Evidence-based gene models for structural and functional annotations of the oil palm genome

Chan Kuang Lim[a,b,#], Tatiana V. Tatarinova[c,#], Rozana Rosli[a,d], Nadzirah Amiruddin[a], Norazah Azizi[a], Mohd Amin Ab Halim[a], Nik Shazana Nik Mohd Sanusi[a], Jayanthi Nagappan[a], Petr Ponomarenko[c], Martin Triska[c,d], Victor Solovyev[e], Mohd Firdaus-Raih[b], Ravigadevi Sambanthamurthi[a], Denis Murphy[d], Leslie Low Eng Ti[a*]

[a] Advanced Biotechnology and Breeding Centre, Malaysian Palm Oil Board, P.O. Box 10620, 50720 Kuala Lumpur, Malaysia

[b] Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

[c] Spatial Sciences Institute, University of Southern California, Los Angeles, CA, 90089, USA

[d] Genomics and Computational Biology research group, University of South Wales, Pontypridd, CF371DL, United Kingdom

[e] Softberry Inc., 116 Radio Circle, Suite 400 Mount Kisco, NY 10549, USA.

* To whom correspondence should be addressed. Tel. +60387694963. Fax. +60389261995. E-mail: chankl@mpob.gov.my

[#] These authors contributed equally to the paper and should be considered joint first authors.

Abstract

The advent of rapid and inexpensive DNA sequencing has led to an explosion of data waiting to be transformed into knowledge about genome organization and function. Gene prediction is customarily the starting point for genome analysis. This paper presents a bioinformatics study of the oil palm genome, including comparative genomics analysis, database and tools development, and mining of biological data for genes of interest. We have annotated 26,059 oil palm genes integrated from two independent gene-prediction pipelines, Fgenesh++ and Seqping. This integrated annotation constitutes a significant improvement in comparison to the preliminary annotation published in 2013. We conducted a comprehensive analysis of intronless, resistance and fatty acid biosynthesis genes, and demonstrated that the high quality of the current genome annotation. 3,658 intronless genes were identified in the oil palm genome, an important resource for evolutionary study. Further analysis of the oil palm genes revealed 210 candidate resistance genes involved in pathogen defense. Fatty acids have diverse applications ranging from food to industrial feedstocks, and we identified 42 key genes involved in fatty acid biosynthesis in oil palm. These results provide an important resource for studies of plant genomes and a theoretical foundation for marker-assisted breeding of oil palm and related crops.

Keywords

oil palm; gene prediction; Seqping; fatty acids; intronless; resistance genes

Abbreviations

CDS, coding sequence; GO, Gene Ontology; IG, intronless gene; R, resistance; CC, coiled-coil; NBS, nucleotide binding site; LRR, leucine-rich repeat; CNL, CC-NBS-LRR; FA, fatty acid; FAD2, oleoyl-phosphatidylcholine desaturase; FAD3, linoleoyl-phosphatidylcholine desaturase; ACP, acyl carrier protein; FATB, acyl-ACP thioesterase; TNL, Toll/interleukin-1 NBS-LRR; Avr, avirulence; STK, serine/threonine protein kinase; ACCase, acetyl-CoA carboxylase; FABF, β-ketoacyl-ACP synthase II; FAB2, Stearoyl-ACP desaturase; FATA, oleoyl-ACP thioesterase

1.   Introduction

Oil palm is a member of the genus *Elaeis* of the family Arecaceae. The genus has two species - *E. guineensis* (African oil palm) and *E. oleifera* (American oil palm). Three fruit forms of *E. guineensis* mainly vary in the thickness of their seed (or kernel) shell - *dura* (thick-shell), *tenera* (thin-shell) and *pisifera* (no shell). The African oil palm is by far the most productive oil crop [1] in the world, with estimated production in year 2015/2016 of 61.68 million tons, of which the Malaysian share was 19.50 million tons [2]. Palm oil constitutes ~34.35% of the world production of edible oils and fats. Globally, palm oil is mainly produced from *E. guineensis,* in the *tenera* fruit form. *E. oleifera* is not used commercially due to its low yield. However, it is more disease-resistant and planted in areas where *guineensis* is practically impossible, e.g., Central-Southern America. Even then, it is mainly planted as a backcross to *guineensis* (interspecific hybrid) to increase the yield. Nevertheless, it has economically valuable traits which plant breeders wish to introgress into *guineensis*, such as a more liquid oil with higher carotenoid and vitamin E contents, disease resistance and slow height increment [1].

The importance of oil palm has resulted in considerable interest to sequence its transcriptomes and genome. Initial work used expressed sequence tags (ESTs) [3], a technique very useful for tagging expressed genes but only providing partial coverage of the coding regions and genome in general. Next, GeneThresher[TM] technology was applied to selectively sequence hypomethylated regions of the genome [4]. With the genome sequenced [5], then using genetic mapping and homozygosity mapping via sequencing, the *SHELL* gene was identified [6]. This discovery facilitated development of an efficient genetic test to distinguish between the *dura, pisifera* and *tenera* fruit forms. Subsequently, the *VIRESCENS* gene, which regulates the fruit exocarp colour [7], and the *MANTLED* gene, which causes tissue culture abnormality [8], were also discovered. Accurate genome annotation was critical for the identification of these genes, and will be crucial for increasing the palm productivity.

First gene prediction pipelines appeared in the 1990s. In 1997, a group of mathematicians from Stanford developed Genscan [9] software, followed by a steady stream of specially designed tools to navigate complexity of various genomes. Combining multiple predictors led to the development of automated pipelines integrating various experimental evidences [10]. A major limitation shared by many approaches is their relatively poor performance in organisms with atypical distribution of nucleotides [11–14]. The GC_3 content of the genes plays an important role, as GC_3-rich genes in grasses can be better predicted by transcriptome-based than by homology-based methods [15]. Accurate gene prediction and the discovery of regulatory elements in promoter sequences are two of the most important challenges in computational biology, for the prediction quality affects all aspects of genomics analysis.

To overcome the lack of precision in many predictive models, we developed an integrated gene-finding framework, and applied it to identify high quality oil palm gene models. The framework is a combination of the Seqping [16] pipeline developed at MPOB, and the Fgenesh++ [17] pipeline by Softberry. Individual components of the framework were trained on known genes of plants closely related to the oil palm, such as the date palm, to identify the most suitable parameters for gene prediction. The best gene model for each locus was selected to

3

establish a representative "high confidence" gene set. This paper describes identification and characterization of the "high confidence" set of 26,059 oil palm genes that have transcriptome and RefSeq support, and present bioinformatics analysis of the genes, including comparative genomics analysis, and database and tools development.

2.   Materials and Methods

2.1   Datasets

We used the *E. guineensis* P5-build scaffold assembly from Singh et al.[5], which contained 40,360 genomic scaffolds (N50 length: 1,045,414 nt; longest length: 22,100,610 nt; and shortest length: 1,992 nt). The *E. guineensis* mRNA dataset is a compilation of published transcriptomic sequences from Bourgis et al. [18], Tranbarger et al. [19], Shearman et al. [20,21], and Singh et al. [6], as well as 24 tissue-specific RNA sequencing assemblies from MPOB, which were submitted to GenBank in BioProject PRJNA201497 and PRJNA345530, and oil palm ESTs downloaded from the nucleotide database in GenBank. The dataset was used to train the Hidden Markov Model (HMM) for gene prediction, and as transcriptome evidence support.

2.2   Fgenesh++ Gene Prediction

Fgenesh++ (Find genes using Hidden Markov Models) [17,22] is an automatic gene prediction pipeline, based on Fgenesh, a HMM-based *ab initio* gene prediction program [23]. We used the oil palm genomic scaffolds to make the initial gene prediction set using Fgenesh gene finder and generic parameters for monocot plants. From this set, we selected a subset of predicted genes that encode highly homologous proteins (using BLAST with E-value < 1.0E-10) to known plant proteins from the NCBI non-redundant (NR) database. Based on this subset, we computed gene-finding parameters, optimal for the oil palm genome, and executed the Fgenesh++ pipeline to annotate the genes in the genomic scaffolds. The Fgenesh++ pipeline took into account all available supporting data, such as known transcripts and homologous protein sequences. To predict genes, we compiled a subset of NR plant transcripts mapped to oil palm genomic sequences providing a set of potential splice sites to the Fgenesh++ pipeline. Palm transcripts were aligned to the palm genomic scaffolds and from this mapping a set of potential splice sites was generated and used in Fgenesh and Fgenesh++ gene prediction programs. Plant proteins were mapped to the oil palm genomic contigs and the high scoring matches were used in generating protein-supported gene prediction by the Fgenesh++ program, so that highly homologous proteins were used in gene identification.

Amino acid sequences from predicted oil palm genes were compared to the protein sequences from plant NR database using the 'bl2seq' routine, and the similarity was considered to be significant if it had a blast percent identity ≥ 50, blast score ≥ 100, coverage of predicted protein ≥ 80% and coverage of homologous protein ≥ 80%. BLAST analysis of the predicted sequences was also carried out against the *E. guineensis* mRNA dataset, using an identify cutoff of >90%. Predictions that have both NR plant RefSeq and *E. guineensis* mRNA support were selected for further analysis.

2.3   Seqping Gene Prediction

4

Seqping [16], a customized gene prediction pipeline based on MAKER2 [24], was developed by MPOB. Full-length open reading frames (ORFs) were identified from the *E. guineensis* mRNA dataset described above, using the EMBOSS *getorf* program. The range of ORF lengths was selected to be 500 - 5000 nt, to minimize potential prediction errors. Using BLASTX [25] search, ORFs with E-values <1E-10 were deemed significantly similar to the RefSeq plant protein sequences. ORFs with BLASTX support were clustered using BLASTClust and CD-HIT-EST [26], and subsequently filtered using the TIGR plant repeat database [27], GIRI Repbase [28] and Gypsy Database [29] to remove ORFs similar to retroelements. The resulting set of ORFs was used as the training set to develop HMMs for GlimmerHMM [30,31], AUGUSTUS [32] and SNAP [33], which were subsequently used for gene predictions. Seqping used MAKER2 [24] to combine the predictions from the three models. All programs were used in the Seqping pipeline with the default parameter settings. The predicted sequences were compared to the RefSeq [34] protein sequences and *E. guineensis* mRNA dataset via BLAST analysis. Predictions that have NR plant RefSeq and *E. guineensis* mRNA support (E-value cutoff: 1E-10) were selected for further analysis.

2.4   Integration of Fgenesh++ and Seqping Gene Predictions

To increase the accuracy of annotation, predictions independently made by the Seqping and Fgenesh++ pipelines were combined into the unified prediction set. All predicted amino acid sequences were compared to protein sequences in the NR database using BLAST (E-value cutoff: 1E-10). ORF predictions with <300 nucleotides were excluded. Predicted genes from both pipelines in the same strand were considered overlapping if the shared length was above the threshold fraction of the shorter gene length. A co-located group of genes on the same strand was considered to constitute a locus if every gene in the group overlaps at least one other member of the same group (single linkage approach). Different overlap thresholds, from 60% to 95% in 5% increments, were tested to determine the best threshold value, simultaneously maximizing the annotation accuracy and minimizing the number of single-isoform loci. Protein domains were predicted using PFAM-A [35,36] (release 27.0) and PfamScan ver. 1.5. The palm coding sequences (CDSs) were also compared to the NR plant sequences from RefSeq (release 67), using *phmmer* function from the HMMER-3.0 package [37,38]. To find the representative isoform and determine its function for each locus, we selected the lowest E-value isoform in each locus and the function of its RefSeq match. We excluded hits with E-values >1E-10, as well as proteins that contained words "predicted", "putative", "hypothetical", "unnamed", or "uncharacterized" in its description, keeping only high-quality loci and the corresponding isoforms. Loci without RefSeq match were discarded. The CDS in each locus with the best match to the RefSeq database of all plant species was selected as the *best representative* CDS for the locus. Gene Ontology (GO) annotations were assigned to the palm genes, using the best NCBI BLASTP hit to *Oryza sativa* sequences from the MSU rice database [39] at a cutoff E-value of 1E-10.

2.5   Intronless Genes

Intronless genes (IG) were identified as mono-exonic genes containing full-length ORFs, as specified by the gene prediction pipeline. The same approach was applied to five other genomes: *A. thaliana* (TAIR10) [40], *O. sativa*

5

(MSU 6.0) [39], *S. bicolor* (Phytozome 6.0), *Z. mays* (Phytozome) and *Volvox carteri* (Phytozome 8.0) [41]. Lists of non-redundant IG from all six genomes were obtained, and the oil palm IG were compared with those in other genomes using BLASTP (E-value cutoff: 1E-5). The protein sequences of the IG were also mapped to all NCBI genes in the archaea, bacteria and eukaryote kingdoms using BLASTP with the same cutoff.

2.6   Resistance (R) Genes

All curated plant resistance (R) genes were downloaded from the database PRGdb 2.0 [42]. A local similarity search of known plant resistance genes and oil palm gene models was done using the BLASTP program with E-value ≤ 1E-5. TMHMM2.0 [43] was used to find predicted transmembrane helices in the known R genes, as well as in the oil palm candidate R genes, and these results were used for classification of the R genes. Domain structures of the known and oil palm candidate R genes were identified using InterProScan. All domains found in the known R genes were used to classify candidate R genes in oil palm according to the PRGdb classification. To be considered an R gene, a candidate was required to contain all the domains found in known R genes of the respective class. Our selection was validated based on the published "resistance" gene motifs [44–48] and each class was further validated via multiple sequence alignment and phylogenetic tree, using the ClustalW [49] and MEGA6 [50] programs, respectively. The same procedure was used to identify R genes in *A. thaliana* [40], *O. sativa* [39], *S. bicolor*, *Z. mays* and *V. carteri* genomes. Distribution of coiled-coil (CC) – nucleotide binding site (NBS) – leucine-rich repeat (LRR) or CNL class R genes across 16 chromosomes of the EG5 genome build [5] was conducted to identify physical clustering. A cluster of R genes is defined as two CNL genes that are located less than 200 kb apart, with no more than eight non NBS-LRR genes between the two CNL genes [51,52].

2.7   Fatty Acid (FA) Biosynthesis Genes

A. thaliana, O. sativa, Z. mays, Glycine max and *Ricinus communis* amino acid sequences corresponding to 11 FA biosynthesis genes were obtained from KEGG [53]. The corresponding amino acid sequences for another three genes (oleoyl-phosphatidylcholine desaturase [FAD2], linoleoyl-phosphatidylcholine desaturase [FAD3], acyl-acyl carrier protein (ACP) thioesterase [FATB]) were obtained from published journals [54–57]. These sequences were compared to the oil palm gene models using Exonerate [58] with the "protein2dna" alignment model parameter. The oil palm gene models were annotated using BLASTX against the RefSeq database. Conserved domains of these genes were identified using InterProScan [59] against the HMMPfam database [35,60]. Corresponding protein sequences of candidate oil palm FA biosynthesis genes and FA biosynthesis genes from other organisms were aligned using the ClustalW program. The catalytic residues and conserved motifs of the amino acid sequences of the corresponding candidate FA biosynthesis genes were identified from literature [61–72]. Sequences of the identified FA genes, having more than one copy, were extracted with additional flanking regions of 10Mb upstream and downstream to check for genome duplication events using the PROmer [73] software with default parameters.

2.8   Expression analysis

To estimate the expression of FA biosynthesis genes, two Illumina HiSeq 2000 libraries each of mesocarp and kernel samples in NCBI BioProject PRJNA245226 [5], were read-mapped to the P5-build of the oil palm genome using the Tuxedo suite [74,75]. Fragments Per Kilobase of exon per Million mapped fragments (FPKM) was calculated with the expression of each gene the mean of measures from two biological replicates. Expressions of genes in root, leaf, leaf apex and flower from BioProject PRJNA201497 were determined by mapping two Roche 454 sequencing transcriptome data for each tissue using the same method.

2.9   Comparative Genomics

To identify orthologs of FA biosynthesis and R genes in oil palm sequences, OrthoMCL2.0 [76] was used with default parameters to construct orthologous groups across three sets of gene models: *E. guineensis*, *A. thaliana* and *Z. mays*. The corresponding protein sequences of these genes were confirmed with BLASTP [25] searches against the NCBI NR database with default parameters. Protein members of the cluster sequences were aligned by two methods, Muscle [77] and MAFFT [78] version 7. Protein domain sequences were identified using Pfam [36], InterPro [79], ScanProsite [80] and NCBI CDD [81]. To get an overview of the relationships between selected orthologous genes, phylogenetic trees were constructed using MEGA6 [50] and MAFFT [82]. All programs were used with their default settings.

3.   Results and Discussion

3.1   Gene models

Prediction and annotation of protein-coding genes are routinely done using a number of software tools and pipelines, such as Fgenesh++ [17], MAKER-P [83], Gramene [84], and Ensembl [85]. In plants, at least three model organisms (*A. thaliana, Medicago truncatula* and *O. sativa*) have been annotated using a combination of evidence-based gene models and *ab initio* predictions [86–88]. The first version of oil palm genome [5] was published in 2013 with assembled sequences representing ~83% of the 1.8 Gb-long genome. Using this assembly, we predicted gene models by combining output from the two pipelines: Fgenesh++ and Seqping [16].

Previous studies of five *ab initio* pipelines (Fgenesh++, GeneMark.hmm, GENSCAN, GlimmerR and Grail) to evaluate gene prediction precision showed that Fgenesh++ produced the most accurate maize genome annotations [22]. Fgenesh++ is a common tool for eukaryotic genome annotation, due to its superior ability to predict gene structure [89–92]. In the oil palm genome, Fgenesh++ predicted 117,832 whole and partial-length gene models at least 500 nt long. A total 27,915 Fgenesh++ gene models had significant similarities to the *E. guineensis* mRNA dataset and RefSeq proteins (Table 1).

To improve coverage and accuracy of gene prediction, and to minimize prediction bias, Seqping, which is based on the MAKER2 pipeline [24], was also used. Seqping is an automated pipeline that generates species-

specific HMMs for predicting genes in a newly sequenced organism. It was previously validated using the *A. thaliana* and *O. sativa* genomes [16], and the pipeline was able to successfully predict at least 95% of the Benchmarking Universal Single-Copy Orthologs's (BUSCO) [93] plantae dataset (BUSCO provides quantitative measures for the assessment gene prediction sets based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs [93]). Seqping demonstrated the highest accuracy compared to three HMM-based programs (MAKER2, GlimmerHMM, and AUGUSTUS) with the default or available HMMs [16]. The pipeline was used to train the oil palm specific HMMs. This was done by identifying 7,747 putative full-length CDS from the transcriptome data and using this set to train the models for GlimmerHMM [30,31], AUGUSTUS [32], and SNAP [33]. These oil palm-specific HMMs were used in MAKER2 to predict oil palm genes. The initial prediction identified 45,913 gene models that were repeat-filtered and scrutinized using the oil palm transcriptome dataset and RefSeq protein support (Table 1).

Gene models from Fgenesh++ and Seqping were then combined to identify the number of independent genomic loci. Since the ratio of the number of single-isoform to multi-isoform loci increased more rapidly above the 85% overlap between two loci (Table 2 and Fig. 1), we set this value as the overlap threshold. Gene models that had an overlap ≥85%, were grouped into a locus. Imposing of this threshold allowed us to minimize false positive in merging loci, while maximizing true positives in joining gene models into one locus. At this threshold, we obtained 31,369 combined loci (Table 2) in 2,915 scaffolds, of which 26,087 genomic loci contained gene models with PFAM domains and RefSeq annotations. Of them, 9,899 contained one ORF, 12,163 two, and 3,981 three or more. For every locus, the CDS with the best match to plant proteins from the RefSeq database was selected as its *best representative* CDS.

These scaffolds were also screened using MegaBLAST search against RefSeq Representative Genome Database (E-value cutoff: 0) to get a set of high-confidence hits. Hits to *E. guineensis* were excluded. If the best BLAST hits were represented by bacterial taxa genomes or plastid plant genomes, such scaffolds were marked as potential contaminants. 43 scaffolds were identified and checked manually. The scaffolds were also compared to the *oleifera* genome, RNA-seq data and the latest Pisifera genome builds that MPOB uses internally. Scaffolds with no support in all three levels were removed from the final dataset. As a result, 24 scaffolds containing 28 loci were removed. The remaining representative CDS for 26,059 genomic loci (the "high quality" dataset) are supported by oil palm transcriptome data.

Analysis of the gene structures showed that 14% were intronless and 16% contained only two exons. 395 genes had more than 20 exons. Further analyses on these genes using BLASTX (E-value cutoff: 1E-5) to determine their identity and exon numbers, showed that 366 of these genes had alignment coverage of above 90% with the RefSeq [34] genes. The number increased to 384 genes when it was reduced to at least 80% coverage. The two oil palm genes with the largest exon number (57 exons) were p5.00_sc00063_p0008 and p5.00_sc00076_p0105. Detailed examination of the gene p5.00_sc00063_p0008 showed that it was similar to serine/threonine-protein kinase TOR from *Musa acuminate*, *Vitis vinifera*, *Citrus sinensis* and *Theobroma cacao* that also have 57 exons.

Interestingly, the oil palm translation activator GCN1 (p5.00_sc00076_p0105), was similar to the genes in *Phoenix dactylifera*, *V. vinifera*, *O. sativa* and *M. acuminate* that have 60 exons. The distributions of exons per gene and CDS lengths are shown in Figure 2A and Figure 2B respectively. Evolutionary conservation of gene structure was previously described for several species and gene families [94,95,96]. For example, it was estimated that between mouse and human, 86% of the orthologous gene pairs have the same number of coding exons [96].

BUSCO analysis of representative gene models showed that 90.44% of the 429 eukaryotic BUSCO profiles were available. By comparing to 956 *plantae* BUSCO profiles, 94.04% of the BUSCO genes were found in the predicted gene models, including 90.48% of them as complete BUSCO genes, thus quantifying the completeness of the oil palm genome annotation. By comparison, the first set of gene prediction by Singh el at. [5] in 2013 had matches to 81.17% of the *plantae* BUSCO profiles, of which only 59.00% were as complete BUSCO genes, indicating a big improvement in the latest gene models. Also, for each gene in the current and 2013 annotation, we found the best match to the plant RefSeq database using NCBI BLASTP program. The new annotation has higher identity to the RefSeq proteins as compared to the old one.

3.2   Nucleotide composition of the oil palm genes

One of important characteristics of a genome, frequency of guanine and cytosine contents in the third codon position - GC_3, is defined as $\dfrac{C_3+G_3}{\left(L/3\right)}$, where L is the length of the coding region, C_3 the number of cytosines, and G_3 the number of guanines in the third position of codons in the coding region [15]. Two types of GC_3 distribution have been described - unimodal and bimodal [15,97,98]. Genes with high and low GC_3 peaks have distinct functional properties [98]: GC_3-rich genes provide more targets for methylation, exhibit more variable expression, more frequently possess upstream TATA boxes and are predominant in stress responsive genes. Different gene prediction programs have variable bias to different classes of genes, with GC_3 rich genes reported to be especially hard to predict accurately [99]. Distribution of GC_3 is bimodal in grasses and warm-blooded vertebrates, and unimodal in other species sequenced up to date [100].

Distribution GC_3 in the oil palm is unimodal with a long tail towards high values of GC_3. Figure 3A shows the distribution of GC_3 in the "high quality" dataset. We ranked all genes by their GC_3 content and designated the top 10% (2,605 ORFs) as GC_3-rich ($GC_3 \geq 0.75$), and the bottom 10% as GC_3-poor ($GC_3 \leq 0.37$). Two of the remarkable features that distinguish GC_3-rich and -poor genes are the gradients of GC_3 and CG_3-skew, defined as $CG_3^{skew} = \dfrac{C_3-G_3}{C_3+G_3}$. Increase in the CG_3^{skew} from 5' to 3' has been linked to transcriptional efficiency and methylation status [15,98,101] of GC_3-rich genes. Figures 3 (C and D) show positional gradients of nucleotide composition: GC_3 content of GC_3-rich genes increases from the 5' to 3' end of the gene, but decreases in GC_3-poor genes. Despite the relatively small number of the GC_3-rich genes in the oil palm genome, there are characteristic

9

patterns of positional gradients (Figures 3, C and D) near the predicted start of translation, as found across other well-annotated genomes [15].

The dinucleotide CG relative abundance (a.k.a. "genomic signature") is defined as $\rho_{CG} = \frac{f_{CG}}{f_C f_G}$, where f_x is the frequency of a (di)nucleotide x [102]. Similar to grasses, and other previously analyzed plant and animal species [15,98], the oil palm genome signature differs for GC_3-rich and GC_3-poor genes (Fig. 3B). The GC_3-rich genes are enriched and the GC_3-poor genes are depleted in the number of CpG sites, that represent potential targets for methylation. Gene ontology analysis shows that many of the GC_3-rich genes are stress-related, while many of the GC_3-poor genes have housekeeping functions (see GO annotation in Table S1). The depletion of CpGs in GC_3-poor genes is consistent with their broad, constitutive expression [15].

We calculated the distribution of nucleotides in the oil palm coding regions. We considered the following models of ORF: *Multinomial* (all nucleotides independent, and their positions in the codon not important), *Multinomial* position-specific and *First order three periodic Markov Chain* (nucleotides depend on those preceding them in the sequence, and their position in the codon is taken into account). Tables S2-S5 show probabilities of nucleotides A, C, G and T in GC_3-rich and -poor gene classes. Notice that both methods predict that GC_3-poor genes have greater imbalance between C and G, than GC_3-rich genes (0.05 vs. -0.1). This is consistent with the prior observation [98] that GC_3-rich genes have more targets for methylation than GC_3-poor genes, and some cytosine can be lost from this deamination.

GC_3-rich and -poor genes differ in their predicted lengths and open reading frames (Table S6): the GC_3 rich genes have gene sequences and ORFs approximately seven times and two times shorter, respectively, than the GC_3-poor genes. This is consistent with the findings from other species [15,97,98]. It is important to note that GC_3-rich genes in plants tend to be intronless [15].

3.3 Intronless Genes (IG)

Intronless genes (IG) are common in single-celled eukaryotes, but represent only a small percentage of all genes in metazoans [103,104]. Across multi-cellular eukaryotes, IG are frequently tissue- or stress-specific, GC_3-rich and their promoters have a canonical TATA-box [15,98,103]. Among the 26,059 representative gene models with RefSeq and transcriptome evidence, 3,658 (14.1%) were IG. The mean GC_3 content of IG is 0.668±0.005 (Fig. 4), while the intron-containing (a.k.a. multi-exonic) genes' mean GC_3 content is 0.511±0.002, in line with estimation for other species. Intronless CDS are, on average, shorter than multi-exonic CDS: 924 ±19 nt *vs*. 1289±12 nt long. On average, there is one intronless gene per 9.5 multi-exonic genes on any scaffold containing intronless genes. There is no difference in nucleotide composition and CpG frequency between short scaffolds that contain intronless genes, multi-exonic genes, and no genes.

10

The distribution of IG in the whole genome is different for various functional groups [15,104]. For example, in the oil palm genome, 29% of the cell-signaling genes are intronless, compared to just 1% of all tropism-related genes (Table S7). The distribution of genes by GO categories is similar to that in *O. sativa*. It has been shown that in humans, mutations in IG are associated with developmental disorders and cancer [104]. Intronless and GC$_3$-rich genes are considered to be evolutionarily recent [15] and lineage-specific [103], potentially appearing as a result of retrotransposon activity [104,105]. It is reported that 8 - 17% of the genes in most animals are IG: ~10% in mice and humans [103] and 3 - 5% in teleost fish. Plants have proportionately more IG than animals: 20% in *O. sativa*, 22% in *A. thaliana* [106], 22% in *S. bicolor*, 37% in *Z. mays*, 28% in foxtail millet, 26% in switchgrass and 24% in purple false brome [107]. We have independently calculated the fraction of IG in *O. sativa*, *A. thaliana*, *S. bicolor* and *Z. mays* using the currently published gene models for each species, with results of 26%, 20%, 23% and 37%, respectively (Table S8). To establish a reference point, we calculated the fraction of IG in the green algae, *V. carteri*, and found 15.8%. High proportion of IG in grasses is not surprising, since they have a clearly bimodal distribution of GC$_3$ composition in their coding region, with the GC$_3$-peak of this distribution dominated by IG [15].

Using BLASTP, we found 543 IG (14.84% of oil palm IG) conserved across all the three domains of life: archaea, bacteria, and eukaryotes (Fig. 5). These genes are likely to be essential for survival [108]. A total of 736 oil palm IG had homologs only with eukaryotic and bacterial genes, while only 40 IG had homologs with eukaryotic and archaea genes (and not with bacteria). We speculate (due to extreme growth conditions of archaea [109,110]) that there are (1) fewer opportunities for horizontal gene transfer from archaea than from bacteria to the oil palm genome; and/or (2) possible ancestral gene loss that occurred on the archaeal branch in process of adaptation. Considering three of the most economically important eukaryotic groups, namely Metazoa (animals), Fungi and Viridiplantae (green plants), we observed 1,373 oil palm IG shared among these groups. A significant portion of the oil palm IG (1,863) was only homologous to Viridiplantae. These proteins may have evolved, or been regained, only in plants, while other organisms have lost their ancestral genes during evolution [106].

Reciprocal BLAST was carried out to verify homologies of the oil palm candidate IG to produce a set of high confidence oil palm IG. We found 2,431 (66.46%) proteins encoded by oil palm IG to have orthologs in *A. thaliana*, *O. sativa* or *Z. mays* that were also intronless, indicating that intronlessness is an ancestral state [111,112]. In conclusion, from our representative gene models, we estimate that about one-seventh of the genes in oil palm are intronless. We hope that this data will be a resource for further comparative and evolutionary analysis, and aid in the understanding of IG in plants and other eukaryotic genomes.

3.4 Resistance (R) Genes

Plants differ from animals in many aspects; one of them is the lack of an antibody-based immune system. Instead, they have developed various protein-based mechanisms to recognize invading pathogens [113–115]. The genes encoding for such proteins are called "resistance", or "R", genes. They play an important role in the plant's early detection and signaling mechanism against biotic and abiotic stresses. Using homology, we identified 210 oil palm

candidate R genes from the 26,059 representative gene models that have RefSeq and transcriptome evidence (Supplementary file 1). This represents approximately 0.80% of the high quality genes identified in the oil palm genome, a similar ratio to an earlier study on the hypomethylated regions of the *E. guineensis* genome, where 52 (0.94%) candidate resistance genes were identified among 5,505 gene models [4]. A similar trend was also observed in *A. thaliana* and *O. sativa* - 0.95% and 0.71% resistance genes, respectively. The oil palm candidate R genes were compared to those in banana (*M. acuminata*) and *O. sativa*, and 693 orthologs (253 in *M. acuminata*, 440 in *O. sativa*) were identified for 204 out of the 210 oil palm R genes.

The candidate genes were divided into six classes based on their protein domain structure [42]. Comparison of the distribution of oil palm candidate R genes with the R genes identified using the same method in other plants showed that the CNL class genes had the highest representation in monocots, with *O. sativa* having the largest percentage (51.8%). *A. thaliana*, which is dicotyledonous, has two additional classes: TNL (Toll/interleukin-1 NBS-LRR) and RPW8-NL; while colonial green algae *V. carteri* is missing most of the R gene classes in its genome. TNL, the most prevalent class in *A. thaliana*, is predominantly found in dicots [116]. The CNL and TNL classes both belong to the NBS-LRR family [117]. TNL can be differentiated from CNL based on the Toll/interleukin-1 (TIR) receptor domain at the N-terminus structure [116].

We did not identify any TNL genes in the analyzed monocot genomes, including the oil palm. This is in line with earlier work of Tarr and Alexander [118] who also did not find TNL genes in monocots. It is therefore assumed that R genes in monocots predominantly contain leucine zipper regions that facilitate formation of the conserved CC structure at the N-terminal of NBS-LRR genes, as previously indicated [119]. The CC domain is required for protein-protein interaction [45] while the LRR domain interacts with the avirulence (Avr) gene product from pathogens to activate the plant defense system [120]. Plants producing specific R genes are resistant to pathogens which produce the corresponding Avr gene products. The fraction of R genes across the plant genome suggest importance of these genes for both monocots and dicots (Fig. 6A).

CNL genes form the most abundant class in the oil palm genome. A total 141 genes were identified, of which 133 have orthologs in other plants. The remaining eight are unique to the oil palm and may be involved in palm-specific interactions with pathogen Avr gene products. Validation using multiple sequence alignments of the oil palm CNL genes and their orthologs showed a conserved kinase-2 motif with the last residue W (Tryptophan) in the NBS domain in most of the oil palm CNL genes. Out of the 141 oil palm CNL genes, only nine sequences do not have the final residue, W. The W residue is highly conserved in non-TIR NBS-LRR genes [116]. The percentage of CNL genes (67.14%) agrees with Staskawicz et al. [121] who reported that the majority of disease resistance genes in plants contain the NBS-LRR domain [122].

Another class of R genes critical for plant defense is the Kinase class. This class contains an intracellular serine/threonine protein kinase (STK) domain which plays an important role in many plant processes, including plant-pathogen interaction [45,47,123,124]. *Pto*, an R gene previously identified in the tomato genome to confer

12

resistance against *Pseudomonas syringae* pv. *tomato* strains, belongs to the Kinase class [46,125]. There are several features defining the *Pto* gene in tomato - Pto activation domain,[123] autophosphorylation sites [45,47,123], P+1 loop [125] and N-myristoylation motif [124]. Seven candidate genes in the oil palm genome have the required features. Sequence alignment between the candidate genes and *Pto* revealed several highly conserved sites in the Pto activation domain. However, the third autophosphorylation site in the activation domain had a threonine to glycine mutation (Fig. S1), which was reported to reduce the plant hypersensitive response [123].

The remaining R genes identified were RLP, RLK and Mlo-like. The high quality oil palm dataset contains three RLP and three RLK genes. Both classes contain the transmembrane and LRR domains [45], but only RLK contains an additional STK domain (Fig. 6B). RLP and RLK genes function as pattern recognition receptors (PRRs) in the transmembrane region, and are activated in the initial detection of pathogen in the plant [114,126]. Other plants, such as *A. thaliana* (9.8% RLK and 4.0% RLP) and *O. sativa* (10.5% RLK and 5.4% RLP), have higher percentages of these genes in their genomes. The difference could be due to the fact that none of the transcriptomes used as evidence in the gene prediction process originated from stress-related tissues. The percentage of these two classes can be higher but only the six identified RLK and RLP genes were expressed in the transcriptomes used. Oil palm also has 13 candidate Mlo-like genes, classified by having the Mlo domain [45]. The first member of this class, *MLO* gene from barley, was expressed in leaf in response to invasion by a fungal pathogen, *Erysiphe graminis* f sp. *Hordei*.[126] MLO (Powdery mildew locus O) is an intrinsic protein with six transmembrane regions [127]. The oil palm MLO-like candidates have six/seven transmembrane regions.

About 70% of the 210 candidate R genes were distributed across 16 oil palm chromosomes of EG5 genome build [5] (Fig. S2). 101 out of the 141 CNL class R genes were found on 14 of the chromosomes, of which 62 formed 23 clusters by chromosomal location. The highest number of clustered CNL class R genes (42%) were found on the chromosome 2. R genes in other plants (such as thale cress, flax, barley, lettuce, maize, potato, rice, soybean and tomato) also form location clusters [128]. Plant resistance is determined by (direct or indirect) interaction of plant R genes with pathogens' Avr genes, and evolve to adapt to the different forms of Avr genes [120,129]. Co-located R genes recognize different pathogens and are hypothesised to share function and pathogen recognition systems [128].

Since R genes are important for plant survival and surveillance system, the R genes-related domains seem to be evolutionary conserved across all sequenced plants, including the oil palm. The high-quality dataset was used to find the necessary domains to classify the R genes into six classes. The identification of these candidate genes would be useful for marker development and gene expression studies during infection, especially for basal stem rot, one of the most devastating oil palm diseases in the South East Asia. Comparing the oil palm genome with those of other monocots, it was possible to identify R genes for further functional characterization, and reveal homologous sequences in related crops.

3.5   FA Biosynthesis Genes

Oil palm is unique in that it produces oils with distinct fatty acid profiles in its mesocarp and kernel. The *E. guineensis* mesocarp oil is ~50% saturated (39.2-45.8% palmitic acid [C16:0], 3.7-5.1% stearic acid [C18:0] and 0.9-1.5% myristic acid [C14:0]), 37.4-44.1% unsaturated (mainly oleic acid [C18:1]) and ~10.5% polyunsaturated (10.2% linoleic acid [C18:2] and 0.3% linolenic acid [C18:3]) [130]. The kernel oil is more saturated, with mainly medium chain fatty acids - lauric ([C12:0], ~48%), myristic (~15%) and palmitic (~8%) acid [131]. The fatty acid compositions also vary noticeably both between and within *E. guineensis* and *E. oleifera* populations [132,133]. *E. oleifera* mesocarp oil is typically less saturated (53.5-68.7% oleic acid, 11.9%-26.9% linoleic acid and 0.0%-1.9% linolenic acid) [133]. Forty-two oil palm genes involved in FA biosynthesis, including two multifunctional acetyl-CoA carboxylases (ACCase), were identified (Supplementary file 1). Figures 7A and 7B show the numbers of oil palm genes in the FA biosynthesis pathway and oil palm fatty acid composition respectively. The conserved catalytic residues were identified via sequence alignment of the corresponding amino acids (Figs. S3-S14). This method was used by Li et al. [64] to study the candidate FA biosynthesis genes of *Arachis hypogaea* L. Twenty seven FA biosynthesis genes were categorized in 10 classes based on the conserved catalytic residues of their corresponding amino acid sequences, while six were identified by their conserved motifs. The remaining nine genes encoding ACCase were mainly classified by homology. Using a 70% identity cut-off, 39 candidate oil palm FA biosynthesis genes had 94 corresponding orthologs in *A. thaliana* (29) and *Z. mays* (65). Overall, these results showed that the classifications were consistent with the annotations of *A. thaliana* and *Z. mays* genes. The three remaining candidate genes, one acyl-ACP thioesterase (*EgFATB_1*) and two stearoyl-ACP desaturases (*EgFAB2_3* and *EgFAB2_4*), were defined as singletons. Closer examination of *EgFAB2_3* indicates that the gene could be truncated, as it had a gap in its genomic region, making it a singleton.

ACCase plays an important role in *de novo* FA biosynthesis as it catalyzes the first committed step in the pathway [134]. Analysis of the ACCase genes showed that oil palm contains both the multi-subunit (CT [3 copies], BCCP [2 copies] and BC [2 copies]), and multifunctional (2 copies) forms. This agrees with Wan Omar et al. [135]. who reported two distinct forms of ACCase in oil palm. After the first committed step, stepwise addition of two-carbon residues from malonyl-ACP continues until palmitoyl-ACP (C16:0-ACP). C16:0-ACP is converted to C18:0-ACP by β-ketoacyl-ACP synthase II (FABF).[136] Biochemical analysis showed that FABF activity, and the level of C18:1, were negatively correlated with the level of C16:0 [131]. FABF activity in *E. guineensis* was <50% of several accessions of *E. oleifera* [131]. Although *E. guineensis* has three copies of *FABF*, expression analysis showed a dominant copy in the mesocarp and kernel. *EgFABF_1* is at least 2.8x and 19.2x more highly expressed in mesocarp and kernel respectively than the other two copies (Fig. 8A), suggesting that the conversion of C16:0-ACP to C18:0-ACP is mainly driven by *EgFABF_1*. Overexpression of this gene copy may drive palm oil to higher oleic acid content. The second copy of *EgFABF_2* is also expressed in both the mesocarp and kernel samples but at lower levels. This is in line with Umi Salamah et al. [137] who reported that *FABF*, which is similar to *EgFABF_2* (93% identity at nucleotides level) was also expressed in both mesocarp and kernel samples at relatively higher levels than other tissues using northern blot analysis. The remaining *EgFABF_3* has very low expression.

Stearoyl-ACP desaturase (encoded by FAB2 [138–140]) which converts C18:0-ACP to C18:1-ACP is important in both triacylglycerol and phosphatidylcholine synthesis [141]. FAB2 plays a crucial role in determining the ratio of saturated to unsaturated C18 fatty acids in plant membranes and storage lipids. Multiple alignment of the corresponding amino acid sequences of the oil palm and other plants *FAB2* genes identified two important motifs (EENRH and DEKRH). In this study, *FAB2* gene has the highest number of copies among all FA biosynthesis genes identified. This is not unexpected as oil palm has a moderately high oleic acid content, ~40% in mesocarp and ~15% in kernel. FAB2 is a very active enzyme in the developing oil palm mesocarp and any effort to increase oleic acid content may not therefore require upregulating the gene(s) expressing stearoyl-ACP desaturase [130]. Ortholog analysis showed that two oil palm FAB2 genes (*EgFAB2_3* and *EgFAB2_4*) are singletons while four (*EgFAB2_1*, *EgFAB2_2*, *EgFAB2_5* and *EgFAB2_6*) are similar to orthologs in *A. thaliana* and *Z. mays*.

EgFAB2_1, *EgFAB2_5* and *EgFAB2_6* were in the same clade as FAB2 genes encoded by AT2G43710 (SSI2), AT5G16240 (S-ACP-DES1) and AT3G02630 (S-ACP-DES5) in *A. thaliana* (Fig. 9). This is interesting because SSI2 is involved in determining the 18:1 pool in *A. thaliana* leaves [142] and has a substrate preference for C18 over C16 fatty acids [142,143]. Surprisingly, *EgFAB2_1* has highest expression in the mesocarp and kernel (Fig. 8B), suggesting that *EgFAB2_1* is the dominant copy of the *FAB2* gene, and largely responsible for conversion of C18:0-ACP to C18:1-ACP in *de novo* FA biosynthesis in the oil palm mesocarp and kernel. *EgFAB2_6* also has a relatively high expression in the mesocarp, and lower in the kernel. The gene can also contribute to the production of C18:1-ACP in the mesocarp, as knocking out *SSI2* in *A. thaliana* only reduced desaturase activity by 90% [142]. *EgFAB2_3* and *EgFAB2_5* are hardly expressed in the mesocarp, but highly in the kernel, indicating tissue specific expression. *EgFAB2_3* and *EgFAB2_5* may play a more important role in C18:1 production in the kernel than mesocarp. *EgFAB2_2* has the highest divergence from the other four genes in the phylogenetic tree, and is orthologous to the *A. thaliana* gene, AT1G43800. Northern analysis of AT1G43800 in *A. thaliana* showed that the gene is not expressed in the leaf, stem, root, flower or silique [142]. This is in line with the oil palm 454-transcriptome data, which showed that the *EgFAB2_2* is not expressed in the leaf, root or stalk, with only slight expression in flower (data not shown). Based on expression analysis, *EgFAB2_2, EgFAB2_3,* and *EgFAB2_5* may play more important roles in C18:1 production in the oil palm kernel than mesocarp. The remaining copy of the *FAB2* gene (*EgFAB2_4*) has very low expression in mesocarp and kernel.

C18:1 may be further desaturated to polyunsaturated fatty acids in the plastid or endoplasmic reticulum (ER). FAD2 and FAD3, localized in the ER, are responsible for the synthesis of C18:2 and C18:3, respectively, in storage oils. *EgFAD2_1* and *EgFAD3_1* are the dominant copies of *FAD2* and *FAD3*, respectively, that probably drive the desaturation of C18:1 in the mesocarp (Fig. 8C-D). The expression data showed higher FAD2 and FAD3 expression in the mesocarp than kernel, consistent with the fact that the mesocarp oil contains C18:2 and some C18:3 which are insignificant in kernel oil.

Acyl-ACP thioesterases terminate *de novo* chain elongation by hydrolyzing the acyl-groups on acyl-ACP fatty acids [144,145]. The unesterified fatty acids released are exported to the ER for modification, such as assembly

15

into triacylglycerols and/or membrane lipids. Thioesterases are either FATA or FATB, depending on their specificity for acyl groups - FATA prefers unsaturated and FATB saturated. Six oil palm acyl-ACP thioesterase genes were identified. The corresponding amino acid sequences of the genes contain two conserved motifs, NQHVNN and YRRECG. However, the conserved YRRECG motif in oil palm and other plants differed from the PFAM HMMLogo (Figs. S13 and S14), in line with Voelker et al. [146], who postulated plant thioesterases as a different class of enzymes from those of animals and bacteria. Multiple alignment, BLAST, and ortholog analysis of the corresponding amino acid sequences (Fig. S15) were able to classify *EgFATA_1* and *EgFATA_2* as oleoyl-ACP thioesterases (*FATA*) genes. *EgFATA_1* and *EgFATA_2* are highly similar to experimentally derived oleoyl-ACP thioesterase AAD28187.1 in *E. guineensis* [147], with 97% and 89% BLASTP identity respectively, and to NP_001292940.1 from *J. curcas* (69% identity, 76% positives) and XP_007049712.1 from *T. cacao* (72% identity, 90% positives). Both these sequences had high homology and formed a clade with other characterized plant FATA genes. The remaining four could not be differentiated via sequence analysis but expression data suggested that they function as FATB to hydrolyze saturated acyl-ACPs. *EgFATB_1* is not expressed in mesocarp but has very high expression in the kernel, indicating that it is mainly involved in fatty acid chain termination in the kernel (Fig. 8F).

As oil palm accumulates 48% C12:0 and 15% C14:0 in its kernel oil, *EgFATB_1* probably encodes for a thioesterase with substrate specificity for medium chain, i.e., lauryl- or myristoyl-, ACP thioesterase. *EgFATB_2* and *EgFATB_3*, only moderately expressed in the mesocarp and kernel, are probably involved in the formation of C16:0 since the acid accumulates to ~44% in the mesocarp oil and 15% in the kernel oil. The remaining acyl-ACP thioesterase (*EgFATB_4*) is only detected at very low levels in both the mesocarp and kernel, and may code for stearoyl-ACP thioesterase as palm oil and palm kernel oil only contain 3.7 - 5.1% [130] and 0.5 - 5% [148] stearic acid, respectively.

Comparison of the genomic location of the FA biosynthesis genes in the oil palm genome showed that three genes, namely *EgFABF*, *EgFABH* and *EgFAD3* showed duplication events (Fig. S16). This is in accordance with the segmental duplications of chromosome arms reported by Singh et al. [5]. The study identified and characterized 42 key genes involved in FA biosynthesis in *E. guineensis*. This is the first study to identify key FA biosynthesis genes in both the oil palm mesocarp and kernel through sequence and gene expression analysis. The comprehensive information will help pave the way to an understanding of the different mechanisms involved in producing the unique fatty acid profiles of palm mesocarp and kernel oils.

4.   Conclusions

We developed an integrated gene prediction pipeline, enabling annotation of the genome of the African oil palm and present a set of 26,059 high quality and thoroughly validated gene models. To achieve this, we conducted an in-depth analysis of several important gene categories: IG, R and FA biosynthesis. The prevalence of these groups was similar across several plant genomes, including *A. thaliana*, *Z. mays*, *O. sativa*, *S. bicolor*, *G. max* and *R. communis*.

Coding regions of the oil palm genome have a characteristic broad distribution of GC_3, with a heavy tail extending to high GC_3 values where it contains many stress-related and intronless genes. GC_3-rich genes in oil palm are significantly over-represented in the following GOslim process categories: responses to abiotic stimulus, responses to endogenous stimulus, RNA translation, and responses to stress. BUSCO analysis showed that our high-quality gene models contained at least 90% of the known conserved orthologs in eukaryotes. We found that approximately one-seventh of the oil palm genes were intronless. 210 R genes grouped into six classes based on their protein domain structures were also identified. Lipid-, especially fatty acid (FA)-related genes, are of particular interest in oil palm where, in addition to their roles in specifying oil yield and quality, also contribute to the plant organization and function as well as having important signaling roles related to biotic and abiotic stresses. We identified 42 key genes involved in FA biosynthesis in oil palm. Our study will facilitate understanding of the plant genome organization, and will be an important resource for further comparative and evolutionary analysis. The study on these genes will facilitate further advances in the regulation of gene function in oil palm and provide a theoretical foundation for marker-assisted breeding of the crop with increased oil yield and elevated contents of oleic and other value-added fatty acids.

Availability

As part of the Malaysian Oil Palm Genome Programme (MyOPGP), and to support post-genomics efforts, we developed the palmXplore system (http://palmxplore.mpob.gov.my), a public archive for the collection of oil palm (*Elaeis guineensis*) gene models hosted by MPOB. The collection includes 26,059 representative gene models and annotation information from different annotation methods: Enzyme Code (KEGG), Gene Ontology (Blast2GO) and PFAM. The gene models are also downloadable from http://genomsawit.mpob.gov.my.

Acknowledgements

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version.

References

[1]   E. Barcelos, S. de A. Rios, R.N. V. Cunha, R. Lopes, S.Y. Motoike, E. Babiychuk, A. Skirycz, S. Kushnir, Oil palm natural diversity and the potential for yield improvement, Front. Plant Sci. 6 (2015) 190. doi:10.3389/fpls.2015.00190.

[2]   MPOB, Malaysian Oil Palm Statistics 2014, 34th ed., MPOB, Malaysia, 2015.

[3]   S. Jouannic, X. Argout, F. Lechauve, C. Fizames, A. Borgel, F. Morcillo, F. Aberlenc-Bertossi, Y. Duval, J. Tregear, Analysis of expressed sequence tags from oil palm (Elaeis guineensis), FEBS Lett. 579 (2005) 2709–2714. doi:10.1016/j.febslet.2005.03.093.

[4]   E.T.L. Low, R. Rosli, N. Jayanthi, A.H. Mohd-Amin, N. Azizi, K.L. Chan, N.J. Maqbool, P. Maclean, R. Brauning, A. McCulloch, R. Moraga, M. Ong-Abdullah, R. Singh, Analyses of hypomethylated oil palm gene space, PLoS One. 9 (2014). doi:10.1371/journal.pone.0086728.

[5]   R. Singh, M. Ong-Abdullah, E.-T.L. Low, M.A.A. Manaf, R. Rosli, R. Nookiah, L.C.-L. Ooi, S.-E. Ooi, K.-L. Chan, M.A. Halim, N. Azizi, J. Nagappan, B. Bacher, N. Lakey, S.W. Smith, D. He, M. Hogan, M. a Budiman, E.K. Lee, R. Desalle, D. Kudrna, J.L. Goicoechea, R. a Wing, R.K. Wilson, R.S. Fulton, J.M. Ordway, R. a Martienssen, R. Sambanthamurthi, Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds., Nature. 500 (2013) 335–9. doi:10.1038/nature12309.

[6]   R. Singh, E.-T.L. Low, L.C.-L. Ooi, M. Ong-Abdullah, N.-C. Ting, J. Nagappan, R. Nookiah, M.D. Amiruddin, R. Rosli, M.A.A. Manaf, K. Chan, M.A. Halim, N. Azizi, N. Lakey, S.W. Smith, M.A. Budiman, M. Hogan, B. Bacher, A. Van Brunt, C. Wang, J.M. Ordway, R. Sambanthamurthi, R.A. Martienssen, The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK., Nature. 500 (2013) 340–4. doi:10.1038/nature12356.

[7]   R. Singh, E.-T.L. Low, L.C.-L. Ooi, M. Ong-Abdullah, R. Nookiah, N.-C. Ting, M. Marjuni, P.-L. Chan, M. Ithnin, M.A.A. Manaf, J. Nagappan, K.-L. Chan, R. Rosli, M.A. Halim, N. Azizi, M. a Budiman, N. Lakey, B. Bacher, A. Van Brunt, C. Wang, M. Hogan, D. He, J.D. MacDonald, S.W. Smith, J.M. Ordway, R. a Martienssen, R. Sambanthamurthi, The oil palm VIRESCENS gene controls fruit colour and encodes a R2R3-MYB., Nat. Commun. 5 (2014) 4106. doi:10.1038/ncomms5106.

[8]   M. Ong-Abdullah, J.M. Ordway, N. Jiang, S.-E. Ooi, S. Kok, N. Sarpan, N. Azimi, A.T. Hashim, Z. Ishak, S.K. Rosli, F.A. Malike, N.A.A. Bakar, M. Marjuni, N. Abdullah, Z. Yaakub, M.D. Amiruddin, R. Nookiah, R. Singh, E.L. Low, K. Chan, N. Azizi, S.W. Smith, B. Bacher, M.A. Budiman, A. Van Brunt, C. Wischmeyer, M. Beil, M.

18

Hogan, N. Lakey, C. Lim, X. Arulandoo, C.-K. Wong, C. Choo, W. Wong, Y. Kwan, S.S.R.S. Alwee, R. Sambanthamurthi, R.A. Martienssen, Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm, Nature. 525 (2015) 533–537. doi:10.1038/nature15365.

[9]   C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA., J. Mol. Biol. 268 (1997) 78–94. doi:10.1006/jmbi.1997.0951.

[10]   M.R. Brent, Genome annotation past, present, and future: How to define an ORF at each locus, Genome Res. 15 (2005) 1777–1786. doi:10.1101/gr.3866105.

[11]   K.W. Berendzen, K. Stüber, K. Harter, D. Wanke, Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves., BMC Bioinformatics. 7 (2006) 522. doi:10.1186/1471-2105-7-522.

[12]   M. Pritsker, Y.C. Liu, M.A. Beer, S. Tavazoie, Whole-genome discovery of transcription factor binding sites by network-level conservation, Genome Res. 14 (2004) 99–108. doi:10.1101/gr.1739204.

[13]   M. Troukhan, T. Tatarinova, J. Bouck, R.B. Flavell, N.N. Alexandrov, Genome-wide discovery of cis-elements in promoter sequences using gene expression, OMICS. 13 (2009) 139–151. doi:10.1089/omi.2008.0034 [doi].

[14]   M. Triska, D. Grocutt, J. Southern, D.J. Murphy, T. Tatarinova, CisExpress: Motif detection in DNA sequences, Bioinformatics. 29 (2013) 2203–2205. doi:10.1093/bioinformatics/btt366.

[15]   T. V Tatarinova, N.N. Alexandrov, J.B. Bouck, K.A. Feldmann, GC3 biology in corn, rice, sorghum and other grasses., BMC Genomics. 11 (2010) 308. doi:10.1186/1471-2164-11-308.

[16]   K.-L. Chan, R. Rosli, T. Tatarinova, M. Hogan, M. Firdaus-Raih, E.-T.L. Low, Seqping: Gene Prediction Pipeline for Plant Genomes using Self-Trained Gene Models and Transcriptomic Data, BMC Bioinformatics. (2017) 10.1186/s12859-016-1426-6. doi:10.1186/s12859-016-1426-6.

[17]   V. Solovyev, P. Kosarev, I. Seledsov, D. Vorobyev, Automatic annotation of eukaryotic genes, pseudogenes and promoters., Genome Biol. 7 Suppl 1 (2006) S10.1-12. doi:10.1186/gb-2006-7-s1-s10.

[18]   F. Bourgis, A. Kilaru, X. Cao, G.-F. Ngando-Ebongue, N. Drira, J.B. Ohlrogge, V. Arondel, Comparative transcriptome and metabolite analysis of oil palm and date palm mesocarp that differ dramatically in carbon partitioning., Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 12527–12532. doi:10.1073/pnas.1106502108.

[19]   T.J. Tranbarger, S. Dussert, T. Joët, X. Argout, M. Summo, A. Champion, D. Cros, A. Omore, B. Nouy, F. Morcillo, Regulatory mechanisms underlying oil palm fruit mesocarp maturation, ripening, and functional specialization in lipid and carotenoid metabolism., Plant Physiol. 156 (2011) 564–584. doi:10.1104/pp.111.175141.

[20]   J.R. Shearman, C. Jantasuriyarat, D. Sangsrakru, T. Yoocha, A. Vannavichit, S. Tragoonrung, S.

Tangphatsornruang, Transcriptome analysis of normal and mantled developing oil palm flower and fruit, Genomics. 101 (2013) 306–312. doi:10.1016/j.ygeno.2013.02.012.

[21]   J.R. Shearman, C. Jantasuriyarat, D. Sangsrakru, T. Yoocha, A. Vannavichit, S. Tangphatsornruang, S. Tragoonrung, Transcriptome Assembly and Expression Data from Normal and Mantled Oil Palm Fruit, Dataset Pap. Biol. 2013 (2013) 1–7. doi:10.7167/2013/670926.

[22]   H. Yao, L. Guo, Y. Fu, L.A. Borsuk, T.J. Wen, D.S. Skibbe, X. Cui, B.E. Scheffler, J. Cao, S.J. Emrich, D.A. Ashlock, P.S. Schnable, Evaluation of five ab initio gene prediction programs for the discovery of maize genes, Plant Mol. Biol. 57 (2005) 445–460. doi:10.1007/s11103-005-0271-1.

[23]   A.A. Salamov, V. V. Solovyev, Ab initio gene finding in Drosophila genomic DNA, Genome Res. 10 (2000) 516–522. doi:10.1101/gr.10.4.516.

[24]   C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects, BMC Bioinformatics. 12 (2011) 491. doi:10.1186/1471-2105-12-491.

[25]   S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402. doi:10.1093/nar/25.17.3389.

[26]   L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data, Bioinformatics. 28 (2012) 3150–3152. doi:10.1093/bioinformatics/bts565.

[27]   S. Ouyang, C.R. Buell, The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants., Nucleic Acids Res. 32 (2004) D360–D363. doi:10.1093/nar/gkh099.

[28]   J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements, Cytogenet. Genome Res. 110 (2005) 462–467. doi:10.1159/000084979.

[29]   C. Llorens, R. Futami, L. Covelli, L. Domínguez-Escribá, J.M. Viu, D. Tamarit, J. Aguilar-Rodríguez, M. Vicente-Ripolles, G. Fuster, G.P. Bernet, F. Maumus, A. Munoz-Pomer, J.M. Sempere, A. Latorre, A. Moya, The Gypsy Database (GyDB) of mobile genetic elements: release 2.0., Nucleic Acids Res. 39 (2011) D70-4. doi:10.1093/nar/gkq1061.

[30]   W.H. Majoros, M. Pertea, S.L. Salzberg, TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders, Bioinformatics. 20 (2004) 2878–2879. doi:10.1093/bioinformatics/bth315.

[31]   J.E. Allen, W.H. Majoros, M. Pertea, S.L. Salzberg, JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions., Genome Biol. 7 Suppl 1 (2006) S9.1-13. doi:10.1186/gb-2006-7-s1-s9.

[32]   M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to

improve de novo gene finding, Bioinformatics. 24 (2008) 637–644. doi:10.1093/bioinformatics/btn013.

[33] I. Korf, Gene finding in novel genomes., BMC Bioinformatics. 5 (2004) 59. doi:10.1186/1471-2105-5-59.

[34] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res. 35 (2007). doi:10.1093/nar/gkl842.

[35] E.L.L. Sonnhammer, S.R. Eddy, E. Birney, A. Bateman, R. Durbin, Pfam: Multiple sequence alignments and HMM-profiles of protein domains, Nucleic Acids Res. 26 (1998) 320–322. doi:10.1093/nar/26.1.320.

[36] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L.L. Sonnhammer, J. Tate, M. Punta, Pfam: The protein families database, Nucleic Acids Res. 42 (2014). doi:10.1093/nar/gkt1223.

[37] L.S. Johnson, S.R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure., BMC Bioinformatics. 11 (2010) 431. doi:10.1186/1471-2105-11-431.

[38] J. Mistry, R.D. Finn, S.R. Eddy, A. Bateman, M. Punta, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions, Nucleic Acids Res. 41 (2013). doi:10.1093/nar/gkt263.

[39] Y. Kawahara, M. de la Bastide, J.P. Hamilton, H. Kanamori, W.R. McCombie, S. Ouyang, D.C. Schwartz, T. Tanaka, J. Wu, S. Zhou, K.L. Childs, R.M. Davidson, H. Lin, L. Quesada-Ocampo, B. Vaillancourt, H. Sakai, S.S. Lee, J. Kim, H. Numa, T. Itoh, C.R. Buell, T. Matsumoto, Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data., Rice (N. Y). 6 (2013) 4. doi:10.1186/1939-8433-6-4.

[40] T.Z. Berardini, L. Reiser, D. Li, Y. Mezheritsky, R. Muller, E. Strait, E. Huala, The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome, Genesis. 53 (2015) 474–485. doi:10.1002/dvg.22877.

[41] D.M. Goodstein, S. Shu, R. Howson, R. Neupane, R.D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, D.S. Rokhsar, Phytozome: A comparative platform for green plant genomics, Nucleic Acids Res. 40 (2012) 1178–1186. doi:10.1093/nar/gkr944.

[42] W. Sanseverino, A. Hermoso, R. D'Alessandro, A. Vlasova, G. Andolfo, L. Frusciante, E. Lowy, G. Roma, M.R. Ercolano, PRGdb 2.0: Towards a community-based database model for the analysis of R-genes in plants, Nucleic Acids Res. 41 (2013). doi:10.1093/nar/gks1183.

[43]   a Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes., J. Mol. Biol. 305 (2001) 567–580. doi:10.1006/jmbi.2000.4315.

[44] A. Barbosa-da-Silva, A.C. Wanderley-Nogueira, R.R.M. Silva, L.C. Berlarmino, N.M. Soares-Cavalcanti, A.M.

Benko-Iseppon, In silico of resistance (R) genes in Eucalyptus transcriptome, Genet. Mol. Biol. 28 (2005) 562–574. doi:10.1590/S1415-47572005000400011.

[45]   G.B. Martin, A.J. Bogdanove, G. Sessa, Understanding the functions of plant disease resistance proteins, Annu. Rev. Plant Biol. 54 (2003) 23–61. doi:10.1146/annurev.arplant.54.031902.135035.

[46]   S. Peraza-Echeverria, A. James-Kay, B. Canto-Canché, E. Castillo-Castro, Structural and phylogenetic analysis of Pto-type disease resistance gene candidates in banana, Mol. Genet. Genomics. 278 (2007) 443–453. doi:10.1007/s00438-007-0262-9.

[47]   W.Y. Song, G.L. Wang, L.L. Chen, H.S. Kim, L.Y. Pi, T. Holsten, J. Gardner, B. Wang, W.X. Zhai, L.H. Zhu, C. Fauquet, P. Ronald, A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21., Science. 270 (1995) 1804–1806. doi:10.1126/science.270.5243.1804.

[48]   C. Yun, Classification and function of plant disease resistance genes, Plant Pathol. J. 15 (1999) 105–111.

[49]   J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence, Nucleic Acids Res. 22 (1994) 4673–4680.

[50]   K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: Molecular evolutionary genetics analysis version 6.0, Mol. Biol. Evol. 30 (2013) 2725–2729. doi:10.1093/molbev/mst197.

[51]   S. Yang, X. Zhang, J.X. Yue, D. Tian, J.Q. Chen, Recent duplications dominate NBS-encoding gene expansion in two woody species, Mol. Genet. Genomics. 280 (2008) 187–198. doi:10.1007/s00438-008-0355-0.

[52]   B.C. Meyers, A. Kozik, A. Griego, H. Kuang, R.W. Michelmore, Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis, Plant Cell Online. 15 (2003) 809–834. doi:10.1105/tpc.009308.

[53]   M. Kanehisa, S. Goto, Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res. 28 (2000) 27–30. doi:10.1093/nar/28.1.27.

[54]   J. Okuley, J. Lightner, K. Feldmann, N. Yadav, E. Lark, J. Browse, Arabidopsis FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis., Plant Cell …. 6 (1994) 147–158. doi:10.1105/tpc.6.1.147.

[55]   X. Lin, S. Kaul, S. Rounsley, T.P. Shea, M.I. Benito, C.D. Town, C.Y. Fujii, T. Mason, C.L. Bowman, M. Barnstead, T. V Feldblyum, C.R. Buell, K. a Ketchum, J. Lee, C.M. Ronning, H.L. Koo, K.S. Moffat, L. a Cronin, M. Shen, G. Pai, S. Van Aken, L. Umayam, L.J. Tallon, J.E. Gill, M.D. Adams, a J. Carrera, T.H. Creasy, H.M. Goodman, C.R. Somerville, G.P. Copenhaver, D. Preuss, W.C. Nierman, O. White, J. a Eisen, S.L. Salzberg, C.M. Fraser, J.C. Venter, Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana., Nature. 402 (1999) 761–768. doi:10.1038/45471.

[56]   G.B. Tilton, J.M. Shockey, J. Browse, Biochemical and Molecular Characterization of ACH2, an Acyl-CoA Thioesterase from Arabidopsis thaliana, J. Biol. Chem. 279 (2004) 7487–7494. doi:10.1074/jbc.M309532200.

22

[57]   S.S. Jha, J.K. Jha, B. Chattopadhyaya, A. Basu, S.K. Sen, M.K. Maiti, Cloning and characterization of cDNAs encoding for long-chain saturated acyl-ACP thioesterases from the developing seeds of Brassica juncea, Plant Physiol. Biochem. 48 (2010) 476–480. doi:10.1016/j.plaphy.2010.02.006.

[58]   G.S.C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison, BMC Bioinformatics. 6 (2005) 31. doi:10.1186/1471-2105-6-31.

[59]   E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, R. Lopez, InterProScan: Protein domains identifier, Nucleic Acids Res. 33 (2005) 116–120. doi:10.1093/nar/gki442.

[60]   E.L.L. Sonnhammer, S.R. Eddy, R. Durbin, Pfam: A comprehensive database of protein domain families based on seed alignments, Proteins Struct. Funct. Genet. 28 (1997) 405–420. doi:10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L.

[61]   B.K. Benson, G. Meades, A. Grove, G.L. Waldrop, DNA inhibits catalysis by the carboxyltransferase subunit of acetyl-CoA carboxylase: implications for active site communication., Protein Sci. 17 (2008) 34–42. doi:10.1110/ps.073186408.

[62]   M.C. Barber, N.T. Price, M.T. Travers, Structure and regulation of acetyl-CoA carboxylase genes of metazoa, Biochim. Biophys. Acta - Mol. Cell Biol. Lipids. 1733 (2005) 1–28. doi:10.1016/j.bbalip.2004.12.001.

[63]   G.L. Waldrop, I. Rayment, H.M. Holden, Three-dimensional structure of the biotin carboxylase subunit of acetyl-CoA carboxylase., Biochemistry. 33 (1994) 10249–10256. doi:10.1021/bi00200a004.

[64]   M.-J. Li, A.-Q. Li, H. Xia, C.-Z. Zhao, C.-S. Li, S.-B. Wan, Y.-P. Bi, X.-J. Wang, Cloning and sequence analysis of putative type II fatty acid synthase genes from Arachis hypogaea L., J. Biosci. 34 (2009) 227–238. doi:10.1007/s12038-009-0027-1.

[65]   K. Haralampidis, D. Milioni, Temporal and transient expression of stearoyl-ACP carrier protein desaturase gene during olive fruit development, J. Exp. Bot. 49 (1998) 1661–1669. doi:10.1093/jxb/49.327.1661.

[66]   J. Shanklin, E. Whittle, B.G. Fox, Eight histidine residues are catalytically essential in a membrane-associated iron enzyme, stearoyl-CoA desaturase, and are conserved in alkane hydroxylase and xylene monooxygenase., Biochemistry. 33 (1994) 12787–12794. doi:10.1021/bi00209a009.

[67]   L. Yuan, B.A. Nelson, G. Caryl, The catalytic cysteine and histidine in the plant acyl-acyl carrier protein thioesterases, J. Biol. Chem. 271 (1996) 3417–3419. doi:10.1074/jbc.271.7.3417.

[68]   S. Brenner, The molecular evolution of genes and proteins: a tale of two serines., Nature. 334 (1988) 528–530. doi:10.1038/334528a0.

[69]   D.A. Rozwarski, C. Vilchèze, M. Sugantino, R. Bittman, J.C. Sacchettini, Crystal structure of the Mycobacterium tuberculosis enoyl-ACP reductase, InhA, in complex with NAD+ and a C16 fatty acyl substrate, J. Biol. Chem.

274 (1999) 15582–15589. doi:10.1074/jbc.274.22.15582.

[70] S. Smith, The animal fatty acid synthase: one gene, one polypeptide, seven enzymes., FASEB J. 8 (1994) 1248–59.

[71] G.M.J. Helmkamp, K. Bloch, Beta-hydroxydecanoyl thioester dehydrase. Studies on molecular structure and active site, J Biol Chem. 244 (1969) 6014–22.

[72] M. Siggaard-Andersen, Conserved residues in condensing enzyme domains of fatty acid synthases and related sequences, Protein Seq Data Anal. 5 (1993) 325–335.

[73] A.L. Delcher, A. Phillippy, J. Carlton, S.L. Salzberg, Fast algorithms for large-scale genome alignment and comparison., Nucleic Acids Res. 30 (2002) 2478–2483. doi:10.1093/nar/30.11.2478.

[74] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions., Genome Biol. 14 (2013) R36. doi:10.1186/gb-2013-14-4-r36.

[75] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn, L. Pachter, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks., Nat. Protoc. 7 (2012) 562–78. doi:10.1038/nprot.2012.016.

[76] L. Li, C.J. Stoeckert, D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes., Genome Res. 13 (2003) 2178–89. doi:10.1101/gr.1224503.

[77] R. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797. doi:10.1093/nar/gkh340.

[78] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780. doi:10.1093/molbev/mst010.

[79] A. Mitchell, H. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. Mcanulla, C. Mcmenamin, G. Nuka, S. Pesseat, A. Sangrador-vegas, M. Scheremetjew, C. Rato, S. Yong, A. Bateman, M. Punta, T.K. Attwood, C.J.A. Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D.A. Natale, C.H. Wu, C. Orengo, I. Sillitoe, H. Mi, P.D. Thomas, R.D. Finn, The InterPro protein families database : the classification resource after 15 years, Nucleic Acids Res. 43 (2015) 213–221. doi:10.1093/nar/gku1243.

[80] C.J.A. Sigrist, E. De Castro, L. Cerutti, A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, I. Xenarios, New and continuing developments at PROSITE, Nucleic Acids Res. 41 (2013) 344–347. doi:10.1093/nar/gks1067.

[81] A. Marchler-bauer, M.K. Derbyshire, N.R. Gonzales, S. Lu, F. Chitsaz, L.Y. Geer, R.C. Geer, J. He, M. Gwadz, D.I. Hurwitz, C.J. Lanczycki, F. Lu, G.H. Marchler, J.S. Song, N. Thanki, Z. Wang, R.A. Yamashita, D. Zhang,

C. Zheng, H. Bryant, CDD□: NCBI's conserved domain database, Nucleic Acids Res. 43 (2015) 222–226. doi:10.1093/nar/gku1221.

[82] S. Kuraku, C.M. Zmasek, O. Nishimura, K. Katoh, aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity., Nucleic Acids Res. 41 (2013) 22–28. doi:10.1093/nar/gkt389.

[83] M.S. Campbell, M. Law, C. Holt, J.C. Stein, G.D. Moghe, D.E. Hufnagel, J. Lei, R. Achawanantakun, D. Jiao, C.J. Lawrence, D. Ware, S.-H. Shiu, K.L. Childs, Y. Sun, N. Jiang, M. Yandell, MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations., Plant Physiol. 164 (2014) 513–24. doi:10.1104/pp.113.230144.

[84] C. Liang, L. Mao, D. Ware, L. Stein, Evidence-based gene predictions in plant genomes, Genome Res. 19 (2009) 1912–1923. doi:10.1101/gr.088997.108.

[85] V. Curwen, E. Eyras, T.D. Andrews, L. Clarke, E. Mongin, S.M.J. Searle, M. Clamp, The Ensembl automatic gene annotation system, Genome Res. 14 (2004) 942–950. doi:10.1101/gr.1858004.

[86] S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R.L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman, R.C. Buell, The TIGR Rice Genome Annotation Resource: Improvements and new features, Nucleic Acids Res. 35 (2007). doi:10.1093/nar/gkl976.

[87] W. Zhu, C.R. Buell, Improvement of whole-genome annotation of cereals through comparative analyses, Genome Res. 17 (2007) 299–310. doi:10.1101/gr.5881807.

[88] D. Swarbreck, C. Wilks, P. Lamesch, T.Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang, E. Huala, The Arabidopsis Information Resource (TAIR): Gene structure and function annotation, Nucleic Acids Res. 36 (2008). doi:10.1093/nar/gkm965.

[89] Y. Ge, Y. Wang, Y. Liu, Y. Tan, X. Ren, X. Zhang, K.D. Hyde, Y. Liu, Z. Liu, Comparative genomic and transcriptomic analyses of the Fuzhuan brick tea-fermentation fungus Aspergillus cristatus, BMC Genomics. 17 (2016) 428. doi:10.1186/s12864-016-2637-y.

[90] S. Calabrese, J. Pérez-Tienda, M. Ellerbeck, C. Arnould, O. Chatagnier, T. Boller, A. Schüßler, A. Brachmann, D. Wipf, N. Ferrol, P.-E. Courty, GintAMT3 – a Low-Affinity Ammonium Transporter of the Arbuscular Mycorrhizal Rhizophagus irregularis, Front. Plant Sci. 7 (2016) 1–14. doi:10.3389/fpls.2016.00679.

[91] Z. Zhong, J. Norvienyeku, M. Chen, J. Bao, L. Lin, L. Chen, Y. Lin, X. Wu, Z. Cai, Q. Zhang, X. Lin, Y. Hong, J. Huang, L. Xu, H. Zhang, L. Chen, W. Tang, H. Zheng, X. Chen, Y. Wang, B. Lian, L. Zhang, H. Tang, G. Lu, D.J. Ebbole, B. Wang, Z. Wang, G.S. Gilbert, C.O. Webb, P. Schulze-Lefert, R. Panstruga, M.E. Woolhouse, D.T. Haydon, R. Antia, I.M. Parker, G.S. Gilbert, S. Raffaele, E.H. Stukenbrock, T. Bataillon, E.H. Stukenbrock,

J. Schirawski, M.C. Heath, S. Kamoun, R.A. Wilson, N.J. Talbot, E.A. Kellogg, B.S. Gaut, D.J. Ebbole, R. Dean, J.L. Maciel, M.L. Farman, S. Kang, J.A. Sweigard, B. Valent, J.A. Sweigard, Y. Tosa, S.A. Leong, J.P. Gustafson, J. Taylor, G. Stacey, Y. Zheng, W. Li, J. Wu, T. Sone, J. Huang, W. Si, Q. Deng, P. Li, S. Yang, S. Zhang, B.C. Couch, L.M. Kohn, B.C. Couch, J. Choi, S. Klaubauf, J. Luo, Y. Dong, R.A. Dean, M. Xue, C. Chen, W. Gao, C.H. Khang, S.Y. Park, Y.H. Lee, S. Kang, C.J. Rehmeyer, W. Li, M. Kusaba, M.L. Farman, T. de Lange, W.K. Chu, I.D. Hickson, C. Rehmeyer, J.W. Taylor, M.L. Berbee, J.E. Galagan, M.R. Henn, L.J. Ma, C.A. Cuomo, B. Birren, N.P. Keller, G. Turner, J.W. Bennett, S. Raffaele, S. Kamoun, B. Cresnar, S. Petric, J.F. Doebley, B.S. Gaut, B.D. Smith, B.A. McDonald, C. Linde, Z. He, Z.M. Li, X.M. Zheng, S. Ge, M. Wang, E. Mace, W. Xie, M.E. Woolhouse, J.P. Webster, E. Domingo, B. Charlesworth, B.R. Levin, R. Chikhi, P. Medvedev, G.S. Slater, E. Birney, T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, A. Krogh, B. Larsson, G. Von Heijne, E.L. Sonnhammer, B. Langmead, S.L. Salzberg, H. Li, A. McKenna, M.A. DePristo, G.A. Van der Auwera, H. Li, K. Tamura, S. Kurtz, P.J. Bradbury, Y. Wang, M. Zytnicki, E. Akhunov, H. Quesneville, G. Abrusan, N. Grundmann, L. DeMester, W. Makalowski, M. Tarailo-Graovac, N. Chen, L. Li, C.J. Stoeckert, D.S. Roos, T. De Bie, N. Cristianini, J.P. Demuth, M.W. Hahn, M.A. Larkin, S.F. Altschul, P. Librado, J. Rozas, D. Wang, Y. Zhang, Z. Zhang, J. Zhu, J. Yu, Z. Yang, R. Nielsen, Directional Selection from Host Plants Is a Major Force Driving Host Specificity in Magnaporthe Species, Sci. Rep. 6 (2016) 25591. doi:10.1038/srep25591.

[92]   C. Van Ghelder, D. Esmenjaud, TNL genes in peach□: insights into the post- LRR domain, BMC Genomics. (2016) 1–16. doi:10.1186/s12864-016-2635-0.

[93]   F.A. Sima, R.M. Waterhouse, P. Ioannidis, E. V Kriventseva, E.M. Zdobnov, BUSCO□: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics. 2015 (2015) 1–3. doi:10.1093/bioinformatics/btv351.

[94]   I.B. Rogozin, A. V. Sverdlov, V.N. Babenko, E. V. Koonin, Analysis of evolution of exon-intron structure of eukaryotic genes, Brief. Bioinform. 6 (2005) 118–134. doi:10.1093/bib/6.2.118.

[95]   S. Schwartz, E. Meshorer, G. Ast, Chromatin organization marks exon-intron structure., Nat. Struct. & Mol. Biol. 16 (2009) 990–995. doi:10.1038/nsmb.1659.

[96]   I.M. Meyer, R. Durbin, Gene structure conservation aids similarity based gene prediction, Nucleic Acids Res. 32 (2004) 776–783. doi:10.1093/nar/gkh211.

[97]   N.N. Alexandrov, V. V. Brover, S. Freidin, M.E. Troukhan, T. V. Tatarinova, H. Zhang, T.J. Swaller, Y.-P.P. Lu, J. Bouck, R.B. Flavell, K.A. Feldmann, Insights into corn genes derived from large-scale cDNA sequencing, Plant Mol. Biol. 69 (2009) 179–194. doi:10.1007/s11103-008-9415-4.

[98]   T. Tatarinova, E. Elhaik, M. Pellegrini, Cross-species analysis of genic GC3 content and DNA methylation patterns, Genome Biol. Evol. 5 (2013) 1443–1456. doi:10.1093/gbe/evt103.

[99]   A. Souvorov, T. Tatusova, L. Zaslasky, B. Smith-White, Glycine max and Zea mays Genome Annotation with Gnomon, in: ISMB/ECCB, 2011.

[100]   E. Elhaik, M. Pellegrini, T. V Tatarinova, Gene expression and nucleotide composition are associated with genic methylation level in Oryza sativa., BMC Bioinformatics. 15 (2014) 23. doi:10.1186/1471-2105-15-23.

[101]   T. Ahmad, G. Sablok, T. V. Tatarinova, Q. Xu, X.-X.X. Deng, W.-W.W. Guo, Evaluation of codon biology in citrus and Poncirus trifoliata based on genomic features and frame corrected expressed sequence tags., DNA Res. 20 (2013) 135–50. doi:10.1093/dnares/dss039.

[102]   S. Kariin, C. Burge, Dinucleotide relative abundance extremes: a genomic signature, Trends Genet. 11 (1995) 283–290. doi:10.1016/S0168-9525(00)89076-9.

[103]   S.M. Agarwal, P.K. Srivastava, Human intronless disease associated genes are slowly evolving., BMB Rep. 42 (2009) 356–360. doi:10.5483/BMBRep.2009.42.6.356.

[104]   E. a. Grzybowska, Human intronless genes: Functional groups, associated diseases, evolution, and mRNA processing in absence of splicing, Biochem. Biophys. Res. Commun. 424 (2012) 1–6. doi:10.1016/j.bbrc.2012.06.092.

[105]   A. a. Ferguson, N. Jiang, Pack-MULEs: recycling and reshaping genes through GC-biased acquisition, Mob. Genet. Elements. 1 (2011) 135–138. doi:10.4161/mge.1.2.16948.

[106]   M. Jain, P. Khurana, A.K. Tyagi, J.P. Khurana, Genome-wide analysis of intronless genes in rice and Arabidopsis, Funct. Integr. Genomics. 8 (2008) 69–78. doi:10.1007/s10142-007-0052-9.

[107]   H. Yan, C. Jiang, X. Li, L. Sheng, Q. Dong, X. Peng, Q. Li, Y. Zhao, H. Jiang, B. Cheng, PIGD: a database for intronless genes in the Poaceae, BMC Genomics. 15 (2014) 832. doi:10.1186/1471-2164-15-832.

[108]   H. Yan, W. Zhang, Y. Lin, Q. Dong, X. Peng, H. Jiang, S. Zhu, B. Cheng, Different evolutionary patterns among intronless genes in maize genome, Biochem. Biophys. Res. Commun. 449 (2014) 146–150. doi:10.1016/j.bbrc.2014.05.008.

[109]   A. Bolshoy, T. Tatarinova, Methods of combinatorial optimization to reveal factors affecting gene length, Bioinform. Biol. Insights. 6 (2012) 317–327. doi:10.4137/BBI.S10525.

[110]   T. Tatarinova, B. Salih, J. Dien Bard, I. Cohen, A. Bolshoy, Lengths of Orthologous Prokaryotic Proteins Are Affected by Evolutionary Factors., Biomed Res. Int. 2015 (2015) 786861. doi:10.1155/2015/786861.

[111]   D. Kordis, Extensive intron gain in the ancestor of placental mammals, Biol Direct. 6 (2011) 59. doi:10.1186/1745-6150-6-59.

[112]   D. Kordis, J. Kokosar, What can domesticated genes tell us about the intron gain in mammals?, Int. J. Evol. Biol. 2012 (2012) 7. doi:10.1155/2012/278981.

[113]   B.C. Freeman, G.A. Beattie, An Overview of Plant Defenses against Pathogens and Herbivores, Plant Heal. Instr. (2008). doi:10.1094/PHI-I-2008-0226-01.

[114]   J.D.G. Jones, J.L. Dangl, The plant immune system., Nature. 444 (2006) 323–9. doi:10.1038/nature05286.

[115]   F. Katagiri, K. Tsuda, Understanding the plant immune system., Mol. Plant. Microbe. Interact. 23 (2010) 1531–1536. doi:10.1094/MPMI-04-10-0099.

[116]   B.C. Meyers, A.W. Dickerman, R.W. Michelmore, S. Sivaramakrishnan, B.W. Sobral, N.D. Young, Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide binding superfamily, Plant J. 20 (1999) 317–332.

[117]   C. Ameline-Torregrosa, B.-B. Wang, M.S. O'Bleness, S. Deshpande, H. Zhu, B.A. Roe, N.D. Young, S.B. Cannon, Identification and characterization of NBS-LRR genes in the model plant Medicago truncatula, Plant Physiol. 146 (2007) pp.107.104588. doi:10.1104/pp.107.104588.

[118]   D.E.K. Tarr, H.M. Alexander, TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders., BMC Res. Notes. 2 (2009) 197. doi:10.1186/1756-0500-2-197.

[119]   Q. Pan, J. Wendel, R. Fluhr, Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes., J. Mol. Evol. 50 (2000) 203–213. doi:10.1007/s002399910023.

[120]   D.A. Jones, J.D.G. Jones, The Role of Leucine-Rich Repeat Proteins in Plant Defences, Adv. Bot. Res. 24 (1997) 89–167. doi:10.1016/S0065-2296(08)60072-5.

[121]   B.J. Staskawicz, F.M. Ausubel, B.J. Baker, J.G. Ellis, J.D. Jones, Molecular genetics of plant disease resistance, Science (80-.). 268 (1995) 661–667. doi:10.1126/science.7732374.

[122]   D. Marone, M. a. Russo, G. Laidò, A.M. De Leonardis, A.M. Mastrangelo, Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: Active guardians in host defense responses, Int. J. Mol. Sci. 14 (2013) 7302–7326. doi:10.3390/ijms14047302.

[123]   G. Sessa, M. D'Ascenzo, G.B. Martin, Thr38 and Ser198 are Pto autophosphorylation sites required for the AvrPto-Pto-mediated hypersensitive response., EMBO J. 19 (2000) 2257–69. doi:10.1093/emboj/19.10.2257.

[124]   L. Shan, V.K. Thara, G.B. Martin, J.M. Zhou, X. Tang, The pseudomonas AvrPto protein is differentially recognized by tomato and tobacco and is localized to the plant plasma membrane., Plant Cell. 12 (2000) 2323–2338. doi:10.1105/tpc.12.12.2323.

[125]   R.D. Frederick, R.L. Thilmony, G. Sessa, G.B. Martin, Recognition Specificity for the Bacterial Avirulence Protein AvrPto Is Determined by Thr-204 in the Activation Loop of the Tomato Pto Kinase, Mol. Cell. 2 (1998) 241–245. doi:http://dx.doi.org/10.1016/S1097-2765(00)80134-3.

[126]   M. Sekhwal, P. Li, I. Lam, X. Wang, S. Cloutier, F. You, Disease Resistance Gene Analogs (RGAs) in Plants, Int. J. Mol. Sci. 16 (2015) 19248–19290. doi:10.3390/ijms160819248.

[127]   R. Büschges, K. Hollricher, R. Panstruga, G. Simons, M. Wolter, a Frijters, R. van Daelen, T. van der Lee, P.

Diergaarde, J. Groenendijk, S. Töpsch, P. Vos, F. Salamini, P. Schulze-Lefert, The barley Mlo gene: a novel control element of plant pathogen resistance., Cell. 88 (1997) 695–705. doi:http://dx.doi.org/10.1016/S0092-8674(00)81912-1.

[128] R.W. Michelmore, B.C. Meyers, Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process, Genome Res. 8 (1998) 1113–1130. doi:10.1101/gr.8.11.1113.

[129] M.A. Gururani, J. Venkatesh, C.P. Upadhyaya, A. Nookaraju, S.K. Pandey, S.W. Park, Plant disease resistance genes: Current status and future directions, Physiol. Mol. Plant Pathol. 78 (2012) 51–65. doi:DOI 10.1016/j.pmpp.2012.01.002.

[130] R. Sambanthamurthi, K. Sundram, Y. Tan, Chemistry and biochemistry of palm oil., Prog. Lipid Res. 39 (2000) 507–558. doi:http://dx.doi.org/10.1016/S0163-7827(00)00015-1.

[131] R. Sambanthamurthi, O. Abrizah, R. Umi Salamah, Biochemical factors that control oil composition in the oil palm, J. Oil Palm Res. (1999) 23–33.

[132] A. Noh, N. Rajanaidu, A. Kushairi, Y. Mohd Rafil, A. Mohd Din, Z.A. Mohd Isa, G. Saleg, M. Rafii, M. Din, M. Isa, Variability in fatty acid composition, iodine value and carotene content in the MPOB oil palm germplasm collection from Angola, J. 0il Palm Res. 14 (2002) 18–23.

[133]   a Mohd Din, N. Rajanaidu, B. Jalani, Performance of Elaeis oleifera from Panama, Costa Rica, Colombia and Honduras in Malaysia, J Oil Palm Res. 12 (2000) 71–80.

[134] R. Sambanthamurthi, J.B. Ohlrogge, Acetyl-coA carboxylase activity in the oil palm, in: J.P. Williams, M.U. Khan, N.W. Lem (Eds.), Physiol. Biochem. Mol. Biol. Plant Lipids, Dordrecht: Kluwer Academic Publishers, 1996: p. 26.

[135] W.S.W. Omar, L.B. Willis, C. Rha, A.J. Sinskey, U.S. Ramli, A.M.M. Yunus, G.K.A. Parveez, R. Sambanthamurthi, Isolation and utilization of acetyl-coa carboxylase from oil palm (elaeis guineensis) mesocarp, J. Oil Palm Res. 2 (2008) 97–107.

[136] R. Umi Salamah, R. Sambanthamurthi, β-keto acyl ACP synthase II I oil palm (Elaeis guineensis) mesocarp, in: J.P. Williams, M.U. Khan, N.W. Lem (Eds.), Physiol. Biochem. Mol. Biol. Plant Lipids, Kluwer Academic Publishers, 1996: pp. 69–71.

[137] R. Umi Salamah, R. Sambanthamurthi, A.R. Omar, G.K.A. Parveez, M.A.A. Manaf, O. Abrizah, A.M.M. Yunus, S.-C. Cheah, S.S.R.S. Alwee, A. Siti Nor Akmar, M.B. Wahid, The isolation and characterisation of oil palm (Elaeis guineensis Jacq.) β-ketoacy-acyl carrier protein (ACP) synthase (KAS) II cDNA, J Oil Palm Res. 24 (2012) 1480–1491.

[138] Y.M. Zhang, C.C. Wang, H.H. Hu, L. Yang, Cloning and expression of three fatty acid desaturase genes from cold-sensitive lima bean (Phaseolus lunatus L.)., Biotechnol. Lett. 33 (2011) 395–401. doi:10.1007/s10529-010-

0432-4.

[139]   G.K.A. Parveez, O.A. Rasid, R. Sambanthamurthi, Genetic engineering of oil palm, in: W. Mohd Basri, Y.M. Choo, K.W. Chan (Eds.), Furth. Adv. Oil Palm Res., 2011: pp. 141–201.

[140]   A. Siti Nor Akmar, S.-C. Cheah, S. Aminah, L.C.-L. Ooi, R. Sambanthamurthi, D.J. Murphy, Characterization and regulation of the oil palm (Elaeis guineensis) stearoyl-ACP Desaturase genes, J Oil Palm Res. Special Is (1999) 1–17.

[141]   R. a. Sidorov, V.D. Tsydendambaev, Biosynthesis of fatty oils in higher plants, Russ. J. Plant Physiol. 61 (2013) 1–18. doi:10.1134/S1021443714010130.

[142]   A. Kachroo, J. Shanklin, E. Whittle, L. Lapchyk, D. Hildebrand, P. Kachroo, The Arabidopsis stearoyl-acyl carrier protein-desaturase family and the contribution of leaf isoforms to oleic acid synthesis, Plant Mol. Biol. 63 (2007) 257–271. doi:10.1007/s11103-006-9086-y.

[143]   P. Kachroo, J. Shanklin, J. Shah, E.J. Whittle, D.F. Klessig, A fatty acid desaturase modulates the activation of defense signaling pathways in plants., Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 9448–53. doi:10.1073/pnas.151258398.

[144]   O. Abrizah, C. Lazarus, T. Fraser, K. Stobart, Cloning of a palmitoyl-acyl carrier protein thioesterase from oil palm., Biochem. Soc. Trans. 28 (2000) 619–622. doi:10.1042/BST0280619.

[145]   O. Abrizah, Isolation and characterization of an acyl-acyl carrier protein (ACP) thioesterase from oil palm, University of Bristol, UK, 2001.

[146]   T.A. Voelker, A.C. Worrell, L. Anderson, J. Bleibaum, C. Fan, D.J. Hawkins, S.E. Radke, H.M. Davies, Fatty Acid Biosynthesis redirected to Medium Chains in Transgenic Oilseed Plants., Science. 257 (1992) 72–74. doi:10.1126/science.1621095.

[147]   O. Asemota, C.T. San, F.H. Shah, Isolation of a kernel oleoyl-ACP thioesterase gene from the oil palm Elaeis guineensis Jacq., Afr J Biotechnol. 3 (2004) 199–201.

[148]   R.H. V Corley, P.B. Tinker, The Oil Palm, Fourth edi, Blackwell Science Ltd, Oxford, 2003.

Tables

Table 1: Number of gene models predicted by Fgenesh++ and Seqping with different prediction support

	Length filtered	Repeat filtered	With oil palm transcriptome annotation	With RefSeq protein annotation	With oil palm transcriptome AND RefSeq protein annotation	Combined loci from two pipelines	With PFAM domains and RefSeq annotation
Fgenesh++	117,832	85,432	26,476	60,042	27,915	31,413	26,059 *
Seqping	45,913	40,528	27,648	17,822	17,680		

* After filtering potential contaminants

Table 2: Number of loci at different overlap thresholds between Fgenesh++ and Seqping gene models

Number of isoforms in locus	Number of loci at different overlap threshold							
	60%	65%	70%	75%	80%	85%	90%	95%
1	17268	17358	17474	17612	17781	18003	18305	18904
2	12881	12873	12837	12788	12722	12636	12531	12290
3	731	709	696	683	672	657	631	603
4	63	61	61	61	60	60	59	52
5	14	14	13	14	14	13	11	10
6	1	1	1	0	0	0	0	0
all loci	30,958	31,016	31,082	31,158	31,249	31,369	31,537	31,859

Figures
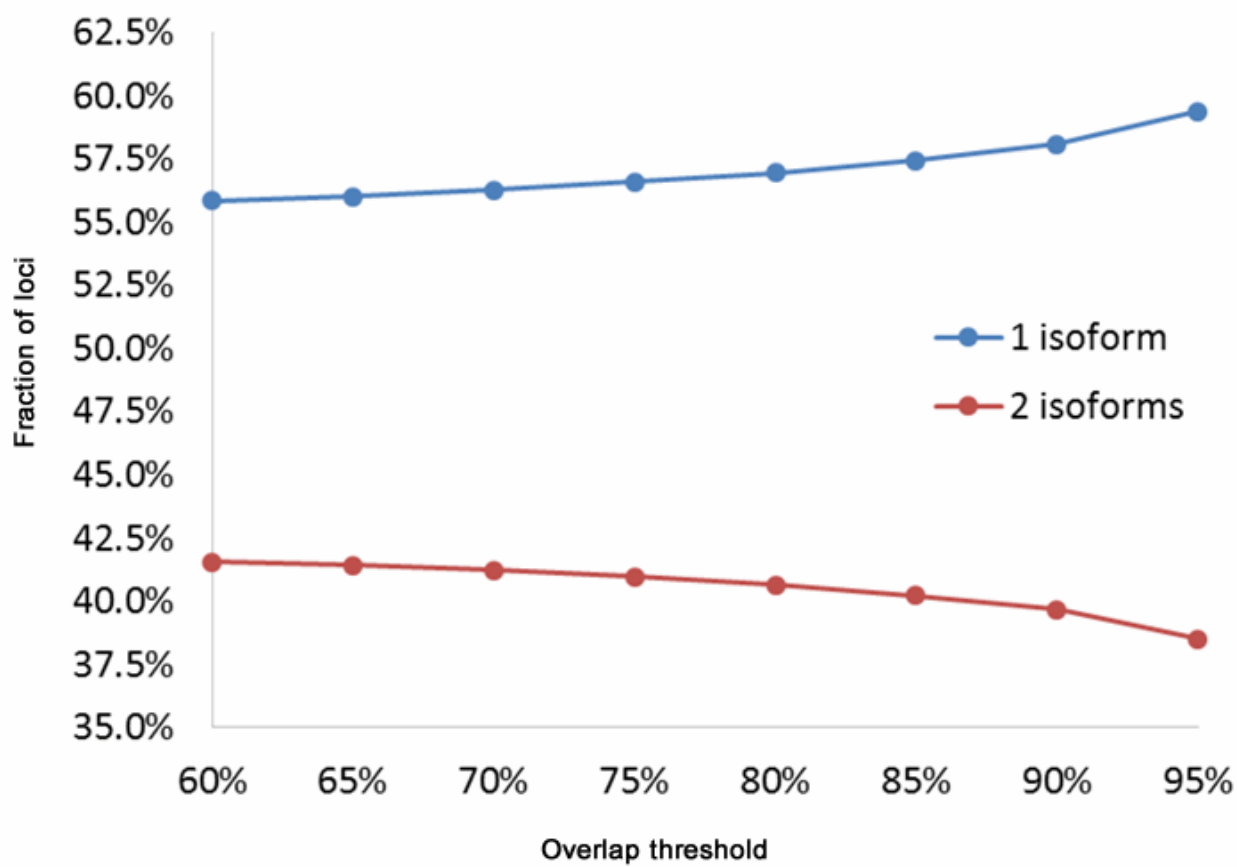
Figure 1: Selection of overlap thresholds using the rate of increase of fraction of single-isoform loci.
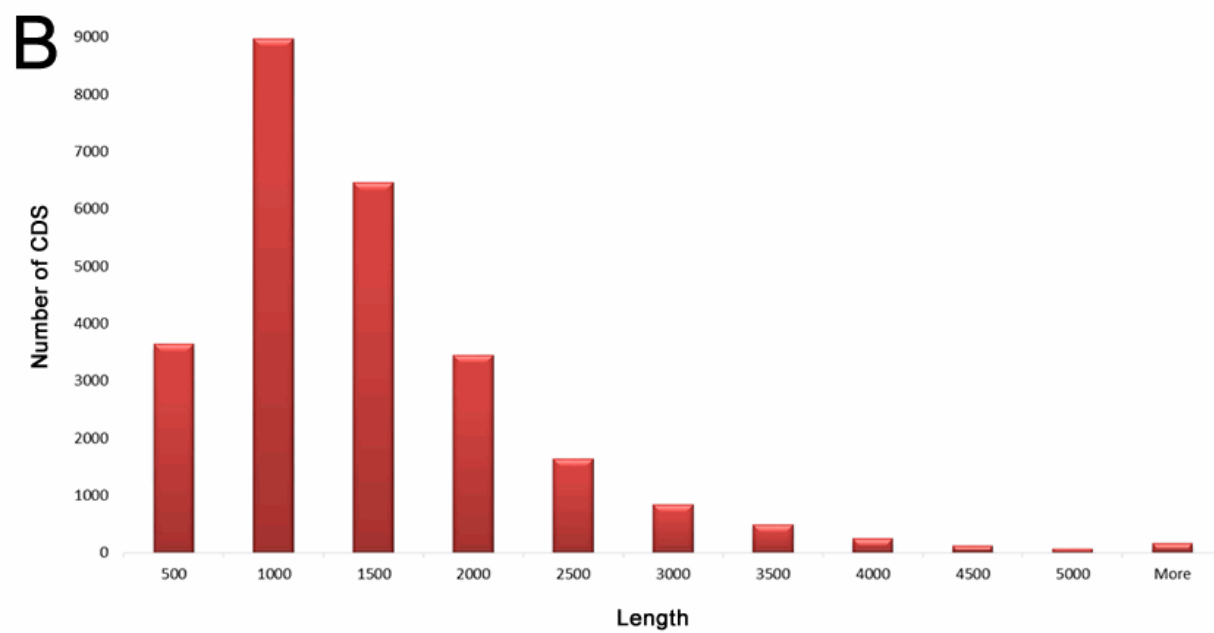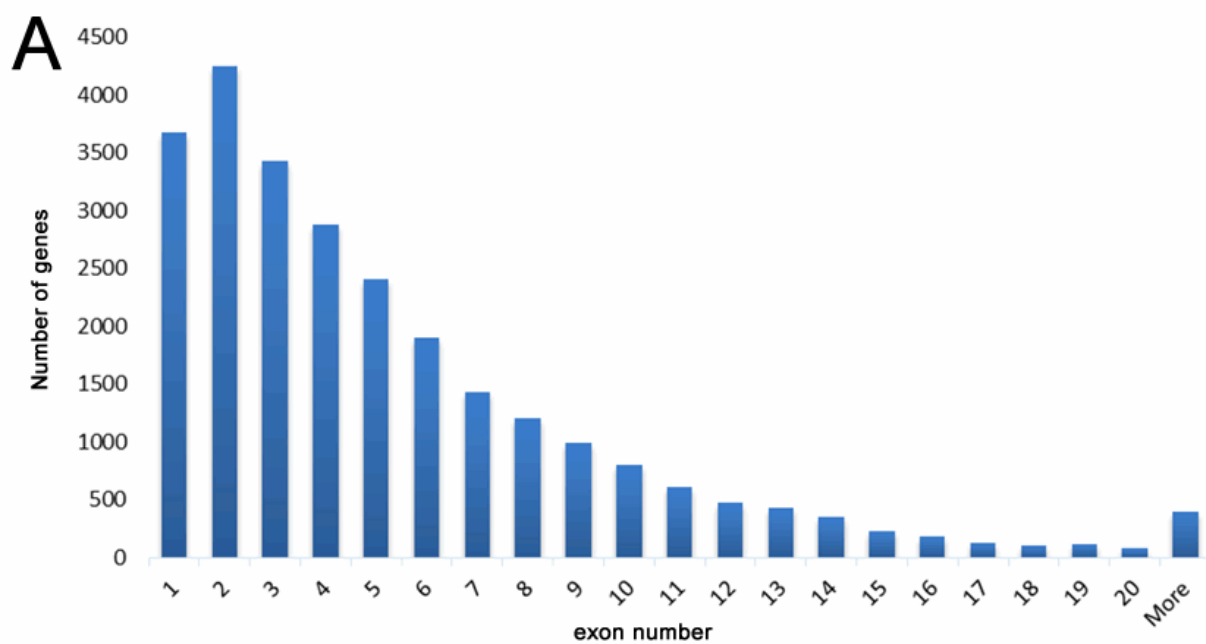
Figure 2: Distribution of oil palm gene models. (A) Number of genes *vs*. number of exons per gene (B) Number of genes *vs* lengths of CDS
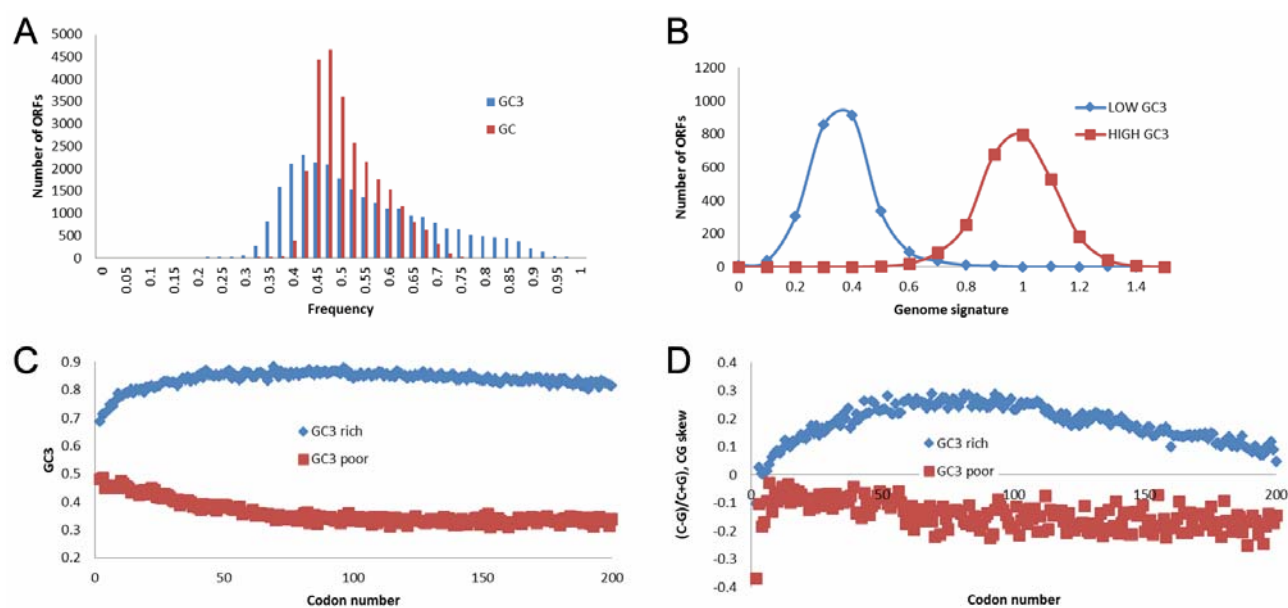
Figure 3: GC$_3$ distribution in oil palm gene models. (A) GC (red) and GC$_3$ (blue) composition of the coding regions of *E. guineensis*. (B) Genome signature for GC$_3$-rich and -poor genes. (C) GC$_3$ gradient along the open reading frames of GC$_3$-rich and -poor genes. (D) CG$_3$ skew gradient along the open reading frames of GC$_3$-rich and -poor genes.
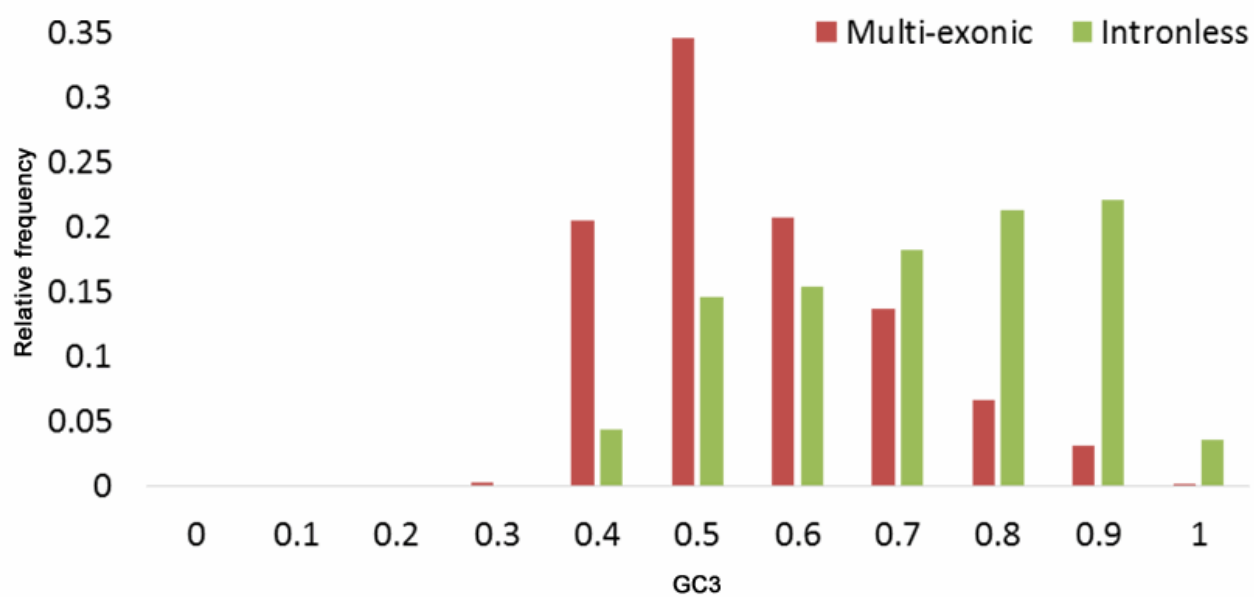
Figure 4: GC_3 contents of oil palm intronless and multi-exonic genes
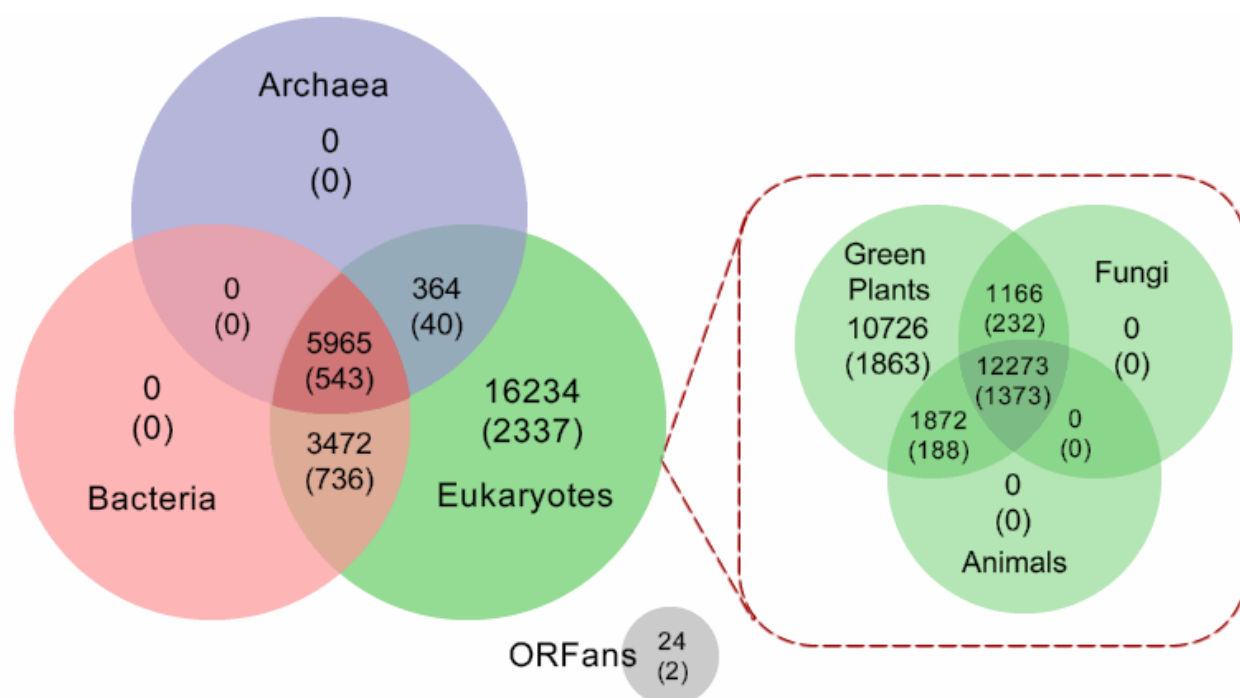
Figure 5: Classification of oil palm intronless genes (IG) in different taxonomy groups. The Venn diagram shows the projections of 26,059 oil palm high quality loci and 3,658 oil palm IG (shown in parenthesis) into three domains of life based on homology, including archaea, bacteria and eukaryotes. The sub-diagram shows the distribution of oil palm IG from the eukaryote domain into three major taxonomy groups of life - Green Plants, Fungi and Animals. ORFans refers to the unique sequence that shares no significant similarity with other organisms.

37
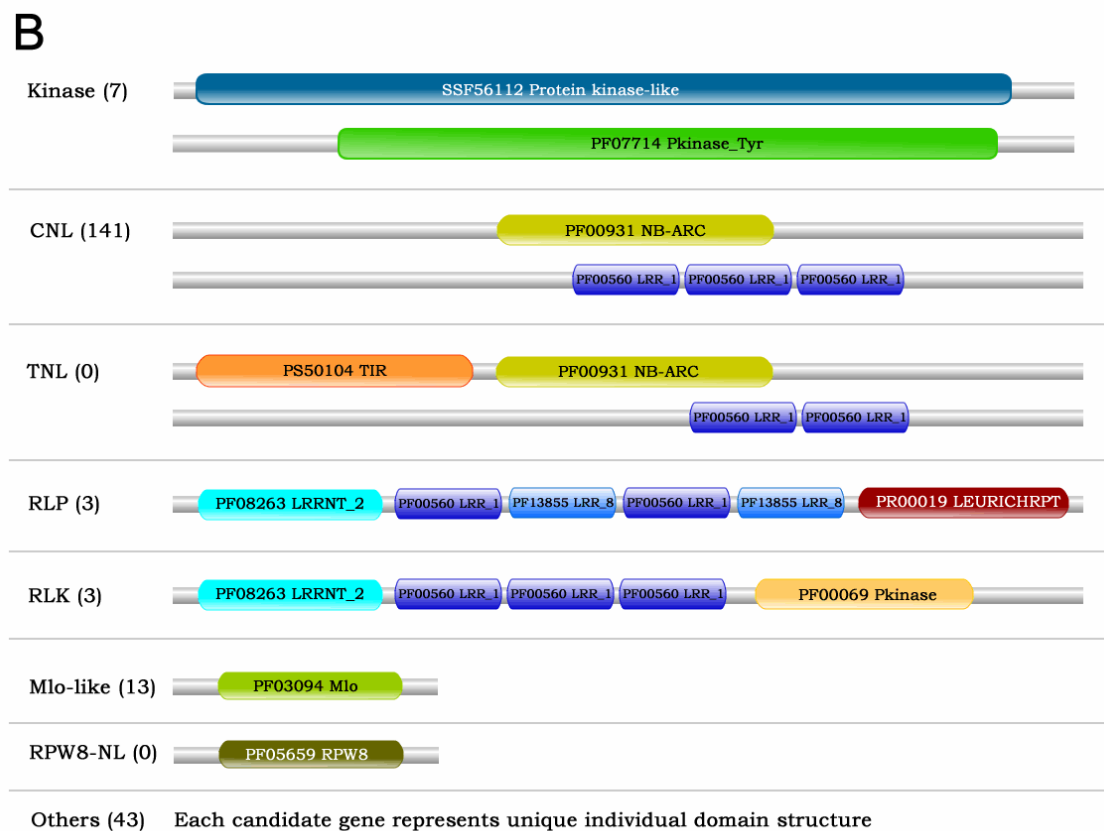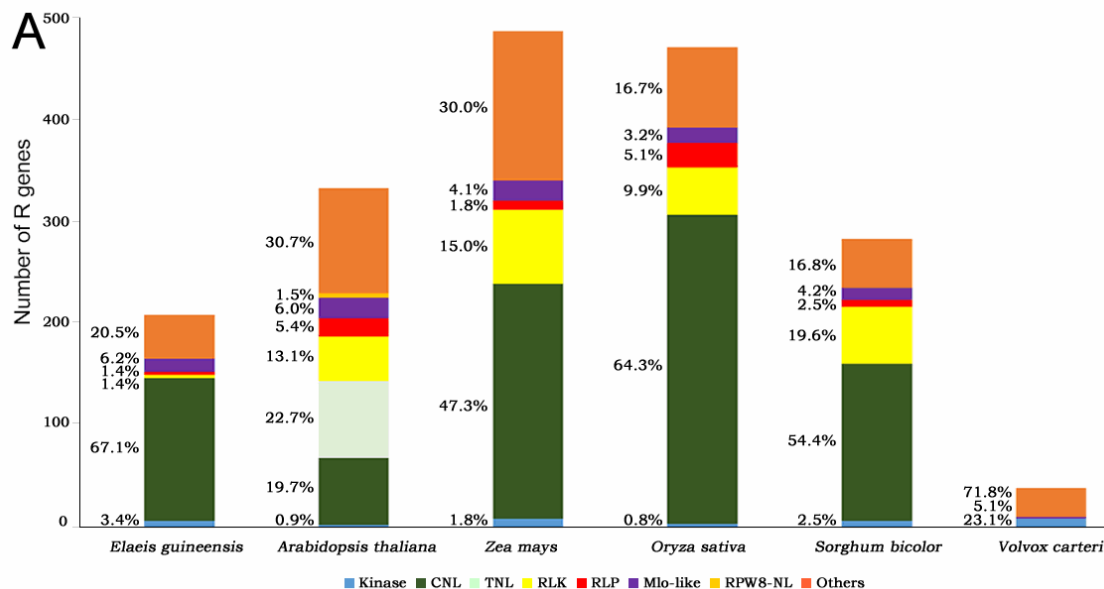
Figure 6: Classification of candidate R genes. (A) Distribution of candidate R genes in oil palm, *A. thaliana*, *Z. mays*, *O. sativa*, *S. bicolor* and *V. carteri* (B) Examples of key domains identified via Interproscan in oil palm candidate R-genes. Number of identified candidate oil palm genes are shown in brackets.
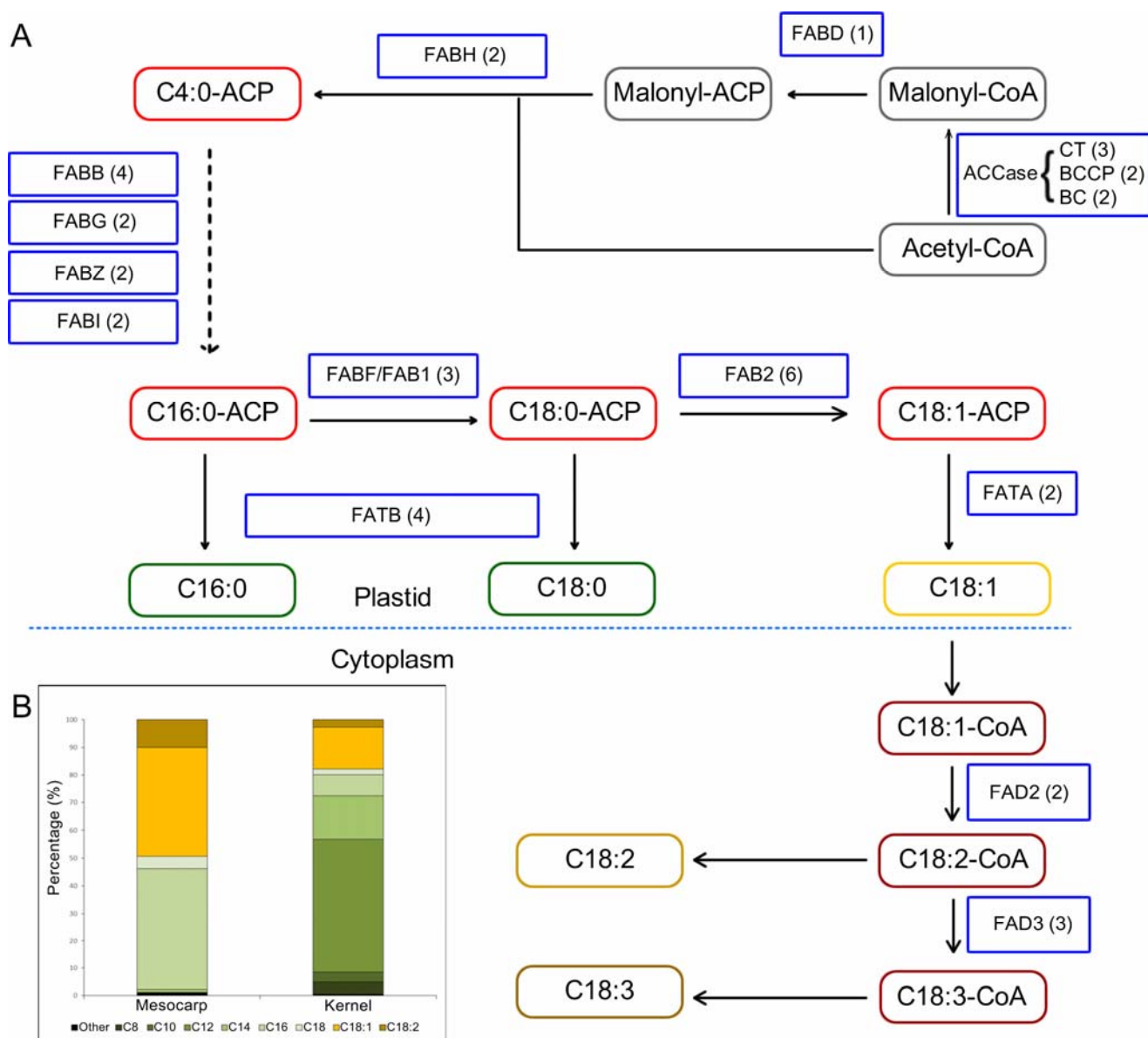
Figure 7: Fatty acid biosynthesis in *E. guineensis* (A) Schematic pathway diagram for fatty acid biosynthesis. Number of identified oil palm candidate genes are shown in brackets. (B) Fatty acid composition in mesocarp and kernel.
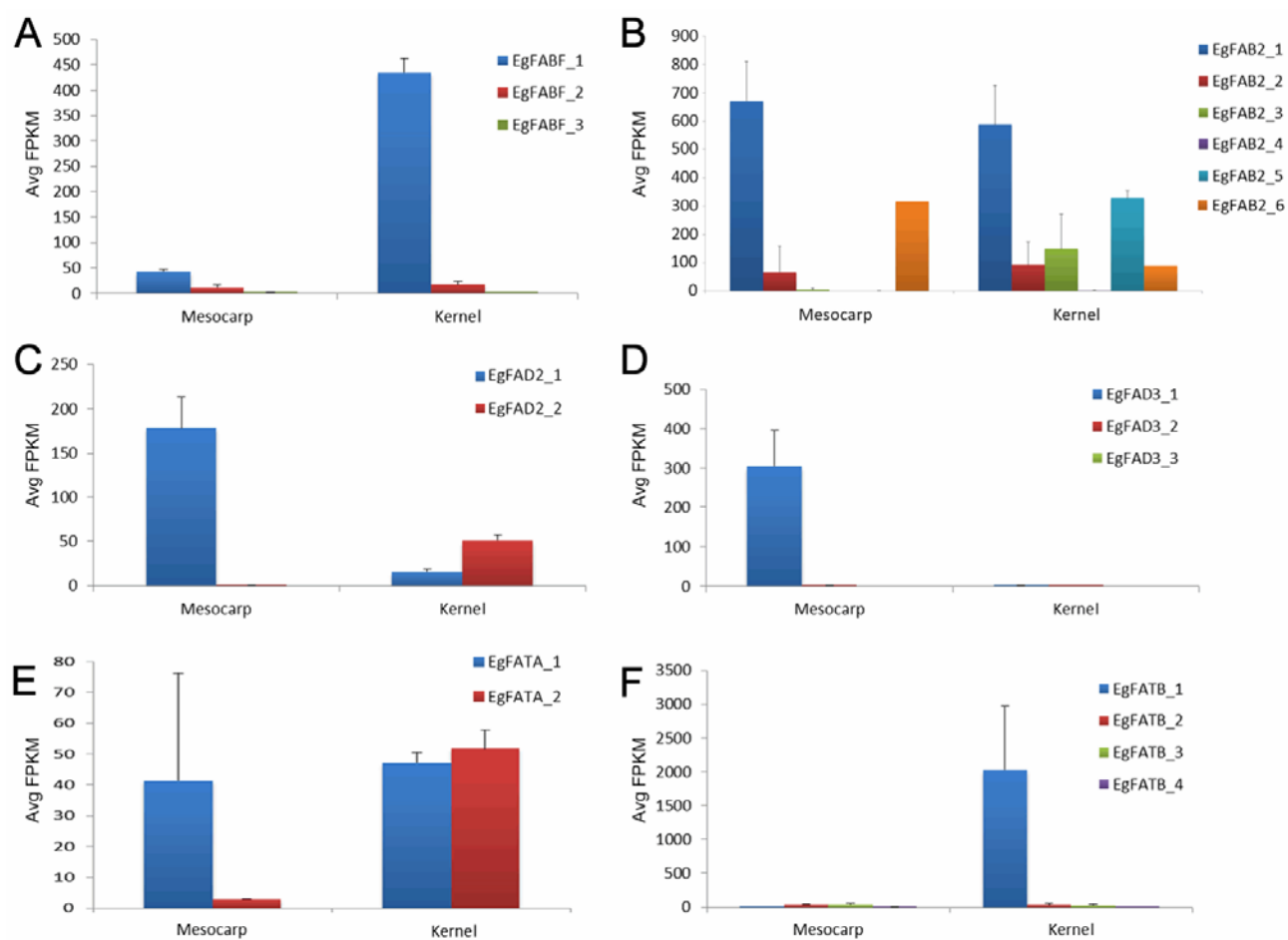
Figure 8: Transcriptome analysis of (A) FABF, (B) FAB2, (C) FAD2, (D) FAD3, (E) FATA and (F) FATB genes in mesocarp and kernel samples
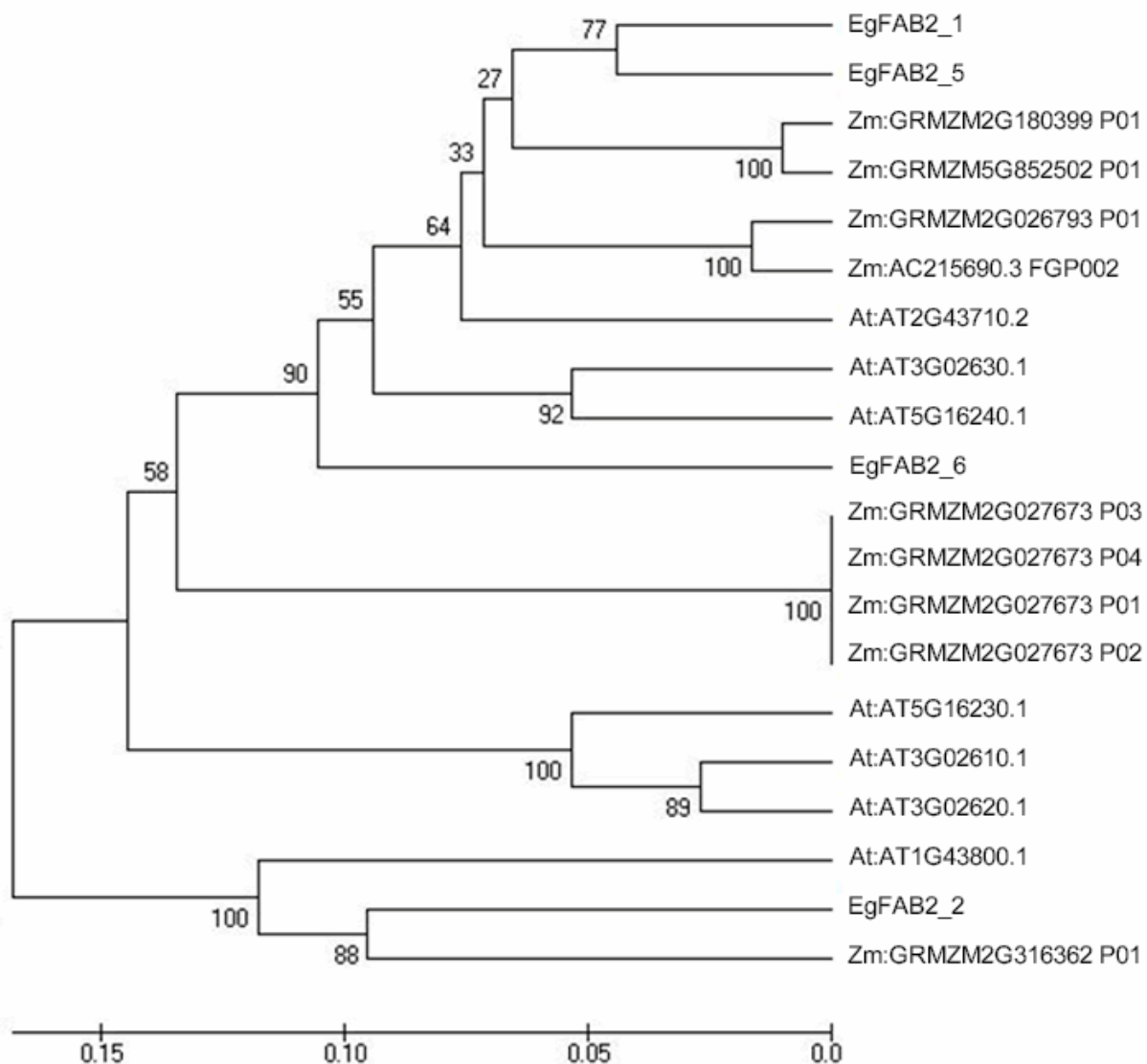
Figure 9: Evolutionary relationship of FAB2 in oil palm (*E. guineensis*), *A. thaliana* and *Z. mays*. Analyses were carried out using UPGMA method in MEGA 6 software. Abbreviations: Eg- *E. guineensis*; At - *A. thaliana*; Zm - *Z. mays*.