Plant Biotechnology 29, 123–130 (2012)
DOI: 10.5511/plantbiotechnology.12.0515a

Original Paper

Upgraded genomic information of *Jatropha curcas* L.

Hideki Hirakawa[1], Suguru Tsuchimoto[2], Hiroe Sakai[2], Shinobu Nakayama[1], Tsunakazu Fujishiro[1], Yoshie Kishida[1], Mitsuyo Kohara[1], Akiko Watanabe[1], Manabu Yamada[1], Tomoyuki Aizu[3], Atsushi Toyoda[3], Asao Fujiyama[3,4], Satoshi Tabata[1], Kiichi Fukui[5], Shusei Sato[1,*]

[1] Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan; [2] Plant Bioengineering for Bioenergy Laboratory Contributed by SEI CSR foundation, Graduate School of Engineering, Osaka University, Suita, Osaka 565-0871, Japan; [3] Center for Genetic Resource Information, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan; [4] Principles of Informatics Research Division, National Institute of Informatics, Tokyo 101-8430, Japan; [5] Laboratory of Dynamic Cell Biology, Department of Biotechnology, Graduate School of Engineering, Osaka University, Suita, Osaka 565-0871, Japan
* E-mail: ssato@kazusa.or.jp Tel: +81-438-52-3923 Fax: +81-438-52-3924

Received January 11, 2012; accepted May 15, 2012 (Edited by T. Demura)

Abstract In order to upgrade the genome sequence information of *J. curcas* L., we integrated *de novo* assembly of a total of 537 million paired-end reads generated from the Illumina sequencing platform into the current genome assembly which was obtained by a combination of the conventional Sanger method and the Roche/454 sequencing platform. The total length of the upgraded genome sequences thus obtained was 297,661,187 bp consisting of 39,277 contigs. The average and N50 lengths of the generated contigs were 7,579 bp and 15,950 bp, both of which were increased fourfold from the previous genome assembly. Along with genome sequence upgrading, the currently available transcriptome data were collected from the public databases and assembled into 19,454 tentative consensus sequences. Based on a comparison between these tentative consensus sequences of transcripts and the predictions of computer programs, a total of 30,203 complete and partial structures of protein-encoding genes were deduced. The number of genes with complete structures was increased about threefold from the previous genome annotation. By applying the upgraded genome sequence and predicted protein-coding gene information, the number and features of the tandemly arrayed genes, syntenic relations between Jatropha and other plant genomes, and structural features of transposable elements were investigated. The detailed information on the updated *J. curcas* genome is available at http://www.kazusa.or.jp/jatropha/.

Key words: *Jatropha curcas*, genome sequencing, transcriptome sequences, tentative consensus sequence, tandem gene duplication, database.

Jatropha curcas L. is a perennial small tree or large shrub that belongs to the Euphorbiaceae family. *J. curcas* is endemic to central America but is distributed throughout the tropics and subtropics of Asia and Africa. *J. curcas* is an important non-edible oilseed crop with great potential for the production of biodiesel fuel. Since *J. curcas* is an undomesticated plant, its positive attributes in terms of breeding and utilization are not fully understood.

In order to accelerate its genetic improvement, it is desirable to understand the genome information of *J. curcas*. With this goal in mind, we have analyzed the genome sequence of *J. curcas* by applying combined sequencing methods, and have made the obtained sequence information available through the public and web databases. The accumulated genome information (JAT_r3.0) was 285,858,490 bp consisting of 120,586 contigs and 29,831 singlets, and this accounted for approximately 95% of the gene-containing regions. A total of 40,929 complete and partial structures of protein-encoding genes have been deduced on the accumulated genome sequences. However, the majority of the predicted genes were partially predicted ones as the contig lengths were relatively short in JAT_r3.0. Further improvement of the genome sequence information is therefore needed.

Along with the genome sequence approach, several transcriptome analyses have been attempted. Natarajan et al. have reported 12,084 ESTs using a normalized cDNA library from developing seeds (Natarajan et al. 2010), and Costa et al. have reported 13,249 ESTs using non-normalized cDNA libraries from developing and germinating endosperm (Costa et al. 2010). In addition to these approaches using the conventional Sanger sequencing method, efforts have been made to accumulate transcriptome data for Jatropha by using a next-generation sequencer. In our previous report, we

This article can be found at http://www.jspcmb.jp/
Published online June 15, 2012

described the information on cDNA sequences from leaf and callus transcriptomes generated by pyrosequencing using a Roche/454 GS FLX sequencer. Recently, two additional cDNA sequence analyses using a Roche/454 sequencer were reported. Natarajan et al. reported a pyrosequencing analysis of bulked inserts of cDNA libraries constructed from RNA extracted from a mixture of roots, mature leaves, flowers, developing seeds, and embryos, and King et al. reported a pyrosequencing analysis of cDNA constructed from RNA extracted from three different stages of developing seeds. A small scale (2,210 reads) set of pyrosequencing reads was also deposited in the NCBI's SRA database during the process of EST-derived SSR marker development (Yadav et al. 2010). In order to fully utilize the accumulated transcriptome information of *J. curcas*, it would be preferable to have these data centralized and assembled into tentative consensus sequences, and to make them available to the research community through a database with a user-friendly interface.

In this study, we upgraded the genome sequence information by integrating the *de novo* assembly of 537 million paired-end reads generated by the Illumina sequencing platform. In addition, we accumulated the currently available transcriptome data of *J. curcas* and used assembled transcriptome sequences to assist in the gene assignment. The updated genome information and organized transcriptome data for Jatropha, which are provided on the renewed web database at http://www.kazusa.or.jp/jatropha/, will enhance both fundamental and applied research on *J. curcas* and related plants.

Materials and methods

Plant materials

The details of the plant material used in this study, a *J. curcas* line originating from the Palawan Island in the Philippines, were described in our previous report (Sato et al. 2011).

Genome sequencing and assembly

The strategy used to accumulate shotgun genomic sequences generated by an Illumina-solexa GAII sequencer is described in the previous report (Sato et al. 2011). The following four paired-end runs, two of which had been used in the previous study and deposited to DDBJ Sequence Read Archive (DRA), were used for the assembly: 36/36 bp (DRA000305), 50/31 bp (DRA000306), 51/51 bp, and 76/76 bp. These Illumina reads were filtered by quality using the fastq_quality_filter program with the parameters (-q 10 -p 10) and were trimmed using quality trimming with the fastq_quality_trimmer program with the parameters (-t 10 -l x; x=21) for the libraries of 36/36 bp, 50/31 bp and 51/51 bp, and (x=31) for the library of 76/76 bp, artifact trimming using the fastx_artifacts_filter program, and adaptor trimming using the fastx_clipper program with the parameters (-n -M 5). These programs were included in the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). After the filtering and trimming, the remaining reads were assembled by using SOAPdenovo 1.05 (Li et al. 2010) with an optimized kmer size of 31 bp. The constructed scaffolds were further assembled with the sequences of JAT_r3.0 released (Sato et al. 2011) by using PCAP.rep (Huang et al. 2006) with the parameters (-m 10 -l 50 -t 95 -v 0 -y 4). In the hybrid assembly, the contigs with coverage ≥50 were defined as repeat sequences.

Assembly of transcript sequences

The following *J. curcas* cDNA sequences generated using a 454 GS FLX sequencer were accumulated from the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra/): 534,137 reads and 456,913 reads of cDNA constructed from the mRNA of leaves and callus, respectively (DRX000446 and DRX000447), 383,937 reads of cDNA sequences generated from a mixture of five major tissue libraries (SRX035761), 195,692 reads of cDNA sequences of a seed transcript fragment library (SRX011411), and 2,210 reads of cDNA constructed from mRNA of leaves (SRX020243). In addition, 46,842 EST sequences of *J. curcas* were accumulated from dbEST in NCBI (http://www.ncbi.nlm.nih.gov/dbEST/). These sequences were assembled by using Newbler 2.6 (Roche Diagnostics, USA).

Gene finding and modeling

For the *ab initio* gene findings, protein-coding regions were predicted automatically using the AUGUSTUS program (Stanke et al. 2004) with the matrix trained by an *A. thaliana* gene set for the draft genomic sequence. To find the genes which were not predicted by AUGUSTUS, the genes assigned in JAT_r3.0, which were predicted by a combination of GeneScan and GeneMark.hmm, and the genes predicted in JAT_r4.5 were compared by BLASTN searches (Altschul et al. 1997). The genes of JAT_r3.0 that were not homologous to the genes predicted by AUGUSTUS were assigned as protein coding genes of JAT_r4.5. In addition, the unigene sets, which were not homologous to the predicted genes by AUGUSTUS but homologous to the NR database, were assigned as protein coding genes of JAT_r4.5. The protein-coding genes assigned in this manner were denoted by IDs with the contig names followed by sequential numbers from one end to another. They were classified into four categories based on the status of coding sequences and the sequence similarity to registered genes: genes with complete structure, pseudo genes, genes with partial structure, and transposons/retrotransposons.

Functional assignment and classification of potential protein-coding genes

To assign gene families, functional domains, GO terms, and GO accession numbers (Ashburner et al. 2000), the predicted genes were searched against InterPro using InterProScan (Hunter et al. 2009) software. Genes with E-values ≤1.0 were selected. GO terms were grouped into plant GO slim categories using the map2slim program (http://www.geneontology.org/GO.slims.

shtml). The predicted genes were also searched against TrEMBL database (Bairoch and Apweiler 1996) using BLASTP program with E-value cut-off of 1E-20. The predicted protein-encoding genes were mapped onto KEGG metabolic pathways (Ogata et al. 1999) using the BLASTP program against the GENES database (Ogata et al. 1999). Thresholds of amino acid sequence identity ≥25% and of length coverage of the query sequence ≥50% with a cut-off E-value ≤1E-10 were applied.

Synteny and tandem gene duplication analysis

Translated amino acid sequences of the products of genes assigned on JAT_r4.5 were compared with those in the reference genomes of *Arabidopsis* (TAIR10), soybean (Glyma1) and caster bean (http://castorbean.jcvi.org/downloads.php), and a BLASTP E-value of less than 1E-20 was considered to be significant. All-against-all similarity searches within the products of genes assigned on JAT_r4.5 were conducted using the BLASTP program to define the gene family.

Synteny blocks were surveyed on the JAT_r4.5 contigs on which five or more genes were predicted. A synteny block was defined as the region where three or more conserved homologs were located within 10 consecutive genes in each of two genomes. Tandem duplicated genes were surveyed on the JAT_r4.5 contigs on which two or more genes were predicted. Tandem duplicated genes were defined as genes in any gene pair that (1) have significant similarity to each other and/or to the same gene product of the reference genomes, and (2) are separated by five or fewer nonhomologous spacer genes.

Analysis of repetitive sequences

Member sequences of each transposable element were collected by BLAST search of the Jatropha genome using the repetitive sequence previously identified by RECON and RepeatMasker (Sato et al. 2011) as a query. Structure of each transposable element was deduced by aligning the member sequences by GENETYX-MAC ver.13 (GENETYX Corporation, Japan).

Results and discussion

Hybrid assembly of JAT_r3.0 and Illumina sequences

In the previous assembly of the *J. curcas* genome sequence (JAT_r3.0), we used the Roche/454 and Sanger reads to produce the assembly, and applied the Illumina reads, which have a different error profile than the Roche/454 data, to increase the accuracy of the genome assembly. In order to upgrade the JAT_r3.0 sequences, we used accumulated Illumina reads (a total of 83 times the genome coverage) to create a *de novo* assembly using the assembler SOAPdenovo 1.05. As a result of the assembly of Illumina sequences using SOAPdenovo 1.05, the total length of the genome sequences constructed was 221,111,674 bp consisting of 107,255 scaffolds (Figure 1). The average, maximum and N50 lengths were 2,062 bp, 208,313 bp and 11,315 bp, respectively. The obtained Illumina sequence assembly was then applied to the hybrid assembly with the JAT_r3.0 sequences by using assemble program PCAP.rep (Figure 1). The statistics of these assemblies are shown in Table 1. By the hybrid assembly, the total length of the genome sequence was 297,661,187 bp consisting of 28,665 super contigs (SCs), which were composed of assembled SOAPdenovo scaffolds and JAT_r3.0 sequences, and 10,612 unassembled contigs (UCs) with an average length of 7,579 bp. The maximum length of the sequences was 277,264 bp, and the N50 was 15,950 bp. The G+C content was 33.7%. Through the hybrid assembly, the numbers of sequence elements were decreased fourfold (from 150,417 contigs and singles in JAT_r3.0 to 39,277 SCs and UCs in JAT_r4.5), and the average and N50 lengths were increased fourfold (Table 1 and Supplementary Figure 1). The updated version of the *J. curcas* genome sequences was named JAT_r4.5. The resulting SCs and UCs were designated Jcr4S and Jcr4U, respectively, followed by a five-digit number.

Assembly of transcript sequences

In order to create tentative consensus sequences (TCs) of *J. curcas* transcripts, all of the available cDNA and EST sequences were collected from public databases. From the NCBI's SRA database, the following sequences generated by a Roche/454 GS FLX sequencer were accumulated: 534,137 reads and 456,913 reads of cDNA constructed from the mRNA of leaves and callus, respectively (DRX000446 and DRX000447), 383,937 reads of cDNA sequences generated from a mixture of five major tissue libraries deposited by SRM University (SRX035761), 195,692 reads of cDNA sequences of a seed transcript fragment library deposited by CNAP (SRX011411), and 2,210 reads of cDNA constructed from the mRNA of leaves deposited by the National Botanical Research Institute (SRX020243). In addition, 46,842 EST sequences generated by the Sanger sequencing method were accumulated from dbEST. These accumulated cDNA/EST sequences were assembled by using the program Newbler 2.6. As a result, a total of 19,454 TCs were obtained (Supplemental Table S1). The average length and the G+C content of TCs were 969 bp and 41.3%, respectively. Among the 19,454 TCs thus obtained, 19,435 TCs (99.9%) have corresponding genome sequences in JAT_r4.5. This indicates that the coverage of gene space in JAT_r4.5 was increased from that in JAT_r3.0 (95%).

Gene modeling

In order to take the advantage of the upgraded genome sequence information and accumulated transcriptome information, we renewed the gene modeling of the Jatropha genome. As an initial step for renewal of gene modeling, we implemented an *ab initio* gene prediction

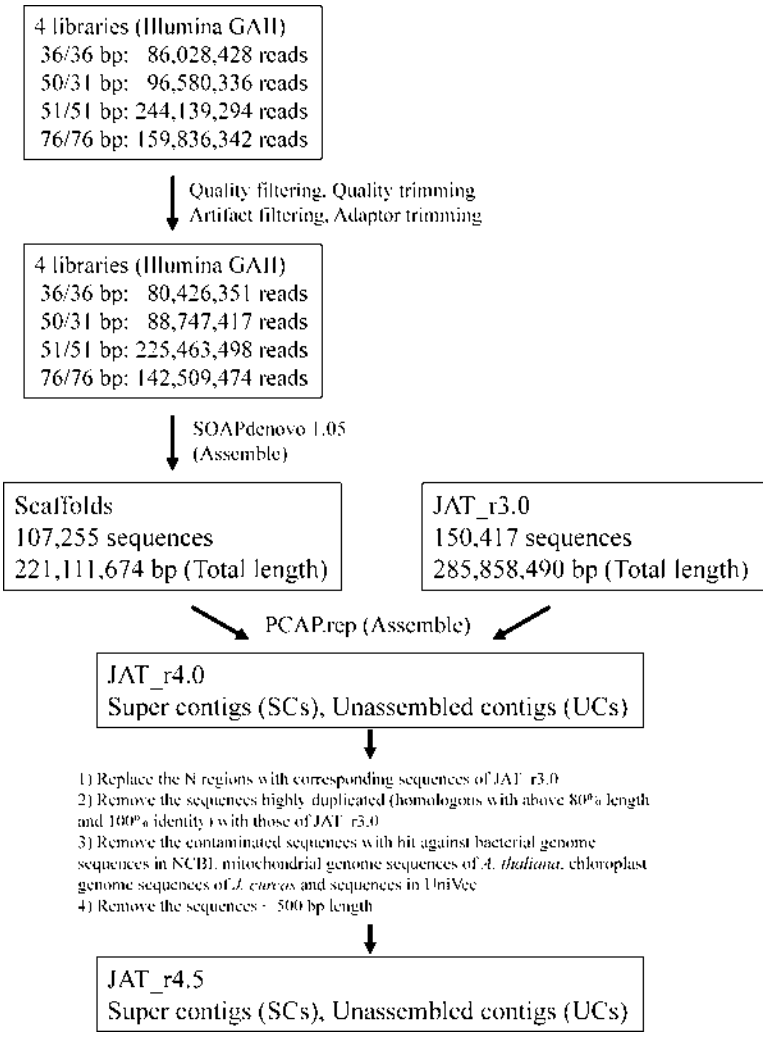

Figure 1. The strategy and status of hybrid assembly.

Table 1. Assemble statistics of previous (JAT_r3.0) and upgraded (JAT_r4.5) Jatropha genome sequences.

	JAT_r3.0	JAT_r4.5
Total length of sequence elements*	285,858,490	297,661,187
Total number of sequence elements	150,417	39,277
Average length of sequence elements	1,900	7,579
Maximum length of sequence elements	29,744	277,264
N50 length of sequence elements	3,833	15,950
G+C content (%)	34.3	33.7

*Contigs and singlets in JAT_3.0, SCs and UCs in JAT_r4.5.

program, AUGUSTUS, which had not been applied in the previous gene prediction on JAT_r3.0. As a result, a total of 50,313 genes were predicted. To complement the gene modeling by AUGUSTUS, a comparison of the transcriptome sequence information and predicted genes in JAT_r3.0 against the AUGUSTUS predictions in JAT_r4.5 was conducted. As a result, 7,124 gene modelings were appended. Finally, a total of 30,203 genes, in addition to the 17,575 transposon related genes and 2,124 putative pseudogenes, were assigned. Complete structures were predicted for 25,433 genes, while only partial structures were predicted for 4,770 genes mainly due to their location on the end regions of the contigs (Table 2). The number of the genes with complete structures was increased about threefold (from 9,870 in JAT_r3.0) (Table 2). The structural features of the genes with complete structures were close to those of the manually annotated genes on the 17 BAC clones described in the previous report (Sato et al. 2011) (Supplemental Table S2). This result indicates that the quality of the upgraded genome sequence in JAT_r4.5 would be close to the level of the completed BAC clone sequences.

Functional assignment and classification of potential protein-coding genes

As the number of genes with complete structure

Table 2. Statistics of predicted genes on previous (JAT_r3.0) and upgraded (JAT_r4.5) Jatropha genome.

	JAT_r3.0	JAT_r4.5
Total number of genes	40,929	30,203
Number of complete genes	9,870	25,433
Number of partially predicted genes*	31,059	4,770
Average length of genes (CDS)	689	1,058
Average length of complete genes (CDS)	780	1,109
Number of genes with GO annotations	35,070 (85.7%)	25,954 (85.9%)
Number of genes with similarity to the genes in TrEMBL database	31,822 (77.7%)	22,088 (73.1%)

* Pseudo-partial genes are included in JAT_r3.0.

prediction was increased, we re-analyzed the functional assignment and classification of potential protein-coding genes, and compared them to those in the previous report (Sato et al. 2011). A similarity search of the translated amino acid sequences of the 30,203 potential protein-encoding genes was performed using the TrEMBL database as a protein sequence library (Bairoch and Apweiler 1996). The results indicated that 22,088 (73.1%) genes had significant (E-value ≤1E-20) sequence similarity to genes in this database. Based on the sequence similarity and domain features assigned by the InterProScan program, the protein-encoding genes assigned in JAT_r4.5 were classified into plant GO slim categories (i.e., Biological Process [BP], Cellular Component [CC] and Molecular Function [MF]) (Carbon et al. 2009), and compared to the classification of genes in JAT_r3.0 along with those in castor bean (*Ricinus communis*) (31,221 genes) (Chan et al. 2010), which belongs to the same family as Jatropha, and *A. thaliana* (35,386 genes) (TAIR10). As a result, the proportion of the assigned genes in almost all GO slim categories was increased from JAT_r3.0, and the resulting proportion became close to that of the genes in the *A. thaliana* genome (Figure 2). The numbers of genes classified into the various GO slim categories are listed in Supplemental Table S3.

The number of genes assigned on the metabolic pathways in the KEGG database was increased from 2,213 in JAT_r3.0 to 2,402 in JAT_r4.5. By comparing the assignments of the genes of *R. communis* and *A. thaliana* on these metabolic pathways, 19 pathways, including "galactose metabolism" in carbohydrate metabolism, "biosynthesis of steroids" in lipid metabolism, "glycosylphosphatidylinositol (GPI)–anchor biosynthesis" in glycan biosynthesis and metabolism, and "retinol metabolism" in metabolism of cofactors and vitamins, contained enzyme(s) on which the genes in the Jatropha genome were solely mapped (Supplemental Table S4).

Tandem gene duplication

It has been hypothesized that gene duplication is an important driving force for adaptive evolution (Hanada et al. 2008). As the average length of the contigs, especially SCs, has been increased in JAT_r4.5, analysis of the gene compositions at the local level has become feasible. Among the 21,125 genes on 5,389 SCs and UCs with multiple gene prediction, 2,606 genes (12.3%) formed tandem arrays of two or more family genes on 1,170 sites in 1,085 contigs (Supplemental Table S5). The gene families frequently observed in these tandemly arrayed genes are summarized in Supplemental Table S6. As described in our previous report (Sato et al. 2011), genes for NBS-LRR (nucleotide-binding site and leucine-rich repeat) proteins and cytochrome P450 proteins were included in the list of frequently observed gene families. In addition, several gene families involved in secondary metabolism were represented in the list of frequently observed gene families, such as the alpha/beta-Hydrolases superfamily protein (42 genes on 20 sites), 2-oxoglutarate and Fe(II)-dependent oxygenase superfamily protein (38 genes on 16 sites), UDP-Glycosyltransferase superfamily protein (38 genes on 19 sites), NAD(P)-binding Rossmann-fold superfamily protein (32 genes on 12 sites), HXXXD-type acyl-transferase family protein (25 genes on 9 sites), and lipid-transfer protein (24 genes on 8 sites) (Supplemental Table S6). Total numbers of genes in these gene families in JAT_r4.5 are larger than those in the genome of *Arabidopsis* except for lipid-transfer protein (Supplemental Table S6). Thus, these genes would expand the variation and efficiency of secondary metabolism pathways in Jatropha. Gene families involved in membrane transport were also presented in the list of frequently observed gene families, such as the ATP-binding cassette transporter protein family (31 genes on 15 sites), major facilitator protein superfamily (25 genes on 12 sites) and MATE efflux protein family (23 genes on 11 sites) (Supplemental Table S6).

Synteny with sequenced plant genomes

As the lengths of contigs have been increased in JAT_r4.5, the syntenic relations between Jatropha and other plant genomes were reanalyzed by the status of conservation of relative gene positions on contigs. Among the 1,267 SCs and UCs with five or more predicted genes, conservation of the relative positions of three or more genes was observed in 1,117 SCs (88%) against genes predicted in the *R. communis* genomic sequences (Chan et al. 2010) (Supplemental Tables S7 and S8). The ratio of contigs with a syntenic relation against *R. communis* genomic sequences increased from the previous report, in which a syntenic relation was

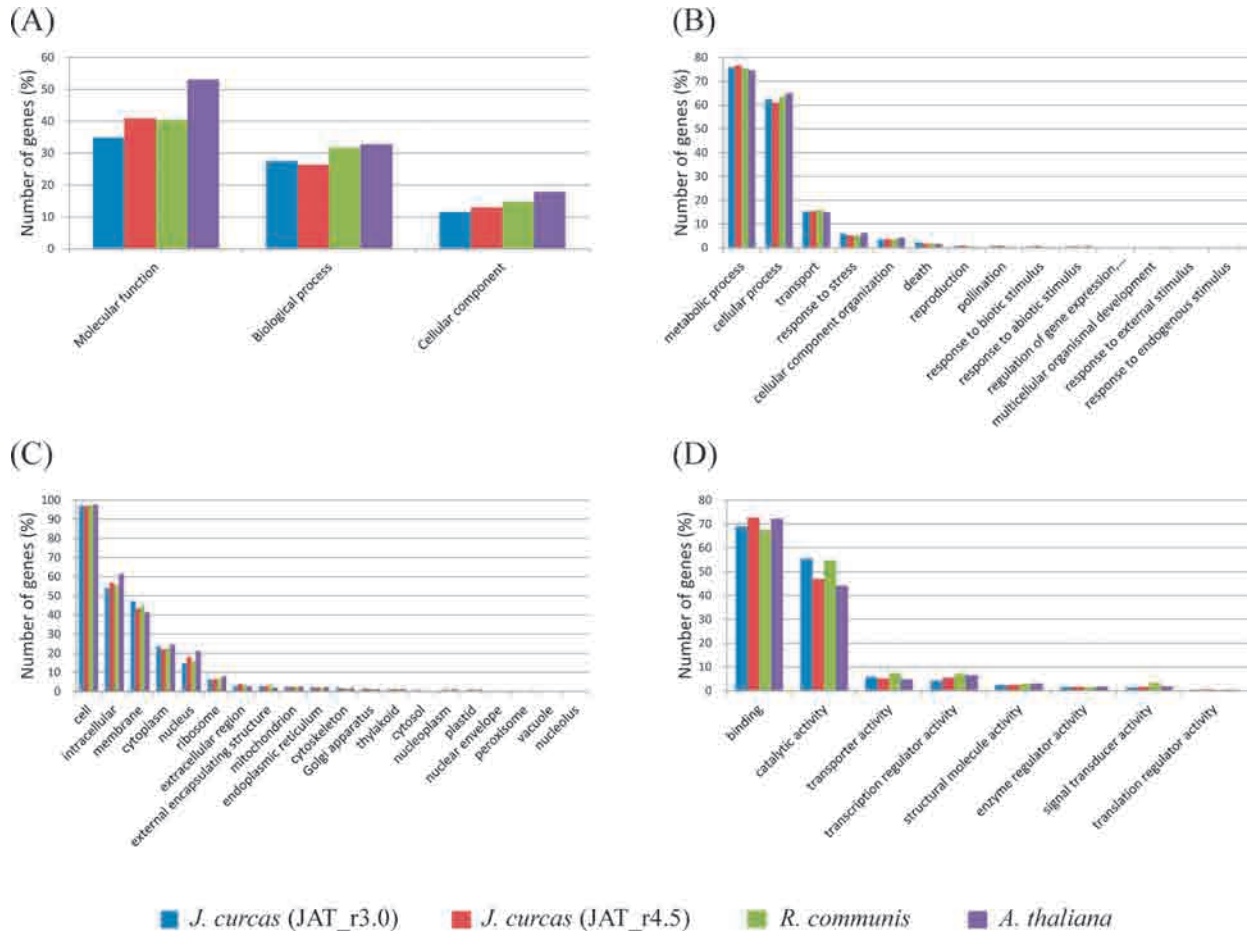

Figure 2. GO category classification. The percentages of genes classified into each GO slim category in *J. curcas* JAT_r3.0, *J. curcas* JAT_r4.5, *R. communis* and *A. thaliana* are respectively shown with blue, red, green and purple bars. (A) GO terms; (B) biological process; (C) cellular component; and (D) molecular function.

detected on 53% of the scaffolds with five or more genes. The main reasons for the increase may be the increases in the length of the contigs and in the number of genes with complete prediction. It appears that a significant degree of synteny can be expected within the family Euphorbiaceae. Moreover, the syntenic relationships between *J. curcas* and *R. communis* may contribute to our understanding of the genome organization of these plants. Actually, 138 SCs of JAT_r4.5 show a syntenic relation against two or more contigs of the *R. communis* genome sequences (Supplemental Table S7). On the other hand, 188 contigs of the *R. communis* genome sequences correspond to multiple SCs of JAT_r4.5 (Supplemental Table S7). These relations could be used to estimate the contig order in both the *J. curcas* and *R. communis* genome sequences.

A syntenic relationship was also detected against the genomes of *A. thaliana* and *G. max* to a lesser degree. Microsyntenic relations have been observed in 613 (48%) and 635 (50%) of the 1,267 SCs of the Jatropha genomic sequences, respectively (Supplemental Table S8). The microsyntenic relationships between these plant species may provide useful information for predicting gene organization in the ancestral genome of dicots.

Repetitive sequences

We have previously classified repeat sequences in the Jatropha genomic sequences, and revealed that transposable elements of class I and class II occupied 32% of the Jatropha genomic sequences (Sato et al. 2011). The upgraded genomic sequences allowed more precise structural analysis of transposable elements mainly because the lengths of the contig sequences have been increased. We analyzed structures of representative elements of both classes.

For class II, we identified two subfamilies of the Jatropha *hAT*-type transposable element (class II), named *JcDT1* (*Jatropha curcas* DNA-type transposon 1) and *JcDT2* (Supplementary Figure S2). The consensus sequence of *JcDT1* members was ca. 3.4 kb long with terminal inverted repeats (TIRs) of 12-bp long (TAGRCATGGCCA), and had 29 copies of a hexamer motif (AACCGG) in the subterminal region. It had an ORF that encodes a polypeptide of 722 amino acids long. The predicted amino acid sequence had high similarity with transposase (TPase) of *DART*, a subfamily of the

rice *hAT*-type transposable element, and had the BED zinc finger as well as the *hAT* family dimerization domain like TPase of *DART* (Fujino et al. 2009). These suggest that the ORF encodes TPase of *JcDT1*. On the other hand, the consensus sequence of *JcDT2* members was ca. 3.3 kb long with TIRs of 11-bp long (TAGGCATGGCC), and had 20 copies of a hexamer motif (AACCGG) in the subterminal region. The homology between the consensus sequences of *JcDT1* and *JcDT2* was 58%. *JcDT2* had an ORF that encodes a polypeptide of 681 amino acids long. The predicted amino acid sequence had high similarity with TPase of rice *DART*, and had the BED zinc finger and the *hAT* family dimerization domain, as in *JcDT1*.

For class I, we identified a subfamily of the Jatropha LINE (Long INterspersed Element), named *JcLINE1* (Supplementary Figure S2). LINE is a non-LTR retrotransposon that belongs to the class I transposable element. The consensus sequence of *JcLINE1* members showed that the length of *JcLINE1* was 5.7 kb or longer. The 3′ terminal sequences of *JcLINE1* members were mostly (A)n. The consensus sequence had two ORFs, named ORF1 and ORF2. ORF1 and ORF2 encode polypeptides of 476 and 1,302 amino acids, respectively. The predicted amino acid sequence of ORF2 had high similarity with ORF2 of *Arabidopsis* LINEs, *ATLN39* and *ATLN43* (Noma et al. 2000), and had domains of reverse transcriptase and endonuclease. These findings suggest that ORF2 encodes a protein that is essential for the retrotransposition of *JcLINE1*.

Databases

Information about the updated genomic sequences (SCs and UCs) is available through international databases (DDBJ/Genbank/EMBL) under accession numbers BABX02000001–BABX02066610 (66,610 entries). Paired-end reads of the genomes of 76/76 bp and 51/51 bp length by a GAII sequencer are available through the DDBJ Sequence Read Archive under accession numbers DRA000501 and DRA000502, respectively.

Detailed information on the updated *J. curcas* genome sequences, SCs and UCs, updated predicted genes and TCs and singleton ESTs generated in this study is also available at the renewed web database for the Jatropha genome information at http://www.kazusa.or.jp/jatropha/. This database contains a BLAST search engine against the updated genome sequences, TCs and singleton ESTs of accumulated transcript sequences, and nucleotide and amino acid sequences of predicted genes. Download of these sequence data sets is also feasible from the ftp site linked from this database. In addition, users can conduct a keyword search against the code and definitions of the IPR domains assigned on the predicted genes by the InterProScan program. The keyword search is also available against the definitions of the top hits for the BLAST searches against NR (http://www.ncbi.nlm.nih.gov), TrEMBL (Bairroch et al. 1996) and predicted gene data sets of *Arabidopsis*, soybean (ftp://ftp.jgi-psf.org/pub/JGI_data/Glycine_max/), *Lotus japonicus* (Sato et al. 2008), *Medicago truncatula* (http://www.medicago.org/genome/downloads/Mt3/), grape (http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/12X/annotation/) and caster bean (http://castorbean.jcvi.org/downloads.php). The database also contains the previous version of the assembly of JAT_r3.0, and the search engine of the corresponding gene codes in JAT_r3.0 and in JAT_r4.5.

Acknowledgements

We thank the Sumitomo Electric Industries, Ltd. for their philanthropic donation of funds to aid the Jatropha genome project celebrating the 110th anniversary of their foundation. This work was also supported by the Kazusa DNA Research Institute Foundation.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25: 25–29

Bairoch A, Apweiler R (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* 24: 21–25

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics* 25: 288–289

Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, Cahoon EB, Gedil M, Stanke M, Haas BJ, Wortman JR, Fraser-Liggett CM, Ravel J, Rabinowicz PD (2010) Draft genome sequence of the oilseed species *Ricinus communis*. *Nat Biotechnol* 28: 951–956

Costa GG, Cardoso KC, Del Bem LE, Lima AC, Cunha MA, de Campos-Leite L, Vicentini R, Papes F, Moreira RC, Yunes JA, Campos FA, Da Silva MJ (2010) Transcriptome analysis of the oil-rich seed of the bioenergy crop *Jatropha curcas* L. *BMC Genomics* 11: 462

Fujino K, Matsuda Y, Sekiguchi H (2009) Transcriptional activity of rice autonomous transposable element Dart. *J Plant Physiol* 166: 1537–1543

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol* 148: 993–1003

Huang X, Yang SP, Chinwalla AT, Hillier LW, Minx P, Mardis ER, Wilson RK (2006) Application of a superword array in genome assembly. *Nucleic Acids Res* 34: 201–205

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A,

Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* 37(Database issue): D211–D215

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20: 265–272

Natarajan P, Kanagasabapathy D, Gunadayalan G, Panchalingam J, Shree N, Sugantham PA, Singh KK, Madasamy P (2010) Gene discovery from *Jatropha curcas* by sequencing of ESTs from normalized and full-length enriched cDNA library from developing seeds. *BMC Genomics* 11: 606

Noma K, Ohtsubo H, Ohtsubo E (2000) ATLN elements, LINEs from *Arabidopsis thaliana*: identification and characterization. *DNA Res* 7: 291–303

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27: 29–34

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, Fujishiro T, Katoh M, Kohara M, Kishida Y, Minami C, Nakayama S, Nakazaki N, Shimizu Y, Shinpo S, Takahashi C, Wada T, Yamada M, Ohmido N, Hayashi M, Fukui K, Baba T, Nakamichi T, Mori H, Tabata S (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res* 15: 227–239

Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N, Takahashi C, Nakayama S, Kishida Y, Kohara M, Yamada M, Tsuruoka H, Sasamoto S, Tabata S, Aizu T, Toyoda A, Shin-i T, Minakuchi Y, Kohara Y, Fujiyama A, Tsuchimoto S, Kajiyama S, Makigano E, Ohmido N, Shibagaki N, Cartagena JA, Wada N, Kohinata T, Atefeh A, Yuasa S, Matsunaga S, Fukui K (2011) Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. *DNA Res* 18: 65–76

Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* 32(Web Server issue): W309-12

Yadav HK, Ranjan A, Asif MH, Mantri S, Sawant SV, Tuli R (2011) EST-derived SSR markers in *Jatropha curcas* L.: development, characterization, polymorphism, and transferability across the species/genera. *Tree Genet Genomes* 7: 207–219