

RESOURCE

Integrated genome sequence and linkage map of physic nut (*Jatropha curcas* L.), a biodiesel plant

Pingzhi Wu^{1,†}, Changpin Zhou^{1,2,†}, Shifeng Cheng^{3,†}, Zhenying Wu^{1,2}, Wenjia Lu^{1,2}, Jinli Han^{1,2}, Yanbo Chen^{1,2}, Yan Chen³, Peixiang Ni³, Ying Wang³, Xun Xu³, Ying Huang³, Chi Song³, Zhiwen Wang³, Nan Shi³, Xudong Zhang³, Xiaohua Fang⁴, Qing Yang⁵, Huawu Jiang¹, Yaping Chen¹, Meiru Li¹, Ying Wang¹, Fan Chen^{4,*}, Jun Wang^{3,6,*} and Guojiang Wu^{1,*}

¹Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China,

²University of Chinese Academy of Sciences, Beijing 100049, China,

³BGI-Shenzhen, Shenzhen 518083, China,

⁴National Center for Plant Gene Research, State Key Laboratory of Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China,

⁵Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Mengla, Yunnan Province 666303, China, and

⁶Department of Biology, University of Copenhagen, Copenhagen, Denmark

Received 26 August 2014; revised 18 December 2014; accepted 6 January 2015; published online 20 January 2015.

*For correspondence (e-mails fchen@genetics.ac.cn; wangj@genomics.org.cn; wugj@srbg.ac.cn).

†These authors contributed equally to this work.

SUMMARY

The family Euphorbiaceae includes some of the most efficient biomass accumulators. Whole genome sequencing and the development of genetic maps of these species are important components in molecular breeding and genetic improvement. Here we report the draft genome of physic nut (*Jatropha curcas* L.), a biodiesel plant. The assembled genome has a total length of 320.5 Mbp and contains 27 172 putative protein-coding genes. We established a linkage map containing 1208 markers and anchored the genome assembly (81.7%) to this map to produce 11 pseudochromosomes. After gene family clustering, 15 268 families were identified, of which 13 887 existed in the castor bean genome. Analysis of the genome highlighted specific expansion and contraction of a number of gene families during the evolution of this species, including the ribosome-inactivating proteins and oil biosynthesis pathway enzymes. The genomic sequence and linkage map provide a valuable resource not only for fundamental and applied research on physic nut but also for evolutionary and comparative genomics analysis, particularly in the Euphorbiaceae.

Keywords: physic nut, *Jatropha curcas* L., genome sequence, linkage map, genome evolution, gene family, oil biosynthesis, starch biosynthesis, toxicant synthesis.

INTRODUCTION

The family Euphorbiaceae includes some of the most efficient biomass accumulators, providing oil [physic nut (*Jatropha curcas* L.), castor bean (*Ricinus communis* L.) and *Plukenetia volubilis* L.], starch (cassava, *Manihot esculenta* Crantz) and rubber (*Hevea brasiliensis*). Crop improvement in Euphorbiaceae for sustainable industrial raw materials and food production requires a substantial increase in research on the various species. Whole genome sequencing and the development of genetic maps of these species are important components in molecular breeding and genetic improvement.

Renewable biofuel has received increasing attention worldwide as a substitute for fossil fuels. Among several biofuel plants, physic nut has been identified as a plant with high potential for biofuel plantations, because of its high seed oil content, easy propagation, rapid growth, short gestation period and adaptation to a wide range of agro-climatic conditions (Kumar and Sharma, 2008). The physic nut plant is a diploid species with a $2n$ chromosome number of 22 (Dehgan, 1984). A recent study has estimated the genome size (1C) to be 416 Mbp (Carvalho

et al., 2008). This small genome makes physic nut an attractive model for the functional genomics of Euphorbiaceae. Generation of primary genome information for the physic nut will provide information on factors controlling oil yield, abiotic stress tolerance and toxicity. The previous reported genome sequence of physic nut was 285.9 Mbp and contained 21 225 unigenes (Sato *et al.*, 2011). For the upgraded physic nut genome sequence, the mean and N50 scaffold lengths are 7.6 and 16.0 kbp, respectively (Hirakawa *et al.*, 2012). Linkage maps for physic nut have also been established (Wang *et al.*, 2011; King *et al.*, 2013). The unique loci of the genetic map established by King *et al.* (2013) is 399, and only 17 Mbp (6%) of the reported assembly could be anchored onto the linkage map. Nevertheless, these data have provided valuable information for establishing a high-density genetic map of physic nut.

Here, we report a more complete physic nut genome assembly. We first generated whole genome shotgun (WGS) paired-end-tag sequence reads from the genomic DNA and an expressed sequence tag (EST) database for leaf, root and seed, and then built a *de novo* assembly. Then we constructed the EST database of leaf, root, stem and flower obtained from pyrosequencing of cDNA from *Jatropha integerrima* to develop simple sequence repeat (SSR), insertion–deletion (InDel), and single nucleotide polymorphism (SNP) markers, and then established the interspecific genetic map using the backcross 1 (BC1) population of physic nut across *J. integerrima*. Next, we anchored the scaffolds of the assembly to this genetic map to obtain information on the complete genome for physic nut. In addition, we examine genomic differences between physic nut and castor bean, two Euphorbiaceae species. The genome analyses presented in this study provide a rich resource of genetic information for physic nut breeding and genetic improvement.

RESULTS AND DISCUSSION

Genome sequencing and annotation

We selected the inbred cultivar GZQX0401 of the physic nut for our genome sequencing project. Libraries with a range of different insertion sizes were constructed in accordance with the recommendations of Li *et al.* (2010). A total of 102.68 Gbp were generated (Table S1 in Supporting Information). After excluding low-quality reads, 78.75 Gbp (189-fold genome coverage) were assembled into contigs and scaffolds using SOAPdenovo (Figure S1, Table S2). The contigs contain 264.4 Mbp with an N50 of 26 746 bp and the scaffolds contain 320.5 Mbp, including embedded gaps, with an N50 of 0.746 Mbp (Tables 1 and S3). The N90 number of scaffold is 653 (Table S3). The mean GC content of this genome is 33.3%, which is close to the genome of castor bean (33.8%) (Table 1, Figure S2).

Table 1 Genome assembly and annotation statistics for the genome sequence of physic nut

Total size of contig (Mbp)	264.366
Longest contig (kbp)	396.144
N50 length of contig (kbp)	26.746
N50 number of contigs	2333
Total number of contigs (>2 kbp)	15 438
Mean contig length (>2 kbp) (kbp)	5.145
Total number of contigs (>100 bp)	72 474
Mean contig length (>100 bp) (kbp)	3.647
Total size of scaffolds (Mbp)	320.546
Total size of scaffolds (>2 kb) (Mbp)	316.420
Longest scaffold (Mbp)	5.289
N50 length of scaffolds (Mbp)	0.746
N50 number of scaffold	97
Total number of scaffolds (>2 kb)	1884
Mean scaffold length (>2 kbp)	167.951
Total number of scaffolds (>100 bp)	23 125
Mean scaffold length (>100 bp) (kbp)	13.861
GC content in genome	33.32%
Protein gene models	27 172
GC content in genes	36.10%
Mean length of mRNA (bp)	3247.35
GC content in mRNA	36.44%
Mean length of coding region (bp)	1097.28
GC content in coding region	43.03%
Mean number of introns	3.80
Mean length of intron (bp)	486.08
GC content in introns	32.43%
Mean number of exons	4.80
Mean length of exons (bp)	291.15
GC content in exons	41.51%
Mean length of intergenic regions (bp)	16111.34
GC content in intergenic regions	29.84%

The size of the physic nut genome, estimated by flow cytometric analysis of isolated nuclei stained with propidium iodide, is about 416 Mbp (Carvalho *et al.*, 2008). A large proportion of the heterochromatic DNA, such as centromeric and pericentromeric DNA, was probably not assembled by the WGS process, but more than 94% of the available EST-cDNA data for physic nut and about 87% of that for *J. integerrima* was covered by the assembled genome (Table S4). Thus, the assembled sequences are likely to represent about 77% of the physic nut genome as a whole and more than 94% of the euchromatic regions. To assess the quality of the genome assembly, we validated it in three ways. First, to validate the scaffolding, we remapped the paired-end (PE) clean reads generated in this project onto the 73 largest assembled scaffolds. More than 95% of the selected scaffolds were strongly supported by PE read relationships (Figure S3). Second, we aligned 17 bacterial artificial chromosomes (BACs) derived from the physic nut (DDBJ AP011961–AP011977) against our scaffolds. Most of these BACs aligned well with the scaffolds generated in this work, exhibiting coverage of over 99% (Figure S4); some collinear blocks contained ‘gaps’ indicative of ambiguous ‘Ns regions’ or ‘lost regions’ in the sequence data, but the order

of the connected contigs was generally correct. Third, we used the LASTZ program to perform whole genome scaffold-to-scaffold alignment of our physic nut genome assembly (632 scaffolds were covered) with that presented by Hirakawa *et al.* (2012). The two genomes exhibited good consistency (Figure S5). The size of the assembled genome generated in this work is similar to that reported by Hirakawa *et al.* (2012) (298 Mbp). However, the mean scaffold length (>100 bp) and the N50 in this study are 13 861 bp and 0.746 Mbp, respectively, which are approximately 1.8 and 46.8 times longer, respectively, than the reported physic nut genome (Hirakawa *et al.*, 2012). As the result of short contigs and scaffolds, too many predicted genes in their study were fragmented. Only 17 Mbp (6%) of their assembly (but about 50% of our assembly) could be anchored onto the linkage map constructed by King *et al.* (2013).

In the plants studied to date, most genomes contain highly repetitive regions with transposable elements interspersed between unique gene-containing regions. The identified highly repetitive DNA in the castor bean genome is about 50.3% (Chan *et al.*, 2010). About 49.8% of this physic nut draft genome was found to consist of repetitive DNA sequences, of which 92.3% are transposable elements (Table S5). We used three gene-prediction methods (*de novo*, homolog and cDNA-EST) to identify protein-coding genes, and then used GLEAN to assemble the results (Table S6). After combining the results of RNA-seq and GLEAN, 27 172 putative protein-coding genes were identified, with a mean coding sequence length of 1097 bp (Tables 1 and S7). On the basis of pair-wise protein sequence similarities we carried out a gene family clustering analysis of all genes from several plants for which a genome sequence was available, using rice as an outgroup (The Arabidopsis Genome Initiative, 2000; Goff *et al.*, 2002; Yu *et al.*, 2002; Tuskan *et al.*, 2006; Jaillon *et al.*, 2007; Ming *et al.*, 2008; Chan *et al.*, 2010; Argout *et al.*, 2011; Shulaev *et al.*, 2011). A total of 7406 gene families were revealed distributed among all nine plants, and 8378 among all eudicots. The physic nut genes consist of 15 268 families, 496 of which are unique to physic nut (Table S8); about 80.4% of the genes either have known homologs or can be functionally classified (Table S9).

Anchoring genome sequences to the genetic map

To build a linkage map for physic nut, we used a BC1 population (*J. curcas* × *J. integerrima*) of 190 individuals created from physic nut GZQX0401 and *J. integerrima* GZBN1, using 1208 SNP, InDel and SSR markers (Data S1a). The total genetic distance covered by this *Jatropha* linkage map was 1655.8 cM, with an average marker density of 1.4 and 2.1 cM for all loci and unique loci, respectively (Figure S6, Table S10). The number of unique loci is 801, which is 402 more than on the map reported by King *et al.* (2013). The order of loci within each of the 11 linkage groups was very

consistent between the interspecific map in this study and the intraspecific map reported by King *et al.* (2013) (Data S1b). We then aligned and oriented the scaffolds of the sequence assembly with the *Jatropha* linkage map (Figure 1). A total of 480 scaffolds were anchored to the map. These scaffolds contain 261.8 Mbp of sequence data (about 81.7% of the total scaffold sequences). The average ratio of genetic-to-physical distance was 1 cM per 251 kbp (calculated assuming the genome size to be 416 Mbp). We assembled the scaffolds into 11 pseudochromosomes, numbered according to the linkage group nomenclature used previously (Wang *et al.*, 2011; King *et al.*, 2013) (Figure 1).

Comparative mapping of markers derived from the physic nut genome sequence against the castor bean draft genomic sequences was conducted. A total of 410 scaffolds from the castor bean draft genome sequences were anchored to the *Jatropha* map. These scaffolds contain 189.8 Mb of sequence data (about 54.1% of the total scaffold sequence) (Data S1b). Given the similar gene arrangements between physic nut and castor bean genomes, we defined 320 well-conserved syntenic blocks that contained 10 760 physic nut genes collinear to castor bean among these scaffolds. The mean block length in physic nut is 380 294 kbp. The mean density of collinear genes is 89.4 and 88.4 in physic nut and castor bean genomes, respectively. These syntenic blocks could serve as a useful resource for the future identification of genes by homology, and assist in the creation of syntenic genetic maps in physic nut and castor bean.

Genome evolution

By examining 1970 single-copy gene families from nine sequenced plant genomes we found that physic nut, castor bean and *Populus trichocarpa* belong to a common subclade and probably diverged approximately 65.6 million years ago (Figures 2a and S7). Physic nut and castor bean diverged from a common ancestor approximately 49.4 million years ago (Figure S7).

Angiosperms have been shown to have evolved through rounds of paleo-polyploidy (Abrouk *et al.*, 2010). An ancient hexaploidization event is common to the dicotyledonous lineage (Jaillon *et al.*, 2007; Argout *et al.*, 2011). In order to investigate the paleo-history of the physic nut genome, we characterized paralogous relationships identified between the 11 physic nut chromosomes, as well as shared paleo-polyploidies based on the orthologous relationships identified among physic nut, cacao and grape. Self-alignment of the physic nut genome sequences on the basis of the 18 927 gene models located in the 11 pseudochromosomes identified 1045 paralogous gene groups. The mean synonymous substitution rate (K_s) in the duplicated genes ($K_s = 1.56$) suggested that an ancient whole-genome duplication occurred in physic nut (Figure S8). Seven blocks of anchored paralogous genes were identified and characterized in physic nut (Figure 2b). A

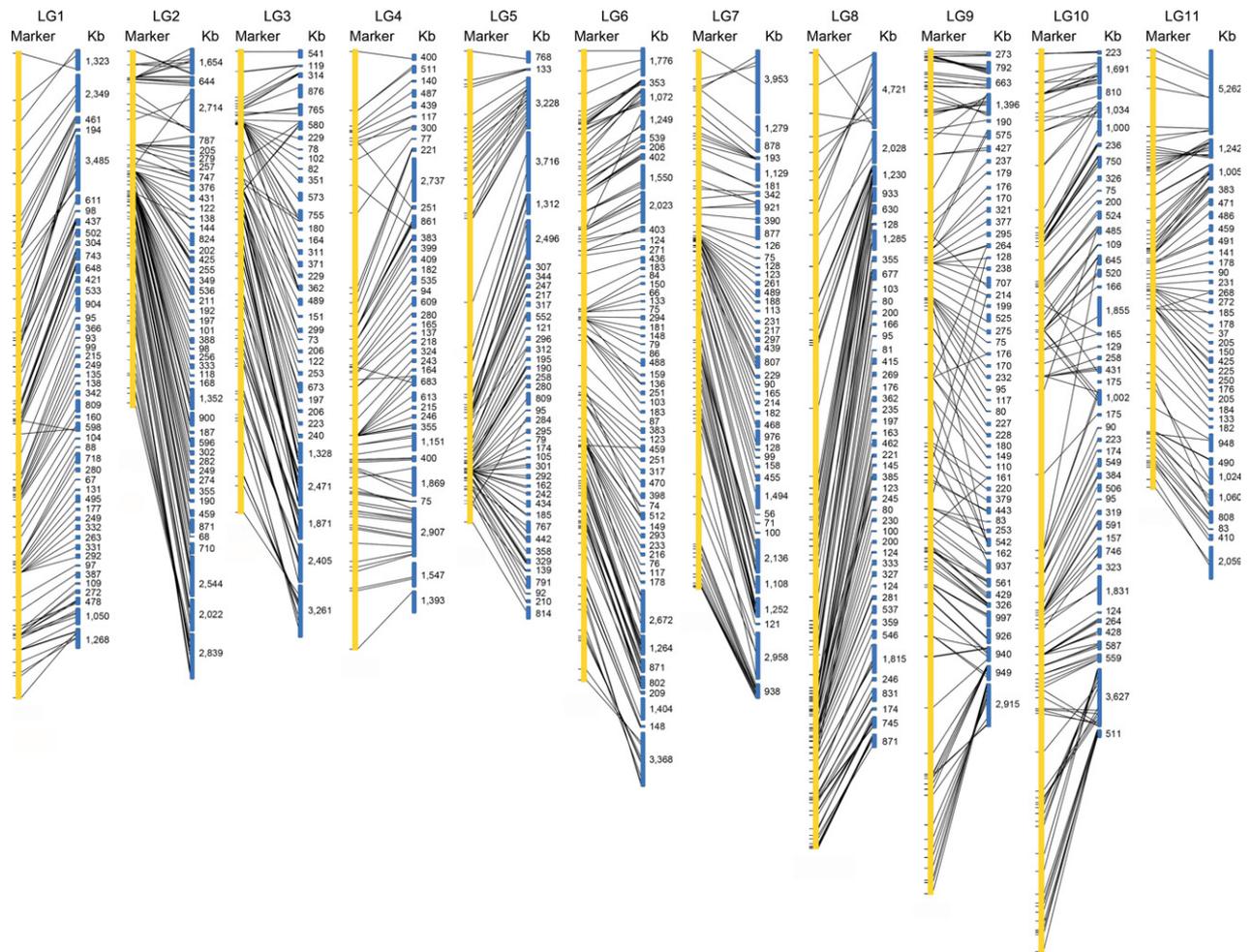


Figure 1. Anchoring the genome sequence assembly to the genetic map of *Jatropha* (*J. curcas* × *J. integerrima*).

Assembled scaffolds (blue) representing 261.8 Mb of genome sequence including embedded gaps (81.7% of the assembled genome sequence) were anchored to the 11 linkage groups (LG1–LG11, yellow) using 1208 genetic markers.

one-to-one orthologous relationship predominated in the physic nut-to-cacao and physic nut-to-grape comparisons, which were carried out using 7305 orthologous relationships which covered 76% of the physic nut genome. We found that the ancestral paleo-polyploidy event shared at orthologous positions on chromosome pair combinations in physic nut was comparable to the seven ancestral triplicated chromosome groups reported in grape (Figures 2c and S9). On the basis of cross-species synteny block boundaries, we estimated that at least 25 chromosome translocation and fusion events have occurred in the 11 physic nut chromosomes during their evolution from the chromosomes of the paleo-hexaploid ancestor that is believed to be common to all eudicots whose genomes have been sequenced to date (Figure 2d).

Gene expansion and neofunctionalization in physic nut

Physic nut and castor bean share a total of 13 887 gene families, which is 91.0 and 88.7% of the gene families in

physic nut and castor bean, respectively. A total of 1082 gene families were lost in the physic nut genome which remain in castor bean and other plant species; the equivalent number for castor bean is 885. After statistical analysis of the evolution of the size of gene families, we detected 149 and 171 families expanded and contracted, respectively, in both physic nut and castor bean, after they diverged from grape. Further, 285 and 295 families in physic nut and 221 and 388 in castor bean, respectively, specifically expanded and contracted after the divergence of physic nut and castor bean (Figure S10).

Tandem gene duplication is one of the major mechanisms of duplication in eukaryotes. We detected 3839 tandem gene duplications among 1442 loci in the physic nut genome. The longest tandem gene array consists of 15 cytochrome P450 (CYP) genes (Table S11). Proteins with kinase active site domains, disease resistance gene products, UDP-glucuronosyl-/glucosyltransferase domain-containing proteins and short-chain dehydrogenase/reductases were among the

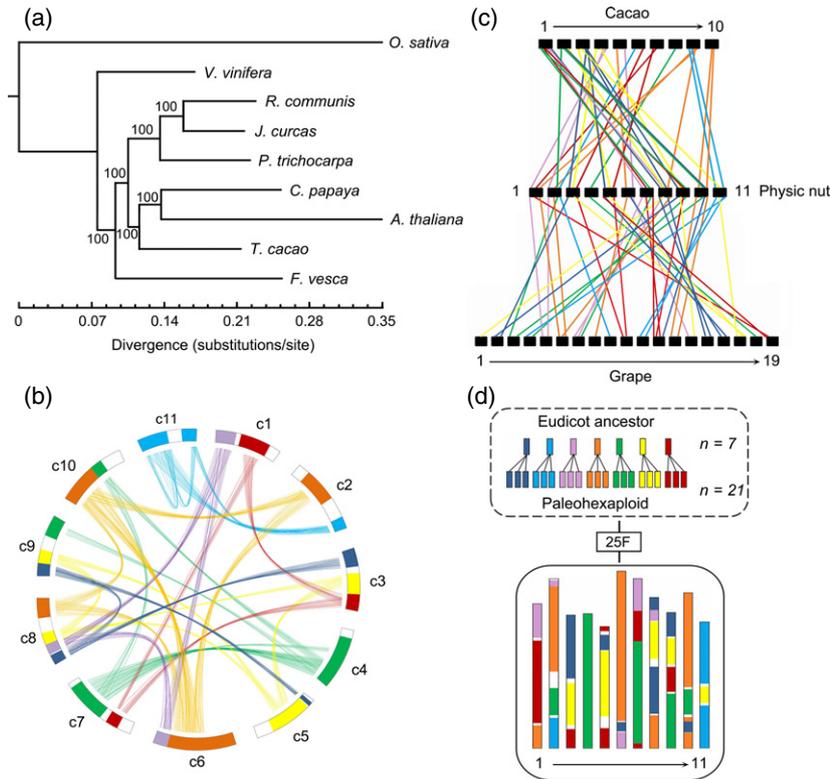


Figure 2. Evolutionary analysis of the physic nut genome. (a) Physic nut phylogeny on the basis of 1970 single-copy genes shared between the nuclear genomes of Arabidopsis, rice, cacao, poplar, grape, strawberry, papaya and castor bean.

(b) Schematic representation of major interchromosomal relationships within 1045 paralogous gene groups in the physic nut genome. Syntenic blocks derived from the seven ancestral protochromosomes are color-coded as indicated in (d).

(c) Physic nut genome synteny. A schematic representation of the orthologs identified between physic nut pseudo-chromosomes (1–11), shown in the center, and the grape (1–19) (lower) and cacao (1–10) (upper) chromosomes. The seven different colors used to represent the blocks reflect their origins from the seven ancestral eudicot linkage groups.

(d) Evolutionary model for the physic nut genome. The eudicot chromosomes are represented using a seven-color code to illustrate the evolution of segments from a common ancestor with seven protochromosomes (top). The shuffling events that have shaped the structure of this genome during its evolution from the common paleo-hexaploid ancestor are indicated by F (fusion of chromosomes).

proteins and domains found most frequently in tandemly repeated genes in the physic nut genome.

The complement of cell wall biosynthesis genes in the physic nut genome is similar to that of Arabidopsis (Kanehisa and Goto, 2000), but whereas Arabidopsis has only two flavonoid *O*-methyltransferase genes, physic nut has 17 (Table S12). Physic nut (433) has fewer putative cell-wall degradation genes than Arabidopsis (459), but more than castor bean (407) (Table S13). The CYPs are a large and diverse group of enzymes that catalyze the oxidation of organic substances (Nelson *et al.*, 2004). Our draft of the physic nut genome reveals at least 319 putative CYP genes, including potential pseudogenes, 62 more than are found in castor bean (257) (Data S2). The nucleotide-binding site (NBS) R genes play important roles in resistance to pathogens and in the cell cycle (DeYoung and Innes, 2006). A total of 172 putative NBS-containing genes were identified in physic nut (Table S14) – more than the 121 found in the castor bean genome (Chan *et al.*, 2010) but fewer than the 207 of Arabidopsis (Meyers *et al.*, 2003).

On the basis of the RNA-seq data, we analyzed the divergence in functional expression of ancient duplicate gene pairs (those arising from the hexaploidization event in dicotyledons) and tandemly duplicated genes. There was a significant reduction in expression of about 7.9, 11.6 and 8.8% of ancient duplicated genes in root, leaf and seed, respectively (Figure S11). These results support the hypothesis that there has been a divergence in functional expression between members of some duplicated gene pairs in physic nut. These ratios were higher among the tandem duplicators (9.9, 13.5 and 11.9 in root, leaf and seed, respectively) (Figure S11). This result might be due to the pseudogenization of some tandemly duplicated genes in physic nut.

Analysis of transcription factor families and genes highly expressed in the early developing stage of seeds

Comparisons with transcription factor (TF) protein models from grape and Arabidopsis (<http://plntfdb.bio.uni-potsdam.de/v3.0/>) using stringent BLAST criteria revealed 1551

TFs in the physic nut genome (Table S15). Many TF families in the physic nut genome, such as the WRKY gene family (Xiong *et al.*, 2013), have not undergone any recent gene duplication. Nevertheless, our draft genome demonstrated the retention of some ancient and recent duplication in various TF subfamilies, including some within the GRAS and MYB families.

Investigation of seed development and size regulation is important for oilseed plants. In physic nut (Jiang *et al.*, 2012) and many other dicots seed development is marked by two main stages. The first stage (the first 27 days after pollination, DAP) is associated with histodifferentiation, rapid increases in seed volume and seed coat dry matter. The second stage is associated with a rapid increase in kernel dry mass. We have previously analyzed global gene expression in developing physic nut seeds 14, 19, 25, 29, 35, 41 and 45 DAP, generating seven different global gene expression profiles labeled S1–S7, respectively (Jiang *et al.*, 2012). It was suggested that accumulation of oil and seed storage protein in the endosperm of physic nut seeds is controlled by a regulatory TF network similar to that found in the embryos of *Arabidopsis* seeds (Jiang *et al.*, 2012). In this work, we endeavored to specifically identify TF genes that are strongly expressed during the first stage of seed development. To this end, we focused on genes whose average number of transcripts per million tags (TPMs) in S1, S2 and S3 were more than three and five times higher than their average TPMs in S5, S6, and S7. A total of 215 TF genes satisfying these criteria were identified, including orthologs of TFs that are known to play important roles in seed development and size regulation in *Arabidopsis* (Sun *et al.*, 2010), such as *AGL11/STK* (JCGZ_07991), *AGL32/TT16* (JCGZ_00038, JCGZ_05961 and JCGZ_05951), *AGL62* (JCGZ_04579), *AGL80/FEM111* (JCGZ_19811, JCGZ_08738 and JCGZ_23162), *APETALA2* (JCGZ_12430), *MINI3* (JCGZ_19224), *ANT* (JCGZ_12574 and JCGZ_15986), *SEPAL-LATA 3* (JCGZ_17987), *TT1* (JCGZ_15802) and *TTG2* (JCGZ_16445) (Data S3). These TF genes provide a starting point for further investigations into the regulation of seed size and coat thickness in physic nut.

Genes involved in the biosynthesis of oils and starches

In light of the importance of physic nut as an oilseed crop, we examined the gene models in physic nut and castor bean genomes that resembled genes known to be involved in the biosynthesis of fatty acids and triacylglycerols (TAGs) in *Arabidopsis* (<http://aralip.plantbiology.msu.edu/pathways>) (Data S4) (Beisson *et al.*, 2003; Li-Beisson *et al.*, 2013). Of genes from the plastid fatty acid biosynthesis pathway, the physic nut and castor bean genomes contain one more carboxyltransferase gene of the heteromeric acetyl-CoA carboxylase type than *Arabidopsis*, along with one more ketoacyl-ACP synthase I gene, one more enoyl-ACP reductase gene and two more FATA/B-like genes (Figure

S12), but one less homomeric acetyl-CoA carboxylase gene, one less hydroxyacyl-ACP dehydrase gene, two fewer ketoacyl-ACP reductase genes and three fewer stearyl-ACP desaturase genes (Data S4). Of genes involved in the biosynthesis and storage of TAG pathway, the physic nut genome contains two more linoleate desaturase-like genes than either *Arabidopsis* or castor bean. Physic nut and castor bean also contain one more phospholipid: diacylglycerol acyltransferase gene but fewer oil-body protein genes than *Arabidopsis*. The physic nut genome contains the same gene subfamilies for each enzyme isoform as the castor bean genome in fatty acid and TAG biosynthesis pathway (Data S4), although their seed oils all have different fatty acid compositions (Jiang *et al.*, 2012). Physic nut has an oleate desaturase-like gene which is grouped with oleate hydroxylase (FAH) that catalyzes the formation of ricinoleic acid from oleoyl-phosphatidylcholine in castor bean (van de Loo *et al.*, 1995) (Figure S13), but not in *Arabidopsis*.

We also manually annotated and summarized the physic nut gene models associated with starch metabolism and carbon flux (Table S16). Two genes were detected that had not previously been reported in higher plants – a novel starch synthase isoform (SSVI, JCGZ_07972) and a starch phosphorylase isoform containing an ACT domain [cd02116] (PHOA, JCGZ_21479) (Figure S14, Table S16). One or both of these genes could be detected in several dicotyledons and in green algae, but not in the Poaceae whose sequences are in public genome databases (Table S17). Both of these genes had different exon–intron arrangements from those of other reported genes in their subfamilies (Tables S18 and S19) and grouped into new clusters on the phylogenetic trees (Figures S15 and S16). These starch synthesis genes thus appear to be of ancient origin and to have been retained in physic nut as well as in some (but not all) other plants during their evolution.

Genes involved in toxicant synthesis

The toxicity of physic nut has been attributed to its content of ribosome-inactivating proteins (RIPs) and diterpenoid phorbol esters (King *et al.*, 2009). Two types of RIPs are present in members of the Euphorbiaceae. Curcin, which is found in physic nut, is a type I RIP, while ricin found in castor bean is a type II RIP. The castor bean genome contains at least seven genes encoding such proteins (Tregear and Roberts, 1992). We detected 11 curcin genes (*JcRIP1* to *JcRIP11*) and a single ricin-like gene (*JcRIP12*) in the physic nut genome (Figure 3). The existence of these 12 RIP genes was confirmed by PCR from genomic DNA and sequencing (Appendices S1 and S2). On two genomic scaffolds (scaffolds 44 and 304, which together contained 11 genes), the RIP genes are clustered and organized in tandem repeats, suggesting that the original RIP gene has recently undergone repeated duplication at two loci. In castor bean, ricin accumulates in the seeds (Tregear and Roberts, 1992).

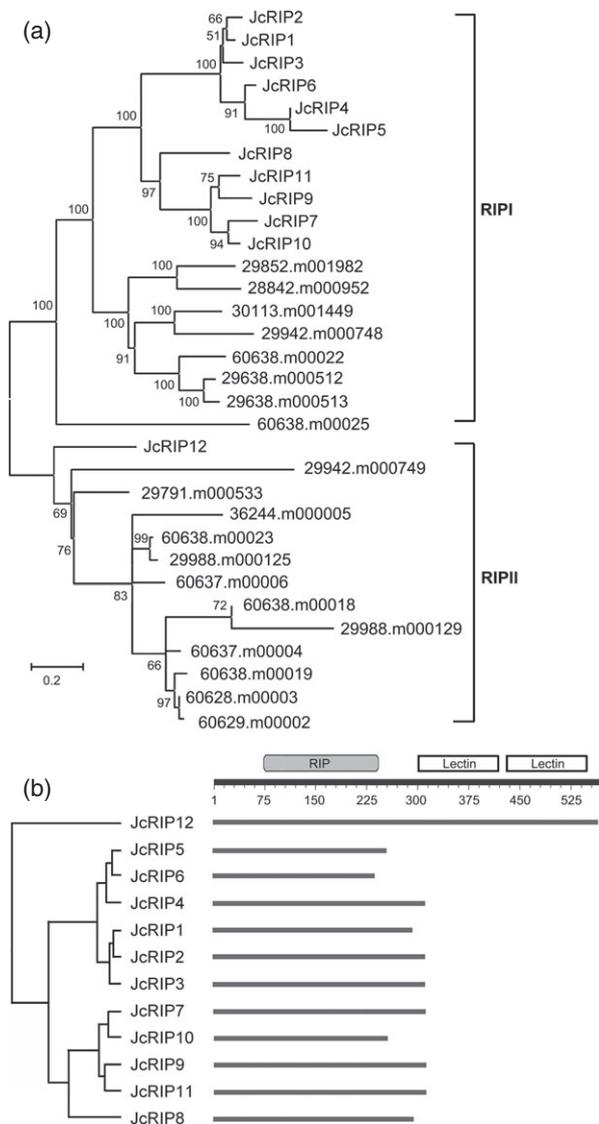


Figure 3. Members of the ribosome-inactivating protein (RIP) gene family in physic nut.

(a) Neighbor-joining tree for RIPs in physic nut and castor bean. The castor bean protein data (29852.m001982, etc) were obtained from Phytosome (<http://www.phytosome.net/ricinus.php>).

(b) Schematic representation of the members of the RIP gene family in physic nut. The RIP domain is represented at the top by gray boxes, while lectin domains are represented by empty boxes.

However, only two RIP genes (*JcRIP1* and *JcRIP3*) were strongly expressed in seeds at the filling stage in the physic nut cultivar GZQX0401. This difference is presumably due to differences in the expansion or contraction of RIP genes in the two species over time, and divergence in their expression.

Terpenoids (isoprenoids) are a very large and structurally diverse group of natural products (Buckingham, 1998). The toxicity of physic nut seeds derives mainly from phorbol esters, members of the tiglane family of

diterpenes (King *et al.*, 2009). The parent skeletons typical of plant terpenoids are formed from five-carbon biosynthetic building blocks and their polymeric derivatives by terpene synthases (TPS) (Davis and Croteau, 2000). A total of 59 putative full-length TPS genes were detected in the physic nut genome (Table S20), which is fewer than in grape (69) (Martin *et al.*, 2010) but more than in *Arabidopsis* (40/32) (Aubourg *et al.*, 2002) and castor bean (41). Previous phylogenetic analyses of plant TPS protein sequences resulted in the delineation of seven TPS subfamilies (a-d, e/f, g and h) (Chen *et al.*, 2011). Five of these subfamilies were represented by at least one physic nut TPS gene, but no genes from the d or h subfamilies were identified. Physic nut (two) has fewer TPS_e (GA2) genes than castor bean (six), but more TPS_a, b and g genes. The expansion of the TPS gene family suggests that physic nut can synthesize a wide range of terpenoid secondary metabolites.

Application of the linkage map to genetic analysis of traits in *Jatropha*

The leaves of *J. integerrima* seedlings are purple. This purple trait appears to be controlled by a single dominant gene, since F₁ plants are purple and there is 1:1 segregation in the BC₁ population (purple offspring: green offspring = 166:162) used in this study. This purple gene was found to lie between markers X5-8 (on scaffold 328) and E523-2 (on scaffold 523) on linkage group 5 (Figure S17). Several R2R3 MYBs (PAP1/PAP2, MYB113/MYB114) have been shown to underlie the expression of genes and the content of proteins implicated in the production of anthocyanins in *Arabidopsis* rosette leaves (Gonzalez *et al.*, 2008). Phylogenetic analysis showed eight orthologs of these four *Arabidopsis* genes in physic nut (Figure S17). These eight genes are located on scaffold 328 in tandem repeats (Figure S17). Two (JCGZ_11817 and JCGZ_11819) of the eight genes were found to be expressed in seeds, according to data in the physic nut EST database, but none was detected in leaves. However, the ortholog of JCGZ_11822 was expressed in the leaves of *J. integerrima*, according to the EST database for the latter species. Quantitative real-time PCR analysis indicated that this gene was strongly expressed in the leaves of purple visible *J. integerrima* and BC₁ plants, but that expression was very low in the leaves of physic nut and of green BC₁ offspring (Figure S18). These results suggest a key role for this PAP1/PAP2-like MYB gene in anthocyanidin synthesis in the leaves of *Jatropha* species. Thus, the identification of the gene responsible for purple leaves indicated that the linkage map and assembled pseudochromosomes reported here could provide a framework for the mapping of agronomic traits in physic nut, and for the development of improved varieties by marker-assisted breeding.

CONCLUSIONS

The Euphorbiaceae family includes some of the most efficient biomass accumulators, providing industrial raw materials and food. The genome sequence information generated in this study accompanied by the published castor bean genomes revealed the coexistence of 13 887 gene families in the two spurge genomes. This genome can advance our knowledge of the evolution of the plant kingdom because of the retention of many ancient gene families in physic nut. The further study of these genes should provide interesting information on their roles in development and metabolism in these plants.

In the work presented here, we have constructed a linkage map of *Jatropha*. Assemblies of 81.7% of the genome were anchored onto the 11 pseudochromosomes. Mapping orthologous markers to the assembled genome sequence of castor bean identified 410 scaffolds (about 54.1% of the total scaffold sequence) and revealed that a total of 320 syntenic blocks were well conserved. Linkage mapping in physic nut reveals a locus controlling the color in leaves and the candidate gene is a MYB family gene. This genetic map and the assembled chromosomes therefore provide a valuable foundation for quantitative trait locus mapping of agronomic traits and the development of better varieties for seeds yield and oil content. These foundation data could also serve as a useful resource for the future identification of genes by homology, and assist in the creation of syntenic genetic maps of physic nut and other Euphorbiaceae crops.

EXPERIMENTAL PROCEDURES

Preparation of DNA and RNA

We selected the physic nut inbred cultivar GZQX0401 for the genome sequencing project. Genomic DNA was isolated from young leaves using a modified SDS method (Peng *et al.*, 2007). Minor changes were made to in order to remove the large amounts of polysaccharide and polyhydroxybenzene that are present in physic nut leaves. A washing step with extraction buffer without SDS was included prior to extraction and a high concentration of NaCl (5 mol L⁻¹) was added before precipitation with isopropyl alcohol.

For transcriptome sequencing of *J. curcas*, roots from 15-day-old seedlings, leaves (half expanded) from 4-year-old plants and seeds from fruits harvested 19–28 DAP were used. Roots, stems, flowers and leaves of *J. integerrima* were collected from 1-year-old seedlings propagated from cuttings. Total RNA was extracted from seeds using an RNeasy Plant Mini Kit (Qiagen, <http://www.qiagen.com/>; cat. no.74 904). Total RNA from other tissues was extracted using the cetyltrimethylammonium bromide method according to Xiong *et al.* (2013) with minor modifications. An equal volume of isopropyl alcohol was used for precipitation, replacing the LiCl. The isolated RNA was subsequently treated with RNase-Free DNase I (Roche, <http://www.roche.com/>).

Mapping populations

An individual physic nut GZQX0401 plant was used as the female parent, and crossed with a *J. integerrima* BN01 individual (male

parent). Hybrid offspring, denoted GZBN1, were generated. The BC1 population, which contained over 300 progeny, was constructed by backcrossing one GZBN1 individual (male) to GZQX0401 (female). From the backcross, 190 offspring were used in the construction of the linkage map; 95 of these individuals had green leaves while the other 95 had purple leaves. The mapping populations and their parental lines were cultivated under standard growth conditions in a greenhouse at the South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China.

Sequencing strategy, data output and genome assembly

In accordance with the strategy of Li *et al.* (2010), we constructed a total of 23 PE sequencing libraries with seven different mean insert sizes: 200 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb and 20 kb. After library construction, we used Illumina GAll and HiSeq to sequence PE reads for each library. In total, we sequenced 49 lanes and generated 102.68 Gb of data. Before sequence assembly we performed an initial analysis of the sequence data, and applied a series of checking and filtering measures to reads generated by the Solexa pipeline to filter out human or technical errors. The initial analysis process included three treatments: image analysis, base calling and sequence analysis. The filtering process included removal of the following: reads in which N or polyA structure constituted more than 10% of the bases, low-quality reads, reads with adapter contamination, reads with a small insert size and PCR duplicates. Finally, we used the cleaned data, which gave 60 × coverage, to assemble the genome by SOAPdenovo (<http://soap.genomics.org.cn>) with K-mer value set at 35 bp.

We used 10 kb non-overlapping sliding windows and calculated the GC content and average depth of coverage in these windows. For sequence data with no serious GC bias, the scatter diagram should exhibit a Poisson-like distribution curve with a peak near the mean GC value for the genome.

In order to confirm the scaffolding, the clean data from 2 and 5 kbp libraries were mapped to the genome using SOAPaligner (version 2.20). The 17 BACs downloaded from GenBank (DDBJ AP011961-AP011977) were mapped against the assembled scaffolds using BLASTALL (2.2.23) to assess the genome assembly. Further, we used the LASTZ program (1.02.00) to confirm the completeness and continuity of the assembly with that presented by Hirakawa *et al.* (2012) with the following parameters: T = 2, C = 2, H = 2000, Y = 3400, L = 6000 and K = 2200.

Transcriptome sequencing and analysis

Samples for transcriptome sequencing were prepared using an Illumina kit (<http://www.illumina.com/>) following the manufacturer's recommendations. Briefly, mRNA was purified from 6 µg of total RNA from each tissue using oligo (dT) magnetic beads. Following purification, the mRNA was fragmented into small pieces using divalent cations at an elevated temperature. The cleaved RNA fragments were used as templates for first-strand cDNA synthesis using reverse transcriptase and random primers. This was followed by second-strand cDNA synthesis using DNA polymerase I and RNase H. The cDNA fragments were then subjected to an end-repair process and ligation of adapters. The products were purified and enriched by PCR to create the final cDNA library. The libraries constructed from different tissues were sequenced on an Illumina GAll. The insert size of each library was approximately 200 bp and both ends were sequenced. Image deconvolution and quality value calculations were performed using the Illumina GAll pipeline 1.3. The raw reads were cleaned by removing adaptor sequences, empty reads and low-quality sequences (reads with unknown sequences flagged as 'N'). After

assembly, sequences were used to produce RNA-seq data sets. For gene expression analysis, the number of expressed tags was calculated and then normalized to RPKM (reads of transcripts per kbp length of coding sequences per million tags).

Development of SSR, InDel and SNP markers

A total of 1120 SSR/InDel markers to be used for mapping were developed from genomic sequences (103), ESTs (919) and publicly available SSR markers (98) after analysis of published data (Wang *et al.*, 2011), as described below. In order to cover as many of the genomic sequences as possible we developed SNP markers for scaffold sequences that were over 70 000 bp long and lacked SSR markers. As a result, a total of 88 SNP markers for 73 scaffolds were obtained.

Genome sequence-based markers (103). We tested the genetic diversity of the 58 available *J. curcas* lines using 725 SSR markers developed from the draft genomic sequence. Primers were designed using PRIMER3PLUS (Untergasser *et al.*, 2007). Only four markers showed polymorphism among the 58 lines. To overcome this limitation, we developed a BC1 population by crossing physic nut with *J. integerrima*. About 14% of the genomic SSR markers exhibited polymorphism between physic nut and *J. integerrima* individuals. The PCR products for genomic SSRs of *J. integerrima* were cloned into the pMD18-T vector (TaKaRa, <http://www.takara.com.cn/>) and sequenced using M13 primers on an ABI 3730 DNA sequencer (Applied Biosystems, <http://www.appliedbiosystems.com/>).

Expressed sequence tag-based markers (919). To provide sufficient markers for the construction of the linkage map, we constructed a database of ESTs from leaf, root, stem and flower obtained by pyrosequencing of cDNA from *J. integerrima*. After alignment of the draft genomic sequence with this EST collection and analysis, 919 EST-SSR and EST-InDel markers were obtained.

Markers from the public domain (98). After analysis, a total of 98 public available SSR markers (Wang *et al.*, 2011) were identified as usable in this study.

Genotyping of SSR, InDels and SNPs

The SSR and InDel PCR products were separated on 8% polyacrylamide gel electrophoresis gels. After electrophoresis, gels were silver stained and genotyping was carried out on the basis of fragment size (Bassam *et al.*, 1991).

The SNP genotyping was based on high-resolution melting (HRM) analysis. The primers were designed to have an annealing temperature of $60 \pm 1^\circ\text{C}$ and to give expected product sizes of 70–150 bp, with a few exceptions. Each PCR reaction (10 μl) consisted of 1.0 μl ; $10 \times$ buffer, 200 μM dNTP, 1.5 mM MgCl_2 , 0.25 μM each of forward and reverse primers, 0.25 U *Taq* polymerase (TaKaRa), about 30 ng DNA and 1.0 μl LC Green PLUS DNA binding dye (BioFire Diagnostics, <http://www.biofiredx.com/>). The HRM properties of PCR products were measured on a 96-well Light Scanner (BioFire Diagnostics) and the data analyzed with LIGHT SCANNER CALL-IT 2.0 (version 2.0.0.1331) (Montgomery *et al.*, 2007). Genotypes were identified by examining normalized, difference and derivative melt plots.

Linkage analysis and genetic mapping

JOIN-MAP version 4.0 was used to build linkage maps (Van Ooijen, 2006). Linkage analysis and marker grouping were performed

using the independence LOD and a minimum threshold LOD value of 4.0. Phases (coupling and repulsion) of the linked marker loci were automatically detected by the software. Map distances in cM were calculated using the maximum likelihood mapping algorithm and the Kosambi mapping function. Linkage maps were drawn using the MAPCHART program.

We used the markers derived from the physic nut genome sequence to perform comparative mapping against the castor bean draft genomic sequences. For molecular markers that were contained within physic nut protein-coding genes we used the corresponding gene sequence to identify the orthologous castor bean gene. For molecular markers located in an intergenic region we used the closest physic nut protein-coding gene within 5 kbp (upstream or downstream) to search for the orthologous castor bean gene. To verify the orthologous status of the genes identified in this way, we required the physic nut gene and its putative castor bean counterpart to encode proteins belonging to the same gene family according to a BLASTP search. In addition, the physic nut gene (A) was required to have a greater degree of amino acid sequence identity with the putative castor bean ortholog (B) than with any other member of the ortholog's gene family in the castor bean genome, and vice versa. The castor bean contigs containing these orthologous genes were anchored onto the genetic map.

Gene prediction and genome annotation

We searched the genome for tandem repeats with the help of TANDEM REPEATS FINDER (Benson, 1999). We used LTR_FINDER (Xu and Wang, 2007), PILER-DF (Edgar and Myers, 2005) and REPEATSCOUT (Price *et al.*, 2005) to build *de novo* libraries of repeats from the genome. Transposable elements in the genome assembly were identified at both the DNA and the protein level. REPEATMASKER (<http://repeatmasker.org>) was applied for DNA-level identification of elements in a custom library (a combination of the Repbase plant repeat database and our genome *de novo* TE library). At the protein level, REPEATPROTEINMASK, a component of the REPEATMASKER package, was used to perform WuBlastX against the database of TE protein sequences. Finally, a repeat library was constructed by combining all the repeats detected as described above. Contaminants and multicopy genes were filtered from the library. REPEATMASKER was run again on this library to identify and classify homologous repeats in the genome.

We used three methods to predict genes. *De novo* gene prediction was performed using AUGUSTUS (Stanke *et al.*, 2006), GENSCAN (Salamov and Solovyev, 2000) and GLIMMERHMM (Majoros *et al.*, 2004), with appropriate parameters for predict the sites of coding genes. For homolog prediction we used homologous protein sequences from seven plant genomes (*Arabidopsis thaliana*, *Carica papaya*, *Cucumis sativus*, *Glycine max*, *P. trichocarpa*, *R. communis* and *Vitis vinifera*). For cDNA/EST prediction we used the coding sequences of physic nut, including ESTs, full-length cDNAs and unigenes, to align against the physic nut genome using BLAT (Kent, 2002) to generate spliced alignments. We then integrated the results of *de novo* gene prediction with spliced alignments of proteins and transcripts to genome sequences using GLEAN to produce consensus gene sets (Elsik *et al.*, 2007). Next we aligned the RNA-seq raw reads against the assembled genome sequence with the software package TOPHAT (mismatch ≤ 1). We used CUFFLINK to predict the splicing positions for each transcript and the abundance of each isoform. BESTORF (using *A. thaliana* as the training set) was used for predicting the open reading frame of each transcript. Lastly, we combined the GLEAN annotation results and the RNA-seq-based annotation results to obtain the final gene set. Coding sequences of lengths less than 150 bp and/or having

translation errors were filtered out. For those gene loci giving more than one transcript, the longest transcript was used for subsequent evolution analysis.

Gene functions were assigned on the basis of the best matches given by BLASTP alignments to the SwissProt and TrEMBL databases (Bairoch and Apweiler, 2000). Motifs and domains in genes were determined by INTERPROSCAN (Zdobnov and Apweiler, 2001) searches against protein databases including Pfam, PRINTS, PROSITE, ProDom and SMART. Gene Ontology IDs for each gene were obtained from the corresponding InterPro entry (Ashburner *et al.*, 2000). All genes were aligned against Kyoto Encyclopedia of Genes and Genomes (KEGG) proteins, and the pathway in which the gene was most likely to be involved was identified from the corresponding genes in KEGG.

Analysis of gene families and their evolution

We used OrthoMCL methodology (Li *et al.*, 2003) to define a gene family as a group of genes descended from a single gene in the last common ancestor of the species. BLASTP was used to compare the putative physic nut protein sequences with a dataset of the selected plant species and those matches with e -values below 1×10^{-5} were retained for further analysis. We used the OrthoMCL method in mode 3 to identify gene families. We used single-copy gene families from the physic nut genome and other sequenced genomes to reconstruct the phylogeny. Multiple alignments were adapted by MUSCLE (version 3.7) (Edgar, 2004). Fourfold degenerate sites were extracted from each family and concatenated to one supergene for each species. The phylogenetic tree was then reconstructed using MRBAYES (Huelsenbeck and Ronquist, 2001) with the HKY model (Hasegawa *et al.*, 1985) on the basis of the super genes. The K_a/K_s ratio was estimated using codem1 in PAML (Yang, 2007) and the transition/transversion rate ratio was estimated as a free parameter. Other parameters were set to their default values. The Bayesian relaxed molecular clock (BRMC) approach was used to estimate species divergence time with the program MULTIDIVTIME, which was implemented using the Thornian Time Traveller (T3) package (Wikström *et al.*, 2001; Crepet *et al.*, 2004). For statistical analysis of the expansion and contraction of gene families, we used computational analysis of gene family evolution (CAFE) (Hahn *et al.*, 2005). Tandemly repeated gene arrays were identified using the BLASTP program with a threshold of 1×10^{-20} . No more than two intervening genes were permitted to be present between two putative tandemly duplicated genes.

For construction of syntenic blocks, we identified blocks of synteny between each pair of candidates (A and B) using the software MCSCAN (<http://chibba.agtec.uga.edu/duplication/mcscan/>). For orthologous syntenic blocks between two species ($A \neq B$), pairwise similarities were extracted from the protein sequences obtained by BLASTP analysis. For paralogous syntenic blocks in one species ($A = B$), pairwise similarities were obtained from the nucleotide sequences given by blasting the genome against itself using BLASTN (e -value $< 1 \times 10^{-5}$) (Tang *et al.*, 2008).

To do the fourfold degenerate site transversion (4DTV) calculation, first, pairwise protein alignments for each gene pair were constructed with MUSCLE (version 3.7) (Edgar, 2004). Second, nucleotide alignments were derived from the protein alignments. The number of transversions and the number of 4D degenerate sites in each pair of similar sequences were counted (their quotient was the raw 4DTV value). Finally, the 4DTV value was calculated from the nucleotide alignments with the HKY substitution model (Hasegawa *et al.*, 1985). K_s values was calculated by yn00 in the PAML software package (Yang, 2007).

For amino acid sequence alignment and phylogenetic tree construction, amino acid sequences were aligned using CLUSTALX version 1.83 (Thompson *et al.*, 1997) and then alignments were manually modified. Phylogenetic trees were constructed by the neighbor-joining method with MEGA 4.0 (Tamura *et al.*, 2007) using the Poisson correction model (Nei and Kumar, 2000) with 1000 bootstrap replications.

Data access

This WGS project has been deposited at DDBJ/EMBL/GenBank under the accession AFEW000000000. The version described in this paper is the first version, AFEW010000000. The physic nut sequence reads generated during genome and transcriptome sequencing have been deposited in the NCBI Sequence Read Archive (SRA) under accession number SRA197144, and the *J. integerrima* transcriptome sequencing reads under SRA197148.

ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program of China (2010CB126600), the Key Innovation Programs of the Chinese Academy of Sciences (KSCX2-YW-G-035), the National Natural Science Foundation of China (31270705, 31070227 and 31200513) and the NSFC-Guangdong Joint Fund of Natural Sciences (U0733005). The authors declare no competing interests.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Sequence depth distribution.

Figure S2. Distribution of GC content in the physic nut and castor bean genomes.

Figure S3. Validation of genome assembly based on paired-end reads.

Figure S4. Assembly assessment by bacterial artificial chromosomes.

Figure S5. Validation of genome assembly based on collinearity blocks generated by LASTZ.

Figure S6. Maps of the 11 linkage groups.

Figure S7. Estimation of divergence time.

Figure S8. Distributions of synonymous substitution rates (K_s) for homologous gene groups in intrachromosome and interchromosome comparisons.

Figure S9. Blocks of synteny detected between the physic nut and grape pseudochromosomes on the basis of gene orthology.

Figure S10. Expansion and contraction of the gene families among three dicotyledons.

Figure S11. Proportion of ancient duplicates (originating in the hexaploidization event that took place in dicotyledons) and tandemly duplicated gene sets differentially expressed in root, leaf and seed.

Figure S12. The neighbor-joining tree of orthologs of FATA-B proteins.

Figure S13. The neighbor-joining tree of orthologs of Arabidopsis FAD2-8 proteins.

Figure S14. Comparison of the domains found within the predicted amino acid sequences of physic nut starch synthase and starch phosphorylase proteins.

Figure S15. The neighbor-joining tree of starch synthase proteins with conserved domains in physic nut and seven other plant species.

Figure S16. The neighbor-joining tree of starch phosphorylase proteins in physic nut and seven other plant species.

Figure S17. The positions of the *Purple* locus and the MYB gene cluster on the physic nut genome.

Figure S18. Quantitative real-time PCR analysis of the expression of JCGZ_11822 in leaves of physic nut, *Jatropha integerrima*, and nine purple-leaf and nine green-leaf offspring of the BC1 population.

Table S1. Summary of raw sequencing data.

Table S2. Summary of filtered sequencing data.

Table S3. Statistics for the assembled sequence.

Table S4. Assessment of coverage of the physic nut genome assembly using expressed sequence tag-cDNA data.

Table S5. Repetitive sequences and transposons.

Table S6. Statistics for predicted protein-coding genes.

Table S7. Combined result of RNA-seq and GLEAN.

Table S8. Gene family clustering.

Table S9. Number of genes with homology or functional classification assigned by each method.

Table S10. Summary of linkage group sizes, marker numbers and densities, and scaffolds anchoring to the *Jatropha* linkage map.

Table S11. Tandem gene clusters in physic nut.

Table S12. Numbers of putative genes in cellulose and lignin biosynthesis-related gene families in physic nut, castor bean and *Arabidopsis*.

Table S13. Number of putative cell-wall degradation related gene families in physic nut, castor bean and *Arabidopsis*.

Table S14. Putative nucleotide-binding site super-family genes detected in the draft physic nut genome.

Table S15. Putative transcription factor genes detected in the draft physic nut genome.

Table S16. Number of predicted genes encoding enzymes of starch metabolism and carbon flux in physic nut, castor bean and *Arabidopsis*.

Table S17. Presence of putative genes encoding the starch synthase isoform VI and the ACT domain-containing starch phosphorylase in different plant species, determined from publicly available genome sequences.

Table S18. Exon arrangements in starch synthase genes in the physic nut genome.

Table S19. The exon arrangement of physic nut starch phosphorylase genes.

Table S20. Manually annotated terpene synthase genes in the draft physic nut genome.

Data S1. The simple sequence repeat, insertion–deletion and single nucleotide polymorphism markers used in this study (a) and the detailed information about DNA markers and scaffolds mapped on the genetic map of *Jatropha* (b).

Data S2. The putative cytochrome P450 monooxygenase (CYP) super-family genes detected in this draft genome.

Data S3. Transcription factor genes highly expressed in the early developing stage seeds.

Data S4. Physic nut genes involved in the synthesis of fatty acid and triacylglycerols.

Appendix S1. Gene sequences of ribosome-inactivating proteins confirmed by PCR from genomic DNA and sequencing.

Appendix S2. The revised sequences of starch synthase genes and starch phosphorylase genes, and sequences of terpenoid synthase genes not predicted by genome annotation.

REFERENCES

- Abrouk, M., Murat, F., Pont, C. *et al.* (2010) Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci.* **15**, 479–487.
- Argout, X., Salse, J., Aury, J.M. *et al.* (2011) The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108.
- Ashburner, M., Ball, C.A., Blake, J.A. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29.
- Aubourg, S., Lechary, A. and Bohlmann, J. (2002) Genomic analysis of the terpenoid synthase (AtTPS) gene family of *Arabidopsis thaliana*. *Mol. Genet. Genomics*, **267**, 730–745.
- Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48.
- Bassam, B.J., Caetano-Anollés, G. and Gresshoff, P.M. (1991) Fast and sensitive silver staining of DNA in polyacrylamide gels. *Anal. Biochem.* **196**, 80–83.
- Beisson, F., Koo, A.J., Ruuska, S. *et al.* (2003) *Arabidopsis* genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. *Plant Physiol.* **132**, 681–697.
- Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.
- Buckingham, J. (1998) *Dictionary of Natural Products on CD-ROM/Version 6.1*. London: Chapman & Hall.
- Carvalho, C.R., Clarindo, W.R., Praca, M.M., Araújo, F.S. and Carels, N. (2008) Genome size, base composition and karyotype of *Jatropha curcas* L., an important biofuel plant. *Plant Sci.* **174**, 613–617.
- Chan, A.P., Crabtree, J., Zhao, Q. *et al.* (2010) Draft genome sequence of the oilseed species *Ricinus communis*. *Nat. Biotechnol.* **28**, 951–956.
- Chen, F., Tholl, D., Bohlmann, J. and Pichersky, E. (2011) The family of terpene synthases in plants, a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212–229.
- Crepet, W.L., Nixon, K.C. and Gandolfo, M.A. (2004) Fossil evidence and phylogeny: the age of major angiosperm clades based on mesofossil and macrofossil evidence from Cretaceous deposits. *Am. J. Bot.* **91**, 1666–1682.
- Davis, E.M. and Croteau, R. (2000) Cyclization enzymes in the biosynthesis of monoterpenes, sesquiterpenes, and diterpenes. *Top. Curr. Chem.* **209**, 53–95.
- Dehgan, B. (1984) Phylogenetic significance of interspecific hybridization in *Jatropha* (Euphorbiaceae). *Syst. Bot.* **9**, 467–478.
- DeYoung, B.J. and Innes, R.W. (2006) Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat. Immunol.* **7**, 1243–1249.
- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797.
- Edgar, R.C. and Myers, E.W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics*, **21**, i152–i158.
- Elsik, C.G., Mackey, A.J., Reese, J.T., Milshina, N.V., Roos, D.S. and Weinstock, G.M. (2007) Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13.
- Goff, S.A., Ricke, D., Lan, T.H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science*, **296**, 92–100.
- Gonzalez, A., Zhao, M., Leavitt, J.M. and Lloyd, A.M. (2008) Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. *Plant J.* **53**, 814–827.
- Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. and Cristianini, N. (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* **15**, 1153–1160.
- Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.
- Hirakawa, H., Tsuchimoto, S., Sakai, H. *et al.* (2012) Upgraded genomic information of *Jatropha curcas* L. *Plant Biotechnol.* **29**, 123–130.
- Huelsbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.
- Jailion, O., Aury, J.M., Noel, B. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

- Jiang, H., Wu, P., Zhang, S., Song, C., Chen, Y., Li, M., Jia, Y., Fang, X., Chen, F. and Wu, G. (2012) Global analysis of gene expression profiles in developing physic nut (*Jatropha curcas* L.) seeds. *PLoS ONE*, **7**, e36522.
- Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30.
- Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.
- King, A.J., He, W., Cuevas, J.A., Freudenberger, M., Ramiarmanana, D. and Graham, I.A. (2009) Potential of *Jatropha curcas* as a source of renewable oil and animal feed. *J. Exp. Bot.* **60**, 2897–2905.
- King, A.J., Montes, L.R., Clarke, J.G. *et al.* (2013) Linkage mapping in the oilseed crop *Jatropha curcas* L. reveals a locus controlling the biosynthesis of phorbol esters which cause seed toxicity. *Plant Biotechnol. J.* **11**, 986–996.
- Kumar, A. and Sharma, S. (2008) An evaluation of multipurpose oil seed crop for industrial uses (*Jatropha curcas* L.): a review. *Ind. Crops Prod.* **28**, 1–10.
- Li, L., Stoekert, C.J. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.
- Li, R., Fan, W., Tian, G. *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.
- Li-Beisson, Y., Shorrosh, B., Beisson, F. *et al.* (2013) *Acyl-lipid metabolism*. *Arabidopsis Book*, **11**, e0161.
- van de Loo, F.J., Broun, P., Turner, S. and Somerville, C. (1995) An oleate 12-hydroxylase from *Ricinus communis* L. is a fatty acyl desaturase homolog. *Proc. Natl Acad. Sci. USA*, **92**, 6743–6747.
- Majoros, W.H., Pertea, M. and Salzberg, S.L. (2004) TigrScan and Glimmer-HMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.
- Martin, D.M., Aubourg, S., Schouwey, M.B., Daviet, L., Schalk, M., Toub, O., Lund, S.T. and Bohlmann, J. (2010) Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.* **10**, 226.
- Meyers, B.C., Kozik, A., Griego, A., Kuang, H. and Michelmore, R.W. (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell*, **15**, 809–834.
- Ming, R., Hou, S., Feng, Y. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, **452**, 991–996.
- Montgomery, J., Wittwer, C.T., Palais, R. and Zhou, L. (2007) Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis. *Nat. Protoc.* **2**, 59–66.
- Nei, M. and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*. Oxford: Oxford University Press.
- Nelson, D.R., Schuler, M.A., Paquette, S.M., Werck-Reichhart, D. and Bak, S. (2004) Comparative genomics of rice and Arabidopsis. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. *Plant Physiol.* **135**, 756–772.
- Peng, J., Wang, H., Haley, S.D., Peairs, F.B. and Lapitan, N.L.V. (2007) Molecular mapping of the Russian wheat aphid resistance gene Dn2414 in wheat. *Crop Sci.* **47**, 2418–2429.
- Price, A.L., Jones, N.C. and Pevzner, P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21**, i351–i358.
- Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522.
- Sato, S., Hirakawa, H., Isobe, S. *et al.* (2011) Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. *DNA Res.* **18**, 65–76.
- Shulaev, V., Sargent, D.J., Crowhurst, R.N. *et al.* (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* **43**, 109–116.
- Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439.
- Sun, X., Shantharaj, D., Kang, X. and Ni, M. (2010) Transcriptional and hormonal signaling control of Arabidopsis seed development. *Curr. Opin. Plant Biol.* **13**, 611–620.
- Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599.
- Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.
- The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882.
- Tregear, J.W. and Roberts, L.M. (1992) The lectin gene family of *Ricinus communis*: cloning of a functional ricin gene and three lectin pseudogenes. *Plant Mol. Biol.* **18**, 515–525.
- Tuskan, G.A., Difazio, S., Jansson, S. *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.
- Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R. and Leunissen, J.A. (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* **35**, W71–W74.
- Van Ooijen, J.W. (2006) *JoinMap® 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations*. Wageningen, Netherlands: Kyazma B.V.
- Wang, C.M., Liu, P., Yi, C. *et al.* (2011) A first generation microsatellite and SNP-based linkage map of *Jatropha*. *PLoS ONE*, **6**, e23632.
- Wikström, N., Savolainen, V. and Chase, M.W. (2001) Evolution of the angiosperms: calibrating the family tree. *Proc. Biol. Sci.* **268**, 2211–2220.
- Xiong, W., Xu, X., Zhang, L., Wu, P., Chen, Y., Li, M., Jiang, H. and Wu, G. (2013) Genome-wide analysis of the WRKY gene family in physic nut (*Jatropha curcas* L.). *Gene*, **524**, 124–132.
- Xu, Z. and Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268.
- Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
- Yu, J., Hu, S., Wang, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.
- Zdobnov, E.M. and Apweiler, R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.