Transcriptome Analysis of Yellow Horn (Xanthoceras sorbifolia Bunge): A Potential Oil-Rich Seed Tree for Biodiesel in China

Yulin Liu1, Zhedong Huang2, Yan Ao3, Wei Li1*, Zhixiang Zhang2*

1 College of Biological Science and Biotechnology, Beijing Forest University, Beijing, China, 2 Laboratory of Systematic Evolution and Biogeography of Woody Plants, College of Nature Conservation, Beijing Forest University, Beijing, China, 3 Academy of Forest, Beijing Forest University, Beijing, China

Abstract

Background: Yellow horn (Xanthoceras sorbifolia Bunge) is an oil-rich seed shrub that grows well in cold, barren environments and has great potential for biodiesel production in China. However, the limited genetic data means that little information about the key genes involved in oil biosynthesis is available, which limits further improvement of this species. In this study, we describe sequencing and de novo transcriptome assembly to produce the first comprehensive and integrated genomic resource for yellow horn and identify the pathways and key genes related to oil accumulation. In addition, potential molecular markers were identified and compiled.

Methodology/Principal Findings: Total RNA was isolated from 30 plants from two regions, including buds, leaves, flowers and seeds. Equal quantities of RNA from these tissues were pooled to construct a cDNA library for 454 pyrosequencing. A total of 1,147,624 high-quality reads with total and average lengths of 530.6 Mb and 462 bp, respectively, were generated. These reads were assembled into 51,867 unigenes, corresponding to a total of 36.1 Mb with a mean length, N50 and median of 696, 928 and 570 bp, respectively. Of the unigenes, 17,541 (33.82%) were unmatched in any public protein databases. We identified 281 unigenes that may be involved in de novo fatty acid (FA) and triacylglycerol (TAG) biosynthesis and metabolism. Furthermore, 6,707 SSRs, 16,925 SNPs and 6,201 InDels with high-confidence were also identified in this study.

Conclusions: This transcriptome represents a new functional genomics resource and a foundation for further studies on the metabolic engineering of yellow horn to increase oil content and modify oil composition. The potential molecular markers identified in this study provide a basis for polymorphism analysis of Xanthoceras, and even Sapindaceae; they will also accelerate the process of breeding new varieties with better agronomic characteristics.

Introduction

Due to the crises of fossil fuel depletion and worsening global environmental conditions, oil-rich seed plants that can be used to produce renewable and environmentally friendly biodiesel have received much attention [1–3]. A variety of vegetable oils that are obtained from rapeseed (canola), soybean, sunflower, peanut, safflower, palm and Jatropha, among others, have been used to produce biodiesel. However, the great majority of these plants are grown on farmland and are used as cooking oil. Under these circumstances, in some developing countries that have limited per capita arable land, the use of food crops to produce biodiesel is not realistic. Therefore, biodiesel production from non-food crops that can be planted in areas that are unsuitable for traditional crops is a realistic option to this problem [4–6].

Yellow horn (Xanthoceras sorbifolia Bunge) is an oil-rich seed shrub that belongs to the Sapindaceae family and has a life span of more than 200 years [7,8]. The seeds of yellow horn contain abundant oil (55–70%), of which 85–93% is unsaturated fatty acids [9,10]. According to previous studies, the molecular composition of yellow horn oil is similar to the ideal fatty ester structure for biodiesel [11,12]. Therefore, it has been identified as a major biodiesel tree species and the Chinese Government provided special support to aid its development because it can produce over 800 gallons of oil per acre of cultivation [13,14]. Unlike other energy-resource trees, such as palm and Jatropha, that cannot survive low temperatures, it can not only grow well in barren, salty and drought soil, it can also survive temperatures as low as −30 to −41°C. In addition, the yellow horn tree has many other uses, including multiple entries in the Chinese Pharmacopoeia, it can assist in eliminating desertification and erosion, and it is grown as an ornamental tree and used as a source of high-level woody natural oil for cooking [15].
Results and Discussion

Longer than 1,000 bp (Table 1, Figure 1C).

Using CD-HIT for isoforms and singlets, a dataset of 51,867 unigenes were assembled using Newbler2.6. Finally, after a last assembly length = 549 bp, median length = 484, GC content = 42.70%)

Length = 874.7 bp, N50 length = 1,116 bp, median length = 732 bp, GC content = 41.40%) and 42,787 singlets (total length = 17,190,382 bp, average length = 401.8 bp, N50 length = 549 bp, median length = 484, GC content = 42.70%) were assembled using Newbler2.6. Finally, after a last assembly using CD-HIT for isoforms and singlets, a dataset of 31,867 unigenes (45 to 10,088 bp with a GC content of 41.37%) was obtained, corresponding to 36.1 Mb with mean, N50 and median lengths of 696, 928 and 570 bp, respectively. Of the unigenes 7,127 (13.74%) were equal to or shorter than 200 bp and 10,960 (21.13%) were longer than 1,000 bp (Table 1, Figure 1C).

Characterisation of Non-redundant Unigenes

To understand their functions, the 51,867 yellow horn unigenes were annotated using BLASTx alignment with an E-value cut-off of 10−5 against the following protein databases: NR, Swiss-Prot, CDD, Pfam, TrEMBL, COG, GO, KEGG and TAIR. A total of 33,924 (65.41%), 33,872 (65.31%), 30,304 (58.81%), 28,460 (54.87%), 26,643 (51.36%), 24,412 (47.07%), 24,258 (46.77%), 11,181 (21.36%) and 7,415 (14.30%) unigenes had significant matches with sequences in the NR, TrEMBL, Pfam, TAIR, GO, Swiss-Prot, CDD, COG and KEGG databases, respectively (Table 2). Of the unigenes, 34,326 (66.18%) were described in at least one database with high homology (Figure 1A) with unigenes from F. tita unifrons (9,580, 18.47%), Ricinus communis (7,900, 15.23%), Populus trichocarpa (7,337, 14.13%), Glicine max (2,052, 3.96%), Medicago truncatula (755, 1.46%), Arabidopsis thaliana (655, 1.22%), and other species (6,067, 11.70%) (Figure 1B). However, the remaining 17,541 (33.82%) were unmatched (70.55% of unigenes <500 bp and 1.90% of unigenes >1,000 bp) (Figure 1C), suggesting that longer sequences were more likely to have BLAST hits and shorter sequences may have been either too short to get hits or lacked a characterised protein domain, which resulted in false-negative results. Because public databases contain little genomic and transcriptomic information for yellow horn, these unmatched unigenes may represent putative tissue-specific novel genes or non-coding regions.

Functional Classification of Unigenes by COG, GO and KEGG

To further evaluate the completeness of our transcriptome library and the effectiveness of our annotation process, we used the annotated unigene sequences to search for genes involved in COG classifications, GO assignments and KEGG pathway assignments to predict and classify their functions.

Overall, 11,181 unigenes were assigned into 24 COG function categories. Among them, the cluster for “general function prediction only” represented the largest group (4,227, 37.81%), followed by “posttranslational modification, protein turnover, chaperones” (2,086, 18.66%), “replication, recombination repair” (1,526, 13.65%), “transcription” (1,497, 13.39%) and “translation, ribosomal structure and biogenesis” (1,459, 13.05%). However, few unigenes were assigned into “cell motility” (36, 0.32%) and “nuclear structure” (3, 0.03%). Additionally, 967 (8.65%) unigenes were assigned into the category “lipid transport and metabolism” (Figure 2).

Of the unigenes, 26,643 were assigned into the three main GO functional categories and then into 50 sub-categories (Figure 3). For the three main categories, 23,978 unigenes (90.03%) were assigned into the largest category, molecular function, followed by biological process (22,187, 83.28%) and cellular component (15,012, 56.35%). The biological process category was assigned into 25 sub-categories; the most abundant was “metabolic process”, which contained 18,891 unigenes (36.42%) of the total, indicating that these genes were enriched in the yellow horn transcriptome libraries. The cellular components category was divided into 11 small groups; the largest sub-category was “cell”, which included 11,855 (22.86%) of the total unigenes. For molecular function, 23,978 unigenes were categorised into 14 GO terms; the majority fell into “binding” and “catalytic activity”, which contained 17,084 and 15,045 unigenes, respectively.

In addition, a total of 7,415 unigenes were assigned to 182 pathways through KEGG pathway assignment. Among them, the largest represented category was “metabolism”, containing 3,862 unigenes, followed by “genetic information processing” and...
Figure 1. Unigenes functional annotation results. (A) Top-hit species distribution for BLASTx matches for yellow horn unigenes using the following order of priority: NR, TrEMBL and Swiss-Prot. (B) E-value distribution of top BLASTx hits for each unigene. (C) Distribution of unigenes in length with BLASTx hits compared with those without hits.

doi:10.1371/journal.pone.0074441.g001
“cellular processes”, which contained 2,736 and 1,070 unigenes, respectively. The pathway “ribosome” involved the largest number of unigenes (475), but both the “primary bile acid biosynthesis” and “nitrotoluene degradation” pathways contained only one unigene. In addition, 548 unigenes were mapped to 15 pathways in the sub-category “lipid metabolism” (Table S1).

### Unigenes Related to FA Biosynthesis

In oil plants, fatty acids are stored as a form of TAG and their biosynthesis pathway can be divided into three steps [34]. The first step is de novo biosynthesis of fatty acids; this process occurs in plastids and is catalysed mainly by the fatty acid synthase complex (FAS). The second step is the synthesis of triacylglycerol (TAG), which occurs in the endoplasmic reticulum (ER), and the last is the formation of oil bodies (OBs), where TAG is combined with oleosin to form an oil body, which is released from the ER into the cytoplasm [35,36].

According to the KEGG pathway assignment and functional annotation of the unigenes, 40 unigenes were annotated as encoding ten key enzymes involved in FAs biosynthesis (Table 3). The reconstructed pathway of FAS biosynthesis was based on these identified enzymes (Figure 4). First, acetyl-CoA carboxylase (ACCase, EC: 6.4.1.2), as a rate-limiting enzyme in the FAs biosynthesis pathway, catalyses acetyl-CoA to form malonyl-CoA [37]; 16 unigenes that encode its four subunits were identified (four for α-carboxyltransferase, three for β-carboxyltransferase, six for acyl-CoA binding protein, and three for acyl carrier protein). Next, a series of condensation reactions of malonyl-CoA with a growing ACP-bound acyl chain are catalysed by FAS, consecutively adding two carbon units per cycle over six or seven cycles to form 16:0-ACP or 18:0-ACP, which can then be catalysed by acyl-ACP desaturase (AAD, EC: 1.14.19.2) to form 16:1-ACP or 18:1-ACP [38]. Nineteen unigenes that encoded the five components of FAS were found, including one that encoded malonyl-CoA-ACP transacylase (MAT, EC: 2.3.1.39), eight that encoded 3-ketoacyl ACP synthase (KAS) for seven for KAS II (EC: 2.3.1.179) and one for KAS III (EC: 2.3.1.180), six that encoded 3-ketoacyl ACP reductase (KAR, EC: 1.1.1.100), two that encoded 3-hydroxyacyl-ACP dehydrase (HAD, EC: 4.2.1.-) and two that encoded enoyl-ACP reductase (EAR, EC:1.3.1.9). Moreover, nine unigenes coded the acyl carrier protein (ACP), an essential cofactor of FAS. After this, under the control of acyl-ACP thioesterase (FAT, EC: 3.1.2.14 3.1.2.-) and palmitoyl-CoA hydrolase (PCH, EC: 3.1.2.2), free FAs are released from the acyl carrier protein (ACP). Two unigenes that encoded FATA, three that encoded FATB and one that encoded PCH which is also involved in FA elongation were identified in this study.

In addition, 16 unigenes that encode long-chain acyl-CoA synthetases (LACS), which catalyse the esterification of free FAs to CoA upon arrival in the cytoplasm [39], and seven that encoded acyl-CoA binding protein (ACBP), which binds medium- and long-chain acyl-CoA esters with very high affinity and may function as an intracellular carrier of acyl-CoA esters [40], were also identified based on the functional annotation of the transcriptome.

As has been reported previously, overexpression of ACCase, a crucial enzyme in fatty acid synthesis, can alter the fatty acid composition of seeds and increase the fatty acid content, which would lead to an increased oleic acid content and seed yield [41,42]. Most research on FAS has concentrated on ACP and KAS. Functional expression of an ACP from Azospirillum brasilense in Brassica juncea can improve the content of 18:1 and 18:2 in seeds, enhance the ratio of monounsaturated (18:1) to saturated fatty acids, increase the ratio of 18:2 to 18:3 and reduce the erucic acid content [22:1] [43]. In plastids, three types of KAS are found: KAS I, KAS II and KAS III. KAS II catalyses 16:0-ACP to elongate to 18:0-ACP and KAS III condenses acetyl-CoA with malonyl-ACP to form 4:0-ACP. Using hairpin RNAi to reduce the activity of KAS II can lead to an increase in 16:0 accumulation, up to about 53% of the total, but some transgenic offspring are deformed during early embryonic development [44]. Overexpression of KAS III can also improve 16:0 accumulation, but the rate of lipid synthesis is reduced [45]. In our study, we did not find any unigenes that encoded KAS I, which is highly active with acyl-ACP with chain lengths from C2 to C14, is far less effective for 16:0-ACP and almost inactive for 18:0-ACP [46].
Unigenes Related to TAG and OBs Biosynthesis

In the suggested pathway for TAG biosynthesis [47,48], a total of 33 unigenes that encode six enzymes were found. As the data in Table 4 and Figure 4 show, initially, one unigene that encodes glycerol kinase (GK, EC: 2.7.1.30), which catalyses glycerol to form glycerol-3-phosphate (G-3-P), an initial substrate in the Kennedy pathway, was detected. Then, 12 unigenes that encode the key component of TAG biosynthesis, glycerol-3-phosphate acyltransferase (ATS1 and GPAT, EC: 2.3.1.15) (three for ATS1, one for GPAT5, three for GPAT6, two for GPAT8 and three for GPAT9), which catalyses the first step of the Kennedy pathway, and 11 unigenes that encode lysophosphatidyl acyltransferase (LPAT, EC: 2.3.1.51) were identified (four for LPAT1, two for LPAT4 and two for LPAT5). Under catalysis by these two enzymes, sequential esterification of acyl chains from acyl-CoA to the positions of sn-1 and sn-2 of G-3-P occur to form lysophosphatic acid (Lyso-PA) and phosphatidic acid (PA), respectively. The next reaction is catalysed by phosphatidate phosphatase (PP, EC: 3.1.3.4), a key regulator of lipid homeostasis, which was encoded by four unigenes; it plays a role in the removal of the phosphate group from PA and forms diacylglycerol (DAG), which is an essential intermediate in the biosynthesis of phosphatidylcholine (PC). Finally, two enzymes, diacylglycerol O-acyltransferase (DGAT, EC: 2.3.1.20) and phospholipid: diacylglycerol acyltransferase (PDAT, EC: 2.3.1.158), which use acyl-CoA and phospholipids as acyl-donors, respectively, transfer an acyl group to the sn-3 position of DAG to produce TAG. Only one unigene that encoded DGAT1 and four that encoded PDAT1 were identified in the yellow horn transcriptome. In addition, compared to the single unigene that encoded GPAT1, the finding of four unigenes that encoded PDAT1 in our study provides further evidence that yellow horn has the potential to channel fatty acids that are incorporated in membrane lipids, such as PC, into TAG biosynthesis.

After biosynthesis, pools of TAGs can be stored as a form of OB surrounded by a membrane composed of a layer of phospholipids embedded with several proteins: oleosin, caleosin and steroleosin in mature seeds [35,54]. According to the annotations of the unigenes, nine encoded oleosin, seven encoded caleosin and four encoded steroleosin (Table S2). Oleosin is the most abundant structural protein in OBs; it helps stabilise OBs through increased space bit resistance and charge repulsion, preventing fusion of OBs [35,55]. Caleosin is not only involved in the synthesis and metabolism of OBs, but may also be associated with plant drought tolerance [55,56]. Steroleosin-like proteins may represent a class of dehydrogenases/reductases that are involved in plant signal transduction regulated by various sterols [57]. In conclusion, the detection of unigenes that are involved in oleosin, caleosin and steroleosin biosynthesis will contribute to future functional studies and improvements in production levels by metabolic engineering of yellow horn.

Unigenes Related to FA Desaturation

Many types of enzyme participate in fatty acid desaturation in plants, and can be divided into two types. One type catalyses the formation of mono-unsaturated fatty acids from saturated fatty acids in plastids (16:0 to 16:1, 18:0 to 18:1); these contain only a soluble enzyme, acyl-ACP desaturase (AAD, EC: 1.14.19.2) [58]. The other type is located on the membranes of the endoplasmic reticulum and chloroplast and introduces double-unsaturated bonds at specifically defined positions (Δ12,
Δ15 or Δ6) in fatty acids that are esterified to a glycerol backbone [59], including Δ12(ω6)-Desaturase (FAD2 and FAD6, EC:1.14.19.-), which desaturates oleic acid (18:1) to form linoleic acid (18:2), Δ15(ω3)-Desaturase (FAD3, FAD7 and FAD8, EC:1.14.19.-), which further desaturates linoleic acid (18:2) to form α-linolenic acid (18:3), etc. In total, six, five, two, one, five and one unigenes that encode AAD, FAD2, FAD3, FAD6, FAD7 and FAD8 were found (Table 3). Oleic (18:1) and
Linoleic acids (18:2) are major constituents of yellow horn oil, according to a previous study, and ideal biodiesel should contain at least one double bond [19]; therefore, AAD, FAD2 and FAD6 are potential biotechnological targets for adjusting yellow horn oil composition.

Unigenes Related to Catabolism Pathways for TAGs and FAs

The complete breakdown of TAGs can be divided into two steps [60]. First, TAGs are metabolised to free FAs; in other words, lipases catalyse the hydrolysis of ester bonds that link fatty acyl chains to the glycerol backbone. During this research, three unigenes that encode triacylglycerol lipase (TGL, EC: 3.1.1.3), which releases fatty acids and intermediate products (DAG or monoacylglycerol) from TAG or DAG, were identified in the transcriptome library (Table 4). In the second step, fatty acids are catabolised to acetyl-CoA, allowing them to be further broken down by oxidation or to follow other metabolic pathways, including re-esterification with glycerol, to form new acylglycerols [61]. Based on the KEGG pathway assignment, we identified 83 unigenes that code for enzymes related to fatty acid catabolism; three key enzymes were acyl-CoA oxidase (ACOX, EC: 1.3.1.3), enoyl-CoA hydratase/3-hydroxyacyl-CoA dehydrase (MEF2, EC: 4.2.1.17 1.1.35 1.1.211) and acyl-CoA acyltransferase (ACAA, EC: 2.3.1.16), which were encoded by 11, 4 and 8 unigenes, respectively (Table 3). Acetyl-CoA generated through fatty acid catabolism is then used to produce energy for the cell via the citrate cycle or may participate in the synthesis of TAG.

TAG and FA catabolism proceeds in a direction opposite that of their synthesis. Therefore, identifying ways to suppress enzymes involved in TAG and FA catabolism may be another method of increasing the accumulation of lipids under conditions that do not affect plant growth. However, suppressing the expression of TGL increases TAG levels but results in severely stunted growth [62].

Detection of Transcription Factors (TFs) Involved in Oil Synthesis

In previous studies, a set of TFs, including LEC1, LEC2, ABI3, FUS3 and WR11 which play key roles in seed oil synthesis and deposition, were identified [63–65]. These TFs were reviewed by Fobert and made available online (http://lipidlibrary.aocs.org/ plantbio/transfactors/index.htm).

To identify TFs that regulate seed oil synthesis, BLASTx was used to search against the AGRIS (Arabidopsis Gene Regulatory Information Server) database with an e-value cut off 10^-5 [66]. The results showed that 3,341 unigenes were annotated with 905 independent coding sequences of Arabidopsis TFs belonging to 49 known TF families. However, none of the unigenes were annotated to the AtRKD family (Table S3). The largest number of unigenes (817) was annotated to the Trihelix family, followed by the C2H2 family (519 unigenes). TF genes in the PHD, CAMTA and JUMONJI families were identified in the yellow horn transcriptome. Among the 3,347 unigenes, 25 encoding eight TFs involved in oil biosynthesis were detected (Table 5), including ABI3, L11, ADOF1, EMF2, HSI2, HSI2-L1, AP2 and G12. However, none of the unigenes showed homology to LEC1, LEC2, FUS3, WR11, PKI, FIE or SWN.
Identification of Molecular Markers Located in Unigenes Related to Oil Biosynthesis and Metabolism

Simple sequence repeats (SSRs), single nucleotide polymorphisms (SNPs) and insertions and deletions (InDels) are valuable tools for genetic analysis because they are highly polymorphic within species. They can be used for molecular marker-assisted selection (MAS), which is a rapid approach to the development of new crop varieties (particularly in perennials and trees) that reduces the assessment time considerably [67,68], association mapping to find genes related to good agronomic characteristics [69], and polymorphism analysis, among other techniques. Molecular markers located in genes will likely be related to the functions of those genes. Some markers that are located in coding regions related to oil synthesis have been developed and applied [70,71].

In this study, a total of 6,707 SSRs distributed in 5,631 unigenes (4.61 Mb) were identified as potential molecular markers, of which 887 sequences contained more than one SSR. The most common SSRs were dinucleotide repeats, occurring at 3,598 loci (53.65%), followed by trinucleotide repeats (2,976, 44.37%). However, tetranucleotide, pentanucleotide and hexanucleotide repeats were found at lower frequencies, occurring only in 72 (1.15%), 17

### Table 3. Enzymes/protein related to FA biosynthesis and metabolism identified by annotation of the yellow horn unigenes.

| Symbol | Enzymes/Protein | EC Number | Number of unigenes |
|--------|-----------------|-----------|--------------------|
| **Fatty acid biosynthesis** | | | |
| accA | α-acetyl-CoA carboxylase carboxyl transferase | EC.6.4.1.2 | 4 |
| accB, bccP | acetyl-CoA carboxylase biotin carboxyl carrier protein | – | 3 |
| accC | acetyl-CoA carboxylase/biotin carboxylase | EC.6.4.1.2 | 6 |
| accD | α-acetyl-CoA carboxylase carboxyl transferase | EC.6.4.1.2 | 3 |
| ACP | acyl-carrier protein | – | 9 |
| MAT | Malonyl-CoA-ACP transacylase | EC.2.3.1.39 | 1 |
| KASII | 3-Ketoacyl ACP synthase II | EC.2.3.1.179 | 7 |
| KAR | 3-Ketoacyl ACP reductase | EC.1.1.1.100 | 6 |
| KASIII | 3-Ketoacyl ACP synthase III | EC.2.3.1.180 | 1 |
| EAR | enoyl-ACP reductase I | EC.1.3.1.9 | 2 |
| HAD | 3R-hydroxymyristoyl ACP dehydrase | EC.4.2.1.1 | 2 |
| FATA | fatty acyl-ACP thioesterase A | EC.3.1.2.14 3.1.2.- | 2 |
| FATB | fatty acyl-ACP thioesterase B | EC.3.1.2.14 3.1.2.- | 3 |
| **Fatty acid elongation** | | | |
| KCS | 3-ketoacyl-CoA synthase | EC.2.3.1.- | 20 |
| KR | α-keto reductase | EC.1.1.1.1 | 3 |
| PHS1 | 3-hydroxy acyl-CoA dehydratase | EC.4.2.1.- | 6 |
| MECR | mitochondrial trans-2-enoyl-CoA reductase | EC.1.3.1.38 | 1 |
| TER | enoyl reductase | EC.1.3.1.- | 7 |
| PCH | palmitoyl-CoA hydrolase | EC.3.1.2.2 | 1 |
| PPT | palmitoyl-protein thioesterase | EC.3.1.2.22 | 3 |
| **Fatty acid desaturation** | | | |
| AAD | acyl-ACP desaturase | EC.1.14.19.2 | 6 |
| FAD2 | Δ12(ω6)-Desaturase (endoplasmic reticulum) | EC.1.14.19.- | 5 |
| FAD3 | Δ15(ω3)-Desaturase (microsomal) | EC.1.14.19.- | 2 |
| FAD6 | Δ12(ω6)-Desaturase (chloroplast) | EC.1.14.19.- | 1 |
| FAD7 | Δ15(ω3)-Desaturase (chloroplast) | EC.1.14.19.- | 5 |
| FAD8 | Δ15(ω3)-Desaturase (chloroplast) | EC.1.14.19.- | 1 |
| **Fatty acid metabolism** | | | |
| ACA1 | acetyl-CoA acyltransferase 1 | EC.2.3.1.16 | 8 |
| at08 | acetyl-CoA C-acyltransferase | EC.2.3.1.9 | 6 |
| ACADM | acyl-CoA dehydrogenase | EC.1.3.99.3 | 7 |
| ACODX | acyl-CoA oxidase | EC.1.3.3.6 | 11 |
| ACOL | long-chain acyl-CoA synthetase | EC.6.2.1.3 | 16 |
| ADH | alcohol dehydrogenase | EC.1.1.1.1 | 15 |
| ALDH | aldehyde dehydrogenase (NAD+) | EC.1.2.1.3 | 16 |
| MFP2 | enoyl-CoA hydratase/3-hydroxacyl-CoA dehydrogenase | EC.4.2.1.17 1.1.1.35 1.1.1.2114 | 14 |

doi:10.1371/journal.pone.0074441.t003
and 44 (0.66%) SSRs, respectively (Table 6). Polymorphisms in these potential markers will be the focus of future research. To increase the authenticity of SNP and InDel identification, we also filtered the results based on stricter multiple criteria, including read depth and allele frequency, compared with previous studies (see materials and methods). In total, 16,925 SNPs distributed across 4,401 different isotig groups that corresponded to 4,234 different unigenes had a total length of 6.10 Mb (Figure 5), resulting in an SNP occurrence rate of 0.003 per base and four SNPs per unigene. Transitions contained 9,225 (54.51%) SNPs and were primarily transversions (7,700 SNPs, 45.49%). The A/G and C/T transition genotypes had similar percentages, but among the four transversion genotypes, C/G transversions were less frequent than the other three (A/T, G/T, A/C). A total of 6,201 Indels were identified from 3,161 isotig groups that corresponded to 3,094 unigenes, with a total length of 4.41 Mb (Figure 5), which indicates there were two variations per unigene. Insertions (2,862, 46.15%) were slightly less common than deletions (3,339, 53.85%). The length of insertions ranged from 1 to 24 bp, and deletions were 1 to 22 bp in length.

We summed the SSRs, SNPs and InDels that were located in the 281 unigenes involved in the FA and TAG biosynthesis and metabolism pathways, which encoded oleosin, caleosin, steroleosin and TFs that regulated seed oil deposition. As a result, 26 SSRs, 194 SNPs and 60 InDels were found in 74 unigenes, which covered 31 enzymes or proteins and two TFs (ABI3 and HSI2-L1) (Table S4, Table S5 and Table S6). The remaining enzymes, such as MAT, EAR, FATA, ADH, ALDH, PCH, FAD6, PP and TGL4, and TFs, such as L1L, ADOF1, EMF2, HSI2, AP2 and GL2, contained no SSRs, SNPs or InDels.

Because we used larger samples from different plant regions and 1.25 runs to obtain the transcriptome, we found 29,833 (SSRs, Table 4.

Table 4. Enzymes related to TAG biosynthesis and metabolism identified by annotation of the yellow horn unigenes.

| Symbol | Enzymes/Protein | EC Number | Number of unigenes |
|--------|----------------|-----------|--------------------|
| TAG biosynthesis |
| GK | glycerol kinase | EC2.7.1.30 | 1 |
| AT51 | glycerol-3-phosphate acyltransferase (chloroplast) | EC2.3.1.15 | 3 |
| GPAT5 | glycerol-3-phosphate acyltransferase 5 | EC2.3.1.15 | 1 |
| GPAT6 | glycerol-3-phosphate acyltransferase 6 | EC2.3.1.15 | 3 |
| GPAT8 | glycerol-3-phosphate acyltransferase 8 | EC2.3.1.15 | 2 |
| GPAT9 | glycerol-3-phosphate acyltransferase 9 | EC2.3.1.15 | 3 |
| LPAT 1 | lysophosphatidyl acyltransferase 1 (chloroplast) | EC2.3.1.51 | 4 |
| LPAT 2 | lysophosphatidyl acyltransferase 2 | EC2.3.1.51 | 2 |
| LPAT 4 | lysophosphatidyl acyltransferase 4 | EC2.3.1.51 | 3 |
| LPAT 5 | lysophosphatidyl acyltransferase 5 | EC2.3.1.51 | 2 |
| PP | phosphatidate phosphatase | EC3.1.3.4 | 4 |
| DGAT1 | diacylglycerol O-acyltransferase 1 | EC2.3.1.20 | 1 |
| PDAT1 | phospholipid:diacylglycerol acyltransferase 1 | EC2.3.1.158 | 4 |
| ACBP | acyl CoA binding protein | | |

Acyl editing

| Symbol | Enzymes/Protein | EC Number | Number of unigenes |
|--------|----------------|-----------|--------------------|
| PLA2 | Phospholipase A2 | EC3.1.1.4 | 1 |
| LPCAT | lysophosphatidylcholine acyltransferase | EC2.3.1.23 2.3.1.67 | 5 |

TAG metabolism

| Symbol | Enzymes/Protein | EC Number | Number of unigenes |
|--------|----------------|-----------|--------------------|
| TGL4 | TAG lipase | EC3.1.1.3 | 3 |

doi:10.1371/journal.pone.0074441.t004

Table 5. Putative transcription factors related to the oil biosynthesis in yellow horn.

| TF Name | TF Family | Unigene number |
|---------|-----------|----------------|
| ABI3 (Abscisic Acid Insensitive 3) | ABI3VP1 | 2 |
| L1L (Leafy Cotyledon 1-Like) | CCAAT-HAP3 | 1 |
| ADOF1 (Arabidopsis Dof Zinc Finger Protein 1) | C2C2-Dof | 1 |
| EMF2 (Embryonic Flower 2) | C2H2 | 5 |
| HSI2 (High-Level Expression of Sugar-Inducible Gene 2) | ABI3VP1 | 5 |
| HSI2-L1/HSIL1 (HSI2-Like 1) | ABI3VP2 | 7 |
| AP2 (APETALA2) | AP2-EREBP | 1 |
| GL2 (GLABRA2) | Homeobox | 3 |

doi:10.1371/journal.pone.0074441.t005

Transcriptome of Yellow Horn
SNPs and InDels potential molecular markers in 9,720 unigenes (18.74%) with a total length of 10.37 Mb, for an average of 3.07 markers per unigene and spaces of 347.58 bp between markers. These markers will play significant roles not only in the production of genetically improved varieties of yellow horn with different oil compositions, increased oil yield and improved agronomic characteristics, but also for studying the evolution and origin of Xanthoceras, and even the Sapindaceae.

Conclusions

In this study, we report the first comprehensive yellow horn sequencing effort using 454 GS FLX. Transcriptome analysis using four tissues (buds, leaves, flowers and developing seeds) of yellow horn found 51,867 unigenes (45 to 10,088 bp), which corresponded to 36.1 Mb and mean, N50 and median lengths of 696, 928 and 570 bp, respectively. These unigenes provide a strong basis for future genomic research to develop microarrays for gene expression assays and can serve as a reference transcriptome for future yellow horn RNA-seq experiments. In addition, 281 unigenes that code for key enzymes and TFs that are involved in reconstructed metabolic pathways for FA and TAG biosynthesis and metabolism were identified. Moreover, a large number of potential molecular markers (6,707 SSRs, 16,925 SNPs and 6,201 InDels) were predicted. Among them, 26 SSRs, 194 SNPs and 60 InDels were identified in 74 unigenes that are related to oil biosynthesis and metabolism. These findings will make a substantial contribution to efforts to improve crop characteristics and will accelerate the breeding of new yellow horn varieties.

Materials and Methods

Collection of Tissues for RNA Extraction

Yellow horn is widely distributed in China, so it has not been listed as an endangered or protected species. In this study, we used yellow horn buds, leaves (young and mature leaves), flowers, and developing seeds (10, 20, 30, 40, 50, 60, and 70 days after pollination) which were collected from 30 plants at the two locations. One is located in Xishan forest farm in Haidian, Beijing, China (E116°04’, N40°03’). This forest farm belongs to Beijing Forestry University and the yellow horn trees in this farm have been used for scientific research for several years. The other one is located in a small hill in Chengde city, Hebei province, China (E117°55’, N40°59’). The yellow horn trees which growing in this location were planted for scientific research by Dr. Ao Yan, a co-authors of this research. These samples were immediately frozen in liquid nitrogen and stored at −80°C. Because our research team is also engaged in the work of conservation and utilization of wild plant resources, we confirm that the field studies did not involve any endangered or protected species.

cDNA Library Construction and 454 Sequencing

Total RNA was extracted separately from the buds, leaves, flowers and developing seeds using RNeasy Plant Mini Kits (Qiagen, Inc., Valencia, CA, USA) following the manufacturer’s protocol. Extracted RNA was qualified and quantified using a Nanodrop ND-1000 Spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA) and all the samples showed a 260/280 nm ratio from 1.9 to 2.1. Poly(A)⁺ RNA was purified from total RNA by Oligotex mRNA Mini Kit (Qiagen, Inc., Valencia, CA, USA) following the manufacturer’s protocol. After that, equal quantities of total RNA from buds, leaves, flowers and seeds were mixed together. A total of 10 μg of total RNA was used for cDNA library construction. cDNA library construction and normalisation were performed using protocols described previously [72]. The resulting library was sequenced

| Repeat motif | Repeat numbers | Number of SSRs | Percentage (%) |
|--------------|----------------|----------------|----------------|
| Di-          | 1,052          | 3,598          | 53.65          |
| Tri-         | 1,687          | 2,976          | 44.37          |
| Tetra-       | 53             | 72             | 1.07           |
| Penta-       | 15             | 17             | 0.25           |
| Hexa-        | 32             | 44             | 0.66           |
| Total        | 1,787          | 6,707          |                |

Percentage (%) 26.64 27.03 14.82 10.27 6.74 4.25 10.24

Table 6. Number of SSRs in yellow horn.

doi:10.1371/journal.pone.0074441.t006

Figure 5. Distribution of putative single nucleotide polymorphisms (SNPs) and insertions and deletions (InDels) in the yellow horn transcriptome.

doi:10.1371/journal.pone.0074441.g005
by means of one and a quarter 454 plate runs on a GS-FLX Titanium platform (Roche, USA).

Analysis of 454 Transcriptome Sequencing Results

The raw reads were trimmed before assembly. First, adaptors and SMART primers that were used in the pyrosequencing reactions were cut using the SeqClean software. Then, we used the LUCY2 software [73] to remove low-quality regions and bases. Trimmed reads that were shorter than 45 bp were discarded and the remaining reads were assembled into isotigs and singlets using Newbler (version 2.6) [40]. Finally, after clustering the isotigs and singlets using CD-HIT (version 4.5.6) [74], the obtained unigenes were used in further analyses.

To understand their functions, the yellow horn unigenes were annotated using BLASTxs alignment with an E-value cut-off of 10\(^{-5}\) against the following protein databases: NCBI non-redundant (NR), Swiss-Prot, Conserved Domain Database (CDD), Pfam protein families database (Pfam), UniProtKB/TrEMBL Protein Database (TrEMBL), Clusters of Orthologous Groups of proteins (COG), Gene Ontology (GO), Kyoto Encyclopaedia of Genes and Genomes (KEGG), and The Arabidopsis Information Resource (TAIR). GO functional classifications and KEGG pathway assignments were performed, as was described previously [75].

Detection of SSRs, SNPs, InDels and TFs

To detect simple sequence repeats (SSRs) in the yellow horn transcriptome, the MISA [http://pgrc.ipk-gatersleben.de/misa/] software was used to identify all 2–6-bp motifs in the unigenes. The minimum repeat unit size was set at six for di-nucleotides and five for tri-, tetra-, penta-, and hexa-nucleotides.

The ssahaSNP software tool [76] was used to identify SNPs and InDels (1 to 100 bp in size) that had high coverage depths. To add the polymorphism and reliability of the SNPs and InDels, as in previous studies [26,77,78], we kept only SNPs and InDels that met the following strict quality criteria: (1) For SNP detection, we identified SNPs that had coverage depths of at least 20 and an alternate allele that was present at a minimum frequency of 20% in all isotigs that contained at least 20 reads. (2) For InDel detection, similar to the standard for SNPs, for insertions or deletions of one base, the coverage depth of the isoig and the InDel allele frequency were set at 20 and 20%, respectively. For insertions or deletions of two or more bases, the coverage depth of the isoig and the InDel allele frequency were set at 10 and 10%, respectively.

To identify TFs, a BLAST search for all unigenes was conducted using Arabidopsis transcription factor database with shared identities >77%, as described previously [23].

Data Deposition

The Roche 454 reads of yellow horn were deposited in the NCBI and can be accessed in the Short Read Archive (SRA) under accession number SRP026671.

Supporting Information

Table S1 Pathway annotation of unigenes from yellow horn. (XLS)

Table S2 Unigenes annotated as olescin, caloecin and stereo- sin. (XLS)

Table S3 Putative transcription factors encoding unigenes in yellow horn. (XLS)

Table S4 SSRs located in unigenes related to oil biosynthesis and metabolism. (XLS)

Table S5 SNPs located in unigenes related to oil biosynthesis and metabolism. (XLS)

Table S6 InDels located in unigenes related to oil biosynthesis and metabolism. (XLS)

Acknowledgments

We thank Dr. Shihui Niu (College of Biological Science and Biotechnology, Beijing Forest University) for assistance in drawing some figures.

Author Contributions

Conceived and designed the experiments: YLL WL ZXZ. Performed the experiments: YLL. Analyzed the data: YLL. Contributed reagents/materials/analysis tools: YLL ZDH YA. Wrote the paper: YLL WL ZXZ.

References

1. Durrett TP, Benning C, Ohlrogge J (2008) Plant triacylglycerols as feedstocks for the production of biofuels. The Plant Journal 54: 593-607.

2. Panzi S, Garcia I, Lopez-Gimenez F, Luque de Castro M, Dorado G, et al. (2009) The ideal vegetable oil-based biodiesel composition: a review of social, economical and technical implications. Energy & Fuels 23: 2325-2341.

3. Dorado M, Cruz F, Palomar J, Lopez F (2006) An approach to the economics of two vegetable oil based biodiesel in Spain. Renewable Energy 31: 1231-1237.

4. Fairless D (2007) Biodiesel: the little shrub that could-maybe. Nature 449: 652-655.

5. Wang F, Xing X, Liu C (2009) Biodiesel in China: opportunities and challenges. In Vitro Cellular & Developmental Biology-Plant 45: 342-349.

6. Huang YH, Wu JH (2008) Analysis of biodiesel promotion in Taiwan. Renewable and Sustainable Energy Reviews 12: 1176–1186.

7. Yu H, Zhou S (2009) Preparation of biodiesel from Xanthoceras sorbifolia Bunge. Journal of Natural Medicines 64: 80-84.

8. Yu H, Zhou S (2009) Preparation of biodiesel from Xanthoceras sorbifolia Bunge. seed oil. China Oils and Fats 3: 43–45.

9. Zhao S, Zou YG, Fu YJ, Luo M, Liu W, et al. (2010) Supercritical carbon dioxide extraction of seed oil from yellow horn (Xanthoceras sorbifolia Bunge) and its anti-oxidant activity. Bioresource technology 101: 2537-2544.

10. Kong WB, Liang JY, Ma ZN, Zhang J (2011) Research advance of xanthoceras sorbifolia Bunge oil. China Oils and Fats. 36 67–72.

11. Yu H, Zhou S (2009) Preparation of biodiesel from Xanthoceras sorbifolia Bunge. Journal of Natural Medicines 64: 80-84.

12. Hartung K (1996) Chemical and physical properties of vegetable oil esters and their effect on diesel fuel performance. Biomass 9: 1–17.

13. Yao Z, Wang L, Qi J (2009) Biosorption of methylene blue from aqueous solution using a bioenergy forest waste. Xanthoceras sorbifolia seed coat. CLEAN-Soil, Air, Water 37: 642-648.

14. Zhang J, Chen G, Sun Q, Li Z, Wang Y (2012) Forest biomass resources and utilization in China. African Journal of Biotechnology 11: 9302-9307.

15. Yao ZY, Qi JH, Yin LM (2013) Biodiesel production from Xanthoceras sorbifolia in China: Opportunities and challenges. Renewable and Sustainable Energy Reviews 24: 57–63.

16. Zhang S, Zou YG, Fu YJ, Luo M, Zhang DY, et al. (2010) Rapid microwave-assisted transesterification of yellow horn oil to biodiesel using a heteropolyacid solid catalyst. Bioresource technology 101: 931–936.

17. Fu YJ, Zou YG, Wang LL, Zhang XJ, Liu W, et al. (2008) Determination of fatty acid methyl esters in biodiesel produced from yellow horn oil by LC. Chromatographia 67: 9-14.

18. Fu YJ, Zou YG, Wang LL, Zhang XJ, Liu W, et al. (2008) Determination of fatty acid methyl esters in biodiesel produced from yellow horn oil by LC. Chromatographia 67: 9-14.

19. Li J, Fu YJ, Qu XJ, Wang W, Luo M, et al. (2012) Biodiesel production from yellow horn (Xanthoceras sorbifolia Bunge.) seed oil using ion exchange resin as heterogeneous catalyst. Bioresource technology 108: 112-118.
46. Shimakata T, Stumpf PK (1983) Purification and characterization of
45. Dehesh K, Tai H, Edwards P, Byrne J, Jaworski JG (2001) Overexpression of 3-
PLOS ONE | www.plosone.org 12 September 2013 | Volume 8 | Issue 9 | e74441
44. Pidkowich MS, Nguyen HT, Heilmann I, Ischebeck T, Shanklin J (2007)
30. Li H, Dong Y, Yang J, Liu X, Wang Y, et al. (2012)
39. Faergeman NJ, Knudsen J (1997) Role of long-chain fatty acyl-CoA esters in the
37. Slabas AR, Fawcett T (1992) The biochemistry and molecular biology of plant
35. Huang AH (1992) Oil bodies and oleosins in seeds. Annual review of plant
26. Sloan DB, Keller SR, Berardi AE, Sanderson BJ, Karpovich JF, et al. (2012)
25. Zhang J, Liang S, Duan J, Wang J, Chen S, et al. (2012)
23. He M, Wang Y, Hua W, Zhang Y, Wang Z (2012)
22. Metzker ML (2009) Sequencing technologies-the next generation. Nature
21. Genovese CA, Rojas A, Butelli E, Tegli A, Emili A, et al. (2012) De novo
20. De novo expression of an Arabidopsis cDNA encoding a diacetylglcyerol acyltransferase
19. Li X, Hea S, Su M, Yang M, Shen S, et al. (2010) Major energy plants and their potential for bioenergy development in China. Environmental management 46:
18. Bates PD, Smythe S, Oehlrogge J (2013) Biochemical pathways in seed oil synthesis. Current opinion in plant biology 16: 338–364.
17. Jain R, Colley M, Lai K, Kumar A, MacKenzie D (2000) Enhancement of seed oil content by expression of glycerol-3-phosphate acyltransferase genes. Biochemical Society Transactions 28: 959–960.
16. Tasaka Y, Nishiyama Y, Mohanty P, Ohba T, et al. (1996) Targeted mutagenesis of acylipid desaturases in Synechocystis: evidence for the important roles of polymethylated acylipid desaturases in growth, respiration and photosynthesis. The EMBO journal 15: 6146.
15. Riimani-Yazdi H, Haznedaroglu B, Bibby K, Peccia J (2011) Transcriptome sequencing and annotation of the microalga Dunaliella tertiolecta: Pathway description and gene discovery for production of next-generation biofuels. BMC genomics 12: 146.
14. Jaworski JG, Clough RG, Barmum SR (1989) A crueltine insensitive short chain 3-ketoacyl-acyl carrier protein synthase in Spinacea oleracea leaves. Plant Physiology 90: 41–44.
13. Pathiak AK, Hopkins MF, Wang TW, McNamara LM, Lo M, et al. (2007) Characterization of a plastid triacylglycerol lipase from Arabidopsis. Plant Physiology 143: 1372–1384.
12. Wang X, Guo J, Lambert KN, Lin Y (2007) Developmental control of Arabidopsis seed oil biosynthesis. Planta 226: 773–783.
11. Baud S, WilliIne S, To A, Rochat C, Lepineux L (2009) Role of WRINKLED1 in the transcriptional regulation of glycolytic and fatty acid biosynthetic genes in Arabidopsis. The Plant Journal 60: 933–947.
10. Tasaka Y, Li S, Chou HH (2004) LUCY2: an interactive DNA sequence quality trimming and filtering tool. Bioinformatics 20: 2088–2089.
9. Palanivannan SK, James S, Sun H, Lamb RS, Davaluri RV, et al. (2006) AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factor networks into regulatory networks. Plant Physiology 140: 818–829.
8. Mazur B, Krehbiel E, Tingey S (1999) Gene discovery and product development for grain quality traits. Science 285: 372–375.
7. O’Malley DM, Meckead SE (1994) Marker assisted selection for breeding value in forest trees. Forest Genetics 1: 207–218.
6. Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. Nature Reviews Genetics 12: 111–122.
5. Hu X, Sullivan-Gilbert M, Gupta M, Thompson SA (2006) Mapping of the loci controlling oleic and linolenic acid contents and development of oil fad2 and fad3 allele-specific markers in canola (Brassica napus L.). Theoretical and Applied Genetics 113: 497–507.
4. Gupte V, Mihukapadyah A, Arumugam N, Sodhi Y, Dental P, et al. (2004) Molecular tagging of erucic acid trait in oilseed mustard (Brassica juncea) by QTL mapping and single nucleotide polymorphisms in FAD2 gene. Theoretical and Applied Genetics 108: 743–749.
3. Wang X, Xu S, Jiang Y, Jiang J, Li X, et al. (2013) De novo Sequence Assembly and Characterization of Leymus arenarius Transcriptome Using GS FLX Titanium Platform of 454 Pyrosequencing. PloS one 8: e60449.
2. Li S, Chou HH (2004) LUCY2: an interactive DNA sequence quality trimming and vector removal tool. Bioinformatics 20: 2982–2983.
1. Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26: 680–682.
0. Wang X, Xu S, Jiang Y, Jiang J, Li X, et al. (2013) De novo Sequence Assembly and Characterization of Leymus arenarius Transcriptome Using GS FLX Titanium Platform of 454 Pyrosequencing. PloS one 8: e60449.
-1. Ning Z, Cox AJ, Mulliken JK (2001) SSAHA: a fast search method for large DNA databases. Genome research 11: 1725–1729.
77. Edwards CE, Parchman TL, Weekley CW (2012) Assembly, gene annotation and marker development using 454 floral transcriptome sequences in *Ziziphus celata* (Rhamnaceae), a highly endangered, Florida endemic plant. DNA research 19: 1–9.

78. Blanca J, Catáizares J, Roig C, Ziarsolo P, Nuez F, et al. (2011) Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). BMC genomics 12: 104.