

Genome wide association study (GWAS) for morphological and yield-related traits in oil palm hybrids (*Elaeis oleifera* x *Elaeis guineensis*)

Jaime Osorio

AGROSAVIA

Gina Garzón

AGROSAVIA

Paola Delgadillo

AGROSAVIA

Silvio Bastidas

AGROSAVIA

Leidy Moreno

AGROSAVIA

Felix Enciso

AGROSAVIA

Omar Comejo

Washington State University

Luz Stella Barrero (✉ lbarrero@agrosavia.co)

AGROSAVIA

Research article

Keywords: Association mapping, *Elaeis guineensis*, *Elaeis oleifera*, genotyping-by-sequencing, plant architecture, yield.

Posted Date: October 16th, 2019

DOI: <https://doi.org/10.21203/rs.2.13122/v2>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Version of Record: A version of this preprint was published on December 3rd, 2019. See the published version at <https://doi.org/10.1186/s12870-019-2153-8>.

Abstract

Abstract Background The genus *Elaeis* has two species of economic importance for the oil palm agroindustry: *Elaeis oleifera* (O), native to the Americas, and *Elaeis guineensis* (G), native to Africa. The work presented herein provides, to our knowledge, the first association mapping study in an interspecific OxG hybrid population of oil palm which presents tolerance to pests and diseases, high oil quality, and acceptable fruit bunch production. **Results** Using genotyping-by-sequencing (GBS), we identified a total of 3,776 single nucleotide polymorphisms (SNPs) that were used to perform a genome-wide association analysis (GWAS) in 378 OxG hybrids for 10 agronomic traits. Twelve genomic regions were located near candidate genes implicated in multiple functional categories, such as tissue growth, cellular trafficking, and physiological processes. **Conclusions** We provide new insights on candidate genes that mapped on genomic regions involved in plant architecture and yield; however, these potential candidate genes need to be confirmed for future targeted functional analysis. The associated markers may be valuable resources for the development of marker-assisted selection in oil palm breeding. **Keywords:** Association mapping, *Elaeis guineensis*, *Elaeis oleifera*, genotyping-by-sequencing, plant architecture, yield.

Background

The oil palm is an important crop with a higher quality oil and yield potential compared to other vegetable oil crops [1]. Colombia is the fourth-largest producer worldwide with 3.8 tons of palm oil per hectare, which positions this country above the world's average yield [2]. Within the *Arecaceae* family, the oil palm species (*Elaeis guineensis*), native to West Africa, is the main source of most vegetable oil [3]. However, another oil palm species, which is native to the tropics of Central and South America, known as *Elaeis oleifera*, is recognized for its high yield production [3]. The palm is a perennial monocot with a lifespan of approximately 25 years [4], which results in slow breeding progress. The Colombian Corporation of Agricultural Research (Agrosavia) breeding program has focused on developing OxG interspecific hybrids (*E. oleifera* x *E. guineensis*). The OxG is characterized by a slow trunk growth [5], as well as tolerance to bud rot [6–9] and red ring diseases [10] in comparison to their parents. Additionally, these hybrid populations inherit the parthenocarpic fruit development of *E. oleifera* species, which allows the production of seedless fruits [11].

Saturated genetic linkage maps are important for the identification of genomic regions associated with major genes and quantitative trait loci (QTLs) controlling agronomic traits. Over the last 20 years, genetic maps of the oil palm have been constructed. The first genetic map for oil palm was generated using restriction fragment length polymorphisms (RFLPs) and amplified fragment length polymorphisms (AFLPs) [12, 13]. Dense genetic maps were subsequently constructed using simple sequence repeats (SSRs) and single nucleotide polymorphism (SNP) markers, which have also been used for QTL identification. Thus, Jeennor and Volkaert [14] identified a QTL associated with bunch weight using a mapping population of 69 accessions and generated a genetic map with 89 SSRs and 101 SNPs. Billotte *et al.* [15], used a multi-parent linkage map generated with 251 SSRs and reported QTLs associated to bunch traits. Similar approaches have enabled the identification of 164 QTLs associated with 21 oil yield components using SSR, AFLP, and RFLP markers [16].

In recent years, advances in next-generation sequencing technology have lowered the cost of DNA sequencing to the point that we can obtain millions of SNPs compared with other technologies [17, 18]. Particularly, genotyping-by-sequencing (GBS) is a rapid, low-cost, and robust approach to screen breeding populations using hundreds or thousands of SNPs [19]. Pootakham *et al.* [20] constructed an oil palm map using an F_2 population and 1,085 SNPs derived from GBS and were able to identify QTLs for height and fruit bunch weight. Similarly, a genome-wide association analysis (GWAS), using a larger number of SNPs (4,031) derived from GBS across a diverse panel of *E. guineensis*, allowed to identify novel QTLs associated with the increase of trunk height [21].

GWAS has been proposed as a robust approach over QTL linkage mapping [22]. The use of a wide range of genetic background in GWAS analyses increases the possibility of detecting QTL regions associated with traits of interest, compared to the limited genetic variation of a bi-parental mapping population [23]. However, GWAS limitations such as the effect of population structure can lead to spurious associations between a candidate marker and a specific phenotypic trait [24]. To solve this, the mixed linear model incorporates the structure data (Q) and the relative kinship effects (K), resulting in the reduction of false-positive associations [25].

Given the food, industrial, and medical purposes, palm oil has experienced rapid growth in economic importance and nowadays is considered the second most traded vegetable oil crop in the world after soybean [26, 27]. The increasing demand for this crop is caused by a shift away from trans-fats to healthier alternatives [28], and because its residues can be processed to produce biofuel [27]. For these reasons, the identification of specific genomic regions and their genes involved in morphological traits such as height, foliar area, and its relationship with productivity, is becoming more critical for this crop.

Although previous studies have identified QTLs controlling morphological and yield-related traits in oil palm, these QTLs were detected using F_1 intraspecific populations. Our study is the first report in which molecular markers were mapped through association analysis in an F_1 interspecific OxG hybrid population. Our study aims to: (i) genotype an OxG oil palm mapping population; (ii) perform GWAS to identify loci or candidate genes involved in morphological and yield-related traits for future use in breeding programs.

Results

Analysis of phenotypic data

Means, standard deviations, and range values of the phenotypic data for the 378 OxG hybrids are shown in Table 1. The principal component analysis (PCA) showed that the first principal component (PC1) explained 45.6% of the total phenotypic variation, where morphological-related traits such as leaf area (LA), foliar area (FA), leaf dry weight (LDW), and trunk height (HT) contributed extensively to this component. Meanwhile, the second principal component (PC2) explained 19.9% of the variance, associated mainly to all yield-related traits (Figure 1A). Moreover, positive correlations were observed between most of the morphological traits ($r = 0.1$ to 0.8), while lower correlation values were found between yield and most of the morphological traits ($r \leq 0.3$) (Figure 1B). Notably,

HT was correlated with morphological-related traits such as FA, LA, LDW and trunk diameter (TD) (r^2 0.6), while yield and bunch number (BN) showed a high correlation between them ($r = 0.91$), but a lower correlation ($r = 0.57$) with bunch weight (BW).

To evaluate the phenotypic similarity among the 378 OxG hybrids, a hierarchical cluster analysis was performed (Figure 2; supplementary Table S1). We found phenotypic differences between the two clusters in agreement to the variability of the morphological-related traits. Overall, the Group II showed the highest mean values for all the morphological-related traits (supplementary Figure S1). For example, OxG from Group II were significantly taller (HT = 269 ± 21 cm) compared to OxG from Group I (HT = 238 ± 28 cm) ($p \leq 0.0001$). However, yield-related traits did not show any significant differences between groups.

SNP calling

A total of 1,058,182,456 raw Illumina sequencing reads from seven Illumina HiSeq lanes were generated for 471 accessions (62 *E. oleifera* (O), 31 *E. guineensis* (G), and 378 F₁ OxG). The genotyping of the collection detected 131,825 SNPs covering the 16 oil palm chromosomes. After filtering, 3,776 SNPs with an average of 236 SNPs per chromosome were retained (supplementary Table S2).

Cluster and association analyses

The neighbor-joining (NJ) of the entire population (471 accessions) (Figure 3A) showed two main groups containing *E. oleifera* and *E. guineensis*, as well as three groups within the hybrid population; one group more similar to *E. guineensis*, another one more similar to *E. oleifera*, and a largest group that showed intermediate similarity between the two parental species. The PCA analysis of the OxG population (378 hybrids) showed that the first three components comprised approximately 15.47% of the total variation and categorized the population in three groups supporting the results observed in the NJ tree in accordance with the hybrid nature of our population and its genetic similarity with each parental species (Figure 3B).

We performed the association analysis with the 378 OxG hybrids and the 3,776 SNPs for seven morphological traits and three yield-related traits (Table 1). Twelve SNPs were the most significantly associated with the traits measured based on p-values across different genomic regions of the oil palm genome before the false discovery rate (FDR) correction (Table 2). Common SNPs for rachis length (RL) and leaflet per leaf (LXL) were observed, as well as for HT and LA and between yield and BN, in accordance with the results from the phenotypic correlations between traits. The Q-Q plots (Figure 4) significantly supported the evidence of SNP associations with the traits ($p \leq 0.005$) and suggested that population stratification in the GWAS model was adequately controlled.

The availability of the oil palm genome sequence [29] enabled to associate genomic regions with specific QTLs into the physical map and to explore potential candidate genes and their possible function. On chromosomes 3, 13, and 15, we identified 10 significant SNPs located on genomic regions harboring genes associated with the morphological-related traits before FDR correction (Figure 4 - Table 2). For yield-related traits, we observed two SNPs into two candidate genes on chromosomes 5 and 10, which were non-significant after FDR correction (Figure 4, Table 2). We evaluated if the SNPs found in association with traits were in chromosomes with a larger number of markers, to assess if our results could have been obtained by biases in genotyping. The associated SNPs found in this study (chromosomes 3, 5, 10, 13 and 15) were not located in the chromosomes with the higher number of SNPs identified by the GBS approach (supplementary Table S2).

The pair-wise linkage disequilibrium (LD) between the SNPs for the chromosomes that presented genomic regions associated to the evaluated traits is illustrated in supplementary Figure S2. The LD blocks were small for all chromosomes presented, which was expected, considering the out-crossing nature of the species and the hybrid population. The R^2 at or near significant SNPs showed that only the genomic region associated on chromosome 15 was in a relatively high LD group ($R^2 = 0.58$).

Discussion

Improving oil quality and increasing yield per hectare in oil palm is a major concern in the oil processing industry. Agrosavia's palm breeding program has focused on developing interspecific hybrids of OxG which present heterosis in traits such as resistance to diseases, fruit number, fruit weight, leaf length, and trunk diameter [30]. To our knowledge, this study represents the first GWAS analysis of an OxG hybrid population.

Phenotypic data

Correlation analysis results among yield-related traits indicated that BN could be a potential and better selection criterion for production than BW in the hybrid population. It is known that the leaf emission rate determines the BN in a palm, which is negatively correlated with the number of leaves [31]. In our study, no significant correlations between yield and leaf-related traits (FA, LA, LDW, LXL, RL) were found, however, a previous study in *E. oleifera* and OxG hybrids found that BN can be higher than the number of leaves just when oil palms produced multiple inflorescences [32]. Increases in BN and BW are also expected to correlate with increased mesocarp and kernel oil yields, as shown in other oil palm germplasm studies [33]. Future studies directed to improve oil yields for the OxG hybrid population should be conducted considering their importance in oil palm breeding.

Association analysis

SNPs detected by GBS have allowed the genetic dissection of genotypic variation in complex traits in many plant species [34–36]. Specifically, in oil palm the GBS technique has been used for identifying candidate genes related to oil bunch [37], bunch weight [20, 37], stem height [21] and oil quality traits studies [38] with the enzymes *ApeKI* and *PstI-MspI*. Here, we used the enzyme *PstI* for the library construction, allowing us to identify relatively high number of SNPs with probably novel genetics variants, which according to Chung *et al.* [39] is one of the most important advantages of GBS.

In the present study, association mapping resulted in the identification of 12 regions related to 10 morphological and yield-related traits (Table 2). However, only five regions associated with LDW, TD, RL, and LXL remained significant ($p \leq 0.05$) after the FDR correction. The small LD blocks in the heatmap analysis could suggest that the causal regions are nearby to the most significant SNPs; however, a more comprehensive study must be conducted to determine the actual SNPs and or genes controlling the trait of interest.

Therefore, we describe the five most significant regions and the genes located within those regions which might represent potential candidate genes involved in the expression of the phenotypic traits evaluated in this study. For morphological traits, a significant association was found for LDW on chromosome 3, explaining 10% of the phenotypic variation. The most significant SNP in this region was located in a mechanosensitive (MS) ion channel protein 10-like (*MSL10*) gene. In plants, MS ion channels have been proposed to play a wide array of roles, from the perception of touch and gravity to the osmotic homeostasis of intracellular organelles [40]. Besides, mechanoperception genes are essential for normal cell and tissue growth and development as well as for the proper response to an array of biotic and abiotic stresses [41]. A second significant region associated with TD on chromosome 15 was identified. Within this region a gene involved in nucleic acid binding and has a C2H2-type Zinc finger domain is located. The C2H2-ZF gene family has been proposed to be involved in the formation of wood, shoot and cambium development in species such as Poplar, as well as playing a role in stress and phytohormone response [42].

For RL and LXL traits, QTLs have been reported on chromosomes 4, 2, 10, and 16 [33]. In our study, three SNPs were associated with three different candidate genes for RL on chromosome 13. The SNP S13_20856724 is the closest to the *AGC3* gene and encodes different G proteins. G proteins had been reported to be involved in a wide range of developmental and physiological processes, having a high potential for yield improvement in crops such as rice [43]. The other significant association was found with the SNP S13_23674227, which is located in an extracellular ribonuclease gene (*RNase* gene). *RNase* genes have been studied for years in plants, playing an important role in plant defense mechanism [44] or plant development due to their ability to modify RNA levels, and thereby influence protein synthesis [45]. The SNP S13_25522088 was also significantly associated with RL and LXL, but further studies are necessary to determine their role, if any, in regulating these traits.

Seven SNPs were not significant after FDR correction, possibly due to reduced sample size. The QTL and association studies are limited by the relatively small mapping population sizes resulting in low statistical power, thus rendering small or even medium effect QTL statistically non-significant and difficult to detect. Such statistically underpowered populations may also suffer from severe inflation of effect size estimates (so-called Beavis effect) [46]. Increasing the population size and marker density is required to enable unbiased beavis effect estimation and to achieve higher statistical power [46–48], yet for perennial populations (long generation time) and limited offspring numbers this would require a considerable investment.

Although these seven regions were non-significant after FDR correction, three of them were located near or alongside genes whose function could be related with the expression of the traits under study. The first of these regions for HT harbors the SNPs S15_22553489 and S15_22553493 located at the *STYK* gene, which is involved in the control of stomatal movement in response to CO₂ [49]. Recent studies also showed the role of *STYK* in stem diameter by increasing the number of xylary fibers in species such as *Bambusa balcooa* [50]. Moreover, Pootakham *et al.* [20] found a region that controls the phenotypic variance of HT near this region on chromosome 15.

The other two regions were observed on chromosomes 5 and 10 for yield-related traits BN and BW. The gene p5.00_sc00003_p0367 potentially associated with BN, codes for a cation/H(+) antiporter. Antiporter proteins function as regulators of monovalent ions, pH homeostasis, and developmental processes in plants [51]. On chromosome 10, the gene p5.00_sc00004_p0097, potentially associated with BW, encodes a zinc finger protein (ZFP). The ZFP is a large protein family, involved in plant development, regulation of plant height, root development, flower development, seed germination, secondary wall thickening, anther development, and fruit ripening [52]. Studies conducted by Wu *et al.* [53] demonstrated that silencing a gene related to ZFP hampered fruit development in *Nicotiana benthamiana* [53]. This ZFP gene might play an important role on yield-related traits in oil palm, as shown in other plants, where overexpression of zinc finger proteins are related with higher yields [54].

In oil palm, harvesting of fruit bunches after a certain age is an arduous labor due to their tallness. For this reason, genotypes with less HT and TD are preferred among oil palm farmers. Likewise, larger foliar (RL and LDW) is related to a greater photosynthetic production which could be involved in higher productivity. But most importantly, increasing the number and weight of fruits means higher productivity per palm and therefore higher incomes for farmers. For this reason, leveraging QTLs or genes related to these traits, as the ones we identify in this study, could contribute to develop plant breeding strategies such as marker-assisted selection, that helps to select promising accessions in earlier stages (greenhouse conditions), and therefore, reduce the breeding cycle. Further work needs to focus on the biological functions of the set of potential candidate genes found in our research, yet the correlations identified in association studies cannot be dubbed as causations.

Conclusions

This is the first study reporting five significant genomic regions associated with morphological and yield-related traits based upon GWAS on an interspecific OxG oil palm population. Genes whose functional annotation are potentially related to the corresponding traits are located within those regions and therefore might represent potential candidate genes for these QTLs. Our results will provide the groundwork for the development of marker-assisted breeding in oil palm and will serve as a strong base for future functional studies to dissect high yield production.

Methods

Plant material

A total of 471 diverse oil palm accessions (62 *E. oleifera* (O), 31 *E. guineensis* (G), and 378 F₁ OxG) were collected from the research centers El Mira and La Libertad of the Colombian Corporation for Agricultural Research (Agrosavia) [55]. The OxG hybrid populations were generated through eight different crossings (eight different *E. oleifera* accessions as female progenitors were crossed with one *E. guineensis* accession as male progenitor); however, the parents of these crossings are already deceased. Details of the crosses and origin of accessions are given in supplementary Table S1. The plant material belongs to the National Germplasm Collection of Colombia, maintained by Agrosavia. All samples were collected following national regulations.

Phenotyping

Oil palm seedlings of the hybrid population were planted in a quincunx or triangular system with 10 m between the plants at the Agrosavia's research center El Mira, Tumaco, Colombia. Plants were randomly distributed using a randomized complete block design with four blocks.

Phenotypic data was collected for the subset of 378 F₁ OxG hybrids. A total of 10 traits distributed into two categories, morphological and yield-related, were evaluated (Table 1): i) For the morphological category: Trunk Diameter (TD, circumference of trunk at medium section), Trunk Height (HT, distance between the lowest green leaves and fruit), Rachis Length (RL, measured on fully expanded leaves), Leaf Dry Weight (LDW, mean dry weight per leaf multiplied by the number of leaves produced), Foliar Area (FA, mean area per leaf multiplied by the number of leaves per palm), Leaf Area (LA, mean area per leaf) and Leaflet per Leaf (LXL, measure of the largest leaflet). ii) For the yield-related category: Bunch Weight (BW, weight of fruits during harvest), Bunch number (BN, the number of fruits per palm during harvest) and Yield per Palm (Yield, Kg per palm per year). Each trait was measured according to the methodology proposed by Corley *et al.* [56] and Breure [57].

Statistical analysis of phenotypic data

The correlations among traits were calculated using the Pearson's correlation coefficient (r) at $p \leq 0.05$. To assess the relationship among the studied traits, a principal component analysis (PCA) was calculated. Finally, a hierarchical cluster analysis using the Ward's method was carried out to analyze the relationship among accessions. Differences between clusters by trait were determined using a t -test at $p \leq 0.0001$. All statistical analysis were performed in R v3.42 software [58].

Genotyping

Genomic DNA of the 471 accessions was extracted from leaf tissues using the DNeasy Plant Mini Kit (QIAGEN, Germany). The DNA quality was estimated using the *Hind*III enzyme and visualized by electrophoresis on 2% agarose gels. The GBS libraries were constructed with the methylation-sensitive restriction enzyme *Pst*I (CTGCAG). Sequencing was performed with 100-bp single-end reads using the Illumina HiSeq 2000 platform (Illumina Inc., United States) at the Institute of Genomic Diversity (Cornell University, Ithaca, NY, United States).

SNP discovery and data processing

Illumina reads were demultiplexed using the standard pipeline from the Tassel v4.5.9 software [59]. Then, reads were mapped to the oil palm reference genome of *E. guineensis* [60] using Bowtie2 [61] with the very-sensitive option. SNP calling was performed using the following parameters: minor allele frequency (MAF) < 5%, minimum locus coverage (mnLCov) of 0.9, minimum site coverage (mnScov) of 0.7 and minimum taxon coverage (mnTCov) of 0.5. Finally, SNPs were filtered to remove 95% of missing data and to retain biallelic SNPs with VCFtools v0.1.13 software [62].

Cluster and marker-trait association analyses

The clustering analysis for all 471 oil palm accessions was performed by a neighbor-joining algorithm using Tassel v4.3.5 [59] and visualized by Figtree v1.4.0 [63]. The population structure for the 378 OxG hybrids was evaluated through PCA using the R package SNPrelate [64]. Associations between molecular markers and phenotypic data were computed using a mixed linear model in the R package GAPIT (Genome Association and Prediction Integrated Tool) [65]. To avoid any possible bias caused by population structure, we included the first five principal components of the PCA and a relatedness (kinship) matrix from GAPIT in the mixed linear model. Quantile-quantile (Q-Q) plots from the observed $-\log_{10}$ p-values and the expected $-\log_{10}$ p-values were generated to study the appropriateness of the GWAS model. A false discovery rate (FDR) [66] was used to correct for spurious associations.

The heatmap of linkage disequilibrium (LD) was generated with a custom script by plotting pairwise R^2 values against the physical distance (base pairs) between markers on the same chromosome.

Potential candidate gene identification

The physical positions of the SNP markers were obtained from the Genomsawit Website of the International Malaysian Oil Palm Genome Programme (http://gbrowse.mpob.gov.my/fgb2/gbrowse/Eg5_1/). Gene annotations under the candidate gene regions were determined using published genome information of *E. guineensis* [56]. To assign putative biological functions of significant SNP markers associated with the traits, the flanking sequences of SNPs were queried against databases, such as: HMMER (<https://www.ebi.ac.uk/Tools/hmmer/>), NCBI (<http://www.ncbi.nlm.nih.gov/>), European Molecular Biology Laboratory (<http://www.ebi.ac.uk/>) and European Nucleotide Archive (<http://www.ebi.ac.uk/ena>).

List Of Abbreviations

FDR, False-Discovery-Rate; GBS, Genotyping-By-Sequencing; GWAS, Genome-Wide Association Studies; LD, Linkage Disequilibrium; PCA, Principal Component Analysis; SD, Standard Deviation; QTL, Quantitative Trait Loci; SNP, Single Nucleotide Polymorphism; RFLPs, Restriction Fragment Length Polymorphisms; AFLPs, Amplified Fragment Length Polymorphisms; SSRs, Simple Sequence Repeats; TD, Trunk Diameter; HT, Trunk Height; RL, Rachis Length; LDW, Leaf Dry Weight; FA, Foliar Area; LA, Leaf Area; LXL, Leaflet Per Leaf; BW, Bunch Weight; BN, Bunch Number.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publish

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare they have no competing interests.

Funding

Publication of this article has been funded by TV-17 supported by the Ministry of Agriculture and Rural Development of the Republic of Colombia. The funding entities had no role in study design, data collection and analysis, interpretation, decision to publish, or preparation of the manuscript.

Authors' contributions

LSB, FEER and SBP conceptualized and conceived the project and its components. SBP, LPD, GAGM, FEER, JAOG, and LPM collected the accessions and the trait data. JAOG and GAGM carried out the genotypic and phenotypic analysis with the supervision of OEC. JAOG and GAGM wrote the manuscript and LSB, OEC and FEER corrected and edited it. All authors reviewed and contributed to draft the manuscript as well as read and approved the final manuscript.

Acknowledgements

The authors would like to acknowledge William Tolosa for his support on sample collection, to Jhon Berdugo for his support in the data analysis, to Marco Antonio Lopez to provide the script for the LD heatmap analysis and Roxana Yockteng for her support in the neighbor-joining analysis. The authors thank Joanna Kelley for assistance in revising the final version of the manuscript. The access to accessions complies with the genetic resource agreement for scientific research without commercial interest No. 74, signed between Agrosavia and the Ministry of Environment and Sustainable Development from Colombia.

References

- Murphy D. Oil palm: Future prospects for yield and quality improvements. 2009.
- Pacheco P, Gnych S, Dermawan A, Komarudin H, Okarda B. The palm oil global value chain: Implications for economic growth and social and environmental sustainability. Bogor, Indonesia; 2017.
- Barcelos E, Rios S de A, Cunha RN V, Lopes R, Motoike SY, Babiychuk E, et al. Oil palm natural diversity and the potential for yield improvement. *Front Plant Sci.* 2015;6:190. doi:10.3389/fpls.2015.00190.
- Srestasathien P, Rakwatin P. Oil Palm tree detection with high resolution multi-spectral satellite imagery. *Remote Sens.* 2014;6:9749–74. doi:10.3390/rs6109749.
- Escobar R, Alvarado A. Estrategias para la producción comercial de semillas y clones de palmas de aceite compactas. *Rev Palmas.* 2004;25:293–305. <https://publicaciones.fedepalma.org/index.php/palmas/article/view/1093>.
- Turner PD. *Oil Palm Diseases and Disorders.* Oxford University Press; 1981. <https://books.google.com.co/books?id=mAnyXwAACAAJ>.
- Amblard P, Billotte N, Cochard B, Durand-Gasselín T, Jacquemard JC, Louise C, et al. El mejoramiento de la palma de aceite *Elaeis guineensis* y *Elaeis oleifera* por el Cirad-CP. *Rev Palmas.* 2002;306–10.
- Zambrano JE. Los híbridos interespecíficos *elaeis oleifera* HBK. x *Elaeis guineensis* Jacq.: una alternativa de renovación para la Zona Oriental de Colombia. *Rev Palmas.* 2004;25:339–49. <http://publicaciones.fedepalma.org/index.php/palmas/article/view/1098>.
- Chinchilla C. Tolerancia y resistencia a las pudriciones del cogollo en fuentes de diferente origen de *Elaeis guineensis*. *Rev Palmas.* 2007;28:273–84.
- Moura J. Manejo integrado das pragas das palmeiras. Ilheus, BA: Centro de Pesquisas do Cacau; 2017.
- Hartley CWS. *The oil palm (Elaeis guineensis Jacq.).* Second. 1967.
- Mayes S, Jack PL, Corley RH V, Marshall DF. Construction of a RFLP genetic linkage map for oil palm (*Elaeis guineensis* Jacq.). *Genome.* 1997;40:116–22.
- Purba AR, Noyer JL, Baudouin L, Perrier X, Hamon S, Lagoda P. A new aspect of genetic diversity of Indonesian oil palm (*Elaeis guineensis* Jacq.) revealed by isoenzyme and AFLP markers and its consequences for breeding. *Theor Appl Genet.* 2000;101:956–61. doi:10.1007/s001220051567.
- Jeenor S, Volkaert H. Mapping of quantitative trait loci (QTLs) for oil yield using SSRs and gene-based markers in African oil palm (*Elaeis guineensis* Jacq.). *Tree Genet genomes.* 2014;10:1–14.
- Billotte N, Marseillac N, Risterucci A-M, Adon B, Brottier P, Baurens F-C, et al. Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* Jacq.). *Theor Appl Genet.* 2005;110:754–65.
- Seng T-YY, Ritter E, Mohamed Saad SH, Leao L-JJ, Harminder Singh RS, Qamaruz Zaman F, et al. QTLs for oil yield components in an elite oil palm (*Elaeis guineensis*) cross. *Euphytica.* 2016;212:399–425. doi:10.1007/s10681-016-1771-6.
- Yadav P, Vaidya E, Rani R, Yadav N, Singh B, K. Rai P, et al. Recent perspective of next generation sequencing: applications in molecular plant biology and crop improvement. 2016.
- Nguyen K Le, Grondin A, Courtois B, Gantet P. Next-generation sequencing accelerates crop gene discovery. *Trends Plant Sci.* 2019;24:263–74. doi:<https://doi.org/10.1016/j.tplants.2018.11.008>.
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS One.* 2011;6:1–10. doi:10.1371/journal.pone.0019379.
- Pootakham W, Jomchai N, Ruang-arereate P, Shearman JR, Sonthirod C, Sangsrakru D, et al. Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics.* 2015;105:288–95. doi:<https://doi.org/10.1016/j.ygeno.2015.02.002>.
- Babu BK, Mathur RK, Ravichandran G, Venu MVB. Genome-wide association study (GWAS) for stem height increment in oil palm (*Elaeis guineensis*) germplasm using SNP markers. *Tree Genet Genomes.* 2019;15:1–8.
- Huang X, Han B. Natural variations and genome-wide association studies in crop plants. *Annu Rev Plant Biol.* 2014;65:531–51. doi:10.1146/annurev-arplant-050213-035715.
- Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods.* 2013;9:29. doi:10.1186/1746-4811-9-29.
- Burghardt LT, Young ND, Tiffin P. A guide to genome-wide association mapping in plants. *Curr Protoc Plant Biol.* 2017;2:22–38. doi:10.1002/cppb.20041.
- Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet.* 2010;42:355. <https://doi.org/10.1038/ng.546>.
- FAO - Trade and market division. *Oilcrops.* 2014. http://www.fao.org/fileadmin/templates/est/COMM_MARKETS_MONITORING/Oilcrops/Documents/Food_outlook_oilseeds/Food_Outlook_May_2014_OILCR
- Kurnia JC, Jangam S V., Akhtar S, Sasmito AP, Mujumdar AS. Advances in biofuel production from oil palm and palm oil processing wastes: A review. *Biofuel Res J.* 2016;3:332–46. doi:10.18331/BRJ2016.3.1.3.
- World Growth. The economic benefit of palm oil to Indonesia. *World Growth Palm Oil Green Dev Campaign.* 2011; February:1–27. http://worldgrowth.org/site/wp-content/uploads/2012/06/WG_Indonesian_Palm_Oil_Benefits_Report-2_11.pdf.
- Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, et al. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature.* 2012;485:635–41.
- Bastidas Perez S. Avances en el desarrollo de materiales genéticos resistentes a la PC. *Rev Palmas.* 2013;34:135–41.
- Breure CJ. Desarrollo de las hojas en la palma de aceite (*Elaeis guineensis*) y determinación de la tasa de apertura de las hojas. *Rev Palmas.* 1996;17. <https://publicaciones.fedepalma.org/index.php/palmas/article/view/563>.
- Bastidas S, Hurtado PYL. Evaluación de palmas prolíficas en la especie *Elaeis oleifera* e híbridos interespecíficos de *E. oleifera* x *E. guineensis*. 1993;55–60.
- Ithnin M, Xu Y, Marjuni M, Serdari NM, Amiruddin MD, Low E-TL, et al. Multiple locus genome-wide association studies for important economic traits of oil palm. *Tree Genet Genomes.* 2017;13:103. doi:10.1007/s11295-017-1185-1.
- Guo D-L, Zhao H-L, Li Q, Zhang G-H, Jiang J-F, Liu C-H, et al. Genome-wide association study of berry-related traits in grape [*Vitis vinifera* L.] based on genotyping-by-sequencing markers. *Hortic Res.* 2019;6:11. doi:10.1038/s41438-018-0089-z.
- Otto L-G, Mondal P, Brassac J, Preiss S, Degenhardt J, He S, et al. Use of genotyping-by-sequencing to determine the genetic structure in the medicinal plant chamomile, and to identify flowering time and alpha-bisabolol associated SNP-loci by genome-wide association mapping. *BMC Genomics.* 2017;18:599. doi:10.1186/s12864-017-3991-0.
- Hu X, Zuo J, Wang J, Liu L, Sun G, Li C, et al. Multi-locus genome-wide association studies for 14 main agronomic traits in Barley. *Front Plant Sci.* 2018;9:1683. doi:10.3389/fpls.2018.01683.
- Babu BK, Mathur RK, Ravichandran G, Anita P, Venu MVB. Genome wide association study (GWAS) and identification of candidate genes for yield and oil yield related traits in oil palm (*Elaeis guineensis*) using SNPs by genotyping-based sequencing. *Genomics.* 2019. doi:<https://doi.org/10.1016/j.ygeno.2019.06.018>.
- Bai B, Wang L, Lee M, Zhang Y, Rahmadsyah, Alfiko Y, et al. Genome-wide identification of markers for selecting higher oil content in oil palm. *BMC Plant Biol.* 2017;17:93. doi:10.1186/s12870-017-1045-z.
- Chung YS, Choi SC, Jun TH, Kim C. Genotyping-by-sequencing: a promising tool for plant genetics research and breeding. *Hortic Environ Biotechnol.* 2017;58:425–31.
- Hamilton ES, Schlegel AM, Haswell ES. United in diversity: mechanosensitive ion channels in plants. *Annu Rev Plant Biol.* 2015;66:113–37. doi:10.1146/annurev-arplant-043014-114700.
- Haswell ES, Peyronnet R, Barbier-Brygoo H, Meyerowitz EM, Frachisse JM. Two MscS homologs provide mechanosensitive channel activities in the *Arabidopsis* root. *Curr Biol.* 2008;18:730–4.
- Liu Q, Wang Z, Xu X, Zhang H, Li C. Genome-wide analysis of C2H2 zinc-finger family transcription factors and their responses to abiotic stresses in poplar (*Populus trichocarpa*). *PLoS One.* 2015;10:1–25.
- Botella JR. Can heterotrimeric G proteins help to feed the world? *Trends Plant Sci.* 2012;17:563–8. doi:10.1016/j.tplants.2012.06.002.
- Sangaev SS, Kochetov A V, Ibragimova SS, Levenko BA, Shumny VK. Physiological role of extracellular ribonucleases of higher plants. *Russ J Genet Appl Res.* 2011;1:44–50. doi:10.1134/S2079059711010060.
- Tvorus EK. Plant ribonucleases.

Sov plant Physiol. 1976; 46. Beavis W. QTL analyses: power, precision, and accuracy. In: Molecular dissection of complex traits. Chicago, IL, USA: American Seed Trade Association; 1998. p. 250–66. 47. Klein RJ. Power analysis for genome-wide association studies. *BMC Genet.* 2007;8:58. doi:10.1186/1471-2156-8-58. 48. Hong EP, Park JW. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inform.* 2012;10:117–22. doi:10.5808/GI.2012.10.2.117. 49. Hashimoto M, Negi J, Young J, Israelsson M, Schroeder JI, Iba K. Arabidopsis HT1 kinase controls stomatal movements in response to CO₂. *Nat Cell Biol.* 2006;8:391–7. doi:10.1038/ncb1387. 50. Ghosh JS, Chaudhuri S, Dey N, Pal A. Functional characterization of a serine-threonine protein kinase from *Bambusa balcooa* that implicates in cellulose overproduction and superior quality fiber formation. *BMC Plant Biol.* 2013;13:128. doi:10.1186/1471-2229-13-128. 51. Ma Y, Wang J, Zhong Y, Geng F, Cramer GR, Cheng Z-M (Max). Subfunctionalization of cation/proton antiporter 1 genes in grapevine in response to salt stress in different organs. *Hortic Res.* 2015;2:15031. <https://doi.org/10.1038/hortres.2015.31>. 52. Zang D, Li H, Xu H, Zhang W, Zhang Y, Shi X, et al. An Arabidopsis zinc finger protein increases abiotic stress tolerance by regulating sodium and potassium homeostasis, reactive oxygen species scavenging and osmotic potential. *Front Plant Sci.* 2016;7:1272. doi:10.3389/fpls.2016.01272. 53. Wu W, Cheng Z, Liu M, Yang X, Qiu D. C3HC4-type RING finger protein NbZFP1 is involved in growth and fruit development in *Nicotiana benthamiana*. *PLoS One.* 2014;9:e99352–e99352. doi:10.1371/journal.pone.0099352. 54. Zhou B, Lin JZ, Peng D, Yang YZ, Guo M, Tang DY, et al. Plant architecture and grain yield are regulated by the novel DHH-type zinc finger protein genes in rice (*Oryza sativa* L.). *Plant Sci.* 2017;254:12–21. doi:<https://doi.org/10.1016/j.plantsci.2016.08.015>. 55. P. SB, R. EAP, C. RR. Genealogía del germoplasma de palma de aceite (*Elaeis guineensis* Jacq.) del proyecto de mejoramiento genético de Corpoica. *Rev Palmas.* 2003;24. <https://publicaciones.fedepalma.org/index.php/palmas/article/view/950>. 56. Corley RH V, Hardon JJ, Tan GY. Analysis of growth of the oil palm (*Elaeis guineensis* Jacq.) I. Estimation of growth parameters and application in breeding. *Euphytica.* 1971;20:307–15. doi:10.1007/BF00056093. 57. Breure CJ. Factors associated with the allocation of carbohydrates to bunch dry matter production in oil palm (*Elaeis guineensis* Jacq.). *Landbouwniversiteit*; 1987. 58. R development core team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2008. <http://www.r-project.org>. 59. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics.* 2007;23:2633–5. doi:10.1093/bioinformatics/btm308. 60. Singh R, Ong-Abdullah M, Low E-TL, Manaf MAA, Rosli R, Nookiah R, et al. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature.* 2013;500:335. <https://doi.org/10.1038/nature12309>. 61. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10:R25. doi:10.1186/gb-2009-10-3-r25. 62. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27:2156–8. doi:10.1093/bioinformatics/btr330. 63. Rambaut A. FigTree: tree figure drawing tool version 1.4.2. 2014. <http://tree.bio.ed.ac.uk/software/figtree>. 64. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A High-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics.* 2012;28:3326–8. doi:10.1093/bioinformatics/bts606. 65. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: genome association and prediction integrated tool. *Bioinformatics.* 2012;28:2397–9. doi:10.1093/bioinformatics/bts444. 66. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc.* 1995;57:289–300.

Tables

Table 1. Mean values, standard deviations (SD) and minimum and maximum values from the phenotypic traits used in this study

Category	Trait	Abbreviation	Unit	Mean	SD	Minimum value	Maximum value
Morphological	Trunk Diameter	TD	cm	88.5	6.0	62.4	102
	Trunk Height	HT	cm	250.3	29.5	133.3	327
	Rachis length	RL	cm	421.5	35.3	275.5	530
	Leaf dry weight	LDW	kg	2.2	0.3	1.3	3.7
	Foliar area	FA	m ²	385	78.2	141.3	617.1
	Leaf area	LA	m ²	8.6	1.3	4.7	12.7
Yield	Leaflet per leaf	LXL	unit	234.8	14.8	184	294
	Bunch weight	BW	kg	6.1	1.8	1	19.5
	Bunch number	BN	unit	8.8	5.0	1.0	27
	Yield per palm	Yield	kg	56.5	39.1	1.8	233

Table 2. Significant marker-trait associations for the 378 OxG population for morphological and yield-related traits using a mixed linear model approach.

Category	Trait	SNP	Chromosome	Position (kb)	p-value	MAF	R ²	FDR Adjusted p-values	Nearby Gene	SNP position relative to the candidate gene	Candidate gene annotation
Morphological	LDW	S3_30467222	3	30467222	1.E-04	0.106	0.107	0.050	p5.00_sc00100_p0017	0	Mechanosensitive ion channel protein 10-like (MSL10)
	TD	S15_21239833	15	21239833	1.98E-05	0.493	0.097	0.010	p5.00_sc00036_p0097	+21.8 kb	Nucleic acid binding
	HT TD	S15_22347191	15	22347191	7.94E-05	0.496	0.098	0.084	p5.00_sc00036_p0145	+0.3 kb	Paired amphipathic helix protein (PAH)
	FA LA	S15_22553489	15	22553489	7.94E-05	0.496	0.098	0.084	p5.00_sc00036_p0152	-0.6 kb	Serine threonine-protein kinase (STYK)
		S15_22553493	15	22553493	8.90E-05	0.495	0.098	0.084			
		S15_23645020	15	23645020	7.94E-05	0.496	0.098	0.084	p5.00_sc00036_p0217	0	Class E vacuolar protein-sorting machinery protein (VPS)
	RL LXL	S13_20856724	13	20856724	3.22E-05	0.499	0.074	0.041	p5.00_sc00035_p0180	-12.0 kb	Guanine nucleotide-binding protein subunit gamma (AGC3)
		S13_23674227	13	23674227	3.22E-05	0.499	0.074	0.041	p5.00_sc00035_p0078	0	Extracellular ribonuclease (RNase)
		S13_25522088	13	25522088	3.22E-05	0.499	0.074	0.041	p5.00_sc00128_p0001	0	Imbr1 domain-containing protein (IMBR1)
		S13_24474516	13	24474516	7.14E-05	0.497	0.070	0.067	p5.00_sc00035_p0043	0	Probable ran guanine nucleotide release factor-like (RANGRF)
Production	Yield BN	S5_41396842	5	41396842	1.E-04	0.132	0.059	0.410	p5.00_sc00003_p0367	+29 kb	Cation h(+) antiporter
	BW	S10_21597426	10	21597426	3.E-04	0.499	0.054	0.316	p5.00_sc00036_p0097	-8.0 kb	Zinc finger protein 8-like (ZFP)

* SNP position relative to the closest candidate gene: SNPs upstream and downstream of candidate genes are specified with "-" and "+", respectively. 0 indicates that SNPs are located within candidate gene.

Additional Files

Additional File Table S1: List of oil palm OxG and parental lines accessions used in the present study.

Additional File Table S2: SNPs identified per chromosome using *E. guineensis* as reference genome.

Additional File Figure S1: Box plots of the two cluster groups for all morphological and yield-related traits. * = significant at $p \leq 0.0001$, ns = non-significant.

Additional File Figure S2: Linkage disequilibrium (LD) heatmap for each chromosome with significant associated SNPs in an OxG population.

Figures

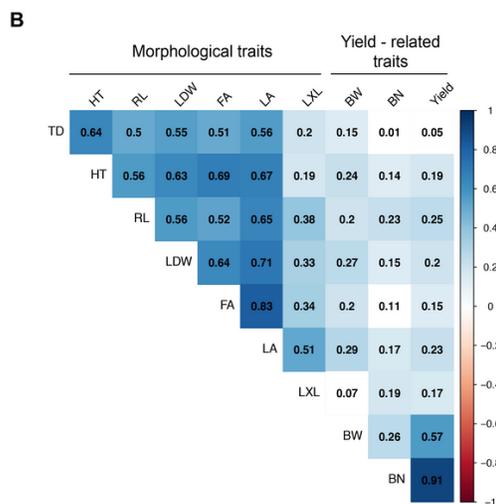
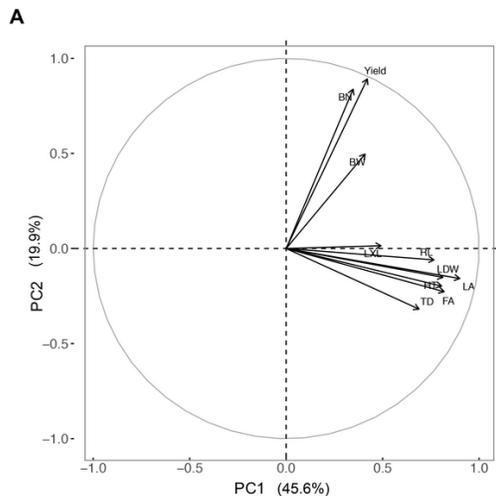


Figure 1

A. Principal component analysis for 10 oil palm morphological and yield-related traits B. Correlation among traits of the whole population. TD = Trunk Diameter, HT = Trunk Height, RL = Rachis Length, LDW = Leaf Dry Weight, FA = Foliar Area, LA = Leaf Area, LXL = Leaflet per Leaf, BW = Bunch Weight, BN = Bunch number and Yield = Yield per Palm). Color boxes indicate significant correlations ($p \leq 0.01$), while white boxes indicate coefficients with a $p > 0.01$.

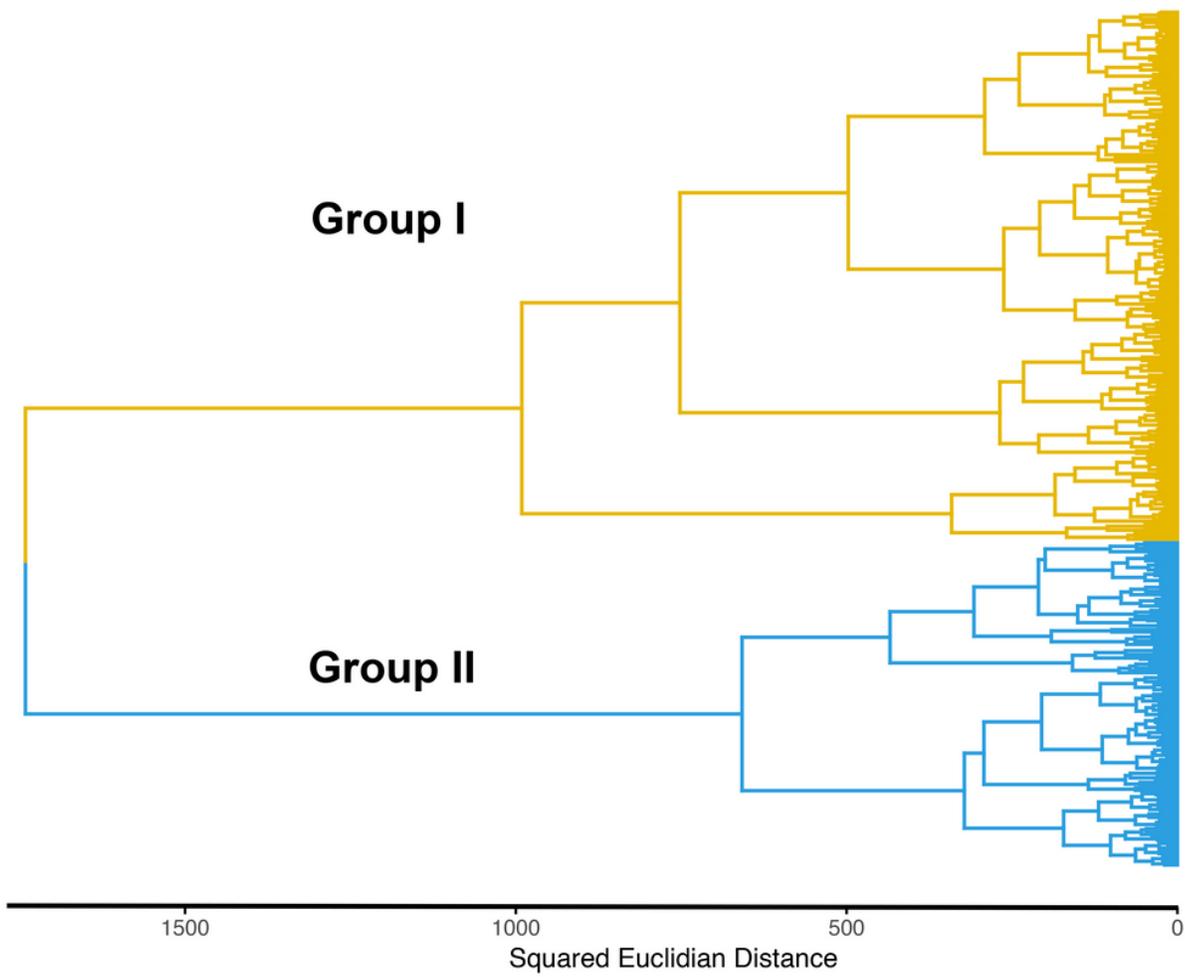


Figure 2

Hierarchical cluster analysis of the 378 OxG population. Accessions were clustered using the Ward's method and the squared Euclidean distance.

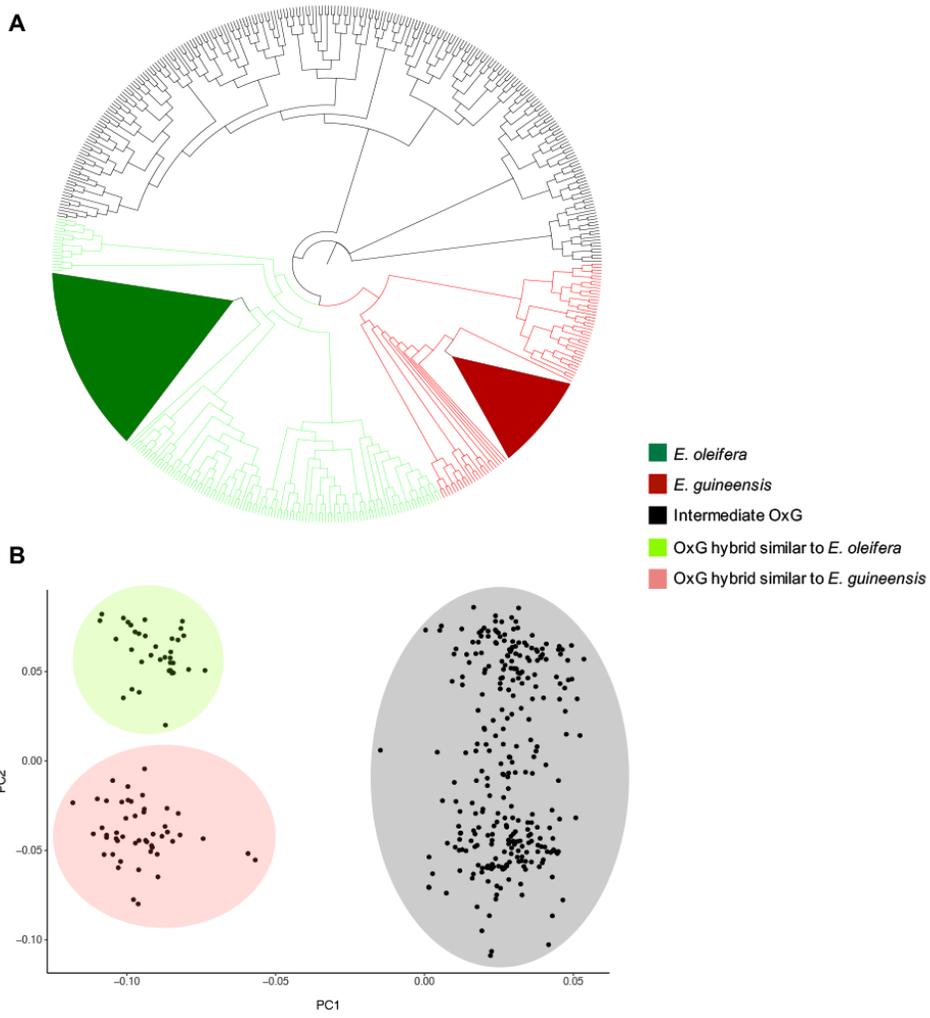


Figure 3

A. Neighbor-Joining tree of the 471 accessions based on Nei's genetic distance of oil palm, B. Principal component analysis (PCA) of the OxG population separating it into two groups.

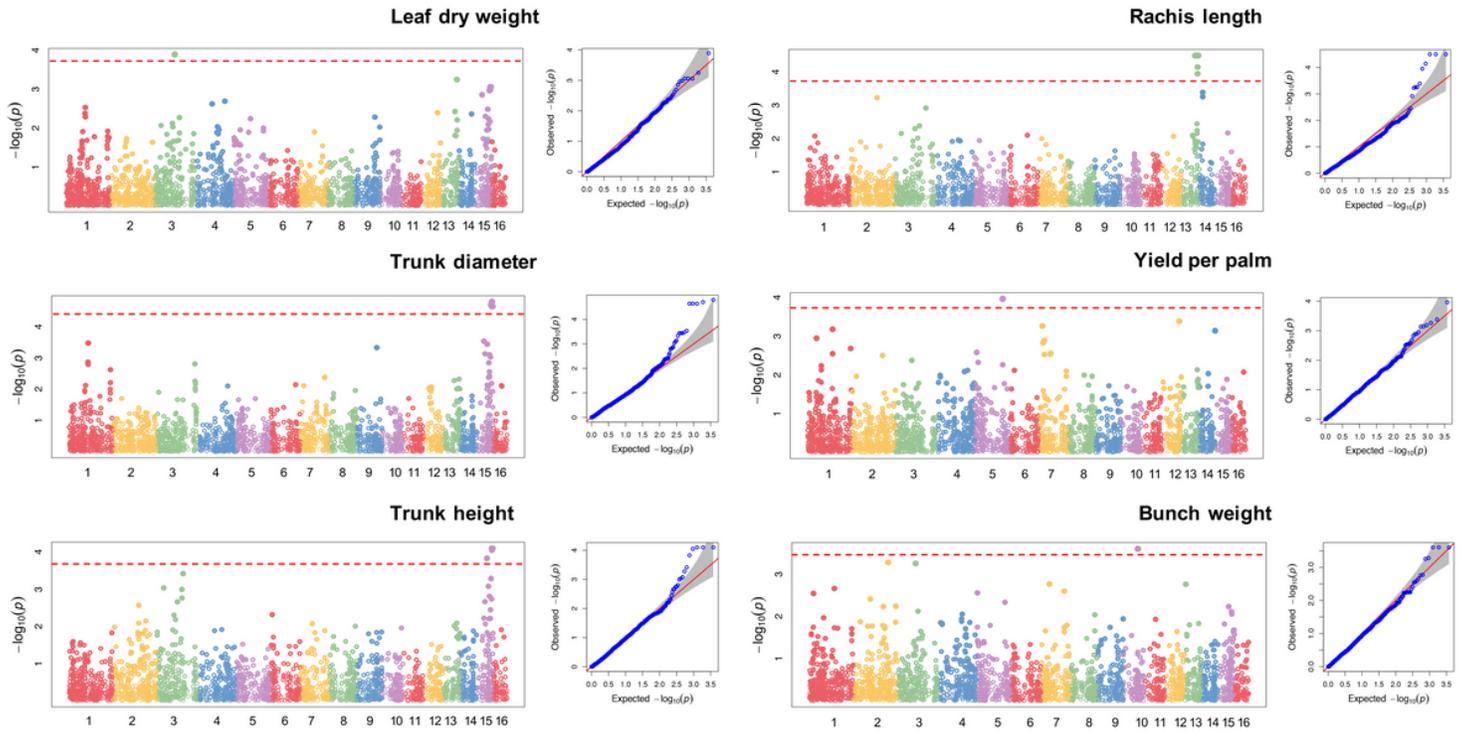


Figure 4

Manhattan and Q-Q plots indicating genomic regions associated with leaf dry weight (LDW), trunk diameter (TD), trunk height (HT), rachis length (RL), yield per palm and bunch weight (BW). The red horizontal line indicates the significant association threshold.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [FigureS2.jpg](#)
- [FigureS1.jpg](#)
- [TableS1.xlsx](#)
- [TableS2.xlsx](#)