Gene 500 (2012) 172–180

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Characterization of the chloroplast genome sequence of oil palm (*Elaeis guineensis* Jacq.)

P. Uthaipaisanwong [a,b], J. Chanprasert [a], J.R. Shearman [a], D. Sangsrakru [a], T. Yoocha [a], N. Jomchai [a], C. Jantasuriyarat [c], S. Tragoonrung [a], S. Tangphatsornruang [a,*]

[a] National Center for Genetic Engineering and Biotechnology, 113 Phaholyothin Road, Klong 1, Klong Luang, Pathumthani 12120, Thailand
[b] Interdisciplinary Graduate Program in Genetic Engineering, The Graduate School, Kasetsart University, Bangkhen Campus, Bangkok 10900, Thailand
[c] Department of Genetics, Faculty of Science, Kasetsart University, Bangkhen Campus, Bangkok 10900, Thailand

ARTICLE INFO

Article history:
Accepted 23 March 2012
Available online 2 April 2012

Keywords:
Oil palm
Chloroplast genome
454 pyrosequencing
RNA editing

ABSTRACT

Oil palm (*Elaeis guineensis* Jacq.) is an economically important crop, which is grown for oil production. To better understand the molecular basis of oil palm chloroplasts, we characterized the complete chloroplast (cp) genome sequence obtained from 454 pyrosequencing. The oil palm cp genome is 156,973 bp in length consisting of a large single-copy region of 85,192 bp flanked on each side by inverted repeats of 27,071 bp with a small single-copy region of 17,639 bp joining the repeats. The genome contains 112 unique genes: 79 protein-coding genes, 4 ribosomal RNA genes and 29 tRNA genes. By aligning the cp genome sequence with oil palm cDNA sequences, we observed 18 non-silent and 10 silent RNA editing events among 19 cp protein-coding genes. Creation of an initiation codon by RNA editing in *rpl2* has been reported in several monocots and was also found in the oil palm cp genome. Fifty common chloroplast protein-coding genes from 33 plant taxa were used to construct ML and MP phylogenetic trees. Their topologies are similar and strongly support for the position of *E. guineensis* as the sister of closely related species *Phoenix dactylifera* in Arecaceae (palm families) of monocot subtrees.

1. Introduction

Chloroplasts are photosynthetic intracellular organelles, which act as the energy source for plants and algae (Vothknecht and Westhoff, 2001). This organelle contains its own genome, which forms a circular double-stranded DNA molecule. A typical cp genome has a conserved quadripartite structure including a large single-copy region (LSC), a small single-copy region (SSC) and a pair of inverted repeats (IRs) that separate the LSC from the SSC. In plants, IRs range in length from 20 to 30 kb (Palmer and Delwiche, 1998). Over 200 cp genomes have been deposited in NCBI GenBank database since the first cp genome of liverwort (*Marchantia polymorpha*) (Ohyama et al., 1986) was reported in 1986 expanding our knowledge on genomic structure, gene regulation and evolution. Most cp genomes contain 110–130 genes (Ravi et al., 2007). The majority of chloroplast coding genes function related to photosynthesis, transcription and translation machineries (Ahlert et al., 2003). Chloroplast gene expression can be controlled at several stages such as transcription (Gruissem and Tonkyn, 1993; Pfannschmidt et al., 1999), post-transcription (del Campo, 2009; Rochaix, 1996, 2001) and translation (Danon, 1997; Marin-Navarro et al., 2007). RNA Editing, which is a post-transcriptional maturation step, takes place by editing single specific nucleotides at the transcript level (Bock, 2000; Schmitz-Linneweber and Barkan, 2007). The most common RNA editing event in chloroplasts is a C-to-U conversion, which occurs predominantly in the first or second position of codons within coding regions of an mRNA. In many cases, editing events cause non-silent mutations and changes in amino acid sequences, generation of initiation or termination codons, and elimination of stop codons. This post-transcriptional modification process creates translatable mature mRNAs that are essential for protein function (Gray, 2009; Maier et al., 1996).

Among flowering plants, there are over 100 monocots families, yet there are currently only 41 monocot cp genomes from nine families that are available in Genbank (Givnish et al., 2010). Twenty seven of these are members of the grass family including rice (Hiratsuka et al., 1989), maize (Maier et al., 1995), wheat (Ogihara et al., 2002), sugarcane (Asano et al., 2004), barley and sorghum (Saski et al., 2007), because these plants are a primary source of nutrition for humans (Wu and Ge,

Abbreviations: A, adenosine; BLAST, The Basic Local Alignment Search Tool; C, cytosine; CsCl, Cesium chloride; DAP, days after pollination; DNA, Deoxyribonucleic acid; DOGMA, The Dual Organellar GenoMe Annotator; G, guanine; U, uracil; G3P, glycerol-3-phosphate; IR, inverted repeats; LSC, large single-copy region; ML, maximum likelihood; MP, maximum parsimony; PA, phosphatidic acid; PAUP, phylogenetic analysis using parsimony; PS, photosystem; SSC, single-copy region; T, thymine; TBR, tree bisection and reconnection; bp, base pair; cDNA, complementary DNA; cp, chloroplast; dNTP, deoxynucleotide triphosphate; g, gram; ha, hectare; kb, kilobase; lnL, log-likelihood; mM, milliMole; min, minute; ng, nanogram; rDNA, ribosomal DNA; s, second; tRNA, transfer ribonucleic acid; x, multiple; ycf, hypothetical chloroplast open reading frame; °C, degree Celsius; µg, microgram.

* Corresponding author. Tel.: +66 2 564 6700; fax: +66 2 564 6584.
E-mail address: sithichoke.tan@biotec.or.th (S. Tangphatsornruang).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2012.03.061

2012; Wu et al., 2009; Zhang et al., 2011). Complete cp genomes in monocots range from 59 kb (*Rhizanthella gardneri*) (Delannoy et al., 2011) to 165 kb (*Lemna minor*) (Mardanov et al., 2008).

Oil palm (*Elaeis guineensis* Jacq.) is a member of the Arecaceae or Palmae family which contains ~200 genera and about 2600 species. So far, the cp genome of *Phoenix dactylifera* (Yang et al., 2010) is the only species that has been reported and characterized in Arecaceae. Oil palm is an important member of the Arecaceae family because it produces one of the highest oil yields compared with other oil producing plants (3.5 ton/ha/year), and accounts for 36% of world production (Bourgis et al., 2011). Palm oil is produced in countries with a tropical climate and high rainfall, is widely used for cooking and is economically important as a potential alternative energy source for the future.

Chloroplasts play an important role in lipid biosynthesis in plants. Fatty acids are synthesized in the plastid stroma and metabolized on the plastid membrane or exported to the cytosol for phosphatidic acid (PA) synthesis, which is important for phospholipid formation within the endoplasmic reticulum (Joyard et al., 1998). The phospholipids are also used for biosynthesis of triacylglycerol, involved in the formation of storage lipid in plant cells (Dubots et al., 2012). Therefore, a potential application of genetic engineering in oil plants by modification of genes related to fatty acid biosynthesis in chloroplasts can increase the production of fatty acids and lipid in plant cells. The availability of the complete chloroplast genome sequence should facilitate the chloroplast transformation technique (Bock, 2001). Several advantages of chloroplast genetic engineering over nuclear transformation are high copy number of a transgene due to the high ploidy levels of the cp genome and abundant transgene transcripts resulting in high levels of accumulated foreign proteins. Polycistronic RNA resulting in multiple transgenes can be inserted and expressed in a single transformation event. Furthermore, there is no report of gene silencing that would interfere with gene expression and cp genes are maternally inherited which means that transgenes cannot spread by pollen dispersal (Daniell et al., 2005). The complete chloroplast genome of *E. guineensis* was sequenced for the purposes of determining the chloroplast genome structure and identifying genes and their functional annotation to provide the basic information required for chloroplast engineering. We used 454 pyrosequencing, which is a fast and cost-effective approach that has been successfully used in applications such as genome sequencing, transcriptomic, and metagenomic studies including organellar genome sequencing (Cronn et al., 2008; Tangphatsornruang et al., 2010, 2011; Yang et al., 2010). Further analysis on the chloroplast post-transcriptional control by RNA editing, and phylogenetic relationships among angiosperms were also included in this study.

2. Materials and methods

2.1. DNA sequencing, genome assembly, and validation

We extracted DNA from 1 g of *E. guineensis* Jacq. fresh leaves using the DNeasy Plant Mini Kit (Qiagen, CA, USA). The DNA (10 μg), was sheared by nebulization, subjected to 454 library preparation and shotgun sequencing using the Genome Sequencer FLX and the GS FLX Titanium platforms (Margulies et al., 2005) at our in-house facility (National Center for Genetic Engineering and Biotechnology, Thailand). The obtained sequence reads were assembled using the Newbler *de novo* assembly version 2.5 (Roche 454). The chloroplast genome contigs were separated from total Genomic DNA by aligning to the cp genome sequences of *P. dactyliferal L.* and *Oryza sativa* Japonica cultiva-group using Sequencher software version 4.10.1. Contig orientation, the junctions between single-copy and IRs regions, lengths of all homopolymers ($n > 7$ bp) and the 5′ region of *cemA* in the chloroplast genome were validated by Sanger sequencing. Primers for sequencing were designed using Primer3 software (Rozen and Skaletsky, 2000). PCR was carried out in total volume of 20 μl containing 10 ng of DNA, 1× buffer, 0.2 mM dNTPs, 1 U Phusion DNA polymerase and 0.5 μM each of forward and reverse primers. Amplification was performed in a GeneAmp PCR 9700 System thermocycle (Applied Biosystems, CA, CUA) programmed as follows: 94 °C for 2 min, followed by 35 cycles of 94 °C for 30 s, 55 °C for 1 min, 72 °C for 3 min and a final extension step at 72 °C for 5 min. Sanger sequencing reactions were performed using BigDye Terminator v3.1 Cycle sequencing kit and were resolved on ABI 3700 DNA Analyzer.

2.2. Genome analysis, repeat structure, and codon usage

The cp genome sequence of oil palm was annotated using The Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al., 2004) with plastid/bacterial genetic code and default conditions. The DOGMA software package performs BLAST searches against completed chloroplast genomes, which include tobacco, rice, and date palm. Comparative analysis between the oil palm and the date palm cp genomes was used to support manual annotation. tRNAScan-SE (Lowe and Eddy, 1997) and tRNADB-CE (Abe et al., 2011) were used for tRNA annotation. The REPuter program (Kurtz et al., 2001) was used to visualize direct repeat and inverted repeats within the oil palm cp genome under the criteria cutoff $n \geq 30$ bp, 90% sequence identities, and non-overlapping regions. The frequency of codon usage was calculated from exon sequences of all protein-coding genes in the oil palm cp genome.

2.3. RNA editing analysis

Total RNAs of 100-DAP fruits were isolated based on a protocol by the cesium chloride method (Morcillo et al., 2006) from five-year-old *E. guineensis*. Double strand cDNA was synthesized with random octamer primers using cDNA synthesis system kit (Roche). The cDNA library was constructed and sequenced according to the GS FLX titanium protocol (Roche). For RNA Editing analysis, cDNA reads were aligned against protein-coding genes of the reference genome using GS Reference Mapper version 2.5 with default conditions. Variations of RNA from the DNA template were detected by Perl script under the condition of ≥25% variable frequency. The results were formatted according to proposed nomenclature (Lenz et al., 2010; Rudinger et al., 2009).

2.4. Phylogenetic tree

We used 50 common chloroplast protein-coding genes consisting of *atpA, atpB, atpE, atpH, atpI, ccsA, cemA, clpP, petA, petB, petD, petG, petN, psaA, psaB, psaC, psaI, psbB, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbM, psbN, psbT, psbZ, rbcL, rpl14, rpl16, rpl2, rpl20, rpl23, rpl32, rpl33, rpl36, rpoA, rpoB, rpoC1, rpoC2, rps11, rps12, rps15, rps2, rps3, rps4, rps7*, and *rps8* to infer the phylogenetic relationships among 33 angiosperm lineages previously published in the GenBank database (Supplementary Table 1) including species closely related to oil palm plus two outgroup gymnosperms (*Pinus* and *Ginkgo*). We used two character-based methods: the maximum likelihood (ML) and maximum parsimony (MP) methods. Nucleotide sequences were aligned by ClustalX version 2.0.12 (Larkin et al., 2007; Thompson et al., 1994) and gaps were removed prior to phylogenetics analyses. Phylogenetic analyses were performed using RAxML v7.0 (Stamatakis, 2006), ML analysis was performed with the GTR + I + G nucleotide substitution model under the best fit parameters determined by Modeltest ver. 3.7 (Posada and Crandall, 2003). The MP analysis was performed using PAUP (Phylogenetic analysis using parsimony) version 4.0b (Swofford, 2002). Starting trees were constructed with 1000 replicates of random taxon addition and a heuristic search using tree bisection and reconnection (TBR) branch swapping (Multrees option off). In these analyses, the bootstrap analysis was performed with 1000 replicates with TBR branch swapping (Efron et al., 1996; Felsenstein, 1985).