Genome properties of key oil palm (Elaeis guineensis Jacq.) breeding populations

Essubalew Getachew Seyum, Ngalle Hermine Bille, Wosene Gebreselassie Abtew, Pasi Rastas, Deni Arifianto, Hubert Domonhédo, Benoît Cochard, Florence Jacob, Virginie Riou, Virginie Pomiès, et al.

Genome properties of key oil palm (Elaeis guineensis Jacq.) breeding populations

Essubalew Getachew SEYUM (g.essu2011@gmail.com)
Jimma University
Ngalle Hermine Bille
University of Yaounde I: Universite de Yaounde I
Wosene Gebreselassie Abtew
Jimma College of Agriculture: Jimma University
Pasi Rastas
University of Helsinki Helsinki Institute of Life Sciences: Helsingin yliopisto Helsinki Institute of Life Sciences
Deni Arifianto
P.T. SOCFINDO Medan
Hubert Domonhédo
INRAB, CRA-PP, Pobè
Benoît Cochard
PALMELIT
Florence Jacob
PALMELIT
Virginie Riou
CIRAD
Virginie Pomiès
CIRAD
David Lopez
CIRAD
Joseph Martin Bell
University of Yaounde I: Universite de Yaounde I
David Cros
CIRAD https://orcid.org/0000-0002-8601-7991

Research Article

Keywords: Genome properties, Genomic selection, Genotyping-by-sequencing, Oil palm, Single nucleotide polymorphisms

Posted Date: February 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1307249/v1

Abstract

Good knowledge about genome properties of the populations helps to optimize breeding methods, particularly genomic selection (GS). In oil palm (*Elaeis guineensis* Jacq.), the first global source of vegetable oil, GS gave promising results. The present study considered two complex oil palm breeding populations, Deli and La Mé, with 943 individuals and 7,324 single-nucleotide polymorphisms (SNPs) from genotyping-by-sequencing. Linkage disequilibrium (LD), haplotype sharing, effective size (N_e), and fixation index (F_{st}) were investigated. A genetic linkage map was constructed, including 4,252 SNPs and spanning 1,778.52 cM, with an average recombination rate of 2.85 cM/Mbp. The LD at r^2=0.3, considered as the minimum to get reliable results for genomic predictions, spanned over 1.05 cM/0.22 Mbp in Deli and 0.9 cM/0.21 Mbp in La Mé. The significant degree of differentiation existing between Deli and La Mé was confirmed by the high Fst value (0.53), the pattern of correlation of SNP heterozygosity and allele frequency among populations, and the decrease of persistence of LD and of haplotype sharing among populations with increasing SNP distance. However, the level of resemblance between the two populations over short genomic distances (correlation of r values between populations >0.6 for SNPs separated by <0.5 cM or <1 kbp and percentage of common haplotypes >40% for haplotypes <3,600 bp or <0.20 cM) likely explains the superiority of GS models ignoring the parental origin of marker alleles over models taking this information into account. The two populations had low N_e (<5). Population-specific genetic maps and reference genomes are needed for future studies.

1. Introduction

Oil palm (*Elaeis guineensis* Jacq.) is a perennial tropical monocot oil-producing plant that belongs to the *Arecaceae* family. It originated from the Gulf of Guinea. It is naturally cross-pollinated, monoecious, allogamous, and diploid, with a chromosome number of 2n = 2x = 32 and having a genome sequence of 1.8 gigabases (Ithnin and Din 2020). The economic life span of oil palm ranges from 25 to 30 years and it is mainly cultivated in humid tropical zones of the world (Barcelos et al. 2015).

Nowadays, the cultivation of oil palm relies on hybrid varieties because they have a high yield per hectare. Group A and Group B are the two heterotic groups involved in the oil palm hybrid cultivar development (Nyouma et al. 2019). Group A mostly consists of the Deli parental population, which is derived from four individuals of an unknown area of Africa planted in 1848 in Indonesia (Hartley 1988). The selection of the Deli population, mainly for yield, started in the early twentieth century. Group B is made up of several African breeding populations. African populations resulted from a limited number of founders collected during the first half of the twentieth century. La Mé population originated collected in the Bingerville region of the Ivory Coast between 1924 and 1930, with three founders the individuals considered here (Cochard et al. 2009). In both A and B groups, inbreeding was commonly used, by using selfing or by mating with related selected individuals (Corley and Tinker 2016).

The total world vegetable oil production is currently around 200 million metric tons (MT), led by oil palm (75 MT), followed by soybean oil (60 MT), rapeseed oil (28 MT), and sunflower oil (19 MT) (Statista 2021). The world demand for oil palm is expected to reach 240 million tons by 2050 (Corley 2009). Oil palm produces an average oil yield of 4 tons per hectare every year, which is approximately 7-10 times higher than soybean (Babu and Mathur 2016; Corley and Tinker 2016; Pirker et al. 2016). Oil palm is an important source of edible oil with over 80% of the products used in the food industry (cooking/frying oil, shortenings, margarine, and confectionery fats), and the rest used in the chemical industry for the formulation of soaps and detergents, pharmaceutical products, cosmetics, biodiesel, etc (Basiron 2007; Corley 2009; Soh et al. 2017).

Despite its wide adaptation and importance, oil palm production and productivity are generally far from their potential due to biotic and abiotic practical constraints. Climate change, land shortage, and labor shortage are among the major factors hindering the current and future yield of oil palm across the world (Corley 2009; Paterson et al. 2013; Barcelos et al. 2015; Pirker et al. 2016; Kwong et al. 2016). The constraints of the conventional methods currently used for oil palm breeding, i.e. long breeding cycle (>15 years) and a limited number of tested individuals, also limit the current palm oil yield (Cros et al. 2015; Jin et al. 2016; Seng et al. 2016). To provide a solution while ensuring a sustainable future, marker-assisted breeding has recently been introduced into oil palm breeding programs (Soh et al. 2017).

Genomic selection (GS) is a marker-assisted selection (MAS) method with a high density of markers on the entire genome so that at least one marker is in linkage disequilibrium with each quantitative trait locus (QTL) (Meuwissen et al. 2001). It is the most effective MAS method to improve quantitative traits (Heffner et al. 2009). Studies on the application of GS in oil palm brought positive results.

Thus, GS could improve oil palm clonal selection (Nyouma et al. 2020) and the selection of parents to use for hybrid crossings (Cros et al. 2017). Generally, GS in oil palm can enhance selection intensity and/or shorten the generation interval, thus increasing the annual genetic gain (Nyouma et al. 2019). However, the method could be optimized, in terms of the prediction model, marker density, etc. This should be done in light of the genome properties of the oil palm populations used in the reciprocal recurrent selection breeding scheme, in particular, linkage disequilibrium (LD), effective size (N_e), haplotype sharing, and fixation index (F_{st}).

Linkage disequilibrium is defined as the nonrandom association of alleles at two or more loci (Weir 1979; Slatkin 2008). The concept of GS relies heavily on LD between QTLs and DNA markers and a good knowledge of LD in the breeding population is necessary to optimize GS (Nakaya and Isobe 2012; Technow et al. 2014; Li and Kim 2015; Bejarano et al. 2018). The LD pattern is shaped by genetic factors, i.e. mutations and historical events that occurred during domestication and population formation, including natural and artificial selection, drift, migration, and non-random mating, as well as from non-genetic factors such as marker ascertainment bias (Flint-Garcia et al. 2003; Gupta et al. 2005; Mackay and Powell 2007; Slatkin 2008).

The number of randomly mating individuals in a population that gives rise to an observed rate of inbreeding is known as effective size (N_e) (Falconer and Mackay 1996). A lower N_e results from higher rates of inbreeding and genetic drift in a population (Lin et al. 2014), making that N_e is one of the major factors influencing LD, and consequently the accuracy of GS (Grattapaglia 2014). There is an inverse relationship between LD and N_e. For a given marker density, training population size, and trait, LD and GS prediction accuracy are higher in populations with low N_e than in populations with high N_e (Solberg et al. 2008; Grattapaglia 2014; Lin et al. 2014). So far, in oil palm, N_e was only estimated in the Deli population (Cros et al. 2014) and there is no information about N_e for La Mé population.

Fixation index (F_{st}) is used to identify loci with divergent allelic frequencies between two or more populations (Wright 1931). It helps to understand the genetic differentiation among groups (Jakobsson et al. 2013). It ranges from 0 (no variation between populations) to 1 (each population is fixed with a different allele). F_{st} analysis has been used to identify regions of the genome associated with domestication and selective sweeps associated with breeding (Yan et al. 2017). It can also improve GS and genome-wide association studies (GWAS). For example, (Chang et al. 2019) showed that prioritizing and weighting SNPs based on their F_{st} values can increase the accuracy of genomic predictions by more than 5%. (Yan et al. 2017) in soybean found that combining GWAS and fixation index analysis helped to identify QTLs for seed weight.

Haplotypes correspond to two or more SNP alleles that tend to be inherited as a unit in the chromosome (Bernardo 2010). Haplotype sharing helps to estimate the genetic resemblance between individuals and is a natural extension of identity by descent (Xu and Guan 2014). Several authors showed that the aggregation of SNPs into haplotypes can increase the prediction accuracy in animals (Calus et al. 2008; Cuyabano et al. 2014; Teissier et al. 2020) and in plant species that were allogamous or with high multiallelism (Matias et al. 2017; Ballesta et al. 2019). Also, consistency of linkage phases between QTL and marker alleles among populations is required to pool them to get a larger population for genetic studies.

The goal of this study is to characterize the genome properties of two major oil palm breeding populations, Deli and La Mé, focusing on key parameters for genomic predictions, namely linkage disequilibrium (LD), haplotype sharing, effective size (N_e), and fixation index (F_{st}).

2. Materials And Methods

2.1. Plant material and experimental design

The plant material used in this experiment consisted of individuals of the Deli and La Mé populations and their hybrid crosses. It comprised 943 genotyped individuals with 423 Deli, 140 La Mé and 380 Deli × La Mé hybrid individuals. Detailed pedigree information of these two populations is known over several generations (Cros et al. 2017). For the construction of the genetic map, all these genotyped individuals were used, as well as the non genotyped individuals comprised in their pedigree. For the other parts of the study, we only used the genotyped individuals of the Deli and La Mé breeding populations. The plant material was located partly in North Sumatra, on the SOCFINDO estate (Indonesia), and partly in Benin, on the INRAB research station of Pobè.

2.2. Genotypic data

Molecular data were obtained by genotyping by sequencing (GBS) (He et al. 2014). DNA extraction, GBS, and SNP calling were performed based on the procedure described in (Cros et al. 2017). The sequence data were processed using Tassel GBS version 5.2.44 (Glaubitz et al. 2014). The reference genome of (Singh et al. 2013) was used for alignment with Bowtie2 software (Langmead and Salzberg 2012). Biallelic SNPs were the only variants kept. SNP data points with depth below 10 were set to missing and only SNPs with less than 50% missing data in the two breeding populations were kept. SNPs with the sum of depth per datapoint above 550,000 and SNPs with 100% heterozygote genotypes were discarded. Individuals with more than 50% missing data were removed. Finally, we obtained 7,324 SNP markers, common to both breeding populations. It included 5,598 SNPs located on the anchored sequences of the genome (i.e. the 16 chromosomes of (Singh et al. 2013). The average percentage of missing data per SNP was 11% in Deli and 13% in La Mé (Supplementary Figure 1).

2.3. Construction of the genetic map

The genetic map was made using LepMAP3 software version 0.4 (Rastas 2017). First, module ParentCall2 was used to call the parental oil palm genotypes, with parameters removeNonInformative=1, to remove the non-informative markers (monomorphic or homozygous in both parents), and halfSibs=1. Second, the Filtering2 module was used to remove SNPs segregating in a non-Mendelian fashion using dataTolerance=0.001. Third, the SeparateChromosomes2 module assigned markers into linkage groups (LGs) by computing all pairwise LOD scores between markers and joined markers with LOD scores higher than the user-given parameter LodLimit, which was set to eight. Fourth, the JoinSingles2All module assigned single markers to the existing LGs by computing LOD scores between every single marker and markers from the existing LGs, using lodLimit=4 and iterate=1. Finally, OrderMarkers2 ordered the markers within each LG by maximizing the likelihood of the data given the order and using the Kosambi mapping function for conversion of recombination frequencies into map distances (centiMorgan, cM) (Rastas 2017). To join the maps of both male and female parents, the sexAveraged argument was set to 1. The individuals that were associated with outlier values in terms of the number of crossing-overs were identified in preliminary analysis and removed before the map construction. The markers which created large gaps at the top or bottom part of the LGs were discarded according to software developer recommendations. This resulted in a genetic map where the 16 largest LGs had largely higher numbers of SNPs than the remaining LGs, which were discarded to keep a genetic map with the number of LGs corresponding to the number of chromosomes of oil palm.

2.4. Comparison of genetic and physical maps

The genetic map and physical map, showing the positions of the reference genome of (Singh et al. 2013), were visualized using the R package LinkageMapView (Ouellette et al. 2018). We used MareyMap (Siberchicot et al. 2017) to plot the genetic position of the molecular markers against their physical position.

2.5. Within population linkage disequilibrium and persistence between populations

Analyses of linkage disequilibrium (LD) were performed in each breeding population using the PLINK software (Purcell et al. 2007). It computed pairwise estimates of LD by the classical measure of the squared correlation of allele frequencies at diallelic loci (r^2) and r. Before the computation of the r^2, the missing data points in the Deli and La Mé individuals were imputed using Beagle5.2 (Browning et al. 2018), independently for each breeding population. For the SNPs located on the assembled parts of the genome, the r^2 values between pairs of SNPs were plotted against physical distances (Mbp). For the SNPs located on the genetic map, the r^2 values were plotted against genetic distances (cM). The LD decay was plotted up to a 0.8 Mbp distance for physical positions and 3 cM for genetic positions. The relation between the r^2 values and distances was modeled by fitting local polynomials with the functions 'locpoly' and 'dpill' of the R package KernSmooth 2.23 (Wand 1995), as done for example in (Yamamoto et al. 2016).

The persistence of LD between populations was measured by the correlation of the r measure of LD between populations given by PLINK (r_{LD}). r_{LD} was computed between the two populations on the SNPs comprised in windows defined along with the genetic and physical maps, over a distance up to 90 cM and 50 Mbp, respectively. r_{LD} values can vary from -1 to 1, with a value close to 1 indicating a similar LD pattern in the two populations for the SNPs located in the genomic window considered.

2.6. Haplotype sharing

This analysis was done with the SNP data phased with Beagle5.2 (Browning et al. 2018). Sliding windows were defined along the chromosomes and linkage groups, with an overlap of 50%. Fifteen window sizes were used for physical distances, from 10 Mbp to

100 bp, and seven window sizes were used for genetic distances, from 10 cM to 0.01 cM. The window sizes were considered by decreasing order and, for each window of given window size, the list of haplotypes existing in each population was made after discarding the haplotypes with an actual length shorter than the next window size. In the end, to avoid redundancy that could result from the overlap between windows, only a single copy of the duplicated haplotypes (i.e. haplotypes identical in sequence and starting at the same position) was kept. Finally, the length of the haplotypes, the percentage of haplotypes common to the two populations and, for the common haplotypes, their frequency in each population were computed. This analysis was done using a custom R script.

2.7. Effective size

The effective size was estimated with the LD method of (Waples and Do 2008) implemented in the NeEstimator 2.1 software (Do et al. 2014). The computation was made separately in each population using the SNPs located on the genetic map and the assumption of random mating. The confidence interval of N_e values was obtained by the Jackknife method on samples.

2.8. Fixation index

Pairwise F_{st} between Deli and La Mé was estimated according to (Wright 1931), using the 7,324 or 5,598 SNPs available with MAF>1% and subsets of 100 random individuals per population, to avoid a bias in computing the F_{st} values between an unequal number of genotyped individuals per population (Gondro et al. 2013). The F_{st} was obtained using the SNPRelate R package (Zheng et al. 2012).

3. Results

3.1. Allele and genotype frequencies

The distribution of minor allele frequency (MAF) in Deli and La Mé oil palm populations showed a reduction in the number of SNPs with the increase of MAF (Figure 1). The average MAF was 0.09 for Deli and 0.14 for La Mé. In both populations, most SNPs had low MAF values. Thus, the percentage of SNPs with MAF <0.05 was 60.5% in Deli and 49.7% in La Mé.

The percentage of heterozygosity per individual ranged from 1.9% (Deli) to 20.9% (La Mé) (Figure 2). Deli was the population with the lowest percentage of heterozygosity (mean 7%, versus 10% for La Mé).

The correlation of heterozygosity per SNPs between the two populations (Figure 3) showed that the majority of SNPs were, in one population, fixed or almost fixed (i.e. concentrated alongside the *x* and *y* axes) while, in the other population, they had a much larger level of heterozygosity.

Similarly, the correlation in the frequency of alternate alleles per SNP between populations demonstrated that most SNPs have distinct segregation patterns among populations, with SNPs largely concentrated alongside the *x* and *y* axes (Figure 4). A large proportion of SNPs thus appeared fixed or almost fixed with the reference allele in one population (i.e. frequency of alternate allele equal or close to zero) while having a significant proportion of the alternate allele in the other population.

3.2. High-density genetic map

The genetic map comprised 4,252 SNPs, spread over 2,782 unique positions (Table 2, Figure 5, Supplementary Figure 2), and spanned 1,778.52 cM. Even coverage of the genome was achieved, with an average mapping interval between adjacent SNPs of 0.67 cM and the largest gaps between adjacent markers ranging from 3.31 cM (LG11) to 6.66 cM (LG14). The size of the LGs ranged from 215.72 cM to 64.75 cM (Table 2).

Table 1
Summary of the physical map (SNPs located on the assembled part of the genome)

Chromosome Name	Number of Markers	Length (bp)	Average Distance of Markers (bp)	Maximum distance of Markers (bp)	Minimum distance of markers (bp)
EG51_1	271	65,071,148	2,409,88.92	4,189,850	1
EG51_2	311	63,345,076	202,765.00	5,820,505	1
EG51_3	318	58,158,439	182,741.28	2,511,138	1
EG51_4	257	42,716,717	163,396.59	7,354,398	1
EG51_5	222	55,995,026	250,540.60	4,155,834	1
EG51_6	154	43,622,229	282,049.04	5,714,930	1
EG51_7	219	51,181,318	232,528.43	3,220,718	1
EG51_8	162	31,376,194	194,283.67	1,759,800	1
EG51_9	79	21,017,043	269,303.63	5,020,104	1
EG51_10	154	39,935,972	260,564.44	2,279,224	1
EG51_11	133	28,384,088	198,092.74	2,810,164	1
EG51_12	132	30,035,350	192,305.95	4,702,868	1
EG51_13	90	37,835,912	418,385.07	4,660,806	1
EG51_14	122	23,067,684	187,621.79	2,011,138	1
EG51_15	66	25,884,061	393,063.51	3,063,683	1
EG51_16	92	23,929,541	237,246.78	1,759,080	1
Sum	**2782**	**641,555,798**			
Mean	**173.87**	**40,097,237.38**			

Table 2
Summary of the genetic map and comparison with the physical map

Linkage group	Number of markers	Length in cM	Average gap size (cM)	Biggest gap size (cM)	Number of unique positions	Corresponding chromosome (Singh et al. 2013)	Number of common markers	Spearman correlation (absolute value)	Recombination rate (cM/Mb)
LG1	554	215.72	0.60	5.20	358	EG51_2	271	0.86	2.19
LG2	436	142.59	0.51	6.42	279	EG51_1	311	0.83	3.41
LG3	432	155.39	0.50	4.75	309	EG51_3	318	0.80	2.67
LG4	326	129.51	0.60	4.91	218	EG51_7	257	0.95	2.53
LG5	312	142.82	0.64	5.09	223	EG51_4	222	0.72	3.34
LG6	278	111.51	0.68	4.35	164	EG51_6	154	0.94	2.56
LG7	277	142.75	0.69	5.04	207	EG51_5	219	0.94	2.55
LG8	220	94.21	0.66	5.70	144	EG51_10	162	0.79	2.36
LG9	225	88.64	0.88	6.04	102	EG51_16	79	0.54	3.70
LG10	216	113.85	0.76	4.92	150	EG51_8	154	0.91	3.63
LG11	204	90.64	0.63	3.31	144	EG51_12	133	0.71	3.02
LG12	185	65.27	0.54	3.80	123	EG51_11	132	0.97	2.30
LG13	163	81.31	0.84	4.95	98	EG51_9	90	0.86	3.87
LG14	158	72.31	0.84	6.66	87	EG51_14	122	0.90	3.13
LG15	136	67.25	0.68	4.40	100	EG51_13	66	0.93	1.78
LG16	130	64.75	0.70	4.54	93	EG51_15	92	0.96	2.50
Sum	**4252**	**1778.52**			**2799**		**2782**		
Mean	**265.75**	**111.15**	**0.67**	**5.00**	**174.93**		**173.875**	**0.85**	**2.85**

3.3. Comparison of genetic and physical maps

The physical and genetic orders were in general agreement, with a Spearman's rank correlation above 0.7 for 15 LGs out of 16 (Table 2, Figure 7). However, upturns of large chromosome segments between the genetic map and the reference genome existed in a few cases, for example in chromosome EG51_16 (Figure 7). Also, punctual disagreements between physical and genetic distances concerning a few SNPs appearing as outliers, i.e. far apart from the regression line, were observed in most chromosomes.

The recombination rate was 2.85 cM/Mbp on average, ranging from 1.78 cM/Mbp (LG15) to 3.87 cM/Mbp (LG13) (Table 2).

3.4. Within population linkage disequilibrium and persistence between populations

The decay of LD between pairs of SNPs according to the genetic distances is shown in Figure 8. The LD reached high values (>0.6) for short distances between SNPs. It was higher in Deli than in La Mé for all distances. For example, considering the r^2 value of 0.3, the corresponding distance between SNPs was 1.05 cM in Deli and 0.9 cM in La Mé (Figure 8). The difference between the two populations was small for short distances and increased with the distance between markers. Similar trends were observed when plotting LD against physical distances (Figure 9), although the r^2 values reached higher levels (i.e. around 0.80), as a consequence of the higher number of markers on the physical map than on the genetic map. The distance corresponding to $r^2=0.3$ was 0.22 Mbp in Deli and 0.21 Mbp in La Mé.

A high correlation of r values between populations (r_{LD}) was observed for close markers, i.e. r_{LD} above 0.6 for SNPs separated by a distance <0.5 cM on the genetic map or <1 kbp on the physical map (Figure 10). The r_{LD} value decreased sharply with the distance between SNPs, and was thus divided by two before 2 cM and 5 Mbp, and -became negligible at distances above 50 cM or 50 Mbp.

3.5. Haplotype sharing

The percentage of shared haplotypes between Deli and La Mé populations according to the length of the genomic window is represented in Figure 11 and Figure 12. A large proportion of haplotypes were common between pairs of populations when considering short distances. Thus, 50% of the haplotypes with length around 30 bp (Figure 11) and 40% of the haplotypes with length around 3,600 bp were common to the two populations, and 40% of the haplotypes with length around 0.20 cM were common to the two populations (Figure 12). As expected, when the length of the haplotypes increased, the percentage of shared haplotypes between populations decreased. The decrease was fast, with the percentage of common haplotypes falling below 20% for haplotypes longer than 300 kbp and 2.5 cM.

The frequency of the common haplotypes coincided to some extent for short haplotypes, while the differences increased for longer haplotypes. Thus, among the common haplotypes identified with a window size of 100 bp, more than one half (51.6%) of the ones with a frequency >90% in Deli also had a frequency >90% in La Mé. This value fell to 25% for haplotypes identified with a window size of 50 kbp and to 14% for a window size of 500 kbp.

3.6. Effective size

The two populations had small N_e values, i.e. 3 for Deli (95% confidence 2.7-3.3) and 3.6 for La Mé (3.0-5.2).

3.7. Fixation index

The F_{st} between Deli and La Mé was 0.53. Figure 14 showed the F_{st} between the two populations at the chromosome level. Depending on the region of the genome considered, there were large variations in the F_{st} among the two pairs of populations. Thus, several regions of the genome had high F_{st} values (>0.6), in particular on chromosomes EG51_2, EG51_8, and EG51_13.

4. Discussion

In this paper, we used GBS data to characterize the genome properties of two key oil palm parental populations of hybrid breeding, i.e. Deli and La Mé. We constructed a high-density genetic map based on these complex populations with numerous families of different sizes and degrees of relatedness. It comprised 4,252 SNPs and spanned 1,778.52 cM, with an average recombination rate of 2.85 cM/Mbp. The LD (r^2=0.3) spanned over 1.05 cM/0.22 Mbp in Deli and 0.9 cM/0.21 Mbp in La Mé. There was a significant level of resemblance between Deli and La Mé over short genomic distances. Thus, for SNPs separated by 100 bp, the persistence of LD and linkage phases between the two populations was high (r_{LD}>0.6, and >40% haplotypes in common, with similarities in terms of haplotype frequencies). This resemblance decreased with the distance between SNPs, with for example the percentage of common haplotypes falling below 20% for haplotypes longer than 300 kbp. Overall, the F_{st}, correlation of heterozygosity per SNP, and correlation of frequency of alternate allele per SNP showed strong genetic differentiation between Deli and La Mé, confirming previous studies on oil palm genetic diversity (e.g. Cochard et al. (Cochard et al. 2009)). The N_e values of the two populations were below 5.

4.1. Within population linkage disequilibrium and persistence between populations

The pattern of LD is one of the utmost factors affecting both GWAS and GS since both methods rely on LD between markers and causal polymorphisms (Sorkheh et al. 2008; Hayes et al. 2009; Yadav et al. 2021). LD is thus one of the major factors that determine the number of markers required (Heffner et al. 2009; Lebedev et al. 2020). r^2 values of 0.3 are considered as a minimum to get reliable results in GS and GWAS studies (Bejarano et al. 2018). Here, when considering the genetic distances, the r^2 value reached 0.3 with SNPs separated by around 1.05 cM in Deli and 0.9 cM in La Mé (Figure 8). As our genetic map spanned 1,778.52 cM, achieving this distance between adjacent SNPs requires around 1,700 SNPs for Deli and 2,000 SNPs for La Mé. When considering the physical distances, the r^2 value of 0.3 was achieved with SNPs separated by around 220 kbp in Deli and 210 kbp in La Mé (Figure 9). As here the genome length covered by SNPs spanned 643 Mbp, achieving these distances between adjacent SNPs would take around 2,900

SNPs in Deli and 3,100 SNPs in La Mé, which can be considered close to the value obtained from the LD decay along with the genetic map. Considering that the goal should be to cover the whole genome and that the oil palm genome spans 1.8 gigabases (Singh et al. 2013), 10,000 SNPs would be enough to reach the r^2 value of 0.3 in the two populations studied here (as this corresponds to around 8,200 SNPs in Deli and 8,600 La Mé). The effect of marker density on the GS accuracy has already been evaluated on oil palm datasets comprising the populations considered here. It showed that, depending on the study and trait, the number of SNPs required to achieve maximum GS accuracy was found to range from 500 to 7,000 (Cros et al. 2017; Nyouma et al. 2020). This is in agreement with the results obtained from the LD analysis.

Our results also revealed that the speed and the magnitude of LD decay varied between the breeding populations. The two populations were submitted to a founding bottleneck of similar magnitude. A bottleneck increases LD and slows down the LD decline (Tenaillon et al. 2008). We can assume the higher value of LD in the Deli population in all genomic distances resulted from the fact that its history was marked by a larger number of generations marked by selection and inbreeding than in La Mé, with the bottleneck event in the Deli history dating back to 1848 against the 1920s in La Mé.

High correlation of r values between populations (r_{LD}>0.6, corresponding to r_{LD}^2>0.25) were obtained considering the markers that were the closest from each other, i.e. with distances <0.5 cM on the genetic map or <1 kbp on the physical map. Similarly, a large proportion of haplotypes was common between Deli and La Mé when considering windows of reduced size, with >40% of haplotypes with lengths below around 3,600 bp or 0.20 cM being common to the two populations. This explains the results of (Nyouma et al. 2020) and (Nyouma et al. 2021), who found, using the same breeding populations and the same genotyping approach (GBS), that for GS predictions in oil palm it was better not to model the parental origin of marker alleles. The superiority of GS models ignoring the parental origin of marker alleles over models considering it does not imply a complete persistence of phases between markers and QTLs among populations. Indeed, models that consider the parental origin of marker alleles are more complex and require the estimation of more parameters, possibly reducing their predictive ability, despite their ability to better depict the genetic differences between the population. The current study and the previous results of (Nyouma et al. 2020) and (Nyouma et al. 2021) indicate that the level of conservation of phases among the Deli and La Mé populations captured with the present marker density is high enough to favor models ignoring the parental origin of marker alleles. To our knowledge, this is the first study investigating the persistence of LD and phases between oil palm populations.

Other studies investigated the pattern of LD in oil palm, in particular (Kwong et al. 2016) and (Teh et al. 2016), using high-density SNP arrays. However, the results are difficult to compare, as the studies involved different populations, in particular inter-group hybrids, against parental populations in our study. However, (Kwong et al. 2016) included in their work two breeding populations, JL×DA and GM×DA, that were mostly of Deli origin. Their LD value decreased by 50% after around 25 kbp to 200 kbp, i.e. in the same range as the value found in our study (around 175 kbp). A previous study considered the same breeding populations as in the present study but used SSR markers (Cochard 2008). The results were however in agreement, with Deli having the highest LD values. The consistency of these results shows that GBS is a suitable approach for LD studies, despite a higher rate of missing values and genotyping errors compared to SNP arrays and SSR, while providing much higher marker density than SSRs.

4.2. Comparison of genetic and physical maps

In oil palm, for the past 20 years, many genetic linkage maps have been constructed (Seyum et al. 2021). The first genetic linkage map was constructed using RFLP markers (Mayes et al. 1997). Since then, both dominant and co-dominant molecular markers have been used for the construction of genetic linkage maps. The construction of a genetic linkage map using SNP markers is now common in oil palm (see, for instance, (Jeennor and Volkaert 2014; Ting et al. 2014; Lee et al. 2015; Bai et al. 2018a, b; Gan et al. 2018; Ong et al. 2019, 2020; Herrero et al. 2020)). Overall, the genetic linkage maps helped to identify genomic regions having major genes and quantitative trait loci (QTLs) that control oil yield (Montoya et al. 2013; Jeennor and Volkaert 2014; Tisné et al. 2015), quality traits (Singh et al. 2009; Pootakham et al. 2015; Ong et al. 2019), vegetative growth (Ukoskit et al. 2014; Lee et al. 2015; Bai et al. 2018b) and resistance to diseases (Tisné et al. 2017; Daval et al. 2021). High-density maps were also used to improve the assembly of previously published genome sequences by assigning scaffolds originally unplaced (Ong et al. 2019, 2020). To our knowledge, the present study involved the largest number of individuals genotyped for the construction of a genetic map in oil palm. Another original aspect of our genetic map is the use of complex plant material including several families with varying degrees of relatedness, several generations, and different populations. By contrast, the previously published oil palm genetic maps were usually constructed from full-sib families (e.g. (Watson et al. 2001; Cochard et al. 2009; Ting et al. 2013; Ukoskit et al. 2014; Ong et al. 2020)),

although (Billotte et al. 2010) used a factorial design. To our knowledge, only (Cochard et al. 2009) and (Daval et al. 2021) constructed genetic maps from populations with similar levels of complexity. However, they used SSR markers and the CRI-MAP software (Green et al. 1990), which seems less efficient than LepMAP3, as it has problems handling large pedigrees with large numbers of bi-allelic markers particularly when there are lots of missing parental and grandparental genotypes.

The map obtained here spanned a total length of 1,778.52 cM, which is higher than the length of previously published genetic maps in oil palm made with SNPs markers and LepMap software. For example, (Herrero et al. 2020) obtained a map spanning 1,370 cM using a Cameroon×Nigeria cross and SNPs from single primer enrichment technology, and (Ong et al. 2019, 2020) obtained maps of 1,151.7 cM, 1,268.26 cM and 1,646.95 cM for Deli×AVROS, Deli Johore Labis×Nigeria and Deli×Nigeria populations, respectively, genotyped with an SNP array. The map of our study is shorter than the map of (Cochard et al. 2015), which reached 1,935 cM and was obtained using a similar oil palm population, SSR markers, and CRI-MAP software (Green et al. 1990). This might be a consequence of the marker type, as it was shown that SSRs led to inflated maps compared to SNPs (Ball et al. 2010).

The linkage map presented here, with an average marker density of one SNP in every 0.67 cM when considering unique positions, had a denser genome coverage compared to most previously published SNPs oil palm genetic linkage maps, like (Ting et al. 2014), with one marker in every 1.40 cM and (Pootakham et al. 2015) with one marker in every 1.26 cM. However, our map is less dense than the genetic linkage maps constructed by (Ong et al. 2019, 2020), with one marker in every 0.04 cM, 0.05 cM and 0.18 cM, depending on the map, and (Bai et al. 2018a), with one marker every 0.29 cM and (Herrero et al. 2020), with one marker in every 0.57 cM. Most of these variations in terms of the marker density of the genetic maps can be explained by differences in genotyping approaches and the size of the populations (Ferreira et al. 2006; Semagn et al. 2006; Seyum et al. 2021). Combining high throughput genotyping and populations with at least 150 individuals appears as an efficient strategy to maximize marker density, as in (Ong et al. 2019, 2020), (Bai et al. 2018a), and the present study.

There were several upturns between the genetic and physical maps (Figure 7). For example, LG 1, 2, 5 and 7 had large upturns for regions of the genome of more than 10 Mbp. Aside from potential genome assembly artifacts, this can be the consequence of genomic rearrangements between populations, as the reference genome was obtained on an individual of the AVROS oil palm population (Singh et al. 2013), which thus differed from the populations used for the genetic mapping. This aspect deserves further investigation, which could be done using population-specific genetic maps and reference genomes. This requires new data, with more genotyped individuals per population and new reference genomes.

The average recombination rate was estimated at 2.85 cM/Mb (Table 2). This value is in general agreement with the ones found by (Ong et al. 2020) considering the same reference genome as in our study, i.e. 1.75 cM/Mb, 2.50 cM/Mb and 1.93 cM/Mb in Deli×AVROS, Deli×Nigeria and Deli Johore Labis populations, respectively. Variations in recombination rate along the chromosomes were noted in some chromosomes (Figure 7). In some cases, e.g. in chromosomes EG51_5, EG51_6, EG51_10, EG51_9 and EG51_15, they led to sigmoidal curves, which are expected under the effect of a lower recombination rate in the centromeric region (Semagn et al. 2006; Ong et al. 2020). For other chromosomes, these variations led to segments with lower SNP density compared to the rest of the chromosome and that corresponded to centromeric regions identified by (Singh et al. 2013). This was for example the case 10 to 15 Mbp region without a marker in chromosome EG51_11 and EG51_12.

4.3. Genetic differentiation between Deli and La Mé

In the F_{st} study, the correlation of heterozygosity per SNP, the correlation of frequency of alternate allele per SNP, and the decrease of persistence of LD and of haplotype sharing with increasing SNP distance showed a significant degree of differentiation among the two oil palm breeding populations. This is in agreement with the result of (Cochard et al. 2009), who concluded that the Deli population derived from a group comprising Benin, Nigeria, Cameroon, Congo and Angola populations, while the populations at the West of Benin were genetically more different from Deli. This result supports the idea that the four founders of the Deli population were collected in Central Africa rather than in West Africa (Cochard et al. 2009).

4.4. Effective size

To our knowledge, there was so far no estimate available of N_e for the La Mé breeding population. The small values obtained here for the Deli and La Mé populations are not surprising given their history, with a small number of founders and under the effect of inbreeding. In (Cros et al. 2014), N_e was estimated for a subset of 104 Deli individuals from the population used here, with 16 SSR

markers chosen on different linkage groups and the LD method of (Waples and Do 2008). This gave a N_e of 5 ± 1.1 (SD), i.e. similar to the result we obtained here. This indicates the robustness of the method against marker type and density.

5. Conclusion

The present study focused on key parameters affecting GS accuracy. A high-density genetic map was constructed from a complex population including several families with varying sizes and levels of relatedness and with different genetic backgrounds. It included 4,252 SNPs from GBS and spanned 1,778.52 cM, with an average recombination rate of 2.85 cM/Mbp. The LD at r^2=0.3, considered as the minimum to get reliable results for genomic predictions, spanned over 1.05 cM/0.22 Mbp in Deli and 0.9 cM/0.21 Mbp in La Mé. In the two populations, 10,000 SNPs would be enough to reach this level of LD. A high correlation of r values of LD between populations (r_{LD}>0.6) was obtained considering the markers separated by short distances, i.e. <0.5 cM on the genetic map or <1 kbp on the physical map. The percentage of common haplotypes was above 40% for short haplotypes (3,600 bp or 0.20 cM). This resemblance decreased with the distance between SNPs, with for example the percentage of common haplotypes falling below 20% for haplotypes longer than 300 kbp. The F_{st}, was high (0.53). Overall, the results showed strong genetic differentiation between Deli and La Mé, but the level of resemblance between them over short genomic distances likely explains the superiority of GS models ignoring the parental origin of marker alleles over models taking this information into account. The N_e values of the two populations were small (<5). Population-specific genetic maps and reference genomes are needed for future studies.

Declarations

Data availability

The datasets are available from the corresponding author on reasonable request and with the permission of PalmElit.

Author Contributions

EGS, DC, NHB and JMB participated in the design of the study. EGS and DC performed the statistical analysis and wrote the manuscript. WGA, NHB, BC, FJ and JMB contributed to the manuscript. PR and DL participated in the construction of the genetic linkage map. DA, HD and BC participated in designing field experiments, producing the plant material and managing field trials. The molecular data were generated by VR and VP. All authors read and approved the final manuscript.

Funding

This work was funded by the GENES Intra-Africa Academic Mobility Scheme of the European Union (EU-GENES:2017-2552/001-001) program and by a grant from PalmElit SAS.

Acknowledgments

The authors acknowledge the GENES program of the Intra-Africa Academic Mobility Scheme of the European Union for financial support (EU-GENES:2017-2552/001-001). The authors acknowledge SOCFINDO (Indonesia), CRAPP (Benin), and PalmElit (France) for authorizing the use of the data for this study. We thank the UMR AGAP genotyping technology platform (CIRAD, Montpellier) and the CIRAD-UMR AGAP HPC data center of the South Green Bioinformatics platform (http://www.southgreen.fr/) for their help.

Conflict of Interest Statement

The authors declare that there is no potential conflict of interest.

References

1. Babu BK, Mathur R (2016) Molecular breeding in oil palm (Elaeis guineensis): Status and Future perspectives. Progress Hortic 48:123–131
2. Bai B, Wang L, Zhang YJ, et al (2018a) Developing genome-wide SNPs and constructing an ultrahigh-density linkage map in oil palm. Sci Rep 8:691. https://doi.org/10.1038/s41598-017-18613-2
3. Bai B, Zhang YJ, Wang L, et al (2018b) Mapping QTL for leaf area in oil palm using genotyping by sequencing. Tree Genet Genomes 14:31. https://doi.org/10.1007/s11295-018-1245-1
4. Ball AD, Stapley J, Dawson DA, et al (2010) A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (Taeniopygia guttata). BMC Genomics 11:1–15
5. Ballesta P, Maldonado C, Pérez-Rodríguez P, Mora F (2019) SNP and Haplotype-Based Genomic Selection of Quantitative Traits in Eucalyptus globulus. Plants 8:331. https://doi.org/10.3390/plants8090331
6. Barcelos E, Rios S de A, Cunha RN, et al (2015) Oil palm natural diversity and the potential for yield improvement. Front Plant Sci 6:190
7. Basiron Y (2007) Palm oil production through sustainable plantations. Eur J Lipid Sci Technol 109:289–295. https://doi.org/10.1002/ejlt.200600223
8. Bejarano D, Martínez R, Manrique C, et al (2018) Linkage disequilibrium levels and allele frequency distribution in Blanco Orejinegro and Romosinuano Creole cattle using medium density SNP chip data. Genet Mol Biol 41:426–433. https://doi.org/10.1590/1678-4685-GMB-2016-0310
9. Bernardo RN (2010) Breeding for quantitative traits in plants, 2nd ed. Stemma Press, Woodbury, Minn
10. Billotte N, Jourjon MF, Marseillac N, et al (2010) QTL detection by multi-parent linkage mapping in oil palm (Elaeis guineensis Jacq.). TAG Theor Appl Genet Theor Angew Genet 120:1673–1687. https://doi.org/10.1007/s00122-010-1284-y
11. Browning BL, Zhou Y, Browning SR (2018) A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet 103:338–348. https://doi.org/10.1016/j.ajhg.2018.07.015
12. Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF (2008) Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178:553–561. https://doi.org/10.1534/genetics.107.080838
13. Chang L-Y, Toghiani S, Hay EH, et al (2019) A Weighted Genomic Relationship Matrix Based on Fixation Index (FST) Prioritized SNPs for Genomic Selection. Genes 10:. https://doi.org/10.3390/genes10110922
14. Cochard B (2008) Etude de la diversité génétique et du déséquilibre de liaison au sein de populations améliorées de palmier à huile (Elaeis guineensis Jacq.)
15. Cochard B, Adon B, Rekima S, et al (2009) Geographic and genetic structure of African oil palm diversity suggests new approaches to breeding. Tree Genet Genomes 5:493–504. https://doi.org/10.1007/s11295-009-0203-3
16. Cochard B, Carrasco-Lacombe C, Pomies V, et al (2015) Pedigree-based linkage map in two genetic groups of oil palm. Tree Genet Genomes 11:. https://doi.org/10.1007/s11295-015-0893-7
17. Corley RHV (2009) How much palm oil do we need? Environ Sci Policy 12:134–139. https://doi.org/10.1016/j.envsci.2008.10.011
18. Corley RHV, Tinker PB (2016) The oil palm, 5th ed. Wiley-Blackwell, Chichester, UK
19. Cros D, Bocs S, Riou V, et al (2017) Genomic preselection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses. BMC Genomics 18:1–17
20. Cros D, Denis M, Sánchez L, et al (2015) Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). Theor Appl Genet 128:397–410. https://doi.org/10.1007/s00122-014-2439-z
21. Cros D, Sánchez L, Cochard B, et al (2014) Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population. Theor Appl Genet 127:981–994. https://doi.org/10.1007/s00122-014-2273-3
22. Cuyabano BCD, Su G, Lund MS (2014) Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15:1171. https://doi.org/10.1186/1471-2164-15-1171
23. Daval A, Pomiès V, Le Squin S, et al (2021) In silico mapping in an oil palm breeding program reveals a quantitative and complex genetic resistance to Ganoderma boninense

24. Do C, Waples RS, Peel D, et al (2014) NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Mol Ecol Resour 14:209–214. https://doi.org/10.1111/1755-0998.12157
25. Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics, 4 edition. Pearson, Harlow
26. Ferreira A, Silva MF da, Silva L da C e, Cruz CD (2006) Estimating the effects of population size and type on the accuracy of genetic maps. Genet Mol Biol 29:187–192. https://doi.org/10.1590/S1415-47572006000100033
27. Flint-Garcia SA, Thornsberry JM, Buckler IV ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374
28. Gan ST, Wong WC, Wong CK, et al (2018) High density SNP and DArT-based genetic linkage maps of two closely related oil palm populations. J Appl Genet 59:23–34. https://doi.org/10.1007/s13353-017-0420-7
29. Glaubitz JC, Casstevens TM, Lu F, et al (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLOS ONE 9:e90346. https://doi.org/10.1371/journal.pone.0090346
30. Gondro C, Werf J van der, Hayes B (eds) (2013) Genome-Wide Association Studies and Genomic Prediction. Humana Press
31. Grattapaglia D (2014) Breeding Forest Trees by Genomic Selection: Current Progress and the Way Forward. In: Tuberosa R, Graner A, Frison E (eds) Genomics of Plant Genetic Resources: Volume 1. Managing, sequencing and mining genetic resources. Springer Netherlands, Dordrecht, pp 651–682
32. Green, P., Falls, K., Crooks, S. (1990) Documentation for CRI-MAP, version 2.4
33. Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Mol Biol 57:461–485. https://doi.org/10.1007/s11103-005-0257-z
34. Hartley CWS (1988) The oil palm (Elaeis guineensis Jacq.). Longman Scientific & Technical ; Wiley, Harlow, Essex, England; New York
35. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443. https://doi.org/10.3168/jds.2008-1646
36. He J, Zhao X, Laroche A, et al (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci 5:484. https://doi.org/10.3389/fpls.2014.00484
37. Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic Selection for Crop Improvement. Crop Sci 49:1–12. https://doi.org/10.2135/cropsci2008.08.0512
38. Herrero J, Santika B, Herrán A, et al (2020) Construction of a high density linkage map in Oil Palm using SPET markers. Sci Rep 10:9998. https://doi.org/10.1038/s41598-020-67118-y
39. Ithnin M, Din AK (eds) (2020) The Oil Palm Genome. Springer International Publishing
40. Jakobsson M, Edge MD, Rosenberg NA (2013) The Relationship Between FST and the Frequency of the Most Frequent Allele. Genetics 193:515–528. https://doi.org/10.1534/genetics.112.144758
41. Jeennor S, Volkaert H (2014) 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 10:1–14. https://doi.org/10.1007/s11295-013-0655-3
42. Jin J, Lee M, Bai B, et al (2016) Draft genome sequence of an elite Dura palm and whole-genome patterns of DNA variation in oil palm. DNA Res Int J Rapid Publ Rep Genes Genomes 23:527–533. https://doi.org/10.1093/dnares/dsw036
43. Kwong QB, Teh CK, Ong AL, et al (2016) Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm. Mol Plant 9:1132–1141. https://doi.org/10.1016/j.molp.2016.04.010
44. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
45. Lebedev VG, Lebedeva TN, Chernodubov AI, Shestibratov KA (2020) Genomic Selection for Forest Tree Improvement: Methods, Achievements and Perspectives. Forests 11:1190. https://doi.org/10.3390/f11111190
46. Lee M, Xia JH, Zou Z, et al (2015) A consensus linkage map of oil palm and a major QTL for stem height. Sci Rep 5:8232. https://doi.org/10.1038/srep08232
47. Li Y, Kim J-J (2015) Effective population size and signatures of selection using bovine 50K SNP chips in Korean native cattle (Hanwoo). Evol Bioinforma 11:EBO-S24359

48. Lin Z, Hayes BJ, Daetwyler HD (2014) Genomic selection in crops, trees and forages: a review. Crop Pasture Sci 65:1177. https://doi.org/10.1071/CP13363
49. Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57–63
50. Matias FI, Galli G, Granato ISC, Fritsche-Neto R (2017) Genomic Prediction of Autogamous and Allogamous Plants by SNPs and Haplotypes. Crop Sci 57:2951–2958. https://doi.org/10.2135/cropsci2017.01.0022
51. Mayes S, Jack PL, Corley RH, Marshall DF (1997) Construction of a RFLP genetic linkage map for oil palm (Elaeis guineensis Jacq.). Genome 40:116–122. https://doi.org/10.1139/g97-016
52. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819–1829
53. Montoya C, Lopes R, Flori A, et al (2013) Quantitative trait loci (QTLs) analysis of palm oil fatty acid composition in an interspecific pseudo-backcross from Elaeis oleifera (H.B.K.) Cortés and oil palm (Elaeis guineensis Jacq.). Tree Genet Genomes 9:1207–1225. https://doi.org/10.1007/s11295-013-0629-5
54. Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? Ann Bot 110:1303–1316. https://doi.org/10.1093/aob/mcs109
55. Nyouma A, Bell JM, Jacob F, et al (2020) Genomic predictions improve clonal selection in oil palm (Elaeis guineensis Jacq.) hybrids. Plant Sci 299:110547. https://doi.org/10.1016/j.plantsci.2020.110547
56. Nyouma A, Bell JM, Jacob F, Cros D (2019) From mass selection to genomic selection: one century of breeding for quantitative yield components of oil palm (Elaeis guineensis Jacq.). Tree Genet Genomes 15:1–16
57. Nyouma, A., Bell, J.M., Jacob, F., Riou, V., Manez, A., Pomiès, V., Nodichao, L., Syahputra, I., Affandi, D., Cochard, B., Durand-Gasselin, T., Cros, D (2021) Improving the accuracy of genomic predictions in an outcrossing species with hybrid cultivars between heterozygote parents: case study of oil palm (Elaeis guineensis Jacq.) (Under Review).
58. Ong A-L, Teh C-K, Kwong Q-B, et al (2019) Linkage-based genome assembly improvement of oil palm (Elaeis guineensis). Sci Rep 9:1–9
59. Ong A-L, Teh C-K, Mayes S, et al (2020) An Improved Oil Palm Genome Assembly as a Valuable Resource for Crop Improvement and Comparative Genomics in the Arecoideae Subfamily. Plants 9:1476. https://doi.org/10.3390/plants9111476
60. Ouellette LA, Reid RW, Blanchard SG, Brouwer CR (2018) LinkageMapView—rendering high-resolution linkage and QTL maps. Bioinformatics 34:306–307. https://doi.org/10.1093/bioinformatics/btx576
61. Paterson R, Sariah M, Lima N (2013) How will climate change affect oil palm fungal diseases? Crop Prot 46:113–120
62. Pirker J, Mosnier A, Kraxner F, et al (2016) What are the limits to oil palm expansion? Glob Environ Change 40:73–81
63. Pootakham W, Jomchai N, Ruang-areerate P, et al (2015) Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). Genomics 105:288–295. https://doi.org/10.1016/j.ygeno.2015.02.002
64. Purcell S, Neale B, Todd-Brown K, et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575
65. Rastas P (2017) Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinforma Oxf Engl 33:3726–3732. https://doi.org/10.1093/bioinformatics/btx494
66. Semagn K, Bjørnstad Å, Ndjiondjop MN (2006) Principles, requirements and prospects of genetic mapping in plants. Afr J Biotechnol 5:. https://doi.org/10.4314/ajb.v5i25.56082
67. Seng T-Y, Ritter E, Mohamed Saad SH, et al (2016) QTLs for oil yield components in an elite oil palm (Elaeis guineensis) cross. Euphytica 212:399–425. https://doi.org/10.1007/s10681-016-1771-6
68. Seyum EG, Bille NH, Abtew WG, et al (2021) Genome Mapping to Enhance Efficient Marker-Assisted Selection and Breeding of the Oil Palm (Elaeis guineensis Jacq.). Adv Biosci Biotechnol 12:407–425. https://doi.org/10.4236/abb.2021.1212026
69. Siberchicot A, Bessy A, Guéguen L, Marais GA (2017) MareyMap Online: A User-Friendly Web Application and Database Service for Estimating Recombination Rates Using Physical and Genetic Maps. Genome Biol Evol 9:2506–2509. https://doi.org/10.1093/gbe/evx178
70. Singh R, Ong-Abdullah M, Low E-TL, et al (2013) Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature 500:335–339. https://doi.org/10.1038/nature12309

71. Singh R, Tan SG, Panandam JM, et al (2009) Mapping quantitative trait loci (QTLs) for fatty acid composition in an interspecific cross of oil palm. BMC Plant Biol 9:114. https://doi.org/10.1186/1471-2229-9-114
72. Slatkin M (2008) Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485. https://doi.org/10.1038/nrg2361
73. Soh AC, Mayes S, Roberts JA (eds) (2017) Oil Palm Breeding: Genetics and Genomics, 1 edition. CRC Press, Boca Raton
74. Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and densities. J Anim Sci 86:2447–2454. https://doi.org/10.2527/jas.2007-0010
75. Sorkheh K, Malysheva-Otto LV, Wirthensohn MG, et al (2008) Linkage disequilibrium, genetic association mapping and gene localization in crop plants. Genet Mol Biol 31:805–814. https://doi.org/10.1590/S1415-47572008005000005
76. Statista (2021) Vegetable oils: production worldwide 2012/13-2020/21, by type.
77. Technow F, Schrag TA, Schipprack W, et al (2014) Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. Genetics 197:1343–1355. https://doi.org/10.1534/genetics.114.165860
78. Teh C-K, Ong A-L, Kwong Q-B, et al (2016) Genome-wide association study identifies three key loci for high mesocarp oil content in perennial crop oil palm. Sci Rep 6:19075. https://doi.org/10.1038/srep19075
79. Teissier M, Larroque H, Brito LF, et al (2020) Genomic predictions based on haplotypes fitted as pseudo-SNP for milk production and udder type traits and SCS in French dairy goats. J Dairy Sci 103:11559–11573. https://doi.org/10.3168/jds.2020-18662
80. Tenaillon MI, Austerlitz F, Tenaillon O (2008) Apparent mutational hotspots and long distance linkage disequilibrium resulting from a bottleneck. J Evol Biol 21:541–550. https://doi.org/10.1111/j.1420-9101.2007.01490.x
81. Ting N-C, Jansen J, Mayes S, et al (2014) High density SNP and SSR-based genetic maps of two independent oil palm hybrids. BMC Genomics 15:309. https://doi.org/10.1186/1471-2164-15-309
82. Ting N-C, Jansen J, Nagappan J, et al (2013) Identification of QTLs Associated with Callogenesis and Embryogenesis in Oil Palm Using Genetic Linkage Maps Improved with SSR Markers. PLOS ONE 8:e53076. https://doi.org/10.1371/journal.pone.0053076
83. Tisné S, Denis M, Cros D, et al (2015) Mixed model approach for IBD-based QTL mapping in a complex oil palm pedigree. BMC Genomics 16:798. https://doi.org/10.1186/s12864-015-1985-3
84. Tisné S, Pomiès V, Riou V, et al (2017) Identification of Ganoderma Disease Resistance Loci Using Natural Field Infection of an Oil Palm Multiparental Population. G3 GenesGenomesGenetics 7:1683–1692. https://doi.org/10.1534/g3.117.041764
85. Ukoskit K, Chanroj V, Bhusudsawang G, et al (2014) Oil palm (Elaeis guineensis Jacq.) linkage map, and quantitative trait locus analysis for sex ratio and related traits. Mol Breed 33:415–424. https://doi.org/10.1007/s11032-013-9959-0
86. Wand M (1995) KernSmooth: Functions for Kernel Smoothing Supporting, Wand & Jones, R package version 2.23–20
87. Waples RS, Do C (2008) ldne: a program for estimating effective population size from data on linkage disequilibrium. Mol Ecol Resour 8:753–756. https://doi.org/10.1111/j.1755-0998.2007.02061.x
88. Watson K, Mayes S, Price Z, et al (2001) Quantitative trait loci for yield components in oil palm (Elaeis guineensis Jacq.). Theor Appl Genet 103:1302–1310. https://doi.org/10.1007/s122-001-8204-z
89. Weir BS (1979) Inferences about Linkage Disequilibrium. Biometrics 35:235–254. https://doi.org/10.2307/2529947
90. Wright S (1931) Evolution in Mendelian Populations. Genetics 16:97–159
91. Xu H, Guan Y (2014) Detecting Local Haplotype Sharing and Haplotype Association. Genetics 197:823–838. https://doi.org/10.1534/genetics.114.164814
92. Yadav S, Ross EM, Aitken KS, et al (2021) A linkage disequilibrium-based approach to position unmapped SNPs in crop species. BMC Genomics 22:1–9
93. Yamamoto E, Matsunaga H, Onogi A, et al (2016) A simulation-based breeding design that uses whole-genome prediction in tomato. Sci Rep 6:1–11
94. Yan L, Hofmann N, Li S, et al (2017) Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. BMC Genomics 18:529. https://doi.org/10.1186/s12864-017-3922-0
95. Zheng X, Levine D, Shen J, et al (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Figures

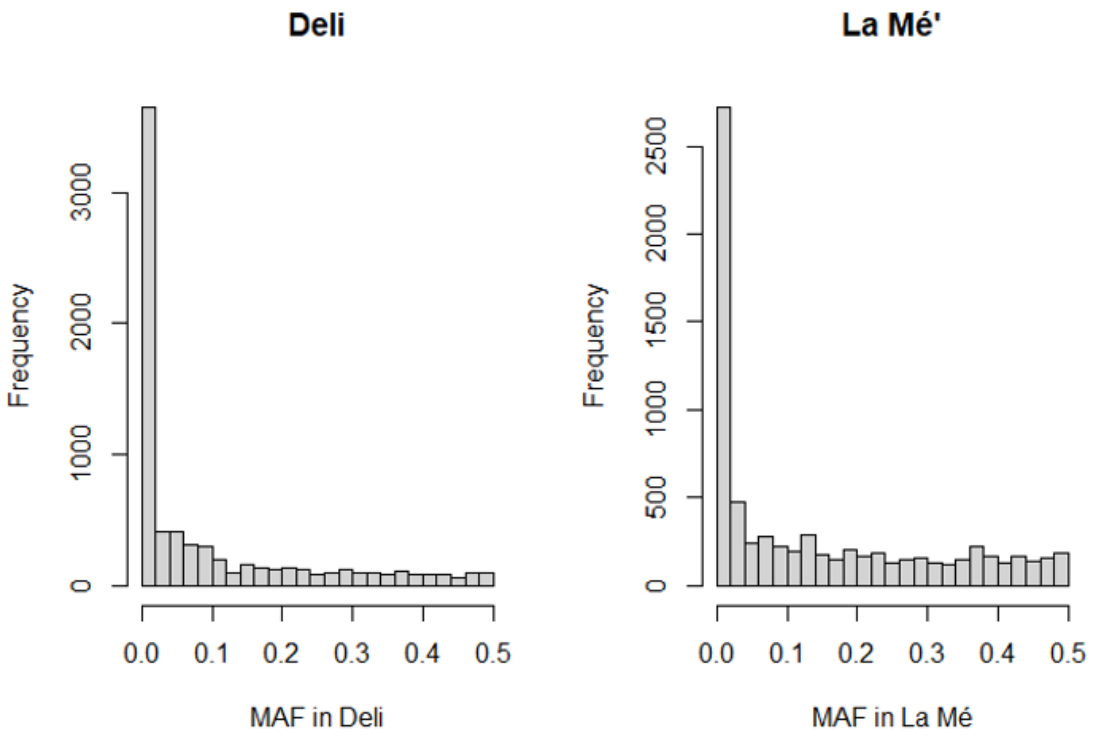

Figure 1

Distribution of minor allele frequency (MAF) in Deli and La Mé oil palm breeding populations.

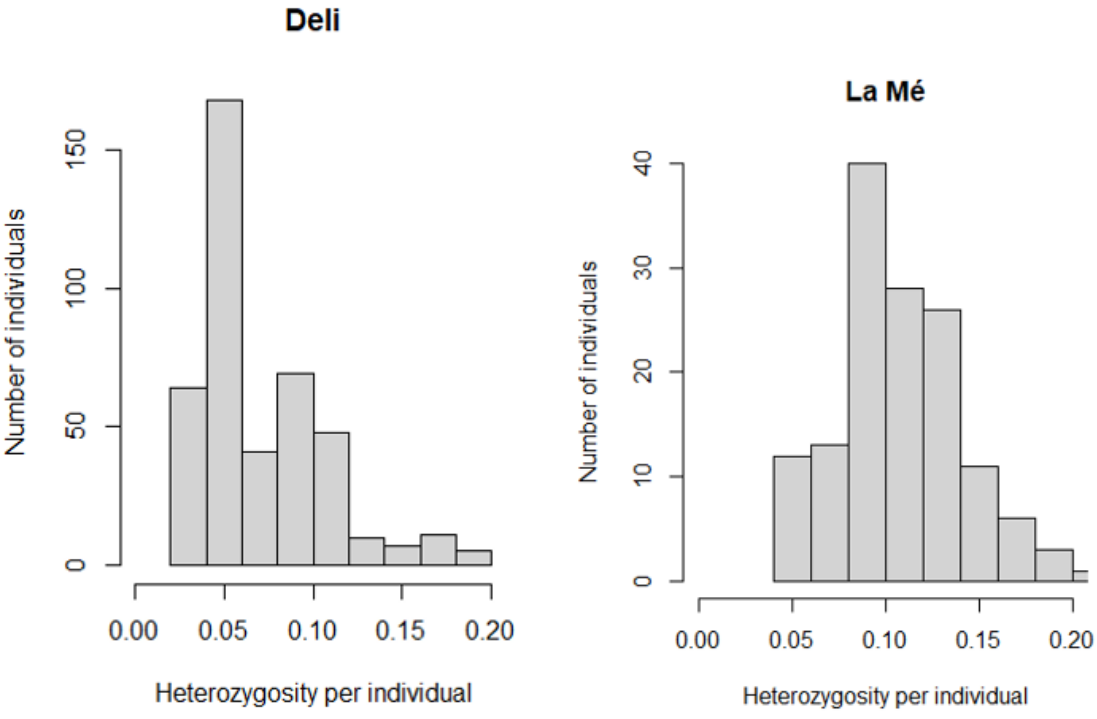

Figure 2

Distribution of the percentage of heterozygosity per individual for Deli and La Mé oil palm breeding populations.

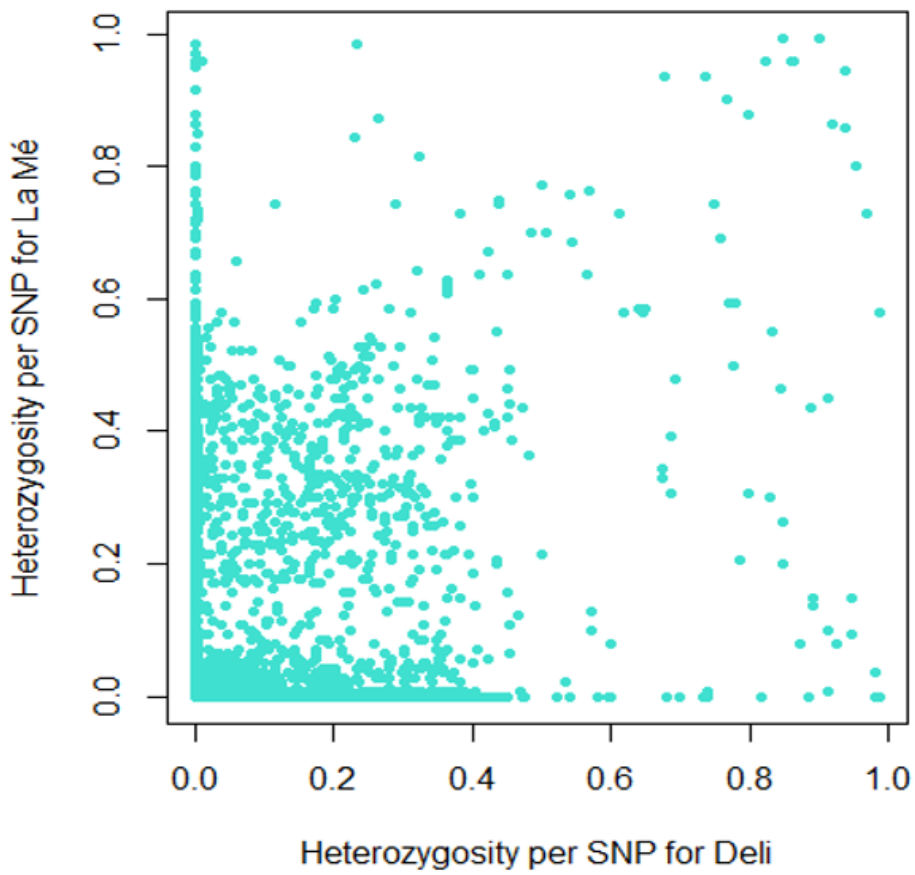

Figure 3

Correlation of heterozygosity per SNPs among Deli and La Mé oil palm breeding populations. Each dot represents an SNP. Color intensity indicates density of overlapping dots.

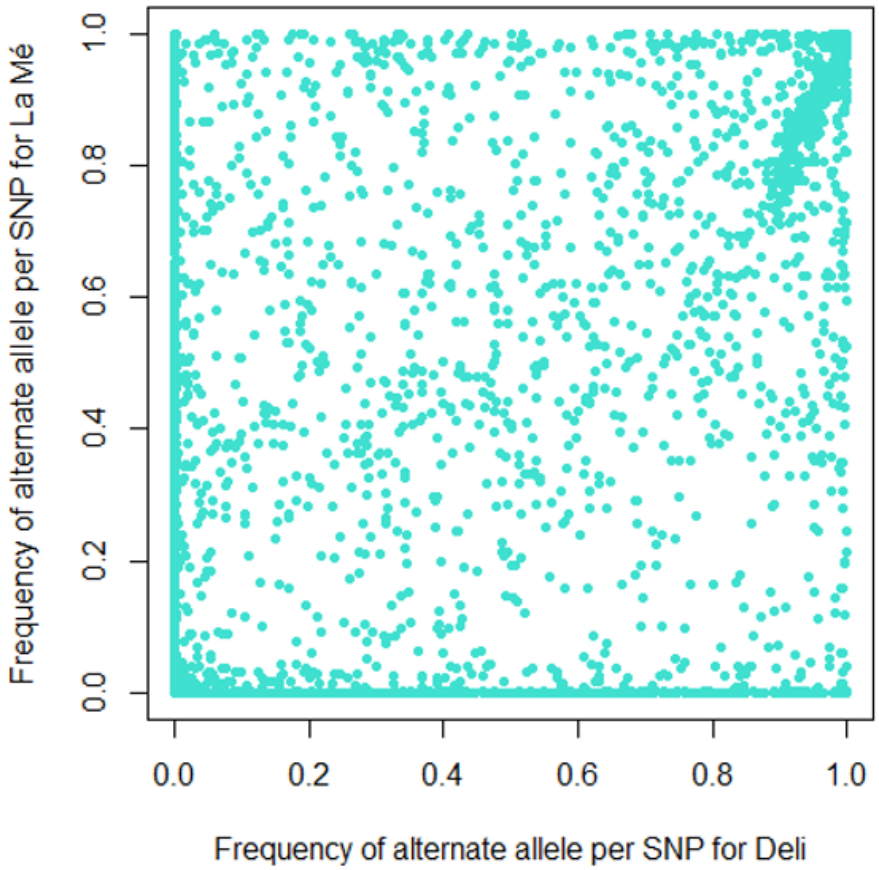

Figure 4

Correlation of frequency of alternate allele per SNP among Deli and La Mé oil palm breeding populations. Each dot represents an SNP. Color intensity indicates density of overlapping dots.

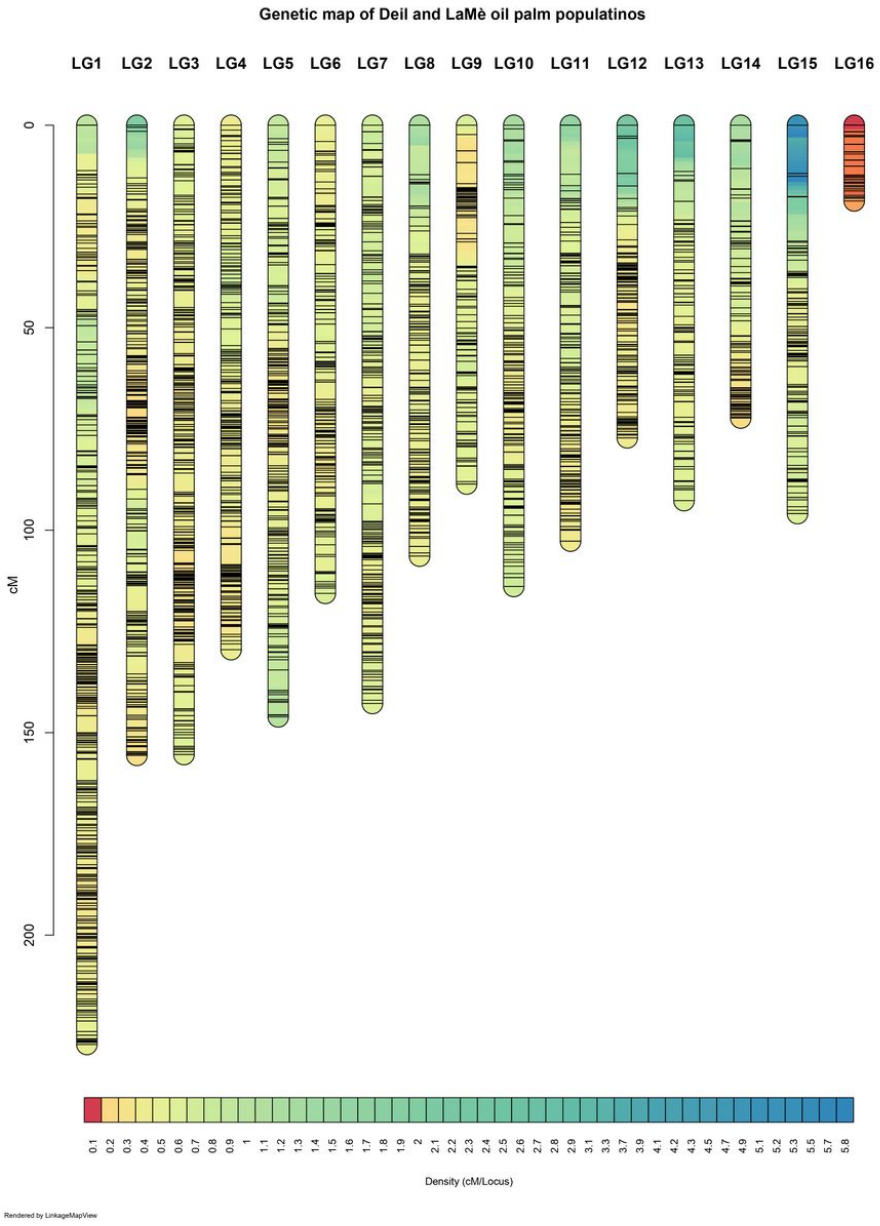

Figure 5

Genetic map of Deli and La Mé oil palm populations with 4,252 SNP markers. The colors indicate the density of markers according to the bottom scale (cM/locus).

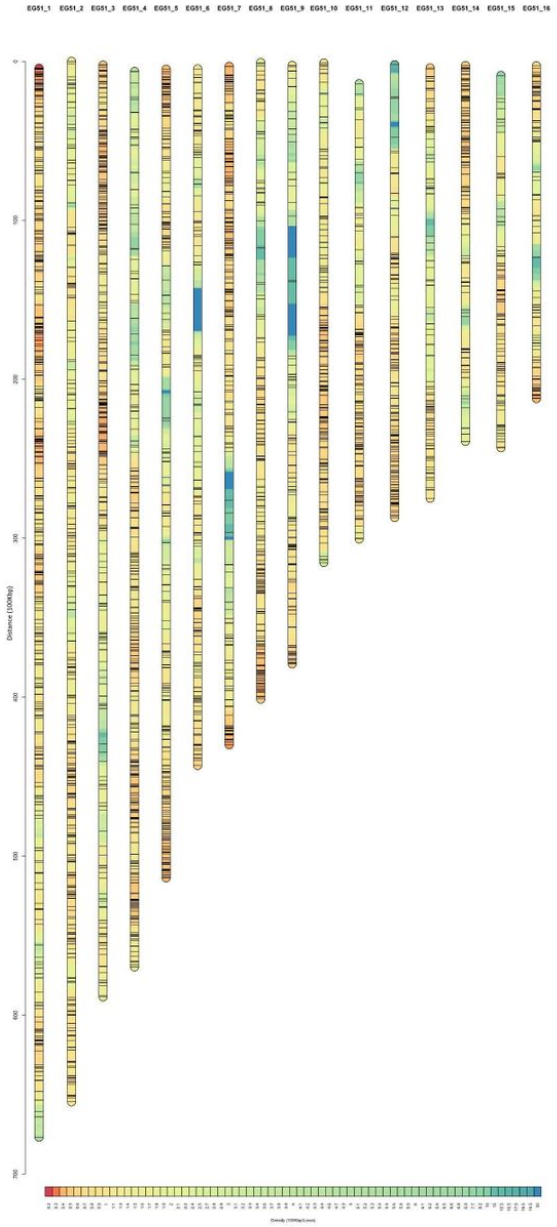

Figure 6

Physical map of Deli and La Mé oil palm populations with 5,598 SNP markers. The colors indicate the density of markers according to the bottom scale (100 kbp/locus).

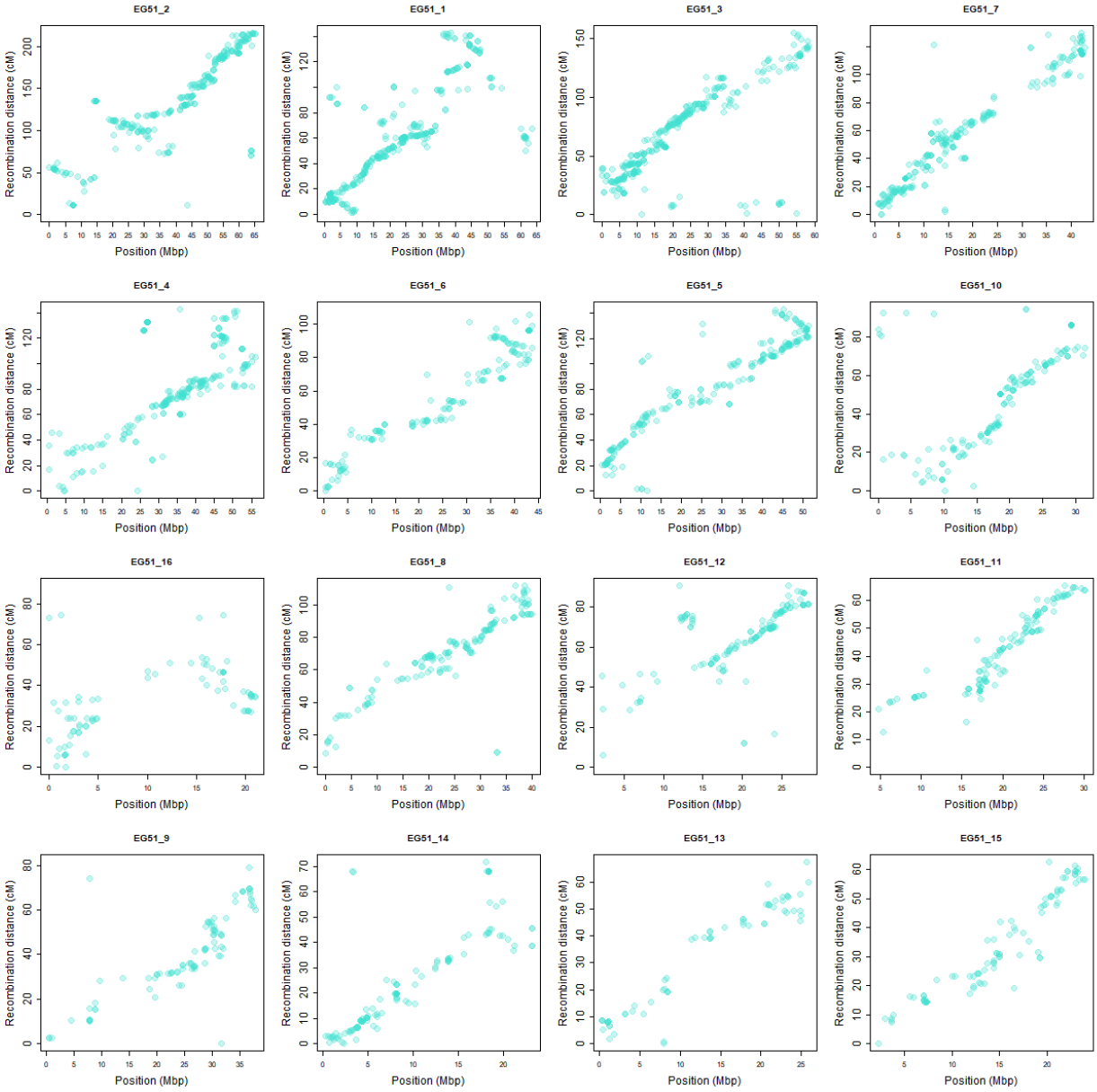

Figure 7

Visualization of marker genetic positions (cM) versus physical positions (Mbp) for each chromosome. The plots are ordered according to linkage groups number. Each dot represents an SNP. Color intensity indicates density of overlapping dots.

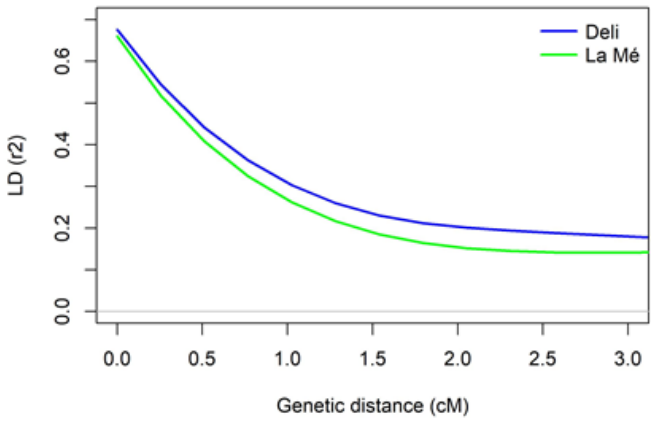

Figure 8

Average genome-wide pattern of linkage disequilibrium (LD) decay between pairs of SNPs (r^2) according to the genetic distance (cM) between SNPs, for Deli and La Mé oil palm breeding populations.

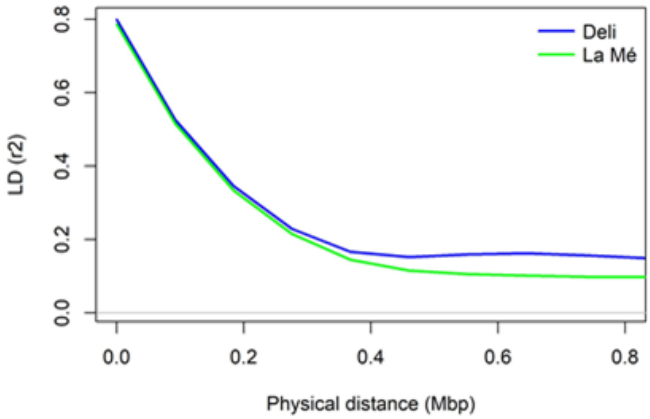

Figure 9

Average genome-wide pattern of linkage disequilibrium (LD) decay between pairs of SNPs (r^2) according to the physical distance (Mbp) between SNPs, for Deli and La Mé oil palm breeding populations

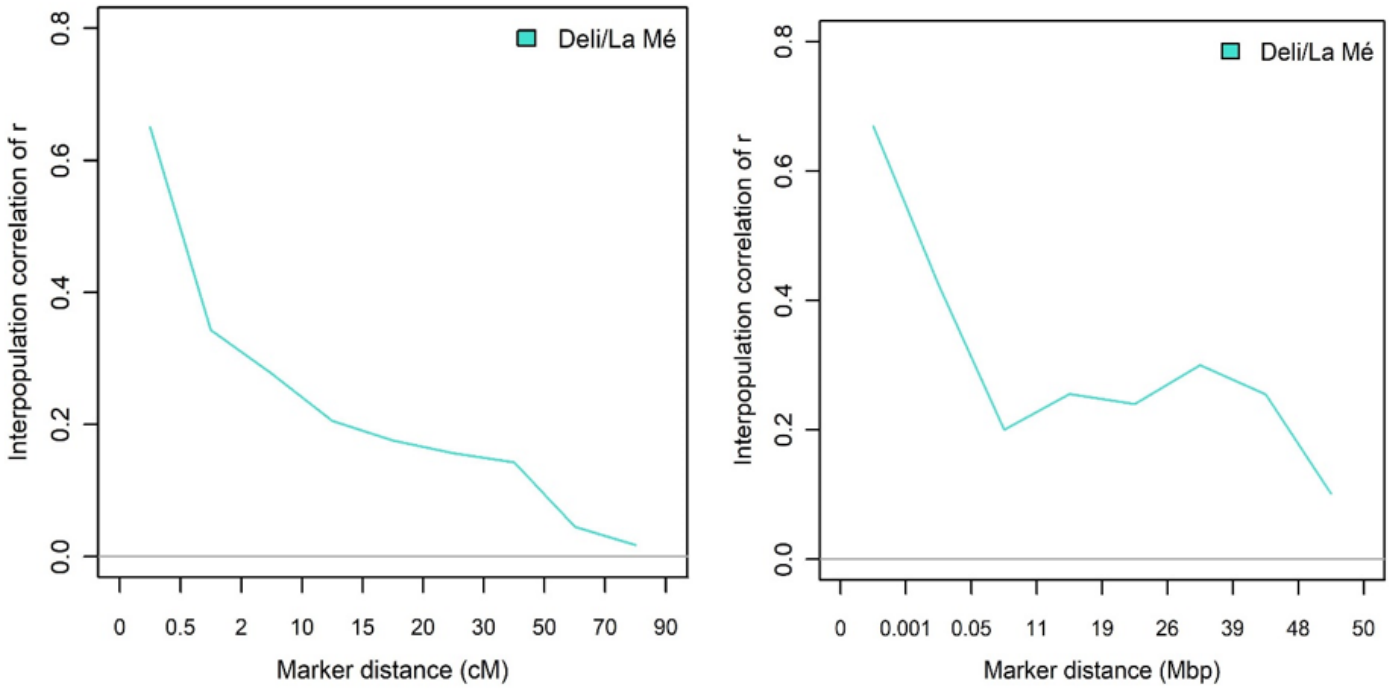

Figure 10

Correlation of the r measure of LD between populations as a function of genomic distance in cM (right) and Mbp (left).

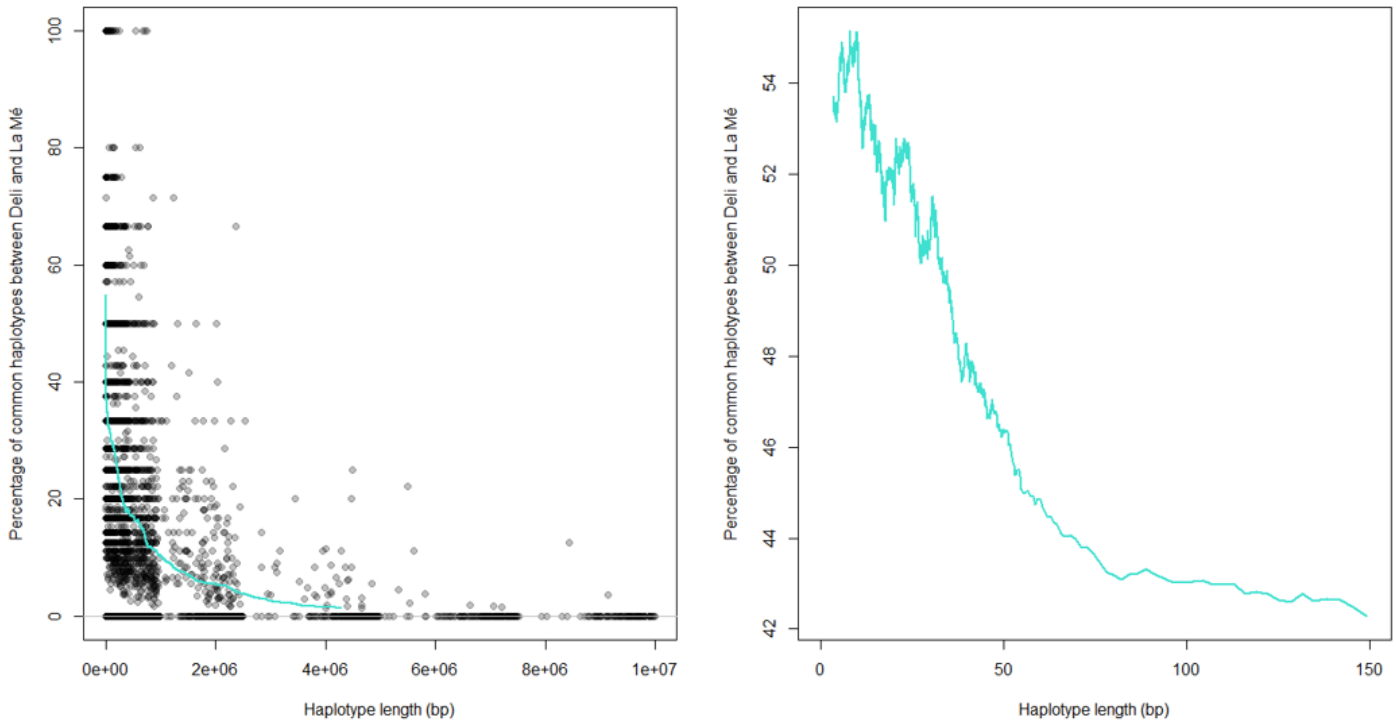

Figure 11

Percentage of common haplotypes between Deli and La Mé oil palm breeding populations according to the haplotype length in bp. Each dot represents a haplotype. Color intensity indicates density of overlapping dots. The smoothing curve in turquoise is the rolling average.

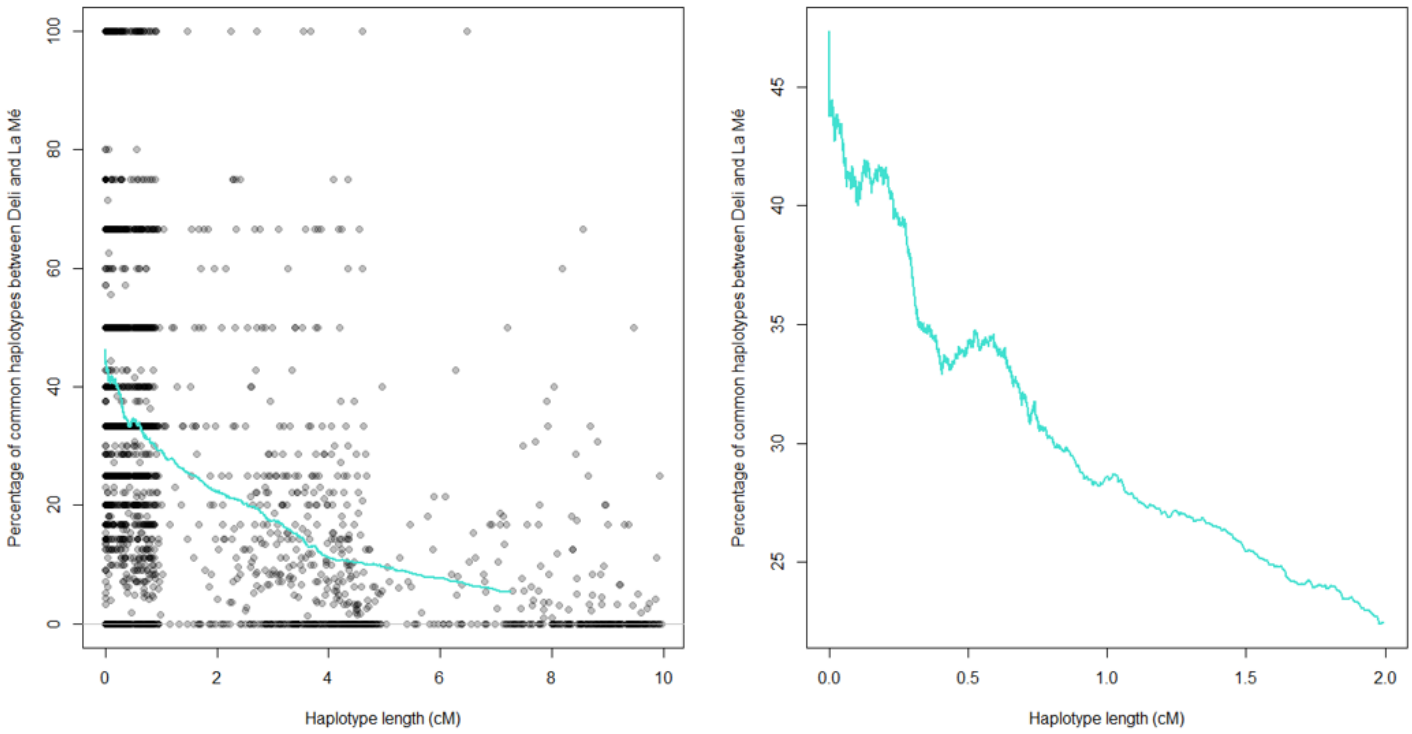

Figure 12

Percentage of common haplotypes between Deli and La Mé oil palm breeding populations according to the haplotype length in cM. Each dot represents a haplotype. Color intensity indicates density of overlapping dots. The smoothing curve in turquoise is the rolling average.

Supplementary Files

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

- SupplementaryFigure2.pdf
- Supplementaryfilesdc3.docx