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

Chan Kuang Lim[1,2#], Tatiana V. Tatarinova[3#], Rozana Rosli[1,4], Nadzirah Amiruddin[1], Norazah Azizi[1], Mohd Amin Ab Halim[1], Nik Shazana Nik Mohd Sanusi[1], Jayanthi Nagappan[1], Petr Ponomarenko[3], Martin Triska[3,4], Victor Solovyev[5], Mohd Firdaus-Raih[2], Ravigadevi Sambanthamurthi[1], Denis Murphy[4], Leslie Low Eng Ti[1*]

1 Advanced Biotechnology and Breeding Centre, Malaysian Palm Oil Board, P.O. Box 10620, 50720 Kuala Lumpur, Malaysia; 2 Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia; 3 Spatial Sciences Institute, University of Southern California, Los Angeles, CA, 90089, USA; 4 Genomics and Computational Biology research group, University of South Wales, Pontypridd, CF371DL, United Kingdom; 5 Softberry Inc., 116 Radio Circle, Suite 400 Mount Kisco, NY 10549, USA.

* To whom correspondence should be addressed. Tel. +60387694504. Fax. +60389261995. E-mail: lowengti@mpob.gov.my

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

Keywords: oil palm, gene prediction, Seqping, intronless, genes

Abstract

The advent of rapid and inexpensive DNA sequencing has led to an explosion of data that must 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 a comparative genomics analysis, database and tools development, and mining of biological data for genes of interest. We annotated 26,087 oil palm genes integrated from two gene-prediction pipelines, Fgenesh++ and Seqping. As case studies, we conducted comprehensive investigations on intronless, resistance and fatty acid biosynthesis genes, and demonstrated that the current gene prediction set is of high quality. 3,672 intronless genes were identified in the oil palm genome, an important resource for evolutionary study. Further scrutiny of the oil palm genes revealed 210 candidate resistance genes involved in pathogen defense. Fatty acids have diverse applications ranging from food to industrial feedstock, and we identified 42 key genes involved in fatty-acid biosynthesis in oil palm mesocarp and kernel. These results provide an important resource for studies on plant genomes and a theoretical foundation for marker-assisted breeding of oil palm and related crops.

1. Introduction

Oil palm belongs to the genus *Elaeis* of the family Arecaceae. The genus has two species - *E. guineensis* (African oil palm) and *E. oleifera* (American oil palm). *E. guineensis* has three fruit forms that 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 tonnes, of which the Malaysian share was 19.50 million tonnes[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* form. *E. oleifera,* is little planted because of its low yield (only 10 - 20% of *guineensis*). However, it is more disease-resistant and planted in areas where *guineensis* is well-nigh impossible, e.g., Central-Southern America. Even then, it is mainly planted as a backcross to *guineensis* (interspecific hybrid) to raise its yield. Nevertheless, it has economically valuable traits which plant breeders drool over 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™ 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 allowed for 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.

de novo gene prediction was first established in the 1990s from modelling. A few years later, the Genscan[9] software predicted multiple and partial genes on both strands, with improved accuracy. Then, a host of programmes just about exploded to navigate the complexity of various genomes. Combining multiple predictors led to the development of automated pipelines to integrate the various experimental evidences.[10] Currently, such pipelines are primarily used for annotating whole genomes. A major limitation of many of the current approaches is their relatively poor performance in organisms with atypical distribution of nucleotides.[11–14] 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[15] pipeline developed at MPOB, and the Fgenesh++[16] pipeline at Softberry. Parameters of the framework were optimized to reliably identify the genes. 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. We identified a high confidence set of 26,087 oil palm genes. This paper presents a bioinformatics analysis of the oil palm genome, including comparative genomics analysis, database and tools development, and characterization of the genes of interest.

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. The *E. guineensis* mRNA dataset is a compilation of published transcriptomic sequences from Bourgis et al.,[17] Tranbarger et al.,[18] Shearman et al.,[19,20] 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 expressed sequence tags (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)[16,21] is an automatic gene prediction pipeline, based on Fgenesh, a HMM-based *ab initio* gene prediction program.[22] We used the P5-build 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) to known plant proteins from the NCBI non-redundant (NR) database. Based on this subset, we computed gene-finding parameters specific to oil palm, and executed the Fgenesh++ pipeline to annotate the genes in our genomic scaffolds. The Fgenesh++ pipeline takes into account all the available supporting data, such as known transcripts and homologous protein sequences. To predict genes, we compiled a subset of NR plant transcripts mapped to genomic sequences providing a set of potential splice sites to the Fgenesh++pipeline. Also, the pipeline (by its *Prot_map* module) aligned all proteins from a compiled subset of NR database with plant proteins to the genomic contigs and found highly homologous proteins in gene identification. All predicted amino acid sequences were compared to protein sequences in the NR database, using BLAST (E-value cutoff: 1E-10). BLAST analysis of the predicted sequences was also carried out against the *E. guineensis* mRNA dataset, using an identify cutoff of >90%.

2.3 Seqping Gene Prediction

Seqping[15], a customized pipeline based on MAKER2,[23] 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 selected was 500 - 5000 nt, to minimize potential prediction errors. Using BLASTX[24] 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[25], and subsequently filtered using the

TIGR plant repeat database,[26] GIRI Repbase[27] and Gypsy Database[28] to remove ORFs similar to retroelements. The resultant ORFs were used as the training set to develop the HMMs for GlimmerHMM,[29,30] AUGUSTUS[31] and SNAP,[32] which were subsequently used for gene predictions. Seqping used MAKER2[23] to combine the predictions from the three models. The predicted sequences were compared to the RefSeq[33] protein sequences and *E. guineensis* mRNA dataset via BLAST analysis.

2.4 Integration of Fgenesh++ and Seqping Gene Predictions

To increase the annotation accuracy, genes predicted by the Seqping and Fgenesh++ pipelines were combined in a unified gene prediction pipeline. ORF predictions with <300 nucleotides were excluded. Predicted genes from both pipelines in the same strand were considered overlapping if the overlap length was above the threshold fraction of the shorter gene length. A co-located group of genes was considered to constitute a locus if every gene in the group overlaps at least one other member of the same group (a single linkage approach). Different overlap thresholds, from 60% to 95% in 5% increments, were tested to determine the best threshold for further analysis, simultaneously maximizing the annotation accuracy and minimizing the number of single-isoform loci.

Gene functions were determined using PFAM-A[34,35] (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.[36,37] To find the representative isoform and function for each locus, we took the lowest E-value isoform in each locus and the function of its RefSeq match. We excluded hits with E-values >1E-10, keeping only high-quality loci and the corresponding isoforms. The ORF 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 match *O. sativa* sequences from the MSU rice database[38] at a cutoff E-value of 1E-10.

2.5 Intronless Genes

Intronless genes (IG) were identified through a set of 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),[39] *O. sativa* (MSU 6.0),[38] *S. bicolor* (Phytozome 6.0), *Z. mays* (Phytozome) and *Volvox carteri* (Phytozome 8.0).[40] Lists of non-redundant IG from all six genomes were obtained, and the oil palm IG compared with those in the 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.[41] 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[42] 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 a certain class. Our selection was validated based on the published "resistance" gene motifs[43–47] and each class validated via multiple sequence alignment and phylogenetic tree, using the ClustalW[48] and MEGA6[49] programs, respectively. The same procedure was used to identify R genes in the five other genomes for comparison.

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++,[16] MAKER-P,[77] Gramene,[78] and Ensembl.[79] 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.[80–82] The first oil palm genome[5] was published in 2013 with assembled sequences representing ~83% of the 1.8Gb-long genome. Gene models were predicted by integrating two pipelines: Fgenesh++ and Seqping[15] (MPOB's in-house tool).

Previous studies of five *ab initio* pipelines (Fgenesh++, GeneMark.hmm, GENSCAN, GlimmerR and Grail) to evaluate the precision in predicting maize genes showed that Fgenesh++ produced the most accurate annotations.[21] Fgenesh++ is a common tool for eukaryotic genome annotation, due to its superior ability to predict gene structure.[83–86] In the oil palm genome, Fgenesh++ predicted 117,832 whole and partial-length gene models at least 500 nt long. A total 23,028 Fgenesh++ gene models had significant similarities to the *E. guineensis* mRNA dataset and RefSeq proteins (Table 1).

To improve the coverage and accuracy of the predicted gene models, and to minimize prediction bias, Seqping, which is based on the MAKER2 pipeline[23], was used. The pipeline was customized to achieve high accuracy predictions for the oil palm genome. It identified 7,747 putative full-length CDS from the transcriptome data used to train the models GlimmerHMM,[29,30] AUGUSTUS,[31] and SNAP.[32] 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).

The gene models were combined to identify the number of independent genomic loci. At an 85% overlap threshold, we obtained 31,413 combined loci (Table 2 and Fig. 1). Next, we chose a subset of these loci that contained PFAM domains and RefSeq annotations, and got 26,087 high quality genomic loci. Of them, 9,943 contained one ORF, 12,163 two, and 3,981 three or more. For each locus, the ORF with the best match to plant proteins from the

RefSeq database was selected as its *best representative* CDS. The analysis of gene structures showed that 14% were intronless and 16% contained only two exons. We identified 395 complex genes with 21 to 57 exons each. The distributions of exons per gene and mRNA and CDS lengths are shown in Figure 2A and Figure 2B, respectively. The highest number of genes was at the 501-1000 nt interval, with 34.40% of mRNA and 30.62% of ORFs. The longest ORF and mRNA were 14,505 and 14,685 nt, respectively. Benchmarking Universal Single-Copy Orthologs (BUSCO)[87] analysis of the representative gene models showed that 90.44% of the 429 eukaryotic BUSCO profiles and 94.04% of the 956 *plantae* BUSCO profiles were available, thus quantifying the completeness of the oil palm genome annotation.

3.2 Fraction of cytosines and guanines in third position of codon, GC_3

GC_3 is defined as $\frac{C_3+G_3}{(L/3)}$, 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.[88] The frequency of guanine and cytosine contents in the third codon position is an important characteristic of a genome.[88] Two types of GC_3 distribution have been described - unimodal and bimodal.[88–90] Genes with high and low GC_3 peaks have distinct functional properties:[90] 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 are variously biased to different classes of genes, with GC_3 rich genes reportedly especially hard to predict.[91]

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% (2,608 ORFs) as GC_3-rich ($GC_3 \geq 0.7535$), 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 $CG_3^{skew} = \frac{C_3-G_3}{C_3+G_3}$). The increase in CG_3^{skew} from 5' to 3' has been linked to transcriptional efficiency and methylation status.[88,90,92] Figures 3C and D show a positional change in nucleotide composition. The 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. In addition, the plot of CG_3 skew increases in the proportion of cytosine in the GC_3-rich group, consistent with the hypothesis of transcriptional optimization of GC_3-rich genes. Despite the small GC_3-rich and GC_3-poor datasets, we observed pronounced peaks of nucleotide frequencies near the predicted start of translation, as also found in other well-annotated genomes.[88]

The genome signature, defined as $\rho_{CG} = \frac{f_{CG}}{f_C f_G}$, where f_x is the frequency of (di)nucleotide, x, shows the distribution of methylation targets. Similar to grasses, and other previously analyzed plant and animal species,[88,90] the 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. 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 Supplementary Table S1). The depletion of CpGs in GC_3-poor genes is consistent with their broad, constitutive expression.[88] Next, we calculated the distribution of all nucleotides in the 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 considered). Supplementary Table S2-S5 shows the 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 (-18.1% vs. 29.4%). This is consistent with the prior observation[90] that GC_3-poor genes are more methylated 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 (Supplementary 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.[88–90] It is important to note that GC_3-rich genes in plants tend to be intronless.[88]

3.3 Intronless Genes (IG)

Intronless genes (IG) are common in single-celled eukaryotes, but are only a small percentage of the genes in metazoans.[93,94] Across multi-cellular eukaryotes, IG are frequently tissue- or stress-specific, GC_3-rich and their promoters have a canonical TATA-box.[88,90,93] Among the 26,087 representative gene models, 3,672 (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 other species. Intronless CDS are, on average, shorter than multi-exonic CDS: 924 ±19 nt *vs.* 1289±12 nt long.

The distribution of IG in the whole genome is different for various functional groups.[88,94] For example, in the oil palm genome, 29% of the cell-signaling genes are intronless, compared to just 1% of all tropism-related genes (Supplementary 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.[94] Intronless and GC_3-rich genes are considered evolutionarily recent[88] and lineage-specific,[93] potentially appearing as a result of retrotransposon activity.[94,95] It is reported that 8 - 17% of the genes in most animals are IG - ~10% in mice and humans[93] and 3 - 5% in teleost fish. Plants have proportionately more IG than animals: 20% in *O. sativa*, 22% in *A. thaliana*,[96] 22% in *S. bicolor*, 37% in *Z. mays*, 28% in foxtail millet, 26% in switchgrass and 24% in purple false brome.[97] 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. (Supplementary Table S8). To establish a reference point, we calculated the fraction of IG in the green algae, *V. carteri*, and found 15.8%. The high 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.[88]

Using BLASTP, we found 555 IG (15% of oil palm IG) conserved across all the three domains of life: Archaea, Bacteria, and Eukaryotes (Fig. 5). These genes are likely essential for their survival.[98] A total of 738 oil palm IG were only homologous to bacterial genes, while 40 IG were only shared with Archaea. This is consistent with the extreme growth conditions of extant Archaea from the typical conditions of Bacteria; meaning that there are

fewer opportunities for horizontal gene transfer from Archaea than from Bacteria to the oil palm genome. Considering the three kingdoms in eukaryotes, we observed 1,384 oil palm IG shared among them. A significant portion of the oil palm IG (1,864) was only homologous to Viridiplantae (green plants). These proteins may have evolved, or been regained, only in plants, while other organisms have lost their ancestral genes during evolution.[96]

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 2,433 (66.26%) 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.[99,100] 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.

Conclusions

We developed an integrated gene prediction pipeline, enabling annotation of the genome of the African oil palm and present a set of 26,087 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 also found that approximately one-seventh of the oil palm genes intronless. We detected a similar number of R genes in the oil palm compared to other sequenced plant genomes. Lipid-, especially fatty acid (FA)-related genes, are of interest in oil palm where, in addition to their roles in specifying oil yield and quality, they 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 studies on 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.

Acknowledgements

We thank Orion Genomics LLC for their assistance and advice in data analysis. Special thanks to the Breeding and Tissue Culture Unit for the supply of palms and RNA extraction for sequencing. Last, but not least, we extend our appreciation to the Director General MPOB, Dr. Ahmad Kushairi Din, for his support and encouragement throughout the project. Dr. Tatarinova was supported by NSF Awards #1456634 and #1622840.

References

1. Barcelos, E., Rios, S. de A., Cunha, R. N. V., et al. 2015, Oil palm natural diversity and the potential for yield improvement. *Front. Plant Sci.*, **6**, 190.

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

3. Jouannic, S., Argout, X., Lechauve, F., et al. 2005, Analysis of expressed sequence tags from oil palm (Elaeis guineensis). *FEBS Lett.*, **579**, 2709–14.

4. Low, E. T. L., Rosli, R., Jayanthi, N., et al. 2014, Analyses of hypomethylated oil palm gene space. *PLoS One*, **9**.

5. Singh, R., Ong-Abdullah, M., Low, E.-T. L., et al. 2013, Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature*, **500**, 335–9.

6. Singh, R., Low, E.-T. L., Ooi, L. C.-L., et al. 2013, The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK. *Nature*, **500**, 340–4.

7. Singh, R., Low, E.-T. L., Ooi, L. C.-L., et al. 2014, The oil palm VIRESCENS gene controls fruit colour and encodes a R2R3-MYB. *Nat. Commun.*, **5**, 4106.

8. Ong-Abdullah, M., Ordway, J. M., Jiang, N., et al. 2015, Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature*, **525**, 533–7.

9. Burge, C., and Karlin, S. 1997, Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

10. Brent, M. R. 2005, Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Res.*, pp. 1777–86.

11. Berendzen, K. W., Stüber, K., Harter, K., and Wanke, D. 2006, 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**, 522.

12. Pritsker, M., Liu, Y. C., Beer, M. A., and Tavazoie, S. 2004, Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res.*, **14**, 99–108.

13. Troukhan, M., Tatarinova, T., Bouck, J., Flavell, R. B., and Alexandrov, N. N. 2009, Genome-wide discovery of cis-elements in promoter sequences using gene expression. *OMICS*, **13**, 139–51.

14. Triska, M., Grocutt, D., Southern, J., Murphy, D. J., and Tatarinova, T. 2013, CisExpress: Motif detection in DNA sequences. *Bioinformatics*, **29**, 2203–5.

15. Chan, K.-L., Rosli, R., Tatarinova, T., Hogan, M., Firdaus-Raih, M., and Low, E.-T. L. 2016, Seqping: Gene Prediction Pipeline for Plant Genomes using Self-Trained Gene Models and Transcriptomic Data. *bioRxiv*, **2016**, 038018.

16. Solovyev, V., Kosarev, P., Seledsov, I., and Vorobyev, D. 2006, Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.*, **7 Suppl 1**, S10.1–12.

17. Bourgis, F., Kilaru, A., Cao, X., et al. 2011, 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**, 12527–32.

18. Tranbarger, T. J., Dussert, S., Joët, T., et al. 2011, Regulatory mechanisms underlying oil palm fruit mesocarp maturation, ripening, and functional specialization in lipid and carotenoid metabolism. *Plant Physiol.*, **156**, 564–84.

19. Shearman, J. R., Jantasuriyarat, C., Sangsrakru, D., et al. 2013, Transcriptome analysis of normal and mantled developing oil palm flower and fruit. *Genomics*, **101**, 306–12.

20. Shearman, J. R., Jantasuriyarat, C., Sangsrakru, D., et al. 2013, Transcriptome Assembly and Expression Data from Normal and Mantled Oil Palm Fruit. *Dataset Pap. Biol.*, **2013**, 1–7.

21. Yao, H., Guo, L., Fu, Y., et al. 2005, Evaluation of five ab initio gene prediction programs for the discovery of maize genes. *Plant Mol. Biol.*, **57**, 445–60.

22. Salamov, A. A., and Solovyev, V. V. 2000, Ab initio gene finding in Drosophila genomic DNA. *Genome Res.*, **10**, 516–22.

23. Holt, C., and Yandell, M. 2011, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, p. 491.

24. Altschul, S. F., Madden, T. L., Schäffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, pp. 3389–402.

25. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–2.

26. Ouyang, S., and Buell, C. R. 2004, The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.*, **32**, D360–3.

27. Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–7.

28. Llorens, C., Futami, R., Covelli, L., et al. 2011, The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.*, **39**, D70–4.

29. Majoros, W. H., Pertea, M., and Salzberg, S. L. 2004, TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–9.

30. Allen, J. E., Majoros, W. H., Pertea, M., and Salzberg, S. L. 2006, JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. *Genome Biol.*, **7 Suppl 1**, S9.1–13.

31. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. 2008, Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**, 637–44.

32. Korf, I. 2004, Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

33. Pruitt, K. D., Tatusova, T., and Maglott, D. R. 2007, NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**.

34. Sonnhammer, E. L. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. 1998, Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–2.

35. Finn, R. D., Bateman, A., Clements, J., et al. 2014, Pfam: The protein families database. *Nucleic Acids Res.*

36. Johnson, L. S., Eddy, S. R., and Portugaly, E. 2010, Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**, 431.

37. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. 2013, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.*, **41**.

38. Kawahara, Y., de la Bastide, M., Hamilton, J. P., et al. 2013, Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice (N. Y).*, **6**, 4.

39. Berardini, T. Z., Reiser, L., Li, D., et al. 2015, The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis*, **53**, 474–85.

40. Goodstein, D. M., Shu, S., Howson, R., et al. 2012, Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, 1178–86.

41. Sanseverino, W., Hermoso, A., D'Alessandro, R., et al. 2013, PRGdb 2.0: Towards a community-based database model for the analysis of R-genes in plants. *Nucleic Acids Res.*, **41**.

42. Krogh, a, Larsson, B., von Heijne, G., and Sonnhammer, E. L. 2001, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–80.

43. Barbosa-da-Silva, A., Wanderley-Nogueira, A. C., Silva, R. R. M., Berlarmino, L. C., Soares-Cavalcanti, N. M., and Benko-Iseppon, A. M. 2005, In silico of resistance (R) genes in Eucalyptus transcriptome. *Genet. Mol. Biol.*, **28**, 562–74.

44. Martin, G. B., Bogdanove, A. J., and Sessa, G. 2003, Understanding the functions of plant disease resistance proteins. *Annu. Rev. Plant Biol.*, **54**, 23–61.

45. Peraza-Echeverria, S., James-Kay, A., Canto-Canché, B., and Castillo-Castro, E. 2007, Structural and phylogenetic analysis of Pto-type disease resistance gene candidates in banana. *Mol. Genet. Genomics*, **278**, 443–53.

46. Song, W. Y., Wang, G. L., Chen, L. L., et al. 1995, A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. *Science*, pp. 1804–6.

47. Yun, C. 1999, Classification and function of plant disease resistance genes. *Plant Pathol. J.*, **15**, 105–11.

48. Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994, CLUSTAL W: improving the sensitivity of progressive multiple sequence. *Nucleic Acids Res.*, **22**, 4673–80.

49. Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. 2013, MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.*, **30**, 2725–9.

50. Kanehisa, M., and Goto, S. 2000, Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

51. Okuley, J., Lightner, J., Feldmann, K., Yadav, N., Lark, E., and Browse, J. 1994, Arabidopsis FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *Plant Cell …*, **6**, 147–58.

52. Lin, X., Kaul, S., Rounsley, S., et al. 1999, Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. *Nature*, **402**, 761–8.

53. Tilton, G. B., Shockey, J. M., and Browse, J. 2004, Biochemical and Molecular Characterization of ACH2, an Acyl-CoA Thioesterase from Arabidopsis thaliana. *J. Biol. Chem.*, **279**, 7487–94.

54. Jha, S. S., Jha, J. K., Chattopadhyaya, B., Basu, A., Sen, S. K., and Maiti, M. K. 2010, Cloning and characterization of cDNAs encoding for long-chain saturated acyl-ACP thioesterases from the developing seeds of Brassica juncea. *Plant Physiol. Biochem.*, **48**, 476–80.

55. Quevillon, E., Silventoinen, V., Pillai, S., et al. 2005, InterProScan: Protein domains identifier. *Nucleic Acids Res.*, **33**, 116–20.

56. Sonnhammer, E. L. L., Eddy, S. R., and Durbin, R. 1997, Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins Struct. Funct. Genet.*, **28**, 405–20.

57. Benson, B. K., Meades, G., Grove, A., and Waldrop, G. L. 2008, DNA inhibits catalysis by the carboxyltransferase subunit of acetyl-CoA carboxylase: implications for active site communication. *Protein Sci.*, **17**, 34–42.

58. Barber, M. C., Price, N. T., and Travers, M. T. 2005, Structure and regulation of acetyl-CoA carboxylase genes of

metazoa. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids*, pp. 1–28.

59. Waldrop, G. L., Rayment, I., and Holden, H. M. 1994, Three-dimensional structure of the biotin carboxylase subunit of acetyl-CoA carboxylase. *Biochemistry*, **33**, 10249–56.

60. Li, M.-J., Li, A.-Q., Xia, H., et al. 2009, Cloning and sequence analysis of putative type II fatty acid synthase genes from Arachis hypogaea L. *J. Biosci.*, **34**, 227–38.

61. Haralampidis, K., and Milioni, D. 1998, Temporal and transient expression of stearoyl-ACP carrier protein desaturase gene during olive fruit development. *J. ...*, **49**, 1661–9.

62. Shanklin, J., Whittle, E., and Fox, B. G. 1994, 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**, 12787–94.

63. Yuan, L., Nelson, B. A., and Caryl, G. 1996, The catalytic cysteine and histidine in the plant acyl-acyl carrier protein thioesterases. *J. Biol. Chem.*, **271**, 3417–9.

64. Brenner, S. 1988, The molecular evolution of genes and proteins: a tale of two serines. *Nature*, **334**, 528–30.

65. Rozwarski, D. A., Vilchèze, C., Sugantino, M., Bittman, R., and Sacchettini, J. C. 1999, Crystal structure of the Mycobacterium tuberculosis enoyl-ACP reductase, InhA, in complex with NAD+ and a C16 fatty acyl substrate. *J. Biol. Chem.*, **274**, 15582–9.

66. Smith, S. 1994, The animal fatty acid synthase: one gene, one polypeptide, seven enzymes. *FASEB J.*, **8**, 1248–59.

67. Helmkamp, G. M., and Bloch, K. 1969, beta-Hydroxydecanoyl thioester dehydrase: studies on molecular structure and active side. *J Biol Chem*, **244**, 6014–23.

68. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. 2013, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

69. Trapnell, C., Roberts, A., Goff, L., et al. 2012, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–78.

70. Li, L., Stoeckert, C. J., and Roos, D. S. 2003, OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–89.

71. Edgar, R. 2004, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–7.

72. Katoh, K., and Standley, D. M. 2013, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–80.

73. Mitchell, A., Chang, H., Daugherty, L., et al. 2015, The InterPro protein families database : the classification resource after 15 years. *Nucleic Acids Res.*, **43**, 213–21.

74. Sigrist, C. J. A., Castro, E. De, Cerutti, L., et al. 2013, New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, 344–7.

75. Marchler-bauer, A., Derbyshire, M. K., Gonzales, N. R., et al. 2015, CDD : NCBI's conserved domain database. *Nucleic Acids Res.*, **43**, 222–6.

76. Kuraku, S., Zmasek, C. M., Nishimura, O., and Katoh, K. 2013, aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. *Nucleic Acids Res.*, **41**, 22–8.

77. Campbell, M. S., Law, M., Holt, C., et al. 2014, MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.*, **164**, 513–24.

78. Liang, C., Mao, L., Ware, D., and Stein, L. 2009, Evidence-based gene predictions in plant genomes. *Genome Res.*, **19**, 1912–23.

79. Curwen, V., Eyras, E., Andrews, T. D., et al. 2004, The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–50.

80. Ouyang, S., Zhu, W., Hamilton, J., et al. 2007, The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res.*, **35**.

81. Zhu, W., and Buell, C. R. 2007, Improvement of whole-genome annotation of cereals through comparative analyses. *Genome Res.*, **17**, 299–310.

82. Swarbreck, D., Wilks, C., Lamesch, P., et al. 2008, The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res.*, **36**.

83. Ge, Y., Wang, Y., Liu, Y., et al. 2016, Comparative genomic and transcriptomic analyses of the Fuzhuan brick tea-fermentation fungus Aspergillus cristatus. *BMC Genomics*, **17**, 428.

84. Calabrese, S., Pérez-Tienda, J., Ellerbeck, M., et al. 2016, GintAMT3 – a Low-Affinity Ammonium Transporter of the Arbuscular Mycorrhizal Rhizophagus irregularis. *Front. Plant Sci.*, **7**, 1–14.

85. Zhong, Z., Norvienyeku, J., Chen, M., et al. 2016, Directional Selection from Host Plants Is a Major Force Driving Host Specificity in Magnaporthe Species. *Sci. Rep.*, **6**, 25591.

86. Ghelder, C. Van, and Esmenjaud, D. 2016, TNL genes in peach : insights into the post- LRR domain. *BMC Genomics*, 1–16.

87. Sima, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V, and Zdobnov, E. M. 2015, BUSCO : assessing

genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **2015**, 1–3.

88. Tatarinova, T. V, Alexandrov, N. N., Bouck, J. B., and Feldmann, K. A. 2010, GC3 biology in corn, rice, sorghum and other grasses. *BMC Genomics*, **11**, 308.

89. Alexandrov, N. N., Brover, V. V., Freidin, S., et al. 2009, Insights into corn genes derived from large-scale cDNA sequencing. *Plant Mol. Biol.*, **69**, 179–94.

90. Tatarinova, T., Elhaik, E., and Pellegrini, M. 2013, Cross-species analysis of genic GC3 content and DNA methylation patterns. *Genome Biol. Evol.*, **5**, 1443–56.

91. Souvorov, A., Tatusova, T., Zaslasky, L., and Smith-White, B. 2011, Glycine max and Zea mays Genome Annotation with Gnomon*ISMB/ECCB*.

92. Ahmad, T., Sablok, G., Tatarinova, T. V., Xu, Q., Deng, X.-X. X., and Guo, W.-W. W. 2013, Evaluation of codon biology in citrus and Poncirus trifoliata based on genomic features and frame corrected expressed sequence tags. *DNA Res.*, **20**, 135–50.

93. Agarwal, S. M., and Srivastava, P. K. 2009, Human intronless disease associated genes are slowly evolving. *BMB Rep.*, **42**, 356–60.

94. Grzybowska, E. a. 2012, Human intronless genes: Functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochem. Biophys. Res. Commun.*, **424**, 1–6.

95. Ferguson, A. a., and Jiang, N. 2011, Pack-MULEs: recycling and reshaping genes through GC-biased acquisition. *Mob. Genet. Elements*, **1**, 135–8.

96. Jain, M., Khurana, P., Tyagi, A. K., and Khurana, J. P. 2008, Genome-wide analysis of intronless genes in rice and Arabidopsis. *Funct. Integr. Genomics*, **8**, 69–78.

97. Yan, H., Jiang, C., Li, X., et al. 2014, PIGD: a database for intronless genes in the Poaceae. *BMC Genomics*, **15**, 832.

98. Yan, H., Zhang, W., Lin, Y., et al. 2014, Different evolutionary patterns among intronless genes in maize genome. *Biochem. Biophys. Res. Commun.*, **449**, 146–50.

99. Kordis, D. 2011, Extensive intron gain in the ancestor of placental mammals. *Biol Direct*, **6**, 59.

100. Kordis, D., and Kokosar, J. 2012, What can domesticated genes tell us about the intron gain in mammals? *Int. J. Evol. Biol.*, **2012**, 7.

101. Freeman, B. C., and Beattie, G. A. 2008, An Overview of Plant Defenses against Pathogens and Herbivores. *Plant Heal. Instr.*

102. Jones, J. D. G., and Dangl, J. L. 2006, The plant immune system. *Nature*, **444**, 323–9.

103. Katagiri, F., and Tsuda, K. 2010, Understanding the plant immune system. *Mol. Plant. Microbe. Interact.*, **23**, 1531–6.

104. Meyers, B. C., Dickerman, A. W., Michelmore, R. W., Sivaramakrishnan, S., Sobral, B. W., and Young, N. D. 1999, Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide binding superfamily. *Plant J.*, **20**, 317–32.

105. Ameline-Torregrosa, C., Wang, B.-B., O'Bleness, M. S., et al. 2007, Identification and characterization of NBS-LRR genes in the model plant Medicago truncatula. *Plant Physiol*, **146**, pp.107.104588.

106. Tarr, D. E. K., and Alexander, H. M. 2009, TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. *BMC Res. Notes*, **2**, 197.

107. Pan, Q., Wendel, J., and Fluhr, R. 2000, Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *J. Mol. Evol.*, **50**, 203–13.

108. Jones, D. A., and Jones, J. D. G. 1997, The Role of Leucine-Rich Repeat Proteins in Plant Defences. *Adv. Bot. Res.*, **24**, 89–167.

109. Staskawicz, B. J., Ausubel, F. M., Baker, B. J., Ellis, J. G., and Jones, J. D. 1995, Molecular genetics of plant disease resistance. *Science (80-.).*, **268**, 661–7.

110. Marone, D., Russo, M. a., Laidò, G., De Leonardis, A. M., and Mastrangelo, A. M. 2013, Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: Active guardians in host defense responses. *Int. J. Mol. Sci.*, pp. 7302–26.

111. Sessa, G., D'Ascenzo, M., and Martin, G. B. 2000, Thr38 and Ser198 are Pto autophosphorylation sites required for the AvrPto-Pto-mediated hypersensitive response. *EMBO J.*, **19**, 2257–69.

112. Shan, L., Thara, V. K., Martin, G. B., Zhou, J. M., and Tang, X. 2000, The pseudomonas AvrPto protein is differentially recognized by tomato and tobacco and is localized to the plant plasma membrane. *Plant Cell*, **12**, 2323–38.

113. Frederick, R. D., Thilmony, R. L., Sessa, G., and Martin, G. B. 1998, 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**, 241–5.

114. Sekhwal, M., Li, P., Lam, I., Wang, X., Cloutier, S., and You, F. 2015, Disease Resistance Gene Analogs (RGAs) in Plants. *Int. J. Mol. Sci.*, **16**, 19248–90.

115. Büschges, R., Hollricher, K., Panstruga, R., et al. 1997, The barley Mlo gene: a novel control element of plant pathogen resistance. *Cell*, **88**, 695–705.

116. Sambanthamurthi, R., Sundram, K., and Tan, Y. 2000, Chemistry and biochemistry of palm oil. *Prog. Lipid Res.*, **39**, 507–58.

117. Noh, A., Rajanaidu, N., Kushairi, A., et al. 2002, Variability in fatty acid composition, iodine value and carotene content in the MPOB oil palm germplasm collection from Angola. *J. 0il Palm Res.*, **14**, 18–23.

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

119. Sambanthamurthi, R., and Ohlrogge, J. B. 1996, Acetyl-coA carboxylase activity in the oil palm In: Williams, J. P., Khan, M. U., and Lem, N. W., (eds.), *Physiology, biochemistry and molecular biology of plant lipids*. Dordrecht: Kluwer Academic Publishers, p. 26.

120. Omar, W. S. W., Willis, L. B., Rha, C., et al. 2008, Isolation and utilization of acetyl-coa carboxylase from oil palm (elaeis guineensis) mesocarp. *J. Oil Palm Res.*, **2**, 97–107.

121. Umi Salamah, R., and Sambanthamurthi, R. 1996, β-keto acyl ACP synthase II I oil palm (Elaeis guineensis) mesocarp In: Williams, J. P., Khan, M. U., and Lem, N. W., (eds.), *Physiology, Biochemistry and Molecular Biology of Plant Lipids*. Kluwer Academic Publishers, pp. 69–71.

122. Sambanthamurthi, R., Abrizah, O., and Umi Salamah, R. 1999, Biochemical factors that control oil composition in the oil palm. *J. Oil Palm Res.*, 23–33.

123. Umi Salamah, R., Sambanthamurthi, R., Omar, A. R., et al. 2012, 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**, 1480–91.

124. Zhang, Y. M., Wang, C. C., Hu, H. H., and Yang, L. 2011, Cloning and expression of three fatty acid desaturase genes from cold-sensitive lima bean (Phaseolus lunatus L.). *Biotechnol. Lett.*, **33**, 395–401.

125. Parveez, G. K. A., Rasid, O. A., and Sambanthamurthi, R. 2011, Genetic engineering of oil palm In: Mohd Basri, W., Choo, Y. M., and Chan, K. W., (eds.), *Further advances in oil palm research (2000-2010)*, pp. 141–201.

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

127. Sidorov, R. a., and Tsydendambaev, V. D. 2013, Biosynthesis of fatty oils in higher plants. *Russ. J. Plant Physiol.*, **61**, 1–18.

128. Lindqvist, Y., Huang, W., Schneider, G., and Shanklin, J. 1996, Crystal structure of delta9 stearoyl-acyl carrier protein desaturase from castor seed and its relationship to other di-iron proteins. *EMBO J.*, **15**, 4081–92.

129. Troncoso-Ponce, M. a., Kilaru, A., Cao, X., et al. 2011, Comparative deep transcriptional profiling of four

developing oilseeds. *Plant J.*, **68**, 1014–27.

130. Allaby, R. G., Peterson, G. W., Merriwether, D. A., and Fu, Y.-B. 2005, Evidence of the domestication history of flax (Linum usitatissimum L.) from genetic diversity of the sad2 locus. *Theor. Appl. Genet.*, **112**, 58–65.

131. Thambugala, D., Duguid, S., Loewen, E., et al. 2013, Genetic variation of six desaturase genes in flax and their impact on fatty acid composition. *Theor Appl Genet*, **126**, 2627–41.

132. Kachroo, A., Shanklin, J., Whittle, E., Lapchyk, L., Hildebrand, D., and Kachroo, P. 2007, The Arabidopsis stearoyl-acyl carrier protein-desaturase family and the contribution of leaf isoforms to oleic acid synthesis. *Plant Mol. Biol.*, **63**, 257–71.

133. Kachroo, P., Shanklin, J., Shah, J., Whittle, E. J., and Klessig, D. F. 2001, A fatty acid desaturase modulates the activation of defense signaling pathways in plants. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 9448–53.

134. Abrizah, O., Lazarus, C., Fraser, T., and Stobart, K. 2000, Cloning of a palmitoyl-acyl carrier protein thioesterase from oil palm. *Biochem. Soc. Trans.*, **28**, 619–22.

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

136. Voelker, T. A., Worrell, A. C., Anderson, L., et al. 1992, Fatty Acid Biosynthesis redirected to Medium Chains in Transgenic Oilseed Plants. *Science*, **257**, 72–4.

137. Davies, H. M., Anderson, L., Fan, C., and Hawkins, D. J. 1991, Developmental Induction, Purification, and further Characterization of 12:0-ACP Thioesterase from Immature Cotyledons of Umbellularia californica. *Arch. Biochem. Biophys.*, **290**, 37–45.

138. Pollard, M. R., Anderson, L., Fan, C., Hawkins, D. J., and Davies, H. M. 1991, A specific acyl-ACP thioesterase implicated in medium-chain fatty acid production in immature cotyledons of Umbellularia californica. *Arch. Biochem. Biophys.*, **284**, 306–12.

139. Patthy, L. 2003, Modular assembly of genes and the evolution of new functions. *Genetica*, **118**, 217–31.

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

141. Corley, R. H. V, and Tinker, P. B. 2003, *The Oil Palm* Fourth edi. Blackwell Science Ltd, Oxford.

142. Kim, H. J., Silva, J. E., Vu, H. S., Mockaitis, K., Nam, J. W., and Cahoon, E. B. 2015, Toward production of jet fuel functionality in oilseeds: Identification of FatB acyl-acyl carrier protein thioesterases and evaluation of combinatorial expression strategies in Camelina seeds. *J. Exp. Bot.*, pp. 4251–65.