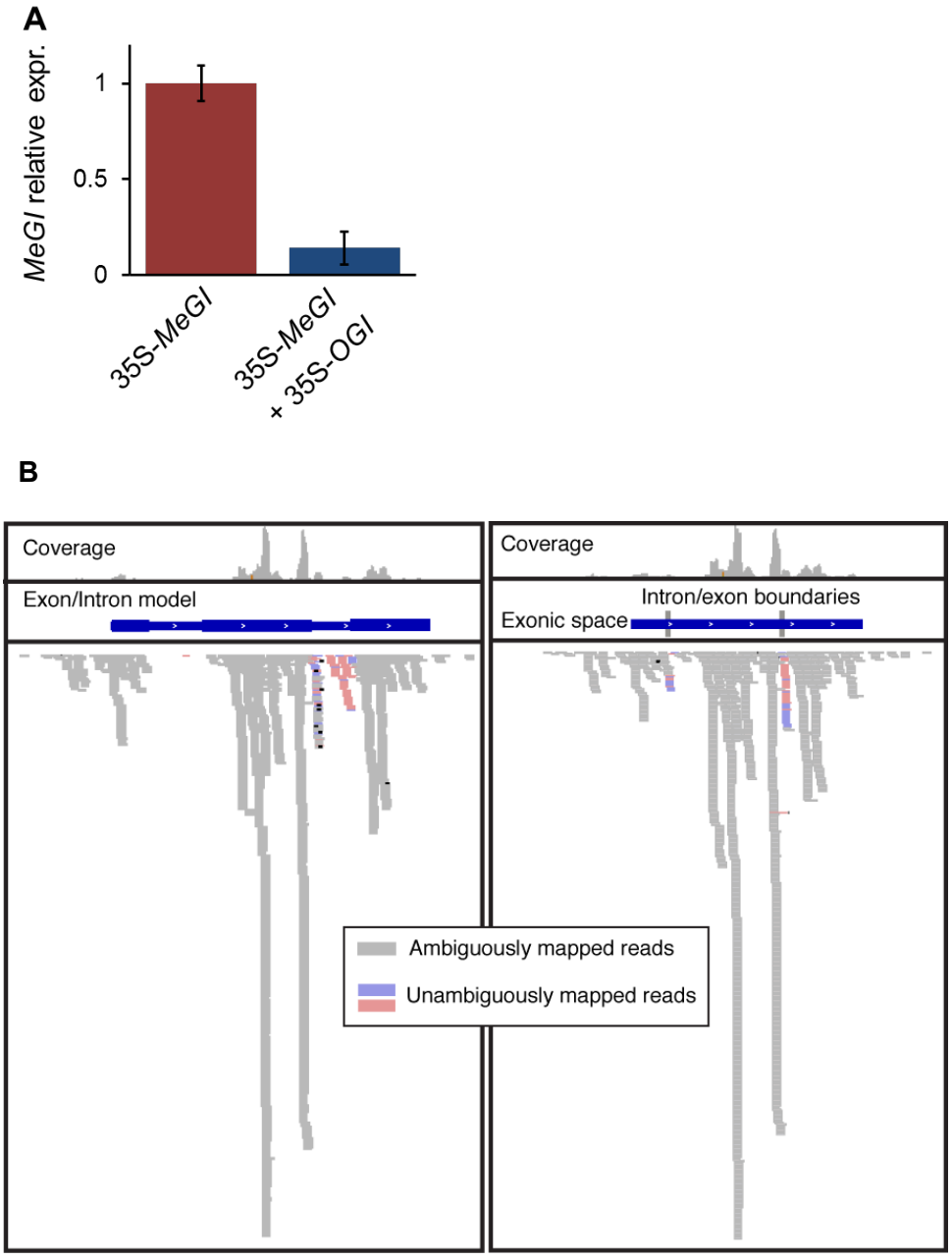


Figure S14. Phenotype of the WT-like hermaphrodite 35S-MeGI transformed Arabidopsis plants. **A**: 4 week-old control and transformed plants. Leaf serration was detected in some of the transformed plants (see Table S4), suggesting potential meristematic growth inhibition by *MeGI* gene expression. The three feminized (and growth-stunted) transformed plants exhibited more severe leaf serration phenotypes (see Fig. 3D), supporting the idea that this effect was dosage-dependent and derived from *MeGI* overexpression. **B**: Flowers produced by the apical meristem exhibited a slight reduction in the numbers of stamens in two of the transformed non-developmentally-delayed plants (see Table S4).

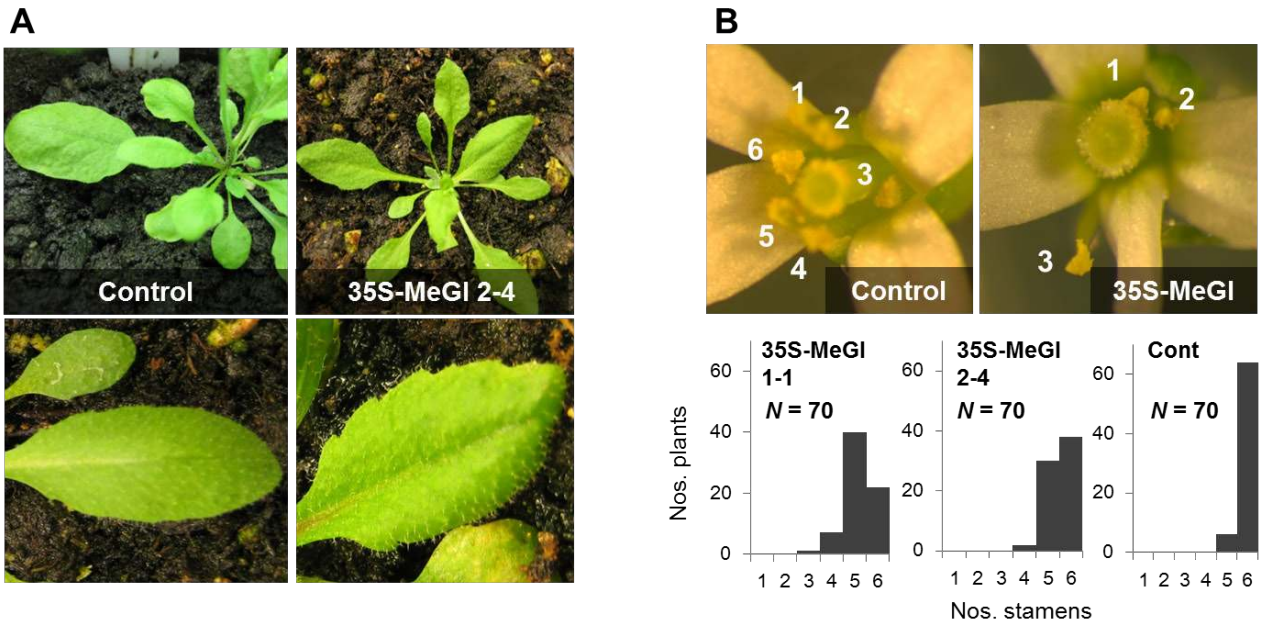
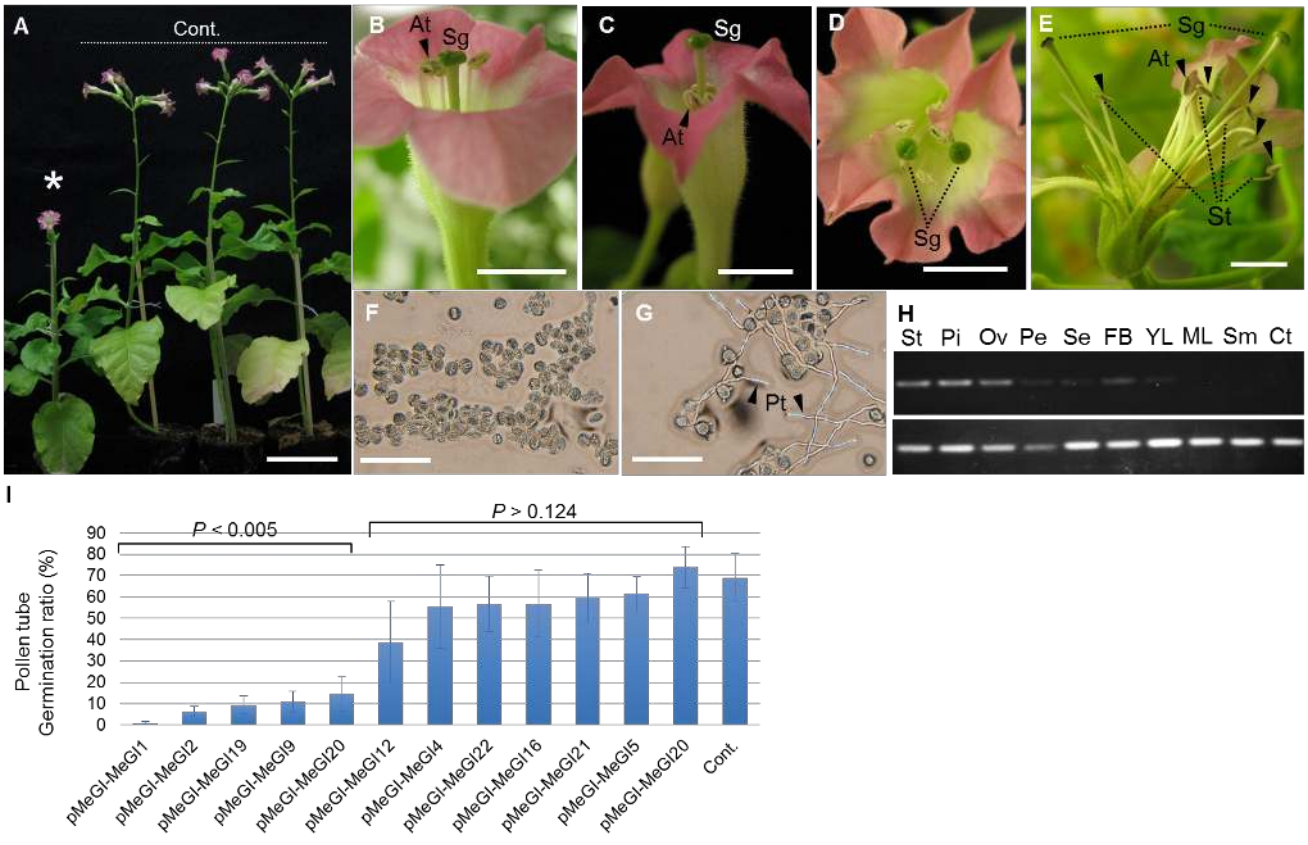


Figure S15. Phenotype of the tobacco plants transformed with MeGI under the control of the native MeGI promoter (pMeGI-MeGI).

We obtained a total of 12 pMeGI-MeGI transgenic planes. **A-G.** Representative features of approximately 2 months-old control (Cont.) and five transformants showing severe phenotypes. The controls were transformed with empty vectors. **A.** The pMeGI-MeGI transgenic plants (marked by asterisks) often exhibited flower abortion in panicle inflorescence, resulting in solitary inflorescence-like structures. This observation is consistent with the architecture of the female flower in *D. lotus* (Fig. S1Q-R). The transgenic plants also exhibited semi-dwarf phenotypes, suggesting potential meristematic growth inhibition by *MeGI* gene expression, as observed in the 35S-MeGI transgenic Arabidopsis plants (Fig. 3). **B.** Control flowers exhibited normal growth of androecia in which anthers (At) can reach to stigma (Sg). **C.** The pMeGI-MeGI transgenic plants showed shorter anthers. **D and E.** The first flower of independent transgenic plants exhibited merged double-flowers with two stigma and generally ten stamens. Stamens (St) were also shorter and anthers did not reach the stigma(s). **F and G.** The transgenic plants produced a high ratio of small and nonfunctional pollen-like grains, which do not have the ability to grow pollen tubes (PT) (**F**), while pollen grains from control plants were fertile (**G**). **H.** Top: Expression patterns of *MeGI* in one pMeGI-MeGI transgenic plant (pMeGI-MeGI1). St: stamen, Pi: pistil, Ov: ovary, Pe: petal, Se: sepal, FB: flower buds, YL: young leaves, ML: matured leaves, Sm: stem, Ct: developing flowers in control plants. Bottom: As an expression reference, we used *actin* gene (EU938079). **I.** Pollen tube germination percentages in all 12 transgenic plants, as well as in control non-transformed plants. Five of 12 transformants exhibited significantly decreased percentages (P=0.001-0.005, Student's T-test), while the others exhibited similar though not-significant tendencies (P>0.124, Student's T-test). Scale bars in A-G indicate 100mm for A, 10mm for B-E, and 0.05mm for F-G. Bars in I indicate standard deviations (SD).



Supplementary Tables:

Table S1. Candidate SD genes identified from the male-specific sequences

Twenty two candidate genes were identified from cDNA fragments located on the MSY (approach (i) and (ii) on RNA-Seq data; refer to Method 4, and Fig. S5). Annotations were given by BLASTX using the TAIR database. Expression ratio between male and female samples, and expression levels, expressed as reads per kb per millions reads (RPKM), are indicated.

Gene/Contig index	Highest hit in TAIR database by BLASTX				Expression (log_2F/M[a])	RPKM	
	Locus ID	Annotation	blast score	e-value		male	female
OGI	AT4G36740	HB-5, ATHB40, HB40 | homeobox protein 40	71	2.00E-21	MS[c]	1.01	0.00
NAM	AT5G64060	anac103, NAC103 | NAC domain containing	176	2.00E-48	-0.69	16.27	10.07
ARR9	AT3G57040	ARR9, ATRR4 | response regulator 9	137	1.00E-38	-0.59	4.73	3.15
ENOD	AT5G15350	ENODL17, AtENODL17 | early nodulin-like protein	168	2.00E-51	0.15	43.71	48.46
BOP	AT2G13690	BTB/POZ domain-containing protein	259	7.00E-77	0.00	25.64	25.58
TLC	AT1G21790	TRAM, LAG1 and CLN8 (TLC) lipid-sensing domain	358	2.00E-121	-0.34	9.44	7.44
DNB	AT5G63960	DNA binding;nucleotide binding	1154	0	-0.22	8.47	7.26
PNA	AT2G39435	phosphatidylinositol N-acetyglucosaminlytransferase subunit P-like protein	107	1.00E-25	-0.28	3.50	2.88
COX	ATMG00160	COX2 | cytochrome oxidase 2	421	2.00E-146	-0.11	76.48	70.96
CAT[b]	AT2G39450	MTP11, ATMTP11 | Cation efflux family	481	8.00E-168	-0.10	27.88	26.07
SLTR[b]	AT5G13550	SULTR4;1 | sulfate transporter 4.1	427	1.00E-135	-5.56	2.73	0.07
RNI[b]	AT4G19210	ATRLI2, RLI2 | RNAse l inhibitor protein	998	0	-0.17	35.97	31.98
CalSyn[b]	AT2G13680	CALS5, GLS2, ATGSL02 | callose synthase	604	0	0.08	2.47	2.61
FTR[b]	AT5G64600	O-fucosyltransferase family protein	301	2.00E-97	-0.10	9.70	9.03
ARM[b]	AT1G77600	ARM repeat superfamily protein	631	0	-0.28	9.68	8.00
contig42779	no hit				-0.20	10.03	8.72
contig60102	no hit				-0.15	3.03	2.73
contig40011	no hit				-3.15	2.74	0.31
contig60218	no hit				0.21	7.10	8.23
contig77762	no hit				-3.34	7.20	0.71
FE-Contig1245	no hit				0.16	11.08	12.35
FE-Contig392	no hit				0.28	38.03	46.25

[a] F/M: Female / Male

[b] the full length ORF sequences have not been identified

[c] male specific expression

Table S2. Genes differentially expressed in male and female (FDR < 0.01, *n* = 10 male and 10 female samples)

Sixty two genes exhibited significantly different expression levels in male vs female developing buds after DESeq analysis. Annotations using the TAIR database, putative functions, expression level (RPKM), statistics for differential expression (FDR), expression ratio between male and female, and linkage to the SD are indicated.

Contig names	Highest hit in TAIR database by BLASTX		Putative functions	Gene name	RPKM		FDR value	Expression (log_2F/M)	Linkage to SD
	Locus ID	Annotation			in male	in female			
Contig333146	no hit		Transposable element		28.52	0.20	0.E+00	-7.11	X
Contig413090	no hit				9.31	0.00	3.E-49	MS	X
Contig202197	no hit		Transposase		5.35	0.12	4.E-38	-5.42	X
Contig173884	no hit				3.17	0.02	2.E-34	-7.28	
Contig318451	no hit				7.29	18.89	7.E-26	1.41	
Contig134463	no hit		Transposable element-like repetitiive		3.52	0.00	2.E-25	MS	X
Contig314662	no hit				10.75	25.21	2.E-23	1.29	
Contig297812	AT1G77610	EamA-like transporter family	Transporter protein		5.32	9.84	9.E-16	0.95	
Contig350312	no hit				13.40	0.16	1.E-14	-6.13	
Contig217628	AT2G39450	MTP11, ATMTP11	Cation efflux family		5.74	0.21	6.E-14	-4.65	X
Contig300284	AT1G71010	FAB1C	phosphatidylinositol-4-phosphate kinase family		4.00	7.74	1.E-13	1.02	
Contig316125	no hit				1.10	2.56	2.E-13	1.32	
Contig313207	AT5G20630	GLP3, GLP3A, GLP3B, ATGER3, GER3	germin-like protein		9.89	21.43	3.E-12	1.22	
Contig214180	AT4G36740	HB-5, ATHB40, HB40	class-I bZIP-Homeodomain transcription factor	***Oppressor of meGI (OGI)***	1.01	0.00	1.E-11	MS	X
Contig398696	no hit				11.29	22.34	1.E-10	1.09	
Contig345837	no hit				2.51	0.22	4.E-10	-3.35	
Contig240110	no hit				0.24	3.54	3.E-09	3.55	
Contig325568	no hit				12.23	5.24	2.E-07	-1.09	
Contig219578	no hit				8.83	2.77	5.E-07	-1.38	
Contig211185	AT4G23160	CRK8	cysteine-rich RLK		0.76	2.08	5.E-07	1.45	
Contig191907	no hit				15.37	5.79	6.E-07	-1.35	
Contig318380	no hit				10.07	22.25	9.E-07	1.19	
Contig174064	AT5G13550	SULTR4;1	sulfate transporter		2.73	0.07	7.E-06	-5.56	X
Contig391186	no hit				0.97	2.57	1.E-05	1.43	
Contig389572	AT4G36740	HB-5, ATHB40, HB40	class-I bZIP-Homeodomain transcription factor	***Male Growth Inhibitor* (*MeGI*)**	26.29	46.07	1.E-05	0.87	
Contig400850	no hit				1.40	3.28	2.E-05	1.36	
Contig229555	no hit				17.67	31.27	2.E-05	0.80	
Contig225969	no hit				2.78	0.06	3.E-05	-5.39	
Contig317713	AT3G12390		Nascent polypeptide-associated		63.39	47.54	4.E-05	-0.37	
Contig282802	no hit				7.33	0.71	5.E-05	-3.28	
Contig336766	no hit				4.62	1.29	7.E-05	-1.76	
Contig187043	no hit				0.58	2.70	1.E-04	2.26	
Contig300341	ATCG00480	ATPB, PB	ATP synthase subunit beta		3.72	5.57	2.E-04	0.57	
Contig188950	no hit				2.37	4.14	2.E-04	0.82	
Contig247290	no hit				0.52	2.59	4.E-04	2.44	
Contig312712	AT1G19250	FMO1	flavin-dependent monooxygenase		8.16	14.76	4.E-04	0.85	
Contig317018	AT3G02460	Ypt/Rab-GAP domain	Transposable element		9.70	6.23	1.E-03	-0.55	
Contig319996	ATCG00740	RPOA	RNA polymerase subunit alpha		31.95	46.71	1.E-03	0.59	
Contig231235	no hit				0.93	3.28	1.E-03	1.85	
Contig230861	no hit				4.58	13.60	1.E-03	1.63	
Contig419223	no hit				2.22	5.91	1.E-03	1.60	
Contig192369	no hit				2.10	4.57	1.E-03	1.29	
Contig298817	no hit				6.29	3.99	1.E-03	-0.63	
Contig196072	no hit				20.85	6.52	2.E-03	-1.51	
Contig190429	AT5G65740		zinc ion binding		3.40	2.28	2.E-03	-0.55	
Contig380632	ATCG00750	RPS11	ribosomal protein S11		12.64	22.39	2.E-03	0.91	
Contig192064	AT3G14270	FAB1B	phosphatidylinositol-4-phosphate 5-kinase family		3.34	6.42	2.E-03	1.09	
Contig286080	no hit				4.52	10.65	2.E-03	1.29	
Contig285237	no hit				18.39	29.00	2.E-03	0.69	
Contig392488	no hit				2.13	0.82	2.E-03	-1.43	
Contig229020	no hit				38.53	50.90	2.E-03	0.46	
Contig283015	no hit				2.96	0.10	3.E-03	-4.56	
Contig255606	no hit				45.63	30.37	3.E-03	-0.63	
Contig201262	no hit				2.16	0.17	4.E-03	-3.38	
Contig363519	no hit				3.26	0.60	4.E-03	-2.55	
Contig319920	no hit				0.83	3.45	5.E-03	2.13	
Contig274121	ATCG01130	YCF1.2	Ycf1 protein		55.94	87.31	5.E-03	0.71	
Contig312848	AT4G24000	ATCSLG2, CSLG2	cellulose synthase		27.07	19.31	6.E-03	-0.40	
Contig265756	ATCG00790	RPL16	ribosomal protein		13.15	19.26	6.E-03	0.58	
Contig115502	AT1G10380		Putative membrane lipoprotein		2.57	0.98	6.E-03	-1.49	
Contig152287	no hit				0.56	3.22	9.E-03	2.36	
Contig107520	AT3G55240		Plant protein 1589 of unknown function		3.55	7.19	1.E-02	1.06	

Table S3. Genetic diversity of the 22 SD candidate genes

Sequence divergence between the X and Y alleles of each gene is indicated. Synonymous (Ks) and non-synonymous (Ka) divergence, evolutionary speed (Ka/Ks), number of segregating sites (*S*), gene length, and substitution ratio (*S*/length) are also indicated.

Annotation	Ks[b]	Ka[b]	Ka/Ks	*S*	length (bp)	*S* /length
X vs Y divergence						
NAM	0.0201	0.0059	0.294	10	1113	0.0090
ARR9	0.0063	0.0071	1.127	5	723	0.0069
ENOD	0.0159	ND	NA	2	525	0.0038
BOP	0.0052	0.0032	0.615	6	1635	0.0037
TLC	0.0050	ND	NA	1	846	0.0012
DNB	0.0053	0.0004	0.075	5	3270	0.0015
PNA	0.0249	0.0131	0.525	31	2655	0.0117
COX	ND	ND	NA	0	912	0.0000
CAT[a]	0.0218	ND	NA	5	967	0.0052
SLTR[a]	0.0228	0.0038	0.166	12	1416	0.0085
RNI[a]	0.0055	ND	NA	2	1547	0.0013
CalSyn[a]	0.0137	0.0013	0.094	4	1001	0.0040
FTR[a]	ND	0.0076	NA	5	866	0.0058
ARM[a]	0.0172	0.0103	0.602	33	3648	0.0090
FE-Contig1245		no ORF annotation[c]		5	1742	0.0029
FE-Contig392		no ORF annotation		14	1726	0.0081
contig42779		no ORF annotation		2	501	0.0040
contig60102		no ORF annotation, weakly repetitive		30	3353	0.0089
contig40011			no ORF annotation, repetitive			
contig60218			no ORF annotation, repetitive			
contig77762			no ORF annnotation, Y-specific[d], repetitive			
OGI			Y-specific			

[a] full length ORF sequences have not been identified.

[b] Jukes and Cantor corrected values are indicated.

[c] no significantly homologous genes in the TAIR/nr databases using BLASTX, and no putative ORF sequences > 500-bp could be identified.

[d] no distinct X allelic counterpart could be found after alignment using BWA, allowing up to 8 nucleotide mismatches per reads.

Table S4. Phenotypic characterization of the 35S-MeGI *A. thaliana* transformed lines

Transformant	Feminization & Developmentally-delayed	Serration	Reduction in stamens[a]	Flowering time (days[b])
35S-MeGI 1-1	−	+	+	36
35S-MeGI 1-2	−	+	−	38
35S-MeGI 1-3	++	+++	ND	56
35S-MeGI 2-1	−	+	−	37
35S-MeGI 2-2	−	−	−	38
35S-MeGI 2-3	−	−	−	38
35S-MeGI 2-4	−	++	+	39
35S-MeGI 2-5	−	+	−	39
35S-MeGI 2-6	−	−	−	42
35S-MeGI 2-7	++	+++	ND	54
35SMeGI m1-1	+++	+++	ND	62
Cont[c]	−	−	−	38.3 ± 1.69

[a] refer to Fig. S14 for statistical significance.

[b] days after germination. For control, average ± SD (N = 3).

[c] transformed by empty vectors.

Table S5. Plant materials

species	origin	cultivar/accession	sexuality	sampling location	annotation	reference
Diospyros lotus	Southwest Asia	Kunsenshi Male	male	Kyoto University, Kyoto, Japan	male parent of the KK F_1 population	
(Caucasian persimmon)		Kunsenshi Female	female	Kyoto University, Kyoto, Japan	female parent of the KK F_1 population	
		L1	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L2	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L3	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L4	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L5	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L6	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L7	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L8	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L9	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L10	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L11	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L12	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L13	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L14	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L15	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L16	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L17	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L18	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L19	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L20	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L21	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L22	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L23	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L24	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L25	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L27	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L28	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L29	female	Kyoto University, Kyoto, Japan	KK F_1 population	Akagi et al. (2014) (*21*)
		L30	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L31	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L32	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L33	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L34	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L35	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L36	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L37	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L38	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L39	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L40	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L41	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L42	male/female[a]	Kyoto University, Kyoto, Japan	KK F_1 population	
		L43	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L46	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L47	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L48	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L49	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L50	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L51	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L52	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L54	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L56	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L57	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L58	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L60	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L61	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		L62	female	Kyoto University, Kyoto, Japan	KK F_1 population	
		L63	male	Kyoto University, Kyoto, Japan	KK F_1 population	
		Mamegaki male	male	NIFTS, Hiroshima, Japan		
		Budougaki	female	Kyoto University, Kyoto, Japan		Akagi et al. (2014) (*21*)
Diospyros oleifera	East Asia	Shirakawagaki	female	Kyoto University, Kyoto, Japan		
(Oil persimmon)						
Diospyros glaucifolia	East Asia	Male no. 10	male	Kyoto University, Kyoto, Japan	Alternatively, *Diospyros japonica*	
(Chekiang persimmon)		Female no. 10	female	Kyoto University, Kyoto, Japan	Alternatively, *Diospyros japonica*	

Diospyros virginiana (American persimmon)	Eastern US	DDIO 69 0003A[b]	male	USDA/ARS, Davis, CA, USA	putative tetraploid or hexaploid	Germplasm Resouces Information Network (GRIN, USDA/ARS) http://www.ars-grin.gov/
		DDIO 69 0001A[b]	female	USDA/ARS, Davis, CA, USA	putative tetraploid or hexaploid	
		Weber (DDIO 51)	female	USDA/ARS, Davis, CA, USA	putative tetraploid or hexaploid	
		Early Golden (DDIO 92)	female	USDA/ARS, Davis, CA, USA	putative tetraploid or hexaploid	
		Meader (DDIO 192)	female	USDA/ARS, Davis, CA, USA	putative tetraploid or hexaploid	
		Mishirazu (DDIO 240)	female	USDA/ARS, Davis, CA, USA	putative tetraploid or hexaploid	
Diospyros rhombifolia (Princess persimmon)	China	Laoya-Shi	female	Kyoto University, Kyoto, Japan	tetraploid	
Diospyros montana	India to South East Asia	MIA9103	female	USDA/ARS, Miami, FL, USA		Germplasm Resouces Information Network (GRIN, USDA/ARS) http://www.ars-grin.gov/
Diospyros digyna	Mexico/Central America	Maher (MIA27138)	female	USDA/ARS, Miami, FL, USA		
		Reineke (MIA26706)	female	USDA/ARS, Miami, FL, USA		
Diospyros mespiliformis	Africa	MIA3483	male	USDA/ARS, Miami, FL, USA		
		MIA1079	male	USDA/ARS, Miami, FL, USA		
Diospyros kaki (Oriental persimmon)	China (East Asia)	Atago	female	Kyoto University, Kyoto, Japan	hexaploid	Fruit Tree Experiment Station of Hiroshima Prefecture (1979) (*44*) Yonemori et al. (1993) (*45*) Akagi et al. (2014) (*21*)
		Fuyu	female	Kyoto University, Kyoto, Japan	hexaploid, occasional male flower	
		Hiratanenashi	female	Kyoto University, Kyoto, Japan	nonaploid	
		Jiro	female	Kyoto University, Kyoto, Japan	hexaploid, occasional male flower	
		Meotogaki	both[c]	Kyoto University, Kyoto, Japan	hexaploid	
		Nishimurawase	both[c]	Kyoto University, Kyoto, Japan	hexaploid	
		Oyotsumizo	female	Kyoto University, Kyoto, Japan	hexaploid	
		Taishu	both[c]	Kyoto University, Kyoto, Japan	hexaploid	
		Taiwanshoshi	both[c]	Kyoto University, Kyoto, Japan	hexaploid	
		Ta-Mo-Pan	female	Kyoto University, Kyoto, Japan	hexaploid	
		Tohachi	both[c]	Kyoto University, Kyoto, Japan	hexaploid	
		Zenjimaru	both[c]	Kyoto University, Kyoto, Japan	hexaploid	

[a] Progeny showing the possibility of sex changes with time. Its sexuality was determined to be male in 2008, and female for the next 5 years in 2009-2014 (*21*). In this study, we considered this individual to be female, and detected female genotypes with no recombination in the regions surrounding the MSY.

[b] Progeny derived from open pollination of a single female parent.

[c] *Diospyros kaki* (oriental persimmon) is a hexaploid species with mostly female or monoecious individuals (*21*). In this case, “both” refers to individuals with male and female flowers (monoecious) while “female” refers to individuals with exclusively female flowers.

Table S6. Index sequences of Illumina adaptors

Individual	Index sequence	Gender	Generation	Library type
First set[a]				
L1	GCCAAT	male	F_1	Genomic
L2	CAGATC	male	F_1	Genomic
L3	CTTGTA	male	F_1	Genomic
L4	ATCACG	male	F_1	Genomic
L5	TTAGGC	male	F_1	Genomic
L6	AGTTCC	male	F_1	Genomic
L7	CGATGT	female	F_1	Genomic
L8	ACTTGA	male	F_1	Genomic
L9	GATCAG	male	F_1	Genomic
L11	ACTGAT	female	F_1	Genomic
L16	ATGAGC	female	F_1	Genomic
L17	CACTCA	female	F_1	Genomic
L18	ACAGTG	female	F_1	Genomic
L20	ATTCCT	female	F_1	Genomic
L21	TAGCTT	male	F_1	Genomic
L22	ATGTCA	male	F_1	Genomic
L23	CAAAAG	female	F_1	Genomic
L24	CAACTA	female	F_1	Genomic
L25	GGCTAC	male	F_1	Genomic
L27	CAGGCG	female	F_1	Genomic
L28	CCGTCC	male	F_1	Genomic
L29	CATGGC	female	F_1	Genomic
L31	GTAGAG	male	F_1	Genomic
L32	GTCCGC	male	F_1	Genomic
L33	CACGAT	female	F_1	Genomic
L34	GTGAAA	male	F_1	Genomic
L35	CATTTT	female	F_1	Genomic
L36	GTGGCC	male	F_1	Genomic
L38	CCAACA	female	F_1	Genomic
L39	CGGAAT	female	F_1	Genomic
L40	GTTTCG	male	F_1	Genomic
L41	CTAGCT	female	F_1	Genomic
L43	CGTACG	male	F_1	Genomic
L46	CTATAC	female	F_1	Genomic
L47	CTCAGA	female	F_1	Genomic
L48	GAGTGG	male	F_1	Genomic
L49	GGTAGC	male	F_1	Genomic
L51	GCGCTA	female	F_1	Genomic
L52	TCATTC	male	F_1	Genomic
L54	TAATCG	female	F_1	Genomic
L56	TACAGC	female	F_1	Genomic
L57	TCCCGA	male	F_1	Genomic
L58	TATAAT	female	F_1	Genomic
L61	TCGAAG	male	F_1	Genomic
L62	CACCGG	female	F_1	Genomic
L63	TCGGCA	male	F_1	Genomic
Second set[b]				
L10	CGATGT	female	F_1	Genomic
L12	TGACCA	male	F_1	Genomic
L13	GCCAAT	female	F_1	Genomic
L14	CAGATC	female	F_1	Genomic
L15	CTTGTA	female	F_1	Genomic

L18 (replicate-1)[c]	ACAGTG	female	F_1	Genomic
L18 (replicate-2)[c]	ATCACG	female	F_1	Genomic
L19	TTAGGC	female	F_1	Genomic
L30	GCGCTA	female	F_1	Genomic
L37	TAATCG	female	F_1	Genomic
L42	TATAAT	male/female[d]	F_1	Genomic
L50	TCATTC	female	F_1	Genomic
L56[c]	TACAGC	female	F_1	Genomic
L60	TCCCGA	female	F_1	Genomic
Third set (RNA-Seq in developing buds)				
L1	TGACCAAT	male	F_1	RNA-Seq
L2	ACAGTGAT	male	F_1	RNA-Seq
L3	GCCAATAT	male	F_1	RNA-Seq
L4	CAGATCAT	male	F_1	RNA-Seq
L5	CTTGTAAT	male	F_1	RNA-Seq
L7	ATGAGCAT	female	F_1	RNA-Seq
L8	ATCACGAT	male	F_1	RNA-Seq
L10	ATTCCTAT	female	F_1	RNA-Seq
L11	CAAAAGAT	female	F_1	RNA-Seq
L13	CAACTAAT	female	F_1	RNA-Seq
L14	CACCGGAT	female	F_1	RNA-Seq
L16	CACGATAT	female	F_1	RNA-Seq
L17	CACTCAAT	female	F_1	RNA-Seq
L18	CAGGCGAT	female	F_1	RNA-Seq
L21	TTAGGCAT	male	F_1	RNA-Seq
L32	GATCAGAT	male	F_1	RNA-Seq
L33	CATGGCAT	female	F_1	RNA-Seq
L43	ACTTGAAT	male	F_1	RNA-Seq
Kunsenshi Female	CATTTTAT	female	Female parent	RNA-Seq
Kunsenshi Male	TAGCTTAT	male	Male parent	RNA-Seq
Fourth set (smRNA-Seq in developing buds)				
Kunsenshi Female	CGATGT	female	Female parent	smRNA-Seq
Kunsenshi Female	TTAGGC	female	Female parent	smRNA-Seq
Kunsenshi Male	ATCACG	male	Male parent	smRNA-Seq
Fifth set smRNA-Seq in developing flowers)				
Kunsenshi Female	TAGCTT	female	Female parent	smRNA-Seq
Kunsenshi Male	GATCAG	male	Male parent	smRNA-Seq

[a] Individuals used for k-mer analysis and genotyping ($N = 46$, Fig. 1A-C).

[b] Individuals used for genotyping. Taken with 1st and 2nd sets, in total 57 individuals were used for genotyping, which called the MSY (Fig. 1D)

[c] Same individual as used in the 1st set.

[d] Progeny showing the possibility of sex changes with time. Its sexuality was determined to be male in 2008, and female for the next 5 years in 2009-2014 (*21*). In this study, we considered this individual to be female, and detected female genotypes with no recombination in the regions surrounding the MSY.

Table S7. Primer sequences

primer	sequences (5'-3')	targeted genes or regions	note
For transformation			
MeGI-LIC26-stF-Gib	CGAGCTAGTTGGAATAGGTTATGACAGCCAACTTTAATCCTCCG	*MeGI* in *D. lotus*	Adapter sequences added to connect to pPLV26 and pPLV2 binary vector (see Method 10)
MeGI-LIC26-spR-Gib	TGCAGTATGGAGTTGGGTTTCATATAAGGTTAACCCATTCCATGCC		
OGI-LIC26-stF-Gib	CGAGCTAGTTGGAATAGGTTCACATATATAAATCATATAAGGTTAACACATTC	*OGI* in *D. lotus*	
OGI-LIC26-spR-Gib	TGCAGTATGGAGTTGGGTTTCCTGGCACACAAAATATTTTCAACCCT		
NatMeGI-LIC2-stF-Gib	GAATTCTAGTTGGAATGGGTTTTGTAATTTCGACCTGCACTCTCTAC	*MeGI* and the surrounding genomic region in *D. lotus*	
NatMeGI-LIC2-spR-Gib	TCCTTATGGAGTTGGGTTTGTGCGAGAGAAGCCTAATGTAATT		
ppa003808m-LIC26-stF	CGAGCTAGTTGGAATAGGTTATGCTCGAAAATTCAAAGTTACCG	ppa003808m in *P. persica*	
ppa003808m-LIC26-spR	TGCAGTATGGAGTTGGGTTTCAGCTCATAATTGCCATTGAAGA		
For PCR in *Diospyros* species			
DlNAC1-R2	AAAAATCCTGCATCTGCCACCTC	*FTR*	Sequences of PCR products were used for the analysis of X/Y allele evolution in a variety of *Diospyros* (see text and Fig. S9 and S12)
DlNAC1-F2	ATATGTGATATGTAAAGTCTTCCAGAAG		
DlRR9-R2	GTTGTGGAGGTTGATCAGGTGA	*RR9*	
DlRR9-F2	AAGGAGGAGCAGAGGAGTTCTTAC		
DlSLTR-F1-rev1	GACGAACAGAAAGAATTATTCAACTCCTC	*SLTR*	
DlSLTR-R1-rev1	GCGCCTTCCAGGCACTACCAT		
DlPNA-F1-rev1	TAAGACATCAGAATCAAGGCCATTTTCCA	*PNA*	
DlPNA-R1-rev1	CTAGATAAGGGAGAAAAACGCCTAAAGCA		
DlFucoT-F1	CATTTTCAGATGTCTTTGATGAAGTTCA	*FTR*	
DlFucoT-R1	CAAAGTTTAAAATGCTGTAGAGATTTCAAAG		
DlARM-F1	GACCAAATGAGGCCGCTCCTAT	*ARM*	
DlARM-R1	GGAGTGGTTGAAGTTGTAAGTTTGAAG		
DlRNI-F1	CAGTTATGTAATCGTCGTGGAGCAT	*RNI*	
DlRNI-R1	GACACAACGCGACTCTTTGCAA		
DlCalSyn-F1	GCCGCATTTTGAAAGCTTCTTC	*CalSyn*	
DlCalSyn-R1	TACTTAAAGGATACAAAGCTGTCAC		
DlCalSyn-F2	GACCTGATTCATATCAATTGTCTGAAG		
DlCalSyn-R2	CCTGAAATACTTAAAGGATACAAAGCTG		
DlDNB-F1	ACTCTAATTGGCAATAAGACAGATTC	*DNB*	
DlDNB-R1	CAGGCAAGGTTTATGCTCTTTAAGA		
DlDNB-F2	CTTTCATCTAGCAGCAACATTCC		
DlDNB-R2	AGTCGGGACTGTCCAATATTCTAC		
DlBOP-F1	TCGATTGACCGATAAACTCCCAT	*BOP*	
DlBOP-R1	GAACAGAGAATGGATCCCAACA		
DlCAT-F2	CAGGAGAATGAGTTCAGCTTGAC	*CAT*	
DlCAT-R2	CAGATATTCTGGAGCAGCTGAT		
DlFE-contig1245-F1	CATGTACAAGAAATCAGGGAGAGA	FE-Contig1245	
DlFE-contig1245-R1	TGCTCCTTTGTAACATTATTGCCCA		
DlDeNovo42779-F1	TCCTTCCCACCCACCCAT	Contig42779	
DlDeNovo42779-R1	TTATGATGATCTCAGCCTCCATGT		
DlDeNovo60218-F1	GGTTCTGCAGGTGGGAATCTT	Contig60218	
DlDeNovo60218-R1	GTTCTTCAATATGTGTCACTCCAAC		
DlDeNovo40011-F1	CAGAGAGTGGCGTTGGCAG	Contig40011	
DlDeNovo40011-R1	CTGCAAACCCATCCCCTCTTCT		
DeNovo40011-F2	GATGATCTGTGTGCCATATTCAGAG		
DeNovo40011-R2	AACCCATCCCCTCTTCCTGTCAA		
DlDeNovo60102-F1	GAAGCAAAAACAAGGATTCCCTGC	Contig60102	
DlDeNovo60102-R1	GCTTGGAGGTGGGGTGTTACAA		
DeNovo60102-F2	ACAAGGATTCCCTGCAAATTGC		
DeNovo60102-R2	ACATATTTCTGAAAATCAGCAGGGTTCG		
DeNovo60102-F3	ACCCAGCCGGAGCTTAATATC		
DeNovo60102-R3	GCATTGGAGCTGAAAAGATGCTTGA		
DeNovo77762-F1	ATACTCCAGTCGAAGCTGTCCG	Contig77762	
DeNovo77762-R1	ATGGTGGCGTAGAGACAGCTTGA		
DeNovo77762-R2	TTATGATGATCTCAGCCTCCATGTAAA		
OGI-candF1	CACAGTAGTCATATATTTTTAGC	*OGI*	OGI specific in a variety of *Diospyros* species [anneal temp. = 50-53℃ (low string.)]
OGI-spR	CTGGCACACAAAATATTTTCAACCCT		
For qPCR analysis and expression test in persimmon organs			
MeGI-ov1stInt-F	GACACCACGGAGAAGTAGTGAT	*MeGI*	Designed to bridge the two introns of *MeGI*, and target fully spliced transcripts only
MeGI-ov2ndInt-R	GTTCTTTGAGCTTTAGCTCCGTTTC		
OGI-RT-F1	GAGCTTTAGCACTTATCCAACACAGTAGTC	*OGI*	
OGI-RT-R1	GATGAGAGATGAGATCCATTTAATGATCT		
DkActin-F	CATGGAGAAAATCTGGCATCATAC	*Actin* (AB473616)	High expression in all organs tested (*46*)
DkActin-R	GAAGCACTGGGTGCTCTTCTG		

References

1. B. Charlesworth, The evolution of chromosomal sex determination and dosage compensation. *Curr. Biol.* **6**, 149–162 (1996). Medline doi:10.1016/S0960-9822(02)00448-7

2. D. Bachtrog, A dynamic view of sex chromosome evolution. *Curr. Opin. Genet. Dev.* **16**, 578–585 (2006). Medline doi:10.1016/j.gde.2006.10.007

3. D. Bachtrog, Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013). Medline doi:10.1038/nrg3366

4. C. Yampolsky, H. Yampolsky, Distribution of the sex forms in the phanerogamic flora. *Bibliotheca Genetica* **3**, 1–62 (1922).

5. R. Ming, J. Wang, P. H. Moore, A. H. Paterson, Sex chromosomes in flowering plants. *Am. J. Bot.* **94**, 141–150 (2007). Medline doi:10.3732/ajb.94.2.141

6. C. Ainsworth, Boys and girls come out to play: The molecular biology of dioecious plants. *Ann. Bot.* **86**, 211–221 (2000). doi:10.1006/anbo.2000.1201

7. R. Ming, A. Bendahmane, S. S. Renner, Sex chromosomes in land plants. *Annu. Rev. Plant Biol.* **62**, 485–514 (2011). Medline doi:10.1146/annurev-arplant-042110-103914

8. D. Charlesworth, Plant sex chromosome evolution. *J. Exp. Bot.* **64**, 405–420 (2013). Medline doi:10.1093/jxb/ers322

9. R. Bergero, S. Qiu, A. Forrest, H. Borthwick, D. Charlesworth, Expansion of the pseudo-autosomal region and ongoing recombination suppression in the *Silene latifolia* sex chromosomes. *Genetics* **194**, 673–686 (2013). Medline doi:10.1534/genetics.113.150755

10. N. Blavet, H. Blavet, R. Cegan, N. Zemp, J. Zdanska, B. Janoušek, R. Hobza, A. Widmer, Comparative analysis of a plant pseudoautosomal region (PAR) in *Silene latifolia* with the corresponding *S. vulgaris* autosome. *BMC Genomics* **13**, 226 (2012). 10.1186/1471-2164-13-226 Medline doi:10.1186/1471-2164-13-226

11. E. Kejnovsky, B. Vyskot, *Silene latifolia*: The classical model to study heteromorphic sex chromosomes. *Cytogenet. Genome Res.* **129**, 250–262 (2010). Medline doi:10.1159/000314285

12. Z. Liu, P. H. Moore, H. Ma, C. M. Ackerman, M. Ragiba, Q. Yu, H. M. Pearl, M. S. Kim, J. W. Charlton, J. I. Stiles, F. T. Zee, A. H. Paterson, R. Ming, A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature* **427**, 348–352 (2004). Medline doi:10.1038/nature02228

13. A. R. Gschwend, Q. Yu, E. J. Tong, F. Zeng, J. Han, R. VanBuren, R. Aryal, D. Charlesworth, P. H. Moore, A. H. Paterson, R. Ming, Rapid divergence and expansion of the X chromosome in papaya. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13716–13721 (2012). Medline doi:10.1073/pnas.1121096109

14. J. Wang, J.-K. Na, Q. Yu, A. R. Gschwend, J. Han, F. Zeng, R. Aryal, R. VanBuren, J. E. Murray, W. Zhang, R. Navajas-Perez, F. A. Feltus, C. Lemke, E. J. Tong, C. Chen, C. Man Wai, R. Singh, M.-L. Wang, X. J. Min, M. Alam, D. Charlesworth, P. H. Moore, J. Jiang, A. H. Paterson, R. Ming, Sequencing papaya X and Y^h chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13710–13715 (2012). Medline doi:10.1073/pnas.1207833109

15. R. Aryal, R. Ming, Sex determination in flowering plants: Papaya as a model system. *Plant Sci.* **217-218**, 56–62 (2014). Medline doi:10.1016/j.plantsci.2013.10.018

16. E. Cherif, S. Zehdi, K. Castillo, N. Chabrillange, S. Abdoulkader, J. C. Pintaud, S. Santoni, A. Salhi-Hannachi, S. Glémin, F. Aberlenc-Bertossi, Male-specific DNA markers provide genetic evidence of an XY chromosome system, a recombination arrest and allow the tracing of paternal lineages in date palm. *New Phytol.* **197**, 409–415 (2013). Medline doi:10.1111/nph.12069

17. B. Charlesworth, D. Charlesworth, A model for the evolution of dioecy and gynodioecy. *Am. Nat.* **112**, 975–997 (1978). doi:10.1086/283342

18. S. Duangjai, B. Wallnöfer, R. Samuel, J. Munzinger, M. W. Chase, Generic delimitation and relationships in Ebenaceae sensu lato: Evidence from six plastid DNA regions. *Am. J. Bot.* **93**, 1808–1827 (2006). Medline doi:10.3732/ajb.93.12.1808

19. D. C. Christophel, J. F. Basinger, Earliest floral evidence for the Ebenaceae in Australia. *Nature* **296**, 439–441 (1982). doi:10.1038/296439a0

20. B. Turner, J. Munzinger, S. Duangjai, E. M. Temsch, R. Stockenhuber, M. H. Barfuss, M. W. Chase, R. Samuel, Molecular phylogenetics of New Caledonian *Diospyros* (Ebenaceae) using plastid and nuclear markers. *Mol. Phylogenet. Evol.* **69**, 740–763 (2013). Medline doi:10.1016/j.ympev.2013.07.002

21. T. Akagi, K. Kajita, T. Kibe, H. Morimura, T. Tsujimoto, S. Nishiyama, T. Kawai, H. Yamane, R. Tao, Development of molecular markers associated with sexuality in *Diospyros lotus* L. and their application in *D. kaki* Thunb. *J. Jpn. Soc. Hortic. Sci.* **83**, 214–221 (2014). doi:10.2503/jjshs1.CH-109

22. T. Komatsuda, M. Pourkheirandish, C. He, P. Azhaguvel, H. Kanamori, D. Perovic, N. Stein, A. Graner, T. Wicker, A. Tagiri, U. Lundqvist, T. Fujimura, M. Matsuoka, T. Matsumoto, M. Yano, Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1424–1429 (2007). Medline doi:10.1073/pnas.0608580104

23. S. Sakuma, M. Pourkheirandish, G. Hensel, J. Kumlehn, N. Stein, A. Tagiri, N. Yamaji, J. F. Ma, H. Sassa, T. Koba, T. Komatsuda, Divergence of expression pattern contributed to neofunctionalization of duplicated HD-Zip I transcription factor in barley. *New Phytol.* **197**, 939–948 (2013). Medline doi:10.1111/nph.12068

24. A. E. Quinn, S. D. Sarre, T. Ezaz, J. A. Marshall Graves, A. Georges, Evolutionary transitions between mechanisms of sex determination in vertebrates. *Biol. Lett.* **7**, 443–448 (2011). Medline doi:10.1098/rsbl.2010.1126

25. T. Kiuchi, H. Koga, M. Kawamoto, K. Shoji, H. Sakai, Y. Arai, G. Ishihara, S. Kawaoka, S. Sugano, T. Shimada, Y. Suzuki, M. G. Suzuki, S. Katsuma, A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature* **509**, 633–636 (2014). Medline doi:10.1038/nature13315

26. D. Burkart-Waco, K. Ngo, B. Dilkes, C. Josefsson, L. Comai, Early disruption of maternal-zygotic interaction and activation of defense-like responses in *Arabidopsis* interspecific crosses. *Plant Cell* **25**, 2037–2055 (2013). Medline doi:10.1105/tpc.112.108258

27. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). Medline doi:10.1093/bioinformatics/btp324

28. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). Medline doi:10.1093/bioinformatics/btp352

29. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011). Medline doi:10.1038/nbt.1754

30. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011). Medline doi:10.1038/nbt.1883

31. X. Huang, A. Madan, CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999). Medline doi:10.1101/gr.9.9.868

32. J. G. Ruby, P. Bellare, J. L. DeRisi, PRICE: Software for the targeted assembly of components of (meta) genomic sequence data. G3 3, 865–880 (2013).

33. S. Anders, W. Huber, Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010). 10.1186/gb-2010-11-10-r106 Medline doi:10.1186/gb-2010-11-10-r106

34. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). Medline doi:10.1093/molbev/mst010

35. P. Librado, J. Rozas, DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452 (2009). Medline doi:10.1093/bioinformatics/btp187

36. M. Gouy, S. Guindon, O. Gascuel, SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010). Medline doi:10.1093/molbev/msp259

37. S. Duangjai, R. Samuel, J. Munzinger, F. Forest, B. Wallnöfer, M. H. Barfuss, G. Fischer, M. W. Chase, A multi-locus plastid phylogenetic analysis of the pantropical genus *Diospyros* (Ebenaceae), with an emphasis on the radiation and biogeographic origins of the New Caledonian endemic species. *Mol. Phylogenet. Evol.* **52**, 602–620 (2009). Medline doi:10.1016/j.ympev.2009.04.021

38. M. A. Beilstein, N. S. Nagalingum, M. D. Clements, S. R. Manchester, S. Mathews, Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 18724–18728 (2010). Medline doi:10.1073/pnas.0909766107

39. S. Sakuma, M. Pourkheirandish, T. Matsumoto, T. Koba, T. Komatsuda, Duplication of a well-conserved homeodomain-leucine zipper transcription factor gene in barley generates a copy with more specific functions. *Funct. Integr. Genomics* **10**, 123–133 (2010). Medline doi:10.1007/s10142-009-0134-y

40. B. De Rybel, W. van den Berg, A. S. Lokerse, C.-Y. Liao, H. van Mourik, B. Moller, C. I. Llavata-Peris, D. Weijers, A versatile set of ligation-independent cloning vectors for functional studies in plants. *Plant Physiol.* **156**, 1292–1299 (2011). Medline doi:10.1104/pp.111.177337

41. X. Zhang, R. Henriques, S. S. Lin, Q. W. Niu, N. H. Chua, *Agrobacterium*-mediated transformation of *Arabidopsis thaliana* using the floral dip method. *Nat. Protoc.* **1**, 641–646 (2006). Medline doi:10.1038/nprot.2006.97

42. M. P. Alexander, Differential staining of aborted and nonaborted pollen. *Biotech. Histochem.* **44**, 117–122 (1969). Medline doi:10.3109/10520296909063335

43. R. B. Horsch, J. E. Fry, N. L. Hoffmann, D. Eichholtz, S. G. Rogers, R. T. Fraley, A simple and general method for transferring genes into plants. *Science* **227**, 1229–1231 (1985). Medline doi:10.1126/science.227.4691.1229

44. Fruit Tree Experiment Station of Hiroshima Prefecture, *Showa 53-nendo Shubyo-tokusei-bunrui-chosa-hokokusho (Kaki)* (Fruit Tree Experiment Station of Hiroshima Prefecture, Akitsu, Hiroshima, Japan, 1979).

45. K. Yonemori, A. Sugiura, K. Tanaka, K. Kameda, Floral ontogeny and sex determination in monoecious-type persimmons. *J. Am. Soc. Hortic. Sci.* **118**, 293–297 (1993).

46. T. Akagi, A. Ikegami, T. Tsujimoto, S. Kobayashi, A. Sato, A. Kono, K. Yonemori, DkMyb4 is a Myb transcription factor involved in proanthocyanidin biosynthesis in persimmon fruit. *Plant Physiol.* **151**, 2028–2045 (2009). Medline doi:10.1104/pp.109.146985