Compendium of Plant Genomes
Series Editor: Chittaranjan Kole

Suguru Tsuchimoto *Editor*

The Jatropha Genome

Springer

Compendium of Plant Genomes

Series Editor

Chittaranjan Kole, Raja Ramanna Fellow, Department of Atomic Energy, Government of India, Kalyani, India

Whole-genome sequencing is at the cutting edge of life sciences in the new millennium. Since the first genome sequencing of the model plant Arabidopsis thaliana in 2000, whole genomes of about 70 plant species have been sequenced and genome sequences of several other plants are in the pipeline. Research publications on these genome initiatives are scattered on dedicated web sites and in journals with all too brief descriptions. The individual volumes elucidate the background history of the national and international genome initiatives; public and private partners involved; strategies and genomic resources and tools utilized; enumeration on the sequences and their assembly; repetitive sequences; gene annotation and genome duplication. In addition, synteny with other sequences, comparison of gene families and most importantly potential of the genome sequence information for gene pool characterization and genetic improvement of crop plants are described.

More information about this series at http://www.springer.com/series/11805

Suguru Tsuchimoto
Editor

The Jatropha Genome

Springer

Editor
Suguru Tsuchimoto
Graduate School of Engineering
Osaka University
Suita, Osaka
Japan

ISSN 2199-4781 ISSN 2199-479X (electronic)
Compendium of Plant Genomes
ISBN 978-3-319-49651-1 ISBN 978-3-319-49653-5 (eBook)
DOI 10.1007/978-3-319-49653-5

Library of Congress Control Number: 2017939319

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This book series is dedicated to
my wife Phullara, and our children Sourav,
and Devleena

Chittaranjan Kole

Preface to the Series

Genome sequencing has emerged as the leading discipline in the plant sciences coinciding with the start of the new century. For much of the twentieth century, plant geneticists were only successful in delineating putative chromosomal location, function, and changes in genes indirectly through the use of a number of 'markers' physically linked to them. These included visible or morphological, cytological, protein, and molecular or DNA markers. Among them, the first DNA marker, the RFLPs, introduced a revolutionary change in plant genetics and breeding in the mid-1980s, mainly because of their infinite number and thus potential to cover maximum chromosomal regions, phenotypic neutrality, absence of epistasis, and codominant nature. An array of other hybridization-based markers PCR-based markers, and markers based on both facilitated construction of genetic linkage maps, mapping of genes controlling simply inherited traits and even gene clusters (QTLs) controlling polygenic traits in a large number of model and crop plants. During this period a number of new mapping populations beyond F_2 were utilized and a number of computer programs were developed for map construction, mapping of genes, and for mapping of polygenic clusters or QTLs. Molecular markers were also used in studies of evolution and phylogenetic relationship, genetic diversity, DNA-fingerprinting and map-based cloning. Markers tightly linked to the genes were used in crop improvement employing the so-called marker-assisted selection. These strategies of molecular genetic mapping and molecular breeding made a spectacular impact during the last one and a half decades of the twentieth century. But still they remained 'indirect' approaches for elucidation and utilization of plant genomes since much of the chromosomes remained unknown and the complete chemical depiction of them was yet to be unraveled.

Physical mapping of genomes was the obvious consequence that facilitated development of the 'genomic resources' including BAC and YAC libraries to develop physical maps in some plant genomes. Subsequently, integrated genetic-physical maps were also developed in many plants. This led to the concept of structural genomics. Later on, emphasis was laid on EST and transcriptome analysis to decipher the function of the active gene sequences leading to another concept defined as functional genomics. The advent of techniques of bacteriophage gene and DNA sequencing in the 1970s was extended to facilitate sequencing of these genomic resources in the last decade of the twentieth century.

As expected, sequencing of chromosomal regions would have led to too much data to store, characterize, and utilize with the-then available computer software could handle. But development of information technology made the life of biologists easier by leading to a swift and sweet marriage of biology and informatics and a new subject was born—bioinformatics.

Thus, evolution of the concepts, strategies and tools of sequencing and bioinformatics reinforced the subject of genomics—structural and functional. Today, genome sequencing has traveled much beyond biology and involves biophysics, biochemistry and bioinformatics!

Thanks to the efforts of both public and private agencies, genome sequencing strategies are evolving very fast, leading to cheaper, quicker and automated techniques right from clone-by-clone and whole-genome shotgun approaches to a succession of second generation sequencing methods. Development of software of different generations facilitated this genome sequencing. At the same time, newer concepts and strategies were emerging to handle sequencing of the complex genomes, particularly the polyploids.

It became a reality to chemically—and so directly—define plant genomes, popularly called whole-genome sequencing or simply genome sequencing.

The history of plant genome sequencing will always cite the sequencing of the genome of the model plant *Arabidopsis thaliana* in 2000 that was followed by sequencing the genome of the crop and model plant rice in 2002. Since then, the number of sequenced genomes of higher plants has been increasing exponentially, mainly due to the development of cheaper and quicker genomic techniques and, most importantly, development of collaborative platforms such as national and international consortia involving partners from public and/or private agencies.

As I write this preface for the first volume of the new series "Compendium of Plant Genomes", a net search tells me that complete or nearly-complete whole-genome sequencing of 45 crop plants, eight crop and model plants, eight model plants, 15 crop progenitors and relatives, and three basal plants are accomplished, the majority of which are in the public domain. This means that we nowadays know many of our model and crop plants chemically, i.e., directly, and we may depict them and utilize them precisely better than ever. Genome sequencing has covered all groups of crop plants. Hence, information on the precise depiction of plant genomes and the scope of their utilization is growing rapidly every day. However, the information is scattered in research articles and review papers in journals and dedicated web pages of the consortia and databases. There is no compilation of plant genomes and the opportunity of using the information in sequence-assisted breeding or further genomic studies. This is the underlying rationale for starting this book series, with each volume dedicated to a particular plant.

Plant genome science has emerged as an important subject in academia, and the present compendium of plant genomes will be highly useful both to students and teaching faculties. Most importantly, research scientists involved in genomics research will have access to systematic deliberations on the plant genomes of their interest. Elucidation of plant genomes is not onlyof interest for the geneticists and breeders, but also for practitioners of an array of plant science disciplines, such as taxonomy, evolution, cytology,

physiology, pathology, entomology, nematology, crop production, biochemistry, and obviously bioinformatics. It must be mentioned that information regarding each plant genome is ever-growing. The contents of the volumes of this compendium are therefore focusing on the basic aspects of the genomes and their utility. They include information on the academic and/ or economic importance of the plants, description of their genomes from a molecular genetic and cytogenetic point of view, and the genomic resources developed. Detailed deliberations focus on the background history of the national and international genome initiatives, public and private partners involved, strategies and genomic resources and tools utilized, enumeration on the sequences and their assembly, repetitive sequences, gene annotation, and genome duplication. In addition, synteny with other sequences, comparison of gene families, and, most importantly, potential of the genome sequence information for gene pool characterization through genotyping by sequencing (GBS) and genetic improvement of crop plants have been described. As expected, there is a lot of variation of these topics in the volumes based on the information available on the crop, model, or reference plants.

I must confess that as the series editor it has been a daunting task for me to work on such a huge and broad knowledge base that spans so many diverse plant species. However, pioneering scientists with life-time experience and expertise on the particular crops did excellent jobs editing the respective volumes. I myself have been a small science worker on plant genomes since the mid-1980s and that provided me the opportunity to personally know several stalwarts of plant genomics from all over the globe. Most, if not all, of the volume editors are my longtime friends and colleagues. It has been highly comfortable and enriching for me to work with them on this book series. To be honest, while working on this series I have been and will remain a student first, a science worker second, and a series editor last. And I must express my gratitude to the volume editors and the chapter authors for providing me the opportunity to work with them on this compendium.

I also wish to mention here my thanks and gratitude to the Springer staff, Dr. Christina Eckey and Dr. Jutta Lindenborn in particular, for all their constant and cordial support right from the inception of the idea.

I always had to set aside additional hours to edit books besides my professional and personal commitments—hours I could and should have given to my wife, Phullara, and our kids, Sourav, and Devleena. I must mention that they not only allowed me the freedom to take away those hours from them but also offered their support in the editing job itself. I am really not sure whether my dedication of this compendium to them will suffice to do justice to their sacrifices for the interest of science and the science community.

Kalyani, India Chittaranjan Kole

Preface to the Volume

Jatropha, *Jatropha curcas* L., a drought-tolerant shrub species of the Euphorbiaceae family, has recently attracted people's attention as a promising biofuel crop. Seeds of jatropha contain non-edible oil that is suitable to produce biodiesel or jet fuel. It is expected to contribute to reduction of CO_2 emission by substituting for fossil fuel, without competing with the food crops. However, because it has been mainly used as a hedge plant for many years, its breeding history as a commercial biofuel crop is not long enough. More extensive studies and breeding efforts would be required for more profitable commercial production. Genome sequence information is an important basis for further genetic studies and breeding. Jatropha is a diploid species, and its estimated genome size is 416 Mb in $n = 11$ chromosomes. Its first draft genome sequence was published in 2011, and since then, more than 300 Mb of the jatropha genome sequence has been determined. Genomic and genetic studies, such as researches on QTL, transcriptional factors, flowering genes, and DNA markers have made progress. On the other hand, metabolomic and physiological studies have also been done, which are essential to improve quality and quantity of the oil, and also to reduce or utilize toxic substances of seeds. To improve agronomic traits, methods to generate genetically modified jatropha have been established, and useful transgenic jatropha plants have been developed based on the genetic and physiological studies. Because jatropha is a commercial crop, translation of the scientific achievements to the practical use is important. Latest results of breeding and studies on practical cultivation would be indispensable for it. The origin of jatropha was deciphered to be Mesoamerica, and only limited genotypes in the Mesoamerican population have been brought to other continents. Tracing dispersal routes of jatropha in the cultural aspect would provide useful information of the genetic resources for breeding. This volume includes achievements in these studies and provides an overview to improve our understanding of jatropha from genomic, metabolomic, genetic, and practical points of view. I thank all the authors for their excellent contributions to this volume and hope that all the chapters will help researchers and breeders to study and improve this promising biofuel plant.

Suita, Japan — Suguru Tsuchimoto

Acknowledgement

I am really grateful for the assistance given by Dr. Chittaranjan Kole, as well as the support by staff members of our lab, in publishing this book.

Contents

Part IV Breeding and Application

Part I
Genome and Molecular Analyses

Genome Analysis

1

Hideki Hirakawa and Shusei Sato

Abstract

In order to accelerate basic and applied researches that involve genetic improvement through molecular breeding, comprehensive analyses of genes and the genome of *Jatropha curcas* have been conducted using both conventional and advanced technologies. The first publicly available draft sequence of the genome of *J. curcas* was reported in 2011, and an updated genome sequence, which is 297 Mb long and covers 99% of the euchromatic regions of the genome, was released in 2012. This genome sequence information has served as a reference for transcriptome analysis and the creation of SSR and SNP markers. The latest genome sequence information with longer scaffold length is now available, and most of the scaffolds have been anchored on the genetic linkage map. The genomic sequence and linkage map provide a valuable resource for basic and applied researches on *J. curcas* as well as comparative genomic analysis.

1.1 Introduction

The tremendous advances in DNA-sequencing technologies and associated bioinformatics and computational processes have allowed us to acquire whole genomes of sequence information on plants of agronomic importance in a relatively short period of time. The publication of the highly accurate sequence of the whole genome of *Arabidopsis thaliana* in 2000 (Arabidopsis Genome Initiative 2000) demonstrated the value of sequence information in plant genetics and genomics for the first time. It also paved the way for sequencing of the whole genomes including draft sequences, which are cost-effective but less accurate; subsequently, genome sequence information was published for a number of plant species, including rice (International Rice Genome Sequencing Project 2005), poplar (Tuskan et al. 2006), grapevine (The French–Italian Public Consortium for Grapevine Genome

H. Hirakawa · S. Sato
Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan

S. Sato (✉)
Graduate School of Life Sciences, Tohoku University, 2-1-1 Katahira, Aoba-Ku, Sendai 980-8577, Japan
e-mail: shuseis@ige.tohoku.ac.jp

S. Tsuchimoto (ed.), *The Jatropha Genome*, Compendium of Plant Genomes,
DOI 10.1007/978-3-319-49653-5_1

Characterization 2007), and *Lotus japonicus* (Sato et al. 2008). The emergence of a massively parallel sequencing system, so-called next-generation sequencing (NGS) technology, has significantly accelerated this trend.

Jatropha curcas is an important oilseed crop with great potential for the production of biodiesel fuel. *J. curcas* is a diploid species having an estimated genome size of 416 Mb (Carvalho et al. 2008) arranged in $n = 11$ chromosomes (Miller and Webster 1966) with an average GC content of 39% (Carvalho et al. 2008). The size and the base composition of the genome make *J. curcas* an attractive target for functional genomics and molecular breeding.

The sequence information accumulated from the cDNA library from seeds at three stages of fruit maturation was reported in 2010 (Gomes et al. 2010). Since then, a large quantity of information on the gene and genome structures has been published. By using a conventional Sanger sequencing method, large-scale expressed sequence tag (EST) data were generated from the cDNA libraries of developing seeds (Natarajan et al. 2010) and of developing and germinating endosperm (Costa et al. 2010). In addition, transcriptome analyses using NGS have been carried out for leaf and callus (Sato et al. 2011), a mixture of roots, mature leaves, flowers, developing seeds, and embryos (Natarajan and Parani 2011), and three different stages of developing seeds (King et al. 2011). EST-derived simple sequence repeat (SSR) markers have been developed by pyrosequencing the mRNAs (Yadav et al. 2011).

Regarding the genome sequence, a bioenergy crop company, SG Biofuels Inc., announced their completion of the *J. curcas* genome sequencing to 100× coverage using the SOLiD system. However, this information has not been made available to the research community. A draft sequence of the whole genome of *J. curcas* was therefore determined by using a combination of conventional and NGS technologies (Sato et al. 2011), and this sequence was further upgraded by the addition of new data in 2012 (Hirakawa et al. 2012). Because this information was widely available through international databases (DDBJ/GenBank/EMBL) and Web databases (http://www.kazusa.or.jp/jatropha), the draft genome sequence of *J. curcas* served as a resource for acceleration of basic and applied research. Recently, a more complete genome assembly with a longer scaffold length was generated by applying paired-end (PE) and mate-pair (MP) sequence reads of a range of different insert sizes (Wu et al. 2015). Since most of the scaffolds were anchored on a genetic linkage map, the genome sequence and linkage map provide a rich resource of genetic information for breeding and genetic improvement.

In this chapter, the current status of large-scale analyses of the genes and genome of *J. curcas* will be reviewed, and their characteristics and examples of their application will be summarized.

1.2 Genome and Transcriptome Sequence Analyses

1.2.1 Genome Assembly

The first published genome sequence of *J. curcas* was reconstructed by using sequence data obtained by a capillary sequencer (Sanger protocol) and next-generation sequencers. The cultivar name applied to the sequencing was Palawan, which originated in the Philippines. The sequence data obtained using an ABI 3730xl sequencer (Applied Biosystems, USA) consisted of 1,025,000 reads of shotgun sequences and 5300 bacterial artificial chromosome (BAC) end sequences, PE reads with insert size 3 kb and cDNA sequences obtained by using a Genome Sequencer (GS) FLX System (Roche Diagnostics, USA), and PE reads obtained by using an Illumina GAII (Genome Analyzer) sequencer (Illumina Inc., USA) with four kinds of read lengths against shotgun libraries of different insert sizes: 36/36 bases, 50/31 bases, 51/51 bases, and 76/76 bases (read 1/ read 2). The draft genome sequence JAT_r3.0 was determined in 2010 (Sato et al. 2011) and was updated to JAT_r4.5 in 2012 (Hirakawa et al. 2012). Illumina GAII reads with read lengths of 36/36 bases

and 50/31 bases were applied in JAT_r3.0, and those with read lengths of 51/51 bases and 76/76 bases were added in JAT_r4.5.

The strategy used to construct the draft genome sequence JAT_r3.0 can be summarized as follows. The PCAP.rep program (Huang et al. 2006) was applied for the assembly of the Sanger reads trimmed by Figaro and Lucy (White et al. 2008) programs, and then, Sanger PCAP contigs and Sanger singlet reads were generated. The MIRA program (Chevreux et al. 1999) was applied for the assembly of the GS FLX reads whose quality was improved by Pyrobayes (Quinlan et al. 2008) and removal of artifacts by Replicate filter (Gomez-Alvarez et al. 2009), and then, 454 MIRA contigs and 454 singlet reads were generated. The contigs and singlets generated by Sanger and GS FLX reads were merged by PCAP.rep. The hybrid contigs were scaffolded by using BAC end sequences using PCAP.rep, and the resultant contigs were further scaffolded by RNA-Seq and cDNA sequences obtained using a GS FLX System and ESTs obtained from NCBI's dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) using GS reference mapper (Roche Diagnostics, USA) and the BLAT program (Kent 2002), followed by generation of the scaffolds, contigs, and singlets. To improve the sequence quality, Illumina GAII PE reads with insert sizes of 36/36 bases and 50/31 bases were assembled by the Velvet program (Zerbino and Birney 2008), and the resulting contigs were mapped onto the scaffolds, contigs, and singlets to correct the InDel errors. Finally, the draft genome sequence (JAT_r3.0) was generated. As a result, the number of contigs and singlets was 150,417, and their total length was 285,858,490 bases. The assembly results are summarized in Table 1.1.

To update the genome information, Illumina GAII PE reads with read lengths of 51/51 bases and 76/76 bases were additionally sequenced. The Illumina GA reads were subjected to quality filtering and adaptor trimming, and the resulting reads were assembled by SOAPdenovo (Li et al. 2010). The scaffolds were merged with the scaffolds of JAT_r3.0 by using PCAP.rep. The merged sequences were named JAT_r4.0. After the assembly, the scaffolds were subjected to the following alterations for analysis: 1) The N regions with sequences corresponding to JAT_r3.0 were replaced; 2) the sequences that largely duplicated (homologous with above 80% length and 100% identity) were removed; 3) the contaminated sequences with hits against bacterial genome sequences in NCBI, mitochondrial genome sequences of *A. thaliana* (accession number: NC_001284.2), chloroplast

Table 1.1 Stastitics of *J. curcas* genome assemblies

Assembly version	JAT_r3.0	JAT_r4.5	JatCur_1.0
Cultivar name	Palawan	Palawan	GZQX0401 (scaffold)
Number of sequences	150,417	39,277	6023
Total length (bases)	286,159,324	297,661,187	318,363,324
Average length (bases)	1902	7579	52,858
Max. length (bases)	29,746	277,264	5,289,327
Min. length (bases)	42	500	200
N50 length (bases)	3832	15,950	746,835
A	17,165,800	97,945,514	87,495,699
T	16,088,187	97,679,966	87,460,404
G	12,723,655	49,793,449	43,659,798
C	10,524,293	49,839,024	43,700,994
N	11,888	2,403,234	56,046,369
Total (ATGC)	56,501,935	295,257,953	262,316,895
G + C% (ATGC)	41.1	33.7	33.3

genome sequences of *J. curcas* (accession number: FJ695500), and sequences in NCBI's UniVec database (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/) were removed; and 4) the sequences <500 bp length were removed. The resulting scaffolds were named JAT_r4.5. As a result, the total length of the genome sequence was 221,111,674 bases and consisted of 107,255 scaffolds. The assembly statistics of JAT_r3.0 and JAT_4.5 are compared in Table 1.1.

1.2.2 Gene Finding

In JAT_r3.0, gene prediction and gene modeling were performed by a combination of the methods based on ab initio gene prediction and similarity searches against the plant proteome data sets. For the ab initio method, GeneMark.hmm (Lukashin and Borodovsky 1998) and GENESCAN (Burge and Karlin 1997) with training set of *A. thaliana* were used. For modeling of exon-intron structures, NetGene2 (Hebsgaard et al. 1996) and SplicePredictor (Brendel and Kleffe 1998) programs were applied. In JAT_r4.5, the genes were newly predicted by the Augustus (Stanke et al. 2004) program with the training set of *A. thaliana*. By comparing the genes in JAT_r3.0 and JAT_r4.5, 7124 genes in JAT_r4.5 were appended to the gene set of JAT_r3.0, and 30,203 genes were consequently assigned to JAT_r4.5. The number of genes with a complete structure in JAT_r4.5 was increased about threefold from that in JAT_r3.0. The statistics of the genes predicted in JAT_r3.0 and JAT_r4.5 are summarized in Table 1.2. Similarity searches were performed against the TrEMBL database (http://www.ebi.ac.uk/uniprot) to assign annotation to the predicted genes in JAT_r4.5. A total of 22,088 of the 30,203 (73.1%) genes had significant similarity against the genes in the database. With regard to the gene structure, 25,433 of 30,203 genes (84.2%) had complete structures, and 4770 genes (15.8%) had partial structures. The average length of the 30,203 genes was 1058 bases and that of the 25,433 complete genes was 1109 bases. The number of the genes with GO annotation was 25,954 (85.9%) and that with similarity against the TrEMBL database was 22,088 (73.1%). The domain searches were performed against InterPro database (Mitchell et al. 2015) with the InterProScan program (Jones et al. 2014). To classify the predicted genes into functional categories, GO slim (http://geneontology.org/page/go-slim-and-subset-guide) analysis was performed based on the results of the InterProScan. As a result, the distributions of the GO slim categories in *J. curcas* were similar to the results for *A. thaliana*. The genes were assigned to the metabolic pathways by means of BLASTP (Altschul et al. 1997) searches against the KEGG GENES database (Ogata et al. 1999). By comparing the genes on the pathways among *J. curcas*, *Ricinus communis*, and *A. thaliana*, the genes of *J. curcas* were solely mapped onto 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.

Table 1.2 Stastitics of *J. curcas* CDSs

Assembly version	TCs	JAT_r3.0	JAT_r4.5	JatCur_1.0
Cultivar name	–	Palawan	Palawan	GZQX0401
Number of sequences	19,454	58,720	57,437	27,172
Total length (bases)	18,845,949	56,513,823	60,346,622	29,873,889
Average length (bases)	969	962	1051	1099
Max. length (bases)	15,641	8107	16,878	16,764
Min. length (bases)	50	20	22	153
G + C%	41.3	41.1	43.3	43.0

1.2.3 Transcript Sequence Assembly

The transcripts of leaf (534,137 reads; DRX000446) and callus (456,913 reads; DRX000447) in *J. curcas* cv. Palawan were sequenced by using the GS FLX System at the Kazusa DNA Research Institute (KDRI). The SRA database of NCBI (http://www.ncbi.nlm.nih.gov/sra) includes the transcript reads from a mixture of five major tissue libraries (383,937 reads; SRX035761), one seed library (195,692 reads; SRX011411), and one leaf library (2210 reads; SRX020243). In addition, 46,842 EST sequences were registered in dbEST. To construct the tentative consensus sequences (TCs) of *J. curcas*, the transcript reads and EST sequences described above were assembled by GS De Novo Assembler v2.6 software (Roche Diagnostics, USA). As a result, 19,454 TCs were generated, and the average length and GC content were 969 bases and 41.3%, respectively (Table 1.2). A total of 19,435 of 19,454 TCs (99.9%) were assigned to the scaffolds of JAT_r4.5, which means that the coverage of gene space in JAT_r4.5 was increased from that in JAT_r3.0 (95%).

1.2.4 Sequencing of the Organelle Genome

The chloroplast genome sequence has been determined (accession number: NC_012224.1 (Asif et al. 2010)). The total genome size was 163.9 kb, the GC content was 35.4%, and the numbers of genes, proteins, rRNAs, and tRNAs were 130, 84, 8, and 37, respectively (Asif et al. 2010). The total genome size and genome arrangement were similar to those of other plant species. According to the phylogenetic tree of 81 proteins in chloroplasts among 64 taxa, *J. curcas* is close to *Manihot esculanta*, which belongs to the same family as Euphorbiaceae. In the chloroplasts of *J. curcas*, the gene functions of *infA* (translation initiation factor 1) and *rps16* (small subunit ribosomal protein) were lost, and inverted repeats were found in the genic region in the *rpl2* (proteins of large ribosomal subunit) gene. The mitochondrial genome sequence of *J. curcas* has not been determined yet.

1.2.5 Genome Sequencing Projects

The draft genome sequence of *J. curcas* cv. Palawan was determined in 2010 and 2012 at KDRI (Sato et al. 2011; Hirakawa et al. 2012). The genome sequence of JAT_r3.0 had a total length of 297.7 Mb and was comprised of 39,277 scaffolds (BioProject accession number: PRJDA52543). The GC content of the genome sequence was 33.8% (Sato et al. 2011). The genome sequence of JAT_r3.0 has been updated to JAT_r4.5 by adding Illumina reads (Hirakawa et al. 2012). In 2015, the genome sequence of *J. curcas* cv. GZQX0401 was determined at the Chinese Academy of Sciences (BioProject accession number: PRJNA63485). The sequence platforms used were Illumina GAII and HiSeq with seven insert sizes: 200 b, 500 b, 800 b, 2 kb, 5 kb, 10 kb, and 20 kb. The current sequence version is called JatCur_1.0. The genome sequence had a total length of 318.4 Mb and was comprised of 6023 scaffolds. The GC content of the genome sequence was 33.3%, and 27,172 genes were predicted from the genome sequence. In the study, 1208 markers were also developed, and 81.7% of them were mapped onto the 11 pseudomolecules (Wu et al. 2015). The statistics of the assemblies of the studies of genome sequencing of *J. curcas* are compared in Table 1.1. The total length of JAT_r4.5 (39,277 scaffolds, 297,661,187 bases) was close to that of JatCur_1.0 (6023 scaffolds, 318,363,324 bases), while the N50 length of JatCur_1.0 (746,835 bases) was much longer than that of JAT_r4.5 (15,950 bases). The GC contents were not largely different (JAT_r4.5: 33.7%; JatCur_1.0: 33.3%). The scaffolds of JAT_r4.5 and JatCur_1.0 were compared by the LAST program (Kiełbasa et al. 2011) with a score of 281 (corresponding to E value = $1E{-}100$); 36,853 of 39,277 scaffolds of JAT_r4.5 were homologous to scaffolds of JatCur_1.0, while 2424 of 39,277 scaffolds of JAT_r4.5 were non-homologous to scaffolds

of JatCur_1.0, whose total length was 2,010,462 bases (0.68%).

To assess the completeness of the gene space, the core eukaryotic genes (CEGs), which are highly conserved among six organisms in eukaryotes (*Homo sapiens*, *Drosophila melanogaster*, *A. thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*), were employed for the scaffolds of JAT_r3.0, JAT_r4.5, and JatCur_1.0. The 248 CEGs were applied to the analysis by using CEGMA (Parra et al. 2007), and the numbers of the genes having complete and partial structures were 110 (44.4%) and 206 (83.1%) in JAT_r3.0, 218 (87.9%) and 242 (97.6%) in JAT_r4.5, 228 (91.9%), and 244 (98.4%) in JatCur_1.0. Based on these results, the completeness of the core genes was similar between JAT_r4.5 and JatCur_1.0.

In addition, to assess the conservation of the CDSs between Palawan and GZQX0401, the CDSs of JAT_r4.5 and JatCur_1.0 were, respectively, mapped onto each other's scaffolds by using the BLAT program. A total of 52,839 of 57,437 (92.0%) CDSs of JAT_r4.5 were mapped onto the scaffolds of JatCur_1.0, whose total, average, and N50 lengths were 59,366,347, 1124, and 1521 bases. In addition, 4598 of 57,437 (8.0%) CDSs of JAT_r4.5 were not mapped, and the total, average, and N50 lengths of these unmapped CDSs were 980,275, 213, and 201 bases. From these results, most of the CDSs of JAT_r4.5 that were frequently found to be non-homologous to JatCur_1.0 were short, and they might have been specific to cultivar Palawan (JAT_r4.5), or the scaffolds that encoded the corresponding short CDSs have not been sequenced yet in cultivar GZQX0401 (JatCur_1.0). On the other hand, 26,339 of 27,172 (96.9%) CDSs of JatCur_1.0 were mapped onto the scaffolds of JAT_r4.5, whose total, average, and N50 lengths were 29,694,129, 1127, and 1479 bases. In addition, 833 of 27,172 (3.1%) CDSs of JatCur_1.0 were not mapped, and their total, average, and N50 lengths were 179,760, 216, and 195 bases. As a result, the CDSs that were frequently found to be non-homologous to JAT_r4.5 were also short, and they might be specific in cultivar GZQX0401 (JatCur_1.0), or the scaffolds encoded the corresponding short CDSs were not sequenced yet in cultivar Palawan (JAT_r4.5).

1.2.6 miRNA Analysis

According to the similarity searches using BLASTN for 2502 miRNAs, 24 new potential miRNAs were identified from 46,862 ESTs and 1569 GSS sequences (Vishwakarma and Jadeja 2013). The transcription factors for the regulation of cell growth and development, signaling, and metabolism in oil synthesis were found in the 78 potential genes categorized into three miRNA families (Vishwakarma and Jadeja 2013). Transcription sequences (RNA-Seq) were obtained from the three libraries from immature, intermediate, and mature seeds, and 180 conserved miRNAs, 41 precursor miRNAs (pre-miRNAs), and 16 novel pre-miRNAs were identified (Galli et al. 2014). The 426 and 356 sequences were obtained from the two small RNA libraries of *J. curcas* from leaves and seeds ranging from 18 to 26 bases. The small RNA sequences were searched against the genome sequences of *A. thaliana*, rice, grape, poplar, *Euphorbia genistoides*, and *J. curcas*, and 52 miRNAs were identified based on the secondary structure prediction. The target genes of the miRNAs were examined by expression patterns. Among them, 10 miRNAs highly expressed in fruits and seeds were thought to be potentially related to the development or synthesis of fatty acids in seed. One of the identified miRNAs, JcumiR004, was considered to be related to the development and formation of fruit, and this miRNA includes the four target genes for modulating significant oil composition (Wang et al. 2012).

1.2.7 EST Analysis

Currently, (March 2015), 46,865 ESTs are registered in the dbEST. There have been several studies using ESTs to search for SSR markers. In one such study, 43,349 ESTs were used to find

SSR markers, and these ESTs were assembled by the CAP3 program (Huang and Madan 1999). SSRs were extracted by the MIcroSAtellite (MISA) program (Thiel et al. 2003) by setting the repeat length and the number of repeats at 10, 6, 5, 5, 5, and 5 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides. As a result, 6108 SSRs were found in the 5175 assembled sequences (Laosatit et al. 2013). The 42,477 ESTs were used for finding SSR markers, and they were assembled to 12,358 contigs and 5730 singlets by the CAP3 program. By searching the microsatellite, 3557 motifs were found from 7.91% of the unigenes. According to the GO analysis, 931 unigenes were related to fatty acid or lipid metabolism pathways, and those over-represented were classified into the GO terms, "Fatty acid metabolic process" and "Fatty acid biosynthetic process" (Grover et al. 2014).

A total of 383,918 reads were obtained using GS FLX Titanium technology (Roche Diagnostics, USA) from the five tissue types: roots, mature leaves, flowers, developing seeds, and embryos of *J. curcas*, and the obtained reads were assembled by GS De Novo Assembler v2.5.2. As a result, 17,457 contigs and 54,002 singlets were generated. The 25,333 ESTs obtained by a capillary sequencer (Sanger protocol) and the assembled contigs from the previous study (Sato et al. 2011) were assembled together, resulting in the generation of 14,327 contigs. Consequently, the sequences of 28,794 unigenes were obtained. Among them, 2320 unigenes were included in the pathways related to oil biosynthesis (Natarajan and Parani 2011).

1.3 Repetitive Sequences of the *J. curcas* Genome

1.3.1 Simple Sequence Repeats (SSRs)

SSRs are tandemly arranged repeats of short DNA motifs (1–6 bp in length) which tend to exhibit length polymorphism due to the variation in the number of repeats. Because of their abundance, codominant nature, and high reproductive nature, SSRs are used as valuable genetic markers applicable for various aspects of molecular genetic studies, such as assessment of genetic diversity and marker-assisted breeding.

In the *J. curcas* genome sequence, more than 40,000 di-, tri-, and tetra-nucleotide SSRs equal to or greater than 15 bp were identified, and thus, the frequency of occurrence of these SSRs can be estimated to be one SSR in every 7.0 kb. The di-, tri-, and tetra-nucleotide SSRs accounted for 46, 34, and 19% of the identified SSRs, respectively. The SSR patterns that appeared frequently were (AT)n, (AAT)n, and (AAAT)n, each representing 71% of di-, 60% of tri-, and 58% of tetra-nucleotide repeat units, respectively. The trinucleotide SSRs, particularly (AAG)n and (AGC)n, were preferentially found in exons. (AT)n, (AG)n, and (AAT)n were enriched in the 5′ and 3′ untranslated regions (UTRs), and (AC)n frequently occurred in introns (Sato et al. 2011; Hirakawa et al. 2012). A similar distribution of SSR proportion has been reported in analysis of the cassava genome (Vásquez and López 2014). Studies are currently underway to create SSR markers and analyze the germplasm diversity, creation of linkage maps, and so on by applying the genome sequence and EST sequences available in the public databases.

1.3.2 Transposable Elements

A search of the *J. curcas* genome sequences using the repeat sequence finding program RECON (Bao and Eddy 2002) unraveled the occurrence of a variety of repeat elements, including the class I and class II transposable element (TE) subfamilies, as well as some elements that were difficult to classify into known subfamilies. The composition of these repeat sequences was analyzed with the RepeatMasker program (http://repeatmasker.org). In total, the identified repetitive sequences accounted for 31.5% of the *J. curcas* genome sequences (Hirakawa et al. 2012). The most abundant repeat category was class I TE (25.2%), in which gypsy-type (16.1%) and copia-type (7.1%) LTR retroelements constituted the major components.

The extensive analysis of gypsy-type retrotransposons in the *J. curcas* genome was reported recently (Alipour et al. 2014). By combining a molecular genetics technique with computer-based mining, this work identified four new gypsy-type retrotransposons, named Jg1-4, which were then grouped into two lineages. Along with the RECON and RepeatMasker analyses, these four retrotransposons bore hit numbers after a BLAST search against *J. curcas* genome sequences. The results of fluorescence in situ hybridization (FISH) revealed that these gypsy-type retrotransposons were accumulated at the pericentromeric region of *Jatropha* metaphase chromosome spreads.

Genome-wide analysis and cytogenetic mapping of copia-type retrotransposons have also been reported (Alipour et al. 2013). The PCR fragments amplified using the degenerated primers for the reverse transcriptase domain of copia-type retroelements were classified into families, which were then grouped into three lineages corresponding to TAR, Angela, and Ale in copia-type families of other plant species. The insertion-site preferences of each family were surveyed by using the *J. curcas* genome sequence, and five of them, which existed in the gene-rich regions, were found to bear potential as appealing candidates for the development of DNA marker systems. The results of FISH analyses on mitotic chromosomes confirmed that the retrotransposons of these families were dispersed throughout all chromosomes with clustering dominantly in the distal part of chromosome arms. These findings indicated that copia-type retrotransposons can be considered a powerful system of molecular markers for elucidating the evolutional and genetic relationships among its various accessions.

With respect to class II transposable elements, significant numbers of miniature inverted-repeat transposable elements (MITEs) were identified in the *J. curcas* genome sequence. MITEs are prevalent in eukaryotic species, including plants, and are believed to be deletion derivatives of DNA transposons. Like autonomous DNA transposons, MITEs usually have terminal inverted repeats (TIRs), flanked by short direct repeats. MITEs are often located in gene-rich euchromatic regions and are associated with genes. The extensive de novo identification of MITEs from 41 plant species using the computer programs MITE Digger, MITE-Hunter, and/or RSPB (Repetitive Sequence with Precise Boundaries) revealed 18,975 elements that were classified into 17 MITE families in the *J. curcas* genome (Chen et al. 2014). Together, these MITE elements covered a total of 4.8 Mb regions of the *J. curcas* genome. The detailed information on these MITE elements is available from the P-MITE database (http://pmite.hzau.edu.cn).

1.4 Protein-Coding Genes

1.4.1 Gene Family Analysis

Based on the obtained *J. curcas* genome sequence information, studies of the gene families with important biological functions have been carried out.

WRKY genes are transcription factors related to development and stress responses, and they have one or two WRKYGQK sequences followed by a zinc finger motif to bind to the W box of target genes. By an intensive search using the WRKY domains of *Arabidopsis* WRKY proteins as query sequences against the *J. curcas* genome (JAT_r3.0) and their own transcriptome sequences, Xiong et al. (2013) identified a total of 58 WRKY genes in the *J. curcas* genome. Comparative analyses of *J. curcas* WRKY genes with *Arabidopsis*, rice, and castor bean WRKY genes revealed that evolutionarily recent WRKY paralogs in the *J. curcas* genome probably arose from early gene duplication events. Among the 58 *J. curcas* WRKY genes, 47 genes were responsible to at least one abiotic stress (Xiong et al. 2013).

In regard to biotic stress, the identification and characterization of disease-resistance genes, including nucleotide-binding site-Leucine-reach repeat (NBS-LRR), is an important approach to accelerate the process of genetic improvement of

disease resistance. In the *J. curcas* genome, a total of 91 NBS-LRR genes have been identified by mapping Pfam domains as well as mapping publicly available NBS-LRR mRNA sequences (Sood et al. 2014). By comparing the NBS-LRR genes identified in the *J. curcas* genome sequence with the NBS-LRR genes identified in the castor bean genome, several genes unique to one or common to both genomes were identified (Sood et al. 2014). Along with the 122 defense response-related transcription factors identified in the *J. curcas* genome sequence (Sood et al. 2014), the genome-wide information on NBS-LRR-resistance genes should provide novel insights about the molecular basis of disease-resistance phenotypes.

1.4.2 Tandem Gene Duplication

Tandem gene duplication is one of the major mechanisms of duplication in eukaryotes. In the first and second draft genome sequence reports, the numbers of tandemly arrayed genes, such as NBS-LRR disease-resistance proteins, were described as a characteristic feature of the *J. curcas* genome (Sato et al. 2011; Hirakawa et al. 2012). In the latest genome with longer scaffold length, 3839 tandem gene duplications among 1442 loci in the *J. curcas* genome were confirmed (Wu et al. 2015). The longest tandem gene array consists of 15 cytochrome P450 (CYP) genes. Proteins with kinase active site domains, disease-resistance gene products, UDP-glucuronosyl/-glucosyltransferase domain-containing proteins, and short-chain dehydrogenase/reductases were among the proteins, and domains found most frequently in tandemly repeated genes in the *J. curcas* genome (Wu et al. 2015).

1.5 Comparative Analysis of the *J. curcas* Genome

The obtained *J. curcas* genome sequence information paved the way for proceeding with comparative genome analysis at several levels.

1.5.1 Comparative Analysis Within *J. curcas* Accessions

Understanding the genetic diversity within the *Jatropha* germplasm is critically important in establishing breeding strategies and designing breeding programs. Genetic variations in natural and cultured populations around the world have been studied by using random amplified polymorphic DNA (RAPD), inter-simple sequence repeat (ISSR), and amplified fragment length polymorphism (AFLP) before the genome sequence became available (Sun et al. 2008; Basha and Sujatha 2009; Sudheer-Pamidimarri et al. 2009; Shen et al. 2010). The focus of the study of genetic diversity is shifting to sequence information-based codominant markers, such as SSR and single nucleotide polymorphism (SNP) markers. When the first draft sequence of the *J. curcas* genome was reported, a total of 88 SSR markers were generated from the obtained genome sequence information. These markers were evaluated using the 12 lines of *J. curcas* corrected from meso-America, Africa, and Asia. The number of alleles per locus ranged from one to four with a mean value of 1.31. The markers showed no polymorphisms; those detecting a single allele were most frequent. Polymorphic information content (PIC) values ranged from 0 to 0.45 with a mean value of 0.06 (Sato et al. 2008). The large number of tested markers detecting no polymorphisms and the low mean value of the PIC indicated that the genetic diversity in *Jatropha* lines is generally narrow. Among the tested accessions, the three lines derived from meso-America regions (Guatemala1, Guatemala2, and Mexico2b) are genetically distinct from the other lines derived from Asia and Africa, whereas no significant difference was observed between the Asian and African lines (Sato et al. 2008). A similar trend of slightly higher genetic diversity in varieties from meso-America regions was observed in the analysis of additional SSR markers designed from genome sequence information (Raposo et al. 2014). By using *J. curcas* accessions from Guatemala as a genotyping population, the average number of alleles per locus was

increased to 6.9 (among 18 polymorphic loci detected by the analysis of twenty SSR markers). PIC values ranged from 0.114 to 0.886 with a mean value of 0.627. SNP markers have also been used for the genetic diversity analysis. Sandoval et al. (2014) applied both SNP markers and SSR markers to their genetic diversity analysis of 70 accessions from around the world. They found that the variance components based on SNP and SSR marker data were similar, and the genetic diversity in the accessions from meso-America regions was higher than that in accessions from Africa, Asia, and South America, which was consistent with the previous studies (Sandoval et al. 2014). Extensive analysis of SNPs in the *J. curcas* accessions was also carried out based on whole-transcriptome re-sequencing of pooled samples of 12 *J. curcas* accessions (Silva-Junior et al. 2011). By mapping of 66.5 million reads on the unigene set of the draft genome sequence, 60.8 million reads were aligned on the 28,110 unigenes with an average coverage of 152×. Only 18,225 SNPs were detected (with no minor allele frequency or flanking sequence filtering), and this very low number indicated the low level of sequence polymorphism in the transcribed regions of *J. curcas* accessions. Another attempt to create large-scale SNP markers was carried out by using high-throughput sequencing on a GS FLX Systems of the complexity-reduced genomic DNA of 61 accessions. From the 871 assembled contigs contacting 26,940 sequences, a total of 2482 informative SNPs were identified with an average frequency of one SNP per 100 bp. Genotyping of selective SNPs among the 148 global collections of *J. curcas* accessions revealed that a narrow level of genetic diversity existed among the indigenous genotypes as compared to the exotic genotypes of *J. curcas* (Gupta et al. 2012). These SNP markers could be very useful in large-scale marker application in molecular breeding.

1.5.2 Comparative Analysis in Euphorbiaceae

Euphorbiaceae, to which *Jatropha* belongs, is a complex family consisting of 229 genera and 6511 accepted species in "The Plant List" database (http://www.theplantlist.org/browse/A/Euphorbiaceae/). *Euphorbiaceae* includes several economically important species in addition to *Jatropha*, such as an important oilseed crop, castor bean (*R. communis*); an essential food source and bioenergy crop, cassava (*M. esculenta*); and a resource for natural rubber, rubber tree (*Hevea brasiliensis*). The availability of several Euphorbiaceae genomes allows us to take advantage of a comparative genomic approach, which is a powerful tool for gaining insights into genome structure and evolution.

1.5.2.1 Castor Bean Genome Information

Castor bean is a tropical perennial shrub of African origin which is now cultivated in many tropical and subtropical regions around the world. It can be self- and cross-pollinated, and studies performed using castor bean collections from around the world have revealed the relatively low genetic diversity among the castor bean germplasm (Allan et al. 2008; Foster et al. 2010).

Castor bean was the first Euphorbiaceae species whose draft genome sequence information was reported. By assembling ~2.1 million high-quality sequence reads from plasmid and fosmid libraries generated by the Sanger sequence method, a draft genome sequence of castor bean ($2n = 20$) composed of 25,828 scaffolds with an N50 of 496.5 kb was obtained (Chan et al. 2010). When 3500 scaffolds larger than 2 kb were considered, the castor bean genome assembly spanned 325.5 Mb, which was consistent with the genome size estimated by flow cytometry (Arumuganathan and Earle 1991). With the aid of 52,165 EST sequences

accumulated from five cDNA libraries, a total of 31,237 potential protein-coding genes were predicted on the obtained genome sequences.

When the first draft genome sequence of *J. curcas* was obtained, the syntenic relation between castor bean and *J. curcas* was analyzed (Sato et al. 2011). A significant level of syntenic relation was detected on 53% of the scaffolds with five or more genes. The ratio of contigs with a syntenic relation between castor bean and *J. curcas* was increased to 88% when the upgraded *J. curcas* genome sequence (JAT_r4.5) was obtained due to the increases in the length of the contigs and in the number of genes with complete prediction (Hirakawa et al. 2012). The syntenic relation was further confirmed by anchoring the castor bean draft genome sequences to the *Jatropha* genetic map (Wu et al. 2015). A total of 410 scaffolds covering 54% of the total scaffold sequences of the castor bean draft genome were anchored to the *Jatropha* genetic map, and 320 well-conserved synteny blocks containing more than 10,000 *J. curcas* genes collinear to castor bean genes on the anchored scaffolds were defined. These syntenic relations could serve as a useful resource for the identification of genes by homology and for information exchange within Euphorbiaceae species.

1.5.2.2 Cassava Genome Information

Cassava is a perennial woody shrub with edible tuberous roots and is grown throughout tropical and subtropical regions of the world, especially Africa, Asia, and the Americas. Its large, starchy roots and edible leaves provide the primary staple food for over 800 million people worldwide (Ceballos et al. 2010). The high starch content (20–40%) makes cassava a desirable energy source both for human consumption and for industrial biofuel applications (Balat and Balat 2009; Schmitz and Kavallari 2009).

The draft genome sequence of cassava was obtained by a whole-genome shotgun strategy using the GS FLX System. A total of 22.4 billion bp of raw sequence data were accumulated, which was ~29 times the estimated size of the genome (770 Mb) (Awoleye et al. 1994), and these reads were assembled into 12,977 scaffolds that spanned a total of 533 Mb, on which 96% of the cassava EST sequences in GenBank were mapped (Prochnik et al. 2012). With the aid of 1.4 million additional EST reads from leaf and root libraries by the GS FLX System, a total of 30,666 genes and 3485 alternative splice forms were predicted on the obtained cassava draft genome sequences. Recently, the draft genome sequences of additional cassava genotypes, W14 (*M. esculenta* ssp. *flabellifolia*), a wild subspecies that shows low storage root yield and low root starch, and KU50, a variety commonly cultivated in Southeast Asia that has sixfold to eightfold higher storage root yield potential, were analyzed by using an integrated assembly strategy combining the sequence reads obtained from Illumina and GS FLX System (Wang et al. 2014). Comparative genomics analysis revealed a considerable amount of genome diversity (SNPs and InDels) in W14 and KU50 when compared with the reference cassava genome (a partially inbred line, AM560). The results of a comparative analysis of the predicted genes from the genomes of W14, KU50, and AM560 revealed that 1584 were unique to W14 or lost in KU50 and AM560, whereas another 1678 genes were specific to the cultivated varieties. The availability of high-quality draft genome sequences for these three genotypes will contribute to the genetic improvement of cassava through a better understanding of its biology.

Detailed information on the reference draft genome sequences of castor bean and cassava can be accessed through the plant comparative genomics portal Phytozome (http://phytozome.jgi.doe.gov). This impressive body of genomic resources will establish the basis of an information exchange for Euphorbiaceae species.

1.5.3 Comparative Mapping

A genetic linkage map is the essential framework for genome-wide identification of associations between DNA markers and traits (Doerge 2002). In addition, a genetic linkage map assists in anchoring the assembled genome sequences to create pseudomolecules and provides a solid

basis for comparative mapping. Comparative mapping, in turn, should help to establish the syntenic relationships against the model plant species, such as *A. thaliana*, which would be beneficial for applying the information generated in the model plant species.

A first-generation linkage map was constructed using a mapping population containing two backcross populations consisting of 93 progenies. The F1 populations were produced by an interspecies cross between two accessions of *J. curcas* as female parents and a single *J. integerrima* individual as male parent, and the two BC1F1 populations obtained by backcrossing *J. curcas* parents to each F1 progeny were used as mapping populations (Wang et al. 2011). The mapping populations were genotyped with two types of codominant DNA marker SSRs on the genome sequences and SNPs on the EST sequences. A total of 506 markers were mapped onto 11 linkage groups, out of which 216 were SSR markers and 290 were SNP markers. The total length of the obtained *Jatropha* genetic map was 1440.9 cM, with an average marker spacing of 2.8 cM (range 1.2–4.3 cM in each linkage group). By a similarity search of the 222 ESTs containing SSR and SNP markers mapped on the linkage map against reference genomes, it was revealed that 96.8, 91.0, and 77.5% of *J. curcas* ESTs were homologous to their counterparts in castor bean, poplar, and *Arabidopsis*, respectively. A comparative map between *Jatropha* and *Arabidopsis* using 192 orthologous markers elucidated 38 syntenic blocks and revealed that small linkage blocks were well conserved, but often shuffled (Wang et al. 2011). The linkage map and the data of comparative mapping provide a solid basis for quantitative trait locus (QTL) mapping of agronomic traits, marker-assisted breeding, and cloning genes responsible for phenotypic variations.

Another high-density linkage map was constructed using the BC1 population of an interspecific cross (*J. curcas* × *J. integerrima*) with 1208 SNP, InDel, and SSR markers (Wu et al. 2015). The total genetic distance covered by this linkage map was 1655.8 cM, with an average marker density of 2.1 cM for unique loci. This linkage map was applied for anchoring the scaffolds of the latest genome assembly. A total of 480 scaffolds covering 261.8 Mb (∼81.7% of the total scaffold sequences) were anchored to the map to produce eleven pseudochromosomes. This high-density genetic linkage map and anchored genomic sequences provide a valuable resource for fundamental and applied researches as well as for evolutionary and comparative genomics analysis (Wu et al. 2015).

The first intraspecific *J. curcas* map was constructed from four F2 mapping populations created from parental lines displaying differences in a range of traits (King et al. 2013). Genotyping assays were performed using both SNP and SSR markers, and the linkage maps for each mapping population were built individually. Then, the genotype information was merged to create an integrated linkage map containing 502 codominant markers, distributed over 11 linkage groups, with a mean marker density of 1.8 cM per unique locus. By using one of the four mapping populations created from G33 (toxic seed) × G43 (non-toxic seed), linkage analysis for loci controlling phorbol ester biosynthesis was carried out. QTL analysis revealed that a single locus at 41 cM on linkage group 8 was associated with phorbol ester biosynthesis. By anchoring the draft genome sequence contigs of *J. curcas* and castor bean, additional markers were created on the target region to fine map this mutation within 2.3 cM. This first intraspecific *J. curcas* map therefore provides a framework for the dissection of agronomic traits in *J. curcas* and the development of improved varieties by marker-assisted breeding.

1.6 Databases

1.6.1 *J. curcas* Genome Database

The genomic information obtained by the genome assembly has been published, and the data can be downloaded from the database (http://www.kazusa.or.jp/jatropha/, Sato et al. 2011). The current version of the genomic information in the Web database is JAT_r4.5 (Hirakawa et al.

2012). In addition, the genomic information of JAT_r3.0 is still available in the database by clicking the banner for JAT_r3.0. Users can browse the general features of the assembly and search against the CDSs, protein sequences, and keywords in the annotation data for both JAT_r3.0 and JAT_r4.5.

1.6.2 TreeTFDB

TreeTFDB (http://treetfdb.bmep.riken.jp/index.pl, Mochida et al. 2013) is a database of transcription factors for six tree crop species: *J. curcas* (JAT_r3.0), papaya (*Carica papaya*; Phytozome v8.0), cassava (*M. esculenta*; Phytozome v8.0; http://www.phytozome.net), poplar (*Populus trichocarpa*; Phytozome v8.0), castor bean (*R. communis*; Phytozome v8.0), and grapevine (*Vitis vinifera*; Phytozome v8.0). In the current version (1.10), annotated TF models were registered for *J. curcas* (1481 TFs), papaya (1552 TFs), cassava (2638 TFs), poplar (3110 TFs), castor bean (1512 TFs), and grape (1493 TFs). In this database, TFs determined in JAT_r3.0 were registered. The genomic locations of the TFs can be browsed on GBrowse (Stein et al. 2002). In the database, about 60 kinds of TF families have been registered, and the sequences of cDNAs, proteins, 500, 1000, and 3000 bp upstream regions, and domains, and cis-motifs can be obtained from the download site. Blast searches can be conducted for CDSs of the six tree species and cDNAs of *A. thaliana*.

1.6.3 TropiTree

TropiTree (http://ics.hutton.ac.uk/tropiTree/) is a database for the assembled transcripts (unigene) for 24 tropical tree species (*Acacia senegal*, *Acrocarpus fraxinifolius*, *Adansonia digitata*, *Albizia lebbeck*, *Calliandra calothyrsus*, *Diospyros mespiliformis*, *Enterolobium cyclocarpum*, *Faidherbia albida*, *Gliricidia sepium*, *Jacaranda mimosifolia*, *J. curcas*, *Leucaena diversifolia*, *Leucaena leucocephala*, *Moringa stenopetala*, *Prunus Africana*, *Samanea saman*, *Senna siamea*, *Sesbania macrantha*, *Sesbania sesban*, *Tephrosia candida*, *Tipuana tipu*, *Warburgia ugandensis*, *Ziziphus mauritiana*). In addition, EST-SSRs detected from the ESTs and primers for these EST-SSRs can be browsed. Users also can search the genes by homology searches for the query sequences and keyword searches. For *J. curcas*, 13,252 unigenes and 1118 primer sets for EST-SSRs (di-, tri-, and tetra-nucleotide) can be browsed. The unigenes of *J. curcas* are built from the published RNA-Seq data (accession number: ERS399695).

1.6.4 KaPPA-View4

KaPPA-View4 (http://kpv.kazusa.or.jp) is a database for the representation of transcriptome and metabolome data on pathway maps (Sakurai et al. 2011). KaPPA-View4 has two systems, KaPPA-View Classic and KaPPA-View KEGG. The former is based on the traditional KaPPA-View map of *A. thaliana*, which contains information on several genome-sequenced plants, i.e., *A. thaliana*, rice, tomato, *L. japonicus*, soybean, barley, poplar, wheat, grape, and maize (10 plant species). The latter system includes other organisms, e.g., human, mouse, rat, *C. elegans*, *D. melanogaster*, *A. thaliana*, rice, poplar, castor bean, sorghum, grape, maize, *Physcomitrella patens*, *Escherichia coli*, and budding yeast (15 species). The genomic information of *J. curcas* (JAT_r3.0) has been registered on the KaPPA-View4 Jatropha site (http://kpv2.kazusa.or.jp/kpv4-jat/). A total of 8058 of 40,929 genes were mapped onto the KEGG pathway maps.

1.6.5 PGDBj (Plant Genome DataBase Japan)

PGDBj (http://pgdbj.jp) is a portal site integrating the databases related to plant omics studies (Asamizu et al. 2014). The information related to DNA markers, QTL, and plant diseases has been collected from the literature for 80 plant species by manual curation. PGDBj has links to the other

databases, such as SABRE2 DB (http://sabre.epd.brc.riken.jp/SABRE2.html, ref_db4) for the resource (clone sequences) of 15 plant species (*A. thaliana*, *Thellungiella halophila*, *Brassica rapa*, *Nicotiana tabacum*, *Solanum lycopersicum*, *L. japonicus*, *Glycine max*, *M. esculenta*, *Striga hermonthica*, *Ipomoea nil*, *Brachypodium distachyon*, *Triticum aestivum*, *Hordeum vulgare*, *Populus nigra*, *Physcomitrella patens*) and KNApSAcK (http://kanaya.naist.jp/KNApSAcK/) for the species–metabolite relationship. Currently, 50,899 metabolites and 111,199 metabolite–species pairs have been registered. Users can search information of metabolites by inputting information such as the organism name, metabolite name, molecular weight, and molecular formula. For oilseed crops, *J. curcas* and *R. communis* have been registered. In *J. curcas*, 563 SSR markers, 3 SCAR markers, and 290 SNP markers have been registered.

1.7 Conclusion

The availability of whole-genome sequences has significantly altered the approach to understanding *J. curcas*. The publicly available draft genome information has led to large-scale genomic analyses, such as transcriptome analyses and analyses for molecular marker creation. Along with the genome sequence information on castor bean and cassava, the data from comparative genome analysis will serve as a basis to transfer knowledge among the Euphorbiaceae species and should ultimately elucidate the genetic systems in Euphorbiaceae and accelerate the breeding process. The new and fine draft genome information on the additional accession will further enhance the DNA marker creation to cover the entire genome and facilitate a higher level of genomic studies, such as genome-wide association study. By putting more effort into the enrichment of accessions with accurate phenotype information, the advanced genome-based strategies will be able to contribute in a significant way to the breeding programs for improving *Jatropha*.

References

Alipour A, Tsuchimoto S, Sakai H, Ohmido N, Fukui K (2013) Structural characterization of copia-type retrotransposons leads to insights into the marker development in a biofuel crop, *Jatropha curcas* L. Biotechnol Biofuels 6:129

Alipour A, Cartagena JA, Tsuchimoto S, Sakai H, Ohmido N, Fukui K (2014) Identification and characterization of novel Gypsy-type retrotransposons in a biodiesel crop, *Jatropha curcas* L. Plant Mol Biol Rep 32:923–930

Allan G, Williams A, Rabinowicz PD, Chan AP, Ravel J, Keim P (2008) Worldwide genotyping of castor bean germplasm (*Ricinus communis* L.) using AFLPs and SSRs. Genet Resour Crop Evol 55:365–378

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

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–218

Asamizu E, Ichihara H, Nakaya A, Nakamura Y, Hirakawa H, Ishii T, Tamura T, Fukami-Kobayashi K, Nakajima Y, Tabata S (2014) Plant Genome DataBase Japan (PGDBj): a portal website for the integration of plant genome-related databases. Plant Cell Physiol 55:e8

Asif MH, Mantri SS, Sharma A, Srivastava A, Trivedi I, Gupta P, Mohanty CS, Sawant SV, Tuli R (2010) Complete sequence and organisation of the *Jatropha curcas* (Euphorbiaceae) chloroplast genome. Tree Genet Genomes 6:941–952

Awoleye F, Duren M, Dolezel J, Novak FJ (1994) Nuclear DNA content and in vitro induced somatic polyploidization cassava (*Manihot esculenta* Crantz) breeding. Euphytica 76:195–202

Balat M, Balat H (2009) Recent trends in global production and utilization of bio-ethanol fuel. Appl Energ 86:2273–2282

Basha SD, Sujatha M (2009) Genetic analysis of *Jatropha* species andinterspecific hybrids of *Jatropha curcas* using nuclear and organelle specific markers. Euphytica 168:197–214

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269–1276

Brendel V, Kleffe J (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. Nucleic Acids Res 26:4748–4757

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic. DNA. J Mol Biol 268:78–94

Carvalho CR, Clarindoa WR, Praça MM, Araújoa FS, Carels N (2008) Genome size, base composition and karyotype of *Jatropha curcas* L., an important biofuel plant. Plant Sci 174:613–617

Ceballos H, Okogbenin E, Pérez JC, López-Valle LAB, Debouck D (2010) Cassava. In: Bradshaw JE (ed) Root and tuber crops, handbook of plant breeding, vol 7. Springer, New York, pp 53–96

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

Chen J, Hu Q, Zhang Y, Lu C, Kuang H (2014) P-MITE: a database for plant miniature inverted-repeat transposable elements. Nucleic Acids Res 42(Database issue):D1176–D1181

Chevreux B, Wetter T, Suhai S (1999) Genome sequence assembly using trace signals and additional sequence information, In: Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB), 99, pp 45–56

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

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3:43–52

Foster JT, Allan GJ, Chan AP, Rabinowicz PD, Ravel J, Jackson PJ, Keim P (2010) Single nucleotide polymorphisms for assessing genetic diversity in castor bean (*Ricinus communis*). BMC Plant Biol 10:13

Galli V, Guzman F, de Oliveira LF, Loss-Morais G, Körbes AP, Silva SD, Margis-Pinheiro MM, Margis R (2014) Identifying microRNAs and transcript targets in Jatropha seeds. PLoS One 9:e83727

Gomes KA, Almeida TC, Gesteira AS, Lôbo IP, Guimarães ACR, de Miranda AB, Van Sluys MA, da Cruz RS, Cascardo JCM, Carels N (2010) ESTs from seeds to assist the selective breeding of *Jatropha curcas* L. for oil and active compounds. Genom Insights 3:29–56

Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. ISME J 3:7–1314

Grover A, Kumari M, Singh S, Rathode SS, Gupta SM, Pandey P, Gilotra S, Kumar D, Arif M, Ahmed Z (2014) Analysis of *Jatropha curcas* transcriptome for oil enhancement and genic markers. Physiol Mol Biol Plants 20:139–142

Gupta P, Idris A, Mantri S, Asif MH, Yadav HK, Roy JK, Tuli R, Mohanty CS, Sawant SV (2012) Discovery and use of single nucleotide polymorphic (SNP) markers in *Jatropha curcas* L. Mol Breed 30:1325–1335

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouzé P, Brunak S (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucleic Acids Res 24:3439–3452

Hirakawa H, Tsuchimoto S, Sakai H, Nakayama S, Fujishiro T, Kishida Y, Kohara M, Watanabe A, Yamada M, Aizu T, Toyoda A, Fujiyama A, Tabata S, Fukui K, Sato S (2012) Upgraded genomic information of *Jatropha curcas* L. Plant Biotechnol 29:123–130

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:77–868

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

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:40–1236

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC (2011) Adaptive seeds tame genomic sequence comparison. Genome Res 21:93–487

King AJ, Li Y, Graham IA (2011) Profiling the developing *Jatropha curcas* L. seed transcriptome by pyrosequencing. Bioenerg Res 4:211–221

King AJ, Montes LR, Clarke JG, Affleck J, Li Y, Witsenboer H, van der Vossen E, van der Linde P, Tripathi Y, Tavares E, Shukla P, Rajasekaran T, van Loo EN, Graham IA (2013) Linkage mapping in the oilseed crop *Jatropha curcas* L. reveals a locus controlling the biosynthesis of phorbol esters which cause seed toxicity. Plant Biotechnol J 11:986–996

Laosatit K, Tanya P, Saensuk C, Srinives P (2013) Development and characterization of EST-SSR markers from *Jatropha curcas* EST database and their transferability across jatropha-related species/genus. Biologia 68:41–47

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

Lukashin A, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26:1107–1115

Miller KI, Webster GL (1966) Chromosome numbers in the *Euphorbiaceae*. Brittonia 18:372–379

Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJ, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I,