

Integration of linkage and physical genome map of oil palm for the exploration of important traits within palm species

Ong Ai Ling

BSc, MSc (Bioinformatics)

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

September 2020

School of Biosciences

Abstract

Perennial crops, such as oil palm (*Elaeis guineensis* Jacq.) have a long breeding cycle (typically 20 years per cycle) and inherently heterogeneous populations. Oil palm is the most productive and traded vegetable oil globally compared to other oil crops such as soybean, rapeseed etc, but it has only undergone four generations of selection since breeding started 80 years ago. No doubt genome-enabled biology in oil palm breeding is crucial yet genetics and crop improvement have been actively pursued in research but with limited deployment in the industry. Further, the oil yield has stagnated for many years, hence the realistic approach of genome-based breeding is needed.

High resolution genetic map maximizes discoveries of recombination events and thus empowers comparative genomic and quantitative trait locus (QTL) studies. The present study aims to develop multiple ultra-dense linkage maps using high density marker panel and integrate these with physical genome map of oil palm. To improve the current genome assembly, more of the previously unplaced genomic scaffolds can now anchored to chromosomes using mapped markers from the genetic maps.

Linkage maps generated using three populations contained 11,421 transferable markers - consolidate each and the total maps span, on average, 1,355.64 cM with the lowest average marker intervals of 0.04 cM. These are by far the densest linkage maps reported in oil palm. The creation of a consensus linkage maps from combinations of different mapping populations provides essential information about the consistency of marker order and genetic distance among oil palm populations, while increasing map precision and resolution. The published oil palm physical genome size is 657.98 Mb while the assembled scaffolds add up to 1.53 Gb. The coverage of the assembled physical genome still leaves more than 60% to be fully covered.

The estimation of genome-wide recombination rates revealed varying degrees of genetic exchange across different populations and between chromosomes. The integration of the high-density linkage maps with the published oil palm genome could potentially improve the current genome for utilisation in oil palm improvement programmes. The outcome of this study has significantly improved the assembly of oil palm genome by anchoring more genome sequence scaffolds to chromosomes, resulting in 1.2 Gb equivalent to extension of 79% of the publicly available genome version.

The creation of the most complete genome map within palm species here, allows comprehensive comparative genomic studies. Applications of genome information in oil palm breeding through genome-wide association study (GWAS) and genomic selection (GS) and their combinations will pave the way for crop improvement not only in oil palm but also the exploration of economically important traits within palm species generally.

Acknowledgement

With deepest appreciation to both of my academic supervisors: Professor Festo Massawe and Associate Professor Sean Mayes from the university as well as to my external supervisor, Dr David Ross Appleton for their dedications and guidance along the journey of my study. Their full commitment and countless time spent to enhance my research directions and manuscript/thesis write up by providing constructive idea were invaluable too. Thanks to my comrades in work from Sime Darby Plantation R&D Centre and Associate Professor Dr Chew Fook Tim from National University of Singapore for the insights and knowledge sharing from their wide exposure in multidisciplinary scientific research areas. Not to forget my spouse and family members for their supports and accompaniments all the while, no matter good or bad times. Last but not least, thanks to my education sponsor from Sime Darby Plantation a.k.a Yayasan Sime Darby for providing full scholarship in my study.

Table of Contents

Abstract ..... **i**

Acknowledgement ..... **ii**

List of Abbreviations ..... **vi**

List of Figures ..... **vii**

List of Tables ..... **x**

Chapter 1: General Introduction ..... **- 1 -**

1.1 . The oil palm ..... - 1 -

1.2 . Genomics and Genetics in agriculture ..... - 2 -

1.3 . Definition of linkage map ..... - 5 -

1.4 . Definition of physical genome map ..... - 7 -

1.5 . Oil palm breeding and molecular technologies ..... - 8 -

1.6 . Problem statement ..... - 9 -

Aims and Objectives ..... **- 12 -**

Project Overview and Thesis Structure ..... **- 14 -**

Chapter 2: Establishment of an ultrahigh-density linkage map to improve the published oil palm genome (Objective one) ..... **- 16 -**

2.1 . Executive summary ..... - 16 -

2.2 . Background information ..... - 17 -

2.3 . General methodology ..... - 20 -

2.3.1 Oil palm mapping population and DNA preparation ..... - 20 -

2.3.2 Acquisition of genotyping data ..... - 20 -

2.3.3 Mining of genic SNP markers and annotations ..... - 20 -

2.3.4 Construction of genetic linkage map ..... - 21 -

2.3.5 Improvement of genome coverage ..... - 22 -

2.4 . Key findings, output, and general contribution to knowledge ..... - 22 -

2.5 . Conclusions and key recommendations for future work ..... - 33 -

2.6 . Published paper contributing to objective one: ..... - 36 -

Chapter 3: Comparison of linkage maps from multiple origins to enhance the integration of linkage and physical maps of oil palm (Objective 2) . - 37 -

3.1 . Executive summary ..... - 37 -

3.2 . Background information ..... - 38 -

3.3 . General Methodology ..... - 40 -

3.3.1 SNP-based marker linkage map construction from multiple origins of oil palm mapping population ..... - 40 -

3.3.2 Construction of linkage map and generation of a comparative map in oil palm . - 41 -

3.3.3 Anchoring of scaffolds to linkage groups ..... - 41 -

3.4 . Key findings, output, and general contribution to knowledge ..... - 41 -

3.5 . Conclusions and key recommendations for future work ..... - 57 -

3.6 . Published paper contributing to objective two: ..... - 59 -

Chapter 4: Application of genome-enabled biology in oil palm breeding (Objective three) ..... - 60 -

4.1. Executive summary ..... - 60 -

4.2. Background information ..... - 61 -

4.3. General methodology ..... - 63 -

4.3.1 Genome and gene comparisons ..... - 63 -

4.3.2 Reconstruction of date palm pseudo-chromosomes ..... - 63 -

4.4. Key findings, output, and general contribution to knowledge ..... - 64 -

4.5. Conclusions and key recommendations for future work ..... - 72 -

Chapter 5: General Conclusion ..... - 77 -

5.1. Exploring synergies and linkages between and across objectives ..... - 77 -

5.2. Conclusion ..... - 77 -

5.3. Recommendation for future work ..... - 78 -

References ..... - 80 -

Appendices ix

Appendix A: Publication 1 entitled ix

"Linkage-based genome assembly improvement of oil palm *(Elaeis guineensis)* ix

Appendix B: Publication 2 entitled ix

"An Improved Oil Palm Genome Assembly as a Valuable Resource for Crop Improvement and Comparative Genomics in the *Arecoideae* Subfamily".....ix

List of Abbreviations

NGS	Next Generation Sequencing
SSR	Simple Sequence Repeats
SNP	Single Nucleotide Polymorphism
GBS	Genotyping-by-sequencing
GWAS	Genome-wide association studies
QTL	Quantitative trait locus/ loci
MPOB	Malaysian Palm Oil Board
NCBI	National Centre for Biotechnology Information
LD	Linkage disequilibrium
GS	Genomic selection
BPRO	Breeding populations of restricted origin
LOD	Logarithm of the odds
cM	centimorgan

List of Figures

Figure 1. Genotyping selection guide depends on markers density and throughput (figure adapted and modified from (Lateef 2015)). ..... - 4 -

Figure 2. Project overview and study materials. ..... - 14 -

Figure 3. An ultrahigh-density Deli dura x AVROS pisifera linkage map with 27,890 SNP markers. The left scale represents the linkage length in unit centiMorgan (cM), whereas the number above each bar represents the linkage group number. ..... - 28 -

Figure 4. The common SNP markers between the genomic and the genic maps which were constructed using Lep-MAP 3 and JoinMap 5, respectively. High concordance of marker ordering was observed for lower density and ultra-dense genetic maps. Chromosomes 16, 14 and 8 are shown as examples. - 29 -

Figure 5: A) A high-density genic SNP-based linkage map of the Deli *dura* x AVROS *pisifera* family with 2,314 SNP markers, B) A pie chart depicting the distribution of the top-ranked pathway membership for the annotated genic SNP markers. ..... - 30 -

Figure 6: Scatter plots of integration between genetic linkage order and physical marker order throughout the oil palm genome. Blue dots represent the *dura* genetic map and red dots represent the *pisifera* genetic map. Grey dot represents the *tenera* combined map. (* Fragmented linkage groups with shorter length, including LG 2.2, LG 5.2 and LG 7.2 were assigned to Chromosome 2, Chromosome 5 and Chromosome 7, respectively.)..... - 31 -

Figure 7. Genetic versus physical locations of markers in parents and progeny. (Blue represent parent 1, red represent parent 2 and grey for progeny). Sex-averaged recombination rate (solid line) for chromosome 16 indicating regions of recombination hotspots and cold spots. ..... - 31 -

Figure 8. Dotted symbols represent the actual position of markers in chromosome assembly, while dotted line indicates the ideal orientation of chromosome scaffolds. ..... - 31 -

Figure 9. A) Comparison of number of scaffolds from previous physical map (R. Singh, Ong-Abdullah, et al. 2013) and the improved physical map developed in this study. B) Total length improvement (Mb) for each of the pseudo-chromosome after addition of new scaffolds to the published genome sequence in this study. ..... - 33 -

Figure 10. Comparative linkage maps for all mapping populations for chromosome 5. Abbreviated populations: *Mapping (Tenera-1), DelixNg (Dura-1), JLxJL (Dura-2), Key_DxD (Key DD), Key_DxP (Key DP)* ..... - 48 -

Figure 11. Synteny between the reference ultrahigh-density map (*Tenera-1*) and the other populations for chromosome 3. Each dot represents a marker ordered according to its position in the linkage map. ** genetic maps for *GMxDA(DxP)* and *GMxGM (DxD)* adapted from *Teh C.K et. al. (2020)*..... - 49 -

Figure 12. Refined assembly of oil palm genome (PMv6). (A) Chromosome length (Mb) comparison between P5-build, Singh et al., (2013), and PMv6 (this study). (B) Number of scaffolds from previous physical map) and the improved physical map developed in this study. (C) Build-up of PMv6 by scaffold anchoring using multiple linkage maps; left panel represents the central combined linkage map from three mapping populations, and the right panel shows the correlation plots of each populations' genetic versus physical order; presented in the order of Deli dura x Nigerian dura (*dura 1*—green), Johore Labis dura (*dura 2*—orange), and Deli *dura* x AVROS *pisifera* (*tenera*—blue) for Chromosome 2 and (D) Chromosome 7 as examples. (E) Genome-wide association studies (GWAS) Manhattan plot for a commercial oil palm population for one of the oil yield traits for the genome published by P5-build. (F) GWAS Manhattan plot for the same commercial oil palm population for one of oil yield trait after assembly improvement in this study, circle in red represent newly assigned and significant SNPs. (G) Statistics of genomic features improvement after placement of scaffolds into chromosomes for P5-build and this study. ..... - 54 -

Figure 13: Genome assembly improvement by scaffolds placement for PMv6 (blue) compared to P5-build (orange) for 16 chromosomes. ..... - 55 -

Figure 14. Integration of linkage with physical maps for oil palm, Chromosome 1 to 16 ..... - 57 -

Figure 15. Synteny analysis including three palm species: oil palm (eg), coconut (cn), and date palm (pd). Chromosome pairs with strong syntenic relations were linked by the same ribbon color. ..... - 65 -

Figure 16. Comparative genome comparisons of coconut against oil palm, both in chromosomal formats with annotated locations of the genes of interests. (cn- coconut; eg- oil palm)..... - 67 -

Figure 17: Date palm chromosome built from genome comparison against oil palm's chromosome 10. Integration of genetic markers from reported gender-linked linkage group (LG12) and LG11 of date

palm shown high correlation of genetic and physical ordering, suggesting this could be sex-determining chromosome of the monoecies oil palm. ..... - 68 -

Figure 18. High conservation of genes among palm species with minimal amino acid changes. (a) Alignments of *SHELL* gene encoding for the layer of shell surrounding the kernel for oil palm and the orthologues from coconut and date palm. (b) Phenogram tree for *LOG* gene orthologues, which play a role in sex determination for date palms; orthologues of the gene among palms and other monocots outgroup taxa were included. ..... - 70 -

Figure 19. Oil-palm-directed date palm genome assembly improvement by length (Mb) across 18 pseudomolecules. Anchoring of scaffolds was performed on the two published assemblies of date palm genome by Al-Dous et al. (2011) and later by Hazzouri et al. (2019). ..... - 71 -

Figure 20. Correlation plots for locations of DNA marker located within date palm chromosome (pd-15) for genome assembly improvement versions by Al-Dous,E.K et al. (2011) and Hazzouri, K. M. et al (2019) ..... - 72 -

Figure 21. Conceptual framework indicating key outcomes and highlights for each thematic area - 77 -

List of Tables

Table 1 Statistics of published oil palm assemblies to-date..... - 11 -

Table 2: Summary of the Deli *dura* x AVROS *pisifera* linkage map using genomic SNP markers, the selected genic SNP markers and estimated recombination rate..... - 26 -

Table 3. Summary of lower density Deli *dura* x AVROS *pisifera* genic SNP-based linkage map annotated from transcriptomic databases..... - 27 -

Table 4. Number of progenies and polymorphic markers for each of the population for linkage mapping. ..... - 40 -

Table 5: Genetic map for the Deli *dura* x Nigerian *dura* (*Dura-1*) based population with the estimation of chromosome and genome recombination rate ..... - 42 -

Table 6: Genetic map for the Johore Labis *dura* x Johore Labis *dura*/ *Key DD* (DxD) population with the estimation of chromosome and genome recombination rate..... - 43 -

Table 7: Genetic map for the Ulu Remis *dura* x AVROS *pisifera*/ *Key DP* (DxP) population with the estimation of chromosome and genome recombination rate..... - 44 -

Table 8: Genetic map for the Johore Labis *dura* x Johore Labis *dura*/ *Dura-2* (DxD) population with the estimation of chromosome and genome recombination rate*..... - 45 -

Table 9. Number of common markers found each linkage group (LG) in selected mapping populations ..... - 46 -

Table 10. Genome assembly improvement by coverage and quality using linkage map for Deli *dura* x AVROS *pisifera* (*Tenera-1*) as reference ..... - 49 -

Table 11. Summary for consensus map generated from multiple linkage maps and statistics for SNP markers and scaffolds used to build genome assembly for PMv6. ..... - 55 -

Table 12. Published genome assembly status of agronomically interesting palm species..... - 62 -

Chapter 1: General Introduction

1.1. The oil palm

Agriculture has been the most important economic sector in Malaysia and has contributed to economic growth in the past five decades. In an official report prepared by the National Statistical Department, agriculture still contributes to 7.3 percent - equivalent to RM99.5 billion - to the country Gross Domestic Product (GDP) for year 2019. There are three sub-sectors under the agriculture sector: crops, livestock, and fisheries. Oil palm is the major contributor, with 37.9 percent of the total GDP in agriculture sector (Department of Statistics Malaysia 2019).

Oil palm has been referred to as the "golden" crop due to its unbeatable yield performance among other oil crops (for instance, sunflower, rapeseed, or soybean). The average oil yield is about 4 tonne/hectare/year, 10 times higher than soy. The crop also has a major cost advantage over other vegetable oils as it uses a quarter less land than other sources for a given unit of oil. The cost per unit of palm oil production is followed by soy (20 percent higher).

Indeed, the crop is the primary agriculture crop for Malaysia and has been selected as the main value chain development of the country and due to diversification into multiple end products that drive economic transformation (Wold Bank Group 2019). There are four product categories derived from the crop: palm oil and kernel-based products, oleochemicals, non-food-based products such as cosmetics or household products and lastly palm biomass that potentially turns into energy sources. Hence, oil palm is the most sustainable oil crop to-date.

1.2. Genomics and Genetics in agriculture

The roles of the genome in agriculture has been kickstarted since the success of human genomics studies. The challenges that global agriculture face include climate change, rising food demands, increasing costs of fertilizer and maintenance of fields. These factors raise the concern of food supply and the public seek for efficient crops that could potentially increase yield in a short period of time, with climate and disease tolerance. Genomics offers unprecedented opportunities to increase crop yield, quality, and stability of production through advanced breeding strategies, enhancing the resilience of major crops to climate variability, increasing productivity, and promoting sustainability (Abberton et al. 2016).

Development of sequencing technologies has exponentially reduced the cost of sequencing, which increased the volume of sequencing data exponentially too ever since the 1,000 human genomes project in year 2008 to 100,000 genomes project launched in year 2012 (Turnbull et al. 2018). The 1,000 Plant Genomes Project (1KP) was an international research effort to establish the most detailed catalogue of genetic variation in plants after the successful implementation of the human genome project (Matasci et al. 2014; G. K. S. Wong et al. 2020). The genetic polymorphism in plant such as rice or maize populations becomes important when mutated regions of the genome are discovered to influence phenotypic changes, such as susceptibility to diseases, enhanced adaptability in certain environments and increased crop yields. To facilitate such discoveries, a reference genome sequence that links DNA (genetic) markers to validated gene models, transcripts, proteins, and other physical genomic features is required to better understand the mutations behind phenotypic changes. The next move from 1KP is towards 10,000 or even more species (Cheng et al. 2018).

Genetic markers can be described as differences in DNA sequence between individuals or species. Classification of markers is generally based on the types of allelic variation and sites of variation in DNA. To select an efficient marker system in plant breeding, several

criteria are considered (Bohar, Chitkineni, and Varshney 2020). For the case of wheat breeding, Gupta et al. (1999) suggested a set of criteria for an efficient marker system applicable to crop improvement: primarily that markers should be highly polymorphic, evenly distributed, preferably codominant, specific, cost-efficient and amenable to automation for high throughput genotyping purposes. There have been multiple molecular marker systems since 90's to-date. Restriction fragment length polymorphisms (RFLPs) random-amplified polymorphic DNAs (RAPDs), amplified fragment length polymorphisms (AFLPs), Diversity Arrays Technology (DArT) and simple sequence repeats (SSRs) have been widely utilized in plant breeding. Among these marker systems, SSR markers have been used extensively, primarily due to their high polymorphism rate and the ability to be assayed with high reproducibility on commonly available lab equipment.

Upon entering the era of Next Generation Sequencing (NGS) technologies, single nucleotide polymorphisms (SNPs) became widely used markers system due to the ease of discovery and abundance in genomes. SNP markers are common and widely used for various genomic applications also due to high-throughput genotyping systems that lowered the cost per data point (Rafalski 2002). Various technologies were developed for SNP based marker system: for small-scale and cost-effective high-throughput, TaqMan, KASP™ (Kompetitive Allele-Specific Polymerase chain reaction) and rhAmp genotyping are the options due to their assay design flexibility, assay design success rate, allele call rate and quality, ease of experimental run and cost per sample (Ayalew et al. 2019). For medium-scale genotyping panels, array-based platforms such as Infinium (Illumina, Inc.) or Axiom (Thermo Fisher) would be the option, as both of these platforms are comparable and have high accuracies and genotyping call-rates (Jiang et al. 2013). The only drawback is the inflexibility to add markers into these array-based genotyping platforms. SNP genotyping panels or SNP arrays have been developed and utilized in several crops including rice, maize, wheat, chickpea and pigeon pea

(Rasheed et al. 2017). Genotyping-by-sequencing (GBS) and its various forms have been optimized for simultaneous SNP discovery and genotyping. GBS is another option for high density genotyping arrays. There are approximately 13 different GBS techniques that have been used in crop plants , each of which has some distinctive features (Scheben, Batley, and Edwards 2017). This raises the question about what kind of platform is suitable for which genomic applications. A detailed comparison of different SNP genotyping technologies is elucidated in Figure 1.

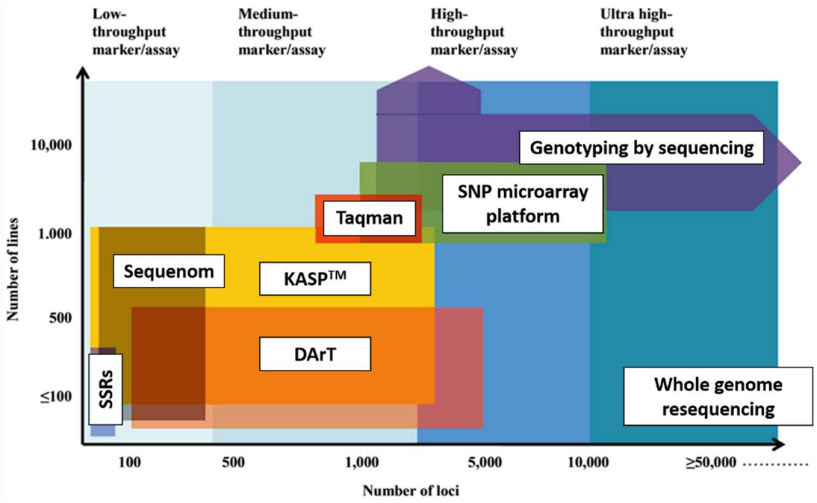


Figure 1. Genotyping selection guide depends on markers density and throughput (figure adapted and modified from (Lateef 2015)).

Genetic marker discovery and screening of large numbers of individuals is no longer a bottleneck due to new technologies. In rice breeding, multiple accessions of rice genome have been sequenced and having close to completion as reference genomes, resulting in ease of mapping of information for breeding traits of interest. The other concern is of narrow genetic background for crops such as rice due to multiple generations of selfing. Thus, genome

sequence information enables the development of informative genetic markers that help to identify genes for desirable traits of interests thus eliminate linkage drag by crossing of different rice varieties. There are more than 20,000 of rice accessions stored in seed banks (Prance et al. 1987). Cataloguing these accessions and comparing to wild varieties or elite germplasm, would allow introgression of alleles from accessions and wild varieties into cultivars for better yield or other performance traits. For example, traits for resistance to more than seven pathogens have been introgressed into elite rice germplasm from wild species, and some introgressed lines are in commercial cultivation (Brar and Khush 1997). The mapping of yield-associated quantitative trait locus (QTL) in interspecific populations have indicated that wild rice species harbour genes or alleles of interest for the improvement of rice cultivation in diverse environments (Cho et al. 2007; Moncada et al. 2001). The same approach is observed in soybean, given that modern lines have extremely narrow genetic backgrounds (Concibido et al. 2003; D. Li, Pfeiffer, and Cornelius 2008; Sebolt, Shoemaker, and Diers 2000; R. J. Singh and Hymowitz 1999). Some of this new variation was introduced through interspecific hybridization. Hence, a catalogue of wild genebank and cultivar genetic resources are essential in plant breeding (Zamir 2001).

1.3. Definition of linkage map

Linkage mapping is a method to determine the order of genetic markers along a chromosome via the estimation of recombination frequencies between genetic markers collected in a mapping (controlled cross) population. The initial step toward constructing linkage maps is to associate each marker with a linkage group which later to be assigned as chromosome. Estimation of genetic length is measured in centimorgans (cM) where 1 cM corresponds to recombination frequency of 1 % within a population with the assumption that homologous recombination events can occur at any point the DNA sequence (Painter 1998).

Chromosome assignments can be determined via reference to the physical map data, when available, and otherwise via reference to previously published linkage maps. Physical map distances can be used to interpolate and estimate linkage map distances, although there is considerable difference between linkage and physical distances, both within species and between species along chromosomes (Kong *et al* 2004). In humans, combined linkage-physical map analysis has been supported by both physical and recombination-based data. The estimation of map positions based on a centimorgan scale can be interpolated by numerous statistical methods (Matise *et al* 2007).

Development of linkage maps has been focused on discovery of QTL of specific traits of interest. The type of genetic markers used to develop linkage maps has switched from multiallelic markers to biallelic SNP markers due to the abundance of SNPs and ease of discovery from sequencing data. The construction of linkage maps using high resolution SNP markers improves the estimates of genetic distances, recombination rates and aids understanding the inheritance from maternal and paternal chromosomes (Kong *et al* 2002). Comparison of linkage maps between different populations can enhance the understanding of chromosome evolution between related species or even within populations of the same species. The creation of integrated/consensus linkage maps from combinations of different mapping populations can provide essential information about the consistency of marker order and genetic distance among oil palm populations, thus increasing map precision and resolution, as well as identifying potential regions of rearrangement which may have implication for breeding programmes (Staub *et al* 1996).

Linkage maps of large datasets would also be helpful in *de novo* genome assembly by facilitating comprehensive genome validation and refinement for better understanding of evolution, by analysis of genome synteny, association studies, quantitative trait locus mapping and so on (Rastas *et al* 2013). A high-resolution genetic map also could help to access the

integrity for genome of interest (Kong *et al* 2002). The establishment of high-density linkage map helps researchers to place, order, and orientate scaffolds along chromosomes and thereby assemble continuous chromosome sequence (Kawakami 2014).

1.4. Definition of physical genome map

The physical mapping or genome referred to genomic DNA, mapping and sequencing of an organism and heavily depends on the development of sequencing technology (Olson et al. 1989). It was also reckoned as a systematic entity with genomic DNA bound to many different proteins and organised into chromosomes. Genome was coded by only four nucleotides but they involved many biological functions and processes such as transcription, DNA replication and repair, recombination and chromosome segregation occurred along the way (Pollack and Iyer 2002). The most important parameter when describing a genome is its size measured in megabases (Mb) per haploid genome following the initial estimation standard applied in human genome. Chromosome arm lengths are critical parameter to be complimented with flow cytometry estimation. Genetic markers do play a role, being the choice of well-spaced loci for linkage tests between different gender or species to identify distance from centromere to telomere to nearest markers for genetic or linkage map development. DNA content of each chromosome is most accurately estimated as proportion of the haploid genome with their relative length obtained by autoradiography, image and flow cytometry (Morton 1991). The interpolation of physical and genetic length for estimation of a genome size has always been discussed, for instance in plants such as rice (Chen et al. 2002), barley (Mayer et al. 2012), maize (Wei et al. 2009) and many more to come that will better facilitate and comprehensive for all aspect of plant biological research.

1.5. Oil palm breeding and molecular technologies

Throughout the decades since the oil palm industry began, breeding of African oil palm in Malaysia has been based on phenotypic selection, using traits such as yield with recording over long selection cycles. Hence, the use of molecular markers has the potential to speed up the selection process and play a significant role when developing materials with specific desirable traits. The application of markers can also be useful to assess the genetic relationships between different populations (Hardon *et al* 1976, Corley and Tinker 2008). The nature of oil palm which is naturally cross-pollinating, with long generation intervals, means that genetic material is generated from crosses between parents of varying heterozygosity. The commercial planting material is known as *tenera* - a hybrid between thick-shelled *dura* palms and shell-less *pisifera* pollen palm. The male parents, *pisifera* are more heterozygous due to less inbreeding while the female parents, *dura* are normally descendants from more generations of selfing or sib-mating from limited gene pools (Seng *et al* 2011). The varying in heterozygosity is due to the main Deli *dura* population (seed parent) is derived from only four palms introduced into the Bogor Botanic Garden in Indonesia in 1848. Offspring of these were grown in the Economic Garden and were used by many plantations for selection, although selection criteria adopted were often different, oil palm breeders use "breeding populations of restricted origin" (BPRO) that can be traced back to distinct, often small groups of wild or unimproved ancestral palms (Rosenquist 1985). On the pollen side, the offspring of a palm called 'Djongo'—the 'best'—was imported from Congo and eventually came to represent the AVROS *pisifera* origin with 75% of the genetic material derived from a single palm. This is now the most important *pisifera* source (Corley and Tinker 2015; Ithnin and Kushairi 2020).

The history of linkage mapping in oil palm started 20 years ago (Mayes *et al* 1997) using RFLP markers. After that, more research on the development of controlled cross linkage maps and QTL mapping have been published in the last 10 years (Billotte *et al* 2005,

Pootakham 2015). It was in quite recent years that consensus maps for oil palm were studied, based on single or very few populations (Seng *et al* 2011, Lee *et al* 2015) with little incorporation of SNP markers (Ting *et al* 2014). Most published linkage maps in oil palm are largely based on microsatellites with the density of at most one thousand markers. While the highest density for a public domain SNP array reported so far is about 4,500 SNP in size. Besides this SNP panel, restriction-site associated DNA sequencing (RAD-seq) technology (Andrews et al. 2016) was used to generate high number of SNP markers, but the call-rates for each of the samples within the populations were rather inconsistent due to PCR artefact effects on allele frequency estimates (Puritz et al. 2014). Linkage maps developed from RAD-seq have the highest density to date with total the of 10,023 markers (Bai et al. 2018). The application of genetic markers in oil palm, which is also known as marker-assisted selection (MAS) is one of potential approaches to improve planting materials by shortening the selection cycles (Soh, Mayes, and Roberts 2017). Recently, efforts for QTL discovery have also been expanded to genome-wide association studies (GWAS) using a high density SNP marker panel (Teh et al. 2016). The findings were later tested in various genomic selection models for estimating the genetic potential of a set of markers for oil yield improvement (Kwong, Ong, et al. 2017; Kwong, Teh, et al. 2017)

1.6. Problem statement

In oil palm breeding, conventional breeding would need more than a decade to complete a breeding trial. Hence, a myriad of genetics applications has been actively pursued in research but with limited deployment in the industry. However, its yield potential was estimated to be achievable at 9 tonne/hectare/year in breeding trials (Rajanaidu et al. 2005). Traits of interest in oil palm are primarily focused on oil yield. Oil yield is derived from multiplication of total weight of fresh fruit bunch (FFB) with oil extraction rate (OER) from the mill. Secondary traits

focus on disease tolerance or climate resilience such as drought or harvestability traits such as long stalk palms or dwarf palms.

There is always an immense need for continued productivity improvement: Productivity in FFB yields and OER have been stagnant for decades. Malaysian government's 2020 target for FFB yield is 35 tons per hectare, while the realised yield has averaged 18.6 tons per hectare since the late 1970s. The oil yield of 3.6 tons per hectare has not changed since the 1960s. The potential yield is estimated at 8.8 tons per hectare (Hameed and Arshad 2007). The government is aware of the need to improve productivity, which has been limited by the narrow genetic base in Malaysia, among other factors, which also risks the palm exposure to pests and disease. To address this issue, the Malaysian Palm Oil Board (MPOB) undertook expeditions that started in the 1970s to collect germplasm from Africa and Latin America; hence they have the largest germplasm collection in the world (Zulkifli et al. 2017). Other contributing factors limiting productivity could be due to labour shortage or lack of quality labour, low replanting rates, limited agricultural land for expansion, monoculture risks causing loss of biodiversity and land fertility; the chain effects causing vulnerability to unforeseen diseases, pests, and climate change impacts; susceptibility to weeds, insects, and increased soil erosion (Sunderland 2011; Wold Bank Group 2019).

The genome of oil palm, *Elaies guineensis* was first published by MPOB (R. Singh, Ong-Abdullah, et al. 2013). Table 1 summarizes the size of the oil palm genome in the public database, National Center for Biotechnology Information (NCBI). The physical map is merely one third of the full expected genome size (657.98 Mb vs. 1.8 Gb) as predicted (Pellicer and Leitch 2020; Rival et al. 1997), while the total length of scaffolds length is only about 1.535 Gb. The other genome available in the database was published by the Seoul National University but mainly for the purpose of SSR marker development (Taeprayoon et al. 2016). There is still a need for oil palm genome improvement. This is complicated by the high percentage of N

nucleotide in the MPOB genome version, ranging for 23 – 33% for the published physical map and scaffolds. The estimation of recombination rates across oil palm has always been of interests as well as for crops that undergo an outbreeding/outcrossing breeding scheme. The challenge of estimating recombination rates is always due to the limited resolutions of genetic maps and the limited availability of genome sequences. An assessment of recombination rates with respect to physical distance will elucidate the relationship between the degree of linkage disequilibrium within different populations or origins in the crop of interest.

There are number of reasons why the improvement of the current oil palm genome is required. One is to improve the coverage of the current published oil palm genome for better understanding of the distribution of genes and enhance the application of genome-wide association mapping (Teh *et al* 2016). The completeness of the oil palm reference map will establish a high-density oil palm integrated linkage map for further fine mapping of quantitative trait-linked markers. To discover the co-linearity or co-segregation of genes and DNA polymorphisms within the oil palm population and to begin to understand the effects of selection pressure acting in different origins of the species is also important. Increasing marker densities and increased initiatives in physical mapping of an organism of interest have necessitated the need to generate a single consensus linkage map in this case for oil palm (Groenen 2000). This is because there will be many markers in common among different maps and this makes it feasible to consolidate scaffold positions on large and intermediate-sized chromosomes. With a good reference map, we can move on to locate the QTLs for phenotypes of interest on the genome through GWAS, which we believe, is one the most useful applications of a genome.

Table 1 Statistics of published oil palm assemblies to-date

Reference /Submitter	**Genome representation**	**Size (Mb)**	**Genome assembly accession**	**Total Length of scaffolds anchored into chromosomes (Mb)**
Dura/Seoul National University	Scaffolds	499	GCA_001672495.1	-
Pisifera/Orion Genomics	Chromosomes and unplaced scaffolds	1,535	GCA_000442705.1	658

Source : https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/2669/

This chapter covered up to date information from the literature. Most of the initial concepts and inspirations have been written in:

Oil palm breeding: Genetics and Genomics (Chapter 9: Molecular genetics and breeding) (Soh, Mayes, and Roberts 2017)

Bioinformatics: A Practical Handbook of Next Generation Sequencing and Its Applications (Chapter 10: Application of NGS data). (Low and Tammi 2017)

Aims and Objectives

Specific objectives of the project are structured around three broad areas (goals/aims):

Objective one:

Establishment of an ultrahigh-density linkage map to improve the published oil palm genome.

Specific objectives:

- to construct an ultrahigh-density reference linkage map for genome improvement using the largest commercial populations reported to date.
- to compare linkage maps of the commercial population to the physical genome.
- to map a genic SNP subset representing important biochemical pathways related to oil yield traits.

Objective two:

Comparison of linkage maps from multiple origins to enhance the integration of linkage and physical maps of oil palm.

Specific objectives:

- to refine three individual linkage maps and obtain a robust consensus genetic map from a comparative study of the individual linkage maps.
- to integrate the linkage and physical map to improve the genome assembly, resulted in a newer version of oil palm genome (PMv6).

Objective three:

Application of genome-enabled biology in oil palm breeding

Specific objectives:

- Application in breeding: GWAS and Genomic Selection (GS).
- To compare the improved oil palm genome with other palm species, based on conserved synteny across the palm family.

Project Overview and Thesis Structure

The thesis is structured around the three objectives with five main sections/chapters as summarised below:

1. Chapter 1 - General Introduction: This section provides a detailed description of oil palm as the crop of interest and summarises the current information on genome enabled biology and its importance in crop improvement. The overview of the project and study materials separated into three phases and the milestones are presented in Figure 2. Study materials includes genotyping data using high density marker panel from multiple oil palm BPROs and genomics data from three palm species (oil palm, coconut and date palm).

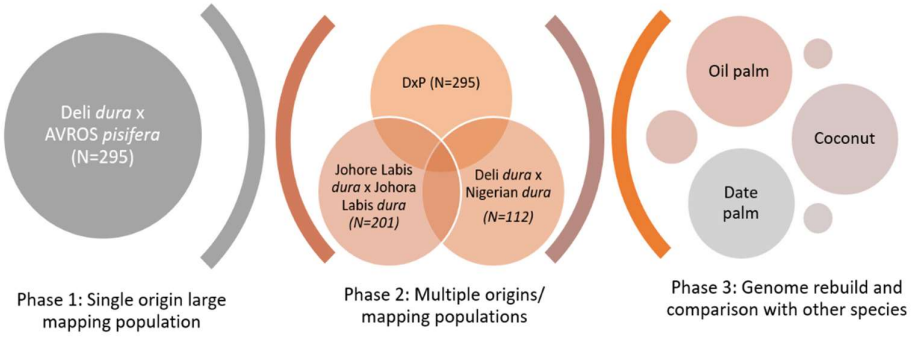


Figure 2. Project overview and study materials.

2. Chapter 2 - Objective one: Development of an ultrahigh-density linkage map and subsequent genic-based linkage map for validation and QTL discovery purposes.
3. Chapter 3 - Objective two: Expand on the work covered in Chapter 2 through inclusion of additional mapping populations to consolidate genetic markers across different oil palm origins and use multiple linkage maps to anchor genomic scaffolds into chromosomes.
4. Chapter 4 - Objective three: Explain the significance of the improved oil palm genome and describe comparative studies undertaken within different palm species.
5. Chapter 5 - General Conclusion: This section summarizes the outcomes of these studies, their implications and provides recommendations for future work.

Chapter 2: Establishment of an ultrahigh-density linkage map to improve the published oil palm genome (Objective one)

2.1. Executive summary

Meiotic crossovers in outbred species, such as oil palm (*Elaeis guineensis* Jacq., 2n = 32) contribute to allelic re-assortment in the genome. Such genetic variation is usually exploited in breeding to combine positive alleles for trait superiority. A good quality reference genome is essential for identifying the genetic factors underlying traits of interest through linkage or association studies. The AVROS *pisifera* genome is the only publicly available for oil palm with assembled chromosomes. Distribution and frequency of crossovers throughout chromosomes in different origins of oil palm are still unclear. Hence, an ultrahigh-density linkage map of a commercial Deli *dura* x AVROS *pisifera* family was constructed using the OP200K SNP array, to evaluate its alignment against the AVROS *pisifera* genome assembly. A total of 27,890 linked SNP markers generated a total map length of 1,151.7 cM and an average mapping interval of 0.04 cM. Nineteen linkage groups represented 16 pseudo-chromosomes of oil palm, with 61.7% SNPs mapped to the published genome. Meanwhile, the physical map was also successfully extended from 658 Mb to 969 Mb by assigning unplaced scaffolds to pseudo-chromosomes. A genic linkage map with a major representation of sugar and lipid biosynthesis pathways was subsequently built for future studies on oil related QTL. This study improves the current genome assembly of oil palm, and provides important insights into its recombination landscape, eventually unlocking the full potential for genome sequence-enabled biology of oil palm.

2.2. Background information

Oil palm *(Elaeis guineensis Jacq.)* is the only major crop member in the *Arecaceae* family producing edible oil (Hartley 1967). Being an efficient oil crop, the average oil yield of oil palm is four metric tons per hectare every year, which is approximately 10 times higher than soy (Glycine max.) (Corley and Tinker 2015; FAO 2017). Palm oil is currently the most traded vegetable oil, supplying one third of global vegetative oil (Byerlee, Falcon, and Naylor 2016; FAO 2017) on less than 6% of the world's agricultural land dedicated to oil crops. Despite yield superiority, most of the commercially grown oil palm materials have undergone only a limited generation of selection since 1920's due to its long selection cycles (12-19 years). In recent decades, the oil palm industry has put major effort into genomics and DNA marker research, primarily to hasten breeding progress. To achieve this, marker-assisted selection (MAS) has been introduced into breeding programmes to allow estimation of the breeding values in palms as early as the nursery stage.

Marker-assisted selection usually requires prior knowledge on the distribution of quantitative trait loci (QTL) for a targeted trait in the genome. Over 20 years, linkage maps from various backgrounds of oil palm have been constructed. The first restriction fragment length polymorphism (RFLP)-based linkage map for oil palm was reported in a self-pollinated palm from the Binga origin in Africa (Mayes et al. 1997). Similar efforts were then continued mainly using microsatellites for higher throughput and reproducibility, followed by QTL localization onto these maps (Billotte et al. 2005; Pootakham et al. 2015; Rance et al. 2001). The family-based mapping approach was found to be useful for detecting major QTLs, however QTLs were not consistent across populations (with a few exceptions). Consensus maps of one or a number of progenies, with incorporation of single nucleotide polymorphism (SNP) markers have been constructed to maximize mapping density.

Genetic maps reflect the actual inheritance of loci into their offspring based on the patterns of recombination during meiosis. Recombination is generally modelled to be a Poisson distribution, with individual chromosome expected to undergo at least one recombination event (as this is needed for correct disjunction), but generally no more than three. These are then converted to centiMorgans (cM) through the use of mapping functions, such as the Haldane or Kosambi (Gogoi, Ali, and Sharmah 2013; Haldane 1919; KOSAMBI 1943). This limitation in the extent of recombination in gametes also restricts the rate at which alleles close together on the same chromosome are separated in the offspring. Recombination is also not randomly distributed across the genome, larger genome species often having high levels of recombination at the ends of the chromosome, but little recombination in the heterochromatic and centromeric regions of the chromosome, despite genes still being present (e.g. in wheat; (Jordan et al. 2018; Sidhu and Gill 2005)). Recombination within an oil palm progeny limits the potential combination of alleles that can be achieved by breeding and, particularly, the number of offspring that need to be generated to be able to identify those with the desired combinations. Conventional oil palm breeding is often based on using test-crosses to evaluate the breeding potential of a parent. To assess parents and develop General Combining Ability (GCA; additive effects) and Specific Combining Ability (SCA; non-additive effects) for parental selection, the number of crosses assessed is more important than the absolute number of test-cross palms recorded (Corley and Tinker, 2015). In practice, 16 to 96 palms per cross in a trial are planted. This limited number of individuals from a single cross has focused attention on genome-wide association studies (GWAS). Teh and colleagues evaluated two multi-parent populations (2,045 tenera palms) for mesocarp oil yield (Teh et al. 2016) using a high-density OP200K SNP array (Kwong et al. 2016). The GWAS was based on the published physical genome of a *pisifera* palm of *E. guineensis* (R. Singh, Ong-Abdullah, et al. 2013). Only 60% (55,054) of the SNPs informative in the cross were mapped onto the reference genome. Furthermore, the

size of the pseudo-molecule physical map is merely one third (657.96 Mb) of the full expected 1.8-Gb genome size; even the total scaffold lengths only added up to 1.53 Gb, which is still incomplete. Hence, it is critical to further improve the reference genome to unlock the full potential of genome-enabled biology for oil palm. Linkage maps can contribute to the placement of scaffolds into the pseudo-chromosomes. The maps also give important information on the relative recombinational distance between loci from different pools of germplasm. In combination with a physical genome sequence, they can indicate the presence of inversions, translocations and duplications which may influence how effectively important gene combinations can be brought together through crossing and marker selection.

For the oil palm industry, the commercial planting materials are dominated by high-yielding *tenera* palms, which are derived from crosses between Deli dura and AVROS *pisifera*. The paternal *pisifera* is more heterozygous due to lower levels of inbreeding (although potentially from an extremely narrow genetic base) while the maternal dura is normally the descendent from more generations of self-pollination or sib-mating from a limited gene pool (Seng et al. 2011). Genetic differences, such as structure variations among the genomes of dura, *tenera* and *pisifera* can confound QTL results, predicted interaction and breeding objectives. Currently, the extent of such variation in oil palm is still unclear. In this study, we aimed 1) to construct an ultrahigh-density reference linkage map for genome improvement using the largest commercial *tenera* population reported to date; 2) to compare linkage maps of progeny to the physical genome; 3) to map a genic SNP subset representing important biochemical pathways to search for QTLs for oil yield traits.

2.3. General methodology

2.3.1 Oil palm mapping population and DNA preparation

The bi-parental population consists of a total of 295 F1 *tenera* palms derived from maternal Deli *dura* and paternal AVROS *pisifera* (278 x TT41/4) families and was selected as the mapping population. Fresh tissue was sampled from the third leaflet of each seedling. The population is currently at juvenile stage and maintained at Sime Darby Plantation, Malaysia. Total genomic DNA was isolated from 0.1g of the leaf tissue using the DNAeasy Plant Mini Kit (Qiagen, Germany). The DNA quality was then quantified on a 0.8% agarose gel using known standards.

2.3.2 Acquisition of genotyping data

The mapping population (genomic DNA= 25ng/µl) was genotyped using the 200K SNPs accommodated on the published OP200K Infinium array (Illumina, USA) (Kwong et al., 2016) on the Infinium iScan platform according to the manufacturer's recommendations. The raw intensity SNP data were analysed with GenomeStudio version 20011.1 with genotyping module version 1.8.4. Using a GenCall score cut-off of 0.15, auto clustering of the SNPs was done. The SNP clustering was confirmed manually by visual inspection. The SNP calls were exported into the PLINK program for minor allelic frequency (MAF) and call-rate filtering (Purcell et al. 2007). There are total of 170,860 SNP markers on OP200K, that are well spaced along the genome. Out of 170,860 SNP markers, only 41.4% of the markers were found in the published physical map, and the rest of the markers are mapped to unplaced scaffolds as reported by Kwong et al (2016).

2.3.3 Mining of genic SNP markers and annotations

All oil palm transcriptome sequences available in GenBank (Clark et al. 2016) were downloaded and combined with an in-house transcriptome database (Wong et al. 2017; Wong

et al. 2014). The transcriptome dataset was then aligned to the published oil palm genome (R. Singh, Ong-Abdullah, et al. 2013) using BLAT (Kent 2002) to identify gene structures. The SNP markers which were found in the published genome, residing within an interval between the first and the last exon of the gene structure were identified as genic. This step was done using an in-house Perl script. The SNP-tagged transcripts were annotated by referring to KEGG (Ogata et al. 1999) and the UniProt database (The UniProt Consortium 2015) via BLAST (BLAST 2013) based on homology search. These genic SNPs were eventually compiled and subjected to mapping analysis.

2.3.4 Construction of genetic linkage map

Linkage maps were developed using Lep-MAP3 (Rastas 2017; Rastas et al. 2013) for large SNP datasets and especially developed for outbred families like oil palm. The program consists of *Filtering* modules for markers' quality check. It is then followed by running modules *SeparateChromosomes* for binning of the markers into specific linkage groups with different logarithm of the odds (LOD) values ranging from 4 to 15 for better separation of the groups. *JoinSingles* assigns single markers to existing linkage groups to maximize the map-ability of the total input markers. Lastly, the module *OrderMarkers* orders the markers using the Kosambi mapping function to determine distances (Kosambi 1943). Kosambi function was used here to estimate the recombination frequencies thus calculating the distance between loci in linkage map and take into consideration of double crossovers event thus minimize erroneous in the estimation compared to Haldane's mapping function (Huehn 2011; Vinod 2011). The option "sexAveraged = 1" was included during execution of *OrderMarkers* when constructing the consensus map, this option will produce average recombination calculated from male and female parents from each of their mapping positions.

Smaller SNP subsets belonging to the polymorphic genic region were subsequently loaded into JoinMap 5 (Van Ooijen 2011) for linkage mapping to compare the performance of

different mapping tools. Mapping results generated from both mapping tools were compared and validated. All linkage maps were visualised using R script with minor modifications (Hussain et al. 2017) and Mapchart (Voorrips 2002).

2.3.5 Improvement of genome coverage

SNP markers from OP200K were mapped against the published oil palm assembled sequences, available at NCBI under BioProject accessions https://www.ncbi.nlm.nih.gov/bioproject/PRJNA434010 and assembly at https://www.ncbi.nlm.nih.gov/assembly/GCA_000442705, for the whole genome shotgun sequencing project with all scaffolds including the unplaced scaffolds in the physical map. The progeny's linkage map was applied to anchor the unplaced scaffolds to the physical map and the improvement was quantified based on the sum of total length of the newly anchored scaffolds. SNP markers that were present in the linkage map were mapped to the oil palm physical map. Genetic locations in the unit of centimorgan (cM) were compared to the actual physical position in base pair (bp) for each linkage groups or chromosomes. Markers that were not mapped to the physical map but found in unplaced scaffolds were later quantified based on the sum of the newly anchored scaffolds and their total length in unit of megabase (Mb) for each of the pseudo-chromosomes. This was to estimate the achieved in genome coverage improvement.

2.4. Key findings, output, and general contribution to knowledge

This report presents the first ultrahigh-density linkage map for a commercially important oil palm breeding population generated from 295 individual with 27,890 SNP markers, of which 2,314 are linked to genes in major oil biosynthesis pathways. For the genomic linkage map, each LG was traced to their published chromosomes (R. Singh, Ong-Abdullah, et al. 2013)

based on 61.7% of linked SNPs that were found in the reference genome (Table 2, Table 3, Figure 3 and Figure 5Figure 5). This will enable localization of specific critical oil biosynthesis pathway genes onto the genetic map. Development of an ultrahigh-density linkage map as a reference will allow robust linkage of complete genome and its improvement for use in full comparative and conserved synteny and genomics studies as well as QTL discovery.

So far, only Lep-MAP and MSTmap (Wu et al. 2008) programmes are able to map more than 10K SNP markers in under an hour. However, the MSTmap is limited to single families and phased data (including BC1, RIL, DH and Hap populations). The Lep-MAP does not have such limitations, so it was chosen to analyse the unphased cross-pollinating (CP) populations of oil palm. By using Lep-MAP3, total length of the genomic SNP-based linkage map was significantly reduced from 2,938.2cM obtained using Lep-MAP2 (Bai et al. 2018) to 1,151.7cM. With 16 chromosomes, a combined linkage distance in the order of 1,600cM would be expected. Lep-MAP3 is able to minimise inflated map length and ordering error when the number of markers is much higher than the mapping population size (Rastas 2017). This is a critical improvement because incorrect map length can greatly affect subsequent recombination estimates and scaffold anchoring.

The oil palm consists of a diploid genome with 16 homologous chromosome pairs (2n=32), with three LGs fragmented from LG 2, LG 5 and LG 7, respectively (Figure 3). A similar LG assignment was also observed in the genic linkage map (Figure 5A). In this case, LG 2.2 and LG 7.2 were absent because none of their SNP was genic, whereas removing non-genic SNPs on LG 11 caused fragmentation to two groups. More markers normally confer a longer map, reflecting better mapping coverage, but also potentially reflecting a reduction in the data quality as more comprehensive use is made of the dataset. The phenomena were not observed in this study. The genic linkage map with only 2,314 SNPs (1,733 cM) was longer than that of the genomic linkage map with 27,890 SNPs (1,151.7 cM). Although linkage

analysis is often problematic when the number of markers is high (approaching 10,000 markers). However, we found that both JoinMap5 and Lep-MAP3 performed equally as reported (Rastas 2017; Rastas et al. 2013). The common SNP markers between the genomic and the genic linkage maps were in synteny (Figure 4), thus Lep-MAP3 is recommended.

The oil palm is the world's most effective oil crop (Sime Darby Plantation 2014). Understanding the candidate genes underlying QTLs for important traits such as oil yield and quality in advanced breeding populations and commercial material at the genome level will assist marker discovery and their deployment in breeding programmes. It will also facilitate functional gene discovery resulting from better coverage of the genome and potentially locate some of the unassigned genetic traits detected so far. In the future, this reference map will assist with the dissection of the QTL differences between populations, with differences in gene order between origins and in the identification of causal genes. The initial genome wide as well as chromosomal wide recombination patterns for oil palm are also reported here for the first time, revealing the missing links between the physical and linkage maps of this important species.

Sugar and lipid-related biosynthesis pathways on the genic linkage map (Figure 5) revealed genetic polymorphisms still presence within the commercial *tenera* family which could be exploited in breeding program. Notably that there is still number of unannotated genes showing polymorphism within this population indicating novel genes might have important functions and to be discovered by further QTL mapping. The published QTLs for mesocarp oil content were located on Chromosome 5 and Chromosome 11 in Deli *dura* x AVROS *pisifera* population (Teh et al. 2016). In the genic linkage map, the QTL on Chromosome 5 was mainly populated by plant hormone signal transduction, glycerophospholipid and glycerolipid metabolism pathways. The plant hormone signal transduction pathways tagged by the most SNPs was closely related to plant floral development, growth and response to the environment (Campos-Rivero et al. 2017). As for the QTL on Chromosome 11, it included glycolysis and

starch-and-sucrose metabolism pathways. Marker intervals bearing double crossovers are not distinguishable from those without crossover, causing underestimation of linkage between marker and trait during QTL mapping. However, double recombination should rarely occur in the genic linkage map with an average mapping interval of 0.75cM, which is much lower than 10cM threshold (Xu, Zou, and Vision 2005) and the mapping function partly compensates for this. Hence, this suggests that the future QTL mapping analyses for oil yield traits in oil palm can be guided using the genic linkage map with 10X reduction of SNP markers, while also assessing polymorphism associated with genes known to be involved in important pathways related to oil production.

The distribution of markers according to linkage and physical locations revealed clear differences in parental recombination patterns. Based on the comparison of linkage versus physical location for each of the markers, most of the telomeric regions for each chromosome were well covered with markers except for chromosome 5, 11 and 15. On top of that, linear mapping relationships were also observed in the telomeric regions for most of the chromosomes validating ordering of the markers for both genetic and the physical maps (Figure 7).

For chromosome 16, Parent 1 representing the *dura* palm had no distinguishable recombination, perhaps suggesting identity-by-descent of this chromosome in the maternal line, whereas parent 2 which is the *pisifera* had undergone active cross over events with steep slopes at both ends of the linkage group/chromosomal regions (Figure 7). Ideally, sigmoidal curve such as dotted line plotted for the case of chromosome 16 (Figure 8) is expected indicating active recombination patterns towards the telomeric (end) regions while almost no recombination events occurring in centromere. However, due to lack of mapping resolution for this region within the genome (possibly caused by incomplete scaffold placement or genome

assembly or recombination events that link the telomere of the chromosome with centromere.

The pattern of cM:Mb is also a good indicator of the location of the centromere.

Table 2: Summary of the Deli *dura* x AVROS *pisifera* linkage map using genomic SNP markers, the selected genic SNP markers and estimated recombination rate

Linkage group	Linkage Length (cM)	No. of linked SNPs	No. of SNPs mapping in published genome	Chromosome size (Mb)	No. of selected genic SNPs	Average Recombination rate (cM/Mb)
1	122.0	2422	1803 (74.4%)	68.44	215	1.8
2	120.0	3533	1535 (43.4%)	65.56	194	2.0
2.2	9.7	248	199 (80.2%)			
3	103.2	2582	1959 (75.9%)	60.06	260	1.7
4	121.1	2235	1203 (53.8%)	57.25	134	2.1
5	33.0	1281	659 (51.4%)	51.96	84	0.8
5.2	7.8	124	75 (60.5%)		18	
6	81.2	1752	1079 (61.6%)	44.36	117	1.8
7	25.8	889	758 (85.3%)	43.46	120	0.6
7.2	12.6	217	198 (91.2%)			
8	89.4	2512	1312 (52.2%)	40.20	178	2.2
9	61.4	1723	889 (51.6%)	38.06	120	1.6
10	74.4	2912	1675 (57.5%)	31.90	310	2.3
11	50.8	1012	783 (77.4%)	30.07	98	1.7

12	84.9	1312	754 (57.5%)	28.80	123	2.9
13	27.4	620	423 (68.2%)	27.81	41	1.0
14	56.8	1553	1037 (66.8%)	24.38	165	2.3
15	42.5	586	541 (92.3%)	24.31	88	1.7
16	27.7	377	328 (87.0%)	21.37	49	1.3
Total	**1151.7**	**27890**	**17210 (61.7%)**	**657.99**	**2314**	
Mean	**60.6**					**1.8**

cM – centiMorgan, the percentage of 200K SNP markers mapped in the reference genome in parentheses.

* Linkage group number follows chromosome number of the published genome (R. Singh, Ong-Abdullah, et al. 2013).

Table 3. Summary of lower density Deli *dura* x AVROS *pisifera* genic SNP-based linkage map annotated from transcriptomic databases.

Linkage group	Linkage Length (cM)	No. of selected genic SNPs
1	162.3	215
2	235.1	194
3	156.6	260
4	146.0	134
5	68.0	84
5.2	13.6	18
6	105.8	117
7	103.9	120
8	119.3	178
9	94.5	120
10	140.9	310
11	47.1	63
11.2	29.9	35
12	69.6	123
13	49.8	41
14	72.6	165
15	72.4	88
16	46.4	49
Total	**1733.0**	**2314**
Mean	**96.3**	

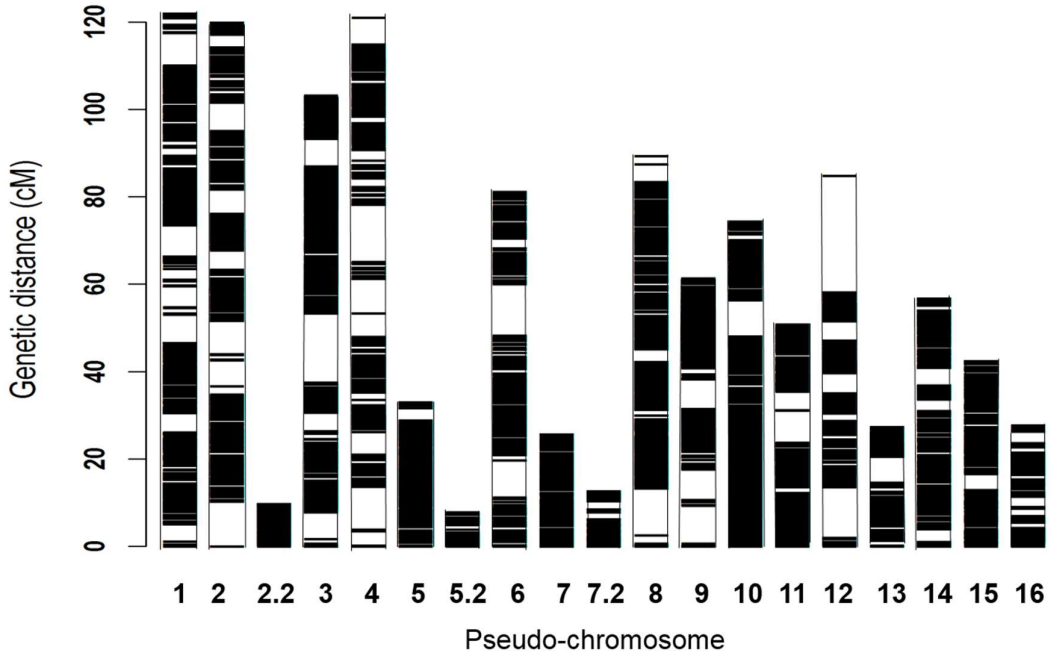


Figure 3. An ultrahigh-density Deli dura x AVROS pisifera linkage map with 27,890 SNP markers. The left scale represents the linkage length in unit centiMorgan (cM), whereas the number above each bar represents the linkage group number.

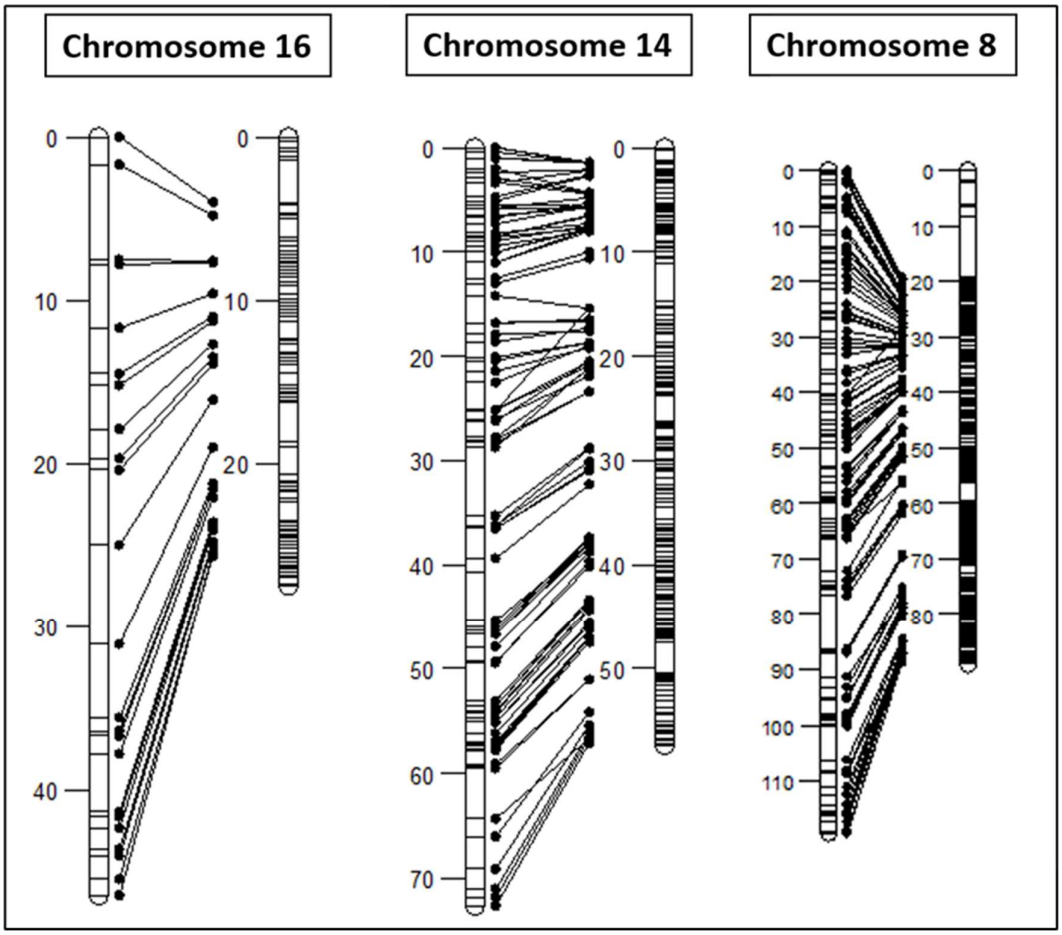


Figure 4. The common SNP markers between the genomic and the genic maps which were constructed using Lep-MAP 3 and JoinMap 5, respectively. High concordance of marker ordering was observed for lower density and ultra-dense genetic maps. Chromosomes 16, 14 and 8 are shown as examples.

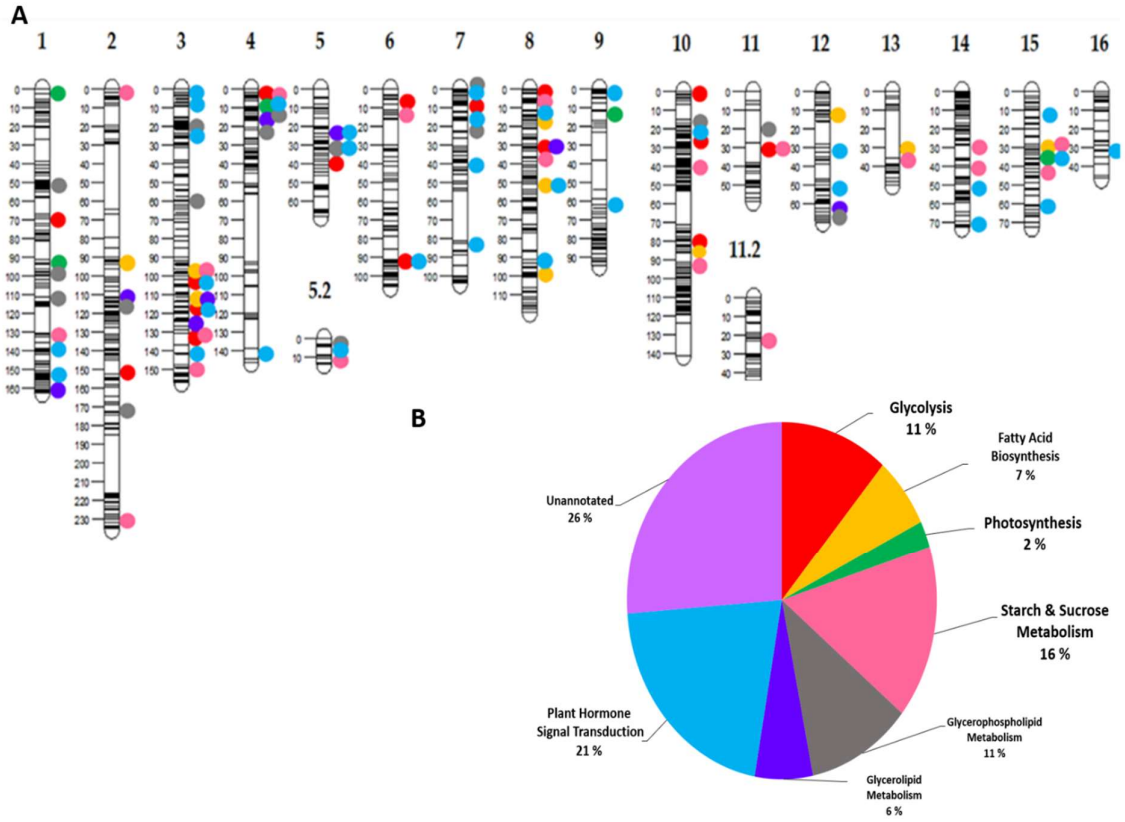


Figure 5: A) A high-density genic SNP-based linkage map of the Deli *dura* x AVROS *pisifera* family with 2,314 SNP markers, B) A pie chart depicting the distribution of the top-ranked pathway membership for the annotated genic SNP markers.

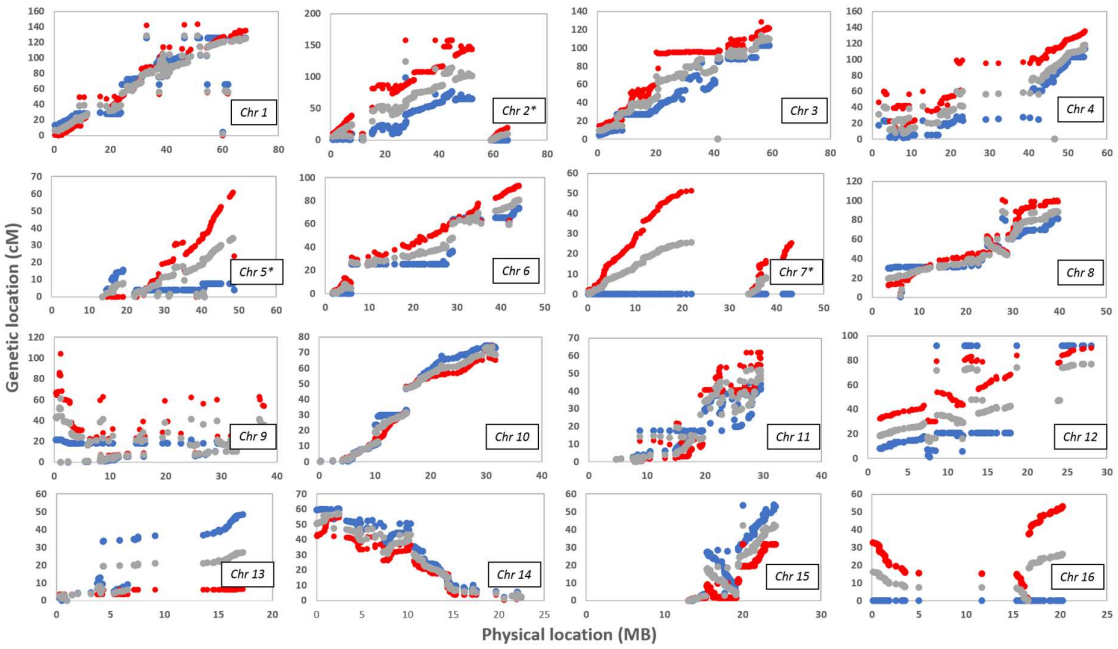


Figure 6: Scatter plots of integration between genetic linkage order and physical marker order throughout the oil palm genome. Blue dots represent the *dura* genetic map and red dots represent the *pisifera* genetic map. Grey dot represents the *tenera* combined map. (* Fragmented linkage groups with shorter length, including LG 2.2, LG 5.2 and LG 7.2 were assigned to Chromosome 2, Chromosome 5 and Chromosome 7, respectively.)

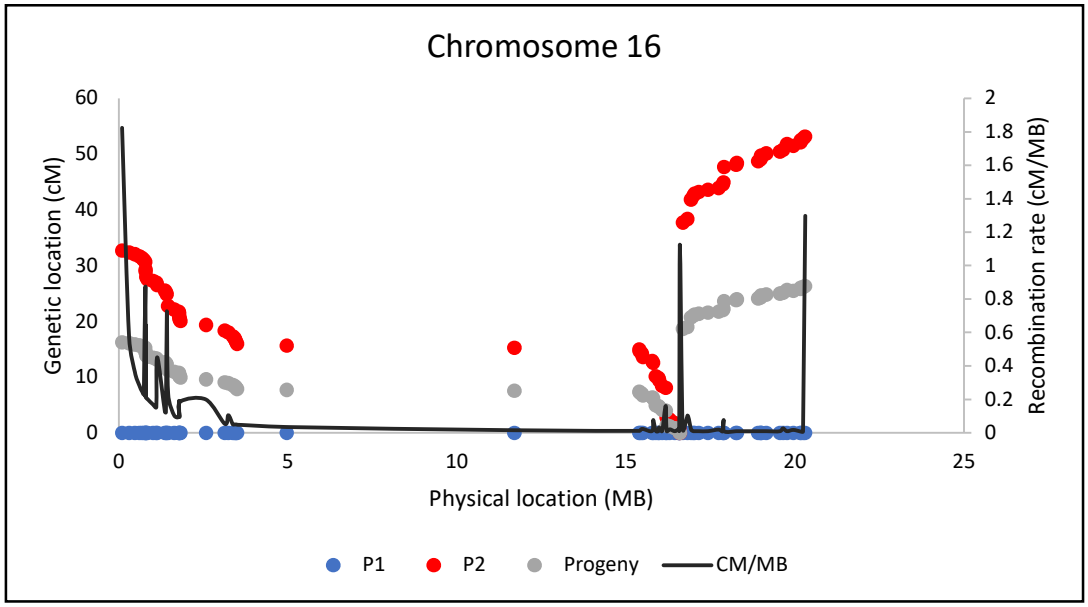


Figure 7. Genetic versus physical locations of markers in parents and progeny. (Blue represent parent 1, red represent parent 2 and grey for progeny). Sex-averaged recombination rate (solid line) for chromosome 16 indicating regions of recombination hotspots and cold spots.

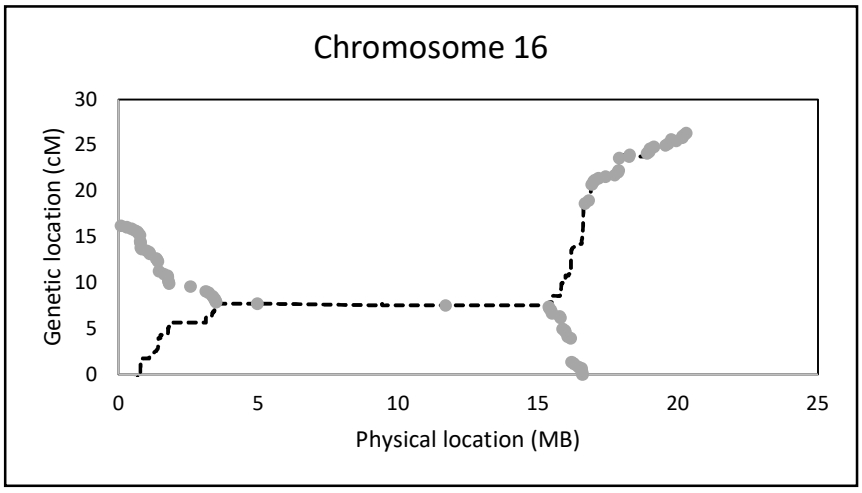


Figure 8. Dotted symbols represent the actual position of markers in chromosome assembly, while dotted line indicates the ideal orientation of chromosome scaffolds.

The average number of scaffolds contributing to genome sequences by Singh et al (2013) are 19 scaffolds in each pseudo-chromosome. The ultrahigh-density linkage map identified a total of 311 Mb in scaffolds in addition to the existing genome assembly that could be added to the linkage groups. In total, 1,323 additional scaffolds were placed into pseudo-chromosome locations (Figure 9A). Figure 9B depicted the total length for each pseudo-chromosome from published genome sequences. Each of the pseudo-chromosome was extended from this study with the most significant extension in chromosome 2 at 56 Mb, while chromosome 15 and 16 also have minimal extensions of less than 1 Mb, hence the improvements were not distinguishable in the plot (Figure 9B). After scaffolds addition, the physical map was successfully extended from 658 Mb to 969 Mb, which is equivalent to a 47% improvement. Still, the improved physical map only covers 54% of the total genome based on the estimated genome size of oil palm of 1.8 Gb.

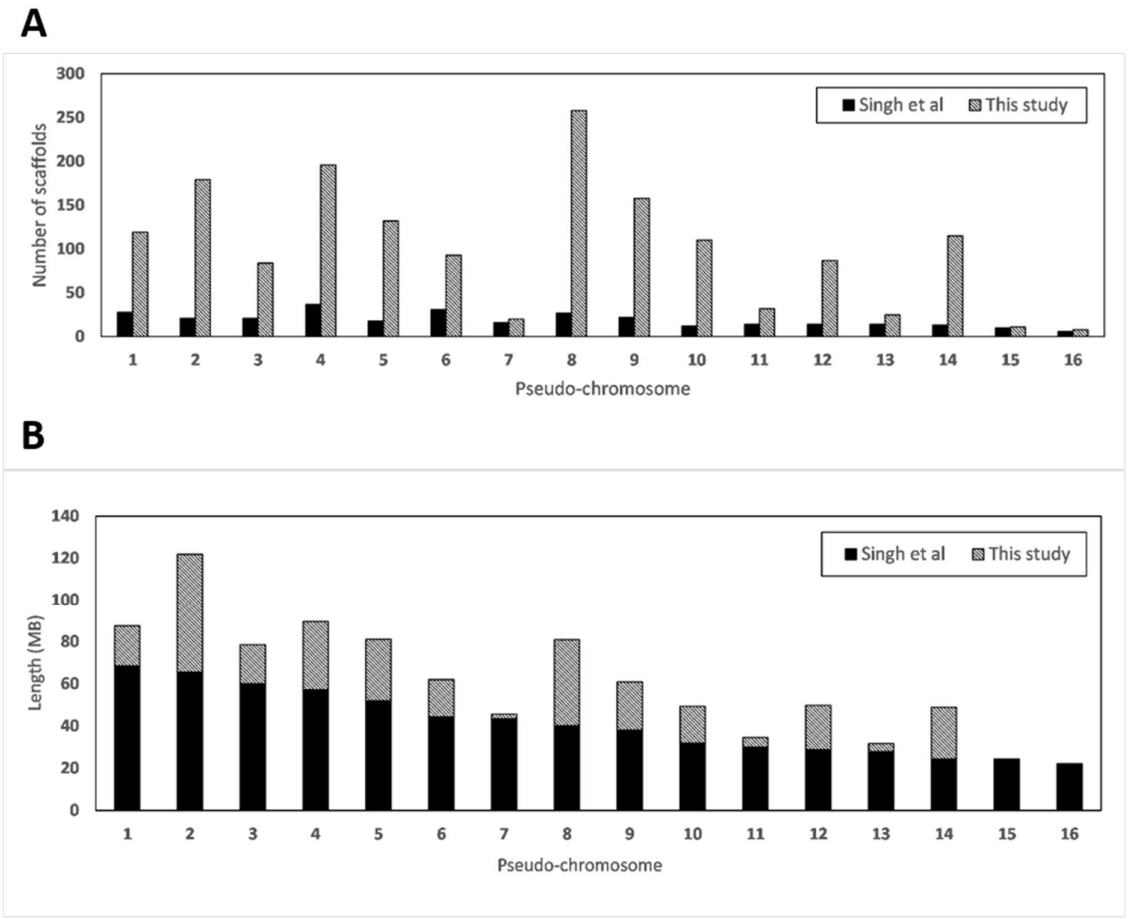


Figure 9. A) Comparison of number of scaffolds from previous physical map (R. Singh, Ong-Abdullah, et al. 2013) and the improved physical map developed in this study. B) Total length improvement (Mb) for each of the pseudo-chromosome after addition of new scaffolds to the published genome sequence in this study.

2.5. Conclusions and key recommendations for future work

An ultrahigh-density linkage map containing 27,890 markers in an oil palm commercial population to improve the current physical genome and to compare the maternal versus paternal patterns of recombination genome-wide in commercial oil palm (Deli *dura* x AVROS *pisifera*). The genetic analysis allowed the physical map to be extended to cover 969 Mb through addition of previously unplaced scaffolds. The goal of constructing a high-density linkage map has been achieved for this phase of the study with an average mapping interval of 0.04 cM. This is largest reported average mapping interval, compared to the previous studies including 1.26cM (Ting

et al. 2014), 1.40cM (Pootakham et al. 2015) and 0.29cM (Bai et al. 2018). In fact, the improvement of the mapping interval was mainly due to more recombination events detected in this larger mapping population, which was doubled compared to previous reported studies. The ratio between linkage (cM) to physical distance (Mb) was estimated based on genome-wide recombination rates. The completion of an ultrahigh-density linkage map will serve as reference map for future breeding programmes as well for QTL fine mapping purposes. The genic SNP-based genetic sub-map that contains critical oil biosynthetic pathway and genes, generates a quality reference map for oil palm *tenera* populations and enables pathway-specific location of gene orthologues which can be compared to yield related QTLs, providing candidate genes for trait components.

The completeness of the oil palm reference genetic map will establish a high-density oil palm integrated linkage map for further fine mapping of quantitative traits to linked markers. To discover the co-linearity or co-segregation of genes and DNA polymorphisms within the oil palm population and to begin to understand the effects of selection pressure acting in different origins of the species. With only ~62% of the total number of linked markers in the consensus map being found in the assembled physical genome, there is room for improvement of the chromosome coverage of the current oil palm genome. New markers identified from the current linkage map could assign and orientate 'orphan' scaffolds when these markers are located to the physical map. Also observed was the breakage of 3 chromosomes/linkage groups into two parts indicatives of regions with insufficient polymorphic markers to link the fragments together. The averaged recombination rate for oil palm was estimated at 1.8 cM/Mb, which means ~2% of recombination occurred within each 1 Mb of the genome. It has been reported that high recombination rates are highly correlated with gene density or distance from the centromere (Mézard 2006). Though not many plant species have been studied at a finer scale, especially in highly heterogenous and heterozygous tree species such as oil palm, we

have provisionally identified several regions in the genome with recombination rates higher than the genome average value, therefore defined as potential recombination hotspots. These hotspots often lay towards the telomeric ends of the chromosomes, which can also be observed from the genic-based markers linkage map (Figure 7 and Figure 8), indicating high recombination rates correlated with gene rich regions.

This is the first report elucidating the relationship between genetic and physical locations of SNP markers at high-density in an oil palm population. Here, we also reported the first ever provisional high-resolution recombination map of oil palm genome. Recombination hotspots and 'cold-spots' can be observed across chromosomal regions. The comprehensive determination of the recombinational rates is somewhat hampered by unplaced scaffolds. Furthermore, the assembled genome sequence of oil palm in chromosomal format was only one third of the expected 1.8 Gb of the size of the genome, indicating missing links between the genetic and physical locations of the SNP markers (Pellicer and Leitch 2020; Rival et al. 1997). Ideally, a sigmoid curve is expected between linkage and physical position for chromosomes with centrally placed centromeres (Nachman 2002). However in oil palm, substantial inversions were observed in certain chromosomes (Figure 7) It could be also due to high percentage of N nucleotide regions that cause disorientation of the scaffold placement onto the physical location, errors in map construction or the existence of inversions and translocations between the genotype used for genome sequencing and the current parental material used to develop the cross and genetic map. The estimation of genome-wide recombination rates revealed differences across parental lines, progenies and between chromosomes. More importantly, it has revealed differences between recombination rates in the maternal and paternal lines. The integration of the ultrahigh-density linkage maps into the published oil palm genome, could potentially improve the current genome sequence for utilisation in oil palm improvement programme.

The work should be extended to more mapping populations to ensure the consistency of markers ordering between different origins, thus increase the confidence of anchoring more genome scaffolds to the physical map of oil palm. Detailed recombination rates can be determined to improve the estimates of genetic distances and to identify mutation hotspots within the genome that could be potentially associated to phenotype variation. Assignment of further scaffolds to the pseudo-chromosomes may allow more extensive genome coverage and could help to assign more genetic effects to specific chromosomes, rather than unlinked assignment, but lack of polymorphism may still exist due to identity-by-descent issues. The detected outliers and inversion of chromosome segments with respect to the reference genome also might reflect mutations, mainly inversion and translocation or even duplications. Long-read sequencing approaches, such as and Pacific Biosciences (PacBio) (Rhoads and Au 2015), Oxford Nanopore technologies (Deamer, Akeson, and Branton 2016) and also optical mapping can be coupled with the linkage map to differentiate the genuine mutations from the errors (Pollard et al. 2018).

2.6. Published paper contributing to objective one:

Linkage-based genome assembly improvement of oil palm (Elaeis guineensis) (Ong et al. 2019) *(Appendix A)*

Chapter 3: Comparison of linkage maps from multiple origins to enhance the integration of linkage and physical maps of oil palm (Objective 2)

3.1. Executive summary

Oil palm is the most efficient and sustainable of oil crops, accounting for 39% of world production of vegetable oil on 6% of the land dedicated to oil crops. As a perennial tree crop, oil palm would greatly benefit from genome-enabled breeding as an essential approach for the continued development and improvement of the crop. High linkage maps allow anchoring of more markers from previously unplaced scaffolds into physical genome. Linkage-based genome assembly improvement discussed in Chapter 2 has been discussed on single ultrahigh-density linkage map. Incorporation of linkage maps from multiple origins increase the confidence and maximise of ordering, and orientations of markers and scaffolds thus enhance completeness of genome assembly. The newly assembled genome of oil palm – reported here as PMv6 - has 79% greater physical length in the 16 pseudo-chromosomes compared to the previously published version. The improvement was made by assigning an additional 1968 originally unplaced scaffolds that were available publicly into the physical genome. By integrating three high-density linkage maps and using them to place genomic scaffolds, 16 mega-scaffolds were ordered and oriented. The improved genome has enhanced the mapping resolution for genome-wide association studies (GWAS) and permitted further identification of candidate genes/protein-coding regions (CDSs) and any non-coding RNA that may be associated with them for further studies

3.2. Background information

Comparative linkage mapping has aided the discovery and understanding of genome evolution, identifying colinear markers or even genome rearrangements with genomic fragments translocated or inverted (Doganlar et al. 2002). Comparing more linkage maps generated from within species, but from different origins will enhance the accuracy of markers ordering within genomic regions and increase the confidence of the genome quality. This homology among linkage groups will result in a consensus map that consolidates several maps into single map (Brondani et al. 2006).

Development of high-density linkage maps not only permits QTL discovery but also enables fine mapping of QTL regions for identification of causal genes of genomic features for traits of interest. Incrementally increasing mapping resolution allows known genomic coordinates to anchor markers for genes or set of genes in orders that might be the cause phenotypic changes (Markelz et al. 2017). A better resolution map also enables genome differentiation and identification of domestication genes for traits of agronomic importance such as disease resistance, yield or tolerance to pest (Y. Li et al. 2020). In fact, every species will need a good genome assembly for understanding the evolutionary conserved sequence motifs, or gene regulators to provide a framework for reconstruction of genome organization, content, and dynamics, to empower new models of disease and heritable phenotypes; and to provide a starting point for the assessment of expansion, contraction, and adaptation of gene families in different evolutionary lineages (Lewin et al. 2009). Anchoring of scaffolds from genetic maps into chromosome-based physical maps empower precise comparative genome organization studies by facilitating the correct ordering of genes and non-genic DNA segments, making whole genome alignment more accurate (Larkin et al. 2009).

A better chromosomal scale physical map also enabled scanning of linkage disequilibrium (LD) that plays an important role in associative mapping for QTL discovery.

Plant breeding mating systems of the species (self-pollination versus outcrossing), and phenomena such as population structure and recombination hot spots, can strongly influence patterns of LD (Flint-Garcia, Thornsberry, and Edward IV 2003). Recombination frequency between marker and causal gene can weaken predictability of a trait. One way to quantify this effect is through genome-wide LD decay estimation, but it is influenced by the pedigree of the material. This parameter differs across origins, or even populations (Bosch et al. 2009; Stich et al. 2013). In oil palm, longer LDs in the commercial populations have been observed, compared to wide germplasm materials (Jin et al. 2016; Kwong et al. 2016; Teh et al. 2016) and the evidence presented in previous study suggests that maternal and paternal recombination patterns also vary, which potentially influences breeding strategies (Ong et al. 2019).

A good genome assembly does not only rely on assembly of contigs into scaffolds but also the placement and contiguity of scaffolds on chromosomes to form the physical map. To summarize the latest status of assembled genomes of oil palm, coconut, and date palm (as available in National Centre for Biotechnology Information; NCBI): 1535.18 Mb of P5-build oil palm genome first achieved a chromosome coverage level with 43% of scaffolds anchored (R. Singh, Ong-Abdullah, et al. 2013), although no improvement has been published since 2013. On the other hand, the 2202.46 Mb of assembled coconut genome and the 772.32 Mb of assembled date palm genome have recently surpassed the oil palm genome, with 46% (Xiao et al. 2017) and 50% (Hazzouri et al. 2019) of scaffolds anchored to the chromosomes. The genome size of oil palm is estimated at about 1.8 Gb (Pellicer and Leitch 2020; Rival et al. 1997). We previously reported on the possibility of further extending the physical map of the current P5-build reference genome using one ultrahigh-density linkage map of a commercial Deli *dura* x AVROS *pisifera* population, and the total length of additional scaffolds anchored in our previous study has improved by 311 Mb (Ong et al. 2019). Hence, our aim in this study

was to maximize the improvement of the oil palm genome assembly by integrating multiple linkage maps from various origins.

3.3. General Methodology

3.3.1 SNP-based marker linkage map construction from multiple origins of oil palm mapping population

Multiple linkage maps were developed from their respective mapping populations using the method as described in section 2.3.1 and 2.3.3. There are five mapping populations in total for this study. Out of the five populations, two were from the hybrid (DxP) (commercial) background while the rest are controlled crossed developed for potential mother palms (DxD) (Table 4).

Table 4. Number of progenies and polymorphic markers for each of the population for linkage mapping.

Populations/Alternative name (Crossing type)	Number of progenies
Deli *dura* x AVROS *pisifera/ Tenera* (DxP)	295
Deli *dura* x Nigerian *dura/ Dura-1* (DxD)	112
Johore Labis *dura* x Johore Labis *dura*/ *Key DD* (DxD)	60
Ulu Remis *dura* x AVROS *pisifera*/ *Key DP* (DxP)	60
Johore Labis *dura* x Johore Labis *dura*/ *Dura-2* (DxD)	201

3.3.2 Construction of linkage map and generation of a comparative map in oil palm

Multiple origins-based linkage maps were built according to section 2.3.4. Comparative linkage maps were visualise using Map chart (Voorrips 2002). Markers were selected based on common/transferable SNPs found within three out of five linkage maps. Pairs of homologs were shown between adjacent linkage maps and connected lines coloured differently.

3.3.3 Anchoring of scaffolds to linkage groups

Linked SNP markers (from OP200K) with flanking sequences located in the oil palm scaffolds (placed and unplaced in the chromosomes) were located using BLASTN (BLAST 2013) with e-value cut-off at 1e-150 and sequence identity at least of 95%. High-density linkage maps built from three mapping populations were merged using ALLMAPS *assembly* module, together with equal weightages among the linkage maps (Tang et al. 2015). Scaffolds were then ordered and oriented according to the merged and weighted linkage maps, thus resulting in a new assembled genome.

3.4. Key findings, output, and general contribution to knowledge

The five mapping populations were from various BPROs for this study with average population size of 146. Mapping densities for most of the populations were comparable, with the total map length ranging from 1,152 to 1,792 cM (Table 5-Table 8). Overall, the SNP markers were well distributed on all chromosomes with ~98% of binned markers in agreement with the chromosome grouping according to the published genome (P5-build). Hence, the linkage grouping reported here was based on the published oil palm genome chromosomes (R. Singh, Ong-Abdullah, et al. 2013). The percentage of mapping errors as observed from each of the tables was low, with less than 2% markers appearing to be derived from different chromosomes

in different crosses. As reported in previous chapter, the percentage of markers mapped to the genome sequence from the linkage maps was on average ~62%. This potentially leaves 38% of the unplaced markers from the 200K array which could be used to place and orient new scaffolds to the physical map, thus improving the genome. The total number of linkage groups for each population are inversely correlated to the number of progenies, indicating insufficient recombination to link small fragments into a single chromosome where population size is relatively small, or polymorphism is limited.

Table 5: Genetic map for the Deli *dura* x Nigerian *dura* (*Dura-1*) based population with the estimation of chromosome and genome recombination rate

Linkage group	Length (cM)	Number of markers	Number of markers mappable to published genome (P5-build)	Percentage (%) of markers from same chromosome	Percentage (%) of markers from different chromosome	Recombination rate estimated from P5-build (cM/Mb)
1	91.0	1328	1234	93%	0.0%	2.1
1.2	52.6	440	360	82%	0.0%	
2	161.9	3382	1771	52%	0.2%	2.5
3	120.6	3015	2104	70%	1.0%	2.0
4	118.9	2422	1403	58%	0.5%	2.1
5	109.9	2359	1678	71%	1.1%	2.1
6	121.5	2450	1249	51%	0.8%	2.7
7	107.1	2675	1636	61%	0.6%	2.5
8	111.5	2181	1267	58%	1.9%	2.8
9	102.5	2129	1281	60%	5.0%	2.7
10	91.9	2148	1374	64%	1.6%	2.9
11	85.7	1658	856	52%	0.4%	2.8
12	54.8	1538	917	60%	0.8%	1.9
13	92.6	1468	725	49%	2.0%	3.3
14	71.1	1244	881	71%	1.1%	2.9
15	70.4	1293	827	64%	0.2%	2.9
16	83.3	820	646	79%	0.0%	3.9

Total	1,647.3	32,650	20,209	62%		2.5
Average					*1.0%*	*2.6*

Table 6: Genetic map for the Johore Labis *dura* x Johore Labis *dura*/ *Key DD* (DxD) population with the estimation of chromosome and genome recombination rate

Linkage group	Length (cM)	Number of markers	Number of markers mappable to published genome (P5-build)	Percentage (%) of markers from same chromosome	Percentage (%) of markers from different chromosome	Recombination rate estimated from P5-build (cM/Mb)
1	26.1	881	655	74.3%	0.1%	
1.2	30.8	446	416	93.3%	0.2%	2.8
1.3	23.3	283	215	76.0%	0.0%	
1.4	110.8	2049	598	29.2%	24.7%	
2	50.0	1027	635	61.8%	0.2%	
2.2	75.4	910	537	59.0%	0.3%	
2.3	55.5	726	335	46.1%	0.0%	3.4
2.4	37.1	306	196	64.1%	1.0%	
2.5	6.7	107	97	90.7%	0.0%	
3	125.9	3514	2274	64.7%	0.4%	2.1
4	32.9	673	509	75.6%	0.0%	1.2
4.2	37.7	177	163	92.1%	0.0%	
5	106.5	2289	1801	78.7%	0.2%	2.1
6	112.3	2600	1167	44.9%	0.9%	2.7
6.2	6.7	86	56	65.1%	0.0%	
7	88.0	939	344	36.6%	0.1%	
7.2	1.7	21	11	52.4%	0.0%	3.8
7.3	75.4	1273	928	72.9%	1.5%	
8	63.6	464	437	94.2%	0.0%	3.9
8.2	93.5	1791	571	31.9%	3.2%	
9	5.0	200	173	86.5%	0.0%	2.4
9.2	87.7	2127	1291	60.7%	4.6%	
10	46.3	2202	1392	63.2%	1.3%	1.5
11	56.4	638	586	91.8%	0.0%	3.4
11.2	46.5	611	193	31.6%	1.0%	
12	88.6	1894	1225	64.7%	0.5%	3.1
13	48.2	1136	735	64.7%	1.1%	1.7
14	79.7	1607	936	58.2%	0.4%	3.3

15	26.1	1086	611	56.3%	0.3%	1.1
16	26.4	299	201	67.2%	0.0%	1.5
16.2	6.0	77	42	54.5%	0.0%	
Total	**1,676.7**	**32,439**	**19,330**			**2.5**
Average					*1.4%*	*2.5*

Table 7: Genetic map for the Ulu Remis *dura* x AVROS *pisifera*/ *Key DP* (DxP) population with the estimation of chromosome and genome recombination rate

Linkage group	Length (cM)	Number of markers	Number of markers mappable to published genome	Percentage (%) of markers from same chromosome	Percentage (%) of markers from different chromosome	Recombination rate estimated from P5-build (cM/Mb)
1	67.7	1231	1149	93.3%	0.2%	
1.2	39.0	395	301	76.2%	0.0%	
1.3	7.6	43	39	90.7%	0.0%	2.1
1.4	7.7	33	22	66.7%	18.2%	
1.5	20.6	21	21	100.0%	0.0%	
2	91.4	1600	992	62.0%	0.0%	
2.2	35.0	83	31	37.3%	0.0%	3.1
2.3	76.1	2221	869	39.1%	0.6%	
2.4	3.4	28	25	89.3%	0.0%	
3	131.5	2951	2058	69.7%	1.3%	2.3
3.2	4.3	14	13	92.9%	0.0%	
4	57.6	854	645	75.5%	0.1%	
4.2	35.2	146	70	47.9%	0.0%	3.2
4.3	91.7	1761	712	40.4%	1.0%	
5	13.5	226	211	93.4%	0.0%	
5.2	7.6	197	190	96.4%	0.0%	1.9
5.3	6.7	74	66	89.2%	1.4%	
5.4	71.1	1901	1152	60.6%	2.8%	
6	111.7	2833	1637	57.8%	0.6%	2.5
7	64.2	961	864	89.9%	0.0%	
7.2	36.1	1546	471	30.5%	2.7%	2.4
7.3	4.2	42	38	90.5%	0.0%	
8	15.2	327	297	90.8%	0.0%	2.2
8.2	71.4	1717	846	49.3%	3.1%	
9	70.1	818	675	82.5%	0.2%	2.4

9.2	21.3	255	195	76.5%	8.6%	
10	76.6	2011	1228	61.1%	1.9%	2.4
11	62.5	1348	331	24.6%	0.6%	
11.2	32.1	413	373	90.3%	0.0%	5.1
11.3	24.9	69	64	92.8%	0.0%	
11.4	34.4	119	38	31.9%	1.7%	
12	90.1	1328	460	34.6%	0.6%	
12.2	10.1	198	175	88.4%	2.0%	3.7
12.3	5.9	111	99	89.2%	0.9%	
13	39.7	616	361	58.6%	0.6%	
13.2	28.6	585	333	56.9%	0.0%	2.7
13.3	7.7	14	11	78.6%	0.0%	
14	10.2	43	36	83.7%	0.0%	3.3
14.2	71.2	1554	1019	65.6%	0.8%	
15	60.3	1218	832	68.3%	0.2%	2.5
16	52.6	286	198	69.2%	0.0%	
16.2	4.2	145	133	91.7%	0.0%	3.6
16.3	6.9	37	36	97.3%	0.0%	
16.4	12.9	16	16	100.0%	0.0%	
Total	**1,792.5**	**32,389**	**19,332**			**2.7**
Average					1.1%	2.8

Table 8: Genetic map for the Johore Labis *dura* x Johore Labis *dura*/ *Dura-2* (DxD) population with the estimation of chromosome and genome recombination rate*.

Linkage group	Length (cM)	Number of markers	Number of markers mappable to published genome	Percentage (%) of markers from same chromosome	Percentage (%) of markers from different chromosome	Recombination rate estimated from P5-build (cM/Mb)
1	57.2	217	214	98.6%	0.0%	
1.2	18.4	71	70	98.6%	0.0%	1.1
1.3	2.8	28	15	53.6%	28.6%	
2	182.7	970	686	70.7%	0.2%	2.8
3	113.1	594	558	93.9%	0.0%	1.9
4	26.3	191	130	68.1%	4.7%	
4.2	37.6	148	108	73.0%	0.0%	1.3
4.3	9.0	82	80	97.6%	0.0%	
5	113.4	672	567	84.4%	1.2%	2.2
6	108.0	523	416	79.5%	1.7%	2.4

7	87.1	443	358	80.8%	0.9%	2.2
7.2	9.0	36	34	94.4%	0.0%	
8	87.6	635	531	83.6%	0.6%	2.2
9	62.2	276	237	85.9%	3.6%	1.6
10	9.5	131	125	95.4%	0.8%	0.5
10.2	8.0	81	80	98.8%	0.0%	
11	73.4	385	280	72.7%	0.3%	2.4
12	56.6	356	325	91.3%	0.6%	2.0
13	57.9	444	349	78.6%	4.1%	2.1
14	10.3	99	90	90.9%	0.0%	1.1
14.2	17.3	88	84	95.5%	0.0%	
15	68.4	333	288	86.5%	0.0%	2.8
15.2	0.8	7	7	100.0%	0.0%	
16	51.9	77	74	96.1%	0.0%	2.4
Total	**1,268.3**	**6,920**	**5,706**			**1.9**
Average					2.0%	1.9

*The mapping population which was genotyped from a custom minimized array especially developed for QTL mapping with the density of 20,000 SNP markers that are in common with OP200K.

Linkage maps developed for both *key DD* and *key DP* populations with only 60 progenies each were highly fragmentize. Hence, only the remaining three mapping populations were subjected for subsequent integration with oil palm scaffolds for genome assembly (Table 9).

Table 9. Number of common markers found each linkage group (LG) in selected mapping populations

Linkage groups (LGs)	*Tenera**	*Dura-1*	*Dura-2*	Total of transferable markers
1	2,422	1,768	288	578
2	3,781	3,382	970	1,423
3	2,582	3,015	594	1,175
4	2,235	2,522	339	742

5	1,405	2,359	672	658
6	1,752	2,450	523	785
7	1,106	2,675	479	639
8	2,512	2,181	668	1,125
9	1,723	2,129	304	797
10	2,912	2,148	212	1,115
11	1,012	1,658	385	421
12	1,312	1,538	356	504
13	620	1,468	444	292
14	1,553	1,244	269	614
15	586	1,293	340	403
16	377	820	77	150
Total	27,890	32,650	6,920	11,421

Figure 11 shows the comparative linkage maps for the four populations with an additional two populations adapted from *Teh et al* (2020) with the reference mapping population. Most the markers were found to be ordered in the same way except for a few regions with potential inversions and breakage between chromosomal segments. Interestingly, in the population GMxDA, possible genome duplication was observed. This can be further investigated by determining the relationships between linkage and physical order of the markers from both maps. Numerous disagreements between the physical and genetic maps have been reported even for the human genome (Kong et al., 2002). Discrepancies in the form of scaffolds including markers from more than one linkage group were indicative of scaffold chimerism in the assembly, hence more linkage maps from different origins were investigated as validation (Kawakami et al. 2014). Theoretically, it is preferable to obtain a consensus order

by combining the data from various sources rather than combining the resulting maps. Although most of the discrepancies arise from a limitation in the quality or quantity of the data, some may result from genuine polymorphism of macro-rearrangements of chromosomes.

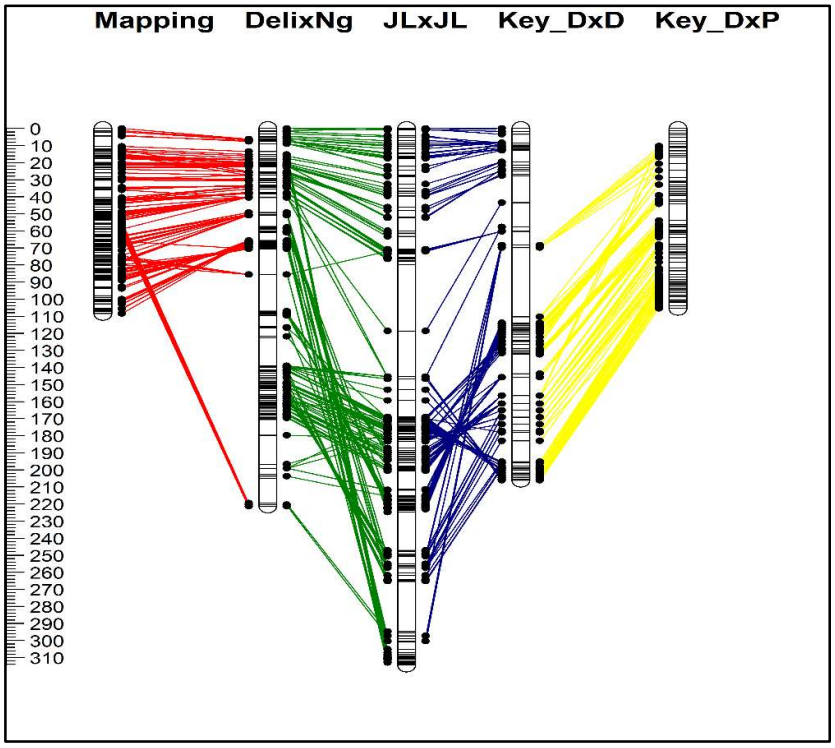


Figure 10. Comparative linkage maps for all mapping populations for chromosome 5. Abbreviated populations: *Mapping (Tenera-1), DelixNg (Dura-1), JLxJL (Dura-2), Key_DxD (Key DD), Key_DxP (Key DP)*

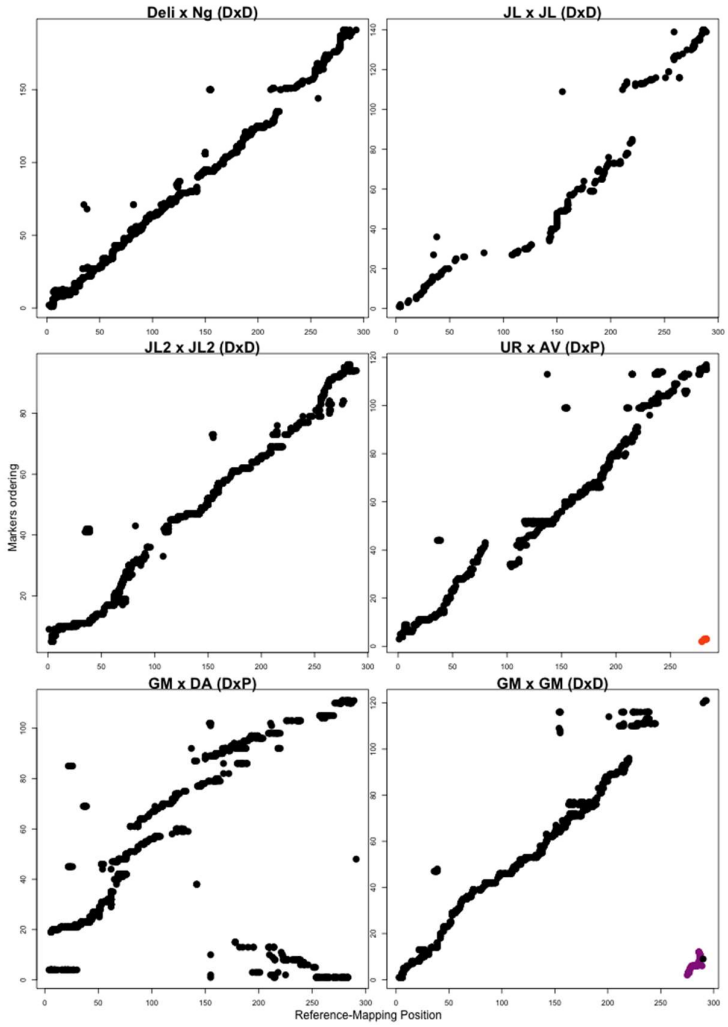


Figure 11. Synteny between the reference ultrahigh-density map (*Tenera-1*) and the other populations for chromosome 3. Each dot represents a marker ordered according to its position in the linkage map. ** genetic maps for *GMxDA(DxP)* and *GMxGM (DxD)* adapted from *Teh C.K et. al. (2020)*

Table 10. Genome assembly improvement by coverage and quality using linkage map for Deli *dura* x AVROS *pisifera* (*Tenera-1*) as reference

Linkage group/ Chromosome	Linkage Length (cM)	Current chromosome size – P5-build (Mb)	Genome version PMv6 (Mb)	Improvement (%)	Average Recombination rate genome P5-build (cM/Mb)	Average Recombination rate genome PMv6 (cM/Mb)
1	122.1	68.4	71.99	22%	1.90	1.47
1.2			11.22			
2	112.0	65.6	142.51	117%	2	0.91
2.2	9.7					
3	103.2	60.1	89.74	49%	1.9	1.15
4	121.1	57.3	94.88	66%	2.1	1.28
5	33.0	52.0	89.32	72%	0.8	0.46
5.2	7.8					
6	81.2	44.4	91.34	106%	1.8	0.89
7	25.7	43.5	75.55	74%	0.9	0.51
7.2	12.6					
8	89.4	40.2	83.14	107%	2.2	1.08
9	61.4	38.1	76.24	100%	1.6	0.81
10	74.4	31.9	59.31	86%	2.3	1.25
11	50.8	30.1	63.68	112%	1.7	0.80
12	84.9	28.8	61.99	115%	2.7	1.37
13	27.4	27.8	55.09	98%	1	0.50
14	56.8	24.4	45.89	88%	2.3	1.24
15	42.5	24.3	45.32	86%	1.7	0.94
16	27.7	21.4	21.72	2%	1.3	1.28
Total	**1151.7**	**658.0**	**1178.93**	**79%**		
Mean					**1.8**	**0.99**

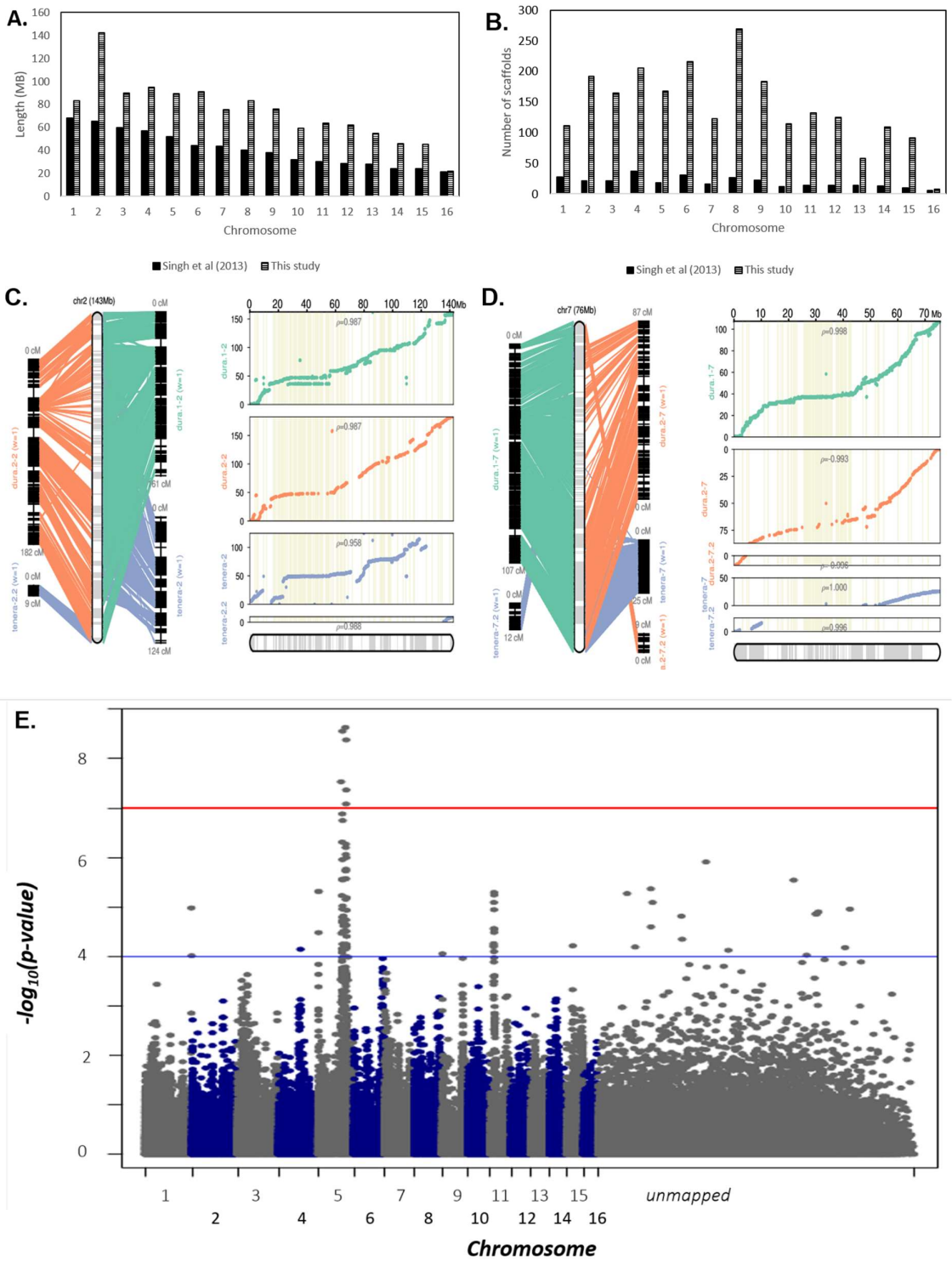
The newly assembled genome now contains 1,179 Mb with N50 of 83.1 Mb with 16 major pseudo-chromosomes plus one unlinked linkage group fragment belonging to chromosome 1 (Figure 13). The new estimated genome wide recombination rate is close to 1 cM/Mb. On average, each pseudomolecule in the PMv6 genome was assembled using 142 scaffolds with an average length of 73.7 Mb, whereas only 19 scaffolds giving an average length of 41.1 Mb were reported in Singh et al., 2013. Chromosome 2 was also extended by 1.2-fold and now is the largest pseudomolecule, accounting for 12% of the total PMv6 assembly length (**Error! Reference source not found.**A-B).

Linkage directed genome assembly improvement has assigned previously unplaced scaffolds onto the pseudochromosomes. **Error! Reference source not found.** shows the correlation of markers by linkage against physical locations in chromosome 2 and 7. The sigmoidal curves were clearly observed distinguishing telomere regions from centromere where almost no recombination events were observed. A previously developed linkage map for a *tenera* population (**Error! Reference source not found.**C: box 3) for these two chromosomes where fragmented into two linkage groups, but here are now linked into a single pseudo-chromosome. The results show that assembly improvement using the commercial Deli *dura* x AVROS *pisifera* map (Ong et al. 2019) (blue) was only limited to one end of the pseudo-molecule, especially on Chromosome 7. Nevertheless, the shortcomings were solved when integrating the semi-wild Deli *dura* x Nigerian *dura* map (green) and the *Johore Labis dura* map (orange). The sigmoidal curves of Chromosome 2 and 7 clearly distinguishing both telomere regions from centromeric regions where limited recombination events were observed. The correlation between linkage and physical positions reached more than 0.90 in each mapping family, and the collinearity between each linkage maps can be observed with consistence marker orderings within each LG. Similar results were observed for the remaining chromosomes and are shown in Figure 14. The details of total SNPs residing within genomic

scaffolds used to build PMv6 genome assembly can be found in Table 11. Using the correlation outputs, the genome-wide recombination rate was estimated to be 1 cM/Mb, which is about half of the previously reported estimation based on the P5-build genome (Table 10). This investigation may infer overestimation of the recombination rate in our previous estimation, mainly due to the length extension of the PMv6 genome assembly reported here. Understanding recombination differences has distinct implications for population structure, gene evolution, and genetic improvement (Kianian et al. 2018).

The improved PMv6 genome was used to evaluate the improvement of mapping resolution of the previously reported GWAS for mesocarp oil content (Teh et al. 2016), reducing the unmapped SNPs from 40% (37,003) (Figure 12E) to only 13% (12,175) (Figure 12F). The major QTLs remain at Chromosome 5 and Chromosome 11, but the significance level on Chromosome 11 was increased to a –log p-value = 5.9 from 5.2 and more SNPs surpassed genome-wide -log p-value cutoff at 4.0, due to SNPs newly assigned to Chromosome 11, improving trait location. From the perspective of genomic features and gene annotations, the PMv6 genome assembly now includes 5769 genes, 7705 protein-coding regions (CDS), and 1563 non-coding RNA, which could not be placed into the P5-build genome (Figure 12G). Interestingly, among the non-coding RNA, 48% of total predicted ribosomal RNA (rRNA) were mapped for the first time in PMv6.

Figure 13 depicted placement of genomic scaffolds for 16 chromosomes with comparison to the current assembly (PMv6) with previously published version (P5-build). Significant improvement was observed for overall placement of 1,996 scaffolds into each chromosome rising from the original 304 scaffolds. Integration of linkage with physical maps for all 16 chromosomes in oil palm as reported in Figure 14.


A.
Length (MB)
Chromosome
B.
Number of scaffolds
Chromosome
Singh et al (2013)
This study
C.
chr2 (143Mb)
D.
chr7 (76Mb)
E.
-log10(p-value)
unmapped
Chromosome

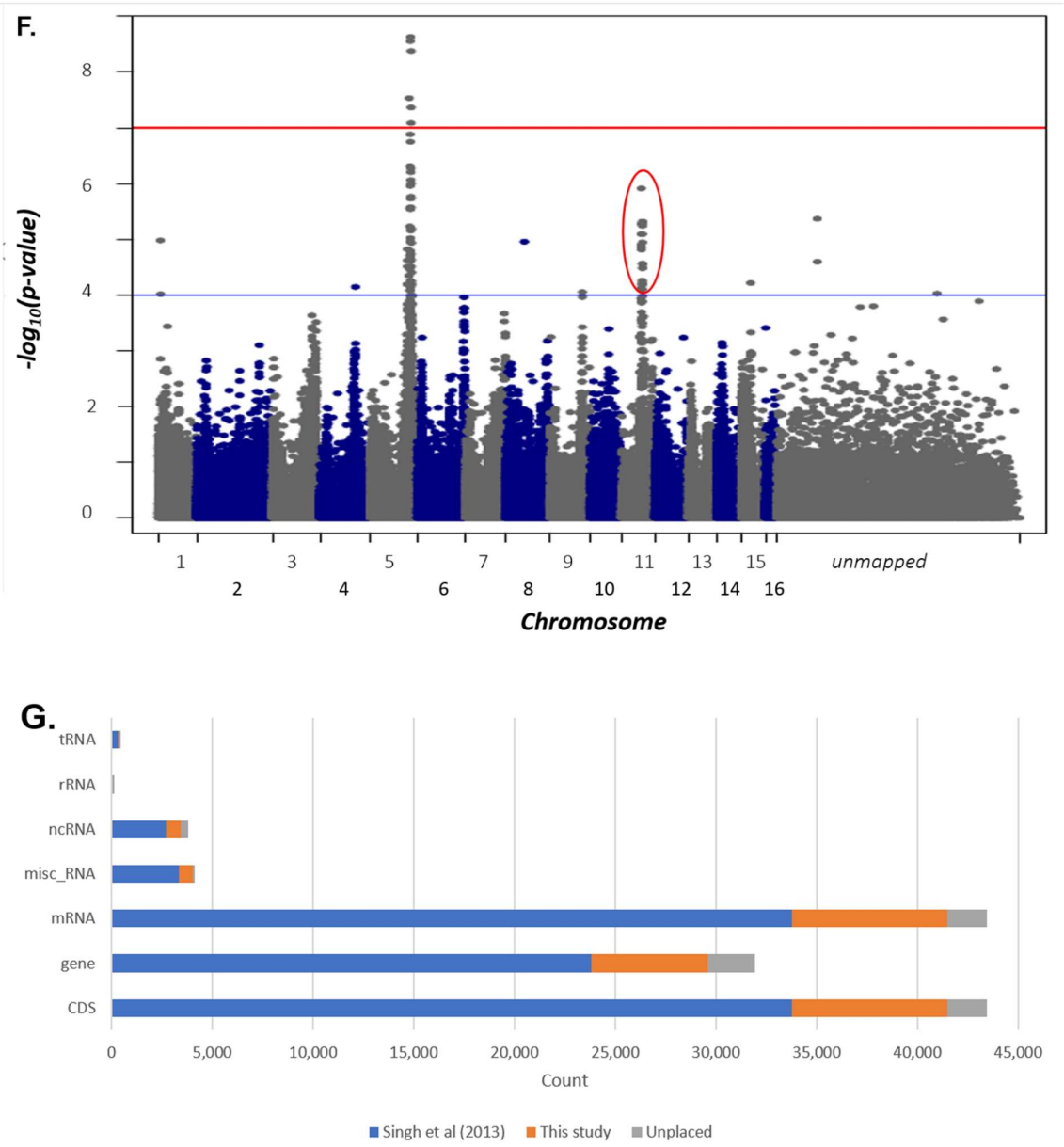


Figure 12. Refined assembly of oil palm genome (PMv6). (A) Chromosome length (Mb) comparison between P5-build, Singh et al., (2013), and PMv6 (this study). (B) Number of scaffolds from previous physical map) and the improved physical map developed in this study. (C) Build-up of PMv6 by scaffold anchoring using multiple linkage maps; left panel represents the central combined linkage map from three mapping populations, and the right panel shows the correlation plots of each populations' genetic versus physical order; presented in the order of Deli dura x Nigerian dura (*dura 1*—green), Johore Labis dura (*dura 2*—orange), and Deli *dura* x AVROS *pisifera* (*tenera*—blue) for Chromosome 2 and (D) Chromosome 7 as examples. (E) Genome-wide association studies (GWAS) Manhattan plot for a commercial oil palm population for one of the oil yield traits for the genome published by P5-build. (F) GWAS Manhattan plot for the same commercial oil palm population for one of oil yield trait after assembly improvement in this study, circle in red represent newly assigned and significant SNPs. (G) Statistics of genomic features improvement after placement of scaffolds into chromosomes for P5-build and this study.

Table 11. Summary for consensus map generated from multiple linkage maps and statistics for SNP markers and scaffolds used to build genome assembly for PMv6.

Summary for consensus map			
	Anchored	Oriented	Unplaced
Markers(unique)	39,616	31,657	347
Markers per Mb	33.6	41.7	11.4
Scaffolds	2,272	434	148
Scaffolds with 1 marker	869	0	99
Scaffolds with 2 markers	335	56	24
Scaffolds with 3 markers	175	24	7
Scaffolds with >=4 markers	893	354	18
Total bases	1,178,731,551	759,750,341	30,463,349
Percentage	97.5%	62.8%	2.5%

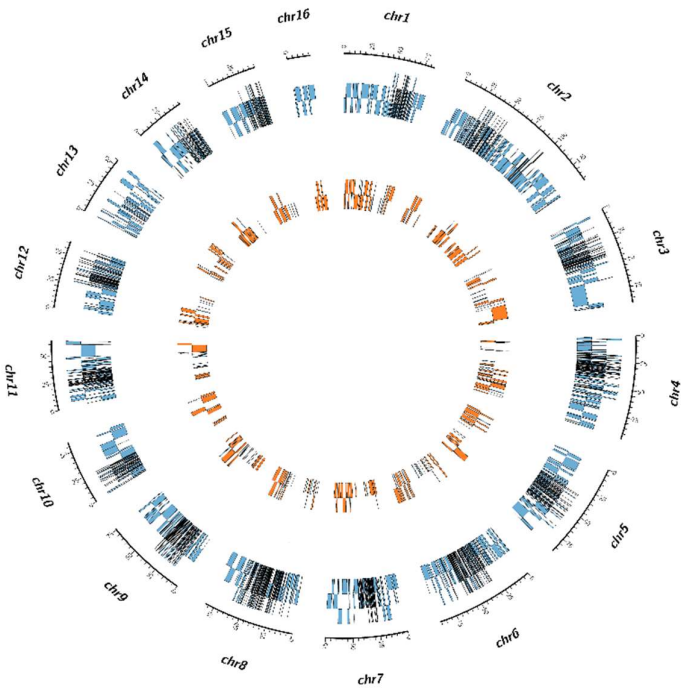
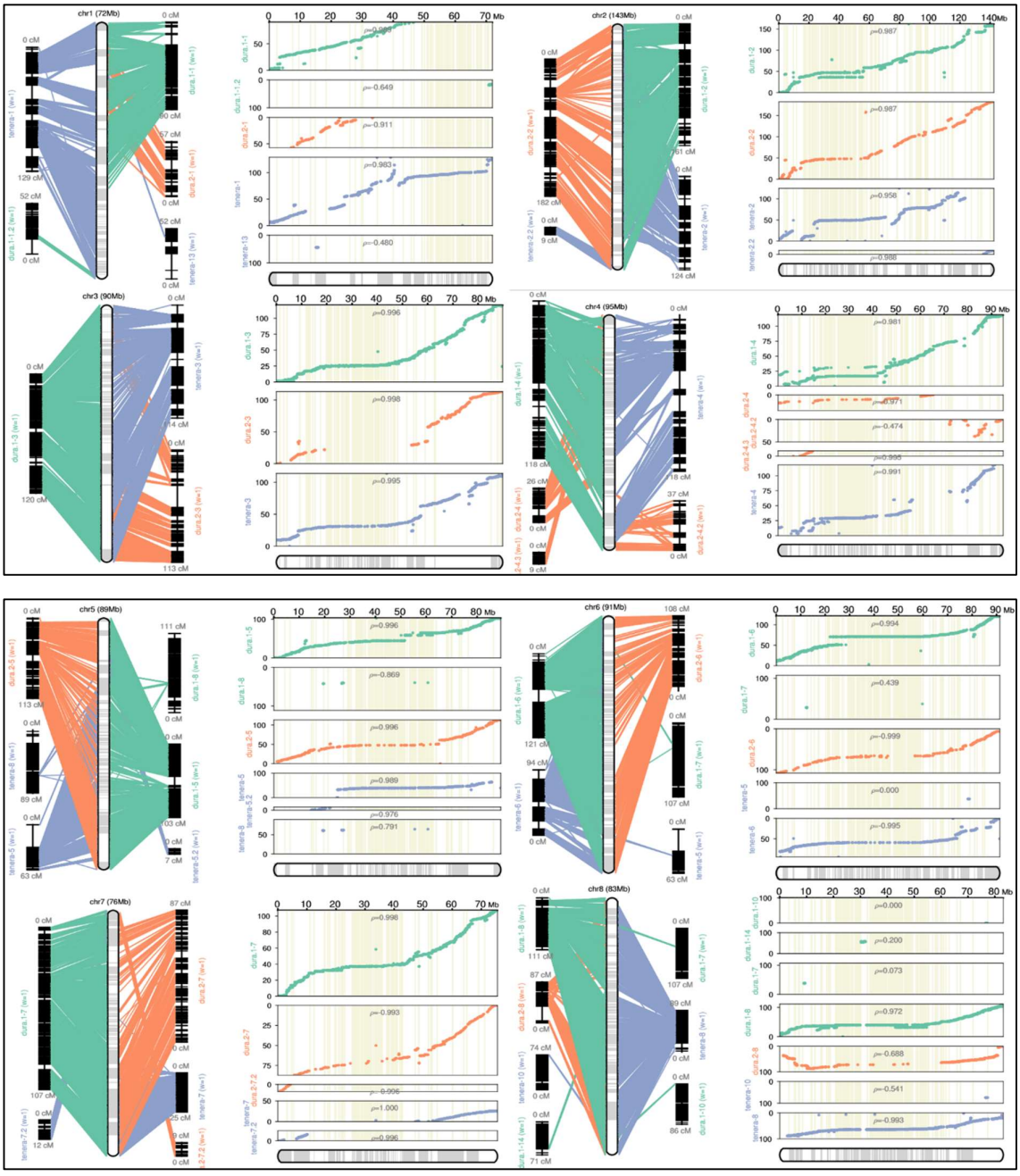


Figure 13: Genome assembly improvement by scaffolds placement for PMv6 (blue) compared to P5-build (orange) for 16 chromosomes.


chr1 (72Mb)
chr2 (143Mb)
chr3 (90Mb)
chr4 (95Mb)
chr5 (89Mb)
chr6 (91Mb)
chr7 (76Mb)
chr8 (83Mb)

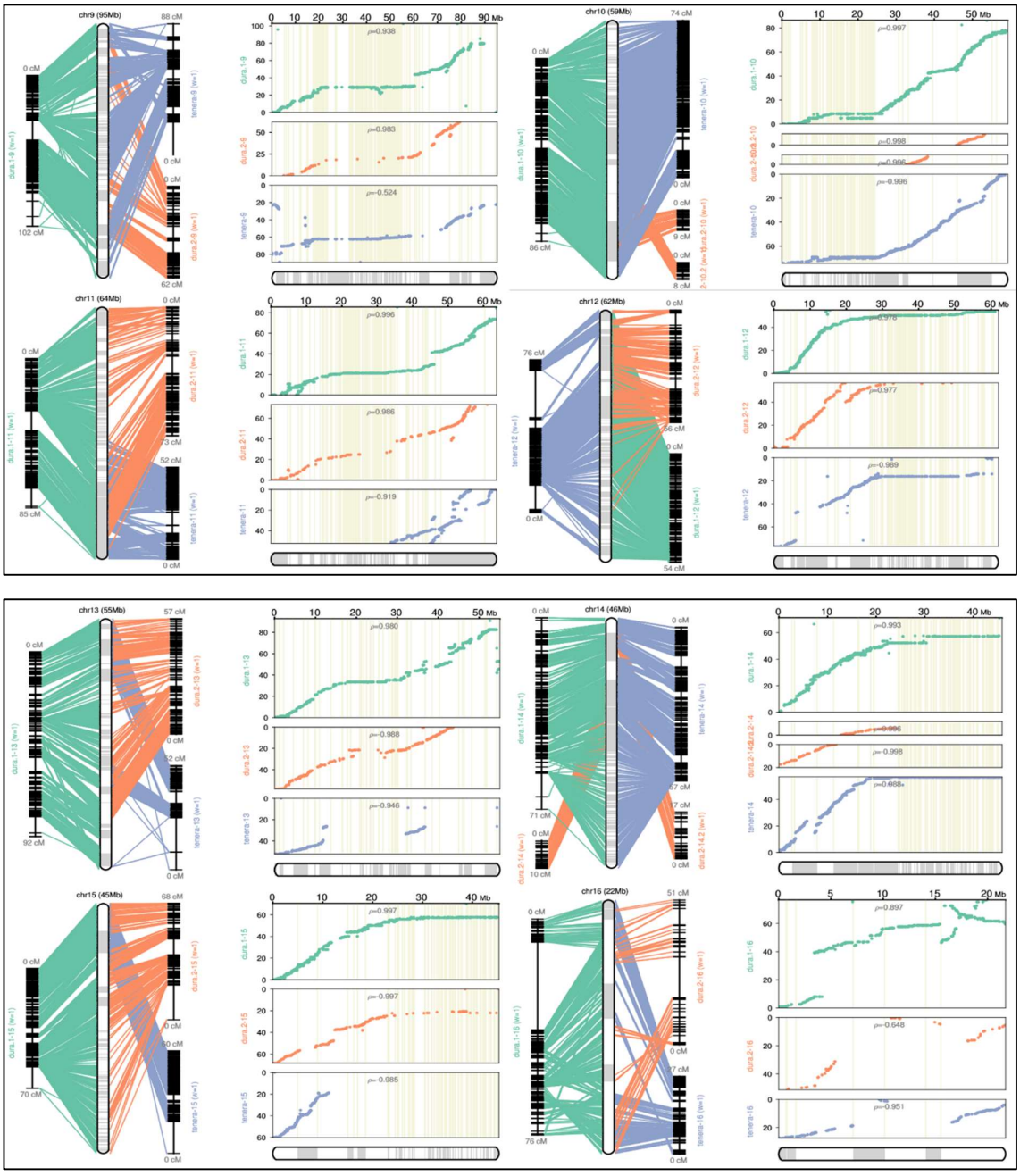


Figure 14. Integration of linkage with physical maps for oil palm, Chromosome 1 to 16

3.5. Conclusions and key recommendations for future work

The newly assembled genome of oil palm – named here PMv6 - has 79% greater physical length in the 16 chromosomes compared to the previously published version (P5-build). By

integrating three high-density linkage maps and using them to place the genome scaffolds, 16 mega-scaffolds were ordered and oriented representing 66% of the total predicted genome size.

By integrating the three linkage maps, a total of 11,421 family-transferable/common SNPs were used to anchor 77% of the assembled 1.5 Gb genome, in order to generate the new oil palm physical map, PMv6 (Figure 13), compared to 43% in the previously published P5-build physical map. Using Chromosome 2 and 7 as examples, the quality of scaffold anchoring based on the integrated linkage maps can be visualized and was validated through a correlation between linkage and physical positions, allowing the recombination landscape throughout each chromosome to be determined. Species with larger genomes often have higher levels of recombination at the telomeres and lower recombination in the heterochromatic and centromeric regions despite genes being present within transposable element (TE) (Milo, R.; Phillips 2016; Shen et al. 2017) clusters, as reported in wheat (Jordan et al. 2018; Sidhu and Gill 2005). In such a scenario, a sigmoidal curve of the recombination rate should be expected, but this was not clearly observed in both chromosomes when assembled from the commercial Deli *dura* x AVROS *pisifera* map (blue), suggesting only a partial improvement. However, the integration of the semi-wild Deli *dura* x Nigerian *dura* map (green) and the *Johore Labis dur*a map (orange) enabled extension of both chromosomes with a sigmoidal curve of recombination from telomere-to-telomere, although whether this represents the full chromosome needs further validation, through techniques such as Fluorescence in situ hybridization (FISH) technique.

The PMv6 genome enabled an increase in the mapping resolution of the reported GWAS for mesocarp oil content (Teh et al. 2016) by locating an additional 65% (24,828) previously unplaced SNPs Figure 12F). Among the total SNP markers mapped in PMv6 assembly, there are 32,552 SNPs located within genic/non-coding RNA regions, equivalent to 25% of the total SNP assigned to scaffolds/genome sequence. Increased marker density allows

better capture of linkage disequilibrium between origins and render GWAS more powerful. Hence, PMv6 enables a more complete set of annotated genes and non-coding RNA to be located that are potentially responsible for phenotypic changes residing within QTL to be identified after conducting GWAS.

Generally when estimating the completeness of a genome assembly, one of the important measurement would be the scaffolds are ordered and anchored into pseudomolecules or linkage groups representing chromosomes (Veeckman, Ruttink, and Vandepoele 2016). To date, this current oil palm assembly has the largest number and coverage of anchored scaffolds into pseudochromosomes. In addition, such high coverage allows a comparative genome approach as a cost-effective way to improve and validate for other palms' genome assemblies.

3.6. Published paper contributing to objective two:

An Improved Oil Palm Genome Assembly as a Valuable Resource for Crop Improvement and Comparative Genomics in the Arecoideae Subfamily (Ong et al. 2020) *(Appendix B)*

Chapter 4: Application of genome-enabled biology in oil palm breeding (Objective three)

4.1. Executive summary

Comparative genomics study against two other important palm species - date palm (*Phoenix dactylifera* L.) and coconut palm (*Cocos nucifera* L.) confirmed the high level of conserved synteny among these palm species. We also used the improved oil palm version as a reference to extend the date palm physical map. As perennial tree crops, a total of 107 palm species belonging to *Arecoideae*, the largest subfamily of *Arecaceae* would benefit from genome-enabled breeding as an essential approach for the continued development and improvement of the crop. The other two major palm species of economic importance are coconut (*Cocos nucifera*) and date palm (*Phoenix dactylifera*), with each crop having their specific breeding objectives. One of prominent application of genome-based breeding for oil palm and date palm was GWAS. The Sime Darby Plantation oil palm research group has taken further step towards deploying GWAS into GS as a refinement of GS models - GWAS directed GS as a cost and effective approach for commercialization. This is an important milestone achieved for oil palm genome-enabled breeding for sustainable production of the oil crop and could be reproduced in other palms species too. Besides, how the improved oil palm assembly, namely using Physical Map version 6 (PMv6) could serve as an important reference for comparative genomic studies among coconut and date palms, with expansion to other *Arecaceae* members in the future was demonstrated as well.

4.2. Background information

The *Arecaceae* (formerly known as *Palmeae*) family of the monocot order *Arecales* is a plant family hosting perennial palm species that mostly grow across the equatorial belt. The family is made of 181 genera with 2600 species (Christenhusz and Byng 2016). Among them, the genus *Elaeis* (oil palm) and genus *Cocos* (coconut) together are the world's largest source of oils and fats (about 35%), while date palm plays a major role in food security and agricultural production in arid and fragile ecosystems worldwide. Genomics resources have been developed for date palm and coconut but less is known about the comparative evolutionary genetics of these perennial crops (Hazzouri et al. 2019). Date palms has 3,000 recognizable varieties, while coconut has two large groups of varieties: tall and short varieties. To-date, the status of genomes assembled within palm species are still limited (Table 12). With only date palm and oil palm genomes having full annotations and both became publicly accessible in 2013 (Al-Mssallem et al. 2013; R. Singh, Ong-Abdullah, et al. 2013). These important milestones have helped to illuminate the research focus towards understanding the genetics of fruit developments and yield improvement.

GWAS was first reported in one of the yield components in oil palm (Teh et al. 2016), and followed by date palm for fruit colour and sugar composition (Hazzouri et al. 2019). The application of the genome in oil palm breeding allowed the transition from GWAS to GS in recent studies – hereby define as GWAS directed GS (Kwong, Ong, et al. 2017; Kwong, Teh, et al. 2017). The studies concluded that models built to obtain genome estimated breeding value (GEBV) for complex traits such as oil yield could be refined by understanding the LD patterns with study population and using GWAS to rank the genetic markers based on their significant association with the phenotypes.

Table 12. Published genome assembly status of agronomically interesting palm species

SPECIES	SCIENTIFIC NAME	SIZE (MB)	PERCENTAGE OF PLACED SCAFFOLDS IN CHROMOSOMES (%)	NUMBER OF GENOME VERSIONS
OIL PALM	*Elaeis guineensis*	1535.18	43%	2
COCONUT	*Cocos nucifera*	2202.46	46%	3
DATE PALM	*Phoenix dactylifera*	772.32	50%	4

Genomic resources for these three palm species also permits comparative phylogenomics since both oil and date palm genomes representing important annotated references for the sequencing of a great diversity of other palm genomes in the future. Phylogenomics studies in palms by Barrett et al mainly utilized the mitochondrial genome, plastome and transcriptome as nuclear genomes were very limited (Barrett et al. 2016). With the availability of the improved oil palm nuclear genome, more detailed comparison is now possible, especially within the largest subfamily of *Arecaceae* to refine the structure of the *Arecaceae* ancestral genome, in a similar way to the *Rosaceae* family was described upon completion of genome sequencing of apple, strawberry and peach (Troggio et al. 2012). Future challenges - just like the *Rosacease* genome research - would be continuation of the development of genomic tools for species across the family. Beside identifying genetic markers differentiating traits of interest, it is also crucial to understand how the genome sequence comprehensively encodes developmental programs and environmental responses, especially adaptability, as palms has high durability in extreme weather.

4.3. General methodology

Comparative studies performed within the palm family, using the improved oil palm genome as the reference to date palm and coconut.

4.3.1 Genome and gene comparisons

Palms genomic sequences were downloaded from NCBI. Alignments of the coconut (Xiao et al. 2017) and date palm (Al-Mssallem et al. 2013) genome against the current oil palm genome PMv6 was done using *Nucmer* and *Promer*, which is a module developed in Mummer (Kurtz et al. 2004; Marçais et al. 2018) an ultrafast sequences aligner, made for genome alignments.

The syntenic relationship between both oil palm and coconut chromosomal alignments were visualised using CIRCOS software (Krzywinski et al. 2009). Chromosome pairs with strong syntenic relations between oil palm and coconut are linked by same ribbon colours. Candidate genes of interest were also aligned using clustalx (Clustalw, To, and Multiple 2003; Ma et al. 2007) via translated protein sequences and orthologs of other closely related monocots as outgroup: asparagus (*Asparagus officinalis*), banana (*Musa acuminata*), sorghum (*Sorghum bicolor*), corn (*Zea mays*), and turf (*Oryza brachyantha*) protein sequences were identified from NCBI. The phenogram was build using neighbor-joining (NJ) with 1000 bootstrap iterations and visualized using a phylogenetic tree viewer, namely Figtree (v1.4.4) (Rambaut 2018).

4.3.2 Reconstruction of date palm pseudo-chromosomes

Promer output from the step above was used to tile the date palm's query scaffolds according to oil palm pseudochromosome order. These were then ordered and orientated into the Genbank agp format and the pseudo-chromosomes were built using *agp* module in ALLMAPS. SNP markers obtained from date palm genetic maps (Mathew et al. 2014) were aligned to new

pseudo-chromosomes using the same BLASTN criteria above. The integration of SNP markers location into genetic map against physical map locations were plotted using ALLMAPs *path* module as well. Chimeric contigs which contained ambiguous markers that mapped to more than two linkage groups were then processed using the *split* module. Ambiguous short contigs were excluded during the second round of rebuilt of pseudo-chromosomes by running ALLMAPs *path* module again.

4.4. Key findings, output, and general contribution to knowledge

The improved oil palm genome, PMv6 was subsequently used for genome comparison with coconut (Xiao et al. 2017) and date palm (Hazzouri et al. 2019). The results confirmed a high level of syntenic block conservation between the oil palm and the other two palm species (Figure 15). For coconut, the 16 homologous chromosome groups (2n = 32) appeared to be highly coherent with the oil palm genome, but we observed possible fusion and fission event differences on oil palm-based Chromosome 2 (eg-2) and 10 (eg-10), respectively. Compared to coconut, a 30 Mb fragment of eg-2 fused with eg-7 accounts for coconut-based Chromosome 2 (cn-2), suggesting a fission/fusion event since the separate evolution of the two species. In a similar comparative way, a 20 Mb fragment of eg-10 has fused with eg-16 to form cn-8 (Figure 15).

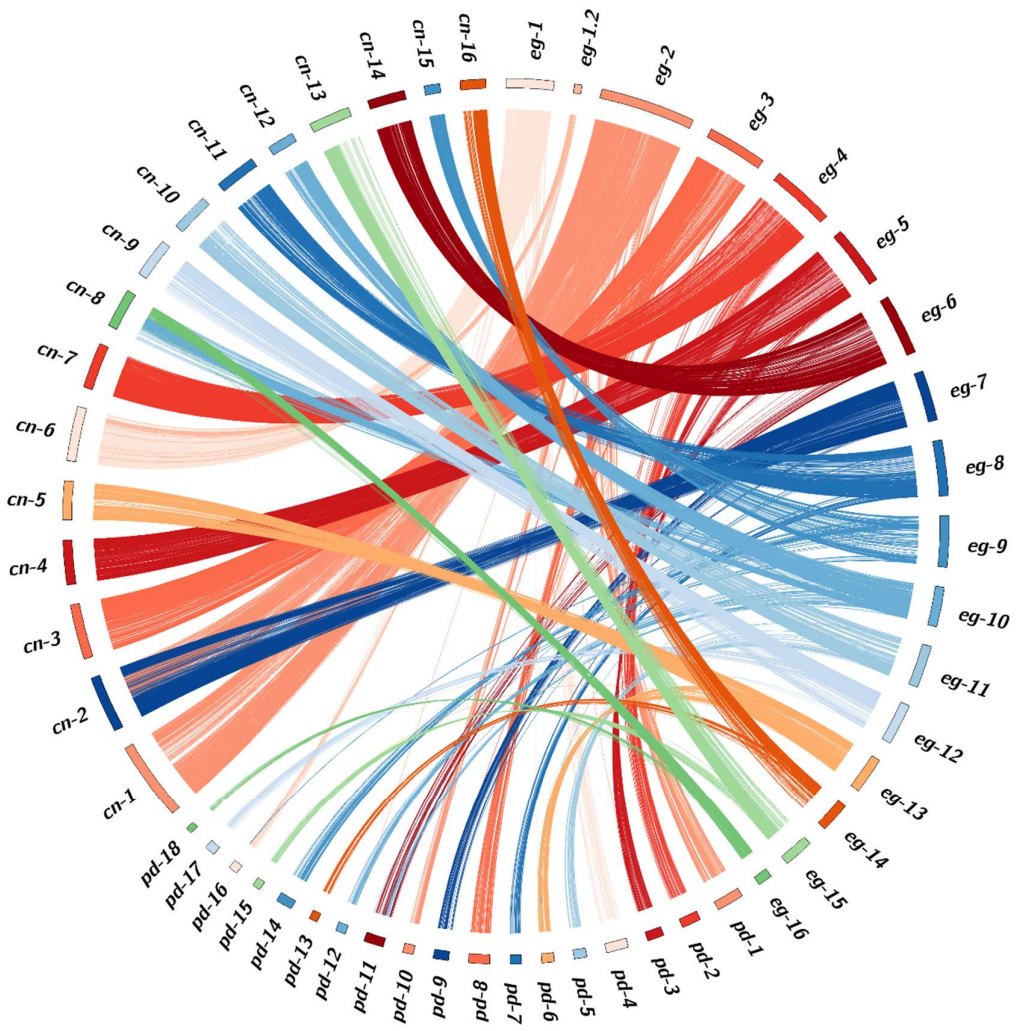


Figure 15. Synteny analysis including three palm species: oil palm (eg), coconut (cn), and date palm (pd). Chromosome pairs with strong syntenic relations were linked by the same ribbon color.

Genomic comparisons of chromosome scale (Figure 16: coconut) as well as scaffolds (Figure 17: date palm) against oil palm showed highly syntenic blocks between these species. Since there is strong macrosynteny between palm species, the oil palm improved genome assembly, could be used as a scaffold/reference to construct a draft genome of date palm, as has been done for sheep and cattle genomes (Dalrymple et al. 2007; Troggio et al. 2012). Thus, initial date palm pseudo-chromosomes were built by placing and ordering scaffolds of date

palm with the oil palm physical map. Assembly improvement for date palm resulted in placement of short scaffolds into pseudochromosomes coupled with ordering of genetic maps for date palm showing major agreement between genetic and physical maps similarly to oil palm (Figure 17). In general, the genome conserved synteny between date palm and oil palm was more complicated. By referring to Ziwen et al. (2015), oil palm and coconut were estimated to have diverged from date palm approximately 62.5 million years ago (He et al. 2015). Chromosome fusion/fission events between date palm-based Chromosome 4 (pd-4) and pd-16 and between pd-1 and pd-10 may have led to formation eg-1 and eg-2 in the oil palm genome, respectively. This possibly explains why date palm (2n = 36) has two additional homologous chromosome groups compared to oil palm. In addition, pd-11 compared to oil palm has two chromosome fragments with one of them fused with pd-12 to form eg-10 and the remaining pd-11 fragment being maintained as eg-6 in the oil palm genome. Figure 17 also shows the integration of the date palm genetic map (Mathew et al. 2014) against the polished date palm pseudo-molecules, based on the oil palm chromosome 10 as template. This work suggests that chromosome 10 of oil palm could be the gender-linked chromosome and location of the same genes in two different chromosomes could be one reason that the date palm is dioecious as reported that LG12 may be the sex chromosome (Mathew et al. 2014). The oil palm directed date palm genome assembly improvement was validated by later publication of an improved date palm genome using long range sequencing method – Pacific Biosciences (Hazzouri et al. 2019).

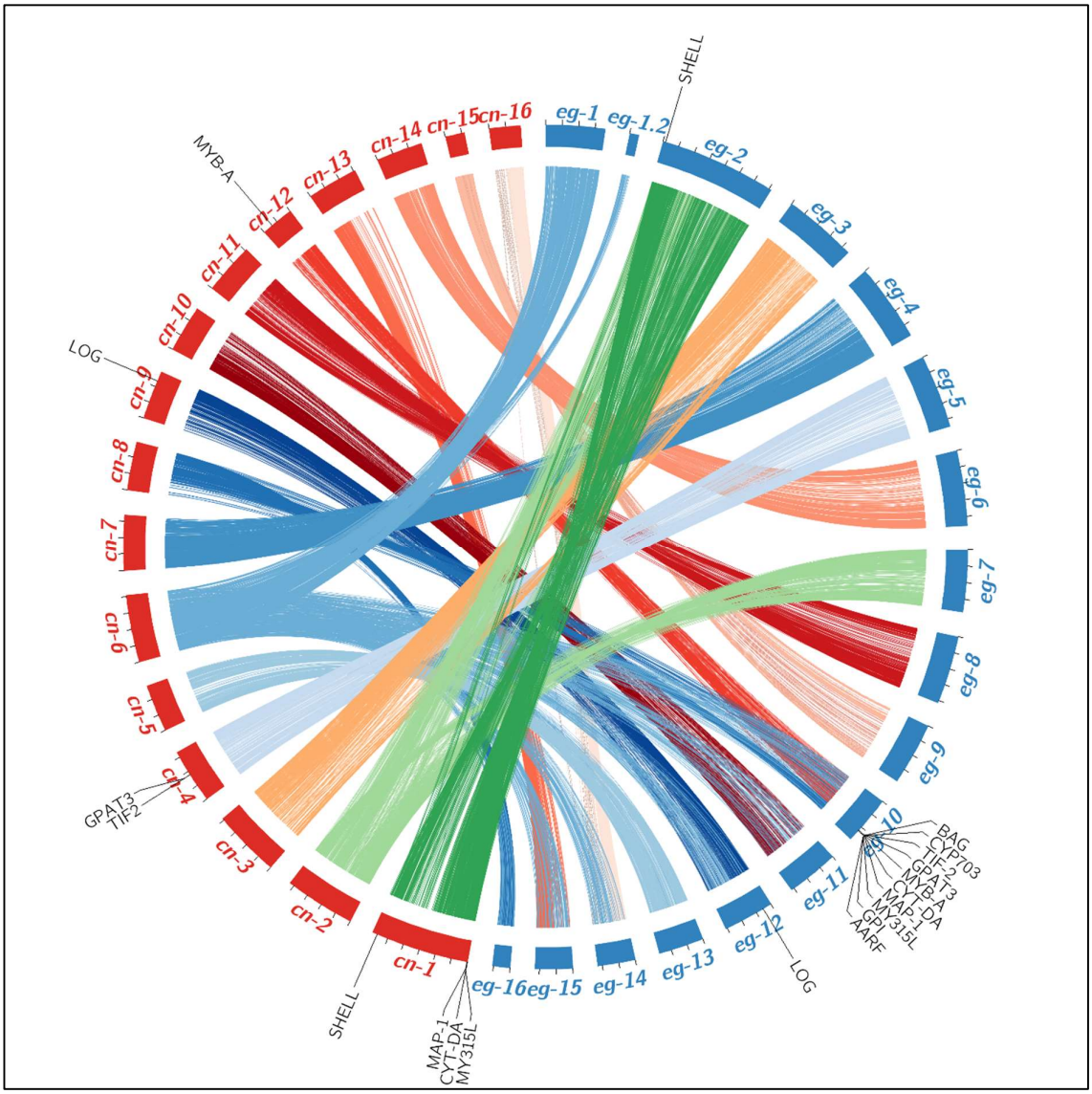


Figure 16. Comparative genome comparisons of coconut against oil palm, both in chromosomal formats with annotated locations of the genes of interests. (cn- coconut; eg- oil palm)

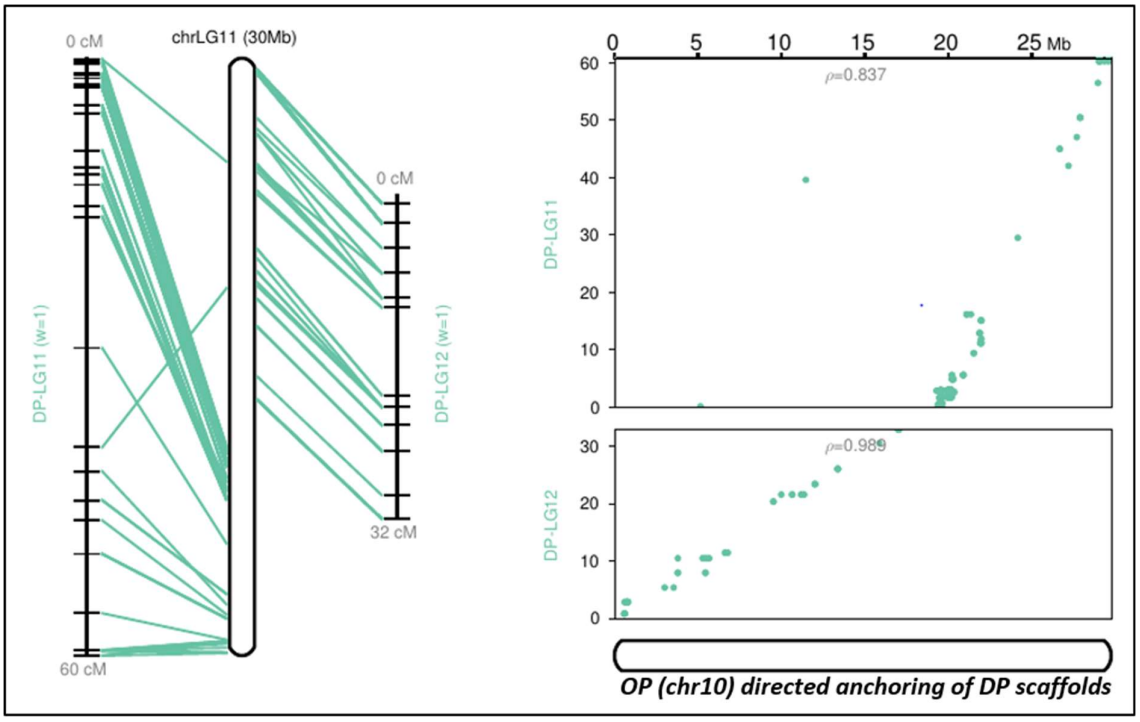


Figure 17: Date palm chromosome built from genome comparison against oil palm's chromosome 10. Integration of genetic markers from reported gender-linked linkage group (LG12) and LG11 of date palm shown high correlation of genetic and physical ordering, suggesting this could be sex-determining chromosome of the monoecies oil palm.

Gene level comparisons was done using SHELL gene, which is a homologue of SEEDSTICK (NCBI Gene ID: LOC105034563)—a type II MADS-box transcription factor reported in oil palm (at 3.05 Mb of eg-2) with the orthologs of coconut (at 116 Mb of cn-1) and date palm (at 38 Mb of pd-1). In the oil palm genome, the *pisifera* allele (sh^{AVROS}) results in an amino acid change from lysine to asparagine, causing the absence of shell surrounding the fruit kernel observed *pisifera* fruit (R. Singh, Low, et al. 2013) (Figure 18a). The analysis was also extended to other candidate genes including Cytochrome P450 *(CYP703)* (LOC105059962), glycerol-3-phosphate acyltransferase 3-like *(GPAT3)* (LOC105059961), Cytidine deaminase-like *(CYT DA)* (LOC105059743), and Lonely-Guy *(LOG)* (LOC105055182) reported to be involved in sex determination of date palm (Torres et al. 2018). All the candidate genes (except

LOG, located at 10.86 Mb of eg-12) were located at 44–45 Mb of eg-10 in the oil palm genome but have yet to have the scaffolds containing them placed in the date palm genome.

In Figure 18b, a phenogram based on the sex-determination-related *LOG* genes revealed that coconut has the shortest evolutionary distance to oil palm followed by date palm, among other monocots outgroup taxa.

(a)

```
CLUSTAL 2.1 multiple sequence alignment


Elaeis_guineensis       MGRGKIEIKRIENTTSRQVTFCKRRNGLLPNAYELSVLCDAEVALIVFSSRGRLYEYANN
Cocos_nucifera          MGRGKIEIKRIENTTSRQVTFCKRRNGLLPKAYELSVLCDAEVALIVFSSRGRLYEYANN
Phoenix_dactylifera     MGRGKIEIKRIENTTSRQVTFCKRRNGLLPKAYELSVLCDAEVALIVFSSRGRLYEYANN
                        ******************************:*****************************

Elaeis_guineensis       SIRSTIDRYKKACANSSNSGATIEINSQQYYQQESAKLRHQIQILQNANRHLMGEALSTL
Cocos_nucifera          SIRSTIDRYKKACANSSSSGATIEINSQQYYQQESAKLRHQIQILQNANRHLMGEALSSL
Phoenix_dactylifera     SIRSTIDRYKKACANSSNSSAAIEINSQQYYQQESAKLRHQIQILQNANRHLMGESLSSL
                        *****************.*.*:**********************************:**:*

Elaeis_guineensis       TVKELKQLENRLERGITRIRSKKHELLFAEIEYMQKREVELQNDNMYLRAKIAENERAQQ
Cocos_nucifera          TVKELKQLENRLERGITRIRSKKHELLFAEIEYMQKREVDLQNDNMYLRAKIAENERAQQ
Phoenix_dactylifera     TVKELKQLENRLERGITRIRSKKHELLFAEIEYMQKKEVELQNDNMYLRAKIAENERAQQ
                        ************************************:**:********************

Elaeis_guineensis       AGIVPAGPDFDALPTFDTRNYYHVNMLEAAQHYSHHQDQTTLHLGYEMKADPAAKNLL
Cocos_nucifera          AGIVQAGPEFDTLPTFDSRNYYHVNMLEAAQHYSHHQDQTTLHLGYEMKADPAAKTLL
Phoenix_dactylifera     AGIVPAGPEFDTLPTFDSRNYYHVNLLEAAQHYSQHQDQTTLHLGYEMKADPAAKNLL
                        **** ***:**:*****:*******:********:********************.**
```

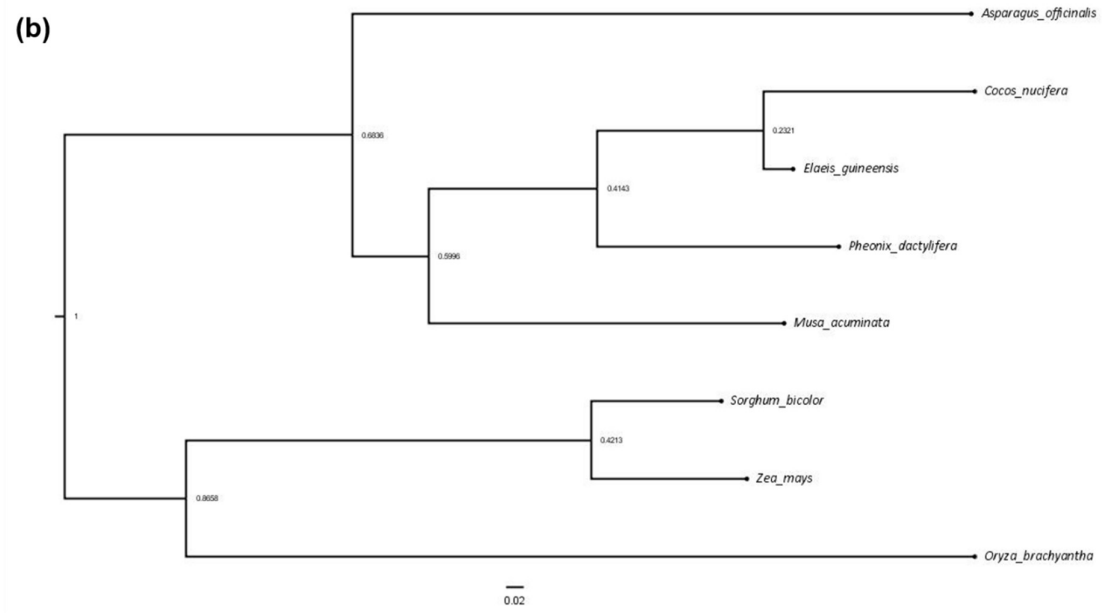


Figure 18. High conservation of genes among palm species with minimal amino acid changes. (a) Alignments of *SHELL* gene encoding for the layer of shell surrounding the kernel for oil palm and the orthologues from coconut and date palm. (b) Phenogram tree for *LOG* gene orthologues, which play a role in sex determination for date palms; orthologues of the gene among palms and other monocots outgroup taxa were included.

Due to the high levels of collinearity of gene order and genomic sequences between the palm species, the PMv6 genome was adopted as a reference to improve the published genome assemblies of date palm. We started with the earlier version, namely PDK_30 (GenBank assembly accession: GCA 000181215.2) (Al-Dous et al. 2011). Genomic scaffolds from date palm were tiled following oil palm ordering for each pair of syntenic chromosomes, then the joined scaffolds were split if required following the published date palm genetic map (Mathew et al. 2014). This resulted in anchoring 78% of scaffolds to the physical map from the total length of the full assembly (381.5 Mb). The next test was performed on the more recent assembly of date palm genome (GCA_009389715.1) (Hazzouri et al. 2019) using the same approach. The second improvement was not as apparent as the first version but did result in a total length of 458 Mb, equivalent to an extra 10% of total scaffold length anchored to chromosomes, compared to the published assembly. The placement of scaffolds in each of the

date palm chromosomes extended their total length except for pd-10 and pd-16 (Figure 19). The assembly improvement for both versions of the date palm genome were further compared using date palm genetic markers (Mathew et al. 2014) located within each assembly. An average correlation of 0.80 was observed for the 18 chromosomes when comparing physical location of the genetic markers between both improved assemblies, ranging from the lowest on pd-16 (r = 0.38) to the highest on pd-4 and pd-15 (equally r = 0.98) (Figure 20).

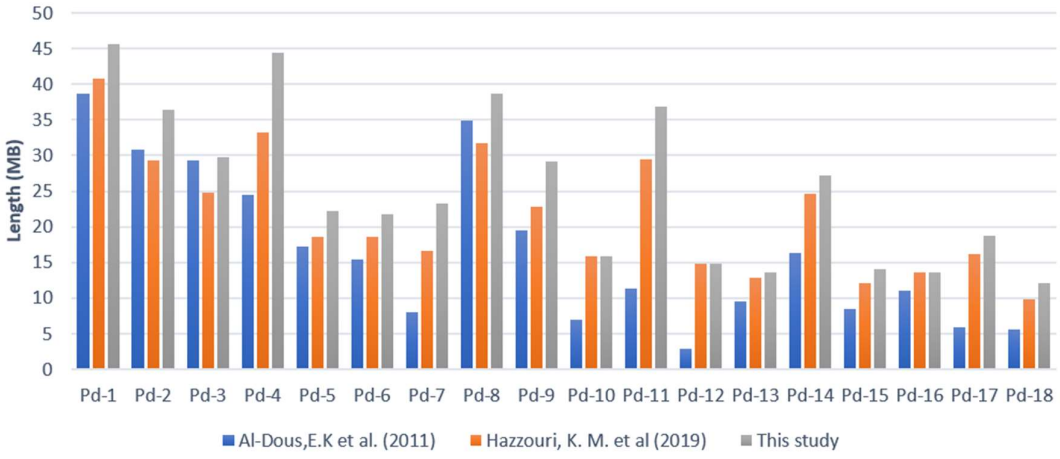


Figure 19. Oil-palm-directed date palm genome assembly improvement by length (Mb) across 18 pseudomolecules. Anchoring of scaffolds was performed on the two published assemblies of date palm genome by Al-Dous et al. (2011) and later by Hazzouri et al. (2019).

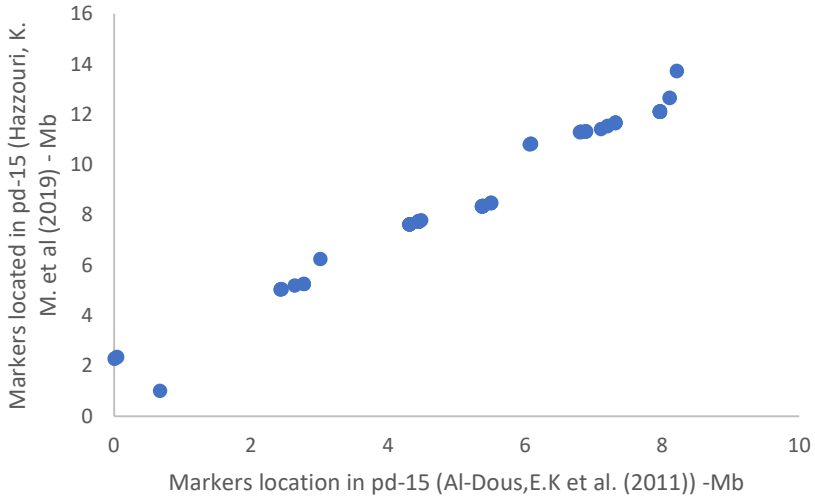


Figure 20. Correlation plots for locations of DNA marker located within date palm chromosome (pd-15) for genome assembly improvement versions by Al-Dous,E.K et al. (2011) and Hazzouri, K. M. et al (2019)

4.5. Conclusions and key recommendations for future work

The improved physical map of oil palm permitted genome comparison within the palm family and revealed high levels of collinearity of genomic sequences, also indicating conserved synteny of location for a range of agronomically important genes that were reported in previous studies. They also share agriculturally desirable traits, such as tall or short varieties in coconut palm, which would be useful in oil palm and date palm. It would be possible to use the causative gene from coconut to screen for allelic differences in the gene orthologues in other species. For instance, the *SHELL* gene in the oil palm genome (eg-2) was found to be highly conserved with the genomic sequence orthologs in coconut (cn-1) and date palm (pd-1). Only the oil palm genome sequence carries sh^{AVROS} variants causing the lysine-to-asparagine amino acid change, which results in disruption of DNA binding within the MADS-box domain, thus no shell is found in this *pisifera* oil palm. Indeed, two more fruit forms including thick-shelled *dura* (sh^{Deli}/sh^{Deli} as wild-type) and thin-shelled *tenera* (sh^{Deli}/sh^{AVROS}) also naturally exist because of haploinsufficiency effect on the SHELL gene (R. Singh, Low, et al. 2013; Teh et al. 2017). The *SHELL* variants are extremely important for oil palm breeding because crossing between *dura* and *pisifera* (usually female sterile) produce *tenera* with fertility restored and 30% more oil than the dura (Corley and Tinker 2015; Hardon, Corley, and Lee 1987). Unlike oil palm, shell-less fruits in coconut and date palm have not yet been reported. The shell layer may be essential to coconut and date palm for maintaining their endosperm moisture content to allow seed germination, especially in arid habitats. Consequently, the mutation probably has been naturally selected out. Nevertheless, we still identified a strong QTL for shell thickness of *tenera* oil palms located within 2.5 Mb to 3.9 Mb sharing the same region with the *SHELL*

gene on eg-2 (Kwong et al. 2016), even when the later was fixed. Hence, the results obtained from oil palm research could lead to the same region in coconut palm being screened for the orthologs responsible for shell variation in coconut within cn-1 and eventually deployed in MAS for selecting thicker- or thinner-shelled coconuts. Coconut shell normally occupies 15% of the total weight of a fruit and approximately 9 million tons of coconut shell is globally produced as an agricultural waste every year (Bellow et al. 2018). Nevertheless, the waste can be turned into activated charcoal and commands a premium (Budi et al. 2018; Mohd Iqbaldin et al. 2013).

Oil palm and coconut are estimated to have diverged from date palm around 62.5 million year ago (He et al. 2015). Fission or fusion of two chromosomes into one or vice versa is known in many lineages (Tang et al. 2008); this event is observed in this study too, whereby date palm has 18 pairs of chromosomes, while oil palm and coconut have only 16 pairs. Torres et al. found that the sex-determination-related genes are now located at 44–45 MB of eg-10 (Torres et al. 2018), where fusion between pd-11 and pd-12 probably occurred, comparative to oil palm. This also suggests that eg-10 contains this cluster of key genes for inflorescence gender and that the split into two chromosomes in the date palm may be an important event leading eventually to a dioecious palm. The finding is supported by a shorter evolutionary distance of *LOG* gene (suppression on female inflorescence) between oil palm and coconut, compared to date palm (Figure 18b). Of the three palms, date palm is the only dioecious palm. The cycle of male and female inflorescences of monoecious oil palm is mainly affected by environmental and genetic factors. Water stress and excessive frond pruning lead to high induction of male inflorescences and abortion of both sexes when the stress becomes extreme (Corley 1973). Understanding the genetic control underlying sex determination in oil palm can provide a new avenue to address abiotic stress tolerance in oil palm. Understanding the basis of inflorescence sex determination may allow breeding for specific environments, to optimize

sex ratio to the target environment, especially in the context of global climate change and of plantations in drought-prone areas. In terms of aptitude for branching, date palm and sago palm can also propagate vegetatively by producing suckers, which would be an extremely valuable trait in oil palm, where this currently does not happen.

Such cases could be useful to infer the location of similar yet undiscovered alleles within these palm species, such as for the presence of the shell layer surrounding palms' kernel, along with genes that differentiate height and sex-determination. These orthologous genes were previously found in unplaced scaffolds of oil palm but have now been assigned and orientated into oil palm's chromosomes. This improved version of the oil palm genome assembly can act as a template for the many of the economically important palm species. This was illustrated through placing an additional 10% of date palm scaffolds into its chromosomes thus improving the physical length of the date palm genome as well. This new oil palm genome assembly allowed one of the orthologous sex-determining regions currently unassigned in the oil palm genome, revealing that chromosome 10 of oil palm is the equivalent chromosome to the date palm gender-linked chromosomes for this monoecious palm species.

Only 19% of the scaffolds were anchored based on a linkage map of the *Khalas* cultivar to the first date palm assembled genome, PDK_30 (Al-Dous et al. 2011; Mathew et al. 2014). Subsequently, the genome has recently been greatly improved to 50% anchored scaffolds of the expected 772.32 Mb genome size using third generation sequencing PacBio long reads by Hazzouri et al. (Hazzouri et al. 2019). In the present study, we have shown that the new oil-palm-guided date palm assembly is still able to anchor a further 10% of scaffolds, and the oil-palm-directed assembly is highly consistent with the latest date palm assembly. Hence, a full utilization of available linkage maps, especially with high mapping density should be incorporated in genome improvement. The oil palm PMv6 genome assembly can serve as a reference to date palm and coconut, and to other palm members in the future. Moreover, based

on the taxonomic distances between oil palm and date palm, we would expect similar results from any palm species with a divergence time less than 62.5 million years (He et al. 2015) ago within *Arecoideae*, a subfamily of *Arecaceae* that consists of 107 species (Barrett et al. 2016; Couvreur, Forest, and Baker 2011), using the oil palm PMv6 genome as a reference. Improved genome sequences can in turn lead to crop improvement strategies for these important food, feed, and fuel sources that will ultimately lead to the higher productivity required for agriculture to meet growing demands in the years to come without the use of more forest land.

Genome comparisons shed light on evolution by understanding conserved synteny whereby set of genes remain in the same relative order across species. Due to different DNA substitution rates and recombination patterns among taxa and species, deviation from collinearity might be a more informative phylogenetic character (Tang et al. 2008). The emergence of new genome sequences helps to explain further the collinearity in evolution by discovering the differential patterns of genes loss or retention between genomes. That is why better resolution of genome or physical map is essential to reveal comparative gene order. Moreover, collinearity in heterochromatic regions in the genome could be disturbed by reshuffling of short DNA segments such as transposable element (TE) causing changes in DNA methylation (Diez, Roessler, and Gaut 2014; Feng and Jacobsen 2011). Tang et al (2008) also highlighted that despite recurring of genome duplications especially in angiosperm species, chromosome numbers did not change as much than their genome sizes. Condensation or fission of two chromosomes into one is known in many lineages; a particular case involved the demonstration that n = 10 (chromosome number) members of the Sorghum genus are ancestral to n = 5 members of the genus. This event is observed in this study too whereby date palm has 18 pairs of chromosomes while oil palm and coconut has only 16 pairs.

Beyond that, to close the genotype to phenotype gap, phenotyping is the next critical focus to be incorporated in plant breeding. The morphology of a plant is the outcome from

interactions between genome with external environment. However, though disruptive technologies are in development in this area, there are still barriers that need to be overcome in order to effectively link phenomics to genomics. One of the barrier is plant epigenomic effects, which are not well understood yet in crops and high throughput methods to accurately assess important phenotypes, to allow both improved selection and decreased selection cycles, to satisfy the breeder equation (Fasoula, Ioannides, and Omirou 2020). In addition, scalability of all these methods is also important to boost the selection intensity in any breeding programs.

Chapter 5: General Conclusion

5.1. Exploring synergies and linkages between and across objectives

The key outcomes for each of the thematic area was depicted in the conceptual framework as below (Figure 21):

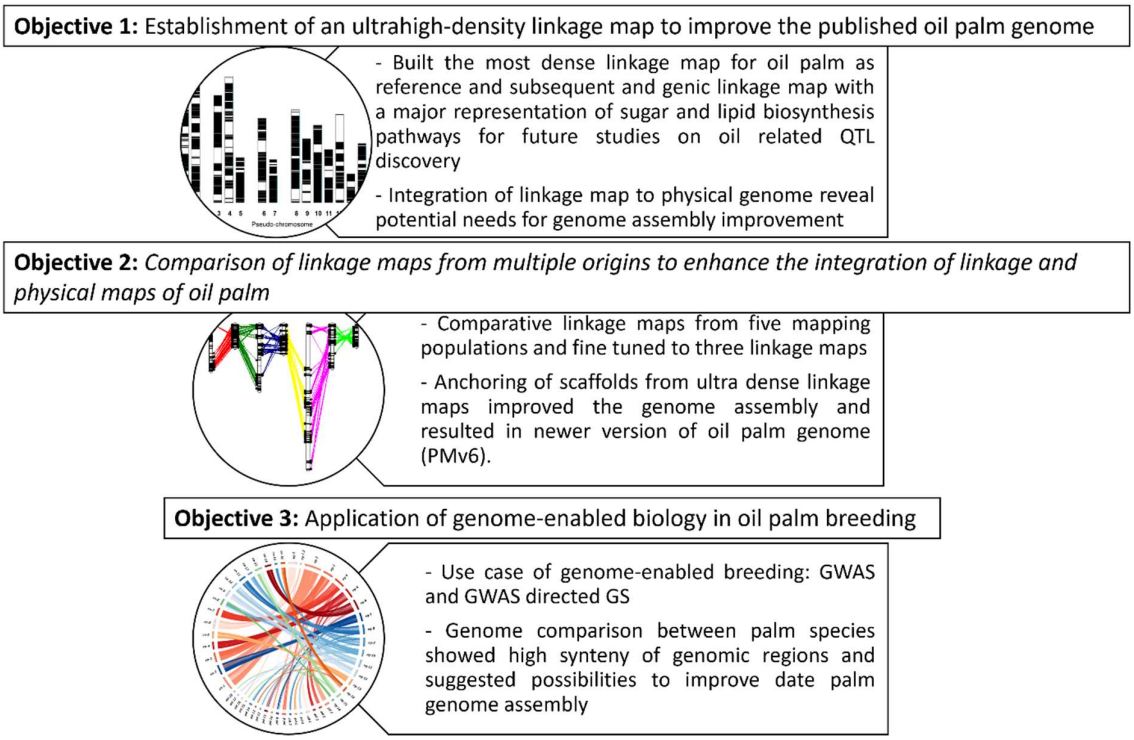


Figure 21. Conceptual framework indicating key outcomes and highlights for each thematic area

5.2. Conclusion

Oil palm is the most productive oil crop with the highest production per growth area. However, the increment in oil yield has stagnated in many countries for a few decades. Genome-enabled breeding should be explored for accurate explore and identify recombinational hotspots within the genome that are potentially associated to phenotypes variation. Knowing the recombination hotspots within a breeding population also allow definition of haplotype blocks that could be used in breeding. Introgression of desired traits between different origins would be possible by

cataloguing the genetic variations of different BPROs as breeders realized the need for new material of sufficient genetic variability for crop improvement (Barcelos et al. 2015).

Improved genome sequences can in turn lead to crop improvement strategies for this important food, feed and fuel source that will ultimately lead to the higher productivity required for agriculture to meet growing demands in the years to come without the use of more forest land. For the case of oil palm, the industry has always been controversial, especially on the topic of deforestation (ASEANPLUS 2020). Part of the reason could be political as oil palm is the most efficient oil crop. Hence sustainable production and practise has been always on the radar for oil palm industry. Genome-enabled breeding will be the hope to ensure continuous production of sustainable palm oil in the market.

5.3. Recommendation for future work

The improved genome assembly still has room for improvement. Combinations of different approaches could cross-validate the assembly, using:

- Karyotying studies for confirmation of the scaffold placement sand locality of markers or genes within the genome. The common application of karyotyping analysis includes fluorescence in situ hybridization (FISH) with probes for 5S and 45S rDNA and telomeric probes or Genomic In Situ Hybridization (GISH) for genomic identification of natural or interspecific hybrids and also useful to analyse chromosomal recombination, both in mitotic and meiotic chromosomes (Castilho, Vershinin, and Heslop-Harrison 2000; Maria, Clyde, and Cheah 1995; de Melo, Souza, and Silva 2017).
- Following date palm genome improvement by generating high contiguity genomic scaffolds, generated from single molecule real time (SMRT) sequencing by Pacific

Biosciences. This approach helps in reducing number of genomic scaffolds and eases the genome assembly (Hazzouri et al. 2019). Optical mapping for chromosomal confirmation for genome scaffolding would be alternative too (Jiao et al. 2017).

References

Abberton, Michael et al. 2016. “Global Agricultural Intensification during Climate Change: A Role for Genomics.” *Plant Biotechnology Journal*.

Al-Dous, Eman K. et al. 2011. “De Novo Genome Sequencing and Comparative Genomics of Date Palm (Phoenix Dactylifera).” *Nature Biotechnology*.

Al-Mssallem, Ibrahim S. et al. 2013. “Genome Sequence of the Date Palm Phoenix Dactylifera L.” *Nature Communications*.

Andrews, Kimberly R et al. 2016. “Harnessing the Power of RADseq for Ecological and Evolutionary Genomics.” *Nat Rev Genet* advance on(2): 81–92.

ASEANPLUS. 2020. “Malaysia’s Sime Darby Seeks Detail after Palm Oil Ban Request.” *The Star News*. https://www.thestar.com.my/aseanplus/aseanplus-news/2020/07/10/malaysia039s-sime-darby-seeks-detail-after-palm-oil-ban-request.

Ayalew, Habtamu et al. 2019. “Comparison of TaqMan, KASP and RhAmp SNP Genotyping Platforms in Hexaploid Wheat.” *PLoS ONE*.

Bai, Bin et al. 2018. “Developing Genome-Wide SNPs and Constructing an Ultrahigh-Density Linkage Map in Oil Palm.” *Scientific Reports* 8(1).

Barcelos, Edson et al. 2015. “Oil Palm Natural Diversity and the Potential for Yield Improvement.” *Frontiers in Plant Science*.

Barrett, Craig F. et al. 2016. “An Introduction to Plant Phylogenomics with a Focus on Palms.” *Botanical Journal of the Linnean Society*.

Bellow, S. A. et al. 2018. “Physical Properties of Coconut Shell Nanoparticles.” *Kathmandu*

University Journal of Science, Engineering and Technology.

Billotte, N. et al. 2005. "Microsatellite-Based High Density Linkage Map in Oil Palm (Elaeis Guineensis Jacq.)." *Theoretical and Applied Genetics* 110(4): 754–65.

BLAST. 2013. "BLAST Basic Local Alignment Search Tool." *Blast Program Selection Guide*: 1–18.

Bohar, Rajaguru, Annapurna Chitkineni, and Rajeev K Varshney. 2020. "Genetic Molecular Markers to Accelerate Genetic Gains in Crops." *BioTechniques*.

Bosch, Elena et al. 2009. "Decay of Linkage Disequilibrium within Genes across HGDP-CEPH Human Samples: Most Population Isolates Do Not Show Increased LD." *BMC Genomics*.

Brar, D. S., and G. S. Khush. 1997. "Alien Introgression in Rice." *Plant Molecular Biology*.

Brondani, Rosana P V, Emlyn R. Williams, Claudio Brondani, and Dario Grattapaglia. 2006. "A Microsatellite-Based Consensus Linkage Map for Species of Eucalyptus and a Novel Set of 230 Microsatellite Markers for the Genus." *BMC Plant Biology* 6.

Budi, E. et al. 2018. "Adsorption and Pore of Physical-Chemical Activated Coconut Shell Charcoal Carbon." In *IOP Conference Series: Materials Science and Engineering*,.

Byerlee, D, W P Falcon, and R L Naylor. 2016. "The Tropical Oil Crop Revolution: Food, Feed, Fuel, and Forests." *The tropical oil crop revolution: food, feed, fuel, and forests*.

Campos-Rivero, Gerardo et al. 2017. "Plant Hormone Signaling in Flowering: An Epigenetic Point of View." *Journal of Plant Physiology*.

Castilho, A., A. Vershinin, and J. S. Heslop-Harrison. 2000. "Repetitive DNA and the Chromosomes the Genome of Oil Palm (Elaeis Guineensis)." *Annals of Botany*.

Chen, Mingsheng et al. 2002. “An Integrated Physical and Genetic Map of the Rice Genome.” *Plant Cell*.

Cheng, Shifeng et al. 2018. “10KP: A Phylodiverse Genome Sequencing Plan.” *GigaScience*.

Cho, Yong Gu et al. 2007. “Identification of Quantitative Trait Loci in Rice for Yield, Yield Components, and Agronomic Traits across Years and Locations.” *Crop Science*.

Christenhusz, Maarten J.M., and James W. Byng. 2016. “The Number of Known Plants Species in the World and Its Annual Increase.” *Phytotaxa*.

Clark, Karen et al. 2016. “GenBank.” *Nucleic Acids Research* 44(D1): D67–72.

Clustalw, Using, Clustalx To, and D O Multiple. 2003. “ClustalW and ClustalX.” *Options*.

Concibido, V. C. et al. 2003. “Introgression of a Quantitative Trait Locus for Yield from Glycine Soja into Commercial Soybean Cultivars.” *Theoretical and Applied Genetics*.

Corley, R H V. 1973. “Oil Palm Physiology-a Review.” In *International Oil Palm ConferenceNovemberKuala Lumpur, Malaysia*,.

Corley, R H V, and P B Tinker. 2015. The Oil Palm: Fifth Edition *The Oil Palm: Fifth Edition*.

Couvreur, Thomas L.P., Félix Forest, and William J. Baker. 2011. “Origin and Global Diversification Patterns of Tropical Rain Forests: Inferences from a Complete Genus-Level Phylogeny of Palms.” *BMC Biology*.

Dalrymple, Brian P. et al. 2007. “Using Comparative Genomics to Reorder the Human Genome Sequence into a Virtual Sheep Genome.” *Genome Biology*.

Deamer, David, Mark Akeson, and Daniel Branton. 2016. “Three Decades of Nanopore Sequencing.” *Nature Biotechnology*.

Department of Statistics Malaysia. 2019. “Selected Agricultural Indicators, Malaysia, 2019.” *Department of Statistics Malaysia*.

Diez, Concepcion M., Kyria Roessler, and Brandon S. Gaut. 2014. “Epigenetics and Plant Genome Evolution.” *Current Opinion in Plant Biology*.

Doganlar, Sami et al. 2002. “A Comparative Genetic Linkage Map of Eggplant (Solanum Melongena) and Its Implications for Genome Evolution in the Solanaceae.” *Genetics* 161(4): 1697–1711.

FAO. 2017. “FAOSTAT.” *Food and Agriculture Organization of the United Nations*.

Fasoula, Dionysia A., Ioannis M. Ioannides, and Michalis Omirou. 2020. “Phenotyping and Plant Breeding: Overcoming the Barriers.” *Frontiers in Plant Science*.

Feng, Suhua, and Steven E. Jacobsen. 2011. “Epigenetic Modifications in Plants: An Evolutionary Perspective.” *Current Opinion in Plant Biology*.

Flint-Garcia, Sherry A., Jeffry M. Thornsberry, and S. Buckler Edward IV. 2003. “Structure of Linkage Disequilibrium in Plants.” *Annual Review of Plant Biology*.

Gogoi, Jugal, Tazid Ali, and Rajeev Sharmah. 2013. “Stochastic Model for Genetic Recombination Because of Crossing over and Chromatid Exchange.” *International Journal of Computer Applications* 80(13).

Gupta, P. K., R. K. Varshney, P. C. Sharma, and B. Ramesh. 1999. “Molecular Markers and Their Applications in Wheat Breeding.” *Plant Breeding*.

Haldane, J B S. 1919. “The Combination of Linkage Values and the Calculation of Distances between the Loci of Linked Factors.” *Journal of Genetics*.

Hardon, J J, R H V Corley, and C H Lee. 1987. *Breeding and Selecting the Oil Palm*.

Improving vegetatively propagated crops. London: Academic Press.

Hartley, C W S. 1967. “The Oil Palm.” *The oil palm.*

Hazzouri, Khaled M. et al. 2019. “Genome-Wide Association Mapping of Date Palm Fruit Traits.” *Nature Communications*.

He, Ziwen et al. 2015. “De Novo Assembly of Coding Sequences of the Mangrove Palm (Nypa Fruticans) Using RNA-Seq and Discovery of Whole-Genome Duplications in the Ancestor of Palms.” *PLoS ONE*.

Huehn, Manfred. 2011. “On the Bias of Recombination Fractions, Kosambis and Haldanes Distances Based on Frequencies of Gametes.” *Genome*.

Hussain, Waseem et al. 2017. “Genotyping-by-Sequencing Derived High-Density Linkage Map and Its Application to QTL Mapping of Flag Leaf Traits in Bread Wheat.” *Scientific Reports*.

Ithnin, Maizura, and Ahmad Kushairi. 2020. *The Oil Palm Genome*. Springer.

Jiang, Lei et al. 2013. “Comparison of the Performance of Two Commercial Genome-Wide Association Study Genotyping Platforms in Han Chinese Samples.” *G3: Genes, Genomes, Genetics*.

Jiao, Wen Biao et al. 2017. “Improving and Correcting the Contiguity of Long-Read Genome Assemblies of Three Plant Species Using Optical Mapping and Chromosome Conformation Capture Data.” *Genome Research*.

Jin, Jingjing et al. 2016. “Draft Genome Sequence of an Elite Dura Palm and Whole-Genome Patterns of DNA Variation in Oil Palm.” *DNA Research*.

Jordan, Katherine W. et al. 2018. “The Genetic Architecture of Genome-Wide

Recombination Rate Variation in Allopolyploid Wheat Revealed by Nested Association Mapping." *Plant Journal*.

Kawakami, Takeshi et al. 2014. "A High-Density Linkage Map Enables a Second-Generation Collared Flycatcher Genome Assembly and Reveals the Patterns of Avian Recombination Rate Variation and Chromosomal Evolution." *Molecular Ecology* 23(16): 4035–58.

Kent, W. James. 2002. "BLAT - The BLAST-like Alignment Tool." *Genome Research* 12(4): 656–64.

Kianian, Penny M.A. et al. 2018. "High-Resolution Crossover Mapping Reveals Similarities and Differences of Male and Female Recombination in Maize." *Nature Communications*.

Kong, Augustine et al. 2002. "A High-Resolution Recombination Map of the Human Genome." *Nature genetics* 31(3): 241–47. http://www.ncbi.nlm.nih.gov/pubmed/12053178.

KOSAMBI, D. D. 1943. "THE ESTIMATION OF MAP DISTANCES FROM RECOMBINATION VALUES." *Annals of Eugenics*.

Krzywinski, Martin et al. 2009. "Circos: An Information Aesthetic for Comparative Genomics." *Genome Research*.

Kurtz, Stefan et al. 2004. "Versatile and Open Software for Comparing Large Genomes." *Genome biology*.

Kwong, Qi Bin et al. 2016. "Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm." *Molecular Plant* 9(8): 1132–41.

Kwong, Qi Bin, Chee Keng Teh, et al. 2017. "Evaluation of Methods and Marker Systems in

Genomic Selection of Oil Palm (Elaeis Guineensis Jacq.)." *BMC Genetics* 18(1).

Kwong, Qi Bin, Ai Ling Ong, et al. 2017. "Genomic Selection in Commercial Perennial Crops: Applicability and Improvement in Oil Palm (Elaeis Guineensis Jacq.)." *Scientific Reports* 7(1).

Larkin, Denis M. et al. 2009. "Breakpoint Regions and Homologous Synteny Blocks in Chromosomes Have Different Evolutionary Histories." *Genome Research*.

Lateef, Djshwar Dhahir. 2015. "DNA Marker Technologies in Plants and Applications for Crop Improvements." *Journal of Biosciences and Medicines*.

Lewin, Harris A., Denis M. Larkin, Joan Pontius, and Stephen J. O'Brien. 2009. "Every Genome Sequence Needs a Good Map." *Genome Research*.

Li, Dandan, T. W. Pfeiffer, and P. L. Cornelius. 2008. "Soybean QTL for Yield and Yield Components Associated with Glycine Soja Alleles." *Crop Science*.

Li, Ye et al. 2020. "Comparative Population Genomics Reveals Genetic Divergence and Selection in Lotus, Nelumbo Nucifera." *BMC Genomics*.

Low, Lloyd, and Martti Tammi. 2017. *Bioinformatics: A Practical Handbook of Next Generation Sequencing and Its Applications*. # N/A. https://www.worldscientific.com/worldscibooks/10.1142/10159#t=aboutBook.

Ma et al. 2007. "ClustalW and ClustalX Version 2.0." *Bioinformatics (Oxford, England)*.

Marçais, Guillaume et al. 2018. "MUMmer4: A Fast and Versatile Genome Alignment System." *PLoS Computational Biology*.

Maria, M., M. M. Clyde, and S. C. Cheah. 1995. "Cytological Analysis of Elaeis Guineensis (Tenera) Chromosomes." *Elaeis*.

Markelz, R. J.Cody et al. 2017. “Using RNA-Seq for Genomic Scaffold Placement, Correcting Assemblies, and Genetic Map Creation in a Common Brassica Rapa Mapping Population.” *G3: Genes, Genomes, Genetics*.

Matasci, Naim et al. 2014. “Data Access for the 1,000 Plants (1KP) Project.” *GigaScience*.

Mathew, Lisa S. et al. 2014. “A First Genetic Map of Date Palm (Phoenix Dactylifera) Reveals Long-Range Genome Structure Conservation in the Palms.” *BMC Genomics*.

Mayer, Klaus F.X. et al. 2012. “A Physical, Genetic and Functional Sequence Assembly of the Barley Genome.” *Nature*.

Mayes, S, P L Jack, R H Corley, and D F Marshall. 1997. “Construction of a RFLP Genetic Linkage Map for Oil Palm (Elaeis Guineensis Jacq.).” *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada* 40(1): 116–22.

de Melo, Cláusio Antônio Ferreira, Margarete Magalhães Souza, and Gonçalo Santos Silva. 2017. “Karyotype Analysis by FISH and GISH Techniques on Artificial Backcrossed Interspecific Hybrids Involving Passiflora Sublanceolata (Killip) MacDougal (Passifloraceae).” *Euphytica*.

Mézard, C. 2006. “Meiotic Recombination Hotspots in Plants.” *Biochemical Society transactions* 34: 531–34.

Milo, R.; Phillips, R. 2016. Garland Science *Cell Biology by the Numbers*. Garland Science.

Mohd Iqbaldin, M. N. et al. 2013. “Properties of Coconut Shell Activated Carbon.” *Journal of Tropical Forest Science*.

Moncada, P. et al. 2001. “Quantitative Trait Loci for Yield and Yield Components in an Oryza Sativa × Oryza Rufipogon BC2F2 Population Evaluated in an Upland Environment.” *Theoretical and Applied Genetics*.

Morton, Newton E. 1991. “Parameters of the Human Genome.” *Proceedings of the National Academy of Sciences of the United States of America*.

Nachman, Michael W. 2002. “Variation in Recombination Rate across the Genome: Evidence and Implications.” *Current Opinion in Genetics and Development* 12(6): 657–63.

Ogata, Hiroyuki et al. 1999. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 27(1): 29–34.

Olson, Maynard, Leroy Hood, Charles Cantor, and David Botstein. 1989. “A Common Language for Physical Mapping of the Human Genome.” *Science*.

Ong, Ai Ling et al. 2019. “Linkage-Based Genome Assembly Improvement of Oil Palm (Elaeis Guineensis).” *Scientific Reports*.

———. 2020. “An Improved Oil Palm Genome Assembly as a Valuable Resource for Crop Improvement and Comparative Genomics in the Arecoideae Subfamily.” *Plants*.

Van Ooijen, J. W. 2011. “Multipoint Maximum Likelihood Mapping in a Full-Sib Family of an Outbreeding Species.” *Genetics Research* 93(5): 343–49.

Painter, Robert H. 1998. “Encyclopedia of Immunology (Second Edition).” In *Encyclopedia of Immunology (Second Edition)*,.

Pellicer, Jaume, and Ilia J Leitch. 2020. “The Plant DNA C-Values Database (Release 7.1): An Updated Online Repository of Plant Genome Size Data for Comparative Studies.” *New Phytologist* 226(2): 301–5.

Pollack, Jonathan R., and Vishwanath R. Iyer. 2002. “Characterizing the Physical Genome.” *Nature Genetics*.

Pollard, Martin O et al. 2018. “Long Reads: Their Purpose and Place.” *Human Molecular*

Genetics.

Pootakham, Wirulda 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(5–6): 288–95.

Prance, Ghillean T. et al. 1987. "Gene Banks and the World's Food." *Brittonia*.

Purcell, Shaun et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81(3): 559–75. http://linkinghub.elsevier.com/retrieve/pii/S0002929707613524.

Puritz, Jonathan B. et al. 2014. "Demystifying the RAD Fad." *Molecular Ecology*.

Rafalski, Antoni. 2002. "Applications of Single Nucleotide Polymorphisms in Crop Genetics." *Current Opinion in Plant Biology*.

Rajanaidu, N et al. 2005. "Oil Palm Planting Materials: Current Developments and Competetiveness with Other Oil Bearing Crops." In *Proceedings of the 2005 International Palm Oil Congress: Technological Breakthroughs and Commercialization—The Way Forward*, , 25–29.

Rambaut, Andrew. 2018. "FigTree v. 1.4.4." *http://tree.bio.ed.ac.uk/software/figtree/*.

Rance, K. a. et al. 2001. "Quantitative Trait Loci for Yield Components in Oil Palm (Elaeis Guineensis Jacq.)." *Theor Appl Genet*.

Rasheed, Awais et al. 2017. "Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives." *Molecular Plant*.

Rastas, Pasi et al. 2013. "Lep-MAP: Fast and Accurate Linkage Map Construction for Large SNP Datasets." *Bioinformatics* 29(24): 3128–34.

———. 2017. “Lep-MAP3: Robust Linkage Mapping Even for Low-Coverage Whole Genome Sequencing Data.” *Bioinformatics* 33(23): 3726–32.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics and Bioinformatics*.

Rival, A. et al. 1997. “Comparative Flow Cytometric Estimation of Nuclear DNA Content in Oil Palm (Elaeis Guineensis Jacq) Tissue Cultures and Seed-Derived Plants.” *Plant Cell Reports*.

Rosenquist, E A. 1985. *The Genetic Base of Oil Palm Breeding Populations*.

Scheben, Armin, Jacqueline Batley, and David Edwards. 2017. “Genotyping-by-Sequencing Approaches to Characterize Crop Genomes: Choosing the Right Tool for the Right Application.” *Plant Biotechnology Journal*.

Sebolt, A. M., R. C. Shoemaker, and B. W. Diers. 2000. “Analysis of a Quantitative Trait Locus Allele from Wild Soybean That Increases Seed Protein Concentration in Soybean.” *Crop Science*.

Seng, Tzer Ying et al. 2011. “Genetic Linkage Map of a High Yielding FELDA Deli×Yangambi Oil Palm Cross.” *PLoS ONE* 6(11).

Shen, Chao, Ximei Li, Ruiting Zhang, and Zhongxu Lin. 2017. “Genome-Wide Recombination Rate Variation in a Recombination Map of Cotton.” *PLoS ONE* 12(11).

Sidhu, Deepak, and Kulvinder S. Gill. 2005. “Distribution of Genes and Recombination in Wheat and Other Eukaryotes.” *Plant Cell, Tissue and Organ Culture*.

Sime Darby Plantation. 2014. Fact sheet - Sime Darby Plantation: Palm Oil Facts & Figures 2014 *Palm Oil Facts & Figures*.

Singh, R. J., and T. Hymowitz. 1999. “Soybean Genetic Resources and Crop Improvement.” In *Genome*,.

Singh, Rajinder, Meilina Ong-Abdullah, et al. 2013. “Oil Palm Genome Sequence Reveals Divergence of Interfertile Species in Old and New Worlds.” *Nature* 500(7462): 335–39. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3929164/.

Singh, Rajinder, Eng Ti Leslie Low, et al. 2013. “The Oil Palm SHELL Gene Controls Oil Yield and Encodes a Homologue of SEEDSTICK.” *Nature*.

Soh, Aik Chin, Sean Mayes, and Jeremy Roberts. 2017. Oil Palm Breeding: Genetics and Genomics *Oil Palm Breeding: Genetics and Genomics*.

Staub, Jack E., Felix C. Serquen, and Manju Gupta. 1996. “Genetic Markers, Map Construction, and Their Application in Plant Breeding.” *HortScience* 31(5): 729–40.

Stich, Benjamin, Claude Urbany, Per Hoffmann, and Christiane Gebhardt. 2013. “Population Structure and Linkage Disequilibrium in Diploid and Tetraploid Potato Revealed by Genome-Wide High-Density Genotyping Using the SolCAP SNP Array.” *Plant Breeding*.

Sunderland, T C H. 2011. “Food Security: Why Is Biodiversity Important? Sécurité Des Aliments: Pourquoi La Biodiverstiré Est-Elle Importante? Seguridad Alimentaria: ¿por Qué Es Importante La Biodiversidad?” *International Forestry Review*.

Taeprayoon, Puntaree et al. 2016. “Genome-Wide SSR Marker Development in Oil Palm by Illumina HiSeq for Parental Selection.” *Plant Genetic Resources: Characterisation and Utilisation*.

Tang, Haibao et al. 2008. “Synteny and Collinearity in Plant Genomes.” *Science*.

———. 2015. “ALLMAPS: Robust Scaffold Ordering Based on Multiple Maps.” *Genome*

Biology.

Teh, Chee Keng et al. 2016. "Genome-Wide Association Study Identifies Three Key Loci for High Mesocarp Oil Content in Perennial Crop Oil Palm." *Scientific Reports* 6.

———. 2017. "Characterizing Haploinsufficiency of SHELL Gene to Improve Fruit Form Prediction in Introgressive Hybrids of Oil Palm." *Scientific Reports*.

———. 2020. "Major Qtls for Trunk Height and Correlated Agronomic Traits Provide Insights into Multiple Trait Integration in Oil Palm Breeding." *Genes*.

The UniProt Consortium. 2015. "UniProt: A Hub for Protein Information." *Nucleic acids research* 43(Database issue): D204-12. http://nar.oxfordjournals.org/cgi/content/long/43/D1/D204.

Ting, Ngoot Chin et al. 2014. "High Density SNP and SSR-Based Genetic Maps of Two Independent Oil Palm Hybrids." *BMC Genomics* 15(1).

Torres, Maria F. et al. 2018. "Genus-Wide Sequencing Supports a Two-Locus Model for Sex-Determination in Phoenix." *Nature Communications*.

Troggio, M. et al. 2012. "Apple, from Genome to Breeding." *Tree Genetics and Genomes*.

Turnbull, Clare et al. 2018. "The 100 000 Genomes Project: Bringing Whole Genome Sequencing to the NHS." *BMJ (Online)*.

Veeckman, Elisabeth, Tom Ruttink, and Klaas Vandepoele. 2016. "Are We There yet? Reliably Estimating the Completeness of Plant Genome Sequences." *Plant Cell*.

Vinod, K. K. 2011. "Kosambi and the Genetic Mapping Function." *Resonance*.

Voorrips, R. E. 2002. "MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs." *Journal of Heredity* 93(1): 77–78. https://academic.oup.com/jhered/article-

lookup/doi/10.1093/jhered/93.1.77.

Wei, Fusheng et al. 2009. “The Physical and Genetic Framework of the Maize B73 Genome.” *PLoS Genetics*.

Wold Bank Group. 2019. “Agricultural Transformation and Inclusive Growth: The Malaysian Experience.” *Global Knowledge and Research Hub in Malaysia*.

Wong, Gane Ka Shu et al. 2020. “Sequencing and Analyzing the Transcriptomes of a Thousand Species Across the Tree of Life for Green Plants.” *Annual Review of Plant Biology*.

Wong, Y.C. et al. 2017. “Differential Gene Expression at Different Stages of Mesocarp Development in High- and Low-Yielding Oil Palm.” *BMC Genomics*.

Wong, Yick et al. 2014. “Expression Comparison of Oil Biosynthesis Genes in Oil Palm Mesocarp Tissue Using Custom Array.” *Microarrays*.

Wu, Yonghui, Prasanna R. Bhat, Timothy J. Close, and Stefano Lonardi. 2008. “Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph.” *PLoS Genetics*.

Xiao, Yong et al. 2017. “The Genome Draft of Coconut (Cocos Nucifera).” *GigaScience*.

Xu, Zongli, Fei Zou, and Todd J. Vision. 2005. “Improving Quantitative Trait Loci Mapping Resolution in Experimental Crosses by the Use of Genotypically Selected Samples.” *Genetics*.

Zamir, Dani. 2001. “Improving Plant Breeding with Exotic Genetic Libraries.” *Nature Reviews Genetics*.

Zulkifli, Y. et al. 2017. “Designing the Oil Palm of the Future.” *Journal of Oil Palm*

Research.

Appendices

Appendix A: Publication 1 entitled

"Linkage-based genome assembly improvement of oil palm *(Elaeis guineensis)*

Appendix B: Publication 2 entitled

"An Improved Oil Palm Genome Assembly as a Valuable Resource for Crop Improvement and Comparative Genomics in the *Arecoideae* Subfamily"

SCIENTIFIC REPORTS

OPEN

Linkage-based genome assembly improvement of oil palm (*Elaeis guineensis*)

Received: 8 October 2018
Accepted: 12 April 2019
Published online: 29 April 2019

Ai-Ling Ong[1,2], Chee-Keng Teh[1,2], Qi-Bin Kwong[1], Praveena Tangaya[1], David Ross Appleton[1], Festo Massawe[2] & Sean Mayes[3]

Meiotic crossovers in outbred species, such as oil palm (*Elaeis guineensis* Jacq., 2n = 32) contribute to allelic re-assortment in the genome. Such genetic variation is usually exploited in breeding to combine positive alleles for trait superiority. A good quality reference genome is essential for identifying the genetic factors underlying traits of interest through linkage or association studies. At the moment, an AVROS *pisifera* genome is publicly available for oil palm. Distribution and frequency of crossovers throughout chromosomes in different origins of oil palm are still unclear. Hence, an ultrahigh-density genomic linkage map of a commercial Deli *dura* x AVROS *pisifera* family was constructed using the OP200K SNP array, to evaluate the genetic alignment with the genome assembly. A total of 27,890 linked SNP markers generated a total map length of 1,151.7 cM and an average mapping interval of 0.04 cM. Nineteen linkage groups represented 16 pseudo-chromosomes of oil palm, with 61.7% of the mapped SNPs present in the published genome. Meanwhile, the physical map was also successfully extended from 658 Mb to 969 Mb by assigning unplaced scaffolds to the pseudo-chromosomes. A genic linkage map with major representation of sugar and lipid biosynthesis pathways was subsequently built for future studies on oil related quantitative trait loci (QTL). This study improves the current physical genome of the commercial oil palm, and provides important insights into its recombination landscape, eventually unlocking the full potential genome sequence-enabled biology for oil palm.

Oil palm (*Elaeis guineensis Jacq.*) is the only major crop member in the *Arecaceae* family producing edible oil[1]. Being an efficient oil crop, the average oil yield of oil palm is 4 metric tons per hectare every year, which is approximately 10 times higher than soy (*Glycine max.*)[2,3]. Palm oil is currently the most traded vegetable oil, supplying one third of global vegetative oil[3,4] on less than 6% of the world's agricultural land dedicated to oil crops. Despite yield superiority, most of the commercially grown oil palm materials have undergone only a limited number of generations of selection since 1920's due to long selection cycles (12–19 years). In recent decades, the oil palm industry has put major effort into genomics and DNA marker research, primarily to hasten breeding progress. To achieve this, marker-assisted selection (MAS) has been introduced into breeding programmes to allow estimation of the breeding values in palms as early as the nursery stage[5,6].

Marker-assisted selection usually requires prior knowledge on the distribution of quantitative trait loci (QTL) for a targeted trait in the genome. Over 20 years, linkage maps from various backgrounds of oil palm have been constructed. The first restriction fragment length polymorphism (RFLP)-based linkage map for oil palm was reported in a self-pollinated palm from the *Binga* origin in Africa[7]. Similar efforts were then continued mainly using microsatellites for higher throughput and reproducibility, followed by QTL localization onto these maps[8–10]. The family-based mapping approach was found to be useful for detecting major QTLs, however QTLs were not consistent across populations (with a few exceptions). Consensus maps of one or a number of progenies, using single nucleotide polymorphism (SNP) markers have been constructed to maximize marker density[11,12].

Genetic maps reflect the actual inheritance of loci into their offspring based on the patterns of recombination during meiosis. Recombination is generally modelled to be a Poisson distribution, with individual chromosomes in each individual gamete expected to undergo at least one recombination event (as this is needed for correct disjunction), but generally no more than three. These are then converted to centiMorgans (cM) through the use

[1]Biotechnology & Breeding Department, Sime Darby Plantation R&D Centre, Serdang, Selangor, Malaysia. [2]School of Biosciences, University of Nottingham Malaysia, Semenyih, Selangor, Malaysia. [3]School of Biosciences, University of Nottingham, Leicestershire, UK. Correspondence and requests for materials should be addressed to A.-L.O. (email: ong.ailing.sdtc@simedarbyplantation.com)

of mapping functions, such as the Haldane or Kosambi[13–15]. This limitation in the extent of recombination in gametes also restricts the rate at which alleles close together on the same chromosome are separated in the offspring. Recombination is also not randomly distributed across the genome, with larger genome species often having high levels of recombination at the ends of the chromosome, but little recombination in the heterochromatic and centromeric regions of the chromosome, despite genes still being present (e.g. in wheat[16,17]). Recombination within an oil palm progeny limits the potential combination of alleles that can be achieved by breeding and, particularly, the number of offspring that need to be generated to be able to identify those with the desired combinations.

Conventional oil palm breeding is often based on using test-crosses to evaluate the breeding potential of a parent. To assess parents and develop General Combining Ability (GCA; additive effects) and Specific Combining Ability (SCA; non-additive effects) for parental selection, the number of crosses assessed is more important than the absolute number of test-cross palms recorded (Corley and Tinker, 2015). In practice, 16 to 96 palms per cross in a trial are planted. This limited number of individuals from a single cross has focused attention on genome-wide association studies (GWAS). Teh and colleagues evaluated two multi-parent populations (2,045 *tenera* palms) for mesocarp oil yield[18] using a high-density OP200K SNP array[19]. The GWAS was based on the published physical genome of a *pisifera* palm of *E. guineensis*[20]. Only 60% (55,054) of the SNPs informative in the cross were mapped onto the reference genome. Furthermore, the size of the pseudo-molecule physical map is merely one third (657.96 Mb) of the full expected 1.8-Gb genome size and even the total scaffold lengths only added up to 1.53 Gb, which is still incomplete. Hence, it is critical to further improve the reference genome to unlock the full potential of genome-enabled biology for oil palm. Linkage maps can contribute to the placement of scaffolds into the pseudo-chromosomes. The maps also give important information on the relative recombination distance between loci from different pools of germplasm. In combination with a physical genome sequence, they can indicate the presence of inversions, translocations and duplications which may influence how effectively important gene combinations can be brought together through crossing and marker selection.

For the oil palm industry, the commercial planting materials are dominated by high-yielding *tenera* palms, which are derived from crosses between Deli *dura* and AVROS *pisifera*. The paternal *pisifera* is more heterozygous due to lower levels of inbreeding (although potentially from an extremely narrow genetic base) while the maternal *dura* is normally the descendent from generations of self-pollination or sib-mating from a limited gene pool[21]. Genetic differences, such as large scale structure variations among the genomes of *dura*, *tenera* and *pisifera* can confound QTL results, predicted genetic interactions and breeding objectives, resulting in inconsistencies of QTL positions between different origins. Currently, the extent of such variation in oil palm is still unclear.

In this study, we aimed (1) to construct an ultrahigh-density reference linkage map for genome sequence improvement using the largest commercial *tenera* population reported to date; (2) to integrate linkage maps of both parents and their progeny to the physical genome; (3) to map a genic SNP subset representing genes in important biochemical pathways for future QTL analyses, especially oil yield traits.

Materials and Methods

Mapping population and DNA preparation. A total of 295 F_1 *tenera* palms derived from a maternal Deli *dura* and paternal AVROS *pisifera* (278 x TT41/4) family were selected as a mapping population. The same origin (Deli *dura* x AVROS *pisifera*) was represented in the pool of 59 accessions used for array design. Fresh tissue was sampled from the third leaflet of each seedling. The population is currently at juvenile stage and maintained at Sime Darby Plantation, Malaysia. Total genomic DNA was isolated from 0.1 g of the leaf tissue using the DNAeasy Plant Mini Kit (Qiagen, Germany). The DNA quality was then quantified on a 0.8% agarose gel using known standards.

SNP identification and genotyping. The mapping population (genomic DNA = 25 ng/μl) was genotyped using the 200,000 SNPs accommodated on the published OP200K Infinium array (Illumina, USA) on the Infinium iScan platform according to the manufacturer's recommendations[19]. The raw intensity SNP data were analysed with GenomeStudio version 2001.1 with genotyping module version 1.8.4. Using a GenCall score cut-off of 0.15, auto-clustering of the SNPs was done. The SNP clustering was done and was manually confirmed by visual inspection. The SNP call were exported into the PLINK program for minor allelic frequency (MAF) and call-rate filtering[22].

Mining of genic SNP markers and annotations. All oil palm transcriptome sequences available in GenBank[23] were downloaded and combined with an in-house developed transcriptome database[24,25]. The transcriptome dataset was then aligned to the published oil palm genome[20] using BLAT[26] to identify gene structures. The SNP markers which were found in the published genome, residing within an interval between the first exon and the last exon of the gene structure were marked as 'genic'. This step was done using an in-house Perl script. The SNP-tagged transcripts were annotated by referring to KEGG[27] and UniProt[28] databases using BLASTN[29] sequence homology search. This allowed us not only to identify SNPs within genes with a potential influence on important traits, but also to construct a reference genic SNP-based linkage map in oil palm.

Construction of genomic, genic SNP-based linkage maps and integration with physical map. Polymorphic SNPs were identified where at least one of the parents was heterozygous. Subsequently, polymorphic SNPs were discarded if call rate <95% or if they had high segregation distortion using False Discovery Rate (FDR) corrected method with cutoff at p-value > 0.05. Next, linkage maps were constructed using Lep-MAP3[30,31] for large SNP datasets, which was especially developed for outbred families like oil palm with an underlying maximum likelihood (ML) method. *Filtering modules* was used for marker quality checking, followed by running the *Separate Chromosomes* module for binning the assayed markers into linkage groups with optimised LOD values ranging from 4 to 30. The *Join Singles* module assigned singular markers to existing

Linkage group	Linkage Length (cM)	No. of linked SNPs	No. of SNPs mapping in published genome	Chromosome size (Mb)	No. of selected genic SNPs	Average Recombination rate (cM/Mb)
1	122.06	2422	1803 (74.4%)	68.44	215	1.8
2	119.96	3533	1535 (43.4%)	65.56	194	2.0
2.2	9.69	248	199 (80.2%)			
3	103.24	2582	1959 (75.9%)	60.06	260	1.7
4	121.09	2235	1203 (53.8%)	57.25	134	2.1
5	33.04	1281	659 (51.4%)	51.96	84	0.8
5.2	7.83	124	75 (60.5%)		18	
6	81.23	1752	1079 (61.6%)	44.36	117	1.8
7	25.67	889	758 (85.3%)	43.46	120	0.6
7.2	12.63	217	198 (91.2%)			
8	89.42	2512	1312 (52.2%)	40.20	178	2.2
9	61.39	1723	889 (51.6%)	38.06	120	1.6
10	74.41	2912	1675 (57.5%)	31.90	310	2.3
11	50.81	1012	783 (77.4%)	30.07	98	1.7
12	84.92	1312	754 (57.5%)	28.80	123	2.9
13	27.37	620	423 (68.2%)	27.81	41	1.0
14	56.76	1553	1037 (66.8%)	24.38	165	2.3
15	42.48	586	541 (92.3%)	24.31	88	1.7
16	27.70	377	328 (87.0%)	21.37	49	1.3
Total	**1151.70**	**27890**	**17210 (61.7%)**	**657.99**	**2314**	
Mean	**60.62**					**1.8**

Table 1. A summary of the Deli *dura* x AVROS *pisifera* linkage map using genomic SNP markers, the selected genic SNP markers and estimated recombination rate. cM – centiMorgan, the percentage of 200,000 SNP markers mapped in the reference genome in parentheses. *The linkage group number follows chromosome number of the published genome[20].

linkage groups to maximize the map-abilities of the total input marker group. Lastly, the *Order Markers* module ordered the binned markers according to the Kosambi mapping function for conversion of recombination frequencies into map distances (in unit centiMorgan, cM). The option "sexAveraged = 1" was selected along with *Order Markers* to join the maps of both parents. The polymorphic genic SNP subset was subsequently loaded into JoinMap 5[32] for linkage mapping to compare the performance of different mapping tools as well as to validate the mapping results. The parameters were set to default, the mapping function used here was ML for estimations of marker distances (in unit cM). All linkage maps were visualised using an R script with minor modifications[33] and Mapchart[34].

Common SNP markers were used to carry out a comparative analysis between the linkage maps in cM distance (*dura* parent, *pisifera* parent and *tenera* progeny) and the physical map in Mb distance, with each chromosome visualised in plots. The scatter plots to visualise the integration of the genomic SNP-based genetic map and the physical map was plotted using an in-house R script. Each chromosome was plotted separately to visualise the comparisons between both parental and progeny (combined) maps against the genome assembly.

Improvement of genome coverage. SNP markers from OP200K were mapped against the published oil palm assembled sequences, available at NCBI under BioProject accessions https://www.ncbi.nlm.nih.gov/bioproject/PRJNA434010 and the assembly at https://www.ncbi.nlm.nih.gov/assembly/GCA_000442705, for the whole genome shotgun sequencing project with all scaffolds including the unplaced scaffolds on the physical map. The progeny (combined) linkage map was applied to bin and anchor the unplaced scaffolds to the physical map. The improvement of genome coverage was quantified based on the sum of the newly anchored scaffolds and their total length in unit of Mb for each of the pseudo-chromosomes.

Results

Identification of genomic and genic SNPs. A total of 34,315 genomic SNPs was identified to be polymorphic in the mapping population, and the number was reduced to 29,057 SNPs after removing the segregation-distorted markers or those with call rate <95%. For linkage analysis, 27,890 SNPs and the subset of 2,314 genic markers were mapped, respectively (Table 1).

Construction of genomic and genic SNP-based linkage maps. Two linkage maps of Deli *dura* x AVROS *pisifera* were constructed using different set of SNPs. The first set consisted of 27,890 genomic SNPs that were successfully assigned to the 19 linkage groups (LGs; representing the 16 chromosomes of oil palm), spanning 1,151.7 cM with an average mapping interval of 0.04 cM (Table 1). Out of 27,890 genomic SNPs, 17,210 SNPs (61.7%) were found in the published genome. The LG number presented in Fig. 1 is ordered by chromosome number for better traceability. The oil palm consists of a diploid genome with 16 chromosome pairs. In the linkage map, we observed three chromosomes (2, 5 and 7) fragmented into two LGs, respectively. By excluding the three

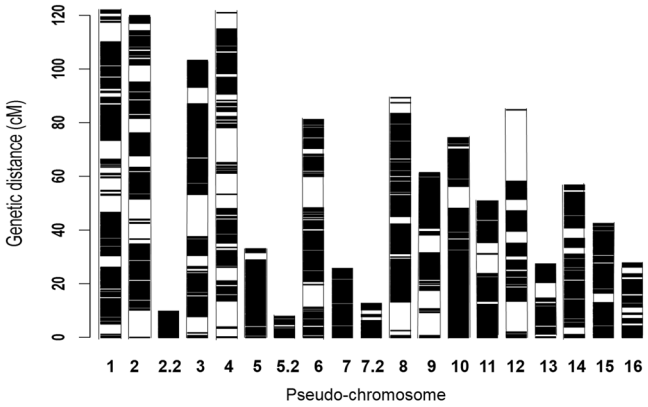


Figure 1. An ultrahigh-density Deli *dura* x AVROS *pisifera* linkage map with 27,890 SNP markers. The left scale represents the linkage length in unit centiMorgan (cM), whereas the number below each bar represents the linkage group number/pseudo-chromosome.

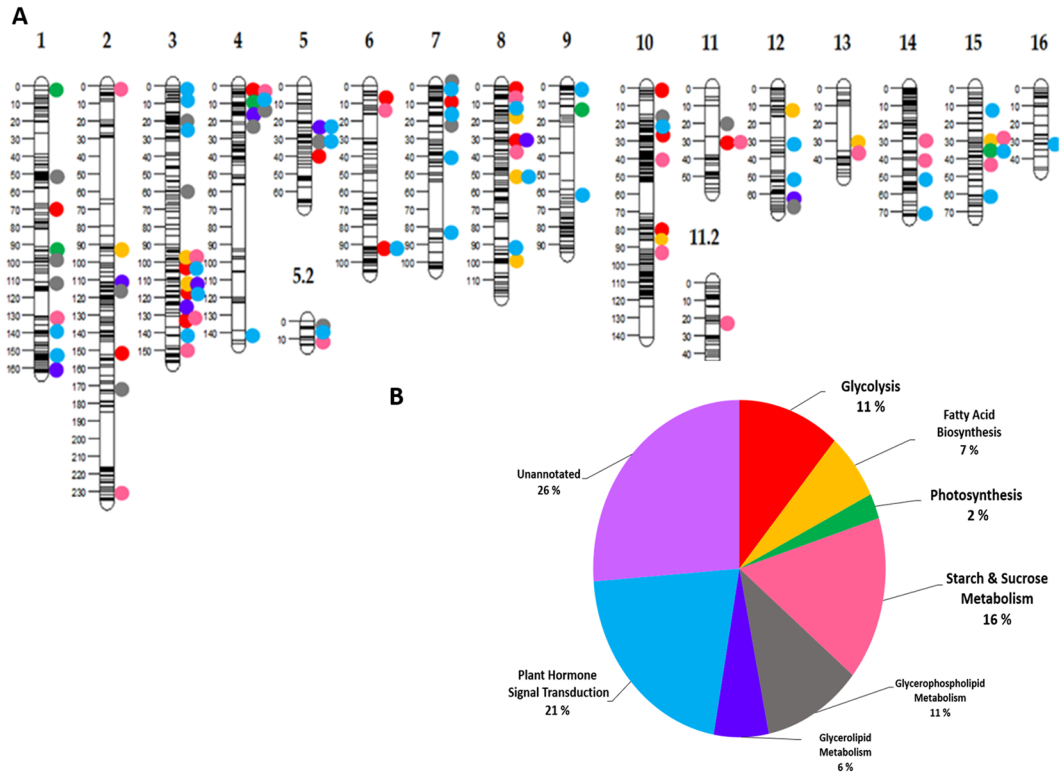


Figure 2. (**A**) A high-density genic SNP-based linkage map of the Deli *dura* x AVROS *pisifera* family with 2,314 SNP markers, (**B**) A pie chart depicting the distribution of the top-ranked pathway membership for the annotated genic SNP markers.

fragments, the longest LG 1 (122.06 cM) and the shortest LG 13 (27.37 cM) which also reflected the similar ranking in terms of their chromosome sizes. On average, 60.62 cM per LG was observed. The ultrahigh-density linkage map is given in Fig. 1. The largest interval between markers was 26.68 cM and located on LG 12.

A set of 2,314 genic markers were selected from the total 27,890 linked SNPs to construct another map with lower complexity, based on genic SNP markers alone, designed for future QTL analyses. The map spanned 1,733 cM with an average mapping interval of 0.75 cM (Supplementary Table 1). The obtained 18 LGs was still higher than the expected 16 LGs for the oil palm genome (Fig. 2A). Two LGs (5 and 11) were found to be fragmented in this map. In Fig. 2B, the annotation results indicate that the genic markers were mainly represented by seven pathways based the annotated genes/pathways, including (1) plant hormone signal transduction (25%), (2) starch and sucrose metabolism (19%), (3) glycolysis (13%), (4) glycerolphospholipid metabolism (13%), (5) fatty acid biosynthesis (8%), (6) glycerolipid metabolism (7%), (7) photosynthesis (3%). The pathway annotations

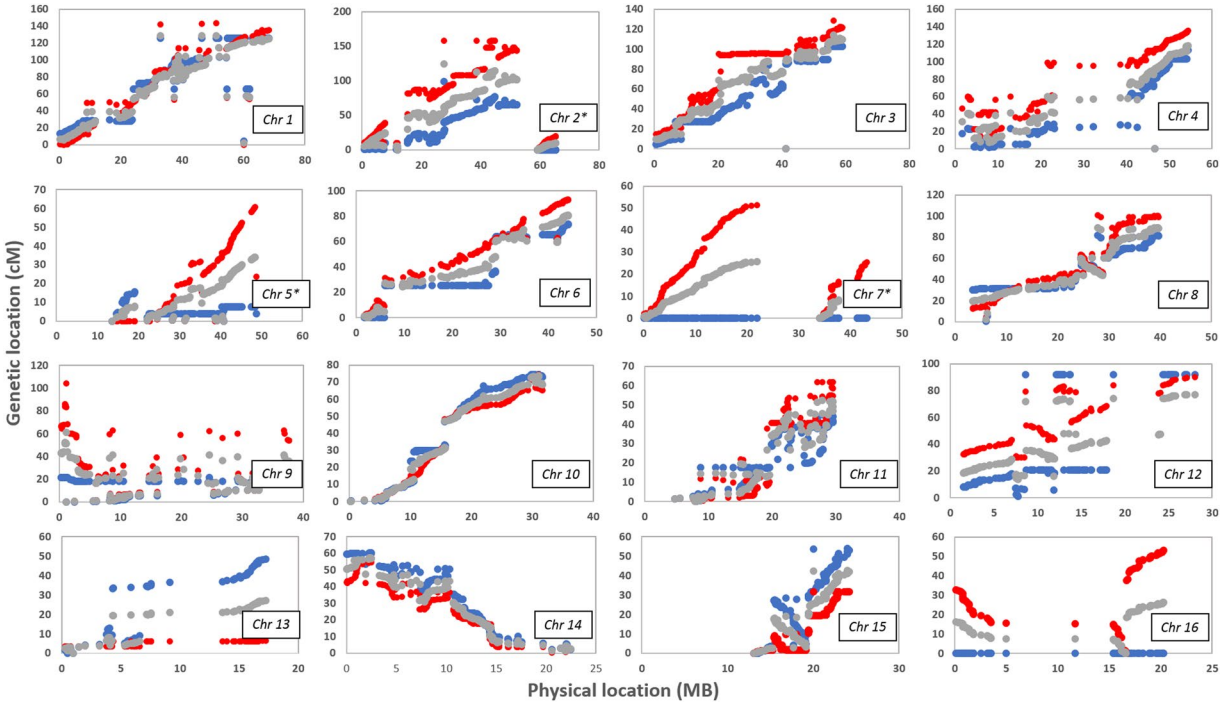


Figure 3. Scatter plots of integration between genetic linkage order and physical marker order throughout the oil palm genome. Blue dots represent the *dura* genetic map and red dots represent the *pisifera* genetic map. Grey dot represents the *tenera* combined map. (*Fragmented linkage groups with shorter length, including LG 2.2, LG 5.2 and LG 7.2 were assigned to Chromosome 2, Chromosome 5 and Chromosome 7, respectively).

were superimposed on the linkage map (Fig. 2A). The genomic SNPs and the subset of genic markers were used to construct maps using Lep-MAP3 and JoinMap 5, respectively. The common SNP markers were compared, showing high concordance of marker ordering (Supplementary Fig. 1).

The integration between the linkage maps and the physical map. The genomic linkage distances of Deli *dura* (Parent 1), AVROS *pisifera* (Parent 2), and their *tenera* progeny were compared to the physical distances. The comparison results across 16 chromosomes are illustrated in Fig. 3. In general, the common SNPs among *dura*, *pisifera* and *tenera* shared the same linkage order due to parallel marker distributions against their physical positions. Lower recombination rates were observed on nine chromosomes for the *dura* parent, compared to the *pisifera* especially on Chromosome 7 and Chromosome 16. Higher rates of recombination for *dura* were only observed on Chromosomes 13 and 15. The combination of the *dura* and *pisifera* linkages map is through combining pairwise distances, so the recombination rate of the *tenera* is intermediate to the parental values. Interestingly, recombination on Chromosomes 7 and 16 of the *tenera* was due to one-sided inheritance of genetic polymorphism from the *pisifera* parent, alone. The order obtained between genetic and physical maps were generally good, except for Chromosomes 9, 11 and 12. Some evidence is seen for potential inversions of chromosome segments between the linkage population and the reference genome, such as on Chromosome 15 (15–20 Mb) and Chromosome 16 (0–4 Mb). Overall, the average recombination rate throughout the oil palm genome was estimated as 1.8 cM/Mb, ranging from 0.8 (Chromosome 5) to 2.9 cM/Mb (Chromosome 12) (Table 1).

Improvement of genome coverage. The average number of scaffolds contributing to genome sequences by Singh *et al.* (2013) are 19 scaffolds in each pseudo-chromosome. The ultrahigh-density linkage map identified a total of 311 Mb in scaffolds in addition to the existing genome assembly that could be added to the linkage groups. In total, 1,323 additional scaffolds were placed into pseudo-chromosome locations (Fig. 4A). Figure 4B depicted the total length for each pseudo-chromosome from published genome sequences. Each of the pseudo-chromosome was extended from this study with the most significant extension in Chromosome 2 at 56 Mb, while Chromosome 15 and 16 also have minimal extensions of less than 1 Mb, hence the improvements were not distinguishable in the plot (Fig. 4B). After scaffolds addition, the physical map was successfully extended from 658 Mb to 969 Mb, which is equivalent to a 47% improvement. Still, the improved physical map only covers 54% of the total genome based on the estimated genome size of oil palm of 1.8 Gb.

Discussion

The published OP200K Infinium array was derived from 59 important and diverse origins[19] and was used in this study. This helps to explain why only 13.9% (27,890) of the assayed SNPs were polymorphic and could be mapped in the advanced commercial Deli *dura* x AVROS *pisifera* cross. Still, the average mapping interval of 0.04 cM is the highest reported, compared to previous studies (1.26 cM[35], 1.40 cM[10] and 0.29 cM[36]). In fact, the improvement of the mapping interval was mainly due to higher recombination detected in this larger mapping population, which

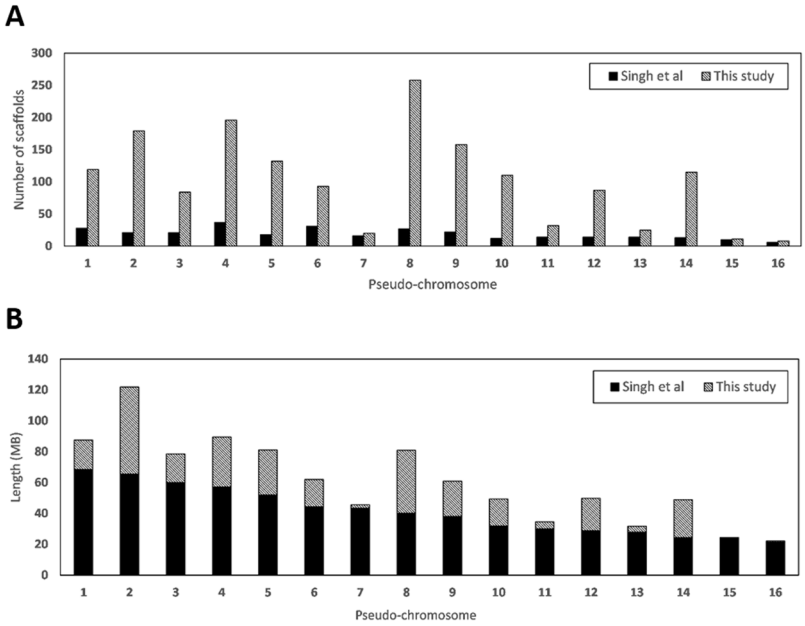


Figure 4. (**A**) Comparison of number of scaffolds from previous physical map[20] and the improved physical map developed in this study. (**B**) Total length improvement (Mb) for each of the pseudo-chromosome after addition of new scaffolds to the published genome sequence in this study.

was double the size of the 153 *tenera* palms used by Bin Bai and colleagues. So far, only Lep-MAP and MSTmap[37] programmes are able to map more than 10,000 SNP markers in under an hour. However, the MSTmap is limited to single families and phased data (including BC1, RIL, DH and Hap populations). The Lep-MAP does not have such limitations, so it was chosen to analyse the unphased cross-pollinating (CP) populations of oil palm. By using Lep-MAP3, the total length of the genomic SNP-based linkage map was significantly reduced from 2,938.2 cM obtained using Lep-MAP2[36] to 1,151.7 cM. With 16 chromosomes, a combined linkage distance in the order of 1,600 cM would be expected. Lep-MAP3 is able to minimise inflated map length and ordering error when the number of markers is much higher than the mapping population size[31]. This is a critical improvement because incorrect map length can greatly affect subsequent recombination estimates.

For the genomic linkage map, each LG was named based to their published pseudo-chromosomes[20] with 61.7% of linked SNP sequences found in the reference genome (Table 1). The oil palm consists of a diploid genome with 16 homologous chromosome pairs ($2n = 2\times = 32$). In this map three LGs fragmented into two parts from LG 2, LG 5 and LG 7, respectively (Fig. 1). Similar LG fragmentation was also observed in the genic linkage map (Fig. 2A). In this case, LG 2.2 and LG 7.2 were absent because none of their SNP was genic, whereas removing non-genic SNPs on LG 11 caused fragmentation to two groups. More markers normally confer a longer map, reflecting better mapping/genome coverage, but also potentially reflecting a reduction in the data quality as more comprehensive use is made of the dataset. The phenomena were not observed in this study. The genic linkage map with only 2,314 SNPs (1,733 cM) was longer than that of the genomic linkage map with 27,890 SNPs (1,151.7 cM). Although linkage analysis is often problematic when the number of markers is high (approaching 10,000 markers). However, we found that both JoinMap 5 and Lep-MAP3 performed equally as reported[30,31]. The common SNP markers between the genomic and the genic linkage maps were in synteny (Supplementary Fig. 1), thus Lep-MAP3 is recommended.

Sugar and lipid-related biosynthesis pathways on the genic linkage map (Fig. 2) revealed genetic polymorphisms still present within the commercial *tenera* family which could be exploited in breeding program. Notably that there is still number of unannotated genes showing polymorphism within this population indicating novel genes might have important functions and to be discovered by further QTL mapping. The published QTLs for mesocarp oil content were located on Chromosome 5 and Chromosome 11 in Deli *dura* x AVROS *pisifera* population[18]. In the genic linkage map, the QTL region on Chromosome 5 was mainly populated by plant hormone signal transduction, glycerophospholipid and glycerolipid metabolism pathway genes. The plant hormone signal transduction pathways tagged by the most SNPs was closely related to plant floral development, growth and response to the environment[38]. As for the QTL region on Chromosome 11, included glycolysis and starch-and-sucrose metabolism pathways. However, double recombination should rarely occur in the genic linkage map with an average mapping interval of 0.75 cM, which is much lower than 10 cM threshold[39] and the mapping function partly compensates for this. For future QTL mapping analyses for oil yield traits in oil palm the genic linkage map with 10-fold fewer markers can be used, while also assessing polymorphism associated with genes known to be involved in important pathways related to oil production and energy.

Recombination frequency between marker and causal gene can weaken predictability of a trait. One way to quantify this effect is through genome-wide LD decay estimation, but it is influenced by the pedigree of the material. This parameter differs across origins, or even populations[40,41]. In oil palm, longer LDs in the commercial

populations have been observed, compared to wide germplasm materials[18,19,42] and the evidence presented here suggests that maternal and paternal recombination patterns also vary, which potentially influences breeding strategies. A good reference genome is also a key enabler to identify and characterise causal genes, based on recombination and to compare differences in genetic order, which could be indicative of structural rearrangements when compared to the physical genome. We further evaluated the assembled reference genome by comparing this with the genomic linkage maps of *dura*, *pisifera* and *tenera*. In general, these three linkage maps were well integrated and consistent with the physical map, although differences in cM/Mb were consistently observed between the two parental maps. Telomeres are hotspots of recombination, with crossover events rarely happening at the centromeric region in the chromosomes of many species[43]. The lack of genetic distance observed in sections of Chromosome 6 and Chromosome 7 for the *dura* parent is likely to be an effect of non-informative markers for linkage in the *dura* parent and could well indicate identity-by-descent on the maternal side. Generally, a lack of polymorphism be due to identity-by-descent issues. More broadly, assignment of further scaffolds to the pseudo-chromosomes may allow more extensive genome coverage and could help to assign more genetic effects to specific chromosomes, rather than 'group 17' (unplaced) assignment, indicating a failure to be localised to a pseudo-chromosome (Fig. 4). The detected outliers and inversion of chromosome segments with respect to the reference genome also might reflect genuine rearrangements, mainly inversion and translocation or even duplications. Long-read sequencing approaches, such as Pacific Biosciences (PacBio)[44] and Oxford Nanopore technologies[45] along with optical mapping could be coupled with linkage mapping to confirm the extension of the genome from newly added scaffolds as well as to differentiate the genuine structural rearrangements from the errors[46].

The evenness of recombination can also be assessed by recombination rates with respect to physical distance[47]. The average recombination rate for the commercial Deli dura x AVROS *pisifera* progeny was estimated at 1.8 cM/Mb (Table 1), which was comparable to 2.0 cM/Mb for a bi-parental recombination inbred lines (RILs) derived from a chickpea (*Cicer arietinum* L.) family[48]. On the other hand, the parameter was higher than that of fourteen maize full-sib families of DH lines with <1.0 cM/Mb[49]. Breeding design and genetic resources that influence the recombination landscape in the genome, usually receive little attention in mapping analysis of oil palm. For oil palm breeding, a variant of reciprocal recurrent selection (RRS) or family and individual selection (FIS) is deployed, producing thick-shelled dura maternal and shell-less *pisifera* paternal pools and hybrids with thin shells and more mesocarp[2]. The superior parents with good combining ability assessed through progeny testing are pollinated to produce thin-shelled *tenera* commercial productions, with each selection cycle taking up to 12–19 years. Thus, selection of mapping population mostly depends on the existing trials, which were originally established for conventional breeding programmes. Furthermore, recombination rates of Deli *dura* and AVROS *pisifera* were also different in this study, further supporting their distinct selection histories in the past. The Deli *dura* are (supposedly) direct descendants of the four palms planted in 1848 in the Bogor Botanical Garden, Indonesia, whereas the AVROS largely originated from the well-known Djongo palms in Congo[2]. We found that the highest rates of recombination in the *tenera* were inherited from the AVROS *pisifera*, but not on all chromosomes (Fig. 3). This may partly be an effect of uninformative meiotic events due to identity-by-descent in the *dura* which appears greater here, but the consistent trend on many chromosomes of higher recombination rates in the *pisifera* parent than the *dura* parent suggests a real effect. The results agree with the presence of limited polymorphism in particular chromosomes, partly reflecting the narrow genetic base of Deli *dura* due to generations of self-pollination or sib-mating originated from the four ancestors[21]. This findings pave the way for beginning to understand the effects of selection pressure in both parent pools and how these might be combined to achieve additional heterosis in oil palm.

This study provides a good foundation to improve the genome coverage in the oil palm genome assembly, to unlock the full potential of genome-enabled biology for oil palm. By using a high-density SNP array and a population of 295 palms, we have been able to assign a further 311 Mb of scaffolds to the pseudo-chromosomes of Singh *et al.*[20]. However, this additional 311 Mb has been assigned according to the linkage order in the Deli *dura* x AVROS *pisifera* cross (278 x TT41/4) used in this study, which may differ from the actual scaffold order (and, ultimately, the genome sequence order) present in the original *pisifera* used to generate the genome sequence. An understanding of these differences will be critical to fully understand GxE and GxG interactions, as well to begin to unravel the underlying basis for some trait QTL having strong effects in one origin, but none in others. The development of a genic-based map also allows breeders to search for QTLs for important oil yield traits with a smaller subset of candidate genes from known pathways. More importantly, the recombination landscape across parents potentially permits understanding of selection pressure in the genome of oil palm and where further improvement may be possible. Appropriate experimental designs and genetic resources are equally important as marker interval and population size to further improve the map quality and QTL searches, both through biparental and population approaches.

References

1. Hartley, C. W. S. The Oil Palm. *oil palm*, https://doi.org/10.1002/9780470750971.ch1 (1967).
2. Corley, R. H. V. & Tinker, P. B. *The Oil Palm: Fifth Edition* (2015).
3. FAO. FAOSTAT. *Food and Agriculture Organization of the United Nations* (2017).
4. Byerlee, D., Falcon, W. P. & Naylor, R. L. The tropical oil crop revolution: food, feed, fuel, and forests. *The tropical oil crop revolution: food, feed, fuel, and forests*, https://doi.org/10.1093/acprof:oso/9780190222987.001.0001 (2016).
5. Wong, C. K. & Bernardo, R. Genomewide selection in oil palm: Increasing selection gain per unit time and cost with small populations. *Theor. Appl. Genet*, https://doi.org/10.1007/s00122-008-0715-5 (2008).
6. Murphy, D. J. The future of oil palm as a major global crop: Opportunities and challenges. *Journal of Oil Palm Research* (2014).
7. Mayes, S., Jack, P. L., Corley, R. H. & Marshall, D. F. Construction of a RFLP genetic linkage map for oil palm (Elaeis guineensis Jacq.). *Genome* **40**, 116–122 (1997).
8. Billotte, N. *et al*. Microsatellite-based high density linkage map in oil palm (Elaeis guineensis Jacq.). *Theor. Appl. Genet.* **110**, 754–765 (2005).

9. Rance, K. A., Mayes, S., Price, Z., Jack, P. L. & Corley, R. H. V. Quantitative trait loci for yield components in oil palm (Elaeis guineensis Jacq.). *Theor Appl Genet*, https://doi.org/10.1007/s122-001-8204-z (2001).
10. Pootakham, W. *et al*. Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics* **105**, 288–295 (2015).
11. Lee, M. *et al*. A consensus linkage map of oil palm and a major QTL for stem height. *Sci. Rep*, https://doi.org/10.1038/srep08232 (2015).
12. Billotte, N. *et al*. QTL detection by multi-parent linkage mapping in oil palm (Elaeis guineensis Jacq.). *Theor. Appl. Genet*, https://doi.org/10.1007/s00122-010-1284-y (2010).
13. Gogoi, J., Ali, T. & Sharmah, R. Stochastic Model for Genetic Recombination because of Crossing over and Chromatid Exchange. *Int. J. Comput. Appl.* **80** (2013).
14. Kosambi, D. D. The Estimation of Map Distances from Recombination Values. *Ann. Eugen*, https://doi.org/10.1111/j.1469-1809.1943.tb02321.x (1943).
15. Haldane, J. B. S. The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics*, https://doi.org/10.1016/j.biortech.2011.07.096 (1919).
16. Sidhu, D. & Gill, K. S. Distribution of genes and recombination in wheat and other eukaryotes. *Plant Cell. Tissue Organ Cult* (2005).
17. Jordan, K. W. *et al*. The genetic architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. *Plant Journal*, https://doi.org/10.1111/tpj.14009 (2018).
18. Teh, C. K., *et al*. Genome-wide association study identifies three key loci for high mesocarp oil content in perennial crop oil palm. *Sci. Rep*. **6** (2016).
19. Kwong, Q. B. *et al*. Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm. *Mol. Plant* **9**, 1132–1141 (2016).
20. Singh, R. *et al*. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature* **500**, 335–339 (2013).
21. Seng, T. Y. *et al*. Genetic linkage map of a high yielding FELDA Deli × Yangambi oil palm cross. *PLoS One* **6** (2011).
22. Purcell, S. *et al*. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
23. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67–D72 (2016).
24. Wong, Y. C. *et al*. Differential gene expression at different stages of mesocarp development in high- and low-yielding oil palm. *BMC Genomics*, https://doi.org/10.1186/s12864-017-3855-7 (2017).
25. Wong, Y. *et al*. Expression Comparison of Oil Biosynthesis Genes in Oil Palm Mesocarp Tissue Using Custom Array. *Microarrays*, https://doi.org/10.3390/microarrays3040263 (2014).
26. Kent, W. J. BLAT - The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
27. Ogata, H. *et al*. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27**, 29–34 (1999).
28. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–12 (2015).
29. BLAST. BLAST Basic Local Alignment Search Tool. *Blast Program Selection Guide* 1–18 (2013).
30. Rastas, P. *et al*. Lep-MAP: Fast and accurate linkage map construction for large SNP datasets. *Bioinformatics* **29**, 3128–3134 (2013).
31. Rastas, P. Lep-MAP3: Robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* **33**, 3726–3732 (2017).
32. Van Ooijen, J. W. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genet. Res. (Camb).* **93**, 343–349 (2011).
33. Hussain, W. *et al*. Genotyping-by-Sequencing Derived High-Density Linkage Map and its Application to QTL Mapping of Flag Leaf Traits in Bread Wheat. *Sci. Rep*, https://doi.org/10.1038/s41598-017-16006-z (2017).
34. Voorrips, R. E. MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *J. Hered.* **93**, 77–78 (2002).
35. Ting, N. C. *et al*. High density SNP and SSR-based genetic maps of two independent oil palm hybrids. *BMC Genomics* **15** (2014).
36. Bai, B. *et al*. Developing genome-wide SNPs and constructing an ultrahigh-density linkage map in oil palm. *Sci. Rep*. **8** (2018).
37. Wu, Y., Bhat, P. R., Close, T. J. & Lonardi, S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet*, https://doi.org/10.1371/journal.pgen.1000212 (2008).
38. Campos-Rivero, G. *et al*. Plant hormone signaling in flowering: An epigenetic point of view. *Journal of Plant Physiology*, https://doi.org/10.1016/j.jplph.2017.03.018 (2017).
39. Xu, Z., Zou, F. & Vision, T. J. Improving quantitative trait loci mapping resolution in experimental crosses by the use of genotypically selected samples. *Genetics*, https://doi.org/10.1534/genetics.104.033746 (2005).
40. Bosch, E. *et al*. Decay of linkage disequilibrium within genes across HGDP-CEPH human samples: Most population isolates do not show increased LD. *BMC Genomics*, https://doi.org/10.1186/1471-2164-10-338 (2009).
41. Stich, B., Urbany, C., Hoffmann, P. & Gebhardt, C. Population structure and linkage disequilibrium in diploid and tetraploid potato revealed by genome-wide high-density genotyping using the SolCAP SNP array. *Plant Breed*, https://doi.org/10.1111/pbr.12102 (2013).
42. Jin, J. *et al*. Draft genome sequence of an elite Dura palm and whole-genome patterns of DNA variation in oil palm. *DNA Res*, https://doi.org/10.1093/dnares/dsw036 (2016).
43. Mézard, C. Meiotic recombination hotspots in plants. *Biochem. Soc. Trans.* **34**, 531–534 (2006).
44. Rhoads, A. & Au, K. F. PB Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, https://doi.org/10.1016/j.gpb.2015.08.002 (2015).
45. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology*, https://doi.org/10.1038/nbt.3423 (2016).
46. Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T. & Sandhu, M. S. Long reads: their purpose and place. *Hum. Mol. Genet*, https://doi.org/10.1093/hmg/ddy177 (2018).
47. Reich, D. E. *et al*. Linkage disequilibrium in the human genome. *Nature*, https://doi.org/10.1038/35075590 (2001).
48. Deokar, A. A. *et al*. Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. *BMC Genomics* **15** (2014).
49. Bauer, E. *et al*. Intraspecific variation of recombination rate in maize. *Genome Biol*, https://doi.org/10.1186/gb-2013-14-9-r103 (2013).

Acknowledgements

We would like to thank the Breeding and Molecular Breeding Units of Sime Darby Plantation R&D for their assistance in sample collection and DNA preparation, and the IT team for high-performance computing-related support. We would also like to thank DNALandmarks Inc., Canada for their sequencing and genotyping service. This study is fully supported by grants from Sime Darby Plantation R&D Centre, Malaysia.

Author Contributions

Conceptualization, C.K.T., A.L.O. and Q.B.K., Methodology, A.L.O., Q.B.K., C.K.T. and P.T.; Software, A.L.O. and Q.B.K.; Formal Analysis, A.L.O., Q.B.K., C.K.T. and P.T.; Investigation, A.L.O., Q.B.K., C.K.T. and P.T.; Data Curation, A.L.O., C.K.T, Q.B.K. and P.T.; Writing – Original Draft, A.L.O. and C.K.T.; Writing – Review & Editing, S.M., F.M. and D.R.A.; Visualization, A.L.O., C.K.T. and P.T.; Supervision, S.M., F.M. and D.R.A.; Project Administration, A.L.O., C.K.T. and Q.B.K.; Funding Acquisition, D.R.A.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-42989-y.

Competing Interests: The authors declare no competing interests.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019

plants

MDPI

Article

An Improved Oil Palm Genome Assembly as a Valuable Resource for Crop Improvement and Comparative Genomics in the *Arecoideae* Subfamily

Ai-Ling Ong [1,2,*], Chee-Keng Teh [1,2], Sean Mayes [2], Festo Massawe [3], David Ross Appleton [1] and Harikrishna Kulaveerasingam [1]

1 Biotechnology & Breeding Department, Sime Darby Plantation R&D Centre, Serdang 43400, Selangor Darul Ehsan, Malaysia; teh.chee.keng@simedarbyplantation.com (C.-K.T.); david.ross.appleton@simedarbyplantation.com (D.R.A.); harikrishna.k@simedarbyplantation.com (H.K.)

2 School of Biosciences, University of Nottingham, Sutton Bonington Campus, Leicestershire LE12 5RD, UK; sean.mayes@nottingham.ac.uk

3 School of Biosciences, University of Nottingham Malaysia, Semenyih 43500, Selangor Darul Ehsan, Malaysia; festo.massawe@nottingham.edu.my

* Correspondence: ong.ailing.sdtc@simedarbyplantation.com

Received: 9 September 2020; Accepted: 21 October 2020; Published: 3 November 2020



Abstract: Oil palm (*Elaeis guineensis* Jacq.) is the most traded crop among the economically important palm species. Here, we report an extended version genome of *E. guineensis* that is 1.2 Gb in length, an improvement of the physical genome coverage to 79% from the previous 43%. The improvement was made by assigning an additional 1968 originally unplaced scaffolds that were available publicly into the physical genome. By integrating three ultra-dense linkage maps and using them to place genomic scaffolds, the 16 pseudomolecules were extended. As we show, the improved genome has enhanced the mapping resolution for genome-wide association studies (GWAS) and permitted further identification of candidate genes/protein-coding regions (CDSs) and any non-coding RNA that may be associated with them for further studies. We then employed the new physical map in a comparative genomics study against two other agriculturally and economically important palm species—date palm (*Phoenix dactylifera* L.) and coconut palm (*Cocos nucifera* L.)—confirming the high level of conserved synteny among these palm species. We also used the improved oil palm genome assembly version as a palm genome reference to extend the date palm physical map. The improved genome of oil palm will enable molecular breeding approaches to expedite crop improvement, especially in the largest subfamily of *Arecoideae*, which consists of 107 species belonging to *Arecaceae*.

Keywords: *Elaeis guineensis*; high-density linkage maps; physical maps; comparative genomics; genome-enabled breeding; *Arecoideae*; pseudomolecules

1. Introduction

The *Arecaceae* (formerly known as *Palmeae*) family of the monocot order *Arecales* is a plant family hosting perennial palm species that mostly grow across the equatorial belt. The family is made of 181 genera with 2600 species [1]. Among them, the genus *Elaeis* (oil palm) and genus *Cocos* (coconut) together are the world's largest source of oils and fats (about 35%), while date palm plays a major role in food security and agricultural production in arid and fragile ecosystems worldwide.

Two species of *Elaeis* including *E. guineensis* Jacq. and *E. oleifera* (HBK) Cortes originating from West Africa and Latin America, respectively, produce edible oil from fruit mesocarp and kernel. However, only the African oil palm is typically commercially planted due to its superior oil yield with an average oil yield in Malaysia of 3.5 tons hectare^{-1} year^{-1}, compared to only 0.4 tons hectare^{-1}

year^{-1} produced by its South American counterpart [2]. The average oil yield of commercial oil palm is also roughly 8 to 10 times higher than other major temperate oil crops [3], making it the most efficient oil crop in the world. However, with increasing demand for fats and oils from a growing population, yield improvement is essential to meet food requirements at the same time as halting deforestation. Crop improvement through breeding and genomics is the basis for increasing yield and therefore productivity without the use of more land.

Although coconut (*C. nucifera* L.) is not as productive as oil palm, it generates more products such as copra, coir, and timber to serve traditional markets, particularly in Asia. Nut yield in Asia is approximately 5.5 tons hectare^{-1} year^{-1}, while the copra yield ranges from 10 to 20 percent of the nut weight, equivalent to around 1 ton hectare^{-1} year^{-1} [4]. The recent rise in demand for coconut products has resulted from coconut water and virgin coconut oil being promoted as an energy drink and health supplement. The market size of coconut products was reported to be USD 11.5 billion in 2018 and is anticipated to reach USD 31.1 billion by 2026 [5]. Another related palm species, the date palm (*Phoenix dactylifera* L.) does not produce oil but stores mostly carbohydrate, which is converted to sugar in its fruit mesocarp [6]. Hence, the crop has been domesticated in the Middle East and Indus Valley as a major staple food, and the earliest cultivation can be traced back to 3700 BC [7]. Date palm thrives in some of the most arid environments on earth and is able to create a local microclimate, which also allows other crops to be grown [8]. Similarly, higher global demand for inherent medicinal, nutritional, and health advantages of date palm products has boosted the market value of the date palm industry from USD 9.2 billion in 2014 to USD 13 billion in 2018 [9].

Within the *Arecaceae*, there are many other palms that could make a significant contribution to economic development, food, and nutritional security and provide ecosystem services for sustainable cultivation, including sago palm, nipa palm, peach palm, etc. However, it is unlikely that companies will be willing to make major investment in developing genetic improvement programs without being sure that there will be sizable returns. To date, oil palm receives the most research attention due to its significant role in the world economy, although coconut and date palm have seen steep increases in research focus with their recent emergence. In general, genomic research is centered on the development of marker-assisted selection (MAS) for expediting breeding selection for these outcrossing perennial palms that typically require more than a decade to complete a breeding cycle [10–14]. In the 90s, the first restriction fragment length polymorphism (RFLP)-based linkage map of oil palm [10] was reported and subsequently deployed for mapping quantitative trait loci (QTL) responsible for fruit characteristics and vegetative traits [15]. The effort was continued and expanded using various DNA marker systems: random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), diversity arrays technology (DArT), simple sequence repeat (SSR), and single nucleotide polymorphism (SNP) [10,16–21]. The reported QTL, however, were only applicable to a few populations closely related to the mapping populations. In oil palm, the Deli *dura* that produces a thick kernel shell, high bunch number, and high mesocarp oil content and is widely used in different programs and subsequently has been developed in breeding populations to different sub-populations, such as *Johore Labis, Ulu Remis*, and *Gunung Melayu.*

Mapping resolution, especially for complex traits (e.g., oil yield and disease tolerance), is limited by insufficient recombination and a lack of statistical power within a small bi-parental family, which usually consists of only 16 to 96 field-planted palms in which traits are recorded in breeding programs. To improve mapping resolution, research efforts gradually shifted towards genome-wide association studies (GWAS) to access total meiotic recombination accumulated in a large admixed population. Nevertheless, the mapping method was of limited power until the reference *E. guineensis* genome (AVROS *pisifera*), namely P5-build, was sequenced and made publicly available in 2013 [22]. We then reported the first GWAS for mesocarp oil content based on 2045 *tenera* oil palms [23] using a high-density OP200K SNP array [24]. However, only 60% of the linked SNPs were located across 16 chromosomes (or pseudomolecules) of the reference oil palm genome, with the P5-build (2n = 32), leaving room for improvement. The same limitations on GWAS were also observed in the date palm [25].

A good genome assembly does not only rely on assembly of contigs into scaffolds but also the placement and contiguity of scaffolds on chromosomes to form the physical map. To summarize the latest status of assembled genomes of oil palm, coconut, and date palm (as available in National Centre for Biotechnology Information; NCBI): 1535.18 Mb of P5-build oil palm genome first achieved a chromosome coverage level with 43% of scaffolds anchored [22], although no improvement has been published since 2013. On the other hand, the 2202.46 Mb of assembled coconut genome and the 772.32 Mb of assembled date palm genome have recently surpassed the oil palm genome, with 46% [26] and 50% [25] of scaffolds anchored to the chromosomes. The genome size of oil palm is estimated at about 1.8 Gb [27,28]. We previously reported on the possibility of further extending the physical map of the current P5-build reference genome using one ultra-dense linkage map of a commercial Deli *dura* × AVROS *pisifera* population, and the total length of additional scaffolds anchored in our previous study has improved by 311 Mb [29]. Hence, our aim in this study was to maximize the improvement of the oil palm genome assembly by integrating multiple linkage maps from various origins. Here, we also report how the improved oil palm assembly, namely using Physical Map version 6 (PMv6), can serve as an important reference for comparative genomic studies among coconut and date palms, with expansion to other *Arecaceae* members in the future.

2. Results

2.1. Construction of Multiple High-Density Oil Palm Linkage Maps

Three bi-parental populations, including the reported commercial Deli *dura* × AVROS *pisifera* (295 *tenera* palms) [29], semi-wild Deli *dura* × Nigerian *dura* (112 *dura* palms), and *Johore Labis dura* × *Johore Labis dura* (214 *dura* palms) were used to construct individual linkage maps. We obtained 17 to 24 linkage groups (LGs; Table 1) with one to six chromosomes being represented by two or three linkage groups representing the expected 16 major LGs for total oil palm chromosomes number. The *Johore Labis dura* population showed the lowest number of linked SNPs (6920 loci; due to the use of refined SNP chip of 20K loci), while the highest number of linked SNPs (32,650 loci) was observed for the Deli *dura* × Nigerian *dura* population. The Deli *dura* × AVROS *pisifera* map remained the densest with a mean inter-marker distance of 0.04 cM, as reported [29], compared to the sparsest *Johore Labis* map (0.18 cM). In this study, a total of 11,421 common SNPs was identified and adopted for integration of the multiple linkage maps.

Table 1. Statistics of linkage maps developed for each oil palm population.

Population	Number of Linked Markers	Number of Linkage Groups (LGs)	Length of Linkage Map (cM)	Marker Interval (cM)	Genome-Wide Recombination Rate of P5-Build (cM/Mb)	Genome-Wide Recombination Rate of PMv6 (cM/Mb)
Deli *dura* × AVROS *pisifera* *	27,890	19	1151.70	0.04	1.75	0.98
Deli *dura* × Nigerian *dura*	32,650	17	1646.95	0.05	2.50	1.40
Johore Labis dura	6920	24	1268.26	0.18	1.93	1.08

* previously reported [29].

2.2. Improvement of the Oil Palm Physical Genome Assembly by Integration of Multiple Oil Palm Linkage Maps

Linkage-directed genome assembly improvement not only anchored previously unplaced scaffolds to the pseudomolecules but also split and reassigned a range of the scaffolds flanked by N bases in the published physical genome, P5-build [22]. The newly assembled genome, termed PMv6 with an N50 of 83.1 Mb, improved the scaffold assignment to oil palm pseudomolecules from 43% to 77% (1.2 Gb, Figure 1). On average, each pseudomolecule in the PMv6 genome was assembled using 142 scaffolds with an average length of 73.7 Mb, whereas only 19 scaffolds giving an average length of 41.1 Mb were

reported in Singh et al., 2013 [22]. Chromosome 2 was also extended by 1.2-fold and now is the largest pseudomolecule, accounting for 12% of the total PMv6 assembly length (Figure 1a,b).

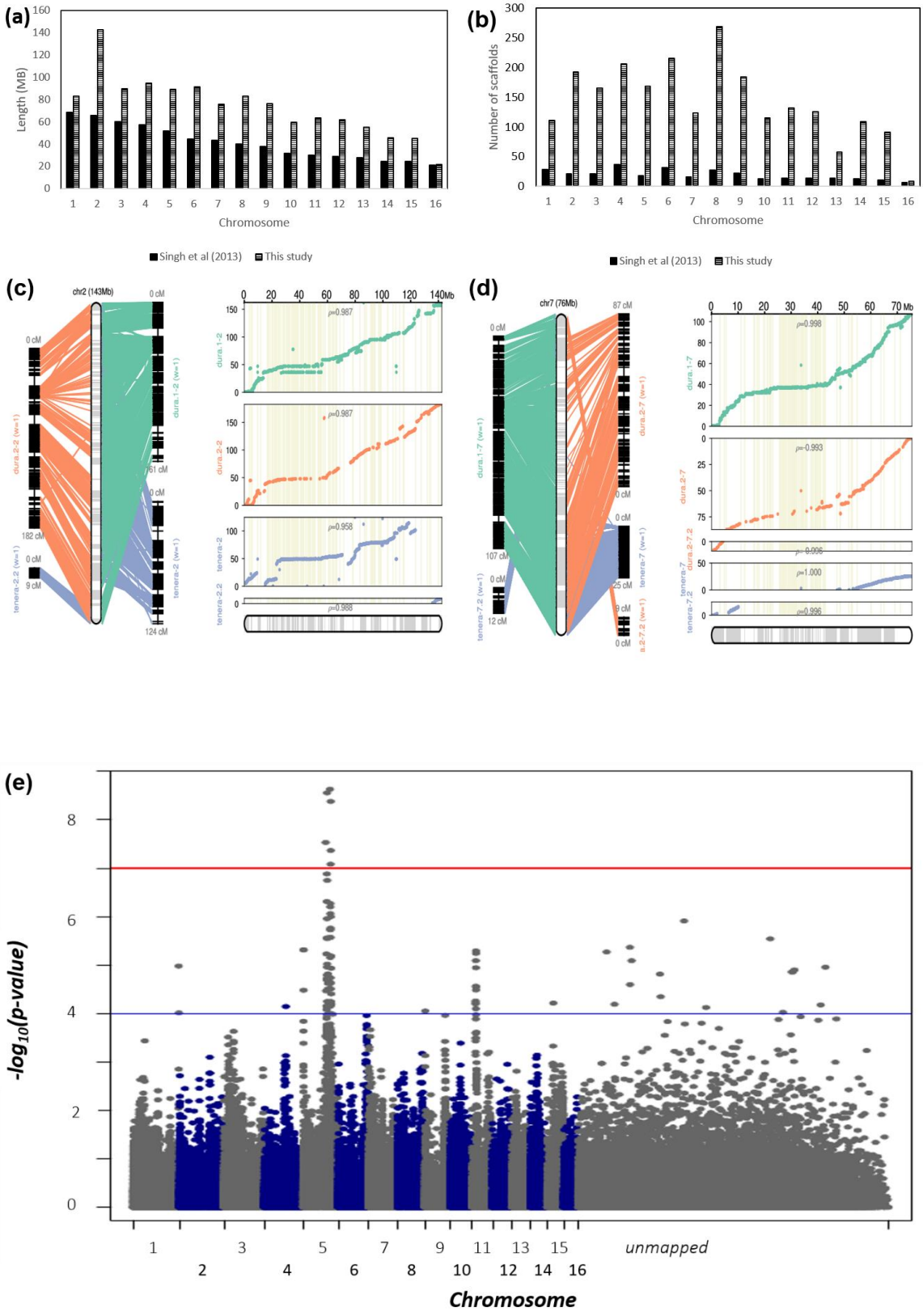


Figure 1. *Cont.*

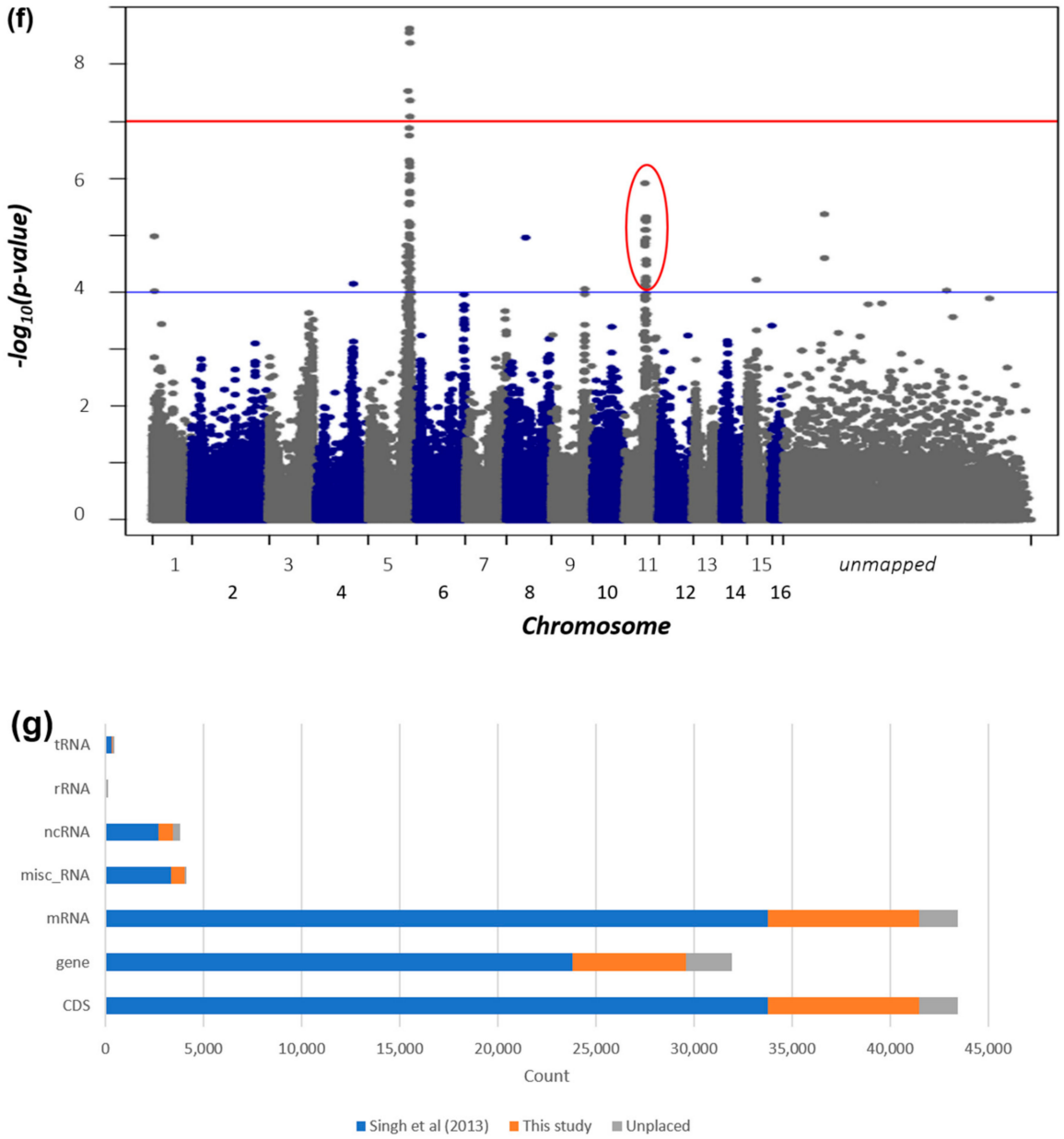


Figure 1. Refined assembly of oil palm genome (PMv6). (**a**) Chromosome length (Mb) comparison between P5-build, Singh et al., (2013), and PMv6 (this study). (**b**) Number of scaffolds from previous physical map [22]) and the improved physical map developed in this study. (**c**) Buildup of PMv6 by scaffold anchoring using multiple linkage maps; left panel represents the central combined linkage map from three mapping populations, and the right panel shows the correlation plots of each populations' genetic versus physical order; presented in the order of Deli *dura* × Nigerian *dura* (*dura* 1—green), *Johore Labis dura* (*dura* 2—orange), and Deli *dura* × AVROS *pisifera* (*tenera*—blue) for Chromosome 2 and (**d**) Chromosome 7 as examples. (**e**) Genome-wide association studies (GWAS) Manhattan plot for a commercial oil palm population for one of the oil yield trait [23] for the genome published by P5-build. (**f**) GWAS Manhattan plot for the same commercial oil palm population for one of oil yield trait after assembly improvement in this study, circle in red represent newly assigned and significant SNPs. (**g**) Statistics of genomic features improvement after placement of scaffolds into chromosomes for P5-build and this study.

In Figure 1c,d, we illustrate how the scaffolds were anchored to Chromosome 2 and 7 based on the integrated linkage maps, and how they were validated by correlation between linkage and physical positions (Figure 1c,d). The results show that assembly improvement using the commercial Deli *dura*

× AVROS *pisifera* map [29] (blue) was only limited to one end of the pseudomolecule, especially on Chromosome 7. Nevertheless, the shortcomings were solved when integrating the semi-wild Deli *dura* × Nigerian *dura* map (green) and the *Johore Labis dura* map (orange). The sigmoidal curves of Chromosome 2 and 7 clearly distinguishing both telomere regions from centromeric regions where limited recombination events were observed. The correlation between linkage and physical positions reached more than 0.90 in each mapping family, and the collinearity between each linkage maps can be observed with consistence marker orderings within each LG. Similar results were observed for the remaining chromosomes and are shown in Supplementary Figure S1. The details of total SNPs residing within genomic scaffolds used to build PMv6 genome assembly can be found in Supplementary Table S1. Using the correlation outputs, the genome-wide recombination rate was estimated to be 1.15 cM/Mb, which is about half of the previously reported estimation based on the P5-build genome [29] (Table 1). This investigation may infer overestimation of the recombination rate in our previous estimation, mainly due to the length extension of the PMv6 genome assembly reported here. Understanding recombination differences has distinct implications for population structure, gene evolution, and genetic improvement [30].

The improved PMv6 genome was used to evaluate the improvement of mapping resolution of the previously reported GWAS for mesocarp oil content [23], reducing the unmapped SNPs from 40% (37,003) (Figure 1e) to only 13% (12,175) (Figure 1f). The major QTLs still remain at Chromosome 5 and Chromosome 11, but the significance level on Chromosome 11 was increased to a –log *p*-value = 5.9 from 5.2 and more SNPs surpassed genome-wide -log *p*-value cutoff at 4.0, due to SNPs newly assigned to Chromosome 11, improving trait location. From the perspective of genomic features and gene annotations, the PMv6 genome assembly now includes 5769 genes, 7705 protein-coding regions (CDS), and 1563 non-coding RNA, which could not be placed into the P5-build genome (Figure 1g). Interestingly, among the non-coding RNA, 48% of total predicted ribosomal RNA (rRNA) were mapped for the first time in PMv6.

2.3. Comparative Genomic Analysis of Three Palm Species

The improved oil palm genome, PMv6 was subsequently used for genome comparison with coconut [26] and date palm [25]. The results confirmed a high level of syntenic block conservation between the oil palm and the other two palm species (Figure 2). For coconut, the 16 homologous chromosome groups (2n = 32) appeared to be highly coherent with the oil palm genome, but we observed possible fusion and fission event differences on oil palm-based Chromosome 2 (eg-2) and 10 (eg-10), respectively. Compared to coconut, a 30 Mb fragment of eg-2 fused with eg-7 accounts for coconut-based Chromosome 2 (cn-2), suggesting a fission/fusion event since the separate evolution of the two species. In a similar comparative way, a 20 Mb fragment of eg-10 has fused with eg-16 to form cn-8.

In general, the genome conserved synteny between date palm and oil palm was more complicated. By referring to Ziwen et al. (2015), oil palm and coconut were estimated to have diverged from date palm approximately 62.5 million years ago [31]. Chromosome fusion/fission events between date palm-based Chromosome 4 (pd-4) and pd-16 and between pd-1 and pd-10 may have led to formation eg-1 and eg-2 in the oil palm genome, respectively. This possibly explains why date palm (2n = 36) has two additional homologous chromosome groups compared to oil palm. In addition, pd-11 compared to oil palm has two chromosome fragments with one of them fused with pd-12 to form eg-10 and the remaining pd-11 fragment being maintained as eg-6 in the oil palm genome.

We then further compared the *SHELL* gene, which is a homologue of *SEEDSTICK* (NCBI Gene ID: LOC105034563)—a type II MADS-box transcription factor reported in oil palm (at 3.05 Mb of eg-2) with the orthologs of coconut (at 116 Mb of cn-1) and date palm (at 38 Mb of pd-1). In the oil palm genome, the *pisifera* allele (sh^{AVROS}) results in an amino acid change from lysine to asparagine, causing the absence of shell surrounding the fruit kernel observed *pisifera* fruit [32] (Figure 3a). The analysis was also extended to other candidate genes including Cytochrome P450 (*CYP703*) (LOC105059962), glycerol-3-phosphate

acyltransferase 3-like (*GPAT3*) (LOC105059961), Cytidine deaminase-like (*CYT DA*) (LOC105059743), and Lonely-Guy (*LOG*) (LOC105055182) reported to be involved in sex determination of date palm [33]. All the candidate genes (except *LOG*, located at 10.86 Mb of eg-12) were located at 44–45 Mb of eg-10 in the oil palm genome but have yet to have the scaffolds containing them placed in the date palm genome.

In Figure 3b, a phenogram based on the sex-determination-related *LOG* genes revealed that coconut has the shortest evolutionary distance to oil palm followed by date palm, among other monocots outgroup taxa (Figure 3b)

2.4. Oil-Palm-Guided Improvement of the Date Palm Genome

Due to the high levels of collinearity of gene order and genomic sequences between the palm species, the PMv6 genome was adopted as a reference to improve the published genome assemblies of date palm. We started with the earlier version, namely PDK_30 (GenBank assembly accession: GCA 000181215.2) [34]. Genomic scaffolds from date palm were tiled following oil palm ordering for each pair of syntenic chromosomes, then the joined scaffolds were split if required following the published date palm genetic map [35]. This resulted in anchoring 78% of scaffolds to the physical map from the total length of the full assembly (381.5 Mb). The next test was performed on the more recent assembly of date palm genome (GCA_009389715.1) [25] using the same approach. The second improvement was not as apparent as the first version but did result in a total length of 458 Mb, equivalent to an extra 10% of total scaffold length anchored to chromosomes, compared to the published assembly. The placement of scaffolds in each of the date palm chromosomes extended their total length except for pd-10 and pd-16 (Figure 4). The assembly improvement for both versions of the date palm genome were further compared using date palm genetic markers [35] located within each assembly. An average correlation of 0.80 was observed for the 18 chromosomes when comparing physical location of the genetic markers between both improved assemblies, ranging from the lowest on pd-16 (r = 0.38) to the highest on pd-4 and pd-15 (equally r = 0.98) (Figure S2).

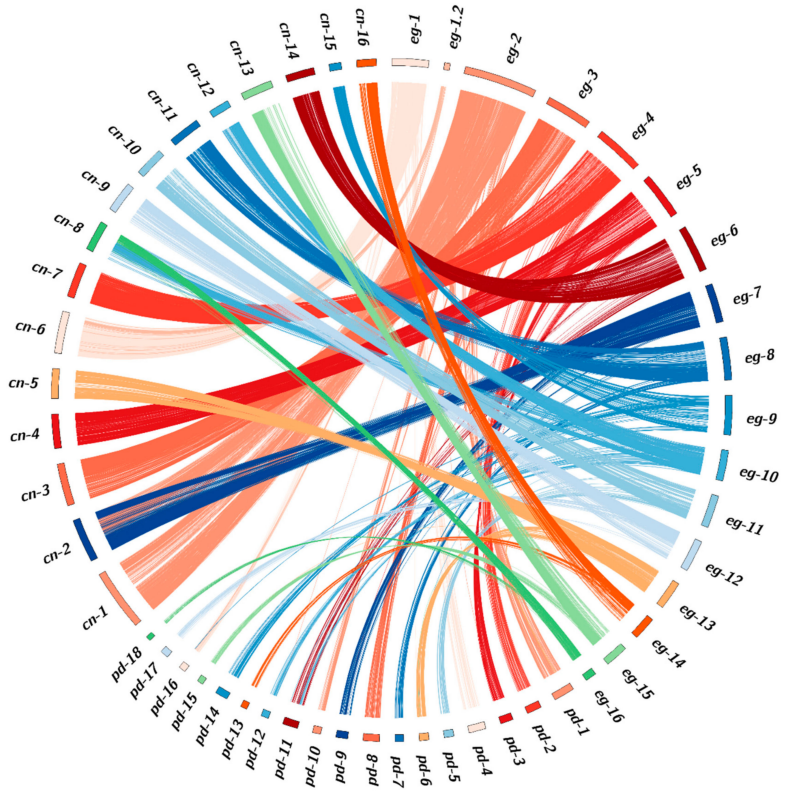


Figure 2. Synteny analysis including three palm species: oil palm (eg), coconut (cn), and date palm (pd). Chromosome pairs with strong syntenic relations were linked by the same ribbon color.

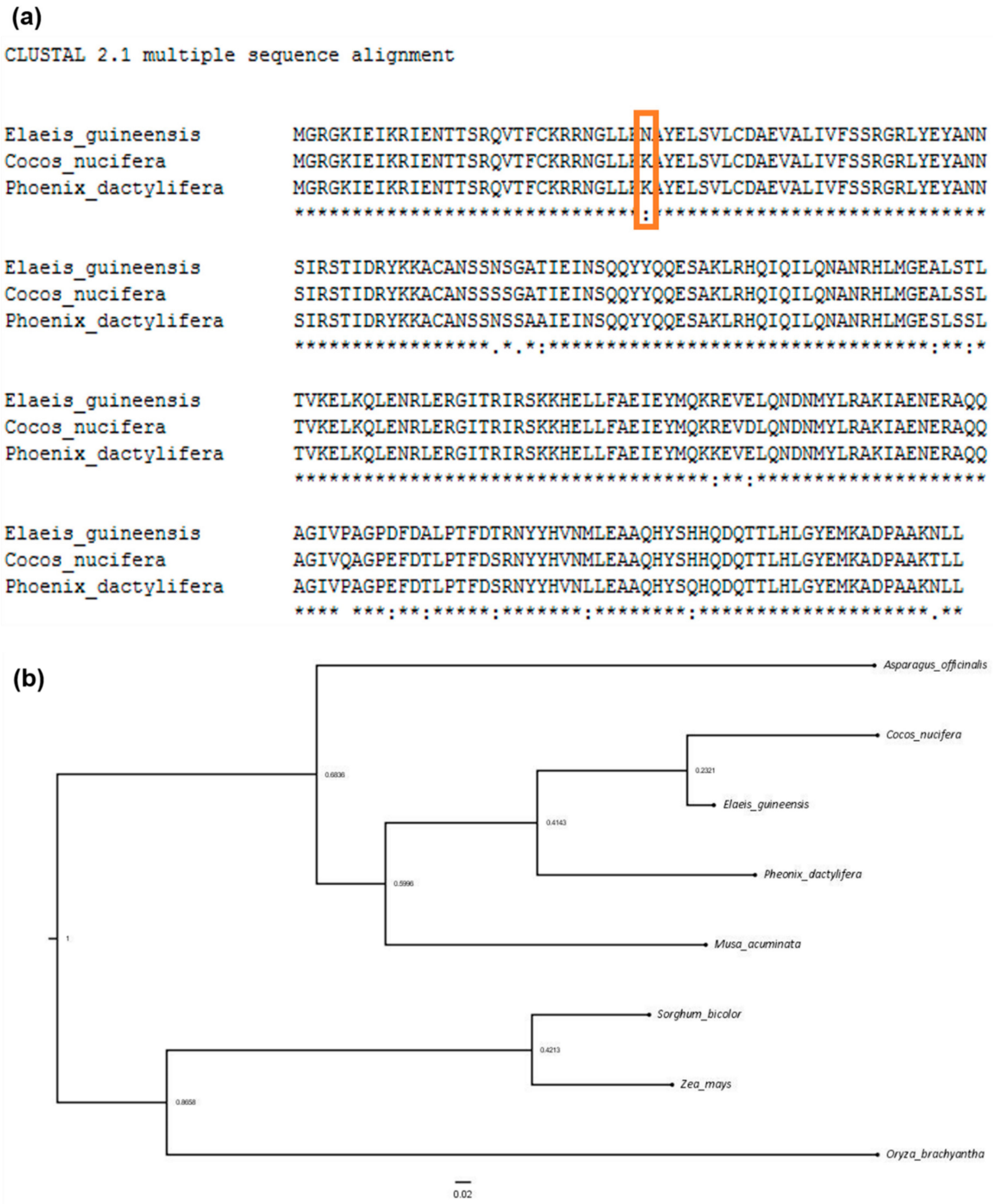


Figure 3. High conservation of genes among palm species with minimal amino acid changes. (**a**) Alignments of *SHELL* gene encoding for the layer of shell surrounding the kernel for oil palm and the orthologues from coconut and date palm. (**b**) Phenogram tree for *LOG* gene orthologues, which play a role in sex determination for date palms; orthologues of the gene among palms and other monocots outgroup taxa were included.

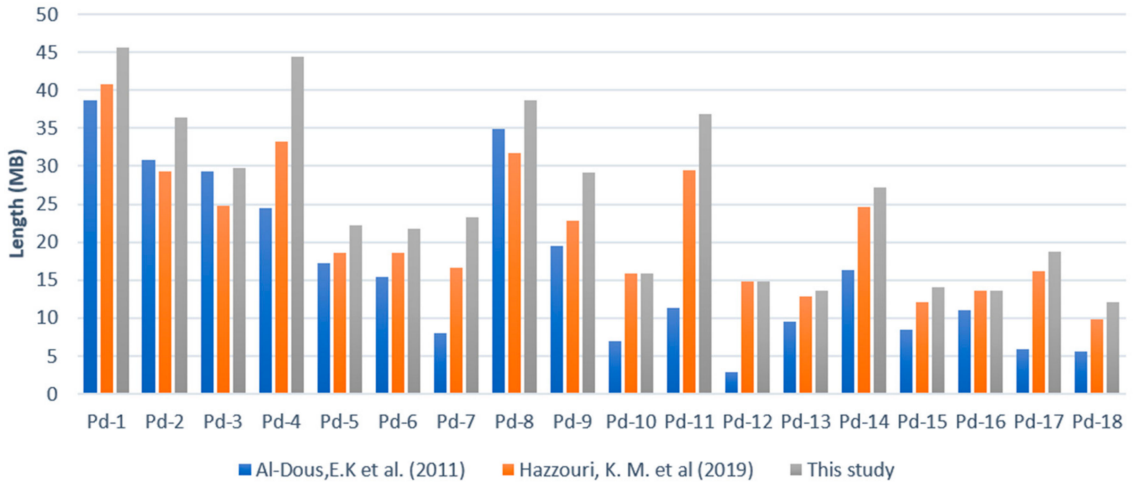


Figure 4. Oil-palm-directed date palm genome assembly improvement by length (Mb) across 18 pseudomolecules. Anchoring of scaffolds was performed on the two published assemblies of date palm genome by Al-Dous et al. (2011) [34] and later by Hazzouri et al. (2019) [25].

3. Discussion

In our recent publication, we demonstrated an extension of 311 Mb of the published oil palm pseudomolecules (P5-Build [22]) using a linkage map of a commercial Deli *dura* × AVROS *pisifera* family (295 palms) [29]. In this study, we constructed two denser linkage maps from a semi-wild Deli *dura* × Nigerian *dura* population (112 palms) and *Johore Labis dura* population (214 palms). The Deli *dura* origin is derived from four palms planted at Bogor Botanical Garden in 1848 [36] and exhibits a thick kernel shell, high bunch number, and high mesocarp oil content. It has become the most important genetic resource for major breeding programs in Southeast Asia. Separate breeding programs using Deli *dura* have led to sub-populations, such as *Ulu Remis* and *Johore Labis* [37]. Breeding history and selection have likely resulted in the lowest number of linked SNPs and the lowest mapping density being observed in the pure *Johore Labis dura* population (Table 1). To avoid inbreeding depression, Sime Darby Plantation R&D also introgressed the Deli *dura* with a Nigerian *dura* to widen the genetic base, reflected by the higher mapping density and number of linked SNPs observed.

By integrating the three linkage maps, a total of 11,421 family-transferable/common SNPs were used to anchor 77% of the assembled 1.5 Gb genome, in order to generate the new oil palm physical map, PMv6 (Figure 1), compared to 43% in the previously published P5-build physical map. Using Chromosome 2 and 7 as examples, the quality of scaffold anchoring based on the integrated linkage maps can be visualized and was validated through a correlation between genetic and physical positions, allowing the recombination landscape throughout each chromosome to be determined (Figure 1c,d). Species with larger genomes often have higher levels of recombination at the telomeres and lower recombination in the heterochromatic and centromeric regions despite genes being present within transposable element (TE) [38,39] clusters, as reported in wheat [40,41]. In such a scenario, a sigmoidal curve of the recombination rate should be expected, but this was not clearly observed in both chromosomes when assembled from the commercial Deli *dura* × AVROS *pisifera* map (blue), suggesting only a partial improvement. However, the integration of the semi-wild Deli *dura* × Nigerian *dura* map (green) and the *Johore Labis dura* map (orange) enabled extension of both chromosomes with a sigmoidal curve of recombination from telomere-to-telomere, although whether this represents the full chromosome needs further validation, through techniques such as Fluorescence in situ hybridization (FISH) technique.

The PMv6 genome enabled an increase in the mapping resolution of the reported GWAS for mesocarp oil content [23] by locating an additional 65% (24,828) previously unplaced SNPs (Figure 1f). Among the total SNP markers mapped in PMv6 assembly, there are 32,552 SNPs located within genic/non-coding RNA regions, equivalent to 25% of the total SNP assigned to scaffolds/genome

sequence. Increased marker density allows better capture of linkage disequilibrium between origins and render GWAS more powerful. Hence, PMv6 enables a more complete set of annotated genes and non-coding RNA to be located that are potentially responsible for phenotypic changes residing within QTL to be identified after conducting GWAS.

In general, most of the chromosomes between the oil palm, coconut, and date palm genomes are highly syntenic (Figure 2). They also share agriculturally desirable traits, such as tall or short varieties in coconut palm, which would be useful in oil palm and date palm. It would be possible to use the causative gene from coconut to screen for allelic differences in the gene orthologues in other species. For instance, the *SHELL* gene in the oil palm genome (eg-2) was found to be highly conserved with the genomic sequence orthologs in coconut (cn-1) and date palm (pd-1). Only the oil palm genome sequence carries sh^{AVROS} variants causing the lysine-to-asparagine amino acid change, which results in disruption of DNA binding within the MADS-box domain [32], thus no shell is found in this *pisifera* oil palm. Indeed, two more fruit forms including thick-shelled dura (sh^{Deli}/sh^{Deli} as wild-type) and thin-shelled *tenera* (sh^{Deli}/sh^{AVROS}) also naturally exist because of haploinsufficiency effect on the *SHELL* gene [32,42]. The *SHELL* variants are extremely important for oil palm breeding, because crossing between dura and *pisifera* (usually female sterile) produce *tenera* with fertility restored and 30% more oil than the dura [43,44]. Unlike oil palm, shell-less fruits in coconut and date palm have not yet been reported. The shell layer may be essential to coconut and date palm for maintaining their endosperm moisture content to allow seed germination, especially in arid habitats. Consequently, the mutation probably has been naturally selected out. Nevertheless, we still identified a strong QTL for shell thickness of *tenera* oil palms located within 2.5 Mb to 3.9 Mb sharing the same region with the *SHELL* gene on eg-2 [24], even when the later was fixed. Hence, the results obtained from oil palm research could lead to the same region in coconut palm being screened for the orthologs responsible for shell variation in coconut within cn-1 and eventually deployed in MAS for selecting thicker- or thinner-shelled coconuts. Coconut shell normally occupies 15% of the total weight of a fruit and approximately 9 million tons of coconut shell is globally produced as an agricultural waste every year [45]. Nevertheless, the waste can be turned into activated charcoal and commands a premium [46,47].

Oil palm and coconut are estimated to have diverged from date palm around 62.5 million year ago [31]. Fission or fusion of two chromosomes into one or vice versa is known in many lineages [48]; this event is observed in this study too, whereby date palm has 18 pairs of chromosomes, while oil palm and coconut have only 16 pairs. Torres et al. found that the sex-determination-related genes are now located at 44–45 MB of eg-10 [33], where fusion between pd-11 and pd-12 probably occurred, comparative to oil palm. This also suggests that eg-10 contains this cluster of key genes for inflorescence gender and that the split into two chromosomes in the date palm may be an important event leading eventually to a diecious palm. The finding is supported by a shorter evolutionary distance of LOG gene (suppression on female inflorescence) between oil palm and coconut, compared to date palm (Figure 3b). Of the three palms, date palm is the only diecious palm. The cycle of male and female inflorescences of monoecious oil palm is mainly affected by environmental and genetic factors. Water stress and excessive frond pruning lead to high induction of male inflorescences and abortion of both sexes when the stress becomes extreme [49]. Understanding the genetic control underlying sex determination in oil palm can provide a new avenue to address abiotic stress tolerance in oil palm. Understanding the basis of inflorescence sex determination may allow breeding for specific environments, to optimize sex ratio to the target environment, especially in the context of global climate change and of plantations in drought-prone areas. In terms of aptitude for branching, date palm and sago palm can also propagate vegetatively by producing suckers, which would be an extremely valuable trait in oil palm, where this currently does not happen.

Only 19% of the scaffolds were anchored based on a linkage map of the Khalas cultivar to the first date palm assembled genome, PDK_30 [34,35]. Subsequently, the genome has recently been greatly improved to 50% anchored scaffolds of the expected 772.32 Mb genome size using third generation

sequencing PacBio long reads by Hazzouri et al. [25]. In the present study, we have shown that the new oil-palm-guided date palm assembly is still able to anchor a further 10% of scaffolds, and the oil-palm-directed assembly is highly consistent with the latest date palm assembly. Hence, a full utilization of available linkage maps, especially with high mapping density should be incorporated in genome improvement. The oil palm PMv6 genome assembly can serve as a reference to date palm and coconut, and to other palm members in the future. Moreover, based on the taxonomic distances between oil palm and date palm, we would expect similar results from any palm species with a divergence time less than 62.5 million years [31] ago within *Arecoideae*, a subfamily of *Arecaceae* that consists of 107 species [50,51], using the oil palm PMv6 genome as a reference. Improved genome sequences can in turn lead to crop improvement strategies for these important food, feed, and fuel sources that will ultimately lead to the higher productivity required for agriculture to meet growing demands in the years to come without the use of more forest land.

4. Materials and Methods

4.1. Mapping Populations and DNA Preparations

Three mapping populations from oil palm breeding populations of restricted origins (BPROs) including the 295 commercial *tenera* palms derived from a Deli *dura* × AVROS *pisifera* family (278 × TT41/4), previously described in Ong et al. [29], were used. The other two populations were maintained as female parental breeding sources, namely the Deli *dura* × Nigerian *dura* and *Johore Labis dura* × *Johore Labis dura* with population sizes of 112 and 214 individuals, respectively. Leaf samples were collected from each palm planted in different breeding trials maintained by Sime Darby Plantation Research and Development Centre, Sime Darby Plantation Sdn Bhd, Malaysia. Fresh leaf tissue was sampled from the third frond of each palm. Samples of genomic DNA were isolated from 0.1 g of the leaf tissue using the DNAeasy Plant Mini Kit (Qiagen, Germany). The DNA concentration and purity were quantified on a 0.8% agarose gel using known standards.

4.2. SNPs Identification and Genotyping

The extracted genomic DNA (25 ng/µL) samples from Deli *dura* × Nigerian *dura* and Deli *dura* × AVROS *pisifera* were genotyped using the 200K SNPs included on the published OP200K Infinium array (Illumina, USA) [24], and the population *Johore Labis dura* × *Johore Labis dura* was genotyped using another 20K SNP panel, selected as a subset of the 200K panel on the Infinium iScan platform according to the manufacturer's recommendations. Raw genotyping data were analyzed with GenomeStudio version 20011.1 with genotyping module version 1.8.4. Using a GenCall score cutoff of 0.15, auto-clustering of the SNPs was done. The SNP clustering was confirmed manually by visual inspection. The SNP calls were exported into the PLINK program for minor allelic frequency (MAF) cutoff at 0.01 [52]. Polymorphic SNPs were identified where at least one of the parents was heterozygous. Subsequently, the polymorphic SNPs were dropped where the call rate <95% or where there was significant segregation distortion, using a false discovery rate (FDR)-corrected method with an effective cutoff at p-value < 0.05.

4.3. Construction of Linkage Maps

All linkage maps were constructed using Lep-MAP3 [53,54] for large SNP datasets, especially developed for outbred families such as oil palm with an underlying maximum likelihood (ML) method. Filtering module was used for marker quality checking, followed by running the SeparateChromosomes module for binning the assayed markers into linkage groups with optimized LOD values ranging from 5 to 15 with intervals of 5. The JoinSingles module assigned singular markers to existing linkage groups to maximize the map ability of the total input marker set. Lastly, the OrderMarkers module ordered the binned markers according to the Kosambi mapping

function for conversion of recombination frequencies into map distances (in unit centiMorgan, cM). The option "sexAveraged = 1" was selected along with OrderMarkers to join the maps of both parents.

4.4. Anchoring of Scaffolds to Linkage Groups

Linked SNP markers (from OP200K) with flanking sequences of 60mers located in the oil palm scaffolds (placed and unplaced in pseudomolecules) were identified using BLASTN [55] with e-value cutoff at 1×10^{-150} and sequence identity at least of 95%. High-density linkage maps built from the three mapping populations were merged using ALLMAPS assembly module, together with equal weighting among the linkage maps [56]. Scaffolds were then ordered and orientated according to the merged and weighted linkage maps, thus resulting in the new genome content. The same parameters have been applied to date palm SNP markers with flanking sequences of 100mers to identify the locations of the markers in two of the date palm genome versions.

4.5. Genome and Gene Comparisons

Palm genomic sequences were downloaded from NCBI. Alignments of the coconut [26] and date palm [57] genome against the current oil palm genome (PMv6) was done using Nucmer, which is a module developed in Mummer [58,59], an ultrafast sequences aligner, made for genome alignments. The syntenic relationship between both oil palm and coconut chromosomal alignments were visualized using the CIRCOS software [60]. Chromosome pairs with strong syntenic relations between oil palm, coconut, and date palm were linked by same ribbon colors. Candidate genes of interest were also aligned using multiple sequence alignment tool, clustalw [61,62] via translated protein and orthologs of other closely related monocots as outgroup: asparagus (*Asparagus officinalis*), banana (*Musa acuminata*), sorghum (*Sorghum bicolor*), corn (*Zea mays*), and turf (*Oryza brachyantha*) protein sequences were identified from NCBI. The phenogram was build using neighbor-joining (NJ) with 1000 bootstrap iterations and visualized using a phylogenetic tree viewer, namely Figtree (v1.4.4) [63].

4.6. Reconstruction of Date Palm Pseudomolecules

Mummer (Nucmer) output from the step above tiled the date palm's query scaffolds according to the sequence order in the newly extended oil palm pseudochromosomes. Those scaffolds were then ordered and orientated into GenBank agp formats, and the pseudomolecules were built using the agp module in ALLMAPS. SNP markers obtained from the published date palm genetic maps [35] were aligned to the new pseudomolecules using the same BLASTN criteria, as above. The integration of SNP marker locations in the genetic map against physical map location was plotted using ALLMAPs path module. Chimeric contigs, which contained ambiguous markers that mapped to more than two linkage groups, were then processed using the split module. Ambiguous short contigs were excluded during the second round of rebuild of pseudomolecules by running the ALLMAPs path module again.

Supplementary Materials: The following are available online at http://www.mdpi.com/2223-7747/9/11/1476/s1, Figure S1: Integration of genetic with physical maps for oil palm, Chromosomes 1 to 16; Figure S2: Correlation plots for locations of DNA marker located within date palm chromosome (pd-15) for genome assembly improvement versions by Al-Dous et al. (2011) and Hazzouri et al. (2019); Table S1: Summary for consensus map generated from multiple linkage maps and statistics for SNP markers and scaffolds used to build genome assembly for PMv6. The assembled genome has been uploaded in NCBI with submission number: SUB7516020/ BioProject: PRJNA636092.

Author Contributions: Conceptualization, A.-L.O., C.-K.T., and S.M.; investigation, A.-L.O., C.-K.T., and S.M.; manuscript writing, C.-K.T. and A.-L.O.; writing—review and editing, S.M., F.M., D.R.A., and H.K.; supervision, F.M., S.M., D.R.A., and H.K.; funding acquisition, D.R.A. and H.K. All authors read and agreed to the published version of the manuscript.

Funding: This research was fully supported by internal grants from Sime Darby Plantation, Malaysia.

Acknowledgments: We would like to acknowledge the contribution of Oil Palm Breeding Section, Sime Darby Plantation R&D Centre for conducting the breeding trials and providing the oil palm materials. We also would like to thank the Molecular Breeding and Bioinformatics team for enhancement of figures in the manuscript. We would like to thank the University of Nottingham Malaysia and Yayasan Sime Darby for the PhD fee scholarship to A.-L.O. and C.-K.T.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Christenhusz, M.J.M.; Byng, J.W. The number of known plants species in the world and its annual increase. *Phytotaxa* **2016**, *261*, 201–217. [CrossRef]
2. Hardon, J.J. Interspecific hybrids in the genus Elaeis II. vegetative growth and yield of F1 hybrids E. guineensis x E. oleifera. *Euphytica* **1969**, *18*, 380–388. [CrossRef]
3. Basiron, Y.; Balu, N.; Chandramohan, D. Palm oil: The driving force of world oils and fats economy. *Oil Palm Ind. Econ. J.* **2004**, *4*, 1–10.
4. Prades, A.; Salum, U.N.; Pioch, D. New era for the coconut sector. What prospects for research? *OCL Oilseeds Fats Crop. Lipids* **2016**, *23*. [CrossRef]
5. LLP, A.A. Coconut Products Market by Type Application and Form: Global Opportunity Analysis and Industry Forecast, 2019–2026. 2019. Available online: https://www.globenewswire.com/news-release/2020/02/21/1988574/0/en/Coconut-Products-Market-Study-2019-2026-World-Market-Projected-to-Cross-31-Billion-by-2026.html (accessed on 22 January 2020).
6. Bourgis, F.; Kilaru, A.; Cao, X.; Ngando-Ebongue, G.F.; Drira, N.; Ohlrogge, J.B.; Arondel, V. Comparative transcriptome and metabolite analysis of oil palm and date palm mesocarp that differ dramatically in carbon partitioning. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 12527–12532. [CrossRef] [PubMed]
7. Munier, P. Le palmier-dattier. In *Munier 1973 Palmier*; Maisonneuve&Larose: Paris, France, 1973; Volume 24.
8. Al-Yahyai, R.; Manickavasagan, A. *Dates: Production, Processing, Food, and Medicinal Values*; CRC Press: Boca Raton, FL, USA, 2012. [CrossRef]
9. Shahbandeh, M. Global Date Palm Industry Value 2014–2023. Available online: https://www.statista.com/statistics/960213/date-palm-market-value-worldwide/ (accessed on 22 January 2020).
10. Mayes, S.; Jack, P.L.; Corley, R.H.; Marshall, D.F. Construction of a RFLP genetic linkage map for oil palm (*Elaeis guineensis* Jacq.). *Genome* **1997**, *40*, 116–122. [CrossRef]
11. Batugal, P.; Bourdeix, R.; Baudouin, L. Coconut breeding. In *Breeding Plantation Tree Crops: Tropical Species*; Springer: Berlin, Germany, 2009; pp. 327–375. ISBN 9780387712031.
12. Batugal, P.; Bourdeix, R. Conventional coconut breeding. *Coconut Genet. Resour.* **2005**, 251.
13. Hadrami, I.E.; Hadrami, A. El; Hadrami, A. El Breeding date palm. In *Breeding Plantation Tree Crops: Tropical Species*; Springer: New York, NY, USA, 2009; pp. 191–216. ISBN 9780387712031.
14. Hardon, J.J. Oil palm breeding, Introduction. *Oil Palm Breed. Introd.* **1976**, 98–107.
15. Rance, K.A.; Mayes, S.; Price, Z.; Jack, P.L.; Corley, R.H.V. Quantitative trait loci for yield components in oil palm (*Elaeis guineensis* Jacq.). *Theor. Appl. Genet.* **2001**, *103*, 1302–1310. [CrossRef]
16. Moretzsohn, M.C.; Nunes, C.D.M.; Ferreira, M.E.; Grattapaglia, D. RAPD linkage mapping of the shell thickness locus in oil palm (*Elaeis guineensis* jacq.). *Theor. Appl. Genet.* **2000**, *100*, 63–70. [CrossRef]
17. Ting, N.C.; Jansen, J.; Mayes, S.; Massawe, F.; Sambanthamurthi, R.; Ooi, L.C.L.; Chin, C.W.; Arulandoo, X.; Seng, T.Y.; Alwee, S.S.R.S.; et al. High density SNP and SSR-based genetic maps of two independent oil palm hybrids. *BMC Genom.* **2014**, *15*. [CrossRef] [PubMed]
18. Gan, S.T.; Wong, W.C.; Wong, C.K.; Soh, A.C.; Kilian, A.; Low, E.T.L.; Massawe, F.; Mayes, S. High density SNP and DArT-based genetic linkage maps of two closely related oil palm populations. *J. Appl. Genet.* **2018**, *59*, 23–34. [CrossRef] [PubMed]
19. Billotte, N.; Marseillac, N.; Risterucci, A.M.; Adon, B.; Brottier, P.; Baurens, F.C.; Singh, R.; Herrán, A.; Asmady, H.; Billot, C.; et al. Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* Jacq.). *Theor. Appl. Genet.* **2005**, *110*, 754–765. [CrossRef]
20. Seng, T.Y.; Saad, S.H.; Chin, C.W.; Ting, N.C.; Singh, R.S.H.; Zaman, F.; Tan, S.G.; Alwee, S.S.R. Genetic linkage map of a high yielding FELDA Deli×Yangambi oil palm cross. *PLoS ONE* **2011**, *6*. [CrossRef] [PubMed]
21. Lee, M.; Xia, J.H.; Zou, Z.; Ye, J.; Rahmadsyah; Alfiko, Y.; Jin, J.; Lieando, J.V.; Purnamasari, M.I.; Lim, C.H.; et al. A consensus linkage map of oil palm and a major QTL for stem height. *Sci. Rep.* **2015**, *5*, 8232. [CrossRef] [PubMed]
22. Singh, R.; Ong-Abdullah, M.; Low, E.-T.L.; Manaf, M.A.A.; Rosli, R.; Nookiah, R.; Ooi, L.C.-L.; Ooi, S.; Chan, K.-L.; Halim, M.A.; et al. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature* **2013**, *500*, 335–339. [CrossRef]

23. Teh, C.K.; Ong, A.L.; Kwong, Q.B.; Apparow, S.; Chew, F.T.; Mayes, S.; Mohamed, M.; Appleton, D.; Kulaveerasingam, H. Genome-wide association study identifies three key loci for high mesocarp oil content in perennial crop oil palm. *Sci. Rep.* **2016**, *6*. [CrossRef]
24. Kwong, Q.B.; Teh, C.K.; Ong, A.L.; Heng, H.Y.; Lee, H.L.; Mohamed, M.; Low, J.Z.B.; Apparow, S.; Chew, F.T.; Mayes, S.; et al. Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm. *Mol. Plant* **2016**, *9*, 1132–1141. [CrossRef]
25. Hazzouri, K.M.; Gros-Balthazard, M.; Flowers, J.M.; Copetti, D.; Lemansour, A.; Lebrun, M.; Masmoudi, K.; Ferrand, S.; Dhar, M.I.; Fresquez, Z.A.; et al. Genome-wide association mapping of date palm fruit traits. *Nat. Commun.* **2019**, *10*, 4680. [CrossRef]
26. Xiao, Y.; Xu, P.; Fan, H.; Baudouin, L.; Xia, W.; Bocs, S.; Xu, J.; Li, Q.; Guo, A.; Zhou, L.; et al. The genome draft of coconut (*Cocos nucifera*). *Gigascience* **2017**, *6*. [CrossRef]
27. Rival, A.; Beule, T.; Barre, P.; Hamon, S.; Duval, Y.; Noirot, M. Comparative flow cytometric estimation of nuclear DNA content in oil palm (*Elaeis guineensis* jacq) tissue cultures and seed-derived plants. *Plant Cell Rep.* **1997**, *16*, 884–887. [CrossRef] [PubMed]
28. Pellicer, J.; Leitch, I.J. The Plant DNA C-values database (release 7.1): An updated online repository of plant genome size data for comparative studies. *New Phytol.* **2020**, *226*, 301–305. [CrossRef]
29. Ong, A.L.; Teh, C.K.; Kwong, Q.B.; Tangaya, P.; Appleton, D.R.; Massawe, F.; Mayes, S. Linkage-based genome assembly improvement of oil palm (*Elaeis guineensis*). *Sci. Rep.* **2019**, *9*, 6619. [CrossRef]
30. Kianian, P.M.A.; Wang, M.; Simons, K.; Ghavami, F.; He, Y.; Dukowic-Schulze, S.; Sundararajan, A.; Sun, Q.; Pillardy, J.; Mudge, J.; et al. High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nat. Commun.* **2018**, *9*, 2370. [CrossRef]
31. He, Z.; Zhang, Z.; Guo, W.; Zhang, Y.; Zhou, R.; Shi, S. De Novo Assembly of Coding Sequences of the Mangrove Palm (*Nypa fruticans*) Using RNA-Seq and Discovery of Whole-Genome Duplications in the Ancestor of Palms. *PLoS ONE* **2015**, *10*, e0145385. [CrossRef] [PubMed]
32. Singh, R.; Low, E.T.L.; Ooi, L.C.L.; Ong-Abdullah, M.; Ting, N.C.; Nagappan, J.; Nookiah, R.; Amiruddin, M.D.; Rosli, R.; Manaf, M.A.A.; et al. The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK. *Nature* **2013**, *500*, 340–344. [CrossRef] [PubMed]
33. Torres, M.F.; Mathew, L.S.; Ahmed, I.; Al-Azwani, I.K.; Krueger, R.; Rivera-Nuñez, D.; Mohamoud, Y.A.; Clark, A.G.; Suhre, K.; Malek, J.A. Genus-wide sequencing supports a two-locus model for sex-determination in Phoenix. *Nat. Commun.* **2018**, *9*, 3969. [CrossRef] [PubMed]
34. Al-Dous, E.K.; George, B.; Al-Mahmoud, M.E.; Al-Jaber, M.Y.; Wang, H.; Salameh, Y.M.; Al-Azwani, E.K.; Chaluvadi, S.; Pontaroli, A.C.; Debarry, J.; et al. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat. Biotechnol.* **2011**, *29*, 521. [CrossRef] [PubMed]
35. Mathew, L.S.; Spannagl, M.; Al-Malki, A.; George, B.; Torres, M.F.; Al-Dous, E.K.; Al-Azwani, E.K.; Hussein, E.; Mathew, S.; Mayer, K.F.X.; et al. A first genetic map of date palm (*Phoenix dactylifera*) reveals long-range genome structure conservation in the palms. *BMC Genom.* **2014**, *15*, 285. [CrossRef]
36. Pamin, K. *A Hundred and Fifty Years of Oil Palm Development in Indonesia from the Bogor Botanical Garden to the Industry*; Medan IOPRI: Medan Maimun, Kota Medan, Indonesia, 1998.
37. Rajanaidu, N.; Kushairi, A.; Rafii, M.; Mohd, D.A.; Maizura, I.; Jalani, B.S. Oil palm breeding and genetic resources. *Adv. Oil Palm Res.* **2000**, *1*, 171–237.
38. Shen, C.; Li, X.; Zhang, R.; Lin, Z. Genome-wide recombination rate variation in a recombination map of cotton. *PLoS ONE* **2017**, *12*. [CrossRef] [PubMed]
39. Milo, R.; Phillips, R. *Cell Biology by the Numbers*; Garland Science: New York, NY, USA, 2016; pp. 334–338. [CrossRef]
40. Sidhu, D.; Gill, K.S. Distribution of genes and recombination in wheat and other eukaryotes. *Plant Cell. Tissue Organ Cult.* **2005**, *79*, 257–270. [CrossRef]
41. Jordan, K.W.; Wang, S.; He, F.; Chao, S.; Lun, Y.; Paux, E.; Sourdille, P.; Sherman, J.; Akhunova, A.; Blake, N.K.; et al. The genetic architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. *Plant J.* **2018**, *95*, 1039–1054. [CrossRef]
42. Teh, C.K.; Muaz, S.D.; Tangaya, P.; Fong, P.Y.; Ong, A.L.; Mayes, S.; Chew, F.T.; Kulaveerasingam, H.; Appleton, D. Characterizing haploinsufficiency of SHELL gene to improve fruit form prediction in introgressive hybrids of oil palm. *Sci. Rep.* **2017**, *7*, 3118. [CrossRef] [PubMed]
43. Corley, R.H.V.; Tinker, P.B. *The Oil Palm: Fifth Edition*; John Wiley&Sons: Hoboken, NJ, USA, 2015.

44. Hardon, J.J.; Corley, R.H.V.; Lee, C.H. Breeding and Selecting the Oil Palm. In *Improving Vegetatively Propagated Crops*; Academic Press: London, UK, 1987.
45. Bellow, S.A.; Agunsoye, J.O.; Adebisi, J.A.; Kolawole, F.O.; Hassan, S.B. Physical properties of coconut shell nanoparticles. *Kathmandu Univ. J. Sci. Eng. Technol.* **2018**, *12*, 63–79. [CrossRef]
46. Budi, E.; Umiatin, U.; Nasbey, H.; Bintoro, R.A.; Wulandari, F.; Erlina, E. Adsorption and Pore of Physical-Chemical Activated Coconut Shell Charcoal Carbon. In *IOP Conference Series: Materials Science and Engineering*; Management Science and Engineering: Stanford, CA, USA, 2018.
47. Mohd Iqbaldin, M.N.; Khudzir, I.; Mohd Azlan, M.I.; Zaidi, A.G.; Surani, B.; Zubri, Z. Properties of coconut shell activated carbon. *J. Trop. For. Sci.* **2013**, *25*, 497–503.
48. Tang, H.; Bowers, J.E.; Wang, X.; Ming, R.; Alam, M.; Paterson, A.H. Synteny and collinearity in plant genomes. *Science* **2008**, *320*, 486–488. [CrossRef]
49. Corley, R.H.V. Oil palm physiology-a review. In Proceedings of the International Oil Palm Conference, Kuala Lumpur, Malaysia, 16–18 November 1973.
50. Couvreur, T.L.P.; Forest, F.; Baker, W.J. Origin and global diversification patterns of tropical rain forests: Inferences from a complete genus-level phylogeny of palms. *BMC Biol.* **2011**, *9*, 44. [CrossRef]
51. Barrett, C.F.; Bacon, C.D.; Antonelli, A.; Cano, Á.; Hofmann, T. An introduction to plant phylogenomics with a focus on palms. *Bot. J. Linn. Soc.* **2016**, *182*, 234–255. [CrossRef]
52. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef]
53. Rastas, P.; Paulin, L.; Hanski, I.; Lehtonen, R.; Auvinen, P.; Brudno, M. Lep-MAP: Fast and accurate linkage map construction for large SNP datasets. *Bioinformatics* **2013**, *29*, 3128–3134. [CrossRef] [PubMed]
54. Rastas, P. Lep-MAP3: Robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* **2017**, *33*, 3726–3732. [CrossRef] [PubMed]
55. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
56. Tang, H.; Zhang, X.; Miao, C.; Zhang, J.; Ming, R.; Schnable, J.C.; Schnable, P.S.; Lyons, E.; Lu, J. ALLMAPS: Robust scaffold ordering based on multiple maps. *Genome Biol.* **2015**, *16*, 3. [CrossRef]
57. Al-Mssallem, I.S.; Hu, S.; Zhang, X.; Lin, Q.; Liu, W.; Tan, J.; Yu, X.; Liu, J.; Pan, L.; Zhang, T.; et al. Genome sequence of the date palm Phoenix dactylifera L. *Nat. Commun.* **2013**, *4*, 2274. [CrossRef]
58. Kurtz, S.; Phillippy, A.; Delcher, A.L.; Smoot, M.; Shumway, M.; Antonescu, C.; Salzberg, S.L. Versatile and open software for comparing large genomes. *Genome Biol.* **2004**, *5*, R12. [CrossRef] [PubMed]
59. Marçais, G.; Delcher, A.L.; Phillippy, A.M.; Coston, R.; Salzberg, S.L.; Zimin, A. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **2018**, *14*, e1005944. [CrossRef]
60. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: An information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [CrossRef]
61. Clustalw, U.; To, C.; Multiple, D.O. ClustalW and ClustalX. *Options* **2003**, 1–22. [CrossRef]
62. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; Mcgettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948. [CrossRef]
63. Rambaut, A. Fig Tree v. 1.4.4. 2018. Available online: http://tree.bio.ed.ac.uk/software/figtree/ (accessed on 12 June 2020).

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).