High Density SNP and DArT-based Genetic Linkage Maps of Two Closely Related Oil Palm Populations

Siou Ting Gan[1,2], Wei Chee Wong[2], Choo Kien Wong[2], Aik Chin Soh[1,2,5], Andrzej Kilian[3], Eng-Ti Leslie Low[4], Festo Massawe[1] and Sean Mayes[1,5,6]

[1]Biotechnology Research Centre, School of Biosciences, The University of Nottingham Malaysia Campus, 43500 Semenyih, Selangor Darul Ehsan, Malaysia

[2]Biotechnology Research Centre, Advanced Agriecological Research Sdn Bhd, 43500 Semenyih, Selangor Darul Ehsan, Malaysia

[3]Diversity Arrays Technology, Bldg 3, Lv D, University of Canberra, Kirinari st., Bruce, ACT 2617, Australia

[4]Malaysia Palm Oil Board, 6, Persiaran Institusi, Bandar Baru Bangi, 4300 Kajang, Selangor Darul Ehsan, Malaysia

[5]Crops For the Future, Jalan Broga, 43500 Semenyih, Selangor Darul Ehsan, Malaysia

[6]Plant and Crop Sciences, Biosciences, The University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire, LE12 5RD, United Kingdom

Corresponding author

Siou Ting Gan

Email: ganst@aarsb.com.my

Tel: (+603) 8727 3275

Fax: (+603) 6151 7081

Abstract

Oil palm (*Elaeis guineensis* Jacq.) is an outbreeding perennial tree crop with long breeding cycles, typically 12 years. Molecular marker technologies can greatly improve the breeding efficiency of oil palm. This study reports the first use of the DArTseq platform to genotype two closely related self-pollinated oil palm populations, namely AA0768 and AA0769 with 48 and 58 progeny respectively. Genetic maps were constructed using the DArT and SNP markers generated, in combination with anchor SSR markers. Both maps consisted of 16 major independent linkage groups ($2n = 2x = 32$) with 1399 and 1466 mapped markers for the AA0768 and AA0769 populations, respectively, including the morphological trait "Shell-thickness" (*Sh*). The map lengths were 1873.7, and 1720.6 cM with an average marker density of 1.34 and 1.17 cM, respectively. The integrated map was 1803.1 cM long with 2066 mapped markers and average marker density of 0.87 cM. A total of 82% of the DArTseq marker sequence tags identified a single site in the published genome sequence, suggesting preferential targeting of gene-rich regions by DArTseq markers. Map integration of higher density focused around the *Sh* region identified closely linked markers to the *Sh*, with D.15322 marker 0.24 cM away from the morphological trait and 5071 bp from the transcriptional start of the published *SHELL* gene. Identification of the *Sh* marker demonstrates the robustness of using the DArTseq platform to generate high density genetic maps of oil palm with good genome coverage. Both genetic maps and integrated map will be useful for quantitative trait loci analysis of important yield traits as well as potentially assisting the anchoring of genetic maps to genomic sequences.

Keywords Elaeis, DArTseq, Genetic linkage map, Integrated map, Shell thickness

Introduction

Oil Palm (*Elaeis guineensis* Jacq) is the leading oil crop in the world with production of 64.5 million tonnes of palm oil, constituting 34.7% of the world's production of major vegetable oils in the year 2016/2017 (Foreign Agricultural Service, United States Department of Agriculture). There are two species of oil palm, the African *E. guineensis* and the South American *E. oleifera*. The most cultivated species is *E. guineensis*, due to its high oil yield (Hardon, 1969). Even though the species originated from Africa, it is mostly cultivated in Southeast Asia, particularly Indonesia and Malaysia, which together account for more than 50% of the total oil palm plantation area in the world (FAS USDA).

Oil palm is a perennial, cross-pollinated, monocotyledonous, diploid ($2n = 2x = 32$) species with a genome size of 1.8 billion base pairs (Singh et al. 2013b). Conventional breeding of oil palm requires 12 years (Mayes et al. 2008) and large planting areas due to the large size of the palms, with standard planting density of 148 palms per hectare (usually a triangularly-spaced planting design with 9 m between palms). In breeding, smaller plot sizes of 10-20 palms planted in 3-6 replicates are commonly used for parental and progeny combining ability evaluations (Soh et al. 1990). Typically, evaluation of 25 parental palms through progeny testcrossing in 4 replications requires 11 hectare of planting area.

Based on the major single Mendelian gene, *SHELL*, oil palm fruits can be divided into three fruit types: *dura* (D), *pisifera* (P) and *tenera* (T). Crossing of thick-shelled *dura* (*Sh*/*Sh*) with the shell-less *pisifera* (*sh*/*sh*) results in 100% of the thin-shelled hybrid *tenera* (*Sh*/*sh*). *Tenera* has a higher mesocarp-to-fruit ratio than the *dura* fruit form and this translates into higher oil yields. Genetic studies have revealed that the shell gene exhibits co-dominant monogenic inheritance that is exploitable in breeding programmes (Beirnaert and Vanderweyen 1941; Singh et al. 2013a). Currently, the product of D x P is almost exclusively used for commercial production with the Deli *dura* x AVROS *pisifera* as the most common commercial hybrid variety (Soh et al. 2006).

Genetic linkage maps are fundamental tools for identification of major genes and quantitative trait loci (QTL) controlling agronomically important traits. Such markers can be directly used in marker-assisted breeding programmes and could facilitate map-based positional cloning. Alongside the rapid development of marker technology, numerous efforts have been made to generate oil palm genetic maps. The first linkage map was constructed from 97 co-dominant RFLP loci which gave 24 linkage groups (LGs) (Mayes et al. 1997). Despite the reliability of RFLP markers, it is very tedious and costly to develop maps based on them; hence later genetic maps have utilised PCR-based molecular markers. The second genetic map of oil palm was constructed using RAPD markers (Moretzsohn et al. 2000). The first high density linkage map of oil palm was created by Billotte et al. (2005) using SSR and AFLP markers. Since then SSR and AFLP markers have been used extensively for the construction of oil palm genetic linkage maps (Singh et al. 2009; Billotte et al. 2010; Seng et al. 2011; Montoya et al. 2013, 2014; Ting et al. 2013, 2014; Jeennor and Volkaert, 2014). With the advance of next-generation sequencing technology, whole-genome profiling using SNP markers has been shown to have great potential for generating high density genetic maps (Sim et al. 2012; Song et al. 2013) that could further facilitate identification

of tightly linked QTL markers for molecular breeding (Moriguchi et al. 2012; Liu et al. 2013; Kujur et al. 2015; Wang et al. 2015). In oil palm, Pootakham et al. (2015) reported the construction of a SNP-based genetic linkage map and the identification of QTL associated with plant height and fruit bunch weight.

Classical Diversity Arrays Technology (DArT) has been successfully applied for various studies (Aitken et al. 2014; Yu et al. 2014; Tadesse et al. 2015; Novoselović et al. 2016; Śliwka et al. 2016) and is currently available for over 130 different organisms, including oil palm, with the array being established using 284 genotypes derived from Malaysian breeding programmes (www.diversityarrays.com). The complexity reduction approach has now been combined with next generation sequencing (NGS) technologies, to generate a method generally termed Genotyping-by-Sequencing (GbS). DArT genotyping-by-sequencing (DArTseq) is a marker platform in which the DArT platform is coupled with Illumina short read sequencing to generate two types of data, dominant silico-DArT markers and co-dominant SNP markers (Sansaloni et al. 2011; Cruz et al. 2013; Ren et al. 2015; Sánchez-sevilla et al. 2015). Silico-DArT markers are scored as "presence/absence" (dominant) markers, while SNP are sequence variants present within the 64 bp sequence tag associated with each marker.

The present study generated high-density genetic linkage maps of two closely related *tenera* self-pollinated oil palm populations using the DArTseq platform, with SSR markers used as anchor loci (Billotte et al. 2005, 2010). To address the issue of small population size, this study also integrated both genetic maps for future analysis of important yield traits in oil palm with higher accuracy.

Materials and methods

Plant materials and DNA isolation

This study used two closely related *tenera* self-pollinated populations, namely AA0768 and AA0769, planted at the Advanced Agriecological Research Sdn. Bhd (AAR) oil palm breeding research station in Paloh Estate, Johore, Malaysia. The F_1 population, AA0228, was generated from a *tenera* (female grandparent) x *pisifera* (male grandparent) cross of Binga x Yangambi-AVROS. Two full-sibs *tenera* palms, AA0228/05 and AA0228/06, were selected and self-pollinated to generate the F_2 populations AA0768 and AA0769, respectively. The mapping populations AA0768 and AA0769 consisted of 48 and 58 progeny, respectively. The fruit type of each progeny was determined phenotypically and used as a morphological marker to allow the location of the *Sh* to be mapped. Genomic DNA was extracted from fresh frond No. 1 leaflets, the youngest fully-opened frond, using the NucleoSpin® Plant II kit according to manufacturer's instruction (Macherey-Nagel, Germany).

Microsatellite genotyping

In this study, an initial set of CIRAD SSR markers with known locations (Billotte et al. 2005, 2010) was pre-screened for amplification efficiency and polymorphism on both the *tenera* mapping parents. The aim was to map at least two polymorphic SSR markers from each chromosome arm as anchoring markers. The PCR amplification was performed in a 20 µl reaction mixture containing 20 ng DNA, 3U *Taq* DNA polymerase (NEB, USA), 1x PCR buffer, 0.8 mM dNTPs mix, 0.02 µM M13-tagged forward primer, 0.2 µM reverse primer, 0.18 µM fluorescent dye-labelled M13 primer. PCR reactions were performed with an initial denaturation of 94 °C for 3

minutes, followed by 35 cycles of denaturation at 94 °C for 1 min., the appropriate annealing temperatures of 50, 53, 56, 59, 62 or 65 °C for 1 min., extension at 72 °C for 2 min., with a final extension of 72 °C for 10 min. Prior to fragment size analysis, pooled PCR products were mixed with 25 µl sample loading solution and size standard mix (1:100 v/v) (Beckman Coulter Inc, Fullerton, USA). The mixture was subjected to capillary electrophoresis on a Beckman CEQ 8000 Genetic Analyzer (Beckman coulter Inc, USA). The fragment sizes of genotypes were scored using the CEQ™ 8000 Fragments Analysis Software Version 8.

DArTseq genotyping

Extracted DNA was sent to Diversity Arrays Technology Pty Ltd, Yarralumla, Australia, for DArTseq™ genotyping using restriction enzyme combinations, as previously reported by Sansaloni et al. (2011) and Cruz et al. (2013). In case of oil palm, the enzyme combination involved PstI and HhaI. A proprietary analytical pipeline, DArTsoft14, was applied to extract marker data from sequenced genomic representations/libraries. The locus designations used by Diversity Array Technology Ltd for "Silico-DArT" (presence/absence) and "SNP" generated were adopted with modification in this study. Prefix "D." or "S." was added to indicate Silico-DArT or SNP, respectively.

Data analysis

The percentage of missing data and allele ratios of DArTseq markers (DArT and SNP) in both the AA0768 and AA0769 populations were calculated. The rate of missing data is the ratio of individuals with missing data to the total number of individuals in the population, while the allele ratio was calculated as the segregation ratio of individual alleles in the population. Subsequently, DArT and SNP markers were selected for linkage mapping based on the following criteria: firstly, markers with less than or equal to 5% missing data were selected; secondly, markers with allele ratio of 0.15-0.85 were selected. Lastly, the genotyping data of the *tenera* parents, AA0228/05 and AA0228/06, was used as a quality control in which inconsistent results between expected segregation patterns based on the parental scores and the observed population scores were eliminated from the dataset. Chi-square analysis was performed for all segregating DArTseq and SSR markers to assess the goodness-of-fit between observed and expected Mendelian segregation ratio. Markers showing very significant distortion ($p<0.0005$) were excluded for further analysis.

Construction of genetic linkage map

JoinMap 4.1 Software (Van Ooijen, 2006) was used to construct the genetic maps for the two F_2 segregating populations. However, the *tenera* and *pisifera* grandparents of both populations no longer exist in the field, so there is no parental data available for phase determination. Because of this, both self-pollinated populations were first analyzed as a Cross Pollinator (CP) for linkage analysis using segregation coding of *hkxhk*. The phase of markers was determined by the software within each LG. The genotyping data was then converted back for analysis as a "F_2" population type using genotype code of *a* (homozygote as the first parent), *b* (homozygote as the second parent, *h* (heterozygote as the F_1), *c* (not genotype a) and *d* (not genotype b). LGs of the F_2 populations were established using an independence logarithm-of-the-odds (LOD) threshold of 4 and they were assigned to

chromosome based on the known location of the SSR markers. Fragmented LGs that belong to the same chromosome were combined. The locus order of markers was established using the regression mapping algorithm at the default value of recombination frequency ≤ 0.4; LOD score ≥ 1; goodness-of-fit jump threshold = 5, ripple value = 1. Haldane's mapping function was used to convert recombination frequencies into map distances in units of centimorgan (cM).

Integration of both genetic maps and shell-thickness (*Sh*) region

Map integration for both genetic maps was performed using the JoinMap 4.1 software. Map calculation was based on mean recombination frequencies and combined LOD scores for each pair of markers in individual maps and regression mapping algorithm was used. Genetic grouping was repeated with all available DArTseq markers around the *Sh* region to obtain high density maps for both AA0768 and AA0769 populations followed by map integration around the *Sh* region.

Alignment of DArTseq markers against the oil palm genome assembly

Homology search using the 64 bp sequence tag associated with each DArTseq marker was performed against the *Elaeis guineensis* genome assembly deposited in GenBank (accession ID: GCA_000442705.1) using BLASTN at a significance threshold of $1e^{-10}$. The final chromosome assignment and orientation of both genetic maps and integrated map were based on genome sequence.

Result

Molecular markers and genotyping

A total of 11675 DArTseq markers, consisting of 6764 DArT and 4911 SNP, were generated from genotyping of the AA0768 and AA0769 mapping populations. The call rate for DArT markers ranged from 0.73-1 with an average of 0.90. Better call rates were attained for SNP markers, ranging from 0.75-1 with a mean of 0.95. The call rate of the *tenera* parents was 82.4% and 98.4% for DArT and SNP markers, respectively. Call rate essentially reflects the percentage of missing data tolerated, indicating that the DArT and SNP markers generated were generally of good quality and exhibited high levels of polymorphism. In total, based on the selection criteria listed, 948 and 958 DArT markers and 719 and 729 SNP markers were chosen from the AA0768 and AA0769 populations, respectively, for subsequent linkage mapping analysis. A total of 102 CIRAD SSR markers were screened using the two mapping populations, out of which 36 polymorphic markers were identified.

The majority of loci in both the AA0768 and AA0769 mapping populations segregated in the expected Mendelian ratios of 1:2:1 for SSR and SNP markers or 1:3 for DArT markers. One hundred and sixty three (9.6% of the total; 6 SSR, 104 DArT and 53 SNP) and 190 (11% of the total; 10 SSR, 99 DArT and 81 SNP) markers were significantly distorted at the 5% significance level for AA0768 and AA0769 populations, respectively (Table 1). Only markers showing very significant distortion ($p<0.0005$) were excluded from further mapping analysis, 11 (3 SSR, 2 DArT and 6 SNP) and 20 (4 SSR, 5 DArT and 11 SNP) markers for the AA0768 and AA0769 controlled

crosses, respectively. No significant deviation was found from the 1:2:1 segregation ratio expected for *dura*: *tenera*: *pisifera* within each cross for the *Sh* major Mendelian gene.

Genetic linkage map construction

Phase determination and linkage analysis was performed using 33 and 32 SSR markers, 946 and 953 DArT markers, 713 and 718 SNP markers, and the *Sh* gene as a morphological marker for AA0768 and AA0769 populations, respectively. Initial marker grouping using a LOD threshold of 4 produced 21 and 17 LGs for the AA0768 and AA0769 populations, respectively. Detailed inspection of the LGs generated by CP and F_2 analyses revealed that the grouping of markers into LGs was exactly the same for both analyses, indicating that the marker conversion step from CP to F_2 was not effecting mapping, as predicted. LGs were initially assigned to chromosomes by using the known location of anchoring SSR markers in the genetic map published by Billotte *et al.* (2010), with the exception of two LGs where no polymorphic SSRs markers were found, leading to these two groups being unassigned. Separate analysis of several LGs that appeared to be fragmented as indicated by the known location SSRs showed that regrouping of the groups could be achieved using lower LODs of 2.5 – 3.9. These LGs were combined together into 16 independent LGs, which corresponded well to the 16 homologous chromosome pairs of oil palm. The final assignment and orientation of LGs were based on oil palm genome sequence assembly (Singh et al., 2013b). LG 2 of AA0768 and LG 4 of AA0769 were analyzed as separate groups of LG 2A and 2B as well as LG 4A and 4B, as map generation failed to combine parts A and B, despite these LG 2A, 2B and LG 4A, 4B being grouped at LOD of 3 and 2.9, respectively.

The genetic maps of AA0768 and AA0769 populations consisted of 1399 (32 SSR, 772 DArT and 594 SNP) and 1466 (32 SSR, 807 DArT and 626 SNP) markers, respectively, including the morphological trait *Sh* (Tables 2 and 3, Online Resources 1, 2 and 4). The total map lengths were 1873.7 and 1720.6 cM with an average marker density of 1.34 and 1.17 cM for AA0768 and AA0769, respectively. Map distances between two consecutive markers varied from 0 to 20 cM for both populations, with only 4.1% (57 out of 1382) and 2.5% (36 out of 1449) intervals greater than 5 cM for the AA0768 and AA0769 populations, respectively. There were a total of 92 and 133 segregation distorted markers mapped into the final genetic map of AA0768 and AA0769, respectively, making up 6.6% and 9.1% of total mapped loci. Both the genetic maps had a total of 717 common markers, consisting of 29 SSR, 196 DArT and 491 SNP.

Integration of genetic maps

The integrated map was 1803.1 cM long and comprised of 2066 markers (35 SSR, 1340 DArT, 690 SNP and *Sh* morphological trait) (Table 4, Online Resources 3 and 4). The number of markers mapped in each LG ranged from 17 to 280, with an average marker density of 0.87 cM, higher density than the individual genetic maps of AA0768 and AA0769 populations. LG 2 was the longest with 280 markers spanning 206.99 cM while LG 16 was the shortest group with only 17 markers and 38.09 cM long. The integrated map had only 1.1% (22 out of 2049) of intervals between two consecutive markers greater than 5 cM, significantly smaller gaps than those reported in AA0768 and AA0769 genetic maps. The number of markers common to both AA0768 and AA0769 populations

ranged between 4 to 121 markers and LG 15 had the highest percentage of common markers, 54.8% (51 out of 93).

Alignment of DArTseq markers to the *E. guineensis* genome sequence assembly

Homology search of the 2307 DArTseq markers mapped in both populations and integrated map against the MPOB *pisifera* genome assembly P5 revealed that despite the short sequence of the marker tags (64 bp), significant homology (E-value $\leqslant 10^{-25}$) were obtained for 2113 (91.6%) markers, while 167 DArTseq markers had no hit. Of the 2113 DArTseq markers, 1654 aligned to the 16 chromosomes while the remaining 459 markers aligned to the additional 40044 small unanchored scaffolds. A total of 1960 markers with significant homology had only a single hit with no sub-alignment score.

Closer inspection of the alignment of the AA0768 and AA0769 genetic linkage maps to the oil palm genome sequence assembly indicates that the overall arrangement of markers on the high density linkage maps was broadly consistent with the genome sequence order, but with considerable local inconsistency. On the other hand, the integrated map showed improved co-linearity of marker order with the genome sequence assembly. The integrated map had a total of 1473 (72.6%) DArTseq markers aligned to the 16 chromosomes, another 405 (20%) DArTseq markers aligned to the small unanchored scaffolds and 152 (7.5%) DArTseq markers had no hit (Table 5). It is interesting to note that LG 4A had 43.4% (33 out of 76) of DArTseq markers aligned to the chromosome and 42.1% (32 out of 76) of markers aligned to the unanchored scaffolds.

In the initial stage of genetic maps construction, LG 4B was treated as LG 6B and grouped with LG 6A at LOD of 2.5 and 4.5 for AA0768 and AA0769 populations. However, sequence analysis of DArTseq markers revealed that these LG 6B markers were wrongly assigned, they belonged to one arm of LG 4. Sixty-six markers (2.9%) were also found to be mapped to the wrong LG based on their sequence hit. The availability of the oil palm genome sequence allowed for the final assignment and orientation of LGs for both genetic maps of AA0768 and AA0769 populations as well as an integrated map to be completed.

Integration of genetic maps around *Sh* region

A higher density integrated map was constructed for the *Sh* region at LG 2 using all available DArTseq markers (Figure 1). Marker order in this higher density integrated map was different from that of the individual maps of the AA0768 and AA0769 populations and markers were more densely arranged on one side of the *Sh* region than the other. Homology search of the markers flanking the *Sh* region within 5 cM revealed that 19 out of 22 markers were located in the same scaffold p5_sc00060 in which *SHELL* gene was identified (Singh et al. 2013a) (Table 6). Closer inspection of the hit region of the DArTseq markers against p5_sc00060 and p5_sc00263 scaffolds revealed that the overall arrangement of markers on the integrated map was broadly consistent with the scaffold sequence order, but with considerable local inconsistency. Integration of genetic maps of AA0768 and AA0769 populations enabled identification of several closely liked *Sh* markers, with D.15322 the closest, only 0.24 cM away from the morphological trait gene. This D.15322 marker was found to be 5139 bp away from the

transcriptional start of the published *SHELL* gene (Singh et al. 2013a) through homology search with the MPOB *pisifera* genome assembly P5.

Discussion

The DArTseq platform is a high-throughput genome profiling method which combines the use of the classical DArT genome complexity reduction method with NGS to generate both dominant DArT markers and co-dominant SNP markers (Sansaloni et al. 2011; Cruz et al. 2013). Genotyping-by-sequencing (GbS) approaches such as DArTseq platform require no reference genome. The consensus of the read clusters across the sequence tagged sites becomes the reference for scoring of SNP markers (Elshire et al. 2011). As the SNPs are scored in a segregating population, they are partially validated markers, particularly if the segregation patterns permits mapping of the marker associated with the sequence tag. Generation of co-dominant SNP markers from DArTseq platform in the present study was accomplished prior to the publication of the oil palm genome sequence by Singh et al. (2013b) and the subsequent published genome had been used to evaluate the markers developed. Therefore the current study has proven that DArTseq platform is suitable for marker development followed by genetic mapping without relying on pre-existing sequence information of the species of interest.

The present study reports the first high density DArT- and SNP-based genetic maps for *E. guineensis* based on the DArTseq platform, with SSR markers from a public database (Billotte et al. 2010) as anchor loci. The average marker density of the individual maps is higher than those previous reported using non-SNP based markers (Billotte et al. 2005, 2010; Seng at al. 2011) as well as the SNP- and SSR-based oil palm genetic maps generated using a dedicated oil palm SNP array (Ting et al. 2014) with an average marker density of 1.4 cM for the DP integrated map. The integrated map presented in this study has further increased the average marker density to 0.87 cM, the higher marker density published for oil palm to date, higher than the GbS-based oil palm linkage map reported by Pootakham et al. (2015) with marker density of 1.26 cM.

Despite a greater overall marker density obtained in the present study, large gaps were observed between adjacent markers in both linkage maps of AA0768 and AA0769 populations, with intervals ranging from 11 to 20 cM. Large gaps were also reported in other oil palm mapping studies (Billotte et al. 2005, 2010; Seng et al. 2011; Ting et al. 2014). Regions of low marker density have previously been reported, even on the ultra-dense genetic linkage map with >10000 loci constructed from a heterozygous diploid potato population in which a gap spanning 14 and 20 cM was found on LG VIII of the maternal and paternal parental maps, respectively (van Os et al. 2006). Large gaps observed between loci could be due to homozygosity of the genome in that particular region or the non-uniform distribution of recombination events (Castiglioni et al. 1999; van Os et al. 2006). Large marker intervals are also likely to be due to the small F_2 population size used in the present study. This small population size effect was partially resolved when both genetic maps were integrated. Significant reduction of large gaps was observed in the integrated map with only 1.1% intervals ranging from 5 to 9.3 cM.

The genetic maps constructed using the AA0768 and AA0769 populations shared a number of similarities, such as grouping of markers in the same LGs, location of markers in terms of chromosomes position (telomere vs centromere), total and average map length as well as average marker density. This is expected due to the full-sib

background of their respective *tenera* parents. However, the linear marker order between the maps was not completely congruent. Map integration greatly improved the linearity of marker order but local inconsistencies were observed. Inconsistencies of marker order are commonly observed in plant species especially when individual maps are integrated into the consensus map and it is believed that this phenomenon is mainly due to differences in recombination frequencies of marker pairs in populations of different sizes and type, probably due to the stochastic nature of recombination, or it could be caused by local rearrangements or segmental duplications of the genome (Mace et al., 2009; Studer et al. 2010; Khan et al. 2012). Despite the relatively small sample sizes of progenies used, this study managed to produce 16 independent LGs with high genome coverage and marker density for both populations as well as the integrated map.

*Pst*I, the most commonly used restriction enzyme in the DArT assay for genome complexity reduction, is a CXG methylation-sensitive enzyme that cuts hypomethylated sequences which are often low-copy and occur primarily in gene-rich regions of the genome (Schouten et al. 2012). The high genome coverage and unambiguous alignment of 85% of DArTseq markers to unique positions in the genome despite being short sequences (64 bp) has proven that DArTseq markers generated in the present study based on the *Pst*I enzyme display a reasonably uniform distribution throughout the genome with preferential targeting of gene-rich regions, similar to those reported in the microarray DArT-based studies (Kullan et al. 2012; Petroli et al. 2012).

LG 2A and 2B in the AA0768 population as well as LG 4A and 4B in the AA0769 population were not mapped into the same LG, most likely due to the small population size or potentially due to a genetic effect, possibly a translocation or a gene-gene interaction effecting viability. More markers, particularly anchoring loci, were possibly needed for these particular chromosomal regions. Meanwhile, the revelation of wrong linkage of part of LG 4 with LG 6 in the initial genetic maps through genome sequence alignment suggested that it is either due to small population size effect or a consequence of the difficulty of separating large numbers of markers on the basis of a single LOD score, with interaction across groups or noise within datasets leading to the requirement for high LODs to fragment false LGs.

As a major step towards application of marker technology in oil palm breeding programme, the reported high density integrated map from two closely related, but relatively small, populations is useful for QTL mapping of important traits, such as bunch number, bunch weight, fresh fruit bunch yield, fruit characters, palm height and leaf area index. The accuracy of genetic maps is vital for fine mapping and for the isolation of genes for traits of interest. In the present study, mapping of all available DArTseq markers followed by map integration around the *Sh* region allowed for the identification of closely linked markers with higher accuracy for potential molecular marker-assisted selection.

Given that majority of the DArTseq markers unambiguously aligned to a unique position in the genome with substantial amount of them aligned to the unanchored scaffolds, the reported genetic maps offer the possibility of assigning unanchored scaffolds to the assembled pseudochromosomes of the published genome assembly (Singh et al. 2013b), although for genomes with substantial scaffold fragmentation, it would require large numbers of progeny to allow fine order mapping to be accurate. The reported high density genetic maps can be further improved. The information from both genetic and physical maps can be combined to correctly order the markers,

particularly those closely linked markers for fine mapping. Framework maps of the AA0768 and AA0769 populations can be established by selecting highly informative markers, particularly co-dominant SNP and SSR markers, which are common to both populations and the map position tested for concordance to the published *pisifera* genome assembly. A full genetic map can then be constructed by fixing the marker order of the framework map to allow addition of more DArTseq markers, without disturbing the best framework order of markers.

Besides QTL mapping, it is believed that the high density, low cost, limited technical requirement of DArTseq genotyping technology as reported here could be a potential approach for both Genomic Selection (GS) and Genome-wide Association Studies (GWAS) in oil palm. By using high density markers covering the whole genome, instead of few significant markers, GS selects favourable individuals based on genomic estimate of breeding value (GEBV) and is expected to address small effect genes that cannot be captured by traditional QTL mapping (Meuwissen et al. 2001; Hayes et al. 2009). Conventional microarray DArT markers have been successfully applied for GS of wheat (Crossa et al. 2010; Rutkoski et al. 2012), sugarcane (Guoy et al. 2013) and forest trees *Eucalyptus* (Grattapaglia et al. 2011; Resende et al. 2012).

Since the publication of a simulation study of GS on oil palm in 2008 (Wong and Bernado 2008), there have been limited reports on the application of GS and GWAS (Kwong et al. 2016; Teh et al. 2016) approaches in oil palm. The first empirical assessment of GS in oil palm breeding using less than 300 SSR markers was reported in year 2015 (Cros et al. 2015). The author commented that higher marker density can increase GS accuracy due to linkage disequilibrium, which could be achieved by DArTseq genotyping. It is believed that GS and GWAS are particularly suited for perennial tree crops with long generation times, such as oil palm, together with the availability of high throughput and low cost genotyping platforms, DArTseq as reported in present study.

The use of high throughput DArTseq genotyping platform and integration of two closely related populations with relatively small population size in this study have allowed high accuracy mapping and identification of useful markers for important yield trait, as reported here the closely linked D.15322 marker that is 5071 bp away from the transcription start of the *SHELL* gene. Conversion of the closely linked DArTseq marker(s) to a PCR-based format has the potential of marker(s) application in oil palm breeding programmes in which fruit type of oil palm seedlings can be distinguished early in the nursery stage, instead of waiting 2-3 years for fruiting, to select for the desired genotypes.

Acknowledgements This study was part of the Ph.D. studentship supported by University of Nottingham Malaysia Campus and it was funded by Advanced Agriecological Research Sdn. Bhd. We would like to thank our Principals, Messr. Kuala Lumpur Kepong Bhd. and Boustead Plantations Bhd. for their permission to publish the article.

Author contribution statement STG performed the experiments, analyzed the data and wrote the paper. CKW, WCW and ACS developed and contributed plant materials. AK performed the DArTseq genotyping. ELL performed the homology search using the *pisifera* genome assembly. FM and SM conceived and designed the study, reviewed the paper and supervised research.

Conflicts of interest The authors declare that they have no conflict of interest.

Reference

Aitken KS, McNeil MD, Hermann S, Bundock PC, Kilian A, Heller-Uszynska K, Henry RJ, Li J (2014) A comprehensive genetic map of sugarcane that provides enhanced map coverage and integrates high-throughput Diversity Arrays Technology (DArT) markers. BMC Genomics 15:152

Beirnaert A, Vanderweyen R (1941) Contribution à l'étude génétique et biométrique des variétés d'*Elaeis guineensis* Jacq. Publications de l'institut national pour l'étude agronomique du Congo Belge, Série Scientifique 27

Billotte N, Jourjon MF, Marseillac N, Berger A, Flori A, Asmady H, Adon B, Singh R, Nouy B, Potier F, Cheah SC, Rodhe W, Ritter E, Courtois B, Charrier A, Mangin B (2010) QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis* Jacq.). Theoretical and Applied Genetics 120:1673-1687

Billotte N, Marseillac N, Risterucci A-M, Adon B, Brottier P, Baurens F-C, Singh R, Herrán A, Asmady H, Billot C, Amblard P, Durand-Gasselin T, Courtois B, Asmono D, Cheah SC, Rohde W, Ritter E, Charrier A (2005) Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis*) Jacq. Theoretical and Applied Genetics 110:754-765

Castiglioni P, Ajmone-marsan P, van Wijk R, Motto M (1999) AFLP markers in a molecular linkage map of maize: co-dominant scoring and linkage group distribution. Theoretical and Applied Genetic 99:425-431

Cros D, Denis M, Sánchez L, Cochard B, Durand-gasselin T, Nouv B, Omoré A, Pomiès V, Riou V, Suryana E, Bouvet JM (2015) Genomic selection prediction accuracy in perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). Theoretical and Applied Genetic 128:397-410

Crossa J, de Los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713-724

Cruz VMV, Kilian A, Dierig DA (2013) Development of DArT marker platforms and genetic diversity assessment of the U. S. Collection of the new oilseed crop Lesquerella and related species. PLoS One 8:e64062

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379

Foreign Agricultural Service, United States Department of Agriculture http://usda.mannlib.cornell.edu/usda/fas/oilseed-trade//2010s/2017/oilseed-trade-01-12-2017.pdf

Grattapaglia D, Resende MDV, Resende MR, Sansaloni CP, Petroli CD, Missiaggia AA, Takahashi EK, Zamprogno KC, Kilian A (2011) Genomic selection for growth traits in Eucalyptus: accuracy within and across breeding populations. BMC Proceedings 5, Supplement 7:O16

Guoy M, Rousselle Y, Bastianelli D, Lecomte P, Bonnal L, Roques D, Efile J-C, Rocher S, Daugrois J, Toubi L, Nabeneza S, Hervouet C, Telismart H, Denis M, Thong-Chane A, Glaszmann JC, Hoarau J-Y, Nibouche S, Costet L (2013) Experimental assessment of the accuracy of genomic selection in sugarcane. Theoretical and Applied Genetics 126:2575-2586

Hardon JJ (1969) Interspecific hybrids in the genus *Elaeis* II. Vegetative growth and yield of F1 hybrids E. *guineensis* x E. *oleifera*. Euphytica 18:380-388

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Genomic Selection in dairy cattle: progress and challenges. Journal of Dairy Science 92:433-443

Jeennor S, Volkaert H (2014) Mapping of quantitative trait loci (QTLs) for oil yield using SSRs and gene-based markers in African oil palm (*Elaeis guineensis* Jacq.). Tree Genetics and Genomes 10:1-14

Khan MA, Han Y, Zhao YF, Troggio M, Korban SS (2012) A multi-population consensus genetic map reveals inconsistent marker order among maps likely attributed to structural variations in the apple genome. PLoS One 7:e47864

Kujur A, Upadhyaya HD, Shree T, Bajaj D, Das S, Saxena MS, Badoni S, Kumar V, Tripathi S, Gowda CLL, Sharma S, Singh S, Tyagi AK, Parida SK (2015) Ultra-high density intra-specific genetic linkage maps accelerate identification of functionally relevant molecular tags governing important agronomic traits in chickpea. Scientific Reports 5:9468

Kullan ARK, van Dyk MM, Jones N, Kanzler A, Bayley A, Myburg AA (2012) High-density genetic linkage maps with over 2,400 sequence-anchored DArT markers for genetic dissection in an F2 pseudo-backcross of *Eucalyptus grandis* x *E. urophylla*. Tree Genetics and Genomes 8:163-175

Kwong QB, Teh CK, Ong AL, Heng HY, Lee HL, Mohamed M, Low JZ, Apparow S, Chew FT, Mayes S, Kulaveerasingam H, Tammi M, Appleton DR (2016) Development and validation of a high-density SNP genotyping array for African oil palm. Molecular Plant 9:1132-1141

Liu L, Qu C, Wittkop B, Yi B, Xiao Y, He Y, Snowdon RJ, Li J (2013) A high-density SNP map for accurate mapping of seed fibre QTL in *Brassica napus* L. PLoS One 8:e83052

Mace ES, Rami J-F, Bouchet S, Klein PE, Klein RR, Kilian A, Wenzl P, Xia L, Halloran K, Jordan DR (2009) A consensus genetic map of sorghum that integrates multiple component maps and high-throughput Diversity Array technology (DArT) markers. BMC Plant Biology 9:13

Mayes S, Hafeez F, Price Z, Macdonald D, Billotte N, Roberts J (2008) Molecular research in oil palm, the key oil crop for the future. In: Moore PH and Ming R (eds) Genomics of Tropical Crop Plants. Springer, pp 371-404

Mayes S, Jack PL, Marshall DF, Corley RHV (1997) The construction of an RFLP linkage map for oil palm. Genome 40:116-122

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829

Montoya C, Cochard B, Flori A, Cros D, Lopes R, Cuellar T, Espeout S, Syaputra I, Villeneuve P, Pina M, Ritter E, Leroy T, Billotte N (2014) Genetic architecture of palm oil fatty acid composition in cultivated oil palm (*Elaeis guineensis* Jacq.) compared to its wild relative *E. oleifera* (H.B.K) Cortés. PLoS One 9:e95412

Montoya C, Lopes R, Flori A, Cros D, Cuellar T, Summo M, Espeout S, Rivallan R, Risterucci A-M, Bittencourt D, Zambrano JR, Alarcón GWH, Villeneuve P, Pina M, Nouy B, Amblard P, Ritter E, Leroy T, Billotte N (2013) Quantitative trait loci (QTLs) analysis of palm oil fatty acid composition in an interspecific pseudo-backcross from *Elaeis oleifera* (H. B. K.) Cortés and oil palm (*Elaeis guineensis* Jacq.) Tree Genetics and Genomes 9:1207-1225

Moretzsohn MC, Nunes CD, Fereirra ME, Grattapaglia D (2000) RAPD linkage mapping of the shell thickness locus in oil palm (*Elaeis guineensis* Jacq). Theoretical and Applied Genetics 100:63-70

Moriguchi Y, Ujino-Ihara T, Uchiyama K, Futamura N, Saito M, Ueno S, Matsumoto A, Tani N, Taira H, Shinohara K, Tsumura Y (2012) The construction of a high-density linkage map for identifying SNP markers that are tightly linked to a nuclear-recessive major gene for male sterility in *Cryptomeria japonica* D. Don. BMC Genomics 13:95

Novoselović D, Bentley AR, Šimek R, Dvojković K, Sorrells ME, Gosman N, Horsnell R, Drezner G, Šatović Z (2016) Characterizing Croatian wheat Germplasm diversity and structure in a European context by DArT markers. Frontiers in Plant Science 7:184

Petroli CD, Sansaloni CP, Carling J, Steane DA, Vaillancourt RE, Myburg AA, da Silva OBJ, Pappas GJJ, Kilian A, Grattapaglia D (2012) Genomic characterization of DArT markers based on high-density linkage analysis and physical mapping to the *Eucalyptus* genome. PLoS One 7:e44684

Pootakham W, Jomchai N, Ruang-areerate P, Shearman JR, Sonthirod C, Sangsrakru D, Tragoonrung S, Tangphatsornruang S (2015) Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). Genomics 105:288-295

Ren R, Ray R, Li P, Xu J, Zhang M, Liu G, Yao X, Kilian A, Yang X (2015) Construction of a high-density DArTseq SNP-based genetic map and identification of genomic regions with segregation distortion in a genetic population derived from a cross between feral and cultivated-type watermelon. Molecular Genetics and Genomics 290:1457-1470

Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA, Pappas GJ, Kilian A, Grattapaglia D (2012) Genomic Selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytologist 194:116-128

Rutkoski J, Benson J, Jia Y, Brown-Guedira G, Jannink J-L, Sorrells M (2012) Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. The Plant Genome 5:51-61

Sánchez-sevilla JF, Horvath A, Botella MA, Gaston A, Folta K, Kilian A, Denoyes B, Amaya I (2015) Diversity Arrays Technology (DArT) marker platforms for diversity analysis and linkage mapping in a complex crop, the octoploid cultivated strawberry (*Fragaria* x *ananassa*). PLoS ONE 10(12):e0144960

Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, Kilian A (2011) Diversity Arrays Technology (DArT) and next generation sequencing combined: genome-wide high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. BMC Proceedings 5:P54

Seng T-Y, Mohamed Saad SH, Chin C-W, Ting N-C, Singh RSH, Zaman FQ, Tan S-G, Syed Alwee SSR (2011) Genetic linkage map of a high yielding FELDA Deli x Yangambi oil palm cross. PLos One 6: e265293

Schouten HJ, van de Weg WE, Carling J, Khan SA, McKay SJ, van Kaauwen MPW, Wittenberg AHJ, van Putten HJJK, Noordijk Y, Gao Z, Rees DJG, van Dyk MM, Jaccoud D, Considine MJ, Kilian A (2012) Diversity arrays technology (DArT) markers in apple for genetic linkage maps. Molecular Breeding 29:645-660

Śliwka J, Soltys-Kalina D, Szajko K, Wasilewicz-Flis I, Strzelczyk-Zyta D, Zimnoch-Guzowska E, Jakuczun H, Marczewski W (2016) Mapping of quantitative trait loci for tuber starch and leaf sucrose contents in diploid potato. Theoretical and Applied Genetics 129:131-140

Sim S-C, Durstewitz G, Plieske J, Wieseke R, Ganal MW, Deynze AV, Hamilton JP, Bueli CR, Causse M, Wijeratne S, Francis DM (2012) Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS One 7:e40563

Singh R, Low E-TL, Ooi LC-L, Ong-Abdullah M, Ting N-C, Nagappan J, Nookiah R, Amiruddin MD, Rosli R, Abdul Manaf MA, Chan K-L, Halim MA, Azizi N, Lakey N, Smith SW, Budiman MA, Hogan M, Bacher B, van Brunt A, Wang C, Ordway JM, Sambanthamurthi R, Martienssen RA (2013a) The oil palm *SHELL* gene controls oil yield and encodes a homologue of SEEDSTICK. Nature 500:340-344

Singh R, Ong-Abdullah M, Low E-TL, Abdul Manaf MA, Rosli R, Nookiah R, Ooi LC-L, Ooi S-E, Chan K-L, Halim MA, Azizi N, Nagappan J, Bacher B, Lakey N, Smith SW, He D, Hogan M, Budiman MA, Lee EK, Desalle R, Kurdna D, Goicoechea JL, Wing RA, Wilson RK, Fulton RS, Ordway JM, Martienssen

RA, Sambanthamurthi R (2013b) Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature 500:335-339

Singh R, Tan SG, Panandam JM, Rahman RA, Ooi LCL, Low E-TL, Sharma M, Jansen J, Cheah SC (2009) Mapping quantitative trait loci (QTLs) for fatty acid composition in an interspecific cross of oil palm. BMC Plant Biology 9:114

Soh AC, Kee KK, Goh KJ (2006) Research and innovation towards sustainable palm oil production. Journal of Science and Technology in the Tropics 2:77-95

Soh AC, Lee CH, Yong YY, Chin CW, Tan YP, Rajanaidu N, Phuah PK (1990) The precision of oil palm breeding trials in Malaysia. In: Soh AC, Rajanaidu N, Basri MNH (eds) International Symposium Application of Statistics to Perennial Tree Crops Research. Palm Oil Research Institute Malaysia, Kuala Lumpur, pp 41-50

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One 8:e54985

Studer B, Kölliker R, Muylle H, Asp T, Frei U, Roldán-ruiz I, Barre P, Tomaszewski C, Meally H, Barth S, Skøt L, Armstead IP, Dolstra O, Lübberstedt T (2010) EST-derived SSR markers used as anchor loci for the construction of a consensus linkage map in ryegrass (*Lolium* spp.). BMC Plant Biology 10:177

Tadesse W, Ogbonnaya FC, Jighly A, Sanchez-Garcia M, Sohail Q, Rajaram S, Baum M (2015) Genome-wide association mapping of yield and grain quality traits in winter wheat genotypes. PLoS One 10(10):e0141339

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

Ting N-C, Jansen J, Nagappan J, Ishak Z, Chin C-W, Tan S-G, Cheah SC, Singh R (2013) Identification of QTLs associated with callogenesis and embryogenesis in oil palm using genetic linkage maps improved with SSR markers. PLoS ONE 8:e53076

Ting N-C, Jansen J, Mayes S, Massawe F, Sambanthamurthi R, Ooi LC-L, Chin C-W, Arulandoo X, Seng T-Y, Syed Alwee SS, Ithinin M, Singh R (2014) High density SNP and SSR-based genetics maps of two independent oil palm hybrids. BMC Genomics 14:309

van Ooijen JW (2011) Multipoint maximum likehood mapping in a full-sib family of an outbreeding species. Genetics Research 93:343-349

van Os H, Andrzejewski S, Bakker E, Barrena I, Bryan GJ, Caromel B, Ghareeb B, Isidore E, de Jong W, van Koert P, Lefebvre V, Millnourne D, Ritter E, van der Voort JNAMR, Rouselle-bougeois F, van Vliet J, Waught R, Visser RGF, Bakker J, van ECK HJ (2006) Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. Genetics 173:1075-1087

Wang X, Yu K, Li H, Peng Q, Chen F, Zhang W, Chen S, Hu M, Zhang J (2015) High-density SNP map construction and QTL identification for the apetalous character in *Brassica napus* L. Frontiers in Plant Science 6:1164

Wong CK, Bernardo R (2008) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theoretical and Applied Genetics 116:815-824

Yu H, Deng Z, Xiang C, Tian J (2014) Analysis of diversity and linkage disequilibrium mapping of agronomic traits on B-genome of wheat. Journal of Genomics 10(2):20-30

Legends to Figures and Tables

Fig. 1 Saturation of the *Sh* region at LG 2 for the AA0768 and AA0769 populations and integration of higher density maps. Marker names are shown to the right of each LG, with map distances (in cM) to the left. Common markers between the two maps were linked. D: DArT marker, S: SNP marker, mEgCIR: *E. guineensis* SSR marker

Table 1 Segregation distortion of markers for both the AA0768 and AA0769 mapping populations at different significance levels

Table 2 Characteristics of the genetic linkage groups of the mapping population AA0768

Table 3 Characteristics of the genetic linkage groups of the mapping population AA0769

Table 4 Characteristics of the integrated map

Table 5 Summary of homology search of the DArTseq markers of integrated map against the MPOB *pisifera* genome assembly P5

Table 6 Homology search of the DArTseq markers close to the *Sh* gene at LG 2 against the MPOB *pisifera* genome assembly P5

Online Resource 1 Genetic linkage maps of AA0768 population

Online Resource 2 Genetic linkage maps of AA0769 population

Online Resource 3 Genetic linkage maps of integrated map

Online Resource 4

Sheet (i) Marker list, position and segregation distortion of genetic linkage maps of AA0768 population

Sheet (ii) Marker list, position and segregation distortion of genetic linkage maps of AA0769 population

Sheet (iii) Marker list, position and sequence of genetic linkage maps of integrated map