Genomics Insights

ORIGINAL RESEARCH

ESTs from Seeds to Assist the Selective Breeding of *Jatropha curcas* L. for Oil and Active Compounds

Kleber A. Gomes[1,2], Tiago C. Almeida[1], Abelmon S. Gesteira[1,3], Ivon P. Lôbo[4], Ana Carolina R. Guimarães[5], Antonio B. de Miranda[5], Marie-Anne Van Sluys[2], Rosenira S. da Cruz[4], Júlio C.M. Cascardo[1] and Nicolas Carels[1,5]

[1]Universidade Estadual de Santa Cruz (UESC), Centro de Biotecnologia e Genética. Laboratório de Genômica e Proteômica, Ilhéus, Bahia, Brazil. [2]Universidade de São Paulo, Instituto de Biociências, Departamento de Botânica, São Paulo, SP, Brazil. [3]Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) Mandioca e Fruticultura Tropical, Cruz das Almas, Bahia, Brazil. [4]Universidade Estadual de Santa Cruz (UESC), Grupo Bioenergia e Meio Ambiente, Ilhéus, Bahia, Brazil. [5]Fundação Oswaldo Cruz (FIOCRUZ), Instituto Oswaldo Cruz (IOC), Laboratório de Genômica Funcional e Bioinformática, Rio de Janeiro, RJ, Brazil. Corresponding author email: nicolas.carels@gmail.com

Abstract: We report here on the characterization of a cDNA library from seeds of *Jatropha curcas* L. at three stages of fruit maturation before yellowing. We sequenced a total of 2200 clones and obtained a set of 931 non-redundant sequences (unigenes) after trimming and quality control, ie, 140 contigs and 791 singlets with PHRED quality ≥10. We found low levels of sequence redundancy and extensive metabolic coverage by homology comparison to GO. After comparison of 5841 non-redundant ESTs from a total of 13193 reads from GenBank with KEGG, we identified tags with nucleotide variations among *J. curcas* accessions for genes of fatty acid, terpene, alkaloid, quinone and hormone pathways of biosynthesis. More specifically, the expression level of four genes (palmitoyl-acyl carrier protein thioesterase, 3-ketoacyl-CoA thiolase B, lysophosphatidic acid acyltransferase and geranyl pyrophosphate synthase) measured by real-time PCR proved to be significantly different between leaves and fruits. Since the nucleotide polymorphism of these tags is associated to higher level of gene expression in fruits compared to leaves, we propose this approach to speed up the search for quantitative traits in selective breeding of *J. curcas*. We also discuss its potential utility for the selective breeding of economically important traits in *J. curcas*.

Keywords: biofuel, *Jatropha curcas*, genomics, fatty acids, terpenes, alkaloids

Genomics Insights 2010:3 29–56

doi: 10.4137/GEI.S4340

This article is available from http://www.la-press.com.

Introduction

Jatropha curcas L. (Plantae; Embryophyta; Spermatopsida; Malpighiales; *Euphorbiaceae*) has recently drawn the attention of the international research community due to its potential as a biodiesel crop.[1] There is an urgent need to diversify oil sources to sustain B5 (5% biodiesel) that is mandatory in Brazil since 2010. In that country, B5 is fueled mainly by soybean (80%), which is a dangerous situation for supply stability in the long term. On the other hand, there is a trend toward the increase of climate dryness in continental areas. For this reason, the availability of an industrial fuel crop well adapted to semiarid conditions would be welcome.

J. curcas is a shrub (3–10 m) whose oldest fossil remains were found by Berry[2] in geological formations from Peru, suggesting that it most probably originates from northern South America. However, establishing a definitive origin of *J. curcas* will depend on the careful analysis of genetic diversity that is currently ongoing in Brazil and Central America (personal communication from B. Galvêas Laviola). Among all of the potential biofuel production crops, *J. curcas* stands out due to its affinity for semiarid to arid growing conditions in tropical and subtropical climates,[3] adaptability to soils of poor quality, including wasted lands, distinctive oil quality and high oil content, short cropping period, low seed cost and ability of biodiesel to persist stable upon storage.

The economic viability of using plant oils as renewable resources is critically dependent on the energetic balance of oil production in this system. Therefore, it is necessary to decipher the regulation of lipogenesis in maturing oilseeds.

Fatty acid synthesis occurs in the plastids. Many seeds accumulate large reservoirs of oil in the form of triacylglycerols (TAG). TAG assembly can be considered as proceeding by two distinct routes.[4] One of these, the Kennedy pathway, relies on the sequential acylation process of fatty acid biosynthesis of a glycerol-3-phosphate (Gly-3-P) backbone provided by GPDH. The second route relies on acyl-exchange between lipids and involves a phospholipid diacylglycerol acyltransferase (PDAT). Basically, TAGs are produced through the condensation of two fatty acid molecules with one molecule of phosphoric acid and one molecule of glycerol. The process of esterification produces various types of phosphatidic acids, which are subsequently dephosphorylated in the endoplasmic reticulum to form diacylglycerol (DAG). Finally, a fatty acid chain is added to the DAG by an acyltransferase, resulting in a TAG that is sequestered in lipid bodies for storage.[5] TAGs are used to support germination and early seedling growth. In this sense, fatty acids act as molecules of energy storage that are used to "fuel" the cell when other energy sources are not available.

During seed maturation, hexose phosphates from photo-assimilates are massively metabolized through the glycolytic pathway before being used for *de novo* fatty acid synthesis in plastids.[6] Fatty acid synthesis starts with the condensation of a malonyl and an acetyl unit to form a four-carbon fragment. To produce the required hydrocarbon chain, this fragment is reduced, dehydrated, and reduced again. The process is catalyzed repeatedly by fatty acid synthase until a C16 fatty acid (palmitate) is synthesized.

Fatty acids chains longer than palmitate are formed by elongation reactions catalyzed by enzymes on the cytosolic side of the endoplasmic reticulum membrane. These reactions add two carbon units (Malonyl CoA) sequentially to the carboxyl ends of both saturated and unsaturated fatty acyl CoA substrates.

To be used as energy fuel, the fatty acids must be activated and transported into mitochondria for degradation (β-oxidation). The fatty acids are then broken down in a step-by-step way into acetyl CoA, which is processed in the citric acid cycle. The conversion of fatty acid into energy (NADH, ATP) occurs in the citric acid cycle.

Acetyl CoA carboxylase plays an essential role in regulating fatty acid synthesis and degradation. Low levels of acetyl-CoA induce fatty acid β-oxidation, which results in ADH2, NADH and acetyl-CoA. The excess of acetyl-CoA results in production of excess citrate (citric acid cycle), which is exported into the cytosol to give rise to cytosolic acetyl-CoA. A high level of citrate means that two-carbon units and ATP are available for fatty acid synthesis. Acetyl-CoA can be carboxylated into malonyl-CoA that is required for (i) synthesis of flavonoids and related polyketides, (ii) elongation of fatty acids to produce waxes, cuticule, and seed oils, and (iii) malonation of proteins and other phytochemicals such as terpenes, steroids and sterols.[7]

The terpenoids constitute the largest class of natural products produced by plants. Terpenoids form groups of related structural types that allow the detection of chemosystematic relations among species. The plant families of *Euphorbiaceae* and *Thymelaeaceae* contain many tumor promoting and irritant diterpenoids. These compounds can all be considered as biogenetically arising from the casbane skeleton. The lathyrane skeleton can be formed by cyclisation of the casbane skeleton and the opening of the cyclopropane ring would lead to the jatrophane skeleton. Further cyclisation of lathyrane skeleton leads to the tigliane skeleton, which can undergo cyclopropane ring opening or rearrangement to give rise to the daphnane and the ingenane skeleton, respectively. Esters of ingol, lathyrane and phorbol, which is a tigliane, are clearly derived from the ingenane skeleton.[8] These esters possess potent biological activities. Phorbol, in particular, interacts with the C1 domain[9] of the protein kinase C super-family,[10] which is conserved among a large number of species. This molecule is used to induce skin tumors in mouse (a model for cancer investigations),[10] and is responsible for strong digestive poisoning upon ingestion. This has raised concerns about the health of people that may come into continuous contact with *J. curcas* oil as well as for animals fed with the oil cake.[1] On the other hand, terpenoids are active components of plant defenses and engineering their elimination[11] could raise additional costs in chemical treatment to ensure crop protection. An economical evaluation of this particular issue is necessary. At first glance, non-toxic varieties could be more economically interesting to small farmers that could use the oil cake to feed animals and toxic varieties to large farmers that would deliver the seed cake back to the field depending on geo-economical constraints.

Despite the attractive features of its oil composition and productivity, *J. curcas* has never been domesticated and its yield is difficult to predict with accuracy. The conditions that best suit its growth are not well-defined and the potential environmental impacts of large-scale cultivation are not yet understood. In addition, other crop features that obviously need to be improved are seed morphology, oil content,[12–14] synchronization of fruit maturation, plant size, toxicity, digestibility, resistance to pests and diseases. Physico-chemical features such as the amount of free fatty acids, unsaponifiables, acid number and carbon residues could also be addressed by selective breeding. However, without understanding the basics of *J. curcas*, a premature push to cultivate it could prove to be very unproductive.

There is a critical need for scientific breeding of Jatropha guided by advanced DNA mapping technologies.[15,16] Individuals of *J. curcas* exhibit high phenotypic interaction with the environment, which makes DNA probes with reproducible polymorphisms essential.[17] Through DNA sequencing techniques, it has become obvious that the genes associated with fatty acid biosynthesis are expressed at a medium to high degree in the developing seed.[18] This property is particularly interesting for the identification of cDNAs involved in fatty acid biosynthesis through EST profiling.[19] In addition, EST sequencing, associated with real-time PCR and even pyrosequencing,[20] allows the investigation of the gene regulation network both in the lipid pathway[21–23] and among genotypes.[24] Even when the metabolic pathway is known, as it is for the fatty acid pathway, the analysis of EST datasets is often a faster process for the identification of corresponding enzymes. The protein family corresponding to a given primer pair can be large, as is the case of P450 in Arabidopsis, which accounts for >250 genes.[24] This makes PCR investigation difficult. Another useful feature of ESTs is that they can probe for key genes involved in complex traits such as oil synthesis, which can be mapped using a combination of EST and AFLP techniques.[25,26]

The identification of quantitative traits that improve agricultural production allows breeders to introduce economically important traits into modern genetic backgrounds and to investigate the molecular mechanisms that regulate their effects.[27] For instance, map-based cloning of a diacylglycerol acyltransferase (DGAT) that catalyzes the final step in the glycerol biosynthetic pathway allowed the selection of a new protein variant influencing oil content and composition in maize seeds.[28]

For this reason, this study aimed to (i) describe the complexity of the oil produced by this plant, (ii) describe the transcriptome complexity during seed maturation through the analysis of a cDNA library constructed from the mRNA of seeds at three different development stages, (iii) tag genes related to the

route of fatty acids and toxic compound metabolism and (iv) measure the difference of expression level between leaves and fruits of four genes (palmitoyl-acyl carrier protein thioesterase, 3-ketoacyl-CoA thiolase B, lysophosphatidic acid acyltransferase and geranyl pyrophosphate synthase) involved in some economically important traits. This information will be useful for the investigation of genetic diversity and for the selective breeding for traits in relation to the optimization of biodiesel production and low-phorbol varieties of *J. curcas*.

Materials and Methods

Construction of the cDNA library

Fruits from *Jatropha curcas* L. were collected at different development stages based on their size, ie, one, two and three cm. These fruit size differences can be associated with corresponding stages of seed development, ie, seed of 5 to 10 mm, 10 to 20 mm and 20 to 40 mm. We collected one gram of seeds corresponding to each of these stages mixed the samples together and extracted the total RNA using the TM kit RNAqueous Phenol Free Total RNA Isolation (Ambion) according to manufacturer recommendations. Total RNA amount was quantified with a GeneQuant Pro spectrophotometer (Amersham Biosciences). The first cDNA strand was generated with the CDS III/3′ PCR Primer and MMLV Reverse Transcriptase (Clontech), according to the protocol of the SMARTTM PCR cDNA Synthesis Kit. The second cDNA strand was synthesized using the CDS III/3′ PCR Primer and 50X Advantage 2 Polymerase Mix. The resulting double-stranded cDNAs were size fractionated according to the average size of 1000 pb with a CHROMA SPIN + TE 1000 (Clontech) column to select larger fragments. We then selected the best three fractions with cDNA size above 500 bp, mixed them and ligated the cDNAs into the pTZ57R/T (Fermentas) expression vector. After this step, the ligation mix was used to transform *E. coli* MegaX DH10BTMT1R eletrocomp cells (Invitrogen). Finally, a cDNA library was constructed, out of which we selected about 2500 recombinant clones. The cells were stored at −80 °C in 96 wells plates with 16% glycerol.

Sequencing of cDNAs

The plates were inoculated with a 96 pins replicator into 96 deepwell plates containing 1.2 ml LB with ampicilin and incubated at 37 °C for 15 hr. DNA from bacterial cultures was extracted and purified according to Sambrook et al.[29] The sequencing reactions were processed according to recommendations of the DYEnamicTM ET Dye Terminator Cycle sequencing kit for MegaBACE 1000 DNA Analysis Systems (Amersham Biosciences). We used a T7 promoter sequencing primer (5′-TAATACGACTCACTATAGGG-3′) for the first reaction.

EST processing

Expressed Sequence Tags (EST) were processed stepwise in order to (i) extract their regions corresponding to PHRED quality ≥10, (ii) trim out vector sequences with CROSS-MATCH,[30] (iii) trim out polyA and polyT tail sequences with home-made Perl scripts, (iv) eliminate sequence redundancy by contig assembling with CAP3,[31] (v) annotate them for putative functions by comparison to GeneOntology (GO, http://www.geneontology.org/) using Blast2GO[32] and selecting homologous pairs with $E \leq 0.0001$, identity ≥ 40% and the homologous region ≥40 amino acids. These 931 ESTs were submitted to the dbEST section of GenBank with accession numbers GT228436-GT229366.

We also download the 13193 ESTs of *J. curcas* from Genbank (Rel. 174, Dec 14, 2009) using ACNUC[33] and assemble them into contigs with CAP3 in order to eliminate sequence redundancy. By this way, we obtained a sequence sample of 5841 ESTs that we compared to the protein sequences of Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/, Dec. 2009) using BLASTX[34] with $E \leq 0.0001$, identity ≥ 40% and the homologous region ≥40 amino acids. EST annotation by function transfer from homologous sequence depends more or less subjectively on the confidence that is given to the quality of the alignment. Many attempts were done to define the score threshold at which function transfer can be performed however when qualitative evaluation is needed it still depends on case to case analysis. Actually, enzyme function can be conserved at 3D level rather than sequence level with the consequence that true positives can be lost by filtering them out in reference to the classification threshold. The attribution of enzymatic function to proteins by experimental analyses is tedious and the only feasible strategy available at moment is to compare sequences to the

KEGG repository that included a total of ~5 million sequences, at time of publication of this article (Dec 1, 2009), whose 955689 include EC annotations (19%). If any false positives would emerge from such comparison, it is always possible to filter them out at some later steps.

For the reasons outlined above, we compared the homology of the 5841 non-redundant ESTs from GenBank to the KEGG accessions including an EC number and selected the homologies that mapped (http://www.genome.jp/kegg/tool/color_pathway.html) to the metabolism of fatty acids (map 61, 62, 71, 590, 591, 592, 1040), lipids (maps 561, 564, 565), terpenes (maps 900, 902, 904, 909), alkaloids (maps 901, 950, 1063, 1064, 1065, 1066), quinones (map 130), drugs (maps 982, 983) and hormones (maps 150, 905, 1070).

We compared (BLASTX) the 5841 non-redundant ESTs from GenBank (Rel. 174, Dec 14, 2009) with the KEGG sequences annotated with EC numbers (955689) used as the reference dataset.

Measure of gene expression by real-time PCR

The level of expression of palmitoyl-acyl carrier protein thioesterase, 3-ketoacyl-CoA thiolase B, lysophosphatidic acid acyltransferase and geranyl pyrophosphate synthase was compared to that of actin in a leaf sample and three seed samples corresponding to the three fruit stages described under "Construction of the cDNA library" (see above). The RNA was extracted from seeds and leaves of *Jatropha curcas* L. as described under "Construction of the cDNA library" (see above) and treated with DNase (Fermentas) to remove contaminating DNA. Approximately 5 µg of total RNA has been used to perform the first strand synthesis by real-time PCR of each of the four samples (seeds and leaves). The five primer pairs were designed according to the conserved domains of each protein and the corresponding annealing temperature calculated with Primer 3.0.

The quality of amplicons was first checked with genomic DNA under conditions suitable to quantitative real-time PCR. The reference gene used as control to measure the constitutive expression was the actin gene from *Ricinus communis* L. The real-time PCR mix was 10 µl SYBR (SYBR Green PCR Master Mix, Applied Biosystems), 4 µl per primer (200 nM), 4 µl cDNA (10 ng/µl) and 2 µl Milli-Q water completing the total volume to 20 µl. The primers were: (i) 5′-GAACTGGAATGGTGAAGGCT-3′ (forward) and 5′-ACATAGGCATCCTTCTGACC-3′ (reverse) for actine; (ii) 5′-GGAAGATTCTACACAGGCGT-3′ (forward)and5′-TGGAGGAAGGTGCTGAGATA-3′ (reverse) for palmitoyl-acyl carrier protein thioesterase; (iii) 5′-GTAGAGATTGTCGCTTCGGA-3′ (forward) and 5′-GGCATTACACAGCTCATCAC-3′ (reverse) for 3-ketoacyl-CoA thiolase B; (iv) 5′-CATACATGCTACCGCCATCT-3′ (forward) and 5′-TGATACGAGCAGCATCTCCT-3′ (reverse) for lysophosphatidic acid acyltransferase and (v) 5′-TCCATCACGATTACGTCGCT-3′ (forward) and 5′-AACAAGCCACTGAACCTCCA-3′ (reverse) for geranyl pyrophosphate synthase. The real-time PCR was carried out with a 7300 Real Time PCR System (Applied Biosystems) under the following conditions: (i) initial denaturation at 95 °C, (ii) 40 cycles of 15 sec at 95 °C followed by one min at 62 °C and (iii) a dissociation step for checking amplicon quality.

The average and standard deviation were obtained from three replicates. The average of actin expression of seeds across the fruit stages was, first, normalized according to that of the leave sample. The linear correction performed according to the ΔΔCt method[35] was, then, applied to find the expression level of each gene in leaves and seeds. Finally, the average level of expression of each gene in seeds was divided by its corresponding value in leaves to obtain the multiplying factor associated to the over-expression of these genes in seeds at the three fruit stages.

Analysis of fatty acid composition of oil from *J. curcas*

In addition to transcriptome characterization, we analyzed the fatty acid composition of *J. curcas* oil in order to better understand the level of complexity of the fatty acids produced, expecting that this information would facilitate the interpretation of fatty acid biosynthetic pathways.

The derivatization of a sample of *J. curcas* oil kindly provided by BIONASA Combustível Natural S.A. was obtained by successive esterification and transesterification to warrant complete conversion of free fatty acid and triacylglycerides (TAG) into methyl-esters of fatty acids. Methyl-esters were then quantified by gas chromatography (GC). For this, we

used a 30 m capillary column of 0.32 mm internal diameter filled with a stationary phase of 0.25 µm polyethylene glycol. Helium was used as carrying gas with a flow of 1.5 ml/min, a pressure of 13 psi and a split flow of 50 ml/min. The oven was set to a fixed temperature of 200 °C. The injector and the flame ionization detector were set to 250 °C. The fatty acids were identified by comparing the total peak area of alkyl-esters to that of the methyl-heptadecane used as a reference according to the formula: $C = (100*A_{ester}*C_{control}*V_{control})/(A_{control}*M_{sample})$, where C is the fatty acid concentration (w/w); A_{ester} is the area of the ester peak of the corresponding fatty acid; $A_{control}$ is the peak area of the control (methyl-heptadecane); $C_{control}$ is the concentration (mg/ml) of methyl-heptadecane; and $V_{control}$ is the solution volume (ml) of methyl-heptadecane.

Results

Assessing the transcriptome and fatty acid composition of *J. curcas* seeds

After analyzing the fatty acid composition of *Jatropha curcas* oil, we found that it is 0.07% myristic (14:0), 14.42% palmitic (16:0), 0.74% palmitoleic (16:1), 6.02% stearic (18.0), 41.13% oleic (18:1), 36.93% linoleic (18:2), 0.18% linolenic (18:3) and 0.08% arachidic (20:0). This shows that the oil complexity and saturation level in *J. curcas* are very low, with 98.6% being C16:0 (14.4%), C18:0 (6.0%), C18:1 (41.2%), C18:2 (37.0%) and the rest being negligible (Fig. 1).

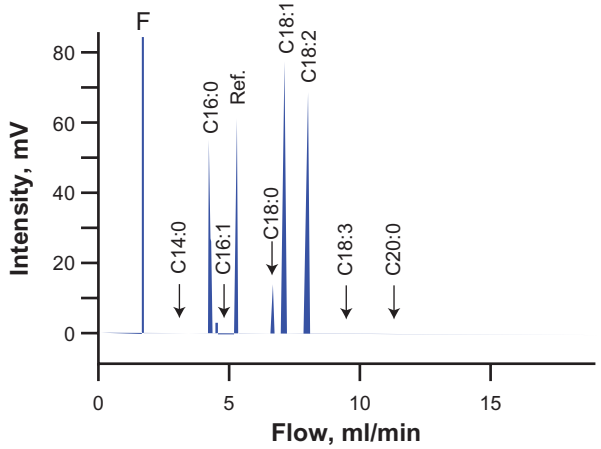

Figure 1. Gas chromatogram of *J. curcas* oil. F represents the perturbation associated with the solvent front. Ref. is for methyl-heptadecane, which is used as a reference.

The double-stranded cDNA fragments obtained from the total RNA extract from seeds at three development stages (Fig. 2A) ranged from 300 to 2000 bp (Fig. 2B), as did the inserted fragments after cloning (Fig. 2C).

We sequenced 2200 cDNA clones, from which we recovered 1337 (60%) reads after quality control and trimming. Among them, 546 were clustered into 140 contigs while 791 remained singlets. Our final sample was therefore 931 non-redundant expressed sequence tags (EST) with PHRED quality higher than 10. The 140 contigs had an average size of 569 bp and the 791 singlets had an average size of 379 bp. The majority (64%) of contigs was made up of only two reads. The rest of the contigs were made up of three to eight reads, except four of them that were made up of a higher read number, ie, 9, 11, 12 and 28. This shows that our library demonstrated a low level of sequence redundancy.

Functional characterization with BLAST2GO

We found 440 ESTs with homology to GO accessions, which allowed them to be grouped into three functional categories: those related to *Biological Processes* (Fig. 3A), *Molecular Functions* (Fig. 3B), and *Cellular Components* (Fig. 3C). The *Biological Process* category could be itself subdivided into 13 categories, with the majority (29%) of ESTs assigned to the "Metabolic Process" subcategory. Considering

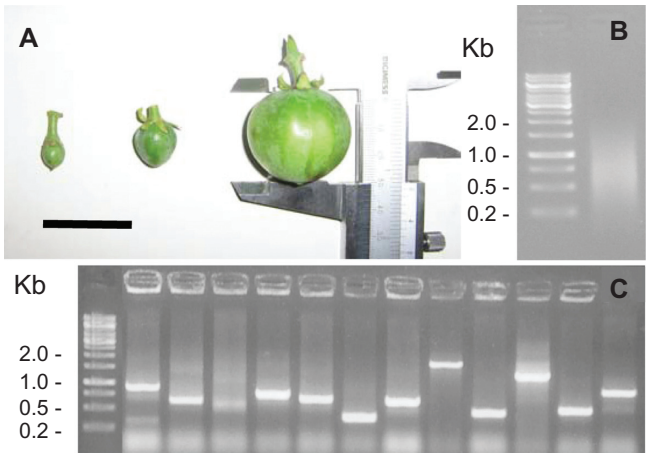

Figure 2. Steps toward library construction. Three different fruit stages were selected for seed RNA extraction. The black bar is three cm (**A**). Range of ds-cDNAs size. The ladder is λ phage digested with *Hind*III + *Eco*RI (**B**). Size distribution of cDNA fragments after cloning and PCR amplification. The ladder is λ phage digested with *Hind*III + *Eco*RI (**C**).

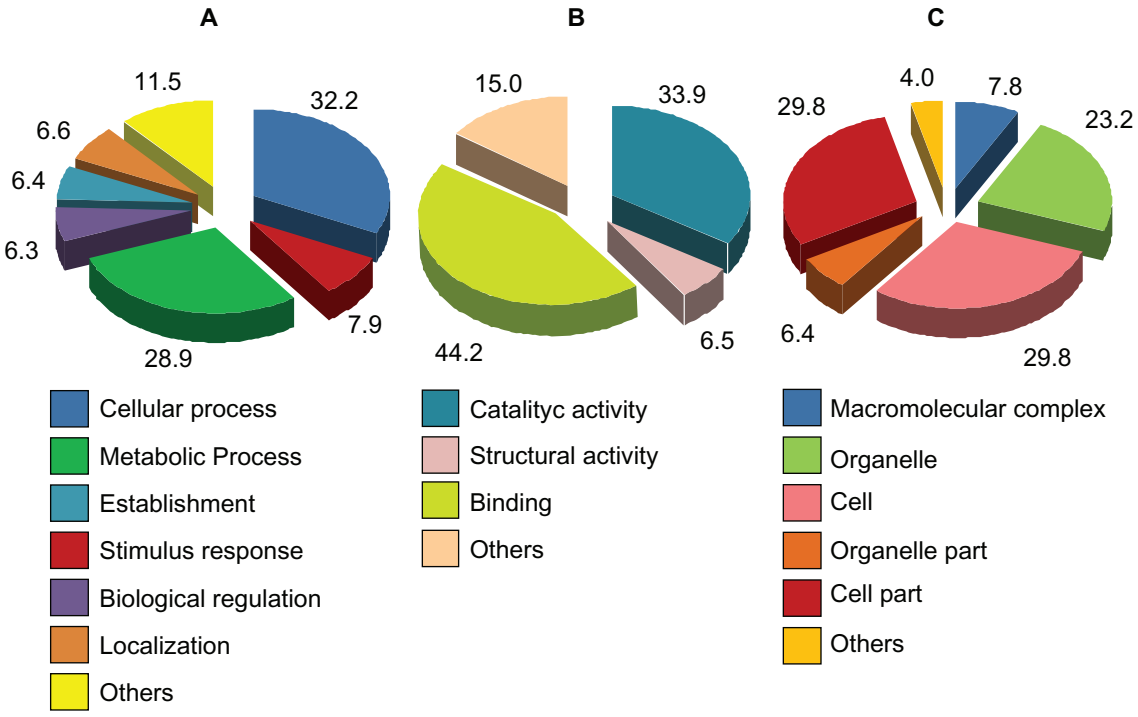

Figure 3. *J.* curcas ESTs BLAST2GO annotation: Biological Process (**A**); Molecular Function (**B**); Cellular Component (**C**). The numbers are proportions (%).

Molecular Functions, the ESTs were assigned to ten categories with the most frequent corresponding to "binding" (44%) and "catalytic activity" (34%). When grouped according to the *Cellular Component* category, the ESTs were assigned to nine subcategories with 30% and 23% covering two GO terms: "Cell Part" and "Organelle", respectively. The ESTs annotated involved a total of 74 metabolic pathways, including those involved in the biosynthesis of fatty acids, steroid biosynthesis, biosynthesis of unsaturated fatty acids and metabolism of α-linolenic acid were identified.

KEGG pathway annotations

Annotation using BLASTX and KEGG allowed the classification of ESTs in agreement with their function in the context of specific metabolic pathways.

We found homologous ESTs for most enzymes involved in the fatty acid biosynthesis (maps 61, 71, Fig. 4), ie, acetyl-CoA carboxylase, 3-oxoacyl-[acyl-carrier-protein] reductase, enoyl reductase, fatty acid synthase (Fig. 4, Table 1 of supplementary materials) and in particular for the enzymes involved in the last steps of oleic, stearic and palmitic acid synthesis (Table 2 of supplementary materials), ie, oleoyl-[acyl-carrier-protein] hydrolase, linoleic acyl-[acylcarrier-protein] (map 61) This was expected since oleic, linoleic and palmitic acids are the three major fatty acids from oil of *J. curcas* (Fig. 1). We also found homologies for (i) fatty acid elongation in mitochondria (map 62, Table 1 of supplementary materials), (ii) double bond hydration/dehydration (map 950), (iii) oxidoreductase (map 951); (iv) glycerol acylation (map 564) and (v) acylglycerol modification (map 565) (Table 3 of supplementary materials).

In addition to fatty acid, lipid and glycerol pathways, the metabolism pathways of active compounds (terpenes, quinones and alkaloids) and hormones are interesting to consider here. On the one hand, understanding the regulation of active compounds is critical for the control of toxic secondary metabolites such as phorbol. On the other hand, understanding biochemical bases of regulatory mechanisms induced by hormones in the processes of organogenesis, defense and fruit maturation is obviously necessary to improve specific agronomical traits.

When considering the pathway of terpenoid backbone synthesis (map 900), we found representations related to enzymatic function downstream of isopentenyl pyrophosphate, suggesting that the activation of this pathway occurs through geranyl pyrophosphate. Some representations were also

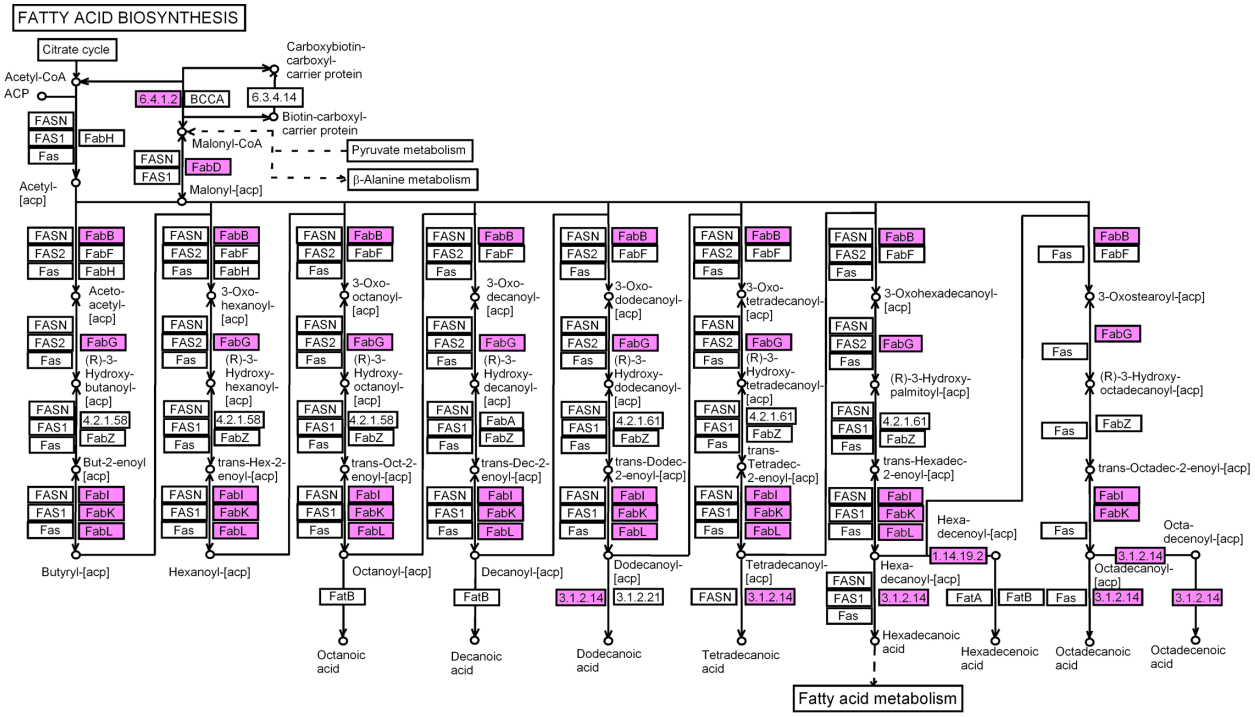

Figure 4. Fatty acids biosynthetic pathway (map 61). The pink boxes are for KEGG ECs that have homologies to *J. curcas* ESTs.

found in the biosynthesis of (i) monoterpenoids (map 902) and diterpenoids (map 904) (Table 4 of supplementary materials), (ii) quinones (map 130) and (iii) alkaloids (maps 901, 950) (Table 5 of supplementary materials). The drug metabolism was also found to be active (maps 982, 983) probably in relation to the metabolism of secondary metabolites (Table 6 of supplementary materials).

We found some ESTs putatively involved in auxine biosynthesis, cytokinin and brassinosteroid synthesis (map 905). Other enzymes involved in regulation were from the androgen and estrogen metabolism (map 150) that are known to regulate many processes of organogenesis and plant defenses (Table 7 of supplementary materials).

Assessing gene expression

Several ESTs from GenBank showed a high level of homology (>80%) over at least 150 bp with palmitoyl-acyl carrier protein thioesterase, 3-ketoacyl-CoA thiolase B, lysophosphatidic acid acyltransferase and geranyl pyrophosphate

Table 1. Features of genes and amplicons assessed by quantitative real-time PCR.

Genes	EC (KEGG)	Map, nb (KEGG)	EST (GenBank)	Size, bp	Hml, aa[1]	Id, %[2]	Amplicon, bp	Ta, °C[3]
Actin	----	----	AY360221.ACT	---	---	----	124	51
Palmitoyl-acyl carrier protein thioesterase	3.1.2.14	61	FM893767, FM891233	419	106	90	183	51
3-ketoacyl-CoA thiolase B	2.3.1.16	1070, 71, 592, 1040, 62	FM892914	460	88	90	104	50
Lysophosphatidic acid acyltransferase	2.3.1.51	564, 561, 565	FM889020	395	54	83	171	48
Geranyl pyrophosphate synthase	2.5.1.30	900	GT229261	417	102	83	185	49

Notes: [1]"Hml" is for the size of the similar region of homologous pairs between GenBank EST and KEG sequences; [2]"Id" is for the level of identity of homologous regions between EST and KEGG sequences; [3]"Ta" is for annealing temperature of primers.

synthase genes (Table 1). This strongly suggests that these ESTs are involved in the same function as their homologous from KEGG. These genes are involved in the pathways of *fatty acid biosynthesis*, *biosynthesis of unsaturated fatty acids*, *glycerolipid metabolism* and *terpenoid backbone biosynthesis*, respectively.

In some cases, EST sequences tagging for one gene were showing nucleotide polymorphism, which is attractive for their use as DNA probe in breeding programs. The level of gene expression detected by quantitative real-time PCR was systematically higher in seeds at the three fruit stages than in leaves (Fig. 5A,B). The maximum of gene expression for geranyl pyrophosphate synthase and lysophosphatidic acid acyltransferase was found in seeds of fruits at stage 1. This was particularly strong for geranyl pyrophosphate synthase whose level o gene expression in seeds of fruits at stage 1 was ~25 times that found in leaves. The level of over-expression of this gene decreased to ~11 in seeds of fruits at stage 2 and to ~6 in seeds of fruits at stage 3. Even if the level of over-expression of lysophosphatidic acid acyltransferase was ~3 in seeds of fruits at stage 1, it was the only gene whose level of expression was lower than that of actin in fruits as well as in leaves. Its level of expression in seeds of fruits at stage 2 and 3 decreased to the same value as that found in leaves. The 3-ketoacyl-CoA thiolase B was over-expressed by a factor ~2 at stages 1 to 2 and ~5 at stage 3. Finally, palmitoyl was over-expressed in fruits by a factor ~5, but the profile of gene expression was flat across the 3 stages.

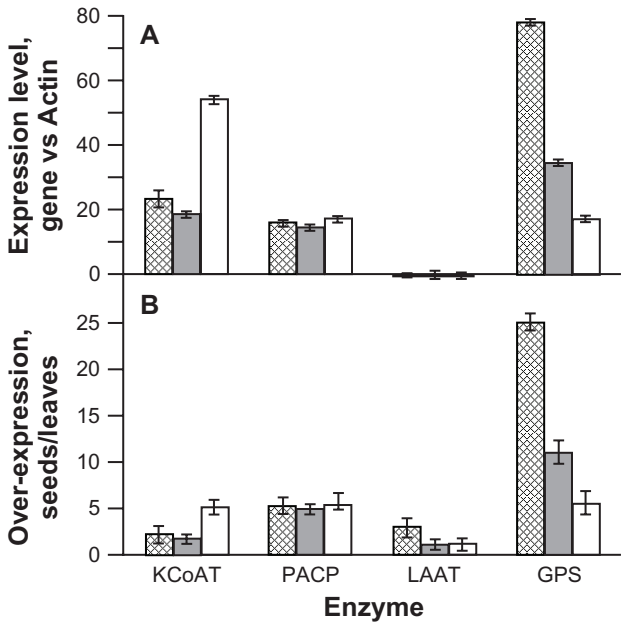

Figure 5. Assessment of the expression profile of four genes by real-time PCR in seeds of *J. curcas* at three stages of fruit maturation. Crossed, gray and white bars are for stages 1, 2 and 3, respectively. The level of expression in relation to actin is given in *y* and the enzyme function in x; KCoAT, PACP, LAAT and GPS are abbreviations for "3-ketoacyl-CoA thiolase B, palmitoyl-acyl carrier protein thioesterase" lysophosphatidic acid acyltransferase and geranyl pyrophosphate synthase, respectively. The level of gene expression assessed by quantitative real-time PCR is given in seeds (**A**) and in seeds compared to leaves (**B**). The bar intervals are for the standard deviations.

Discussion

As previously mentioned, *Jatropha curcas* L. is a promising crop; however, this species will need a minimum of 15 years of breeding before reaching a level of domestication comparable to other industrial crops. Among the traits that can be considered for selective breeding, characteristics such as energy storage, fatty acids synthesis, disease resistance, toxic compound synthesis, flowering synchronization, dioecy, apomixia, fruit size, tree branching, and tree size are obviously of priority.

The low genetic variability that has been found until now in *J. curcas* using RAPD and AFLP is surprising, especially if one generally considers that trees have larger genetic diversity within their populations than herbaceous plants.[36] It has been proposed that the center of origin of the species has not been identified yet. An alternative proposition was that *J. curcas* could perform autogamy despite allowing allogamy. The low variability that is found with RAPD probes (Jaccard coefficient > 0.85)[37] suggests such a type of sexuality. However, this hypothesis is not supported by the experiments of controlled pollination with *J. curcas*. Juhász et al[38] found that when self-pollination was processed manually, the rates of fruit weight, seed size and seed number per fruit was not significantly different from those observed with natural cross-pollinated plants. However, in the natural conditions of Brazil the same authors found that the rate of fruit formation was reduced by a factor two when natural self-pollination was forced simply because masculine and feminine flowers rarely open at the same time in the same inflorescence. Similar conclusions were reached by Abdelgadir et al[39] in African conditions. Actually, *J. curcas* could simply have experienced a population bottle-neck in its past history. This possibility is supported by the fact

that low genetic variability is observed all over the world.[11,17,37]

SSR could be an alternative source of genetic variability. Using SSRIT, we found that ~23% of our EST sample contains short repeats distributed in tandem di-, tri-, tetra-, penta- and hexa-nucleotides, with the largest amount represented by the di-nucleotides $TC_{(n)}$ and $AG_{(n)}$. The polymorphism associated with these repeats is under investigation (data not shown).

Despite its small size, the metabolic coverage of our EST sample was large at least in the first steps of fruit maturation and characterized by a low level of sequence redundancy. The 5841 non-redundant ESTs of GenBank account for ~20% of the total sample of coding sequences; even if it is still a small proportion, it is enough to start looking for correlations with quantitative trait loci (QTL).

The link between protein function and ESTs that has been established in this study could be questioned since false positives are common to this type of methodology. However, the filter used (Expected ≤0.0001, identity ≥40% and homology size ≥40 aa) is standard when describing the proportions among metabolic activities of a transcriptome at a given stage. Actually, a homologous region larger than 40 amino acids is generally significant when associated to both expected rate lower than 0.0001 and identity larger than 40%. As can be seen from data of supplementary materials, homologies are often larger than 80% similarity. Of course, paralogous genes may always appear as false positives, but would ultimately be eliminated if not associated to a QTL.

An advantage of probes derived from mRNA is that they can be generated from different tissues at various development stages and therefore are highly effective for identifying genes that are differentially expressed during the life-cycle of an organism.[40,41]

Of course, a method is needed for selecting and mapping candidate loci associated with particular ESTs. Amplified fragment length polymorphism (AFLP)[42] combined with cDNA libraries can be applied to yield highly informative transcript-derived fragments (TDF) for mapping traits whose expression is time-dependent.

In a first step, Suárez et al[25] introduced the sequencing of ESTs from such TDFs and their locations within a genetic map from cassava. Later on, Quin et al[26] showed how to detect AFLP bands from the pattern of restriction of ESTs. This technique offers the advantage of allowing the exploitation of existing EST resources. More simply, SNPs can be investigated by *Denaturing Gradient Gel Eletrophoresis* (DGGE),[43] *Single-Stranded Conformational Polymorphism* (SSCP)[44,45] or directly by sequencing and would allow the screening of informative ESTs associated to economically relevant agronomic traits. On must also considers the non-coding regions associated to these traits since they normally bring the regulatory motives associated to these traits. In that respect, tagging BACs with EST probes in order to sequence them will be the next important issue of the ongoing work.

The use of the techniques just outlined should be effective in assisting breeding programs for the selection of qualitative as well as QTLs.[46] Here, we showed that quantitative real-time PCR allows the detection of genes that are significantly over-expressed in favorable tissues such as is the case of geranyl pyrophosphate synthase that is over-expressed by a factor ~25 in forming seeds and that is, therefore, a good candidate to tag a QTL associated to phorbol synthesis.

The chemical composition of *J. curcas* oil is rather simple, as it consists mainly of C18 (~84%) chains including one or two double bounds. The results that we found are consistent with those published by other authors, that is: 1.4% myristic (C14:0), 10.5%–15.6% palmitic (C16:0), 2.3%–9.7% stearic (C18:0), 40.8%–48.8% oleic (C18:1), 32.1%–44.4% linoleic (C18:2) and 0.4% arachidic (C20:0) acids.[47–52] Significant variation of oil content was described in Indian accessions.[11] Seed oil content is typically a QTL and genes from fatty acid biosynthesis are expected to tag this trait. The use of probes for fatty acid pathways and other pathways associated with energy management should help in the optimization of *J. curcas* selective breeding for oil quality and content, or at least to follow the rate of that trait when breeding for other traits of agronomical interest such as those listed before. In that respect, 3-ketoacyl-CoA thiolase B is a candidate to help in tagging QTLs associated to oil. Actually, its profile of expression matches the profile of fatty acid accumulation as described by NMR.[53] Following Annarao

et al,[53] the oil content raises in two major steps with a very clear transition occurring around stage IV when the oil content raises from 3 to 18% and TAGs from 30 to >90%. Fresh weight reaches its maximum (~1000 mg) at stage V, ie, stage 3 of our experiment, and then decreases until ~640 mg during stage VI and VII. These last two stages match fruit ripening, which is accompanied by a color change from green to yellowish. However, we found that the basal activity of this enzyme in leaves is also relatively high, which shows that it is involved in alternative functions that could interfere on the general performance of breeding individuals. By contrast, lysophosphatidic acid acyltransferase is not expected to be a marker of any utility for QTL tagging given its low activity level and its profile of expression that is inverted compared to the one expected, ie, that of 3-ketoacyl-CoA thiolase. The case of palmitoyl-acyl carrier protein thioesterase is particular due to its flat profile that does not match that of fatty acid accumulation in fruits. However, its over-expression along the three fruit stages shows that it is actively recruited during all the process of fruit maturation.

Fatty acids differ according to three characteristics: (i) the size of carbon chain, (ii) the unsaturation number and (iii) chemical moeities. The larger the size of the carbon chain, the larger the cetane number and the lubricity, but the higher the viscosity and risk of injector choking. On the other hand, the greater the degree of unsaturation, the larger the cetane number, but also the molecule instability and therefore the risk of polymerization and choking. However, unsaturation promotes soot emission and the abatement of soot emission due to oxygen (~10%) naturally present in the biodiesel might be counterbalanced by the rate of alkyl ester unsaturation (double bonds).[54] A large fraction of plants oils have fatty acid compositions similar to that of *J. curcas*, but with different relative proportions.[52] Sunflower, corn and soybean oils have higher proportion of C18:2 in comparison to C18:1. In *J. curcas*, the proportion between these two is slightly in favor of C18:1 because it mainly combines palmitate (16:0, ~14%) oleate (C18:1, ~41%) and linoleate (C18:2, ~37%), which is a good composition for esterification and transesterification into biodiesel.[55] C16:0, C18:0, C18:1, C18:2, C18:3 have melting points at 64, 70, 13, –9, –17 °C, respectively;[56] therefore, breeding for lower linoleate (C18:2) and higher oleate (18:1) fatty acids can be favored under tropical climates.

Map-based cloning of a diacylglycerol acyltransferase (DGAT) that catalyzes the final step in the glycerol biosynthetic pathway allowed the selection of a new protein variant affecting oil content and composition in maize seeds.[28] QTLs for C16:0, C18:0, C18:1, C18:2, C18:3, C20:1 and C22:1 were also described in rapeseed.[57] There is no reason why such approaches could not be applied to other aspects of this study. However, not all gene functions detected in this work are expected to be useful therefore screening and evaluation will be necessary to guide future studies. However, such screening can be based on some speculation concerning gene function, which is not the case with blinded probes. Among the interesting features of this investigation, we found several ESTs associated with ECs from the biosynthesis pathway of terpenes. An example of this is geranyl pyrophosphate synthase that is a key enzyme upstream the pathway of terpene biosynthesis and that we found over-expressed by a factor ~25 in forming seeds. This over-expression occurs at a fruit stage where terpene precursors are, indeed, expected to form. One may expect that selecting accessions of *J. curcas* with low rate of expression for this enzyme would correlate with low phorbol levels in the corresponding accessions.

Acknowledgements

K. Gomes is grateful to Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) for providing a student fellowship. N. Carels is grateful to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação Oswaldo Cruz (FIOCRUZ) for providing a research fellowship from the Centro de Desenvolvimento Tecnológico em Saúde (CDTS). This work received financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil (no. 471214/2006-0). We thank Dominique Garcia for helping with plant material and cDNA library preparation as well as Fernanda Amatto Gaiotto for help with microsatellite analysis. We regret the unexpected death of Prof. Julio Cascardo while he was at the top of his career. We are grateful for his continuous dedication to what he thought to be the best for everybody.

Author Contributions

KAG did the cDNA library and cDNA sequencing under the supervision of ASG and JCMC. KAG did the qRT-PCR experiments under the supervision of MAVS. IPL did the oil characterization under the supervision of RSC. NC wrote the project, did the EST processing and managed the project with JCMC. All the authors participated to the project.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Carels N. *Jatropha curcas*: A Review. In Advances in Botanical Research. Edited by: Kader JC, Delseny M: Elsevier 2009;50:39–86.
2. Berry EW. An eogene tropical forest in the Peruvian desert. *Proc Natl Acad Sci U S A*. 1929;15:345–6.
3. Heller J. Physic nut *Jatropha curcas L*. Promoting the conservation and use of underutilized and neglected crops, Institute of Plant Genetics and Crop Plant Research, Gatersleben, International Plant Genetic Resources Institute, Rome. no 01.1996. 1996.
4. Napier JA. The production of unusual fatty acids in transgenic plants. *Annu Rev Plant Biol*. 2007;58:295–319.
5. Cahoon EB, Shockey JM, Dietrich, CR, et al. Engineering oilseeds for sustainable production of industrial and nutritional feedstocks: Solving bottlenecks in fatty acid flux. *Curr Opin Plant Biol*. 2007;10:236–44.
6. Schwender J, Ohlrogge J, Shachar-Hill Y. A flux model of glycolysis and the oxidative pentose phosphate pathway in developing *Brassica napus* embryos. *J Biol Chem*. 2003;32:29442–53.
7. Fatland BL, Nikolau BJ, SyrkinWurtele E. Reverse genetic characterization of cytosolic acetyl-CoA generation by ATP-citrate lyase in Arabidopsis. *The Plant Cell*. 2005;17:182–203.
8. Hill RA. Terpenoids. In: The Chemistry of Natural Products. Edited by: Thompson RH, Chapman and Hall, London, UK, 1993, pp. 106–39.
9. Mellor H, Parker PJ. The extended protein kinase C superfamily. *Biochem J*. 1998;332:281–92.
10. Blumberg PM. Protein kinase C as the receptor for the phorbol ester tumor promoters: Sixth Rhoads Memorial Award Lecture. *Cancer Research*. 1988; 48:1–8.
11. Popluechai S, Breviario D, Mulpuri S, et al. Narrow genetic and apparent phenetic diversity in *Jatropha curcas*: initial success with generating low phorbol ester interspecific hybrids, 2009, hdl:10101/npre.2009.2782.1.
12. Raina AK, Gaikwad BR. Chemobotany of *Jatropha* species in India and further characterisation of curcas oil. *Journal of Oil Technology of India*. 1987;19:81–5.
13. Kaushik N, Kumar K, Kumar S, et al. Genetic variability and divergence studies in seed traits and oil content of Jatropha (*Jatropha curcas* L.) accessions. *Biomass and Bioenergy*. 2007;31:497–502.
14. Sunil N, Varaprasad KS, Sivaraj N, et al. Assessing *Jatropha curcas* L. germplasm in situ—A case study. *Biomass and Bioenergy*. 2008;32:198–202.
15. Sudheer Pamidiamarri DVN, Pandya N, Reddy MP, et al. Comparative study of interspecific genetic divergence and phylogenic analysis of genus *Jatropha* by RAPD and AFLP. *Mol Biol Rep*. 2008a, doi:10.1007/s11033-008-9261-0. http://www.springerlink.com/content/m642ml6440657117/.
16. Sudheer Pamidimarri DVN, Singh S, Mastan SG, et al. Molecular characterization and identification of markers for toxic and non-toxic varieties of *Jatropha curcas* L. using RAPD, AFLP and SSR markers. *Mol Biol Rep*. 2008b, doi:10.1007/s11033-008-9320-6.
17. Ranade SA, Srivastava AP, Rana TS, et al. Easy assessment of diversity in *Jatropha curcas* L. plants using two single-primer amplification reaction (SPAR) methods. *Biomass and Bioenergy*. 2008;32:533–40.
18. van de Loo FJ, Broun P, Turner S, et al. An oleate 12 hydroxylase from *Ricinus communis* L. is a fatty acyl desaturase homolog. *Proc Natl Acad Sci U S A*. 1995;92:6743–7.
19. Cahoon EB, Kinney AJ. The production of vegetable oils with novel properties: using genomic tools to probe and manipulate plant fatty acid metabolism. *Eur J Lipid Sci Technol*. 2005;107:239–43.
20. Alagna F, D'Agostino N, Torchia L, et al. Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics*. 2009;10:399, doi:10.1186/1471-2164-10-399.
21. Holter NS, Maritan A, Cieplak M, et al. Dynamic modeling of gene expression data. *Proc Natl Acad Sci U S A*. 2000;98:1693–8.
22. White JA, Benning C. Genomic approaches towards the engineering of oil seeds. *Plant Physiol Biochem*. 2001;39:263–70.
23. Chen GQ, Turner C, He X, et al. Laudencia-Chingcuanco, D. Expression profiles of genes involved in fatty acid and triacylglycerol synthesis in castor bean (*Ricinus communis* L.). *Lipids*. 2007;42:263–74.
24. Nelson DR, Schuler MA, Paquette SM, et al. Comparative genomics of rice and Arabidopsis. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. *Plant Physiol*. 2004;135:756–72.
25. Súarez MC, Bernal A, Gutiérrez J, et al. Developing expressed sequence tags (ESTs) from polymorphic transcript-derived fragments (TDFs) in cassava (*Manihot esculenta* Crantz). *Genome*. 2000;43:62–7.
26. Quin L, Prins P, Jones JT, et al. GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP. *Nucleic Acids Res*. 2001;29:1616–22.
27. Zamir D. Plant breeders go back to nature. *Nat Genet*. 2008;40:269–70.
28. Zheng P, Allen WB, Roesler K, et al. A phenylalanine in DGAT is a key determinant of oil content and composition in maize. *Nat Genet*. 2008; 40:367–72.
29. Sambrook J, Fitsch EF, Maniatis T. Molecular cloning: a laboratory manual. Cold Spring Harbor, Cold Spring Harbor Press 1989.
30. Ewing B, Green P. Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 1998;8:186–94.
31. Huang X, Madan A. CAP3: A DNA Sequence Assembly Program. *Genome Res*. 1999;9:868–77.
32. Conesa A, Gotz S, Garcia-Gomez JM, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005;21:3674–6.
33. Gouy M, Gautier C, Attimonelli M, et al. ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Comput Appl Biosci*. 1985;1:167–72.
34. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–402.
35. Bookout AL, Mangelsdorf DJ. Quantitative real-time PCR protocol for analysis of nuclear receptor signalling pathways. *Nuclear Receptor Signaling*. 2003;1:1–7.
36. Petit RJ, Hampe A. Consequences of being a tree. *Annu Rev Ecol Evol Syst*. 2006;37:187–214.
37. Sun QB, Li LF, Wu GJ, et al. SSR and AFLP markers reveal low genetic diversity in the biofuel plant *Jatropha curcas* in China. *Crop Sci*. 2008; 48:1865–70.
38. Juhász ACP, Pimenta S, Soares BO, et al. Floral biology and artificial polinization in physic nut in the north of Minas Gerais state, Brazil. *Pesq Agropec Bras*. 2009;44:1073–7.
39. Abdelgadir HA, Johnson SD, Van Staden J. Approaches to improve seed production of *Jatropha curcas* L. *South African Journal of Botany*. 2008; 74:359.
40. Bachem CWB, Van der Hoeven RS, de Brujin SM, et al. Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: Analysis of gene expression during potato tuber development. *Plant J*. 1996;9:745–53.

41. Andersen JR, Lübberstedt T. Functional markers in plants. *Trend Plant Sci*. 2003;8:554–60.
42. Vos P, Hogers R, Bleeker M, et al. AFLP: A new technique for DNA fingerprinting, *Nucleic Acids Res*. 1995;23:4407–14.
43. Fischer S, Lerman L. DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: Correspondence with melting theory. *Proc Natl Acad Sci U S A*. 1983;80:1579–83.
44. Orita M, Iwahana H, Kanazawa H, et al. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc Natl Acad Sci U S A*. 1989;86:2766–70.
45. Hongyo T, Buzard G, Calvert R, et al. Cold SSCP: a simple, rapid and non-radioactive method for optimized single-strand conformation polymorphism analyses. *Nucl Acids Res*. 1993;21:3637–42.
46. Boventius H, Weller JI. Mapping and analysis of dairy cattle quantitative trait loci by maximum likelihood methodology using milk protein genes as genetic markers. *Genetics*. 1994;137:267–80.
47. Foidl N, Foidl G, Sanchez M, et al. *Jatropha curcas* L. as a source for the production of biofuel in Nicaragua. *Bioresour Technol*. 1996;58:77–82.
48. Nahar NM, Waris A, Azam MM. Prospects and potential of fatty acid methyl esters of some non-traditional seed oils for use as biodiesel in India. *Biomass and Bioenergy*. 2005;29:293–302.
49. Martínez-Herrera J, Siddhuraju P, Francis G, et al. Chemical composition, toxic/antimetabolic constituents, and effects of different treatments on their levels, in four provenances of *Jatropha curcas* L. from Mexico. *Food Chem*. 2006;96:80–9.
50. Adebowale KO, Adedire CO. Chemical composition and insecticidal properties of the underutilized *Jatropha curcas* seed oil. *African Journal of Biotechnology*. 2006;5:901–6.
51. Kumar A, Sharma S. An evaluation of multipurpose oil seed crop for industrial uses (*Jatropha curcas* L.): A review. *Industrial Crops and Products*. 2008;28:1–10.
52. Singh SP, Singh D. Biodiesel production through the use of different sources and characterization of oils and their esters as the substitute of diesel: A review. *Renewable and Sustainable Energy Reviews*. 2010;14:200–16.
53. Annarao S, Sidhu OP, Roy R, et al. Lipid profiling of developing *Jatropha curcas* L. seeds using 1H NMR spectroscopy. *Bioresource Technology*. 2008;99:9032–5.
54. Klein-Douwel RJH, Donkerbroek AJ, van Vliet AP, et al. Soot and chemiluminescence in diesel combustion of bio-derived, oxygenated and reference fuels. *Proceedings of the Combustion Institute*. 2009;32:1–17, doi:10.1016/j.proci.2008.06.140.
55. Refaat AA. Correlation between the chemical structure of biodiesel and its physical properties. *Int J Environ Sci Tech*. 2009;6:677–94.
56. Cahoon EB, Schmid KM. Metabolic engineering of the content and fatty acid composition of vegetable oils. *Advances in Plant Biochemistry and Molecular Biology*. 2008;1:161–200.
57. Zhang JF, Qi CK, Pu HM, et al. QTL identification for fatty acid content in rapeseed (*Brassica napus* L.). *Acta Agron Sin*. 2008;34:54–60.

Supplementary Tables

Table S1. ESTs that tag for enzymatic functions putatively involved in the metabolism of saturated fatty acid of *J. curcas*.

Pathway	EST	Reads/contig	KEGG homolog	Definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
Fatty acid biosynthesis (00061)									
	*Contig376	FM891097, FM892258, FM891175, FM893271, FM892235, FM893697	rcu:RCOM_1431360	Short chain dehydrogenase	1.1.1.100	1040, 61	293	181	88
	Contig397	FM893767, FM891233	rcu:RCOM_0925410	Palmitoyl-acyl carrier protein thioesterase	3.1.2.14	61	419	106	90
	*Contig895	FM896100, FM888313, FM889937	rcu:RCOM_0251360	Ketoacyl-ACP Synthase III	2.3.1.41	61	400	158	91
	Contig1093	GH295922, FM888387, GH296315	rcu:RCOM_0633300	Ketoavyl-ACP synthase	2.3.1.41	61	469	78	78
	FM887191		rcu:RCOM_0633300	50 kDa ketoavyl-ACP synthase	2.3.1.41	61	469	97	73
	FM887714		rcu:RCOM_1403260	Malonyl-CoA : ACP acyltransferase	2.3.1.39	61	400	124	75
	FM888027		rcu:RCOM_0710230	46 kDa ketoavyl-ACP synthase	2.3.1.41	61	554	52	98
	FM889551		rcu:RCOM_1636080	2,4-dienoyl-CoA reductase	1.1.1.100	1040, 61	285	109	88
	FM891538		rcu:RCOM_1431360	Short chain dehydrogenase	1.1.1.100	1040, 61	293	56	80
	FM895510		rcu:RCOM_0097860	Enoyl-ACP reductase precursor	1.3.1.9	61	359	121	76
	FM895824		pop:POPTR_555276	Enoyl-[acyl-carrier protein reductase	1.3.1.9	61	394	162	82
	GO247153		olu:OSTLU_12103	3-oxoacyl-[acyl-carrier protein reductase	1.1.1.100	1040, 61	273	44	77
	GT228627		rcu:RCOM_1081890	Ketoacyl-ACP Reductase	1.1.1.100	1040, 61	328	188	75
Fatty acid elongation in mitochondria (00062)									
	Contig332	FM890615, FM896803, FM892730	ath:AT3G06860	Multifunctional protein MFP2	1.1.1.211	1070, 71, 1040, 62	725	258	84
	FM892914		ec00062	3-ketoacyl-CoA thiolase B	2.3.1.16	1070, 71, 592, 1040, 62	460	88	90
	GT229218		rcu:RCOM_0696050	Enoyl-CoA hydratase, mitochondrial precursor	4.2.1.17	1070, 71, 592, 1040, 62	389	195	78

Fatty acid metabolism (00071)								
Contig332	FM890615, FM896803, FM892730	ath:AT3G06860	Multifunctional protein MFP2	1.1.1.211	1070, 71, 1040, 62	725	258	84
*Contig433	FM891676, FM891199	rcu:RCOM_0914000	Acyl-CoA synthetase	6.2.1.3	71	694	74	78
*Contig516	FM892437, FM890803, FM892732, GO247603, FM893815, FM891552	npu:Npun_F3727	Ferredoxin 2Fe-2S	1.18.1.3	71	99	96	68
Contig1015	GH295632, GH295820, GH296400, GH296214, GH295687, GH296081, GH296508, GH296024, GH295630,GH295630, GH296026, GH295757, GH296151	pmj:P9211_14071	Ferredoxin	1.18.1.3	71	99	98	61
*Contig1435	GT229265, GO247630, FM894566, FM890476, GO247018, FM889828	dre:436918	Dodecenoyl-CoA delta-isomerase	5.3.3.8	71	357	68	47
FM888666		rcu:RCOM_1032290	Long-chain-fatty-acid CoA ligase	6.2.1.3	71	697	127	83
FM892914		rcu:RCOM_1020840	3-ketoacyl-CoA thiolase B	2.3.1.16	1070, 71, 592, 1040, 62	460	88	90
FM894424		rcu:RCOM_1437260	Cytochrome P450	1.14.14.1	1063	632	180	86
FM894637		vvi:100257352	Glutaryl-CoA dehydrogenase	1.3.99.7	71	446	155	89
FM894979		pop:POPTR_832274	Acyl-CoA oxidase	1.3.3.6	1070, 71, 592, 1040	689	196	92
GO247595		dre:436918	Dodecenoyl-CoA delta-isomerase	5.3.3.8	71	357	68	47
GT229218		rcu:RCOM_0696050	Enoyl-CoA hydratase, mitochondrial precursor	4.2.1.17	1070, 71, 592, 1040, 62	389	195	78

Notes: [1]Sz is for the size of the KEGG homologous protein in amino acids; [2]Homol is for the size of the homologous region in amino acids; [3]Id% is for the pourcentage of identity of amino acid homologous pairs; *A contig with an asterisk in front means that its reads show nucleotide polymorphism.

Table S2. ESTs that tag for enzymatic functions putatively involved in the metabolism of unsaturated fatty acid of *J. curcas*.

Pathway	EST	Reads/contig	KEGG homolog	Gene definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
Biosynthesis of unsaturated fatty acids (01040)									
	Contig332	FM890615, FM896803, FM892730	ath:AT3G06860	Multifunctional protein MFP2	1.1.1.211	1070, 71, 1040, 62	725	258	84
	*Contig376	FM891097, FM89225, FM891175, FM893271, FM892235, FM893697	rcu:RCOM_1431360	Short chain dehydrogenase	1.1.1.100	1040, 61	293	181	88
	*Contig533	FM892613, FM894348, FM895006, FM890906, FM892529	rcu:RCOM_0472100	Short chain alcohol dehydrogenase	1.1.1.100	1040, 61	280	277	78
	FM889551		rcu:RCOM_1636080	2,4-dienoyl-CoA reductase	1.1.1.100	1040, 61	285	109	88
	FM891538		rcu:RCOM_1431360	Short chain dehydrogenase	1.1.1.100	1040, 61	293	56	80
	FM892914		rcu:RCOM_1020840	3-ketoacyl-CoA thiolase B	2.3.1.16	1070, 71, 592, 1040, 62	460	88	90
	FM894979		pop:POPTR_832274	Acyl-CoA oxidase	1.3.3.6	1070, 71, 592, 1040	689	196	92
	GO247153		olu:OSTLU_12103	3-oxoacyl-[acyl-carrier protein reductase	1.1.1.100	1040, 61	273	44	77
	GT228627		rcu:RCOM_1081890	Ketoacyl-ACP Reductase	1.1.1.100	1040, 61	328	188	75
	GT229218		rcu:RCOM_0696050	Enoyl-CoA hydratase, mitoc. precursor	4.2.1.17	1070, 71, 592, 1040, 62	389	195	78
Linoleic acid metabolism (00591)									
	FM888497		rcu:RCOM_0597840	Desacetoxyvindoline 4-hydroxylase	1.14.11.20	1063, 901	361	166	69
	FM889983		rcu:RCOM_0603110	Polyneuridine-aldehyde esterase precursor	3.1.1.78	901	260	144	60
	GO246771		rcu:RCOM_0602630	Polyneuridine-aldehyde esterase precursor	3.1.1.78	901	263	152	67
Alpha-Linolenic acid metabolism (00592)									
	Contig332	FM890615, FM896803, FM892730	ath:AT3G06860	Multifunctional protein MFP2	4.2.1.17	1070, 71, 592, 1040, 62	725	258	84
	*Contig369	FM891024, FM895943, FM893216, FM892025	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	63	77
	*Contig682	FM894273, FM896105, FM89466	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	176	80

*Contig859	FM895802, FM894994, FM889185	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	153	80
*Contig938	FM896585, FM893098, FM889131, FM889465	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	98	84
FM887427		pop:POPTR_550801	Lysine-specific histone demethylase 1	1.-.-.-	592, 950	795	47	68
FM887948		ath:AT3G59840	Hypothetical protein	1.-.-.-	592, 950	97	75	61
FM892914		rcu:RCOM_1020840	3-ketoacyl-CoA thiolase B	2.3.1.16	1070, 71, 592, 1040, 62	460	88	90
FM894979		pop:POPTR_832274	Acyl-CoA oxidase	1.3.3.6	1070, 71, 592, 1040	689	196	92
FM895499		cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	135	77
FM896436		rcu:RCOM_1016730	Cytochrome P450	4.2.1.92	1070, 592	496	140	80
FM896469		cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	87	65
FM896520		xtr:595069	Fa2h; fatty acid 2-hydroxylase	1.-.-.-	592, 950	371	139	52
GR209288		rcu:RCOM_0721980	Amine oxidase, lysine-specific histone demethylase 1	1.-.-.-	592, 950	961	63	80
GT228436		cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	183	68
GT229029		rcu:RCOM_1593180	Cytochrome P450	4.2.1.92	1070, 592	518	46	89
GT229218		rcu:RCOM_0696050	Enoyl-CoA hydratase, mitoc. precursor	4.2.1.17	1070, 71, 592, 1040, 62	389	195	78
Arachidonic acid metabolism (00590)								
*Contig326	FM890550, FM893788, GT229087, FM888006, FM895876, FM894387, FM889334, FM895717, FM890427	tgo:TGME49_067680	Microneme protein	1.14.99.1	590	2182	48	54
*Contig369	FM891024, FM895943, FM893216, FM892025	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	63	77

(*Continued*)

Table S2. (*Continued*)

Pathway	EST	Reads/contig	KEGG homolog	Definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
	*Contig429	FM891645, FM889741, GO246578, FM889668	rcu:RCOM_0902850	Short-chain dehydrogenase	1.1.1.184	590	253	156	70
	*Contig682	FM894273, FM896105, FM89466	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	176	80
	*Contig859	FM895802, FM894994, FM889185	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	153	80
	*Contig938	FM896585, FM893098, FM889131, FM889465	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	98	84
	*Contig1156	GO246787, FM895785, FM896209, FM890569, FM894435, GH295599, FM895852, FM893178, GH295993, GH296534, FM894252, FM888871, FM890857, FM894719	pop:POPTR_749200	Glutathione peroxidase	1.11.1.9	590	251	238	76
	FM894424		rcu:RCOM_1437260	Cytochrome P450	1.14.14.1	1063	632	180	86
	FM895499		cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	135	77
	FM896469		cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	87	65
	GT228436		cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	183	68
	GT229226		rcu:RCOM_1503180	Carbonyl reductase	1.1.1.189	590	315	110	80

Notes: [1]Sz is for the size of the KEGG homologous protein in amino acids; [2]Homol is for the size of the homologous region in amino acids; [3]Id% is for the pourcentage of identity of amino acid homologous pairs; *A contig with an asterisk in front means that its reads show nucleotide polymorphism.

Table S3. ESTs that tag for enzymatic functions putatively involved in the lipid metabolism of *J. curcas*.

Pathway	EST	Reads/ contig	KEGG homolog	Definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
Glycerolipid metabolism (00561)									
	Contig1275	GO247602, GO246458	rcu:RCOM_ 1454540	Alpha-galactosidase/ alpha-n-Acetylgalacto-saminidase	3.2.1.22	52, 561	408	175	72
	FM888117		ath:AT4G31780	1,2-diacylglycerol 3-beta-galacto-syltransferase	2.4.1.46	561	533	53	64
	FM888241		rcu:RCOM_ 0814190	Aldo-keto reductase	1.1.1.21	51, 40, 620, 52, 561	339	70	87
	FM889020		rcu:RCOM_ 0090060	Lysophosphatidic acid acyltransferase	2.3.1.51	564, 561, 565	395	54	83
	FM889814		ath:AT3G56310	Alpha-galactosidase, putative/melibiase/ alpha-D-galactoside galactohydrolase	3.2.1.22	52, 561	437	147	72
	FM892599		rcu:RCOM_ 1454590	Alpha-galactosidase/ alpha-n-Acetylgalacto-saminidase	3.2.1.22	52, 561	412	66	57
	FM895453		rcu:RCOM_ 1158010	Aldo-keto reductase	1.1.1.21	51, 40, 620, 52, 561	301	189	78
	FM895690		rcu:RCOM_ 0577310	ER Phosphatidate Phosphatase	3.1.3.4	564, 561, 565	316	129	76
	FM895817		rcu:RCOM_ 1255930	Diacylglycerol kinase, alpha	2.7.1.107	564, 561	526	140	77
	GO247581		rcu:RCOM_ 0512200	Triacylglycerol lipase 2 precursor	3.1.1.3	561	485	63	49
Glycerophospholipid metabolism (00564)									
	*Contig369	FM891024, FM895943, FM893216, FM892025	cre:CHLREDRAFT_ 185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	63	77
	*Contig682	FM894273, FM896105, FM89466	cre:CHLREDRAFT_ 185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	176	80
	*Contig859	FM895802, FM894994, FM889185	cre:CHLREDRAFT_ 185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	153	80
	*Contig938	FM896585, FM893098, FM889131, FM889465	cre:CHLREDRAFT_ 185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	98	84

(*Continued*)

Table S3. (*Continued*)

Pathway	EST	Reads/ contig	KEGG homolog	Definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
	*Contig1087	GH295893, GH295840, GH295893, FM891128, GH296234, GH296420, GH296287, GH296473	ath:AT1G13560	Ethanolaminephospho-transferase	2.7.8.1	564, 565	389	87	91
	FM887047		rcu:RCOM_ 0907340	Ethanolamine-phosphate cytidylyltransferase	2.7.7.14	564	377	182	89
	FM887959		pop:POPTR_ 596420	Glycerophosphoryl diester phospho-diesterase	3.1.4.46	564	399	131	83
	FM888369		rcu:RCOM_ 1610540	Phosphatidate cytidyly-ltransferase	2.7.7.41	564	404	42	95
	FM888558		rcu:RCOM_ 0899520	Phopholipase d alpha	3.1.4.4	564, 565	808	97	92
	FM888761		rcu:RCOM_ 0161380	Phosphoethanolamine n-Methyltransferase	2.1.1.103	564	492	49	85
	FM889020		rcu:RCOM_ 0090060	Lysophosphatidic acid acyltransferase	2.3.1.51	564, 561, 565	395	54	83
	FM890193		rcu:RCOM_ 0744550	Cdp-diacylglycerol—glycerol-3-phosphate 3-phosphatidyl-transferase	2.7.8.5	564	359	175	47
	FM895499		cre:CHLREDRAFT_ 185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	135	77
	FM895690		rcu:RCOM_ 0577310	ER Phosphatidate Phosphatase	3.1.3.4	564, 561, 565	316	129	76
	FM895817		rcu:RCOM_ 1255930	Diacylglycerol kinase, alpha	2.7.1.107	564, 561	526	140	77
	FM896258		rcu:RCOM_ 0531780	Glycerol-3-phosphate dehydrogenase	1.1.1.8	564	380	172	84
	FM896469		cre:CHLREDRAFT_ 185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	87	65
	GT228436		cre:CHLREDRAFT_ 185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	183	68
Ether lipid metabolism (00565)									
	*Contig369	FM891024, FM895943, FM893216, FM892025	cre:CHLREDRAFT_ 185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	63	77

*Contig682	FM894273, FM896105, FM89466	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	176	80
*Contig859	FM895802, FM894994, FM889185	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	153	80
*Contig938	FM896585, FM893098, FM889131, FM889465	cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	98	84
*Contig1087	GH295893, GH295840, GH295893, FM891128, GH296234, GH296420, GH296287, GH296473	ath:AT1G13560	Ethanolaminephospho transferase	2.7.8.1	564, 565	389	87	91
FM888558		rcu:RCOM_0899520	Phopholipase d alpha	3.1.4.4	564, 565	808	97	92
FM889020		rcu:RCOM_0090060	Lysophosphatidic acid acyltransferase	2.3.1.51	564, 561, 565	395	54	83
FM895499		cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	135	77
FM895690		rcu:RCOM_0577310	ER Phosphatidate Phosphatase	3.1.3.4	564, 561, 565	316	129	76
FM896469		cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	87	65
GT228436		cre:CHLREDRAFT_185967	FTT1; 14-3-3 protein	3.1.1.4	564, 590, 592, 565, 591	259	183	68

Notes: [1]Sz is for the size of the KEGG homologous protein in amino acids; [2]Homol is for the size of the homologous region in amino acids; [3]Id% is for the pourcentage of identity of amino acid homologous pairs; *A contig with an asterisk in front means that its reads show nucleotide polymorphism.

Table S4. ESTs that tag for enzymatic functions putatively involved in terpene biosynthesis in *J. curcas*.

Pathway	EST	Reads/contig	KEGG homolog	Definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
Terpenoid backbone biosynthesis (00900)									
	*Contig1203	GO247043, GH295921, FM896506, GH296298, GH295905, GH296485, GH296314, GH296501	swo:Swol_1344	4-hydroxy-3-methylbut-2-enyl diphosphate reductase/ S1 RNA-binding domain protein	1.17.1.2	1070, 1062, 1066, 900	687	54	51
	FM888319		pop: POPTR_418371	1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7	1070, 1062, 1066, 900	657	64	93
	FM889171		pop: POPTR_646543	Hydroxymethylglutaryl-CoA synthase	2.3.3.10	1070, 1062, 1066, 900	464	112	92
	FM894459		rcu: RCOM_1431350	Mevalonate kinase	2.7.1.36	1070, 1062, 1066, 900	386	60	83
	GO246785		rcu: RCOM_0679170	1-deoxyxylulose-5-phosphate synthase	2.2.1.7	1070, 1062, 1066, 900	720	81	90
	GT229261		rcu: RCOM_0747390	Geranylgeranyl pyrophosphate synthase	2.5.1.30	900	417	102	83
Monoterpenoid biosynthesis (00902)									
	FM887640		rcu: RCOM_1544790	R-limonene synthase	4.2.3.20	1062, 902	596	45	80
	FM888507		rcu: RCOM_1732250	R-limonene synthase	4.2.3.20	1062, 902	250	54	44
	FM888632		rcu: RCOM_1593840	Cytochrome P450	1.3.3.9	1062, 1066, 902	613	64	65
	FM889984		rcu: RCOM_1544790	R-limonene synthase	4.2.3.20	1062, 902	596	72	75
Diterpenoid biosynthesis (00904)									
	FM887587		rcu: RCOM_1013800	3′-N-debenzoyl-2′-deoxytaxol N-benzoyltransferase	2.3.1.167	1062, 904	446	86	89
	FM889243		rcu: RCOM_1607300	Cytochrome P450	1.14.13.79	1070, 1062, 904	298	44	84
	FM891970		rcu: RCOM_1574750	Casbene synthase, chloroplast precursor	4.2.3.8	1062, 904	555	138	69
	FM896109		ath: AT3G48360	BT2; BT2 (BTB and TAZ domain protein 2; protein binding/transcription factor/transcription regulator)	1.14.-.-	1070, 1062, 100, 904, 130, 905	364	87	50
Sesquiterpenoid biosynthesis (00909)									
	FM889102		rcu:RCOM_1045160	Delta-cadinene synthase isozyme A	4.2.3.13	1062, 909	547	112	59

Notes: [1]Sz is for the size of the KEGG homologous protein in amino acids. [2]Homol is for the size of the homologous region in amino acids. [3]Id% is for the pourcentage of identity of amino acid homologous pairs; *A contig with an asterisk in front means that its reads show nucleotide polymorphism.

Table S5. ESTs that tag for enzymatic functions putatively involved in alkaloid biosynthesis in *J. curcas*.

Pathway	EST	Reads/contig	KEGG homolog	Definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
Indole alkaloid biosynthesis (00901)									
	FM888497		rcu:RCOM_0597840	Desacetoxyvindoline 4-hydroxylase	1.14.11.20	1063, 901	361	166	69
	FM889983		rcu:RCOM_0603110	Polyneuridine-aldehyde esterase precursor	3.1.1.78	901	260	144	60
	GO246771		rcu:RCOM_0602630	Polyneuridine-aldehyde esterase precursor	3.1.1.78	901	263	152	67
Isoquinoline alkaloid biosynthesis (00950)									
	*Contig794	FM895278, FM895972, FM895925, FM895522, FM895639	rcu:RCOM_1052360	Aspartate aminotransferase	2.6.1.1	1070, 1064, 710, 950	404	276	90
	Contig872	FM895919, FM896552	rcu:RCOM_0770830	Reticuline oxidase precursor	1.21.3.3	1063, 950	524	251	72
	*Contig930	FM896500, FM896046	rcu:RCOM_1173140	Tyrosine aminotransferase	2.6.1.5	130, 950, 1063, 1064	419	131	83
	*Contig1301	GR209293, GR209295	rcu:RCOM_0536450	Amine oxidase	1.4.3.4	950, 982	491	150	88
	FM887321		rcu:RCOM_1682160	Cytochrome P450	1.14.13.71	950	501	83	57
	FM887370		rcu:RCOM_0651460	Tyrosine aminotransferase	2.6.1.5	130, 950, 1063, 1064	415	110	75
	FM887427		pop:POPTR_550801	Lysine-specific histone demethylase 1	1.-.-.-	592, 950	795	47	68
	FM887948		ath:AT3G59840	Hypothetical protein	1.-.-.-	592, 950	97	75	61
	FM888061		rcu:RCOM_1173140	Tyrosine aminotransferase	2.6.1.5	130, 950, 1063, 1064	419	83	79
	FM888095		vvi:100242364	Aspartate aminotransferase	2.6.1.1	1070, 1064, 710, 950	462	220	91
	FM890598		pop:POPTR_896773	Aspartate aminotransferase	2.6.1.1	1070, 1064, 710, 950	406	84	86
	FM890776		rcu:RCOM_0922300	Aspartate aminotransferase	2.6.1.1	1070, 1064, 710, 950	464	46	97
	FM893208		reh:H16_A1025	Rossmann fold nucleotide-binding protein/ lysine decarboxylase family protein	4.1.1.25	1063, 950	197	94	52
	FM894218		rcu:RCOM_1574460	S-N-methylcoclaurine 3′-hydroxylase isozyme	1.14.13.71	950	318	162	66
	FM896388		rcu:RCOM_0984440	Aspartate aminotransferase	2.6.1.1	1070, 1064, 710, 950	484	104	80

(*Continued*)

Table S5. (*Continued*)

Pathway	EST	Reads/contig	KEGG homolog	Definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
	FM896520		xtr:595069	Fa2h; fatty acid 2-hydroxylase	1.-.-.-	592, 950	371	139	52
	FM896620		rcu:RCOM_1335460	Reticuline oxidase precursor	1.21.3.3	1063, 950	539	165	72
	GO246518		rcu:RCOM_0802970	Aspartate aminotransferase	2.6.1.1	1070, 1064, 710, 950	440	121	85
	GR209288		rcu:RCOM_0721980	Amine oxidase, lysine-specific histone Demethylase 1	1.-.-.-	592, 950	961	63	80
	GT228598		rcu:RCOM_1081680	3′-N-debenzoyl-2′-deoxytaxol N-benzoyltransferase	2.3.1.150	1063, 950	437	45	75
	GT229047		rcu:RCOM_1335910	Reticuline oxidase precursor	1.21.3.3	1063, 950	548	186	73
Ubiquinone and other terpenoid-quinone biosynthesis (00130)									
	*Contig270	FM889998, FM890294	rcu:RCOM_1328840	AMP dependent ligase	6.2.1.26	130	556	108	80
	*Contig930	FM896500, FM896046	rcu:RCOM_1173140	Tyrosine aminotransferase	2.6.1.5	130, 950, 1063, 1064	419	131	83
	Contig1079	GH295872, GH296266, GH296452	rcu:RCOM_0999850	AMP dependent ligase	6.2.1.26	130	564	81	87
	FM887370		rcu:RCOM_0651460	Tyrosine aminotransferase	2.6.1.5	130, 950, 1063, 1064	415	110	75
	FM888061		rcu:RCOM_1173140	Tyrosine aminotransferase	2.6.1.5	130, 950, 1063, 1064	419	83	79
	FM889157		rcu:RCOM_1192240	4-hydroxyphenylpyruvate dioxygenase	1.13.11.27	130	441	171	53
	FM893861		rcu:RCOM_1328840	AMP dependent ligase	6.2.1.26	130	556	158	81
	FM896109		ath:AT3G48360	BT2; BTB and TAZ domain protein 2 (protein binding/transcription factor/ transcription regulator)	1.14.-.-	130, 1070, 1062, 100, 904, 905	364	87	50
	GO246531		rcu:RCOM_1495910	AMP dependent CoA ligase	6.2.1.26	130	521	80	56

Notes: [1]Sz is for the size of the KEGG homologous protein in amino acids; [2]Homol is for the size of the homologous region in amino acids; [3]Id% is for the pourcentage of identity of amino acid homologous pairs; *A contig with an asterisk in front means that its reads show nucleotide polymorphism.

Table S6. ESTs that tag for enzymatic functions putatively involved in the drug metabolism of *J. curcas*.

Pathway	EST	Reads/contig	KEGG homolog	Definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
Drug metabolism—cytochrome P450 (00982)									
	*Contig612	FM893499, FM890999, GO247659, FM892688, GO247172, GT228753, FM890592	rcu:RCOM_0324140	Glutathione-s-transferase theta, gst	2.5.1.18	982	214	210	87
	*Contig730	FM894684, GO246483, GO246481, GO246669, FM887600, GT228501, FM895945	dme:Dmel_CG10067	Act57B; Actin 57B	2.4.1.17	500, 40, 983, 150, 982	376	229	86
	Contig798	FM895299, FM895222	rcu:RCOM_1600930	Dimethylaniline monooxygenase	1.14.13.8	982	517	201	76
	Contig897	FM896130, FM895498	rcu:RCOM_1439520	Glutathione-s-transferase theta, gst	2.5.1.18	982	217	199	79
	*Contig1169	GO246868, GO247111, FM890810, FM893607, FM895462, FM887782	rcu:RCOM_1439530	Glutathione-s-transferase	2.5.1.18	982	213	214	70
	*Contig1301	GR209293, GR209295	rcu:RCOM_0536450	Amine oxidase	1.4.3.4	950, 982	491	150	88
	FM887133		rcu:RCOM_0764870	Microsomal glutathione s-transferase	2.5.1.18	982	146	90	78
	FM889542		rcu:RCOM_0927500	Monooxygenase	1.14.13.8	982	421	61	44
	FM893192		osa:4332456	Glutathione S-transferase	2.5.1.18	982	243	60	55
	FM894424		rcu:RCOM_1437260	Cytochrome P450	1.14.14.1	1063	632	180	86
Drug metabolism—other enzymes (00983)									
	Contig151	FM888672, FM892761, GT228812	bps:BPSS0882	IMP dehydrogenase	1.1.1.205	1065, 983	154	63	44

(Continued)

Table S6 *(Continued)*

Pathway	EST	Reads/contig	KEGG homolog	Definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
	*Contig730	FM894684, GO246483, GO246481, GO246669, FM887600, GT228501, FM895945	dme:Dmel_ CG10067	Act57B; Actin 57B	2.4.1.17	500, 40, 983, 150, 982	376	229	86
	Contig917	FM896312, FM894984	rcu:RCOM_ 1022280	Inosine triphosphate Pyrophosphatase	3.6.1.19	983	284	141	85
	FM887557		pop:POPTR_ 563741	mRNA-decapping enzyme subunit 2	3.-.-.-	983	322	137	95
	FM888766		rcu:RCOM_ 1247680	GMP synthase	6.3.5.2	1065, 983, 620, 61	180	106	88
	FM889652		rcu:RCOM_ 1120040	Thymidine kinase	2.7.1.21	983	234	44	70
	FM891852		pop:POPTR_ 815577	Uridine kinase	2.7.1.48	983	481	73	97
	FM893638		rcu:RCOM_ 1155570	Sigma factor sigb regulation protein RSBQ	3.1.1.1	983	269	156	67
	FM893677		rcu:RCOM_ 0220610	Acyl-protein thioesterase	3.1.1.1	983	258	42	92
	FM894424		rcu:RCOM_ 1437260	Cytochrome P450	1.14.14.1	1063	632	180	86
	FM895004		rcu:RCOM_ 0036520	Sigma factor sigb regulation protein RSBQ	3.1.1.1	983	279	139	48
	FM895184		pop:POPTR_ 1074034	GMP synthase (glutamine-hydrolysing)	6.3.5.2	1065, 983, 620, 61	534	56	96
	FM895463		dme:Dmel_ CG10927	CG10927 gene product	3.-.-.-	983	360	43	60
	FM895583		rcu:RCOM_ 0897920	Acyl-protein thioesterase	3.1.1.1	983	258	77	84
	FM896706		rcu:RCOM_ 1619070	Uridine cytidine kinase I	2.7.1.48	983	657	103	72
	FM896764		pop:POPTR_ 672427	Cytidine deaminase	3.5.4.5	983	292	128	83

Notes: [1]Sz is for the size of the KEGG homologous protein in amino acids; [2]Homol is for the size of the homologous region in amino acids; [3]Id% is for the pourcentage of identity of amino acid homologous pairs; *A contig with an asterisk in front means that its reads show nucleotide polymorphism.

Table S7. ESTs that tag for enzymatic functions putatively involved in hormone biosynthesis in *J. curcas*.

Pathway	EST	Reads/contig	KEGG homolog	Definition	EC	Map	Sz[1]	Homol[2]	Id%[3]
Biosynthesis of plant hormones (01070)									
Auxine									
	FM888386		rcu:RCOM 0642840	4-hydroxybenzoate octaprenyltransferase	2.5.1.-	1070	387	194	74
Cytokinin									
	Contig539	FM892725, FM893185, FM893296, FM895320, FM896321, FM892466, FM896208, FM891696, GO247571, FM892573, GH295577, GH296276, GH295882, GH296462, GH295971	bmy:Bm1 04640	Phosphatidylinositol glycan, class B	2.4.1.-	1070	612	58	58
	GO247621		bmy:Bm1 04640	Phosphatidylinositol glycan, class B	2.4.1.-	1070	612	53	47
Brassinosteroid									
	FM896109		ath:AT3G48360	BT2; BT2 (BTB and TAZ domain protein 2; protein binding/transcription factor/transcription regulator)	1.14.-.-	1070, 1062, 100, 904, 130, 905	364	87	50
Androgen and estrogen metabolism (00150)									
	*Contig730	FM894684, GO246483, GO246481, FM894684, GO246669, FM887600	dme:Dmel CG10067	Act57B; Actin 57B	2.4.1.17	500, 40, 983,150, 982	376	229	86
	Contig933	FM896543, FM894024, GT228698	rcu:RCOM 1082860	Steroid dehydrogenase	1.1.1.62	150	320	167	90
	FM893027		edi:EDI 043370	Copine	2.8.2.2	150	277	80	57
	FM894424		rcu:RCOM 1437260	Cytochrome P450	1.14.14.1	1063	632	180	86
	GO246967		vvi:100243811	3-oxo-5-alpha-steroid 4-dehydrogenase 3	1.3.99.5	150	336	94	64

Notes: [1]Sz is for the size of the KEGG homologous protein in amino acids; [2]Homol is for the size of the homologous region in amino acids; [3]Id% is for the pourcentage of identity of amino acid homologous pairs; *A contig with an asterisk in front means that its reads show nucleotide polymorphism.