Howard Hughes Medical Institute

Published as: *Nature*. 2013 August 15; 500(7462): 335–339.

Oil palm genome sequence reveals divergence of interfertile species in old and new worlds

Rajinder Singh[1,*], **Meilina Ong-Abdullah**[1,*], **Eng-Ti Leslie Low**[1,*], **Mohamad Arif Abdul Manaf**[1], **Rozana Rosli**[1], **Rajanaidu Nookiah**[1], **Leslie Cheng-Li Ooi**[1], **Siew–Eng Ooi**[1], **Kuang-Lim Chan**[1], **Mohd Amin Halim**[1], **Norazah Azizi**[1], **Jayanthi Nagappan**[1], **Blaire Bacher**[2], **Nathan Lakey**[2], **Steven W Smith**[2], **Dong He**[2], **Michael Hogan**[2], **Muhammad A Budiman**[2], **Ernest K Lee**[3], **Rob DeSalle**[3], **David Kudrna**[4], **Jose Louis Goicoechea**[4], **Rod Wing**[4], **Richard K Wilson**[5], **Robert S Fulton**[5], **Jared M Ordway**[2], **Robert A Martienssen**[6,7], and **Ravigadevi Sambanthamurthi**[1,7]

[1]Malaysian Palm Oil Board, 6, Persiaran Institusi Bandar Baru Bangi, 43000 Kajang Selangor, Malaysia

[2]Orion Genomics, 4041 Forest Park Avenue, St. Louis, MO 63108, USA

[3]Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA

[4]Arizona Genome Institute, University of Arizona, Tucson AZ 85705, USA

[5]The Genome Institute at Washington University, Washington University School of Medicine, Saint Louis, MO 63108, USA

[6]Howard Hughes Medical Institute-Gordon and Betty Moore Foundation, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

Abstract

Oil palm is the most productive oil-bearing crop. Planted on only 5% of the total vegetable oil acreage, palm oil accounts for 33% of vegetable oil, and 45% of edible oil worldwide, but

[7]For correspondence: martiens@cshl.edu, raviga@mpob.gov.my.
[*]These authors contributed equally to this work.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions. R.S., M.O.A., E-T.L.L and R.S.M. conceptualized the research programme. R.S., M.O.A., E-T.L.L., R.N., N.L., J.M.O, S.W.W, R.K.W., R.S.F., R.A.M. and R.S.M. designing experiments and coordinated the project. R.S., M.O.A., E-T.L.L., M.A.A.M., L.C.L.O., S-E.O., J.N., B.B., M.A.B., S.W.S., J.M.O. and R.S.M. conducted laboratory experiments and were involved in data analysis. E-T.L.L., R.R., K.L.C., M.A.H., N.A., S.W.S., D.H., M.H. performed bioinformatics analyses. E.K.L. and R.DeS. carried out Phylogenetic analysis. D.K., J.L.G. AND R.W. built BAC libraries and carried out BAC end sequencing. R.S., M.O.A., E-T.L.L., R.N., N.L., S.W.S., J.M.O., R.A.M. and R.S.M. participated in preparing and revising the manuscript. R.S.M. and R.A.M. supervised data generation and analysis.

The *E. guineensis* and *E. Oleifera* Bioprojects have been registered under PRJNA 192219 and PRJNA 183707. GenBank accession numbers will be provided at the copy editing stage.

R.S.F., R.K.W. and R.A.M. are consultants for Orion Genomics, LLC. R.K.W. serves on the Board of Directors for Orion Genomics, LLC. Readers are welcome to comment on the online version of this article at www.nature.com/nature.

increased cultivation competes with dwindling rainforest reserves. We report the 1.8 gigabase (Gb) genome sequence of the African oil palm *Elaeis guineensis*, the predominant source of worldwide oil production. 1.535 Gb of assembled sequence and transcriptome data from 30 tissue types were used to predict at least 34,802 genes, including oil biosynthesis genes and homologues of *WRINKLED1 (WRI1)*, and other transcriptional regulators[1], which are highly expressed in the kernel. We also report the draft sequence of the S. American oil palm *Elaeis oleifera*, which has the same number of chromosomes (2n=32) and produces fertile interspecific hybrids with *E. guineensis*[2], but appears to have diverged in the new world. Segmental duplications of chromosome arms define the palaeotetraploid origin of palm trees. The oil palm sequence enables the discovery of genes for important traits as well as somaclonal epigenetic alterations which restrict the use of clones in commercial plantings[3], and thus helps achieve sustainability for biofuels and edible oils, reducing the rainforest footprint of this tropical plantation crop.

The genus *Elaeis* (tribe *Cocoseae*) is in the family *Arecaceae*[4], one of the oldest families of flowering plants, with fossils from the Cretaceous[5]. The genus comprises two species, *E. guineensis* from West Africa[6] and *E. oleifera* from Central and South America. *E. guineensis* is higher yielding, but *E. oleifera* has higher unsaturated fatty acid content, lower height and resistance to disease[7]. Commercial cultivation of oil palm commenced on the West African coast in the early twentieth century [8]. In SE Asia, where it is one of the most important commercial crops, the first recorded oil palm was brought from Africa via Mauritius and Amsterdam in 1848[9], when four seedlings were planted as ornamentals in the Bogor Botanical Gardens in Java. Commercial cultivation also began in the early twentieth century and despite the long breeding cycle (10–12 years) and large land requirement for field trials[10], high yield breeding materials (up to 12t/ha/yr[9]) have been developed in less than 100 years. As such, the largely undomesticated oil palm is an ideal candidate for genomic-based tools including Expressed Sequence Tags (ESTs)[11–13] and transcriptome sequencing of the oil palm fruit during development, maturation and ripening[1,14] to harness the potential of this remarkably productive crop.

We sequenced the ~1.8 Gb *E. guineensis* genome (AVROS *pisifera* fruit-form) to high coverage with a combination of Roche/454 and Sanger BAC end sequencing (Supplementary Fig. 1–3, Supplementary Tables 1–5 and Methods). The combined total length of the assembly (P5 Build) is 1.535 Gb. Comparison of the P5 Build to genetic linkage maps (Supplementary Fig. 4 and Methods) resulted in 16 genetic scaffolds (one per chromosome) including 40,072 scaffolds with an N50 of 1.26 Mb (42% of the P5 scaffold assembly) (Supplementary Table 4) and 74% of RefSeq supported genes. For comparison, the genome of *E. oleifera* was also sequenced with a combination of fragment and linker libraries (Supplementary Table 2 and Methods). Thirty transcriptome libraries were sequenced and assembled producing 4,528–18,936 isotigs per library (Supplementary Fig. 5 and Supplementary Table 6). We sequenced 298,039 reads from methylation-filtered genomic libraries from Deli *dura* and *pisifera* genotypes of *E. guineensis*, and *E. oleifera* (Methods). Methylation-filtered libraries included 90% of the gene models and were enriched 5.6x for genes with “gene space” between 300 and 400 Mb, comparable to that of rice and maize[15].

G+C content (37%), is similar to other plant genomes, including date palm[16], but genes were conspicuous for having a much higher G+C content (50%). Gene finding algorithms (Methods) predicted 158,946 gene candidates, comprising 92 Mb of exonic gene space (5% of the 1.8 Gb genome sequence) (Supplementary Tables 4 and 5). Of these candidates, 34,802 were similar to known proteins at the peptide sequence level with 96% observed in transcriptome data (Methods). Of the remaining 124,144 candidates, 15,311 were identified in transcriptome data (Supplementary Table 4). Known retroelements (67,169) and other transposons (41,662) made up the remaining 108,833 candidates. Comparison to all repetitive element classes resulted in the identification of 775,703 independent genomic regions matching repetitive sequence elements, corresponding to 282 Mb of sequence (or 18% of the P5 Build), with 39% G+C content (Supplementary Table 4). Repeat content of the unmapped and unassembled contigs was far higher, as expected, and estimated to be approximately half of the 1.535 Gb P5 Build, or 57% of the 1.8 Gb *E. guineensis* genome.

The 16 EG-5 psuedochromosomes (Fig. 1), were numbered according to size and compared with previous mapping (Supplementary Table 7) and fluorescence *in situ* hybridization studies (FISH)[17]. Gene density (Fig. 1a) was unevenly distributed: five of the smallest six chromosomes had one gene-rich arm, and one repeat-rich arm, as previously noted by FISH[17]. Known repeat classes matching the TIGR grass repeat database and REPBASE were distributed in gene-poor, methylated regions (Fig. 1b, c), while simple di- and tri-nucleotide repeats[17] were mostly within genic regions (Fig. 1d). Potential centromeric regions were identified using an internally repetitive pericentromeric repeat[17] (Fig. 1e, purple), while highly conserved TTTAGGG telomeric repeat arrays were identified at the extreme ends of seven of the 32 chromosome arms (Fig. 1e, green). The most prominent 5S rRNA cluster (Fig. 1e, orange) was found on the largest chromosome, while the only telocentric chromosome was one of the two smallest, as previously described[17]. Two interstitial telomere repeat arrays on Chr2 and Chr14 were embedded within putative centromeric regions. Robertsonian fusions of telocentric chromosome ends may have given rise to these two chromosomes[18], and date palm has 18 chromosomes[16], consistent with this hypothesis. Typical of monocot genomes[19], the most abundant repetitive elements were *copia* (33%) and *gypsy* (8%) retroelements, as well as other LTR retrotransposons (6%). Interestingly, 47% of all repeats observed were previously uncharacterized, with 73% absent from *E. oleifera*, and 99% absent from *M. acuminata* (banana) (Supplementary Fig. 6 and Supplementary Methods). The distribution of repeats in methylation-filtered reads indicated that RIRE1 and other *copia* elements are especially heavily methylated.

Comparison of *E. guineensis* chromosomes to each other revealed that oriented homeologous duplicated sequences (segmental duplications) are strikingly abundant (Fig. 1 and Supplementary Fig. 7). Analysis of conserved gene order revealed that the duplications were retained in *E. oleifera*, so that segmental duplications pre-dated the divergence of the African and S. American oil palm (Supplementary Fig. 8a). Given that most of the genome is represented by segmental duplications, and not triplications, we conclude that oil palm is a palaeotetraploid, in line with speculation based on cytogenetics and RFLP mapping[17,20]. These duplications do not span the putative pericentromeric regions (Fig. 1), indicating that most centromeres arose after polyploidization, consistent with extensive chromosome

restructuring. Homologues of 94.4%, 83.5% and 80.2% of the genes from *P. dactylifera*, *M. acuminata* and Arabidopsis respectively, were found in *E. guineensis* (Fig. 2a, Supplementary Table 8, Methods). Each *E. guineensis* duplication matched unique scaffolds in date palm (Supplementary Fig. 8b), indicating that date palms have most of the segmental duplications found in oil palm. Polyploidization has been inferred by chromosome counts in only a limited number of *Arecaceae*[21], and a likely scenario is that the progenitor of both palms arose as a polyploid. We performed a similar analysis of the banana genome[22] and found extensive synteny between each oil palm chromosome and several chromosomes from banana (Supplementary Fig. 9), confirming that duplication events in the *M. acuminata* genome occurred after the *Musa/Elaeis* split, as previously proposed[22].

The 34,802 sequence similarity gene predictions (Supplementary Table 4) were annotated for Gene Ontology (GO) terms (Fig. 2b, c), with a focus on oil biosynthesis (Methods). Oil synthesis in the kernel commences at 11–12 weeks after anthesis (WAA) and is complete by 15–16 WAA at which stage mesocarp oil synthesis starts, reaching a peak at 20WAA[23]. In plants, *de novo* fatty acid synthesis (FAS) is compartmentalized in plastids while triacylglycerol (TAG) synthesis occurs in the cytoplasm. Although the oil palm accumulates markedly higher TAG than the date palm[1], the number of genes involved in TAG biosynthesis is strikingly similar in both palms. In contrast, FAS genes have higher representation in the oil palm genome (Fig. 2b). Apart from the transcriptomes of 30 tissues (Fig. 3a, b, Supplementary Fig. 5 and Supplementary Table 6), in-depth sequencing of kernel and mesocarp (Fig. 3c and Supplementary Table 9) indicated that the transcriptome signatures are similar in both tissues with plastidial FAS genes upregulated compared to TAG. Thus, the enzymes of the Kennedy pathway for TAG assembly must cope with the high flux of *de novo* fatty acids in oil palm. FAS genes were markedly up-regulated just before the onset, and then declined during the peak of lipid accumulation. This contrasts with previous reports that FAS transcripts continue to increase in the mesocarp[1]. This may reflect the *tenera* fruit-form used in this study, as *dura* has a longer maturation period (22WAA)[24] than *tenera* (20WAA)[23,25]. WRI1 regulates oil accumulation in the oil palm mesocarp[1] and we found highest mesocarp expression of WRI1 just before lipid accumulation onset. However, kernel showed 75-fold higher WRI1 expression compared to mesocarp (Fig. 3c and Supplementary Table 9), implying its pivotal role in kernel oil synthesis. LEAFY COTYLEDON1 (LEC1), LEC2, ABSCISIC ACID INSENSITIVE3 (ABI3) and FUSCA3, which activate WRI1 in oilseeds, were not found in mesocarp transcriptomes, but were well-represented in the kernel, with LEC1 and ABI3 showing highest expression at both start and completion of oil accumulation. Interestingly, the transcriptional regulator PICKLE (PKL) was expressed in both the mesocarp and kernel (Supplementary Table 9).

Genes involved in sucrose degradation and the oxidative pentose phosphate pathway were more represented in oil palm than date palm (Fig. 2b). Pentose phosphates are recycled into glucose 6-phosphate to fuel glycolytic pathways, and import of these cytosolic metabolites requires specific transporters on the plastid envelope. Although these transporter genes are strongly upregulated in oil palm[1], they are similarly represented in date and oil palm genomes (Fig 2b). Thus, channeling of sugars destined for oil synthesis is regulated at the

transcriptome level in oil palm. Additional insights important to TAG biosynthesis, fruit ripening and abscission are provided in Supplementary Notes.

To place palms on the evolutionary tree, evidence-based gene models from each species were combined with a previous seed plant dataset[26] to form a matrix of 1,685 gene partitions (858,954 patterns) and 107 taxa. *P. dactylifera*, *E. guineensis* and *E. oleifera* are present in 1,206, 1,229 and 1,190 partitions, respectively. All three were well separated from other monocots (Fig. 4), including nearest neighbors *Musa* (banana), *Curcuma* (turmeric) and *Zingiber* (ginger). Phylogenetic dating using conservative constraints (Supplementary Methods) predicted 65 Mya divergence between date and oil palm, and 51 Mya between *E. oleifera* and *E. guineensis*. This is comparable with divergence between old and new world relatives such as African *S. bicolor* (sorghum) and American *Z. mays* (maize) panicoid grasses (26 Mya). Unlike maize and sorghum, however, *E. guineensis* and *E. oleifera* give rise to fertile hybrids[2], consistent with the vicariant hypothesis for phylogeographical divergence, in which geographically isolated species are under no selective pressure to evolve reproductive isolation[27].

The genome sequence of oil palm will be a rich resource for oil palm breeders, geneticists and evolutionary biologists alike. It has revealed that palms are ancient tetraploids, and that the African and S. American species likely diverged in the old and new worlds. Over-represented genes in lipid and carbohydrate metabolism are differentially expressed in mesocarp and kernel, accounting for the different properties of palm fruit and palm kernel oils[14]. The genome sequence will also allow mapping of somaclonal epigenetic alterations which restrict the use of clones in commercial plantings. The dense representation of sequenced scaffolds on the genetic map will facilitate identification of genes responsible for important yield and quality traits. The genome sequence of this tropical plantation crop is an important step in achieving the goals critical to the sustainability challenges associated with growing demands for biofuels and edible oils.

METHODS

Genome assembly

The assembly of *E. guineensis* (AVROS, *pisifera* fruit form) genome build 5 (P5) was constructed from a total of 148 linker libraries and 81 fragment libraries generated using the Roche 454 GS FLX genome sequencer. Reads were generated from genomic DNA fragment libraries (53.5 million reads), BAC pool DNA fragment libraries (3.6 million reads), and from a series of genomic (89.1 million reads) and BAC (8.6 million reads) paired end linker libraries. In total, 46.8 billion bases of raw sequence were generated, representing approximately 26-fold raw sequence coverage of the 1.8 Gb oil palm genome. Sequence data were assembled using the Newbler algorithm[30] (Supplementary Table 4). The assembly of the *E. oleifera* genome build 7 (O7) was constructed from a total of 127 linker libraries and 68 fragment libraries. In total, 130 million *E. oleifera* reads, representing approximately 25-fold raw sequence coverage, were generated (Supplementary Table 2).

High information content restriction fragment (HICF) fingerprints were generated for 124,286 BAC clones from the *E. guineensis* genome with an average size of 150Kb. BAC

fingerprints were used to construct a physical map of the reference genome (Supplementary Fig. 2 and 3). In addition, BAC ends from this library were sequenced using Sanger sequencing to generate 235,613 paired reads with an average read length of approximately 600bp. BAC end reads were used to create the shotgun assembly, as well as to locate individual BACs within the assembled genome.

Prior to assembly, all 454 sequence runs were screened for quality based on average read length, linker library efficiency and library redundancy using a custom pipeline based on the SSAHA program[31]. For linker positive reads, library insert sizes were estimated by aligning both ends to a draft assembly of oil palm, and measuring the intervening sequence in cases where both ends matched a single draft scaffold. In order to minimize the negative impact of chimeras in the linker libraries, we removed all identical reads due to library redundancy, not independent observations. In addition to the 454 data, a set of 235,613 BAC end reads from the Origen_1 BAC library were included in the P5 assembly, with an average BAC size of 150Kb (Supplementary Table 3). These data were assembled using the Newbler algorithm[30] on a Dell PowerEdge R910 server with 512Gb of RAM, and 32 cores/64 threads running Ubuntu Linux 10.04. The P5 assembly took 15 days 5 hours to complete.

Genetic map construction

Two genetic maps were constructed. The first mapping population was derived from the self-pollination of the Nigerian *tenera* palm T128. The mapping population and map construction methodology are described elsewhere in this issue[32]. The second mapping population comprised 87 palms obtained from a cross between Ulu Remis Deli *dura* (ENL48) and Yangambi *pisifera* (ML161) grown at the FELDA Research Station at Jerantut, Malaysia. Linkage analysis for the *dura* x *pisifera* (P2) cross was performed using both JoinMap® 4.0 and GenStat 14th edition. JoinMap was used to examine markers and identify loci showing distorted segregation (X^2test). JoinMap was initially used to construct the two parental framework maps at recombination frequency (rf) ≤ 0.2 and a nearest neighbor stress (N.N. Stress) value of ≤ 4 (cM) using the maximum likelihood (ML) mapping algorithm as described[33]. The density of the linkage maps was later increased by mapping additional co-dominant markers into the parental framework maps using the ML mapping algorithm in GenStat 14. The integrated map was built using the ML mapping method in GenStat14 by combining data from markers on both the two parental maps. Comparison between the integrated map and the parental maps was visualized using MapChart 2.2[34].

Genetic map integration and chromosome sequence construction

The 1,511 markers used in the generation of the T128_codominant and P2_DxP maps were compared to scaffolds from the P5 build using the exonerate[35] program (-m ungapped –percent 97). Markers that did not uniquely map to P5 scaffolds were discarded, and one scaffold/marker ordering was created for each of the two maps. After reviewing shared scaffolds between the two maps, a final order was determined for ordering 169 scaffolds, and ordering and orienting 124 scaffolds based on multiple markers. The scaffold sequences were then concatenated in order and reversed/complemented as needed to create 16 linkage group based chromosome sequences. During map integration, LG15 (T128_codominant

naming convention) appeared significantly shorter than a previous integration based on the P4 assembly with the P2_regression map. This was due to map instability introduced into the P2_DxP map generation where the P2_regression was more stable. After review, the LG15 chromosome sequence was extended based on mapping of P5 scaffolds onto the P4/P2_regression version.

Gene identification and annotation

Based on the longest set of scaffolds representing 10% of the original P1 build, we used the SNAP[36] gene finder to identify initial candidate gene predictions for the *E. guineensis* genome. Initial SNAP runs were performed using the rice (Os.hmm) gene model. Initial genes discovered were compared with the RefSeq35 database as well as the TIGR Gramineae repeat database in order to remove retroelements and pseudo-genes. The remaining transcripts were then screened for missing start and stop codons, as well as other warnings from SNAP. The remaining set was then used as input to the SNAP programs FATHOM and FORGE according to the SNAP documentation to create a new HMM with greater specificity to *E. guineensis.* The same screening process was applied again and four iterations if the training were used resulting in the final "pisif_2_22_11.hmm" gene model.

For the Glimmer[37] predictions, assembled *E. guineensis* transcriptome sequences (groups A, C, D and G in Supplementary Table 6) were translated from start to stop with a size selection ranging from 500 to 5,000 nucleotides. These were then compared to Magnoliophyta complete coding sequences from Genbank using BLASTX (Eval $\leq 10^{-10}$). Transcripts with significant homology starting at position one of the Magnoliophyta targets were selected for further analysis. The transcripts were then screened using BLASTClust (NCBI) and CD-HIT[38] to reduce the number of genes to meet Glimmer's training requirements. Exon boundaries were determined by mapping to the previous P4 build, and were used to create an *E. guineensis* GlimmerHMM.

Transcriptome analysis

454 reads from each library (Fig. 3a, b and Supplementary Table 6) were assembled into isotigs individually and totally. The isotigs from each library were blasted on *A. thaliana* gene models with a threshold of E < 10–5. The best hit *A. thaliana* gene model was assigned to the homolog of the query isotig. Based on Bourgis's annotation[1], copy numbers were given to each gene in each category group. The final copy numbers of each functional group were scaled on the number of genes in each group. To estimate expression level of genes in mesocarp and kernel tissues, 454 reads from each library were mapped to assembled isotigs from all *E. guineensis* reads by using BWA. Gene group expression levels were reads coverage, which was the number of mapped 454 reads on each isotig divided by the total number of isotigs, times 100,000, and scaled on the number of genes in each categorical group. Both copy number and read coverage were the mean of measures from two biological replicates. Data were analyzed as described above for 454 data, except that expression levels were calculated as transcripts per million tags.

Methylation filtered library analysis

Methylation-filtered ("GeneThresher") genomic DNA libraries were constructed as described [15,39] to select unmethylated clones (depleted of most repetitive sequences[40]) by propagation in *Mcr*BC$^+$ strains of *E.coli*. Briefly, nuclear DNA was individually extracted from young leaves of Deli *dura* and *AVROS pisifera* of *E. guineensis* and from *E. oleifera*. For each of the three DNA populations, genomic shotgun libraries were constructed as described [39]. Sequences were generated from one end of each cloned insert by ABI 3730 sequencing (Life Technologies), generating 298,039 reads (73,390 from Deli *dura*, 101,327 from *AVROS pisifera* and 123,322 from *E. oleifera*).

Segmental duplication analysis

Pseudo chromosomes from the EG-5linked assembly were screened in a self-self comparison using the MUMmer3 set of tools[41]. Final alignments were performed on chromosome pairs using the PROmer program (-d 0.5 -c 200), alignments were visually reviewed with the mummerplot program, and approximate boundaries for segmental duplications were recorded (Supplementary Figure 7 and Supplementary Table 7).

To test for the existence of observed segmental duplications in other genomes, we performed comparative genomics of each half of the 16 segmental duplications in *E. guineensis* with the the *E. oleifera* and *P. dactylifera* scaffold sets. Comparisons were performed using a custom analysis pipeline based on the mummer program (mummer -n -l 30 -b -c -L). Mummer output was summarized as an offset-sorted overlap plot showing where query scaffolds share local alignment with a reference chromosome. Output was reviewed to verify that each half of the proposed segmental duplication was present in the query genome, and that the scaffolds matching were different for each half of the duplication in *E. guineensis*.

Genome comparison by gene models

NCBI TBLASTN program was used to compare *A. thaliana, O. sativa, P. dactylifera, and E. guineensis* predicted proteins on all four species' genomes respectively with a threshold of $E < 10^{-5}$. At this level of conservation, matches represent shared gene families between the query and target genomes. By comparing predicted proteins to genome sequence, biases introduced by gene prediction methods are minimized as compared to a direct gene level comparison. Comparisons between the gene models from one species and its own genome are close to, but less than 100% due to the limit of sensitivity of TBLASTN at this e-value cutoff. Public database sources for comparative genome sequences and gene models are provided in Supplementary Methods.

Gene clustering and Venn diagram

CD-HIT clustering algorithm[38] was used to look for homologous protein sequences among *A. thaliana, O. sativa, P. dactylifera, and E. guineensis* at 40% similarity level. This algorithm avoids all-versus-all BLAST search by using a short word filter.

Gene Ontology (GO) annotation

E. guineensis gene model protein sequences were queried against *A. thaliana* annotated gene model protein sequences by using BLASTP[42] with a threshold of $E < 10^{-5}$. The blast output file was then loaded into Blast2GO[43]. Blast2GO performed GO annotation by using an annotation rule (AR) on the found ontology terms. The most specific annotations were assigned on each sequence with default parameters in AR.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge the contributions of Noh Ahmad, Marhalil Marjuni and Norziha Abdullah of MPOB for sampling of oil palm materials. We thank MoGene, GeneWorks, Beijing Genome Institute, Tufts University Core Facility and Macrogen for sequencing services. We thank Dennis Stevenson (New York Botanical Gardens, NY) for useful discussions. We appreciate the unflagging support from Datuk Dr. Choo Yuen May, Director General of MPOB as well as the Ministry of Plantation Industries and Commodities, Malaysia. E.L., R. DeS., and R.M. are supported by a grant from NSF 0421604 “Genomics of Comparative Seed Plant Evolution”. R.M. is supported by the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation.

References

1. Bourgis F, et al. 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. 2011; 108:12527–32. [PubMed: 21709233]
2. Hardon JJ, Tan GY. Interspecific hybrids in the genus Elaeis I and II. Euphytica. 1969; 18:372–380.
3. Jaligot E, et al. Epigenetic imbalance and the floral developmental abnormality of the in vitro-regenerated oil palm Elaeis guineensis. Ann Bot. 2011; 108:1453–62. [PubMed: 21224269]
4. Dransfield, J.; Uhl, NW.; Royal Botanic Gardens, K. Genera Palmarum: the evolution and classification of palms. Kew Pub; 2008.
5. Purseglove, JW. Monocotyledons. Longman; London: 1972. p. 607
6. Zeven AC. The origin of the oil palm. Journal of the Nigerian Institute of Oil Palm Research. 1965; 4:218–225.
7. Cochard B, Amblard P, Durand-Gasselin T. Oil palm genetic improvement and sustainable development. Oleagineux, Corps Gras, Lipides. 2005; 12:141–7.
8. Hartley, C. The Oil Palm. Hartley, C., editor. Longman; 1988. p. 47-94.
9. Corley, RHV.; Tinker, PB. The Oil Palm. Blackwell Science; Oxford: 2003. p. 1-26.
10. Mayes S, Jack PL, Corley RH, Marshall DF. Construction of a RFLP genetic linkage map for oil palm (Elaeis guineensis Jacq.). Genome. 1997; 40:116–22. [PubMed: 18464812]
11. Jouannic S, et al. Analysis of expressed sequence tags from oil palm (Elaeis guineensis). FEBS Lett. 2005; 579:2709–14. [PubMed: 15862313]
12. Ho C-L, et al. Analysis and functional annotation of expressed sequence tags (ESTs) from multiple tissues of oil palm (Elaeis guineensis Jacq.). BMC Genomics. 2007; 8:381. [PubMed: 17953740]
13. Low ET, et al. Oil palm (Elaeis guineensis Jacq.) tissue culture ESTs: identifying genes associated with callogenesis and embryogenesis. BMC Plant Biol. 2008; 8:62. [PubMed: 18507865]
14. Tranbarger TJ, et al. Regulatory mechanisms underlying oil palm fruit mesocarp maturation, ripening, and functional specialization in lipid and carotenoid metabolism. Plant Physiol. 2011; 156:564–84. [PubMed: 21487046]
15. Palmer LE, et al. Maize genome sequencing by methylation filtration. Science. 2003; 302:2115–7. [PubMed: 14684820]

16. Al-Dous EK, et al. De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). Nat Biotechnol. 2011; 29:521–7. [PubMed: 21623354]
17. Castilho AM, Vershinin AV, Heslop-Harrison JS. Repetitive DNA and the Chromosomes in the Genome of Oil Palm (Elaeis guineensis). Annals of Botany. 2000; 85:837–844.
18. Richards EJ, Chao S, Vongs A, Yang J. Characterization of Arabidopsis thaliana telomeres isolated in yeast. Nucleic Acids Res. 1992; 20:4039–46. [PubMed: 1508688]
19. Du J, et al. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J. 2010; 63:584–98. [PubMed: 20525006]
20. Singh, Rea. Identification of cDNA RFLP markers and their use for molecular mapping in oil palm. Asia Pacific Journal of Molecular Biology & Biotechnology. 2008; 16:53–63.
21. Wood TE, et al. The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci U S A. 2009; 106:13875–9. [PubMed: 19667210]
22. D'Hont A, et al. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature. 2012; 488:213–7. [PubMed: 22801500]
23. Sambanthamurthi R, Sundram K, Tan Y. Chemistry and biochemistry of palm oil. Prog Lipid Res. 2000; 39:507–58. [PubMed: 11106812]
24. Bafor ME, Osagie AU. Changes in Lipid Class and Fatty Acid Composition During Maturation of Mesocarp of Oil Palm (Elaeis guineensis) Variety Dura. Journal of Science in Food Agriculture. 1986; 37:825–832.
25. Shaarani SM, Cárdenas-Blanco A, Amin MHG, Soon NG, Hall LD. Monitoring development and ripeness of oil palm fruit (Elaeis guneensis) by MRI and bulk NMR. International Journal Agriculture & Biology. 2010; 12:101–105.
26. Lee EK, et al. A functional phylogenomic view of the seed plants. PLoS Genet. 2011; 7:e1002411. [PubMed: 22194700]
27. Riggins CW, Seigler DS. The genus Artemisia (Asteraceae: Anthemideae) at a continental crossroads: molecular insights into migrations, disjunctions, and reticulations among Old and New World species from a Beringian perspective. Mol Phylogenet Evol. 2012; 64:471–90. [PubMed: 22580463]
28. Chiu JC, et al. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics. 2006; 22:699–707. [PubMed: 16410324]
29. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22:2688–90. [PubMed: 16928733]
30. Margulies M, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–80. [PubMed: 16056220]
31. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res. 2001; 11:1725–9. [PubMed: 11591649]
32. Singh R, Low L, Ooi LCL. The oil palm Shell gene controls oil yield and encodes a homologue of SEEDSTICK. submitted.
33. Jansen J. Construction of linkage maps in full-sib families of diploid outbreeding species by minimizing the number of recombinations in hidden inheritance vectors. Genetics. 2005; 170:2013–25. [PubMed: 15944349]
34. Voorrips RE. MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered. 2002; 93:77–8. [PubMed: 12011185]
35. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005; 6:31. [PubMed: 15713233]
36. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004; 5:59. [PubMed: 15144565]
37. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004; 20:2878–9. [PubMed: 15145805]
38. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22:1658–9. [PubMed: 16731699]
39. Bedell JA, et al. Sorghum genome sequencing by methylation filtration. PLoS Biol. 2005; 3:e13. [PubMed: 15660154]

40. Ouyang S, Buell CR. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. 2004; 32:D360–3. [PubMed: 14681434]

41. Kurtz S, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5:R12. [PubMed: 14759262]

42. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–402. [PubMed: 9254694]

43. Conesa A, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005; 21:3674–6. [PubMed: 16081474]

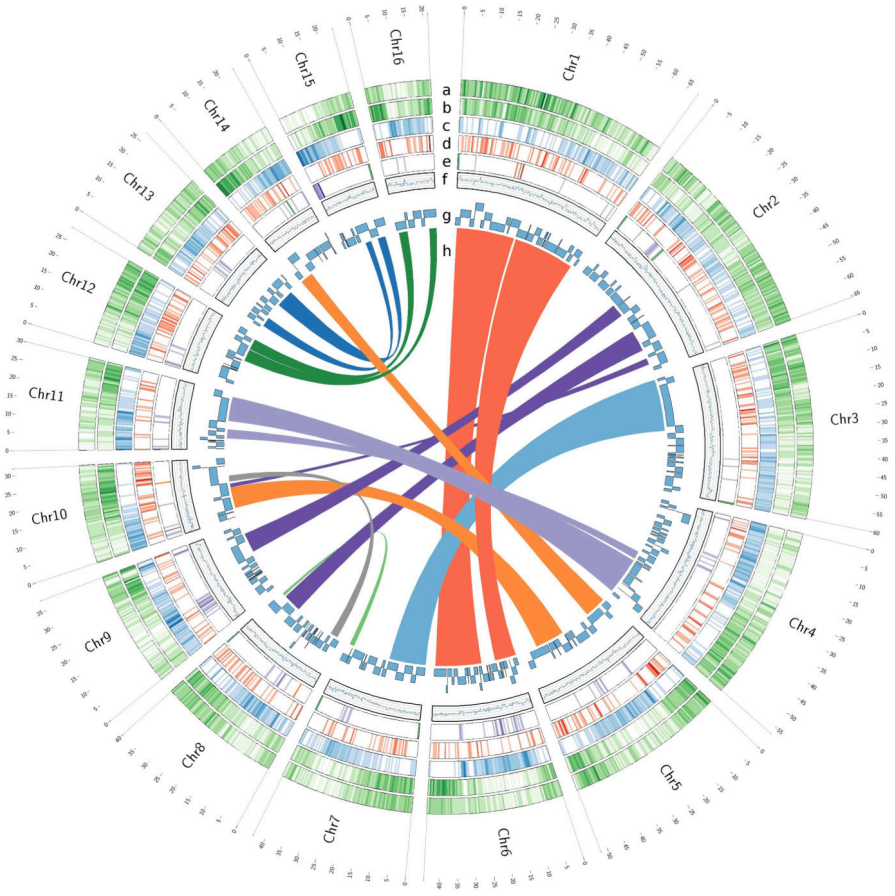

Fig. 1. The chromosomes of Oil Palm

E. guineensis has 16 chromosome pairs, ordered by size, which correspond to 16 linkage groups identified by genetic mapping (Supplementary Table 7). Tracks displayed are **a,** gene density, **b,** methyl-filtered read density, **c,** retroelement density, **d,** SSR repeats, **e,** low copy number repetitive elements including telomere repeat TTTAGGG (green), 5S rRNA (orange) and pericentromeric repeats (purple), **f,** regional G+C content (range 0.3–0.45), **g,** genetically mapped scaffolds from the P5 Build and **h,** segmental duplications. Densities for telomere repeats are exaggerated for visual clarity.

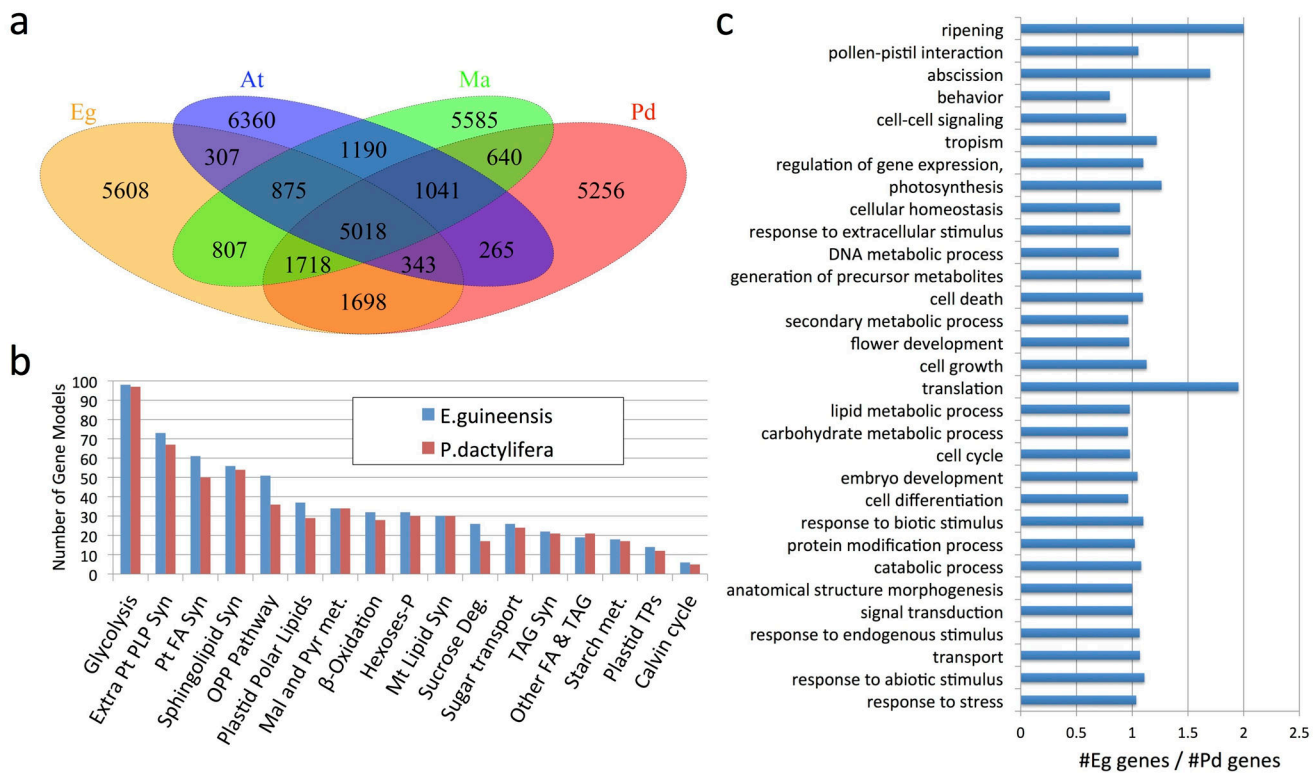

Fig. 2. Gene model comparisons

a, Venn Diagram illustrating proportion of shared gene family clusters in oil palm (Eg), banana (Ma), Arabidopsis (At) and date palm (Pd). Number of genes (clusters) compared were At: 27,416 (15,609) Ma: 36,529 (16,874) Pd: 34,804 (16,407) and Eg: 28,882 (16,802). **b,** GO classifications of oil palm (blue) and date palm (red). **c,** Ratio of gene number (oil palm:date palm) in each GO classification. Abbreviations: Pt PLP Syn, plastidial phospholipid synthesis; Pt FA Syn, plastidial fatty acid synthesis; OPP, oxidative pentose phosphate; Mal and Pyr met., Malate and Pyruvate Metabolism; Hexoses-P, hexose phosphate pathway; Mt Lipid Syn, mitochondrial lipid synthesis; Deg., degradation; TAG, triacylglycerol; TPs, transporters.

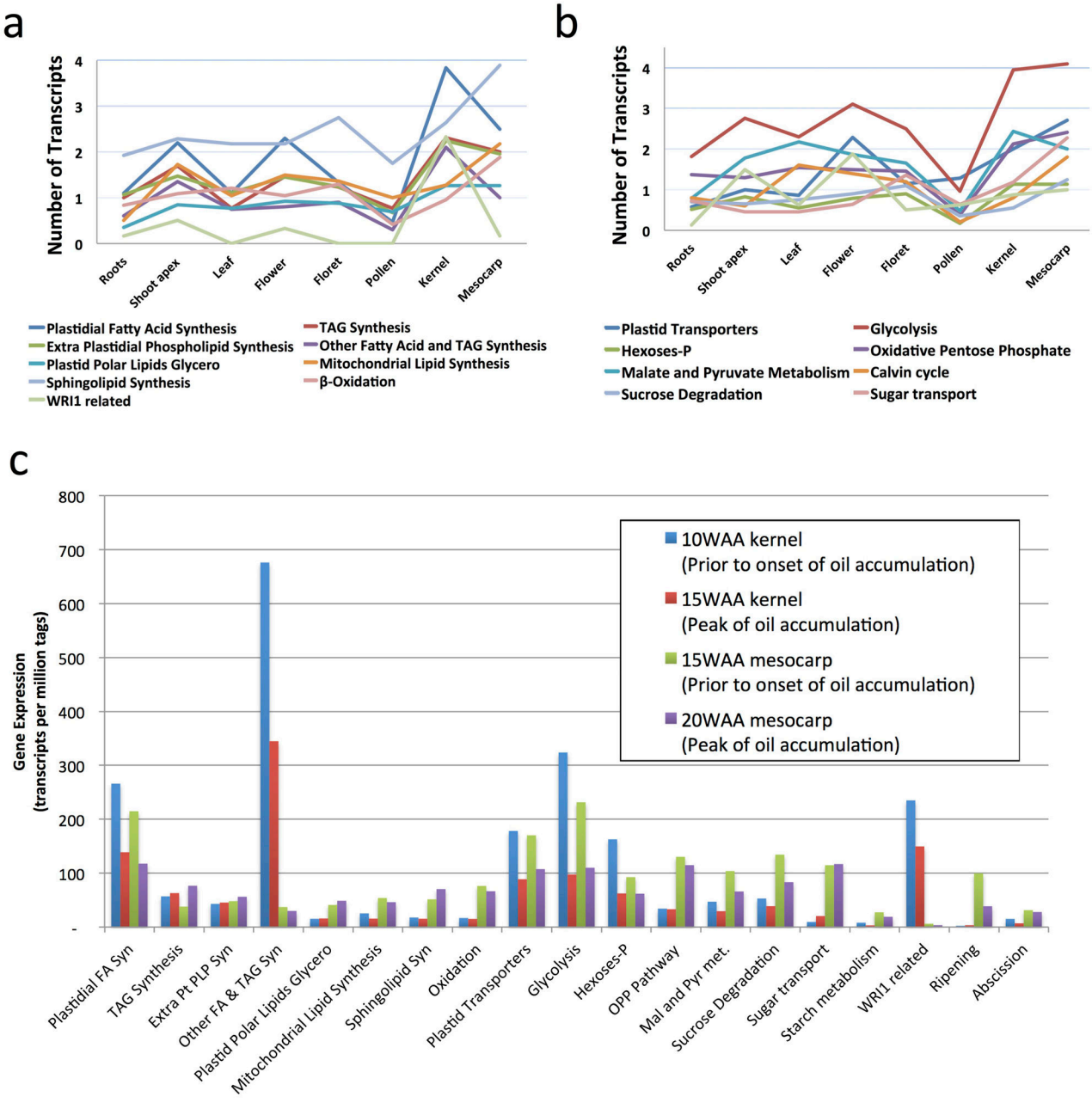

Fig. 3. Lipid and Carbohydrate metabolism in oil palm fruits

Number of **a**, lipid synthesis related and **b**, carbohydrate synthesis related gene transcripts in different tissues. **c,** Comparison of gene expression levels between kernel and mesocarp tissue prior to and at peak of oil accumulation. Abbreviations: FA Syn, fatty acid synthesis; TAG, triacylglycerol; Pt PLP Syn, plastidial phospholipid synthesis; Glycero, Galacto, Glycero, and Sulfo Lipids; OPP, Oxidative Pentose Phosphate; Mal and Pyr met., Malate and Pyruvate Metabolism.

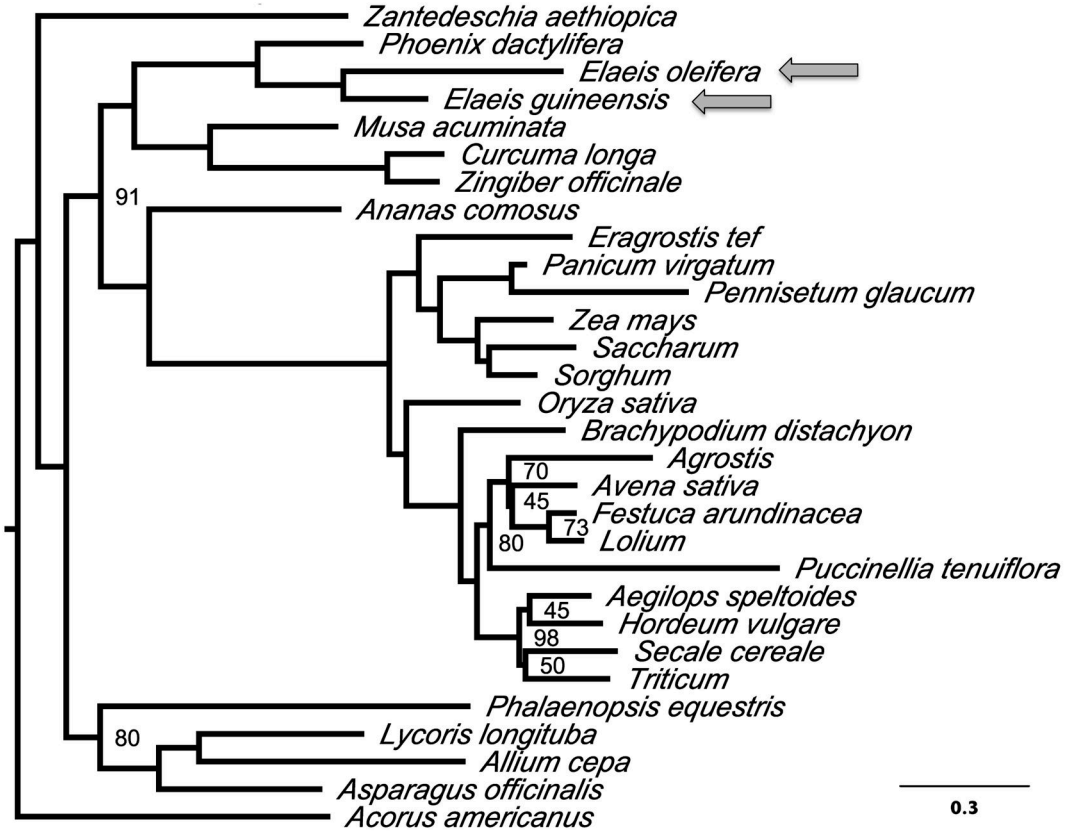

Fig. 4. Phylogenetic analysis

A carefully annotated subset of proteins from *E. guineensis*, *E. oleifera* and *P. dactylifera* were included in a matrix of 1,685 gene partitions (858,954 patterns) and 107 taxa. This matrix is extracted from partitions with at least 30 taxa present in a much larger matrix. A maximum likelihood tree of monocotyledonous taxa is shown, along with bootstrap values when less than 100 (Methods Summary). Scale bar indicates the mean number of substitutions per site.