Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com

Research Article

Genomic diversity and genome-wide association analysis related to yield and fatty acid composition of wild American oil palm

Maizura Ithnin [a, 1], Wendy T. Vu [b, 1], Min-Gyoung Shin [b], Vasantika Suryawanshi [b], Katrina Sherbina [b], Siti Hazirah Zolkafli [a], Norhalida Mohamed Serdari [a], Mohd Din Amiruddin [a], Norziha Abdullah [a], Suzana Mustaffa [a], Marhalil Marjuni [a], Rajanaidu Nookiah [a], Ahmad Kushairi [a], Paul Marjoram [c], Sergey V. Nuzhdin [b], Peter L. Chang [b, *], Rajinder Singh [a, *]

[a] Malaysian Palm Oil Board, 6, Persiaran Institusi, Bandar Baru Bangi, 43000 Kajang, Selangor, Malaysia
[b] Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA
[c] Department of Preventative Medicine, University of Southern California, Los Angeles, CA 90089, USA

ARTICLE INFO

Keywords:
E. oleifera
SNP discovery
Genome diversity
Genome-wide association

ABSTRACT

Existing *Elaeis guineensis* cultivars lack sufficient genetic diversity due to extensive breeding. Harnessing variation in wild crop relatives is necessary to expand the breadth of agronomically valuable traits. Using RAD sequencing, we examine the natural diversity of wild American oil palm populations *(Elaeis oleifera)*, a sister species of the cultivated *Elaeis guineensis* oil palm. We genotyped 192 wild *E. oleifera* palms collected from seven Latin American countries along with four cultivated *E. guineensis* palms. Honduras, Costa Rica, Panama and Colombia palms are panmictic and genetically similar. Genomic patterns of diversity suggest that these populations likely originated from the Amazon Basin. Despite evidence of a genetic bottleneck and high inbreeding observed in these populations, there is considerable genetic and phenotypic variation for agronomically valuable traits. Genome-wide association revealed several candidate genes associated with fatty acid composition along with vegetative and yield-related traits. These observations provide valuable insight into the geographic distribution of diversity, phenotypic variation and its genetic architecture that will guide choices of wild genotypes for crop improvement.

1. Introduction

Oil palm (*Elaeis guineensis*) is the most efficient vegetable oil crop in the world, producing four to ten times more metric tons of oil per hectare compared to other oil crops [1]. Palm oil is one of the healthiest, versatile and widely used vegetable oils, representing 40 % of total worldwide consumption [2] and utilized in food, cosmetic and biofuel industries. However, concerns over the conservation and sustainability of oil palm cultivation [3,4] are growing due to climate change and increasing global demand. Thus, breeding oil palm cultivars with traits that increase yield, nutritional value, and adaptation to extreme conditions while reducing land-use requirements is much needed. Wild germplasm collections are a valuable source of genetic variation necessary to expand the range of desirable traits within cultivated oil palm varieties [5].

The discovery and characterization of genes that underlie novel and desirable traits require analysis of genetic variation in wild populations of crop relatives. Such programs have been implemented in a variety of cultivar species, including tomato, rice, potato, wheat, chickpea and sunflower [6–9]. So far, genetic variation from wild germplasm has been underutilized in oil palm breeding programs. Our study aims to identify alleles associated with agriculturally relevant traits in a collection of wild germplasm from a reproductively compatible sister species of cultivated oil palm.

Oil palm belongs to the genus *Elaeis*, which consist of two species. *E. guineensis* found naturally in Central Africa has been the long-standing oil palm of commerce in Asia, Latin America and Africa while *E. oleifera* inhabits wild groves in Central and South America. Despite being up to 50 million years diverged [10], the two species produce viable offsprings. Wild *E. oleifera,* also known as American oil palm, ex-

[*] Corresponding authors.
Email addresses: peter.chang@usc.edu (P.L. Chang); rajinder@mpob.gov.my (R. Singh)
[1] Maizura Ithnin and Wendy T. Vu should be considered joint first author.

https://doi.org/10.1016/j.plantsci.2020.110731
Received 30 May 2020; Received in revised form 13 October 2020; Accepted 15 October 2020
Available online xxx
0168-9452/ © 2020.

hibits favorable and distinct traits not found in the commercial *E. guineensis,* including tolerance to diseases such as bud rot and lethal wilt [5]. The American oil palm also produce oil with higher levels of unsaturated fatty acids, carotene, vitamin E and sterol contents [11]. While American oil palm show broad environmental adaptability [12], it suffers from extremely low oil yield (0.5 t ha^{-1} yr^{-1}) compared to the African oil palm which produce 3–4 t ha^{-1} yr^{-1} [11]. Therefore, interspecific hybrids have been developed to introduce traits of *E. oleifera* into cultivated *E. guineensis*.

The discovery of quantitative trait loci (QTLs) associated with traits of interest and identification of favorable alleles unnoticed or undetectable previously allow more efficient marker-assisted-selection (MAS) [13]. Considering the long breeding cycle of oil palm (10 years) [14], any undesirable genotype can be identified and removed at early developmental stages, reducing the time and cost for phenotyping, especially for traits that are expressed later in development. This, in turn, results in more efficient use of land space. With MAS, the duration to develop new oil palm planting materials can be shortened by half needed through conventional breeding methods.

American oil palms are widely distributed across Central America and in the northern regions of South America extending into the Upper Amazonian Basin. Comprehensive collection of wild *E. oleifera* seeds began in 1967 in Costa Rica, Panamá, Colombia, Suriname, Honduras, Perú, Ecuador, Nicaragua and the Amazon basin of Brazil [15,16]. We used a genotype-by-sequencing approach (RAD-seq) to characterize this panel of wild *E. oleifera* germplasm along with four cultivated *E. guineensis* palms. We assessed the patterns of genetic diversity and implemented genome-wide association (GWAS) on a panmictic population encompassing Honduras, Costa Rica, Panama, and Colombia. Our GWAS analysis utilized not only a single-locus model but also multi-loci and multi-trait models [17,18], focusing on different aspects of genotypic and phenotypic characteristics and combining different model analyses which increases the power to detect various types of causal loci. We then identified several candidate genes associated with fatty-acid composition as well as vegetative and yield-related traits.

2. Materials and methods

2.1. Sampling wild material and phenotype data collection

The wild germplasm collected from South America were field-planted in four different experimental plots at the MPOB research station in Kluang, Johore, Malaysia. The field trials for genetic material from Brazil, Honduras, Colombia, Panama and Costa Rica were laid down in two experimental plots, respectively in 1984 and 1986, using Completely Randomized Design. Genetic material from Peru were field-planted in 2006 in Progeny Row; those from Ecuador were laid down in 2009 in Completely Randomized Block. For the current study, depending on palm availability in the field plots, one to eight palms were sampled per population. We analyzed a total of 192 palms from 37 populations (Table S1A). All field trials were established at the MPOB Research Station located at Kluang, Johor, Malaysia. Oil palm field data collection began 36 months after planting. From the third to eighth year, 3–5 individual fruit bunches were sampled per palm and analyzed in the laboratory to measure bunch traits and fatty acid composition (FAC).

2.2. Bunch Analysis and trait measurements

Bunch analysis [19] was used to record 18 bunch traits. Measurements of vegetative traits [20] were carried out on frond number 17 harvested from eight-year-old palms. Palm height was also measured from ground level to the base of frond number 41. FAC was characterized by a method described by [21]. Oil extracted from ripe fruits was analyzed using Perkin-Elmer Gas Chromatography System. The results obtained in the form of peaks for individual fatty acids are converted into percentage using a data integrator available in the analysis software provided by Perkin-Elmer.

2.3. DNA sequencing

DNA extraction was carried out on 3 g of leaf samples [22]. DNA concentration was quantified using a NanoDrop Spectrophotometer (NanoDrop Technologies Inc.) and stored at 4 °C until needed. Samples were sequenced using a set of 96 barcoded adapters custom designed using a barcode generator [23] with *HindIII* restriction site overhangs. The common adaptor was designed with a *NlaIII* restriction site overhang. Each 96-multiplex library was sequenced across two lanes on an Illumina HiSeq 2000. All sequencing data can be found in the NCBI BioProject PRJNA434010.

Illumina reads were mapped to the *Elaeis guineensis* 9.1 reference (EG9, an updated version of the oil palm genome assembly [10]) using BWA MEM 0.7.9a-r786 [24]. Variants were called using the GATK pipeline through the HaplotypeCaller program, identifying 5,052,463 single nucleotide polymorphisms (SNPs) segregating among the 196 samples. This set was reduced to 3,649,035 SNPs when only using genotype calls supported by at least 8 reads. Aside from 15 outlier samples, there were 102,189 species-informative SNPs between *E. oleifera* and cultivated *E. guineensis*, defined as loci where the frequency of the major allele in *E. oleifera* samples was at least 0.97 and different from the allele called in the *E. guineensis* draft reference along with the 4 *E. guineensis* GBS samples. Diversity and phylogenetic analyses were based on a set of 15,404 SNPs under Hardy Weinberg equilibrium ($p < 0.05$) with minor allele frequency (MAF) of at least 0.03 and whose genotypes were called in at least 90 % of the 196 individuals. For association studies, only a subset of samples with phenotypic data were included. Association analysis for fatty acid traits was carried out on 77 samples from Honduras, Costa Rica, Panama, and Colombia using a set of 2439 SNPs with MAF of at least 0.03 where the genotypes were successfully called in 90 % of samples analysed. Association analysis for vegetative traits was carried out on 144 samples using a set of 2272 SNPs filtered using the same criteria as above.

2.4. Population structure

STRUCTURE [25] was used to assess admixture and population structure using the admixture and correlated allele frequency model. Ten independent runs of 10,000 burn-in MCMC iterations followed by 50,000 iterations were performed for 2–8 clusters (K = 2–8). Nei's genetic distance was calculated using the 'poppr' R package [26], and hierarchical clustering was implemented using Ward's method [27] with the 'fastcluster' R package [28]. Principal component analysis was implemented using the 'SNPRelate' R package [29].

2.5. Diversity and phylogenetic analysis

Fst, Tajima's D, theta pi and theta W estimates were calculated in vcftools using window sizes of 1Mb [30]. Phylogenetic trees were generated using SNPhylo [31]. SNPhylo was also run on a larger set of SNPs with a minimum 50 % individual call rate, as missing data could be driven by high divergence from the reference genome. Results were extremely similar, and topology was identical, indicating that divergence between samples and the reference did not reduce the accuracy of phylogenetic relationships and that including missing data did not result in higher resolution of evolutionary relationships.

2.6. Linkage disequilibrium

Linkage disequilibrium was estimated by calculating pairwise correlation coefficient (r^2) values between all SNP pairs with a minimum distance of 100bp to minimize effects of physically linked SNPs. A non-linear model based on the Hill and Weir formula [32] was used to fit the decay rate of r^2 as a function of physical distance.

2.7. Genome-wide association study (GWAS) : Linear mixed models

Three separate models were implemented to identify trait associated loci. First, a single-locus linear mixed model was implemented in FaST-LMM [17]. Second, a multi-phenotype association model was implemented in GEMMA to identify pleiotropic loci associated with multiple correlated phenotype clusters [33]. Five vegetative trait clusters and three FAC clusters were analyzed by GEMMA. Finally, a multi-locus linear mixed model was used to identify associated markers, using both forward inclusion and backward exclusion steps. Only steps that had covariates that passed the Bonferroni threshold were considered. When there were more than one valid step, the step with the maximum number of covariates was chosen. The model was implemented in MLMM as an R package [34]. A kinship matrix calculated based on all SNPs was incorporated into each model as a random covariate to minimize false positives caused by population structure in the data. A false-discovery rate of 0.05 was used for the single-locus linear mixed model. The Bonferroni-corrected threshold of 0.05 was used for the remaining models as proposed by the programs. Since the multi-locus mixed model requires that no markers are missing, imputation was done using Beagle version 4.1 [35].

2.8. Validating allelic calls of significant trait-associated SNPs

We selected one trait-associated SNP (14:6491498) to confirm its polymorphisms. Detailed information of the SNP and samples used for validation are presented in Table S1B. Forward and reverse primers were designed to amplify targeted regions containing the SNP. To design primers, five thousand nucleotides upstream and downstream of the SNP position were retrieved from EG9 reference. Sequences were analyzed in the Primer 3 software to design forward and reverse primers. To ensure specificity, the primers were checked using Primer-BLAST. PCR products were separated in 1% agarose gels and purified using QIAGEN gel extraction kit. The purified fragments were cloned using TOPO-TA Cloning. For each palm, two to four colonies were picked and inoculated individually in LB broth. Plasmids were purified using QIAGEN plasmid isolation kit, and Sanger sequenced using reverse and forward primers. Each sequence was mapped to the reference build to confirm location. Sequences that mapped to correct position were aligned and alleles were scored. In addition, alleles were also scored from the sequence chromatograms.

2.9. Homologous gene annotation

Gene sequences were retrieved from the EG9 reference. Retrieved sequences were aligned to the *Arabidopsis thaliana* TAIR10 reference using BLASTX. Homologous genes with the lowest E-values were retrieved.

2.10. Strategic conservation of core Wild American germplasm

Palms from Honduras, Costa Rica, Panama and Colombia were analyzed with the intention of identifying a subset of the germplasm representing 90 % of the total genetic diversity observed in the population. N individuals were randomly drawn across different sample sizes (N = 10–154 individuals) from a total of 154 sequenced palms. For each sample size, the average fraction of polymorphic sites (number of polymorphic loci/total number of loci) across 100 bootstrapped simulations was calculated as a measure of genetic diversity. A genetic diversity index was calculated by taking the ratio of this diversity to the total genetic diversity in the full set of 154 individuals (mean fraction of polymorphic sites of sample size N / fraction of polymorphic sites N = 154).

3. Results

3.1. Evaluation of phenotypic data

To examine phenotypic variation in Central American oil palms, we analyzed 29 different traits across four different countries (Table S2A). Oil palms exhibited differences in fatty acid content (Kruskal-Wallis Test $P < 0.005$), such as for C16:0 where Honduras samples had the highest values (median: 22.1) while having lowest values recorded for C18:1 (median: 48.5). On the other hand, Colombia showed the opposite trend and had the highest value for C18:1 (median: 59.9) and lowest for C16:0 (median: 17.2). We assessed the relationship of all traits using Pearson correlation estimates and illustrated these relationships using hierarchical clustering (Fig. S1). Vegetative traits showed various degrees of correlation and five trait clusters were identified with strong pairwise correlations (> 0.6). Fatty acid traits were represented as three clusters.

3.2. SNP variation

We genotyped a total of 196 palms, including four cultivated *E. guineensis* and 192 wild *E. oleifera* palms sampled from seven countries in Central and South America: Honduras (26), Costa Rica (45), Panamá (78), Colombia (20), Brazil (7), Ecuador (8) and Peru (8). After removing genotype calls supported by less than 8 reads and markers with less than a 90 % call rate, 15,404 SNPs were identified for further population diversity analysis, resulting in an average SNP density of 1 SNP per 180 kb (Fig. S2). Of 64,085 annotated genes, 2122 were tagged with at least one SNP within the gene, and 20,312 and 21,066 genes had at least 1 SNP within 10 kb upstream and downstream, respectively (Fig. S3 shows the distribution of SNPs based on functional annotations across the genome). Aside from the 15 outlier samples, there were 102,189 species-informative SNPs between *E. oleifera* and cultivated *E. guineensis* (Appendix B).

3.3. Population structure and geographic differentiation

To assess the structuring of wild American populations, we estimated genetic relatedness and population divergence among the 196 palms. Principal component analysis (PCA) revealed five distinct clusters (Fig. 1A), with PC1 and PC2 explaining 40 % and 23 % of the genetic variance, respectively. Although palms from Ecuador and Peru clustered strongly in the PCA when all countries were analyzed together, they remained distinct groups when only the two countries were analyzed (Fig. S4), suggesting that these two populations have some minimal genetic structure. Palms from Honduras, Costa Rica, Panama, and Colombia clustered tightly together, except for 15 outlier palms. Model-based clustering revealed that these outlier palms showing patterns of mixed ancestry (Fig. 1C), distributed along PC2 in the PCA and positioned intermediate between cultivated *E. guineensis* and *E. oleifera* palms (Fig. 1A). Based on species-informative SNPs, these outliers possessed interspecific hybrid genotypes that encompass a combination of *E. guineensis* and *E. oleifera* alleles (Fig. S5). When excluding the 15 outlier palms, populations from Honduras, Costa Rica, Panama and Colombia show no apparent genetic structure, and F_{ST} be-

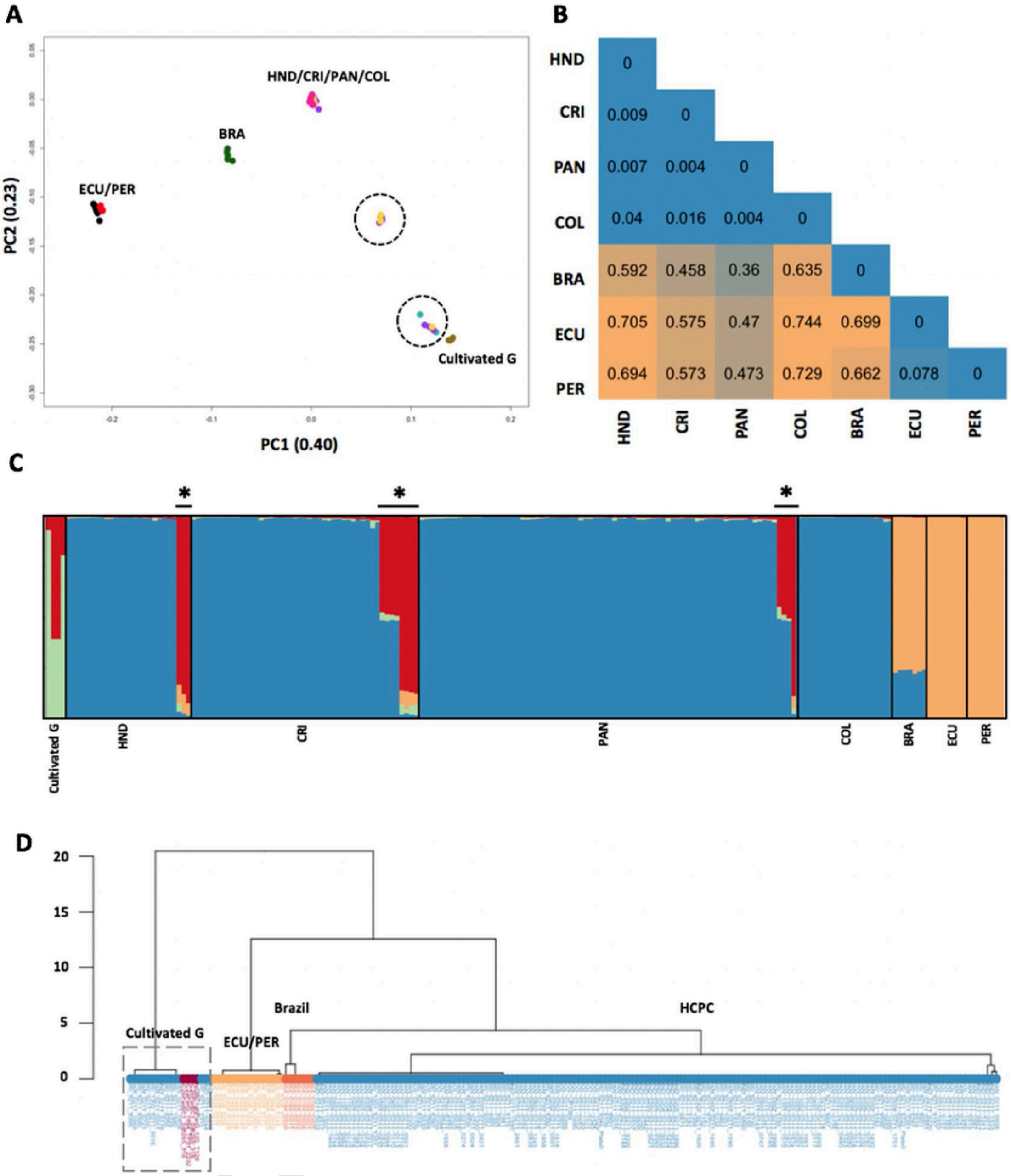

Fig. 1. Genetic structure and divergence of wild American oil palm populations from Honduras (HON), Costa Rica (CRI), Panama (PAN), Colombia (COL), Brazil (BRA), Ecuador (ECU) and Peru (PER). A) PCA, B) FST, C) STRUCTURE plot. Asterisks indicate 15 outlier palms sampled from Honduras, Costa Rica, and Panama. D) Hierarchical clustering of Nei's genetic distance: blue corresponds to palms sampled in HON/CRI/PAN/COL; yellow, ECU/PER; orange, BRA; red, cultivated E. guineensis palms. Dotted circles and boxes contain the 15 outlier palms from Honduras, Costa Rica and Panama.

tween these populations is substantially low (Fig. 1B,C). Therefore, palms from Honduras, Costa Rica, Panama, and Colombia were considered as a single population group (HCPC). Hierarchical clustering of Nei's genetic distance showed that population differentiation was based on geographic distance, where HCPC palms are most genetically similar to Brazil followed by Ecuador/Peru and least similar to cultivated *E. guineensis* (Fig. 1D).

3.4. Genomic patterns of diversity

To investigate the distribution of genetic variation and genomic signatures of diversification in wild American oil palm populations, we quantified heterozygosity, allelic richness and genome-wide nucleotide variation. Since more individuals from HCPC populations were sampled and available for sequencing relative to Brazil (7), Ecuador (8) and Peru (8), there was more variation found here, with 10,030, 372 and 1093 polymorphic sites identified within HCPC, Brazil, and Ecuador/Peru populations, respectively. Despite the smaller sample size of South American palms, palms from Brazil, Ecuador, and Peru have higher levels of heterozygosity and allelic diversity at the individual level compared to HCPC palms, suggesting that these populations are more diverse (Table 1) and should be prioritized for future studies.

There is an excess of rare alleles segregating in HCPC, with singletons representing more than 40 % of all SNPs (i.e., only one individual carries the minor allele), and 33 % of SNPs having minor allele frequencies (MAF) between 0.01 and 0.05. On average, genome-wide LD decays to a baseline $r^2 = 0.20$ and $r^2 = 0.10$ within 4 kb and 10 kb, respectively (Fig. S6). Overall average Tajima's D is negative for the

Table 1
Genetic diversity estimates in Elaeis oleifera.

Country	N	H_O	H_E	F	N_A
HND	23	0.014	0.017	0.86	1.017
CRI	37	0.022	0.024	0.82	1.024
PAN	74	0.021	0.025	0.81	1.025
COL	20	0.017	0.021	0.84	1.020
BRA	7	0.033	0.034	0.49	1.033
TAI	8	0.040	0.042	0.54	1.041
PER	8	0.056	0.057	0.30	1.056

HCPC population (-1.8), with an average nucleotide diversity of $\theta\pi = 3.35e^{-07}$ and $\theta_W = 2.76e^{-06}$, in which θ_W being significantly larger than $\theta\pi$ reflects the excess of rare alleles segregating in these populations. The excess of rare allele frequency indicates either population expansion after a bottleneck or positive selection. The shape of the maximum-likelihood phylogenetic tree built from the HCPC populations indicates patterns of population admixture (Fig. 2B). Furthermore, F values indicate that HCPC populations show a higher level of inbreeding relative to the other populations (Table 1). Combined with field observations that note extreme population fragmentation within the HCPC populations, these results indicate that the genetic patterns observed here likely reflect a genetic bottlene ck followed by population expansion and admixture.

3.5. GWAS

We identified SNPs associated with fatty acid composition (FAC) and vegetative and reproductive traits in the panmictic HCPC populations (Fig. 3D,E, Fig. S7). In addition to the single-marker linear mixed model (LMM), we also implemented a multi-variate LMM and multi-loci LMM to increase the statistical power of detecting SNPs associated with polygenic traits and SNPs that are pleiotropic. We identified 8 loci associated with vegetative traits, 2 loci associated with reproductive traits and 23 loci associated with fatty acid traits (Table S2B). Here, we also presented the locations of the loci and the effect of the significant SNPs on their associated traits. The phenotypic variance explained by correlated SNPs of FAC was typically larger than those from vegetative and yield traits. Most of the trait-associated SNPs were found in intergenic regions, while eight SNPs with synonymous mutations were found within genic regions.

We also identified several pleiotropic SNPs that affect multiple correlated phenotypes, suggesting that these traits share common genes and/or biological pathways. For instance, SNP 1:28380657 is associated with highly correlated leaf area (LA), leaf area index (LAI) and F traits ($r > 0.6$), while SNP 10:17957276 is associated with both mean bunch number (MBNO) and mean fresh fruit bunch (MFFB, $r = 0.9$). For the case of FAC, certain saturated and unsaturated fatty acids may be genetically correlated by sharing the same causal markers. Positively correlated saturated (C14:0, C16:0) and unsaturated (C18:2, C18:3) fatty acids share four common SNPs (3:34413159, 4:140445547, 5:85517312, 13:27681245). Furthermore, SNP 3:34413159 is associated with two sets of correlated fatty acid phenotypes with different saturation types (C18:1/IV and C14:0/C16:0/C18:2/C18:3, Fig. S7C). From the single-locus LMM and the multi-trait LMM, we found eight common SNPs associated with both C18:1 and iodine value (IV), a measure of the degree of unsaturation in oil. As expected, the most abundant unsaturated fatty acid C18:1, accounting for 40 % of the total fatty acid composition, is positively correlated with IV, as it is a major contributor to oil unsaturation level in oil palm collections. The negative correlation between C18:1 and C18:2 suggests that any increase in C18:1 would likely spill over into C18:2. Based on the pattern of FAC correlations, these eight common SNPs are likely associated with genes that primarily regulate C18:1 and thus drive changes in IV as a consequence. Furthermore, SNP 3:34413159 is associated with unsaturated fatty acids C18:1, C18: 2, and C18: 3 as well as the level of unsaturation, IV. The same SNP also influences the saturated fatty acids (C14:0 and C16:0), which are negatively correlated with C18:1, a key driver of the IV level. As one of the main objectives of interspecific hybrid breeding is to increase the level of unsaturation (IV), the appropriate SNP genotype could be used to select for palms with higher IV at the expense of saturated fatty acids.

3.6. Population differentiation of yield traits and associated markers

While genetic data suggests that HCPC populations are panmictic and unstructured, there is some phenotypic and genetic variation within these countries. The distributions of five traits, MFFB, MBNO, C16:0, C18:1, and C18:2, differed between countries (Kruskal-Wallis Test, $P < 0.05$, Table S2A). MFFB- and MBNO-associated marker frequencies differ among HCPC populations, reflecting potential adaptive divergence. Consistent with previous observations [11], Panama palms had the highest median values for both MFFB and MBNO (Table S2A, Fig. 4A), with two markers showing a significant excess of G and T alleles at SNPs 3:82372832 and 10:17957276, respectively (chi-square test, $P < 0.0005$). The presence of these alleles increase the trait values

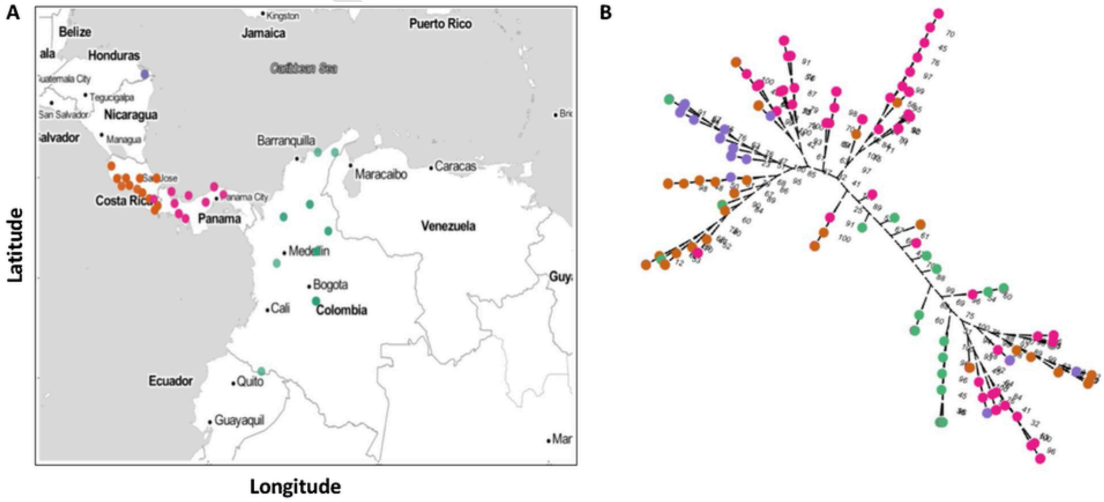

Fig. 2. Distribution and phylogenetic relationships of wild HCPC populations. A) Geographical distribution of sampled palms across Honduras (purple), Costa Rica (orange), Panama (pink), and Colombia (green). B) Unrooted neighbor-joining MLE phylogenetic tree.

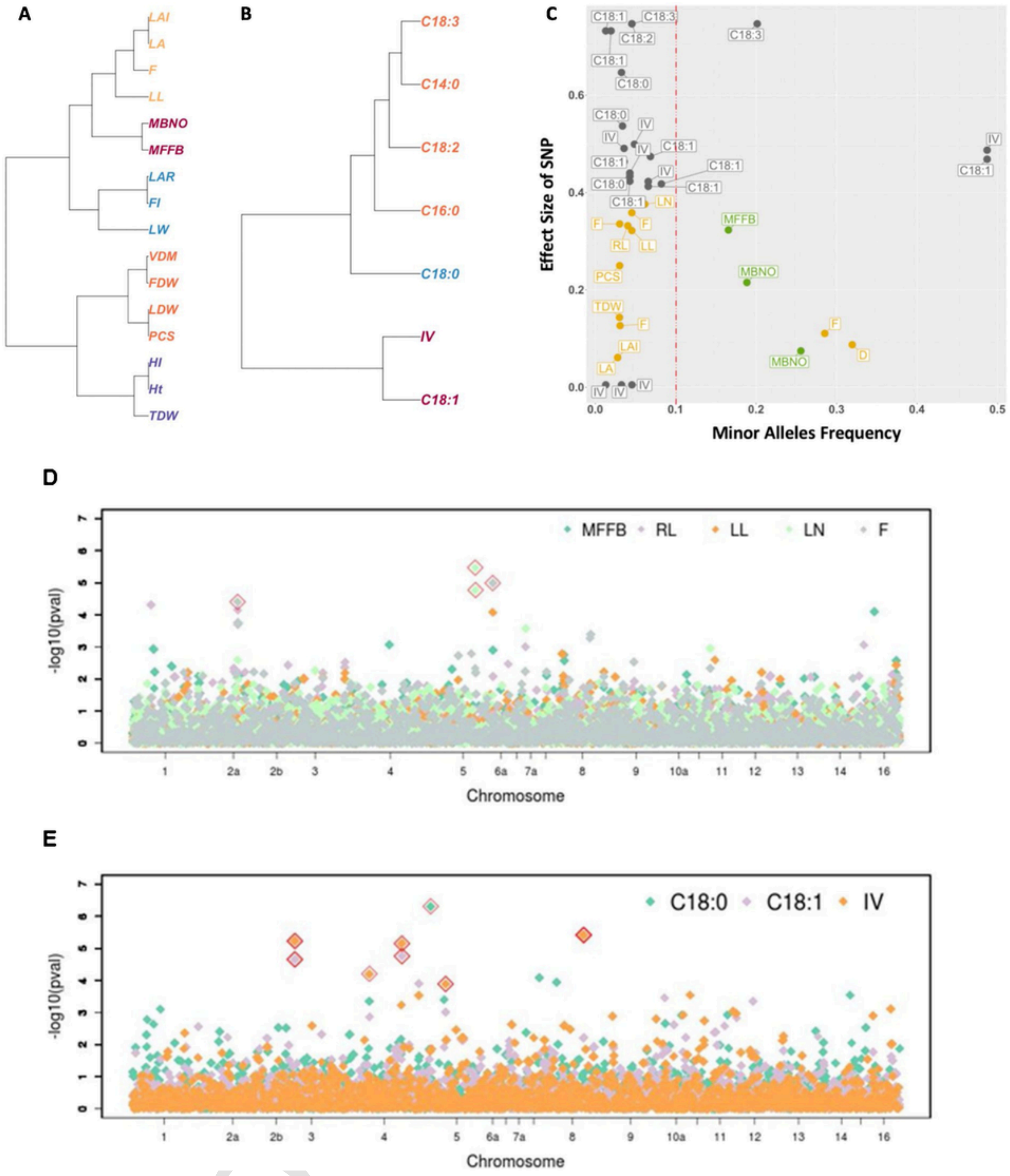

Fig. 3. Phenotypic and genetic architecture of HCPC traits. A) Hierarchical clustering of correlated vegetative/reproductive traits resulted in five distinct clusters. Cluster I: leaf length (LL), fractional interception (F), leaf area (LA), leaf area index (LAI); Cluster II: mean fresh fruit bunch (MFFB) and mean bunch number (MBNO); Cluster III: leaf width (LW), frond index (FI), and leaf area ratio (LAR); Cluster IV: petiole cross section (PCS), leaf dry weight (LDW), frond dry weight (FDW), and vegetative dry matter (VDM); Cluster V: trunk dry weight (TDW), height (Ht), and height increment (HI). B) Hierarchical clustering of correlated FAC traits resulted in three clusters. Cluster I: C18:2, C18:3, C14:0 and C16:0; Cluster II: C18:1 and IV; Cluster III: C18:0 and C14:0. C) Distribution of effect size and minor allele frequency of trait-associated SNPs. Grey corresponds to FAC traits; yellow, vegetative traits; green, reproductive traits. Clusters were determined based on a minimum pairwise Pearson correlation coefficient $r > 0.6$, $P < 0.05$. D) Manhattan plot of GWAS single-loci model analysis in vegetative and reproductive traits. E) Manhattan plot of GWAS single-loci model analysis in fatty acid traits. Different colors indicate different traits. Associated markers (FDR corrected p-value < 0.05) are noted using a red triangle symbol on each point.

for MFFB and MBNO (Fig. 4), suggestive of patterns of local adaptation arising from divergent selection among the HCPC populations.

3.7. Functional annotations of candidate genes

GWAS has been used to discover causal loci that regulate important traits in many staple crops [36,37]. We identified a total of 55 trait-associated SNPs and some interesting candidate genes within 20 kb of these SNPs (Table S2B). Among the genes associated with vegetative and yield-related traits is *proline-rich receptor-like protein kinase (PERK8)* linked to MBNO-associated SNP 3:82372832. Transgenic Arabidopsis *PERK* mutants form flowers with an abnormal appearance and impaired fertility, suggesting that the *PERK8* gene may have an important role in flower development and consequent development of oil palm bunches [38]. Another SNP associated with Leaf Area (13:29551843) is linked to a gene homologous to an *auxin glycosyltransferase protein, UGT74D1*, and overexpression of this gene in Arabidopsis causes a curled leaf phenotype [39], indicating that it is involved in regulating leaf morphology. We find several interesting genes associated with FAC traits, some of which are involved in fatty acid biosynthesis pathways.

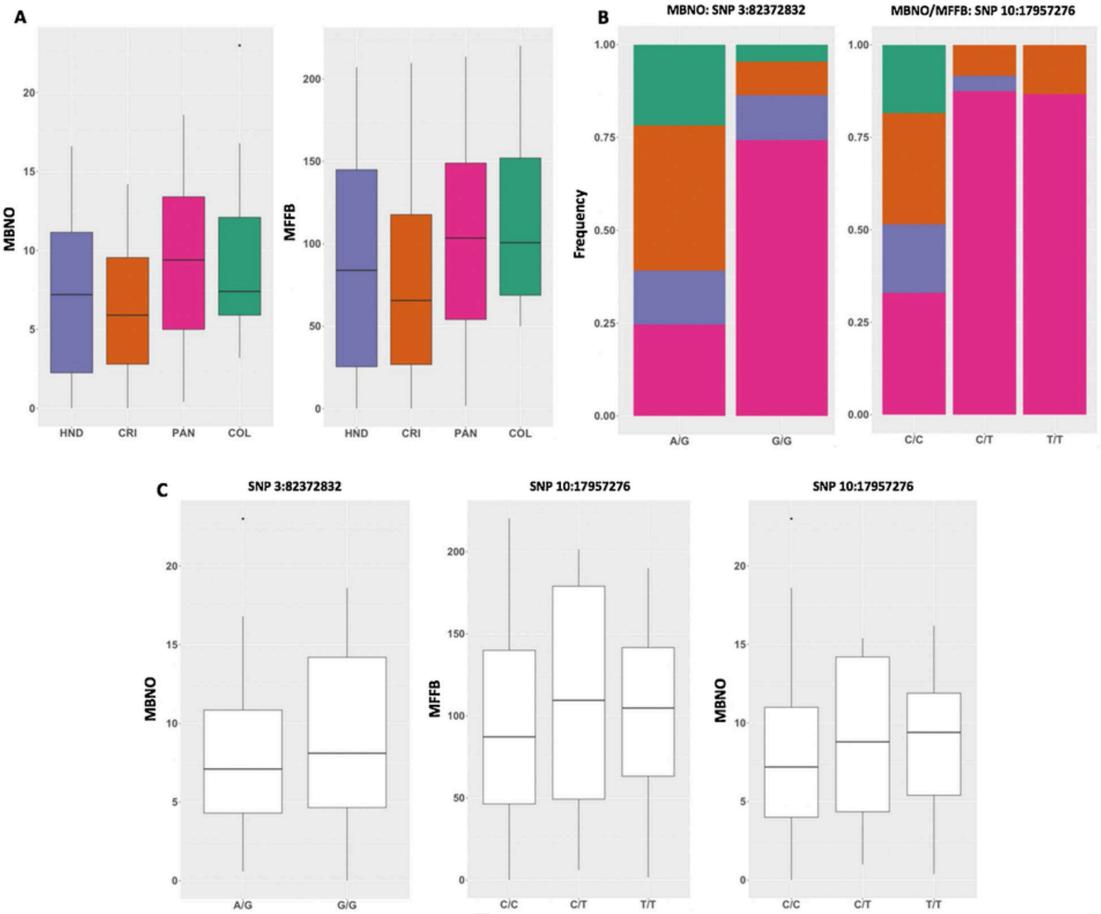

Fig. 4. MFFB and MBNO variation in wild HCPC populations. A) Population phenotypic divergence of reproductive traits (MBNO and MFFB). B) Genotype frequency divergence of MBNO/MFFB-associated SNPs 3:82372832 and 10:17957276. C) Distribution of genotype effects on MBNO/MFFB trait value. Purple corresponds to Honduras; orange, Costa Rica; pink, Panama; green, Colombia.

For instance, IV-associated SNP 5:21628085 is linked to a gene homologous to the well-characterized Arabidopsis *TRANSPARENT TESTA8* (*TT8*) gene, known to not only be involved in flavonoid biosynthesis and seed coat color but also the accumulation of C16 and C18 fatty acids [40]. Similarly, genes that are responsible for seed coat pigmentation in sesame were also found to affect variation in oil content and composition [41]. A SNP associated with C18:0 (14:6491498) is linked to a gene homologous to *glutathione transferase 7* (*GSTU7*), which belongs to a family of *GSTU* proteins involved in the accumulation of fatty acyl derivatives with chain lengths that vary in length from C_6 to C_{18} [42]. SNP 5:3306822 is in a gene homologous to *PEX4* and was associated with two saturated fatty acids: C18:0 and C14:0. *PEX4* is known to regulate peroxisomes, which are sites of fatty acid β-oxidation [43]. The relationship between peroxisomal β-oxidation and long chain fatty acids has been well studied in other organisms [44].

3.8. *Strategic conservation of core HCPC populations*

Maintaining genetic resources in the field collection requires a large land area and incurs high cost, especially for perennial tree species such as *E. oleifera*. Considering the long breeding cycle of oil palm, genetic material that shows good potential for improvement, particularly those that are incorporated into breeding programs, need to be conserved to ensure long-term availability for future exploitation. Because HCPC populations are highly inbred, with most trait variation driven by low frequency alleles, it is important to reduce genetic redundancy while maximizing genetic diversity. Based on computer simulations, sampling approximately 70 individuals (out of a total sample size of 154) will likely capture at least 90 % of the total genetic diversity observed in wild HCPC populations (Fig. 5). A sample size of 70 individuals shows marginal sampling variance of nucleotide diversity, indicating a saturation point at 70 individuals at which sampling more individuals will not influence diversity estimates (Fig. S8).

4. Discussion

4.1. *Distribution of genetic diversity*

In this study, we examined natural diversity in wild American oil palms originating from Honduras, Costa Rica, Panama, Colombia, Brazil, Ecuador and Peru. While the species center of origin has not been identified, it has been suggested that the American oil palm was introduced into the upper Amazon Basin from Central America. This inference was based on observations that Amazonian oil palms are generally found in anthropic soils within regions occupied by humans [45–47]. Our analyses rule out Central America as the species center of origin. Individuals from Central America show much lower levels of genetic diversity compared to individuals within the Amazon Basin, including Brazil, Ecuador, and Peru, (Table 1), consistent with a previous RFLP marker study that showed similar results [48]. Ecuador and Peru individuals have the highest allelic richness, with heterozygosity that is twice that of HCPC individuals. Furthermore, HCPC populations have a relatively high degree of inbreeding and are strongly marked by recent bottleneck episodes that make it unlikely that HCPC populations repre-

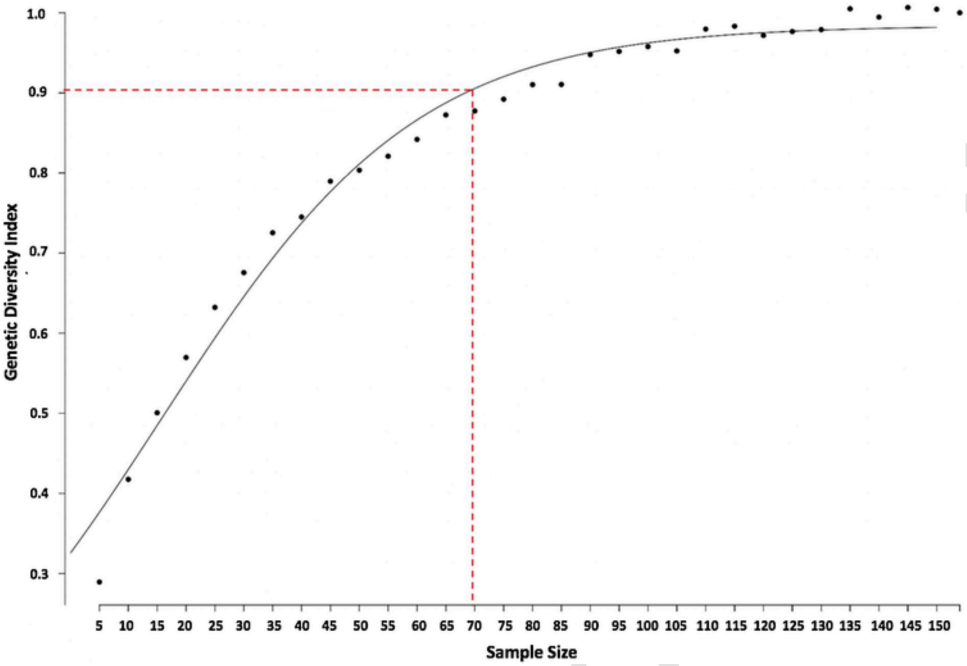

Fig. 5. Conservation of core HCPC populations. Genetic diversity index across different sample sizes based on computer simulations. Genetic diversity index indicates the estimated proportion of genetic variation captured for each sample size. Each dot represents the mean value of 100 iterations per sample size. A regression curve showing variation patterns of genetic diversity of each sample size against those of the total sample size (N = 154). Genotype missing rate = 5%.

sent the species center of origin. Based on our results, we suspect that the migration was in the opposite direction, and that HCPC populations originated from the Amazon basin and dispersed northward from South America to establish populations in Central America and Northern Colombia.

Dispersal into Central America must have occurred after the formation of the Panama isthmus that bridged North and South America. This event led to one of the largest biological exchanges between the previously disconnected landmasses, known as the Great American Biotic Interchange [49]. There is evidence that plants were the first organisms to migrate between North and South America [50,51]. Although the timing of the complete closure of the Panama isthmus remains controversial, it has been estimated to have occurred between 3–23 Mya [51,52]. This time frame is much later than the estimated 50 My divergence time of the American *E. oleifera* and African *E. guineensis* species [10] but is consistent with our genetic distance and diversity metrics. Further analyses of coalescence time, migration rate and effective population size could estimate this divergence as well as confirm the American oil palm center of origin.

Based strictly on SNPs, H_o recorded in this study is lower than H_o across *E. guineensis* breeding populations (0.220–0.260) [53] and *E. guineensis* Angola material ($H_o = 0.400$) [54] but inline with previous observations in wild *E. oleifera* [55]. Among the molecular markers available to-date, SNPs generate highly reliable demographic evidence particularly in small-sized populations of non-model species [56]. Our simulation analyses show the reliability of diversity estimates using SNPs in sample sizes as low as 25. This implies the importance and significance of our findings towards creating informed choices to oil palm breeders and plantation managers for effective use of wild *E. oleifera* genetic resources.

The current research revealed genome-wide LD decays to a baseline $r^2 = 0.20$ and $r^2 = 0.10$ within 4 kb and 10 kb, respectively (Fig. S6). In cultivated *E. guineensis* crosses, LD extended over larger distances, between 20 kb - 120 kb at $r^2 = 0.20$ [57,58]. These differences are likely due to a combination of selection, mutation, migration, population size and mating patterns [59]. Our study revealed that LD decay more rapidly in *E. oleifera*, an outcrossing species as compared to self-pollinated plants such as tea, potato and tomato [60–62].

4.2. Phenotypic diversity and trait architecture

We observed considerable phenotypic variation segregating among these populations and identified several alleles associated with key agricultural traits (Table S2). Phenotypic variation in most traits appeared to be driven by low frequency alleles (Fig. 4C), likely due to 80 % of SNPs having a minor allele frequency less than 0.1. It has been documented that variation in complex traits is enriched for low-frequency alleles [36,63]. We found that alleles associated with FAC explain a larger percentage of phenotypic variance (53 %) compared with vegetative and yield-related traits (33 %). In maize, the phenotypic variation explained by FAC-associated markers reached up to 83 %, which contrasted with 5% of the explained variance for vegetative traits [36]. Yield-related traits are usually highly polygenic and controlled by numerous alleles, as seen in studies of crops and domesticated animals where hundreds of SNPs were found to be associated with yield-related traits [36,64,65]. These patterns indicate that FAC traits may have a simpler mode of inheritance relative to vegetative and yield traits and thus are likely influenced by fewer genes with large effects, making FAC traits relatively more straightforward to breed for. We note that there are no significant correlations between FAC and yield-related traits (MBNO/MFFB/MABW, Fig. S1), indicating that it is possible to breed for high-yielding palms with different saturation profiles for commercial plantations. For example, oil palms that possess high stearic acid are desirable as it opens opportunities for its utilization as cocoa butter substitute and other products, i.e massage oils, shaving cream and lotions [66]. We confirmed the allelic calls of a SNP associated with stearic acid content (C18:0) (SNP 14:6491498) using an independent Sanger experiment (Fig. S9). This associated SNP (14:6491498) potentially has important implications in oil palm breeding and genetic engineering programs.

4.3. Relationship between yield and sexual reproduction in oil palm

MBNO and MFFB are major economic traits that determine oil palm crop productivity because they reflect the number of female flowers produced [67]. The sex-ratio defined by the proportion of female to total inflorescence ultimately defines the number of bunches produced in oil palm. Sex ratios vary significantly among individuals, with some trees found to be nearly all male or female depending on the genotype and environment. Trees grown in conditions of high moisture and nutrients tend to produce more female flowers, while drought stress and defoliation can trigger an increase in male flowers [68]. Selection for high-yielding genotypes inevitably confer an increase in female flower production, and when grown in favorable conditions, the production of male flowers is close to zero [69,70]. Alternatively, breeders have selected ‘supermale’ genotypes to provide sufficient pollen to pollinate high-yielding varieties with high female-male ratios. Several QTLs associated with sex-ratio variation have been previously identified in *E.* guineensis [70].

5. Conclusions

Our study documents the genetic and phenotypic diversity of a panel of wild American oil palm populations. The genetic characterization of the American oil palm germplasm will ensure that the resources are adequately conserved and screening for useful traits will provide opportunity for greater use by breeders. Our results provide a catalog of SNPs associated with agronomically valuable traits that will aid in marker-assisted selection and molecular transgenic breeding programs. Future studies focused on understanding variation in sex-ratios as well as genes associated with responses to environmental stress, such as disease, drought, higher temperatures and nutrient deficiency, will be important for breeding productive palms in the face of climate change.

Author contributions

Wendy Vu: Made the RAD libraries for sequencing, performed data analysis, as well as wrote and edited the manuscript

Maizura Ithnin: selected samples for the experiment, managed sample collection, wrote and edited the manuscript

Min-Gyoung Shin : carried out GWAS analysis and wrote the manuscript

Vasantika Suryawanshi: carried out relevant bioinformatics analysis

Katrina Sherbina: carried out relevant bioinformatics analysis

Siti Hazirah Zolkafli: performed validation experiment of significant trait-associated SNPs

Norhalida Mohamed Serdari: carried out DNA extraction and organized samples for RAD sequencing

Mohd Din Amiruddin: oversaw the field trial and data collection

Norziha Abdullah: organized the FAC analysis of the samples

Suzana Mustaffa : coordinated vegetative measurements

Marhalil Marjuni : organized bunch analysis measurements

Rajanaidu Nookiah: reviewed the manuscript

Ahmad Kushairi : reviewed the manuscript

Paul Marjoram : assisted with GWAS analysis

Sergey V. Nuzhdin : designed the experiment and reviewed the manuscript

Peter L. Chang : designed the experiment, led analysis of sequence data, wrote and edited the manuscript

Rajinder Singh : designed the experiment, selected samples for the experiment and reviewed the manuscript

Funding

The work was supported by MPOB through the Oil Palm Genome Program, grant number R009511000.

Data availability

All sequencing data can be found in the NCBI BioProject PRJNA434010

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

We acknowledge the Director General of MPOB for permission to publish the research findings.

We thank the Breeding group for maintaining the *E. oleifera* germplasm. We also acknowledge MPOB’s support through the Oil Palm Genome Program (R009511000) in funding the research work.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.plantsci.2020.110731.

References

[1] Y. Basiron, Palm oil production through sustainable plantations, Eur. J. Lipid Sci. Technol. 109 (2007) 289–295, https://doi.org/10.1002/ejlt.200600223.

[2] A. Kushairi, R. Singh, M. Ong-Abdullah, Oil palm economic performance in Malaysia and R&D progress in 2017, J. Oil Palm Res. 30 (2) (2018) 163–195.

[3] N. Gilbert, Palm-oil boom raises conservation concerns, Nature. 487 (2012) 14–15.

[4] K.M. Carlson, R. Heilmayr, H.K. Gibbs, P. Noojipady, D.N. Burns, D.C. Morton, N.F. Walker, G.D. Paoli, C. Kremen, Effect of oil palm sustainability certification on deforestation and fire in Indonesia, Proc. Natl. Acad. Sci. U.S.A. 115 (1) (2018) 121–126.

[5] R.H.V. Corley, P.B. Tinker, The Oil Palm, fifth edition, Wiley Online Library, 2015.

[6] R. Hajjar, T. Hodgkin, The use of wild relatives in crop improvement: a survey of developments over the last 20 years, Euphytica. 156 (1) (2007) 1–13.

[7] G. Bauchet, M. Causse, Genetic diversity in tomato (solanum lycopersicum) and its wild relatives, Genetic Diversity in Plants, IntechOpen, 2012https://doi.org/10.5772/33073, Available from: https://www.intechopen.com/books/genetic-diversity-in-plants/genetic-diversity-in-tomato-solanum-lycopersicum-and-its-wild-relatives.

[8] M.L. Warburton, S. Rauf, L. Marek, M. Hussain, O. Ogunola, J. de Jesus Sanchez Gonzalez, The use of crop wild relatives in maize and sunflower breeding, Crop Sci. 57 (2017) 1227–1240.

[9] E.J. von Wettberg, P.L. Chang, F. Başdemir, N. Carrasquila-Garcia, L.B. Korbu, S.M. Moenga, G. Bedada, A. Greenlon, K.S. Moriuchi, V. Singh, M.A. Cordeiro, Ecology and genomics of an important crop wild relative as a prelude to agricultural innovation, Nat. Comm. 9 (2018) 649.

[10] R. Singh, M. Ong-Abdullah, E.T. Low, M.A. Manaf, R. Rosli, R. Nookiah, L.C. Ooi, S.E. Ooi, K.L. Chan, M.A. Halim, N.R. Azizi, et al., Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds, Nature. 500 (7462) (2013) 335–339.

[11] A. Mohd Din, R. Nookiah, S. Jalani, Performance of Elaeis oleifera from Panama, Costa rica, Colombia and Honduras in Malaysia, J. Oil Palm Res. 4 (2000) 146–155.

[12] R.D. Cunha, R. Lopes, R.D. Rocha, W.D. Lima, P.C. Teixeira, E. Barcelos, M.D. Rodrigues, S.D.A. Rios, Domestication and breeding: amazonian species, in: A. Borém, M.T.G. Lopes, C.R. Clement, H. Noda (Eds.), Domestication and Breeding: Amazonian Species, Independent Production,, 2012, pp. 275–296.

[13] E. Francia, G. Tacconi, C. Crosatti, D. Barabaschi, D. Bulgarelli, E. Dall’Aglio, E. Vale, Marker assisted selection in crop plants, Plant Cell, Tissue and Org, Culture 82 (2015) 317–342, https://doi.org/10.1007/s11240-005-2387-z.

[14] S. Mayes, P.L. Jack, D.F. Marshall, R.H.V. Corley, Construction of a RFLP genetic linkage map for oil palm (Elaeis guineensis Jacq.), Genome. 40 (1) (1997) 16–122.

[15] R. Escobar, A. Alvarado, Strategies in production of oil palm seed varieties and clones for high density planting, ASD Oil Palm Papers. 27 (2004) 13–26.

[16] R. Nookiah, M.M. Ainul, Conservation of oil palm and coconut genetic resources, in: M.N. Normah, H.F. Chin, B.M. Reed (Eds.), (Eds.), Conservation of Tropical Plant Species, Springer Science & Business Media, New York,, 2013, pp. 189–212.

[17] C. Lippert, J. Listgarten, Y. Liu, C.M. Kadie, R.I. Davidson, D. Heckerman, FaST linear mixed models for genome-wide association studies, Nat. Methods 8 (10) (2011) 833.

[18] P.F. O'Reilly, C.J. Hoggart, Y. Pomyen, F.C. Calboli, P. Elliott, M.R. Jarvelin, L.J. Coin, MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS, PLoS One7 (5) (2012), e34861.

[19] G. Blaak, L. Sparnaaij, T. Menedez, Breeding and inheritance in the oil palm (Elaeis guineensis Jacq.) II. Methods of bunch quality analysis, J. West African Ins. Oil Palm Res. 4 (1963) 46–155.

[20] R.H.V. Corley, C.J. Breure, Measurements in Oil Palm Experiments (Internal Report), Unilever Plantation Group, London, 1981.

[21] A. Kuntom, N. Aini Idris, N. Amri Ibrahim, M. Thin Sue, S. Wai Lin, T. Yew Ai, MPOB Test Methods a Compendium of Tests on Palm Oil Products, Palm Kernel Products, Fatty Acids, Food Related Products and Others, Ministry of Plantation Industries and Commodities Malaysian Palm Oil Board, Malaysia, 2005.

[22] J.J. Doyle, Isolation of plant DNA from fresh tissue, Focus 12 (1990) 13–15.

[23] R.J. Elshire, J.C. Glaubitz, Q. Sun, J.A. Pol, K. Kawamoto, E.S. Buckler, A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species, PLoS One 6 (5) (2011).

[24] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform, Bioinformatics. 26 (5) (2010) 589–595.

[25] M.J. Hubisz, D. Falush, M. Stephens, J.K. Pritchard, Inferring weak population structure with the assistance of sample group information, Mol. Eco. Res. 9 (5) (2009) 1322–1332.

[26] Z.N. Kamvar, J.C. Brooks, N.J. Grünwald, Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality, Front. Genet. 6 (2015), Article208.

[27] F. Murtagh, P. Legendre, Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?, Int. J. Nurs. Terminol. Classif. 31 (2014) 274–295.

[28] D. Müllner, Fastcluster: fast hierarchical, agglomerative clustering routines for R and Python, J. Statistical. Software 9 (2013) 1–18.

[29] X. Zheng, D. Levine, J. Shen, S.M. Gogarten, C. Laurie, B.S. Weir, A high-performance computing toolset for relatedness and principal component analysis of SNP data, Bioinformatics. 28 (24) (2012) 3326–3328.

[30] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean, The variant call format and VCFtools, Bioinformatics. 27 (15) (2011) 2156–2158.

[31] T.H. Lee, H. Guo, X. Wang, C. Kim, A.H. Paterson, SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data, BMC Genomics15 (2014) 162.

[32] W.G. Hill, B.S. Weir, Variances and covariances of squared linkage disequilibria in finite populations, Theor. Pop. Biol. 33 (1) (1988) 54–78.

[33] X. Zhou, M. Stephens, Efficient multivariate linear mixed model algorithms for genome-wide association studies, Nat. Methods11 (2014) 407.

[34] V. Segura, B.J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren, Q. Long, M. Nordborg, An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations, Nat. Genet. 44 (7) (2012) 825–830.

[35] B.L. Browning, S.R. Browning, Genotype imputation with millions of reference samples, Am. J. Hum. Genet. 98 (1) (2016) 116–126.

[36] Y. Xiao, H. Tong, X. Yang, S. Xu, Q. Pan, F. Qiao, M.S. Raihan, Y. Luo, H. Liu, X. Zhang, N. Yang, Genome-wide dissection of the maize ear genetic architecture using multiple populations, New Phytol. 210 (3) (2016) 1095–1106.

[37] M.G. Shin, S.V. Bulyntsev, P.L. Chang, L.B. Korbu, N. Carrasquila-Garcia, M.A. Vishnyakova, M.G. Samsonova, D.R. Cook, S.V. Nuzhdin, Multi-trait analysis of domestication genes in Cicer arietinum–Cicer reticulatum hybrids with a multidimensional approach: Modeling wide crosses for crop improvement, Plant Sci. 285 (2019) 122–131.

[38] Y. Haffani, N. Silva-Gagliardi, S. Sewter, M.G. Aldea, Z. Zhao, A. Nakhamchik, R. Cameron, D. Goring, Altered expression of PERK receptor kinases in Arabidopsis leads to changes in growth and floral organ formation, Plant Sig. Behav. 1 (5) (2006) 251–260.

[39] S.H. Jin, X.M. Ma, P. Han, B. Wang, Y.G. Sun, G.Z. Zhang, Y.J. Li, B.K. Hou, UGT74D1 is a novel auxin glycosyltransferase from Arabidopsis thaliana, PLoS One8 (4) (2013), e61705.

[40] M. Chen, L. Xuan, Z. Wang, L. Zhou, Z. Li, X. Du, E. Ali, G. Zhang, L. Jiang, TRANSPARENT TESTA8 inhibits seed fatty acid accumulation by targeting several seed development regulators in Arabidopsis, Plant Physiol. 165 (2) (2014) 905–916.

[41] X. Wei, K. Liu, Y. Zhang, Q. Feng, L. Wang, Y. Zhao, D. Li, Q. Zhao, X. Zhu, X. Zhu, W. Li, Genetic discovery for oil production and quality in sesame, Nat. Comm. 6 (2015) 8609.

[42] D.P. Dixon, R. Edwards, Selective binding of glutathione conjugates of fatty acid derivatives by plant glutathione transferases, J. Biol. Chem. 284 (32) (2009) 21249–21256.

[43] Y.T. Kao, W.A. Fleming, M.J. Ventura, B. Bartel, Genetic interactions between PEROXIN12 and other peroxisome-associated ubiquitination components, Plant Physiol. 172 (3) (2016) 1643–1656.

[44] G. Cassin-Ross, J. Hu, Systematic phenotypic screen of arabidopsis peroxisomal mutants identifies proteins involved in β-oxidation, Plant Physiol. 166 (3) (2014) 1546–1559.

[45] J.G.W. Boer, The Indigenous Palms of Suriname, vol 5, Brill Archive, 1965.

[46] G. Morcote-Ríos, R. Bernal, Remains of palms (Palmae) at archaeological sites in the new world: a review, The Botanical Rev. 67 (3) (2001) 309–350.

[47] N. Smith, Palms and People in the Amazon, Springer International Publishing, 2015.

[48] E. Barcelos, P. Amblard, J. Berthaud, M. Seguin, Genetic diversity and relationship in American and African oil palm as revealed by RFLP and AFLP molecular markers, Pesqui. Agropecu. Bras. 37 (8) (2002).

[49] M.O. Woodburne, The great american biotic interchange: dispersals, tectonics, climate, sea level and holding pens, J. Mammalian Evo. 17 (4) (2010) 245–264.

[50] S. Cody, J.E. Richardson, V. Rull, C. Ellis, R.T. Pennington, The great American biotic interchange revisited, Ecography. 33 (2010) 326–332.

[51] C.D. Bacon, D. Silvestro, C. Jaramillo, B.T. Smith, P. Chakrabarty, A. Antonelli, Biological evidence supports an early and complex emergence of the Isthmus of Panama, Proc. Natl. Acad. Sci. U.S.A. 112 (19) (2015) 6110–6115.

[52] A. O'Dea, H.A. Lessios, A.G. Coates, R.I. Eytan, S.A. Restrepo-Moreno, A.L. Cione, L.S. Collins, A. De Queiroz, D.W. Farris, R.D. Norris, R.F. Stallard, Formation of the isthmus of Panama, Sci. Adv. 2 (8) (2016) e1600883.

[53] W. Xia, T. Luo, W. Zhang, A.S. Mason, D. Huang, X. Huang, X.W. Tang, Y. Dou, C. Zhang, Y. Xiao, Development of high-density snp markers and their application in evaluating genetic diversity and population structure in Elaeis guineensis, Front. Plant Sci. 10 (2019) 130.

[54] P.W. Ong, I. Maizura, N.A.P. Abdullah, M.Y. Rafii, L.C.L. Ooi, E.T.L. Low, R. Singh, Development of SNP markers and their application for genetic diversity analysis in the oil palm (Elaeis guineensis), Genet. Mol. Res. 14 (4) (2015) 12205–12216.

[55] M. Ithnin, C.K. Teh, W. Ratnam, Genetic diversity of Elaeis oleifera (HBK) Cortes populations using cross species SSRs: implication's for germplasm utilization and conservation, BMC Genet. 18 (2017) 37.

[56] C. García, E. Guichoux, A. Hampe, A comparative analysis between SNPs and SSRs to investigate genetic variation in a juniper species (Juniperus phoenicea ssp. turbinata), Tree Genet. Genome 14 (2018) 87.

[57] Q.B. Kwong, C.K. Teh, A.L. Ong, H.Y. Heng, H.L. Lee, M. Mohamed, J.Z.B. Low, S. Apparow, F.T. Chew, S. Mayes, H. Kulaveerasingam, Development and validation of a high-density SNP genotyping array for African oil palm, Mol. Plant9 (8) (2016) 1132–1141.

[58] C.K. Teh, A.L. Ong, Q.B. Kwong, S. Apparow, F.T. Chew, S. Mayes, M. Mohamed, D. Appleton, H. Kulaveerasingam, Genome-wide association study identifies three key loci for high mesocarp oil content in perennial crop oil palm, Sci. Rep. Ist. Super. Sanita 6 (2016) 19075.

[59] S.A. Flint-Garcia, J.M. Thornsberry, E.S. Buckler, Structure of linkage disequilibrium in plants, Ann. Rev. Plant Biol. 54 (2003) 357–374.

[60] S. Niu, Q. Song, H. Koiwa, D. Qiao, D. Zhao, Z. Chen, X. Liu, X. Wen, Genetic diversity, linkage disequilibrium, and population structure analysis of the tea plant (Camellia sinensis) from an origin center, Guizhou plateau, using genome-wide SNPs developed by genotyping-by-sequencing, BMC Plant Biol. 19 (1) (2019) 328.

[61] P.G. Vos, M.J. Paulo, R.E. Voorrips, R.G. Visser, H.J. Van Eck, F.A. Van Eeuwijk, Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato, Theor. Appl. Genet. 130 (1) (2017) 123–135.

[62] Y.P. Lin, C.Y. Liu, K.Y. Chen, Assessment of genetic differentiation and linkage disequilibrium in Solanum pimpinellifolium using genome-wide high-density SNP markers, G3 Genes, Genomes, Genetics. 9 (2019) 1497–1505.

[63] A. Gusev, G. Bhatia, N. Zaitlen, B.J. Vilhjalmsson, D. Diogo, E.A. Stahl, P.K. Gregersen, J. Worthington, L. Klareskog, S. Raychaudhuri, R.M. Plenge, Quantifying missing heritability at known GWAS loci, PLoS Genet. 9 (12) (2013), e1003993.

[64] M.E. Goddard, B.J. Hayes, Mapping genes for complex traits in domestic animals and their use in breeding programmes, Nat. Rev. Genet. 10 (2009) 381–391.

[65] Y. Xing, Q. Zhang, Genetic and molecular bases of rice yield, Ann. Rev. Plant Biol. 61 (2010) 421–442.

[66] G.K.A. Parveez, O.A. Rasid, M.Y.A. Masani, R. Sambanthamurthi, Biotechnology of oil palm: strategies towards manipulation of lipid content and composition, Plant Cell Rep. 34 (4) (2015) 533–543.

[67] M.N. Okoye, C.O. Okwuagwu, M.I. Uguru, Population improvement for fresh fruit bunch yield and yield components in Oil Palm (Elaeis guineensis jacq.), american-eurasian jour, Sci. Res. 4 (2) (2009) 59–63.

[68] H. Adam, M. Collin, F. Richaud, T. Beulé, D. Cros, A. Omoré, L. Nodichao, B. Nouy, J.W. Tregear, Environmental regulation of sex determination in oil palm: current knowledge and insights from other species, Ann. Bot. 108 (8) (2011) 1529–1537.

[69] A. Rival, Breeding the oil palm (Elaeis guineensis Jacq.) for climate change, Oilseeds Fats, Crop Lipids. 24 (1) (2017) D107.

[70] S. Somyong, S. Poopear, S.K. Sunner, K. Wanlayaporn, N. Jomchai, T. Yoocha, K. Ukoskit, S. Tangphatsornruang, S. Tragoonrung, ACC oxidase and miRNA 159a, and their involvement in fresh fruit bunch yield (FFB) via sex ratio determination in oil palm, Mol. Genet. Genomics291 (3) (2016) 1243–1257.