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

Essubalew Getachew Seyum1,2,3 · Ngalle Hermine Bille1 · Wosene Gebreselassie Abtew2 · Pasi Rastas4 · Deni Arifianto5 · Hubert Domonhéd6 · Benoit Cochard7 · Florence Jacob7 · Virginie Riou8,9 · Virginie Pomiès8,9 · David Lopez8,9 · Joseph Martin Bell1 · David Cros8,9

Received: 28 January 2022 / Revised: 26 May 2022 / Accepted: 4 June 2022 / Published online: 13 June 2022 © The Author(s), under exclusive licence to Institute of Plant Genetics Polish Academy of Sciences 2022

Abstract

A good knowledge of the genome properties of the populations makes it possible to optimize breeding methods, in particular genomic selection (GS). In oil palm (*Elaeis guineensis* Jacq), the world’s main source of vegetable oil, this would provide insight into the promising GS results obtained so far. The present study considered two complex breeding populations, Deli and La Mé, with 943 individuals and 7324 single-nucleotide polymorphisms (SNPs) from genotyping-by-sequencing. Linkage disequilibrium (LD), haplotype sharing, effective size (*N*<sub>e</sub>), and fixation index (*F*<sub>st</sub>) were investigated. A genetic linkage map spanning 1778.52 cM and with a recombination rate of 2.85 cM/Mbp was constructed. The LD at *r*<sup>2</sup>=0.3, considered the minimum to get reliable GS results, 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 *F*<sub>st</sub> 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/1 kbp and percentage of common haplotypes >40% for haplotypes <3600 bp/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*<sub>e</sub> (<5). Population-specific genetic maps and reference genomes are recommended for future studies.

Keywords Genome properties · Genomic selection · Genotyping-by-sequencing · Oil palm · Single nucleotide polymorphisms
Introduction

The African 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, monoeccious, 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).

The total world vegetable oil production is currently around 200 million metric tons (MT), led by oil palm (37.5%), followed by soybean oil (30%), rapeseed oil (14%), and sunflower oil (9.5%) (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).

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 development of hybrid cultivars of African oil palm (Nyouma et al. 2019). Group A mostly consists of the Deli parental population, which is derived from four individuals from 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 from individuals collected in the Bingerville region of the Ivory Coast between 1924 and 1930, with three founders for the La Mé 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).

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, labor shortage, and diseases (in particular vascular wilt, ganoderma, and bud rot) 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; Kwong et al. 2016; Pirker et al. 2016). In addition, genetic improvement through the conventional method in oil palm is constrained by several factors, in particular a long breeding cycle (>15 years) and a limited number of tested individuals (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) with, for instance, the detection of quantitative trait loci (QTLs) controlling oil yield, quality, vegetative growth, and resistance to diseases (Pootakham et al. 2015; Tisné et al. 2015; Ithnin et al. 2017; Bai et al. 2018b; Daval et al. 2021) and genomic selection (Nyouma et al. 2019).

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 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). The different studies carried out have provided a significant amount of information concerning the conditions of implementation of GS in this species. For example, in Deli and La Mé, GS has been implemented with relatively small training populations (<150) and low marker density (<2000 SNPs) (Cros et al. 2017; Nyouma et al. 2020); and models ignoring the parental origin of marker alleles were found to be more accurate than models accounting for this information (Nyouma et al. 2020, 2022). To better understand the GS results in the populations involved, a detailed study of their genome properties would be of interest, particularly regarding linkage disequilibrium (LD), effective size ($N_e$), haplotype sharing, and fixation index ($F_{st}$), which are known to affect GS accuracy.

Linkage disequilibrium is defined as the non-random 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 by non-genetic factors such as marker ascertainment bias (Flint-Garcia et al. 2003; Gupta et al. 2005; Mackay and Powell 2007; Slatkin 2008).
Effective size ($N_e$) is defined as the size of an idealized Wright-Fisher population that would give rise to the same extent of random genetic drift or rate of inbreeding as the actual population (Wright 1931; Caballero 1994; Falconer and Mackay 1996). Low $N_e$ leads to high rates of genetic drift and inbreeding in a population, making $N_e$ one of the major factors influencing LD, and consequently the accuracy of GS (Grattapaglia 2014). $N_e$ can be inferred from LD (Corbin et al. 2012), as 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; Daetwyler et al. 2010; Wientjes et al. 2013; 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.

Haplotypes correspond to two or more SNP alleles that tend to be inherited as a unit in the chromosome (Bernardo 2010). Haplotype sharing helps 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 (De Roos et al. 2009; Technow et al. 2012).

The fixation index ($F_{st}$) is the correlation between gametes chosen randomly from within the same sub-population relative to the entire population (Wright 1931; Jakobsson et al. 2013; Weir and Goudet 2017). It is used to identify loci with divergent allelic frequencies between two or more populations. It helps to understand the genetic differentiation among groups. It ranges from 0 (no correlation, i.e., gametes within sub-populations are no more similar than gametes among sub-populations) to 1 (each sub-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 identify QTLs for seed weight.

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 LD, haplotype sharing, $N_e$, and $F_{st}$.

**Materials and methods**

**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. The Deli and La Mé populations used here were complex, involving several families with varying sizes and levels of relatedness. Thus, the Deli individuals belonged to 89 families of full-sibs with a mean size of 4.8 individuals (ranging from one to 60 individuals). The La Mé individuals belonged to 24 families of full-sibs with a mean size of 5.8 individuals (ranging from one to 31 individuals). Detailed pedigree information of these two populations is known over several generations (Cros et al. 2017). The Deli × La Mé hybrid individuals were obtained crossing 67 and 63 of these Deli and La Mé individuals, respectively, according to an incomplete factorial design. The hybrid individuals belonged to 101 crosses comprising on average 3.8 individuals (ranging from one to 10). For the construction of the genetic map, all the genotyped Deli, La Mé, and Deli × La Mé individuals were used, as well as the non-genotyped individuals comprised in their pedigree, for a total of 1788 individuals. 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 Pobé.

**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 7324 SNP markers, common to both breeding populations. It included 5598 SNPs located on the anchored sequences of the genome (i.e.,
the 16 chromosomes of Singh et al. (2013) (Table 1). The average percentage of missing data per SNP was 11% in Deli and 13% in La Mé, with median values below 5% (Supplementary Figure 1).

**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 the 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 were 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.

**Comparison of genetic and physical maps**

The genetic map and physical map, showing the positions on 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.

**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.

| 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              | 37,835,912  | 418,385.07                       | 4,660,806                       | 1                                |
| EG51_11         | 133              | 30,035,350  | 192,305.95                       | 4,702,868                       | 1                                |
| EG51_12         | 132              | 37,835,912  | 418,385.07                       | 4,660,806                       | 1                                |
| EG51_13         | 90               | 23,067,684  | 187,621.79                       | 2,011,138                       | 1                                |
| EG51_14         | 66               | 23,929,541  | 237,246.78                       | 1,759,080                       | 1                                |
| EG51_15         | 92               | 641,555,798 | Mean 173.87                      | 2,001,138                       | 1                                |
| EG51_16         | 77               | 30,035,350  | 192,305.95                       | 4,702,868                       | 1                                |

Table 1 Summary of the physical map (SNPs located on the assembled part of the genome)
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 (De Roos et al. 2009) 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.

### 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 to 0.01 cM. The window sizes were considered by decreasing order and, for each window of a given window size, the list

---

**Fig. 1** Distribution of minor allele frequency (MAF) in Deli and La Mé oil palm breeding populations

**Fig. 2** Distribution of the percentage of heterozygosity per individual for Deli and La Mé oil palm breeding populations
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.

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 selecting the “random mating” option of the software. The confidence interval of \( N_e \) values was obtained by the Jackknife method on samples.

Fixation index

Pairwise \( F_{st} \) between Deli and La Mé was estimated according to Wright (1931), using the 7324 or 5598 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).

Results

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 (Fig. 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) (Fig. 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 (Fig. 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 (Fig. 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 0), while having a significant proportion of the alternate allele in the other population.

**High-density genetic map**

The process of construction of the genetic map yielded a set of 16 linkage groups (LGs). The genetic map comprised 4252 SNPs, spread over 2782 unique positions (Table 2 and Fig. 5, Supplementary Figure 2), and spanned 1778.52 cM. Even coverage of the genome was achieved, with an average mapping interval between adjacent positions of 0.67 cM and the largest gaps between adjacent positions ranging from 3.31 cM (LG11) to 6.66 cM (LG14). The size of the LGs ranged from 215.72 cM (LG1) to 64.75 cM (LG16) (Table 2). The number of unique SNP positions mapped to each linkage group ranged from 87 (LG 14) to 358 (LG1), with a mean of 174.93 per linkage group. The biggest gap size between SNPs ranged from 3.31 cM (LG11) to 6.66 cM (LG14).

**Comparison of genetic and physical maps**

The physical and genetic orders were in general agreement, with a Spearman rank correlation above 0.7 for 15 LGs out of 16 (Table 2 and Fig. 6). 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 (Fig. 6). 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 (Fig. 7).

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). The decay of LD between pairs of SNPs according to the genetic distances is shown in Fig. 8. The LD reached high values (>0.6) for short distances between SNPs. It was higher in Deli than that 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 (Fig. 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 (Fig. 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.

---

**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           | 229               | 94.21        | 0.66                  | 5.70                  | 144                        | EG51_10                                    | 620                        | 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                      |
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 (Fig. 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.

**Haplotype sharing**

The percentage of shared haplotypes between Deli and La Mé populations according to the length of the genomic window is represented in Figs. 11 and 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 (Fig. 11) and 40% of the haplotypes with length around 3600 bp were common to the two populations, and 40% of the haplotypes with length around 0.20 cM were common to the two populations (Fig. 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.

**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).

**Fixation index**

The $F_{st}$ between Deli and La Mé was 0.53. Supplementary Figure 3 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.

**Discussion**

In this paper, we characterized the genome properties of two key oil palm breeding populations, Deli and La Mé, using SNP data obtained by GBS. This genotyping approach, despite the higher rate of missing data and errors than SNP array and SSR, appears relevant for the present study, as indicated by previous articles on other species where genetic maps and LD profiles were obtained from GBS, for instance in coconut (Rajesh et al. 2021), hevea (de Souza et al. 2018), and tea (Niu et al. 2019). Also, the Lep-MAP3 software used here to generate the genetic map is particularly robust against missing data, as it does not rely on two-point analysis, and against genotypes of lower reliability due to low coverage sequencing (Rastas 2017).

**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 a minimum to get reliable results in GS studies and GWAS (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 (Fig. 8). As our genetic map spanned 1778.52 cM, achieving this distance between adjacent SNPs requires around 1700 SNPs for Deli and 2000 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 (Fig. 9). As here the genome length covered by SNPs spanned 643 Mbp, achieving these distances between adjacent SNPs would take around 2900 SNPs in Deli and 3100 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 8200 SNPs in Deli and 8600 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 7000 (Cros et al. 2017; Nyouma et al. 2020). This is in agreement with the results obtained from the LD analysis.
Fig. 6 Physical map of Deli and La Mé oil palm populations with 5598 SNP markers. The colors indicate the density of markers according to the bottom scale (100 kbp/locus)
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 of 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 correlations 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 3600 bp or 0.20 cM being common in the two populations. This explains the results of Nyouma et al. (2020, 2022), 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, 2022) 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. A similar conclusion was reached by Technow et al. (2012) in maize, to explain the cases where this type of model outperformed the population-specific allele models. To further investigate this aspect, it would be interesting to study the correlation of marker effects obtained by GS models between Deli and La Mé populations, as done for maize in Technow et al. (2014). 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 Teh et al. (2016) and Kwong 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% from around 25 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
Fig. 11 Percentage of unique haplotypes shared 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 pink is the rolling average.

Fig. 12 Percentage of unique haplotypes shared 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 pink is the rolling average.
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 (Supplementary Figure 1) and genotyping errors compared to SNP arrays and SSR, while providing much higher marker density than SSRs.

**Genetic differentiation between Deli and La Mé**

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 showed a significant degree of differentiation among the two oil palm breeding populations. Reciprocal recurrent selection (RRS) certainly contributed to this differentiation, as suggested by results obtained in the heterotic groups used in maize RRS. Thus, the Iowa stiff stalk and corn borer synthetic populations diverged along the heterotic groups used in maize RRS. Reciprocal recurrent selection certainly contributed to this differentiation, as suggested by results obtained in the heterotic groups used in maize RRS. Thus, the Iowa stiff stalk and corn borer synthetic populations diverged along the heterotic groups used in maize RRS. Our results are also in agreement with the oil palm study 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 west of Benin were genetically more different from those of Deli. This 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). The variation found in the $F_{st}$ profile, which reached high values (>0.6) in some genomic signatures, suggests that $F_{st}$ is likely to be of interest for studying signatures of selection. This could help identify candidate genes, especially for traits with contrasting phenotypic values between breeding populations, such as bunch number and bunch weight between A and B groups. However, a higher SNP density seems necessary to obtain clearer profiles with more pronounced peaks (Porto-Neto et al. 2013) that could be linked to genes of interest based on the available information on oil palm annotation.

**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 $5 \pm 1.1$ (SD), i.e., similar to the result we obtained here. This indicates the robustness of the method against marker type and density.

The small $N_e$ values obtained here also explain the fact that GS can be implemented with small training populations and low marker density. Thus, in previous studies, GS models trained with data from only 108 Deli and 102 La Mé individuals were efficient enough to replace phenotypic selection before clonal trials (Nyouma et al. 2020) while GS accuracy plateaued with only 500 to 2000 SNPs, depending on the trait (Cros et al. 2017).

**Comparison of genetic and physical maps**

The construction of genetic linkage maps using SNP markers is 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; Herrero et al. 2020). The genetic linkage maps helped identify genomic regions having major genes and QTLs that control oil yield (Montoya et al. 2013; Jeennor and Volkaert 2014; Tisné et al. 2015), palm oil fatty acid composition (Singh et al. 2009; Montoya et al. 2013), vegetative growth (Ukoskit et al. 2014; Lee et al. 2015; Bai et al. 2018b; Teh et al. 2020), 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. In 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), although Billotte et al. (2010) used a factorial design. To our knowledge, only Cochard et al. (2015) 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 of our study is shorter than the map of Cochard et al. (2015), which reached 1935 cM and was obtained using a similar oil palm population. 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 SNP oil palm
A high-density genetic map was constructed from a complex mating genetic parameters affecting GS accuracy. A high-density linkage map (Ong et al. 2019, 2020), 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 (Fig. 7). For example, LG 1, 2, 5, and 16 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 (Eg5.1) 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 (Deli and La Mé). To further investigate this aspect, we compared the position of our SNPs on Eg5.1 with their position on EgPMv6, a new version of Eg5.1 improved through the use of a high-density linkage map (Ong et al. 2020), but that was made available after the beginning of the present study (Supplementary Figure 4). We found that, although some upturns existed (in particular for the smallest chromosomes), the positions on the two genomes are in general agreement. Consequently, there are still disagreements between the genetic positions obtained here and the physical positions, even considering the improved assembly. For example, LG 2 had 100% of its SNP located on the same chromosome according to Eg5.1 and EgPMv6 (Eg5.1_1 and GK000077.1, respectively), and almost identical SNP order between the two assemblies, while large upturns existed between the genetic and physical positions (Fig. 7). This aspect deserves further study, 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.

**Conclusion**

The present study focused on two key populations used for hybrid breeding in oil palm, Deli and La Mé, and estimated genetic 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 4252 SNPs from GBS and spanned 1778.52 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 (3600 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 \( F_\text{ST} \) 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 \( N_r \) values of the two populations were small (<5). Population-specific genetic maps and reference genomes would be of interest for future studies.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s13353-022-00708-w.

**Acknowledgements** 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.

**Author contribution** 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 the 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.

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

**Declarations**

**Conflict of interest** The authors declare no competing interests.
References

Babu BK, Mathur R (2016) Molecular breeding in oil palm (Elaeis guineensis): status and Future perspectives. Progress Hortic 48:123–131

Bai B, Wang L, Zhang YJ et al (2018a) Developing genome-wide SNPs and constructing an ultraligh-density linkage map in oil palm. Sci Rep 8:691. https://doi.org/10.1038/s41598-017-18613-2

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

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

Ballesta P, Maldonado C, Pérez-Rodriguez 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

Barcelos E, SDA R, Cunha RN et al (2015) Oil palm natural diversity and the potential for yield improvement. Front Plant Sci 6:190

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

Bejarano D, Martínez R, Manrique C et al (2018) Linkage disequilibrium levels and allele frequency distribution in Blanco Ore-jinegro 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

Bernardo RN (2010) Breeding for quantitative traits in plants, 2nd edn. Stemma Press, Woodbury, Minn

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

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

Caballero A (1994) Developments in the prediction of effective population size. Heredity 73:657–679

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

Chang L-Y, Toghiani S, Hay EH et al (2019) A weighted genetic relationship matrix based on Fixation Index (FST) prioritized SNPs for genomic selection. Genes 10. https://doi.org/10.3390/genes10110922

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.).

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

Cochard B, Carrasco-Lacombe C, Pominés 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

Corbin LJ, Liu A, Bishop S, Woolliams J (2012) Estimation of historical effective population size using linkage disequilibrium with marker data. J Anim Breed Genet 129:257–270

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

Corley RHV, Tinker PB (2016) The oil palm, 5th edn. Wiley-Blackwell, Chichester

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

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

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

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

Daetwyler HD, Pong-Wong R, Villaneuva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031. https://doi.org/10.1534 genetics.110.116855

Daval A, Pomiés 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

De Roos A, Hayes B, Goddard M (2009) Reliability of genomic predictions across multiple populations. Genetics 183:1545–1553
de Souza LM, dos Santos LHB, Rosa JRB et al (2018) Linkage disequilibrium and population structure in wild and cultivated populations of rubber tree (Hevea brasiliensis). Front Plant Sci 9. https://doi.org/10.3389/fpls.2018.00815

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

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Subsequent edition. Benjamin-Cummings Pub Co, Harlow

Ferreira A, da SMF, da CE SL, 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/S1415-47522006000100003

Flint-Garcia SA, Thornsberry JM, Buckler ES IV (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Gan ST, Wong WC, Wong CK et al (2018) High density SNP and DAIR-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

Gerke JP, Edwards JW, Guill KE et al (2015) The genomic impacts of drift and selection for hybrid performance in maize. Genetics 201:1201–1211

Glaubitz JC, Castseveas 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

Gondro C, Werf J van der, Hayes B (eds) (2013) Genome-wide association studies and genomic prediction. Humana Press

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: vol 3. Springer, Netherlands, Dordrecht, pp 651–682

Green P, Falls K, Crooks S (1990) Documentation for CRI-MAP, version 2.4

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
Hartley CWS (1988) The oil palm (Elaeis guineensis Jacq.). Longman Scientific & Technical; Wiley, Harlow, Essex
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
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
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
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
Ithnin M, Din AK (eds) (2020) The oil palm genome. Springer International Publishing
Ithnin M, Xu Y, Marjuni M et al (2017) Multiple locus genome-wide association studies for important economic traits of oil palm. Tree Genet Genomes 13:1–14
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
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
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
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
Labate JA, Lamkey KR, Lee M, Woodman WL (1999) Temporal changes in allele frequencies in two reciprocally selected maize populations. Theor Appl Genet 99:1166–1178
Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
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
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
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
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
Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57–63
Matias FL, Galli G, Granato ISC, Fritsche-Neto R (2017) Genomic prediction of autogamous and alloagamous plants by SNPs and haplotypes. Crop Sci 57:2951–2958. https://doi.org/10.2135/cropsj2017.01.0022
Meeuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
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
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
Niu S, Song Q, Koiwa H et al (2019) 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 19:1–12
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
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
Nyouma A, Bell JM, Jacob F et al (2022) Improving the accuracy of genomic predictions in an outcrossing species with hybrid cultivars between heterozygote parents: a case study of oil palm (Elaeis guineensis Jacq.). Mol Genet Genomics 297:523–533
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
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
Ouellette LA, Reid RW, Blanchard SG, Brouwer CR (2018) LinkageM-avView—rendering high-resolution linkage and QTL maps. Bioinformatics 34:306–307. https://doi.org/10.1093/bioinformatics/btx576
Paterson R, Sariah M, Lima N (2013) How will climate change affect oil palm fungal diseases? Crop Prot 46:113–120
Parker J, Mosnier A, Kraxner F et al (2016) What are the limits to oil palm expansion? Glob Environ Change 40:73–81. https://doi.org/10.1016/j.gloenvcha.2016.06.007
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.genet.2015.02.002
Porto-Neto LR, Lee SH, Lee HK, Gondor C (2013) Detection of signatures of selection using F ST. In: Genome-wide association studies and genomic prediction. Springer, pp 423–436
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
Rajesh MK, Gangurde SS, Pandey MK et al (2021) Insights on genetic diversity, population structure, and linkage disequilibrium in globally diverse coconut accessions using genotyping-by-sequencing. Omics J Integr Biol 25:796–809
Rastas P (2017) Lep-MAPS: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinforma Oxf Engl 33:3726–3732. https://doi.org/10.1093/bioinformatics/btx494
Semagn K, Bjaernstød A, Ndjiondjop MN (2006) Principles, requirements and prospects of genetic mapping in plants. Afr J Biotechnol 5. https://doi.org/10.4314/ajb.v5i25.56082
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
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

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/exv178

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

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

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

Soh AC, Mayes S, Roberts JA (eds) (2017) Oil palm breeding: genetics and genomics, 1 edition. CRC Press, Boca Raton

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

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/S1415-47522008005000005

Statista (2021) Vegetable oils: production worldwide by type, 2012/3-2020/21

Technow F, Riedelsheimer C, Schrag TA, Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125:1181–1194

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

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

Teh C-K, Ong A-L, Mayes S et al (2020) Major QTLs for trunk height and correlated agronomic traits provide insights into multiple trait integration in oil palm breeding. Genes 11:826

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

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

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

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

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

Ukoskit K, Chanroy 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

Wand M (1995) KernSmooth: functions for kernel smoothing supporting. Wand Jones R Package Version 2:23–20

Waples RS, Do C (2008) Idrw: a program for estimating effective population size from data on linkage disequilibrium. Mol Ecol Resour 8:753–756. https://doi.org/10.1111/j.1755-0998.2007.00613.x

Watson K, Mayes S, Price Z et al (2001) Qualitative 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

Weir BS (1979) Inferences about linkage disequilibrium. Biometrics 35:235–254. https://doi.org/10.2307/2529997

Weir BS, Goudet J (2017) A unified characterization of population structure and relatedness. Genetics 206:2085–2103

Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:621–631. https://doi.org/10.1534 genetics.112.146290

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Xu H, Guan Y (2014) Detecting local haplotype sharing and haplotype association. Genetics 197:823–838. https://doi.org/10.1534/ genetics.114.164814

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

Yamamoto E, Matsunaga H, Onogi A et al (2016) A simulation-based approach to position unmapped SNPs in crop species. BMC Genomics 17:971. https://doi.org/10.1186/s12866-016-1074-6

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.