



- (51) **International Patent Classification:**
C12Q 1/68 (2006.01) *A01H 1/04* (2006.01)
- (21) **International Application Number:**
PCT/MY2015/050034
- (22) **International Filing Date:**
14 May 2015 (14.05.2015)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
PI2014001391 14 May 2014 (14.05.2014) MY
- (71) **Applicant:** ACGT SDN BHD [MY/MY]; 24th Floor, Wisma Genting, Jalan Sultan Ismail, Kuala Lumpur, 50250 (MY).
- (72) **Inventors:** CHEAH, Suan Choo; A506, Puteri Palma Condominium, IOI Resort, Putrajaya, 62502 (MY). LEE, Ying Wah; 141, Jalan Bukit Jed 3, Taman Bukit Jed, Seremban, Negeri Sembilan, 70200 (MY). BOON, Soo Heong; No.12, Jalan Puteri 12/23, Bandar Puteri, Puchong, Selangor, 47100 (MY). LEE, Weng Wah; 43, Jalan PP3/17, Taman Putra Prima, Puchong, Selangor, 47130 (MY). CHUA, Kia Ling; 6, Jalan 3/38, Seksyen 3, Bandar Baru Bangi, Selangor, 43650 (MY). LEE, Chong Hee; 2, Jalan Sg. Ramal 32/53, Bukit Rimau, Shah Alam, Selangor, 40460 (MY).
- (74) **Agent:** SPRUSON & FERGUSON (M) SDN BHD; A-32-3A, Level 32, Menara UOA Bangsar, No.5 Jalan Bangsar Utama 1, Kuala Lumpur 59000 (MY).

- (81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))
- with sequence listing part of description (Rule 5.2(a))

(54) **Title:** METHOD OF PREDICTING OR DETERMINING PLANT PHENOTYPES IN OIL PALM

(57) **Abstract:** Methods of identifying polymorphic genetic markers of a plant species linked to a phenotype of interest are described. Polymorphic markers of interest determined by association mapping, and methods of predicting plant phenotypes are further determined.



METHOD OF PREDICTING OR DETERMINING PLANT PHENOTYPES IN OIL PALM

TECHNICAL FIELD

5 The present invention relates generally to the field of molecular biology. In particular, the present invention relates to plant genetics for detecting and using single nucleotide polymorphisms for selection purposes in oil palms.

BACKGROUND

10 The oil palm (*Elaeis guineensis*) industries in Malaysia and Indonesia were first established using *Deli duras* as the commercial planting materials. The *Deli duras* was derived from four palms planted at the Botanic Garden, Bogor in 1848. The oil palm, initially planted as avenue ornamental palms, became a commercial crop with the establishment of the first oil palm estate in 1917 at Tennamaran Estate in Selangor. The Malaysian oil palm industry has
15 expanded rapidly over the past 50 years, with areas planted with oil palm having grown from 54,000 hectares in 1960 to 5 million hectares in 2010.

 After the inheritance of shell thickness was reported in 1941 (Beirnaert and Vanderweyen, Congo Belge, Ser. Sci, 27:1-101, 1941), the commercial oil palm planting materials progressively shifted from *duras* to *dura x pisifera* (or DxP crosses) or *teneras*.
20 Compared to the *duras*, the DxP planting materials have higher mesocarp to fruit ratios (from 60% to more than 80%), which resulted in about 30% increase in oil yield.

 To produce high yielding DxP planting materials, oil palm breeders are improving the *dura* and *pisifera* populations separately, and selecting elite *duras* and *pisifera* palms for seed production. For example, in the case of *duras*, *duras* having desirable traits are selected and
25 used to generate *dura x dura* (or DxD) crosses. The seeds generated are planted in the nursery for a year and subsequently field planted. The palms come into bearing between 2 ½ to 3 years depending on the location at which they are planted. The yields of these palms are then recorded for five years to determine its yield potential for another cycle of selection and improvement.

30 In the *pisifera* improvement programme, selected *pisiferas* are crossed with a selected *tenera* to generate a *tenera x pisifera* (or TxP) cross. Alternatively, selected *teneras* can be used to generate *pisiferas* in a *tenera x tenera* (or TxT) cross. The seeds are planted in the nursery and subsequently field planted as in the *duras*, the only difference being that the *pisiferas* are female sterile or abortive, and hence cannot be yield recorded. Yield recording can only be
35 carried out in the *dura* or *tenera* sibs.

Oil palm breeding has improved the yield potential of oil palm planting materials over the years. For example, Lee et al. (Lee et al., Proc. Workshop 'Progress of oil palm breeding populations' (1990) p.p 81-89.) found a significant increase in oil yield of *dura* populations after four generations of selection compared to unselected *dura* palms (5.0 t oil/ha/year vs 3.1 t oil/ha/year, which represents a yield increase of 61.3%). Similarly, oil palm yield of DxP planting materials had over the years improved significantly as shown in the table below.

Table 1: Yield Performance of Planting Materials in Malaysia

Planting Material	Year Planted	FFB Yield (t/ha/yr)	O/B	Oil Yield (t/ha/yr)	Source
D x CI/UAC/SP	1962	22.6	21.9	4.9	Lee & Toh, Planter (1991) 67:317-324.
D x AVROS	1964	31.0	23.5	7.3	Lee & Toh, Planter (1991) 67:317-324.
D x AVROS	1968	31.1	22.1	6.9	Lee & Toh, Planter (1991) 67:317-324.
D x AVROS	1970	31.6	24.2	7.6	Lee & Toh, Planter (1991) 67:317-324.
D x Dy AVROS	1979	33.3	25.8	8.6	Lee & Toh, Planter (1991) 67:317-324.
D x AVROS	1979	34.5	25.8	8.9	Lee & Toh, Planter (1991) 67:317-324.
D x Yangambi	1988	35.1	26.0	9.5	Sharma & Tan, Proc. of the Oil Palm Planting Materials for Local and Overseas Joint Ventures, pp: 118-135. PORIM (1996)
D x Yangambi	1991	34.8	31.1	10.8	Chin & Shuhaimi, Paper Presented at Intl. Symp. 'Oil Palm Genetic Resource And Utilization (2000)
D x Nigeria	1991	36.1	31.9	11.5	Jalani, Proc. 1998 International Oil Palm Conf. 'Commodity of the past, today and the future, p.p 93-109. (1998)

Source: Jalani *et al* Paper Presented at 2001 Int. Palm Oil Congr. Malaysia (2001)

10

The highest commercial yield achieved to-date was reported from a field of 32.15 hectares with a fresh fruit bunch (FFB) yield of 47.65 tonnes FFB/hectare/year (Pang Oil Palm: Sabah's Future, Paper Presented in PIPOC (2005)). Laboratory analyses on bunches harvested from the commercial fields of these planting materials was reported to be above 30% oil/bunch, while the highest oil mill extraction was 25.01%.

15

Notwithstanding the improvements in oil palm yield since establishment of the first oil palm estate, continuous effort in improving oil palm yield still faces many challenges.

Most of the oil palm breeding in Malaysia is based on conventional means of selection by recording the phenotype in the field and selecting the best oil palm for breeding based on the traits exhibited, performance, and combining ability. This method has several limitations: (a) there is no way of avoiding linkage drag due to the association of desirable genes to undesirable phenotypes; (b) the time required to improve the oil palm is very long, with conventional breeding typically requiring at least 12 years from selection and crossing to the end of yield recording; and (c) in cases where a complex trait such as yield is being selected for, it is difficult to assess the net impact of selection accurately as there may be interaction between the different components of yield – for example, an increase in one component may have a negative effect on another.

In addition, the heterozygosity of the parent lines used for the production of DxP planting materials result in variability in the yield of individual palm. Typically, a yield improvement of about 15% or more can be expected from good oil palm clones.

In an attempt to resolve the problems encountered in conventional breeding and selection methods and further improve oil palm yield, molecular based methods have been used to accelerate development of superior oil palm planting materials. For example, marker-assisted selection has been used, where genomic data are mined (based on the field data available on the performance of specific progenies) to identify specific genetic sequences which consistently correlate to a specific characteristic, such as high FFB yield. These genetic sequences can then be used as molecular markers to select, for example, for high FFB yield even at the nursery stage, thereby reducing the cycle time in a breeding program from 12 years to 6 years.

Hence, there is a need for genetic markers, with which a prediction of the plant phenotype, as well as a selection process of progeny plants, could be expedited, based on the results of the comparison of genetic analysis between genetic markers that are associated with desired phenotypes, thus facilitating, in a more time and cost effective manner, plant breeding programs.

SUMMARY OF INVENTION

In a first aspect, the present invention refers to a method of predicting or determining a plant phenotype of interest, wherein the method comprises detecting the presence or absence of one or more polymorphic genetic markers selected from the group consisting of SEQ ID NOs. 1 to 115.

In a second aspect, the present invention refers to use of one or more polymorphic genetic markers according to any one of the preceding claims as criteria for the selection of a plant,

wherein the selection is based on the presence or absence of one or more polymorphic genetic markers.

DEFINITIONS

5 The following words and terms used herein shall have the meaning indicated:

The term "polymorphism" as used herein refers to a variation or difference in the sequence of a genetic region that arises in some of the members of a species. Variant sequences can be defined with reference to an arbitrary or non-arbitrary standard sequence for the species. A polymorphism is thus said to be "allelic," in that, due to the existence of the polymorphism, 10 some members of a species may have the "standard" sequence (i.e. the standard "allele") whereas other members may have a variant sequence (i.e., a variant "allele"). Thus, as used herein, an allele is one of two or more alternative versions of a gene or other genetic region at a particular location on a chromosome. In the simplest case, only one variant sequence may exist, and the polymorphism is thus said to be bi-allelic. In other cases, the species' population may 15 contain multiple alleles, and the polymorphism is termed tri-allelic, etc.

A single gene or genetic region may have multiple different unrelated polymorphisms. For example, it may have one bi-allelic polymorphism at one site, another bi-allelic polymorphism at another site and a multi-allelic polymorphism at yet another site. When all the sequences for a group of alleles at a chromosomal locus in a plant are the same, the alleles are 20 said to be "homozygous" at that locus. When the sequence of any allele at a particular locus in a plant is different, the population of alleles is said to be "heterozygous" at that locus.

Phenotypic traits can vary due to environmental and/or genetic factors. For example, polymorphisms at a particular chromosomal locus can affect the phenotypic trait associated with that locus.

25 The expression "phenotypic trait" as used herein relates to any particular characteristic or "trait" exhibited by a plant, whether naturally occurring or otherwise, that is capable of being inherited. Moreover, the phenotypic trait of interest may, for example, be transient, permanent or only present when the plant or part thereof is subjected to environmental stimuli or challenge. A phenotypic trait of interest may be a desired or positive trait. In other cases the 30 phenotypic trait of interest may be an undesired or negative trait.

Many traits of agronomic and horticultural interest are controlled by a single gene and fall into a few distinct phenotypic classes. These types of phenotypes are called discontinuous traits and can be used to predict the genotypes of the individuals.

35 Other traits do not fall into discrete classes. Rather, when a segregating population is analysed for these traits, a continuous distribution is found. For example, if a trait does not fall into a tight distribution, but exhibits a bell-shaped curve for a normal distribution, this type of

trait is called a continuous trait and cannot be analysed in the same manner as discontinuous traits. Because continuous traits are often given a quantitative value, they are often referred to as quantitative traits. Furthermore, the loci controlling these traits are called quantitative trait loci or QTL.

5 Furthermore, phenotypic traits are not limited to visible traits. While the phenotypic trait may be any trait, preferred traits of interest are those that have agricultural significance. Examples of agricultural traits include those that affect a component of yield, those that provide disease or chemical resistance, and those that affect developmental traits such as pollen or ovule production, etc., and those that affect composition of plants or plant parts, including seed
10 proteins or oils, starch or sugar composition, nutrient content and the like. In a further example, the trait may comprise but is not limited of traits that affect a component of yield such as fresh fruit bunch yield, bunch number, trunk size, or combinations thereof. In yet another example, the yield component may comprise, but is not limited to, fresh fruit bunch weight, bunch size, fruit size, oil extraction rate, bunch numbers or combinations thereof.

15 An example of a method for predicting or determining a plant genotype of interest may include, but is not limited to, the following steps, starting with taking a sample of a plant, seed or sapling to be analysed. This genetic analysis of the plant would show, for example, which single nucleotide polymorphisms (SNPs) are present and where. The result of this type of genetic analysis or mapping for this sample may then be compared with a list of known SNPs,
20 which had previously been correlated with one or more characteristics for that plant, e.g. plant height, fruit yield etc. Thus, this comparison between the genetic analysis of the sample and the list of known SNPs would be able to infer a statistical probability with which the tested plant would display the desired characteristic upon maturity. In order for this to be possible, at least one SNP must be present in the sample.

25 As explained above, many phenotypic traits are the result of multiple genes or genetic factors, for example, a phenotypic trait that is the result of a quantitative trait allele. An allele of a quantitative trait locus (QTL) can comprise multiple genes or other genetic factors even within a contiguous genomic region or linkage group.

“Locus” refers to a specific chromosome location in the genome of a species where a
30 specific gene can be found. The term “quantitative trait locus” as used herein relates to a region of DNA containing or linked to the genes that underlie a quantitative trait. Mapping regions of the genome that contain genes involved in specifying a quantitative trait is done using molecular tags such as AFLP or, more commonly SNPs. This is an early step in identifying and sequencing the actual genes underlying trait variation. Quantitative traits refer to phenotypes
35 (characteristics) that vary in degree and can be attributed to polygenic effects, i.e., product of

two or more genes, and their environment. It therefore may be defined as a gene affecting the phenotypic variation in continuously varying traits.

As used herein “quantitative trait locus (QTL) analysis” is a statistical method that links two types of information—phenotypic data (trait measurements) and genotypic data
5 (usually molecular markers) in an attempt to explain the genetic basis of variation in complex traits. Accordingly, QTL analysis allows linking certain complex phenotypes to specific regions of chromosomes. The goal of this process is to identify the action, interaction, number, and precise location of these regions.

As used herein, an “allele of a quantitative trait locus” can therefore encompass more
10 than one gene or other genetic factor where each individual gene or genetic component is also capable of exhibiting allelic variation and where each gene or genetic factor also has a phenotypic effect on the quantitative trait in question.

The term “chromosome segment” designates a contiguous linear span of genomic DNA that resides in plants on a single chromosome. The genetic elements or genes located on a
15 single chromosome segment are physically linked. In the context of the present invention the genetic elements located within a chromosome segment are also genetically linked, typically within a genetic recombination distance of less than or equal to 10 centimorgan (CM). That is, two genetic elements within a single chromosome segment undergo recombination during meiosis with each other at a frequency of less than or equal to about 10%.

As used herein, a “marker” is an indicator for the presence of at least one
20 polymorphism. A marker is preferably a nucleic acid molecule. It is understood that a marker can, for example, be an oligonucleotide probe or primer. Molecular markers are preferred for genotyping, because these markers are unlikely to affect the trait of interest. In some examples, the marker may comprise but is not limited to single nucleotide polymorphisms (SNPs), simple
25 sequence repeats (SSRs or microsatellites), amplified fragment length polymorphisms (AFLPs), random amplification of polymorphic DNAs (RAPDs), restriction fragment length polymorphisms (RFLPs), and transposable element positions and combinations thereof.

The term “single nucleotide polymorphism” as defined herein is a DNA sequence variation that occurs when a single nucleotide (A, T, C, or G) in the genome sequence is altered
30 or differs between members of a biological species or paired chromosomes in a human. Thus, as used herein a SNP is any polymorphism characterized by a different single nucleotide at a particular physical position in at least one allele. Each individual in a given population has many single nucleotide polymorphisms that together create a unique DNA pattern for that individual. As used herein, “simple sequence repeats (SSRs)” also known as microsatellites, or
35 short tandem repeats (STRs), are repeating sequences of 2-6 base pairs of DNA. It is a type of variable number tandem repeat (VNTR). SSRs are typically co-dominant.

The term “Amplified Fragment Length Polymorphisms (AFLPs)” as used herein relates to differences in restriction fragment lengths caused by SNPs or INDELs that create or abolish restriction endonuclease recognition sites. The AFLP technique is based on the selective PCR amplification of restriction fragments from a total digest of genomic DNA.

5 As used herein, the term “Random Amplified Polymorphic DNA (RAPD)” markers are DNA fragments from PCR amplification of random segments of genomic DNA with a single primer of arbitrary nucleotide sequence. Unlike traditional PCR analysis, RAPD does not require any specific knowledge of the DNA sequence of the target organism: the identical 10-mer primers will or will not amplify a segment of DNA, depending on positions that are
10 complementary to the primers' sequence. For example, no fragment is produced if primers annealed too far apart or 3' ends of the primers are not facing each other. Therefore, if a mutation has occurred in the template DNA at the site that was previously complementary to the primer, a PCR product will not be produced, resulting in a different pattern of amplified DNA segments on the gel. RAPDs are dominant in the sense that the presence of a RAPD band
15 does not allow distinction between heterozygous and homozygous states. The term “Restriction Fragment Length Polymorphism (RFLP)” as used herein refers to differences in restriction fragment lengths caused by SNPs or INDELs that create or abolish restriction endonuclease recognition sites. RFLP assays are performed by hybridizing a chemically labelled DNA probe to a Southern blot of DNA digested with a restriction endonuclease.

20 A “nucleic acid marker” as used herein means a nucleic acid molecule that is capable of being a marker for detecting a polymorphism.

The term “associated with” or “associated,” when referring to a nucleic acid (e.g., a genetic marker) and a phenotype in the context of the present invention, refers to a nucleic acid and a phenotypic trait that are in linkage phase disequilibrium. The term “linkage phase
25 disequilibrium” or “linkage disequilibrium” refers to a non-random segregation of genetic loci. This implies that such loci are in sufficient physical proximity along a length of a chromosome that they tend to segregate together with greater than random frequency.

The term “genetically linked” refers to genetic loci (including genetic marker loci) that are physically close enough to each other on the same chromosome such that they have a
30 recombination frequency of less than 0.5. When referring to the relationship between two genetic elements, such as a genetic element contributing to yield and a proximal marker, “coupling” phase linkage indicates the state where the “favourable” allele at the yield locus is physically associated on the same chromosome strand as the “favourable” allele of the respective linked marker locus.

35 In coupling phase, both favourable alleles are inherited together by progeny that inherit that chromosome strand. In “repulsion” phase linkage, the “favourable” allele at the yield locus

is physically linked with an “unfavourable” allele at the proximal marker locus, and the two “favourable” alleles are not inherited together (i.e., the two loci are “out of phase” with each other).

The term “physically linked” is used to indicate that two genetic loci, e.g., two marker loci, a marker locus and a locus contributing to variation in a phenotype, are physically present on the same chromosome. Typically, the two loci are located in close proximity, such that recombination between homologous chromosome pairs does not occur between the two loci with high frequency. That is, recombination between two physically linked loci typically occurs with a frequency of less than about 10%, favourably with a frequency of less than 5%, more favourably with a frequency of 2% or less or a frequency of 1% or less. Thus, two loci that are localized to the same chromosome, and at such a distance that recombination between the two loci occurs at a frequency of less than 10% are said to be “proximal to” each other.

A “genetic map” is a description of the genetic linkage relationships among loci on one or more chromosomes (or linkage groups) within a given species, generally depicted in a diagrammatic or tabular form. “Mapping” is the process of defining the linkage relationships of loci through the use of genetic markers, populations segregating for the markers, and standard genetic principles of recombination frequency. A “map location” is an assigned location on a genetic map relative to linked genetic markers where a specified marker can be found within a given species.

A “genotype” is the genetic constitution of an individual (or group of individuals) at one or more genetic loci. Genotype is defined by the allele(s) of one or more known loci that the individual has inherited from its parents. A “haplotype” is the genotype of an individual at a plurality of genetic loci. Typically, the genetic loci described by a haplotype are physically and genetically linked, i.e., on the same chromosome segment.

An individual is “homozygous” if the individual has only one type of allele at a given locus (e.g., a diploid individual with two copies of the same allele at a locus). An individual is “heterozygous” if more than one allele type is present at a given locus (e.g., a diploid individual with one copy each of two different alleles). The term “homogeneity” indicates that members of a group have the same genotype at one or more specific loci. In contrast, the term “heterogeneity” is used to indicate that individuals within the group differ in genotype at one or more specific loci.

A “line” or “strain” is a group of individuals of identical parentage that are generally inbred to some degree and are generally homozygous and homogeneous at most loci.

An “elite line” or “elite strain” is a genetically superior line that has resulted from many cycles of breeding and selection for superior agronomic performance. Numerous elite lines are available and known to those of skill in the art of soybean breeding. An “elite

population” is an assortment of elite lines that can be used to represent the state of the art in terms of agronomically superior genotypes of a given crop species, such as oil palm.

The term “oligonucleotide” as used herein refers to short nucleic acid molecules useful, e.g. for hybridizing probes, nucleotide array elements or amplification primers. Oligonucleotide molecules are comprised of two or more nucleotides, i.e. deoxyribonucleotides or ribonucleotides, preferably more than five and up to 30 or more. The exact size will depend on many factors, which in turn depend on the ultimate function or use of the oligonucleotide. Oligonucleotides can comprise ligated natural nucleic acid molecules or synthesized nucleic acid molecules and comprise between 10 to 150 nucleotides or between about 12 and about 100 nucleotides which have a nucleotide sequence which can hybridize to a strand of polymorphic DNA, e.g. to permit detection of a polymorphism.

Such oligonucleotides may be nucleic acid elements for use on solid arrays (e.g. synthesized or spotted). In preferred aspects of the invention such oligonucleotides can comprise as few as 12 hybridizing nucleotides, e.g. for assays where the oligonucleotide also comprises a detectable label. In other preferred aspects of the invention the oligonucleotide can comprise as few as about 15 hybridizing nucleotides, e.g. for single base extension assays.

Such oligonucleotides may also be primers for use in polymerase chain reaction (PCR) or other reactions. The term “primer” as used herein refers to a nucleic acid molecule, preferably an oligonucleotide whether derived from a naturally occurring molecule such as one isolated from a restriction digest or one produced synthetically, which is capable of acting as a point of initiation of synthesis when placed under conditions in which synthesis of a primer extension product which is complementary to a nucleic acid strand is induced, i.e., in the presence of nucleotides and an agent for polymerization such as DNA polymerase and at a suitable temperature and pH. The primer is preferably single stranded for maximum efficiency in amplification, but may alternatively be double stranded. If double stranded, the primer is first treated to separate its strands before being used to prepare extension products. Preferably, the primer is an oligodeoxyribonucleotide. The primer must be sufficiently long to prime the synthesis of extension products in the presence of the agent for polymerization. The exact lengths of the primers will depend on many factors, including temperature and source of primer. For example, depending on the complexity of the target sequence, the oligonucleotide primer typically contains at least 15, more preferably 18 nucleotides, which are identical or complementary to the template and optionally a tail of variable length which need not match the template. The length of the tail should not be so long that it interferes with the recognition of the template. Short primer molecules generally require cooler temperatures to form sufficiently stable hybrid complexes with the template.

The primers herein are selected to be “substantially” complementary to the different strands of each specific sequence to be amplified. This means that the primers must be sufficiently complementary to hybridize with their respective strands. Therefore, the primer sequence need not reflect the exact sequence of the template. For example, a non-complementary nucleotide fragment may be attached to the 5' end of the primer, with the remainder of the primer sequence being complementary to the strand. Alternatively, non-complementary bases or longer sequences can be interspersed into the primer, provided that the primer sequence has sufficient complementarity with the sequence of the strand to be amplified to hybridize therewith and thereby form a template for synthesis of the extension product of the other primer. Computer generated searches using programs such as Primer3 (www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi), STSPipeline (www-genome.wi.mit.edu/cgi-bin/www-STSPipeline), or GeneUp, for example, can be used to identify potential PCR primers. Exemplary primers include primers that are 18 to 50 bases long, where at least between 18 to 25 bases are identical or complementary to at least 18 to 25 bases of a segment of the template sequence.

This invention also contemplates and provides primer pairs for amplification of nucleic acid molecules in order to detect polymorphisms. As used herein “primer pair” means a set of two oligonucleotide primers based on two separated sequence segments of a target nucleic acid sequence. One primer of the pair is a “forward primer” or “5' primer” having a sequence which is identical to the more 5' of the separated sequence segments (+ strand). The other primer of the pair is a “reverse primer” or “3' primer” having a sequence which is the reverse complement of the more 3' of the separated sequence segments (+ strand). A primer pair allows for amplification of the nucleic acid sequence between and including the separated sequence segments. Optionally, each primer pair can comprise additional sequences, e.g. universal primer sequences or restriction endonuclease sites, at the 5' end of each primer, e.g. to facilitate cloning, DNA sequencing, or re-amplification of the target nucleic acid sequence.

The term “cloning” used in the context of oil palm refers to a process in which identical copies of a selected palm (ortet), are reproduced by developing plantlets from the leaf tissue of *tenera* oil palms with desirable characteristics.

As used herein, a “mapping population” is a collection of plants capable of being used with markers to map the genetic position of traits.

As used herein, a “polymorphic marker” is a marker capable of detecting one or more polymorphisms.

The present invention provides nucleic acid molecules which are markers, i.e. capable of detecting polymorphisms that are distributed throughout the genome of a mapping population.

As used herein, a “characterized polymorphism” is a polymorphism whose physical position on a genome is known. In one example, the physical position of a characterized polymorphism on an isolated nucleic acid molecule, such as a bacterial artificial chromosome comprising oil palm genomic DNA, is known. Thus the present invention also provides nucleic acid molecules capable of detecting characterized polymorphisms throughout a genome.

In one example, a characterized polymorphism is any polymorphism where the nucleic acid sequences of at least two of the polymorphisms present in an oil palm mapping population are known (sequenced characterized polymorphism).

The polymorphisms capable of detection by nucleic acid molecules of the present invention are distributed throughout the genome of the mapping population in a manner that allows the efficient identification of a genomic region associated with a phenotypic trait. In one example, the polymorphisms are distributed throughout the genome where 60%, preferably 70%, more preferably 80%, even more preferably 90%, 95% or 100% of the genome has a characterized polymorphism at a density of higher than one polymorphism per 100 kb, more preferably higher than one polymorphism per 50 kb, and even more preferably higher than one polymorphism per 25 kb, 10 kb, 7 kb, 5 kb or 3 kb. In another example, the polymorphisms are distributed throughout the genome where 60%, preferably 70%, more preferably 80%, even more preferably 90%, 95% or 100% of genome has a characterized polymorphism at a density of higher than one polymorphism per 3.5 cM, more preferably higher than one polymorphism per 3.25 cM, and even more preferably higher than one polymorphism per 3.0 cM, 2.75 cM, 2.5 cM, 2.0 cM, 1.5 cM, 1.0 cM or 0.5 cM.

In one example, the efficient identification of a genomic region associated with a phenotypic trait, e.g. a QTL or a single gene, is disclosed, where the genomic region is less than 100 kb, more preferably less than 50 kb, and even more preferably less than 25 kb, 10 kb, 7 kb, 5 kb or 3 kb from a characterized polymorphism. In another example the efficient identification of a genomic region associated with a phenotypic trait where the genomic region is less than 3.5 cM, more preferably less than 3.25 cM, and even more preferably less than 3 cM, 2.75 cM, 2.5 cM, 2.0 cM, 1.5 cM, 1.0 cM or 0.5 cM from a characterized polymorphism.

It is understood that the distribution of polymorphisms need not be uniform in a genome as certain regions will exhibit a higher average density of polymorphisms (e.g. non-centromeric regions) and certain regions will exhibit a lower average density of polymorphisms (e.g. centromeric regions).

In one example, the efficient identification of a genomic region associated with a phenotypic trait of interest will be obtained by a simultaneous screening for the presence of 25 or more, more preferably 50 or more, even more preferably 75 or more, 100 or more, 150 or more, 200 or more, 250 or more, 300 or more, 400 or more or 500 or more, 1,000 or more,

2,000 or more, 3,000 or more, 4,000 or more polymorphisms. When high throughput assays are employed, e.g. with microarrays, it may be feasible to screen for the presence of 5,000 or more polymorphisms, e.g. at least 10,000 or even 15,000 polymorphisms. In another example, the efficient identification of a genomic region associated with a phenotypic trait of interest will be obtained by a simultaneously screening for the presence of 25 or more, more preferably 50 or more, even more preferably (where appropriate) 100 or more, or 250 or more etc. of the polymorphisms.

“Marker Assisted Selection” or “MAS” refers to the practice of selecting for desired phenotypes among members of a breeding population using genetic markers.

The phrase “hybrid plants” refers to plants which result from a cross between genetically different individuals.

The term “crossed” or “cross” in the context of this invention means the fusion of gametes, e.g., via pollination to produce progeny (i.e., cells, seeds, or plants). The term encompasses both sexual crosses (the pollination of one plant by another) and selfing (self-pollination, i.e., when the pollen and ovule are from the same plant).

In another example, the efficient identification of a genomic region associated with a phenotypic trait of interest will be obtained by screening for the presence of 25 or more, more preferably 50 or more, even more preferably 75 or more, 100 or more, 150 or more, 200 or more, 250 or more, 300 or more, 400 or more or 500 or more, 1,000 or more, 2,000 or more, 3,000 or more, 4,000 or more polymorphisms during a single assay. In yet another example the efficient identification of a genomic region associated with a phenotypic trait of interest will be obtained by screening for the presence of 25 or more, more preferably 50 or more, even more preferably (where appropriate) 100 or more or 250 or more etc. of the polymorphisms during a single assay. A single assay can comprise many steps. One or more of these steps can occur sequentially.

In some examples, the assay may be carried out using a high throughput system. A high throughput system may involve a solid phase array. In one example, the solid phase array may comprise a microarray.

In the assays below, a collection of markers for polymorphisms can comprise from a few up to millions of different nucleic acid molecules. For example, using simple dot-blot hybridization methods, membranes with many nucleic acid molecules can be generated for screening. The solid-phase techniques described below and known in the art can be adapted for high-throughput monitoring of polymorphisms. In such methods different immobilized nucleic acid molecule probes can be placed on a solid support at microarray densities of up to millions of nucleic acid molecules per square inch. Similarly, very large sets of nucleic acid molecules can be immobilized for simultaneous screening against one or more probes.

As used herein, the terms "increase" and "decrease" refer to the relative alteration of a chosen characteristic in a subset of a population in comparison to the same characteristic as present in the whole population. An increase thus indicates a change on a positive scale, whereas a decrease indicates a change on a negative scale. The term "change", as used herein, also refers to the difference between a chosen characteristic of an isolated population subset in comparison to the same characteristic in the population as a whole. However, this term is without valuation of the difference seen.

The word "substantially" does not exclude "completely" e.g. a composition which is "substantially free" from Y may be completely free from Y. Where necessary, the word "substantially" may be omitted from the definition of the invention.

Unless specified otherwise, the terms "comprising" and "comprise", and grammatical variants thereof, are intended to represent "open" or "inclusive" language such that they include recited elements but also permit inclusion of additional, un-recited elements.

As used herein, the term "about", in the context of concentrations of components of the formulations, typically means +/- 5% of the stated value, more typically +/- 4% of the stated value, more typically +/- 3% of the stated value, more typically, +/- 2% of the stated value, even more typically +/- 1% of the stated value, and even more typically +/- 0.5% of the stated value.

Throughout this disclosure, certain embodiments may be disclosed in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of the disclosed ranges. Accordingly, the description of a range should be considered to have specifically disclosed all the possible sub-ranges as well as individual numerical values within that range. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed sub-ranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual numbers within that range, for example, 1, 2, 3, 4, 5, and 6. This applies regardless of the breadth of the range.

Certain embodiments may also be described broadly and generically herein. Each of the narrower species and sub-generic groupings falling within the generic disclosure also form part of the disclosure. This includes the generic description of the embodiments with a proviso or negative limitation removing any subject matter from the genus, regardless of whether or not the excised material is specifically recited herein.

DETAILED DESCRIPTION OF THE PRESENT INVENTION

A molecular based method is association mapping, which allows for the detection of marker-linked traits without requiring the use of progeny plants. In contrast to linkage mapping

where only meiosis events occurring during the transmission from parents to progeny are taken into account, association mapping advantageously takes into account events that have already created association in the distant past, including many generations of meiosis. This allows association mapping to provide a higher resolution or detection power for identifying the link
5 between marker and phenotype.

Association mapping has been used in humans to discover disease related markers, in some plants such as maize (*Zea mays*, L.), soybean (*Glycine max* (L.) Merr.), barley (*Hordeum vulgare* L.), wheat (*Triticum aestivum* L.), tomato (*Lycopersicon esculentum* Mill.), sorghum (*Sorghum bicolor* (L.) Moench), and potato (*Solanum tuberosum* L.), as well as in some tree
10 species, such as aspen (*Populus tremula* L.) and loblolly pine (*Pinus taeda* L.).

In the present disclosure, a population of oil palm may comprise a population of oil palm having different genetic background. For example, a population of oil palm may comprise, but is not limited to, breeding populations of restricted origins (BPRO) such as AVROS, Deli, Yangambi, Cameroon, Ekona, Calabar and combinations thereof. The
15 population may be selected according to its structure, owing to geography, natural selection or artificial selection. For association mapping, a given sample may fall in one of five categories defined by population structure associated with local adaptation or diversifying selection and familial relatedness from recent co-ancestry. Ideally, samples with minimal population structure or familial relatedness result in the greatest statistical power, provided that the trait of
20 interest is well distributed.

The duration of phenotype data collection is given as an example below and it is conceivable that other time periods of data collection can be used for this process. For example, the phenotype data collection period may be 3 months or more, more preferably 6 months or more, even more preferably 9 months or more, 12 months or more, 18 months or more, 24
25 months or more, 36 months or more, 45 months or more, 60 months or more for measuring, collecting and analysing phenotype data. In some examples, the phenotype data collection is 45 months. In further examples the phenotype data collection is 54 months.

In some examples, the phenotype used for analysis and linkage to one or more polymorphic genetic markers in plants such as oil palm, may comprise, but is not limited to,
30 fresh fruit bunch yield, bunch number, trunk height, disease resistance, seed number, height increment, oil composition, oil content, carotene content, tocopherol/ tocotrienol, vegetative measurements and such. Any measureable quantitative or qualitative trait can be used as phenotype data.

In some examples, there is described a method as described herein, wherein the method
35 further comprises the step of recording phenotype data from at least one group of plants of the

plant species for the phenotype of interest and genotyping the genetic material of the plants for which phenotype data have been recorded.

For quantitative traits, the Quantitative Transmission Disequilibrium Test (QTDT) is well known to the person skilled in the art for association mapping. Samples of increased size as well as broader allelic diversity in a species often contain population structure or include familial relationships with structured population.

This method can be applied for any type of phenotype that can be observed or measured. In the case where a reduction of the trait of interest is required instead of an increase, markers with negative values for the effect estimate will be used instead.

The present invention further describes the development of a set of genetic markers that can be used individually or in combinations to predict or determine phenotypic traits in plants, such as in oil palm. This early prediction, or determination, of not yet obvious phenotypes is performed in order to aid the selection plants of the determined, advantageous phenotypic type, thus making it possible to reduce the breeding time required and also to save costs of the breeding program.

In some examples, a method is described, wherein plant material used for identifying polymorphic markers is obtained from at least two, at least three, or at least four or more different plant populations of the same species with different genetic backgrounds. Advantageously, the inventors found that with the inclusion of different genetic backgrounds, the ability of the analysis to discern markers linked to traits is enhanced.

In one example, there is described a method of identifying polymorphic genetic markers of a plant species linked to a phenotype of interest, wherein the method comprises identifying a statistically significant link between polymorphic markers identified for the plant and the phenotype of interest by association mapping; selecting markers which were found under a) to be linked to the phenotype of interest which have either the most significant negative or positive effect on the phenotype of interest by determining effect size for each of the markers identified under a) and selecting those with the most positive or negative effect on the phenotype of interest.

In another example, there is disclosed a method for predicting or determining a plant phenotype of interest, wherein the method comprises detecting the presence or absence of a one or more polymorphic genetic marker including, but not limited to SEQ ID NOs. 1 to 115. In another example the method comprises detecting the presence or absence of two or more, three or more, four or more, five or more, six or more, seven or more, eight or more, nine or more, the or more, or eleven or more or all polymorphic genetic marker(s) including, but not limited to SEQ ID NOs. 1 to 115. In another example, the genetic markers include, but are not limited to Markers nos. 1 (SEQ ID NO. 10), 2 (SEQ ID NO. 20), 3 (SEQ ID NO. 13), 4 (SEQ ID NO.

8), 5 (SEQ ID NO. 19), 6 (SEQ ID NO. 34), 7 (SEQ ID NO. 90), 8 (SEQ ID NO. 105), 9 (SEQ ID NO. 106), 10 (SEQ ID NO. 99) and 11 (SEQ ID NO. 80).

As used herein, the term "predicting" or "prediction" is defined as to state, or make something known in advance, especially using inference or special knowledge. Here, this means that with the knowledge available to a person skilled in the art in regards to the location and amount of polymorphic genetic markers in a specific plant would enable this person to predict, if this specific plant would show the deduced phenotype upon maturity.

As used herein, the term "determining" or "determination" is defined as ascertaining or establishing something exactly by research or calculation. In this case, the determination of a plant phenotype of interest is a direct result of the genetic analysis performed on a plant sample.

In a further example, there is disclosed the method as previously described, wherein the plant is an oil palm. Furthermore, in another example the oil palm, as previously disclosed, is of from the genus *Elaeis*. In another example, the genus, as previously disclosed, may consist of the forms *dura* and *tenera*. In another example, the genus, as previously disclosed, may consist of the variation *pisifera*. As used herein, the biological term "form" is defined as a "secondary" taxonomic rank, which is below that of variety, which in turn is below that of species. When used in the art, a form usually designates a group with a noticeable but minor deviation. For instance, white-flowered forms of species that usually have coloured flowers can be named a "f. alba". As used herein, the biological term "variety" is defined as a taxonomic rank below that of species, but above that of a form. A variety will appear distinct from other varieties, but, when brought into contact with other varieties will freely hybridize and provide offspring. In general, varieties are geographically separated from each other.

The ultimate goal of any breeding program is to combine as many favourable alleles as possible into elite varieties of germplasm that are genetically superior (with respect to one or more agronomic traits) to their ancestors. The markers described herein identify chromosome segments, i.e., genomic regions, and alleles (allelic forms) that have been favoured by long-term selection for yield. Accordingly, these markers can be used for marker assisted selection of oil palm plants with superior agronomic performance. For example, in a cross between parents that complement favourable alleles at the target loci, progeny can be selected that include more favourable alleles than either parent. Such progeny are predicted to be phenotypically superior to either parent. For example, in practice, after breeding, the method disclosed herein can be used for testing progeny plants to see if they would exhibit the selected phenotype by taking a sample from said plant, analysing for the presence or absence of polymorphic genetic markers and based a comparison of those markers with previously identified markers and their known phenotypes, make a prediction if this particular plant will exhibit the desired phenotype during growth or at maturity.

With limited progeny palms, the resolution of detection of markers linked to traits is reduced and the marker distance which is given in centimorgans (cM) becomes larger, thereby making the predictive power of the markers weaker. Inclusion of more markers will increase the chance that a marker will fall within the resolution limit but will not reduce the resolving power itself. This is due to the intrinsic limitation of the linkage mapping method, which detects genetic crossover through meiosis events and compares it between the progenies and the parent palms.

To overcome this problem, advanced intercross lines, such as a sixth or higher generation cross, have been applied in crop improvement programs in other plants. However, these crop improvement programs are not feasible in oil palm because it will take hundreds of years to generate such advanced intercross lines.

Employing Linkage mapping in oil palms has resulted in the development and identification of genetic markers that were statistically linked to traits of interest and were used for screening. This method facilitates breeding by enabling screening to be carried out at an earlier stage and increases the precision by which predictions on traits can be made, thus shortening cycle time and reducing the cost of breeding programs.

Marker assisted selection (MAS), employing the markers of the present invention, and the chromosome segments they identify are useful in the context of, e.g. an oil palm breeding program to increase efficiency in yield improvements. Phenotypic screening for a trait of interest, such as yield, for large numbers of samples can be expensive, as well as time consuming. In addition, phenotypic screening alone is often unreliable due to the effects of epistasis and non-genetic (e.g., environmental) contributions to the phenotype. MAS offers the advantage over field evaluation that it can be performed at any time of year regardless of the growing season or developmental stage. In addition, MAS facilitates evaluation of organisms grown in disparate regions or under different conditions.

A breeder of ordinary skill, desiring to breed oil palm plants with increased yield, can apply the methods for MAS described herein, using, e.g., the exemplary markers described herein or linked markers localized to the chromosome segments identified by markers listed in Tables 1 and 2 below, to derive oil palm lines with superior agronomic performance.

Genetic marker alleles, linked markers, QTL, identifying the chromosome segments encompassing genetic elements that are important for yield, are used to identify plants that contain a desired genotype at one or more loci, and that are expected to transfer the desired genotype, along with a desired phenotype to their progeny. Marker alleles (or QTL alleles) can be used to identify plants that contain a desired genotype at one locus, or at several unlinked or linked loci (e.g., a haplotype), and that would be expected to transfer the desired genotype, along with a desired phenotype to their progeny. Similarly, by identifying plants lacking the

desired allele, plants with an undesirable phenotype, e.g., plants with poor yield, can be identified, and, e.g., eliminated from subsequent crosses. It will be appreciated that for the purposes of MAS, the term "marker" can encompass both marker and QTL loci as both can be used to identify plants with a desired trait.

5 After a desired phenotype and a polymorphic chromosomal locus, e.g., a marker locus or QTL, are determined to segregate together (i.e., are determined to be in linkage disequilibrium), alleles corresponding to the desired phenotype are selected. In brief, a nucleic acid corresponding to the marker nucleic acid is detected in a biological sample from a plant to be selected. This detection can take the form of hybridization of a probe nucleic acid to a
10 marker, e.g., using allele-specific hybridization, Southern analysis, Northern analysis, *in situ* hybridization, hybridization of primers followed by PCR amplification of a product including the marker, or the like. After the presence (or absence) of a particular marker in the biological sample is verified, the plant is selected and, optionally, crossed to produce progeny plants.

 When a population is segregating for multiple loci affecting one or multiple traits, e.g.,
15 multiple loci involved in resistance to single disease, or multiple loci each involved in resistance to different diseases, the efficiency of MAS compared to phenotypic screening becomes even greater because all the loci can be processed in the lab together from a single sample of DNA. Thus, use of marker information for each of the traits in the breeding process is facilitated.

20 It will be appreciated that plants, positive for a marker of the invention, can be selected and crossed according to any breeding protocol relevant to the particular breeding program. Accordingly, progeny can be generated from a selected plant by crossing the selected plant to one or more additional plants selected on the basis of the same marker or a different marker, e.g., a different marker correlating with superior agronomic performance, or a different
25 phenotype of interest, e.g., resistance to a particular disease. Alternatively, a selected plant can be back crossed to one or both parents.

 Backcrossing is usually done for the purpose of introgressing one or a few loci from a donor parent, into an otherwise desirable genetic background from the recurrent (typically, an elite) parent. The more cycles of backcrossing that are performed, the greater is the genetic
30 contribution of the recurrent parent to the resulting variety. A selected plant can also be outcrossed, e.g., to a plant or line not present in its genealogy. Such a plant can be selected from among a population subject to a prior round of analysis, or may be introduced into the breeding program *de novo*. A plant positive for a desired marker can also be self-crossed ("selfed") to create a true breeding line with the same genotype.

35 In some examples, the method is as described herein, wherein the polymorphic markers are polymorphic markers from oil palms. The markers used for this invention are single

nucleotide polymorphism (SNP) markers but the methods can be applied to any genetic markers familiar to the person skilled in the art.

Some examples of genetic markers are described herein. In some examples, there is described the method as described herein wherein polymorphic genetic markers are SNPs (single nucleotide polymorphisms), SSRs (simple sequence repeats), AFLPs (amplified fragment length polymorphisms), RAPDs (random amplification of polymorphic DNAs) and combinations thereof.

In one example, there is described the method as described herein wherein polymorphic genetic markers are SNPs (single nucleotide polymorphisms).

In one example, the presence or absence of one or more polymorphic genetic markers is determined in a plant sample containing genetic material, wherein the sample material is selected from a group consisting of leaves, spear leaves, stem, saplings, roots, buds, flowers, seed and fruit.

Alternative methods such as using different markers paired with the appropriate analysis steps can also be performed to identify markers that show a clustering pattern that can be used for further analysis. Examples of these are SSR markers, AFLP markers, RAPD markers and other known genetic markers.

Once the population and markers have been selected, there is described a means to carry out the QTL analysis. In some examples, the parental strains may be crossed, resulting in heterozygous (F_1) individuals, and these individuals are then crossed as described above. Finally, the phenotypes and genotypes of the derived (F_2) population are scored. Markers that are genetically linked to a QTL influencing the trait of interest will segregate more frequently with trait values, whereas unlinked markers will not show significant association with phenotype.

For traits controlled by tens or hundreds of genes, the parental lines need not actually be different for the phenotype in question; rather, they must simply contain different alleles, which are then resorted by recombination in the derived population to produce a range of phenotypic values. Consider, for example, a trait that is controlled by four genes, wherein the upper-case alleles increase the value of the trait and the lower-case alleles decrease the value of the trait. Here, if the effects of the alleles of the four genes are similar, individuals with the *AABBccdd* and *aabbCCDD* genotypes might have roughly the same phenotype. The members of the F_1 generation (*AaBbCcDd*) would be invariant and would have an intermediate phenotype. However, the F_2 generation, or the progeny from a backcross of an F_1 individual with either parent, would be variable. The F_2 offspring would have anywhere from zero to eight upper-case alleles; the backcross progeny would have anywhere from four to eight upper-case alleles.

Thus, as used herein, the term “effect size” relates to the name given to one or two or a family of indices that measure the magnitude of a treatment effect. Unlike significance tests, these indices are independent of sample size. Effect size measures are the common currency of meta-analysis studies that summarize the findings from a specific area of research. Effect size
5 can be measured for example as the standardized difference between two means, such as the standardised mean difference between the two groups, or as the correlation between the independent variable classification and the individual scores on the dependent variable. This latter correlation is called the “effect size correlation”. In other words, an effect size is a measure of the strength of a phenomenon.

10 An effect size calculated from data is a descriptive statistic that conveys the estimated magnitude of a relationship without making any statement about whether the apparent relationship in the data reflects a true relationship in the population. In that way, effect sizes complement inferential statistics such as *p*-values. In the genetics studies, the effect size may be defined as a coefficient (β) for a SNP when its association with the outcome is modelled
15 through a regression model, such as linear regression for a quantitative trait or logistic regression for a qualitative trait, assuming a linear trend per copy of an allele. The regression coefficients for quantitative traits may be presented in units of standard deviation (SD) of the trait so that they are comparable across traits.

A principal goal of QTL analysis has been to answer the question of whether
20 phenotypic differences are primarily due to a few loci with fairly large effects, or to many loci, each with minute effects. It appears that a substantial proportion of the phenotypic variation in many quantitative traits can be explained with few loci of large effect, with the remainder due to numerous loci of small effect. Once QTL have been identified, molecular techniques can be employed to narrow the QTL down to candidate genes (a process described later in Example 1
25 and 2 below). It may be possible using the method as described herein using the analysis to predict the role of regulatory genes, or genes that code for transcription factors and other signalling proteins.

As with all statistical analyses, sample size is a critical factor. Small sample sizes may fail to detect QTL of small effect and result in an overestimation of effect size of those QTL
30 that are identified. There has been suggested a method for comparing detected QTL to a distribution of expected values in order to estimate how many loci might have been missed.

In some examples, the method, as described herein, describes association mapping, wherein association mapping comprises applying a general statistical linear model (GLM) combined with mixed statistical linear model (MLM).

35 When looking at QTL across traits and taxa, it has been noted that phenotypes are frequently affected by a variety of interactions (e.g. genotype-by-environment, and epistatic

interactions between QTL), although not all QTL studies are designed to detect such interactions. In addition, significant dominance, epistatic, and genotype-by-environment effects have also been reported for life span.

Similarly, QTL studies examining plant architecture differences between plants have repeatedly shown significant epistatic interactions. It is also possible to perform QTL analysis on un-manipulated natural populations using hybrids, sibships (half-sibling or full-sibling families), and/or pedigree information.

Thus there are many way to statistically analyse genetic markers in association mapping. In some examples, association mapping as described herein may be analysed using methods comprising, but not limited to, simple regression based for example on maximum-likelihood, analysis of variance (ANOVA) at the marker loci, methods of least square, Wright's F statistics, structured association, genomic control, quantitative transmission disequilibrium test (QTDT), population structure and relative kinship. In genomic control, random markers are used to estimate and adjust the inflation of test statistics generated by population structure. Genomic control and structured association are common methods to control for false positives (type I error) caused by population structure.

In some examples, there described a method, wherein association mapping comprises applying a general statistical linear model (GLM) combined with mixed statistical linear model (MLM). The general statistical linear model incorporates a number of different statistical models such as ANOVA, ANCOVA, MANOVA, MANCOVA, ordinary linear regression, t-test and F-test. The general linear model is a generalization of multiple linear regression models to the case of more than one dependent variable.

The mixed statistical linear model may be used to account for multiple levels of relatedness simultaneously as detected by random genetic markers. Advantageously, the inventors by applying the combination of GLM and MLM may predict association with more accuracy.

In some examples below, the inventors carried out analysis using STRUCTURE. STRUCTURE is software used for inferring population structure using genotype data. Detection of the number of clusters of individuals is obtained based on the posterior probability of the data for a given population using a Bayesian approach (Evanno, G., Regnaut, S., Goudet, J.; *Molecular Ecology*; 2005; 14: 2611-2620). The relative kinship matrix, defining the degree of genetic covariance between pairs of individuals may be calculated using SPAGeDi software. Both structured association approach (Pritchard, J. K., Stephens, M., Rosenberg, N. A., Donnelly, *Am. J. Hum. Genet.*; 2000; 67: 170-181) and unified mixed model method are used to reduce the false positives by taking account of population structure and family relatedness within populations (Yu, et. al., *Nature Genetics*, 2006, vol.38, no. 2, pp.203). These spurious

associations can occur when the associated markers are not linked to the causative loci but biased towards particular subpopulations when any marker allele is present with a high frequency in said subpopulations.

5 The use of co-factor data can also be added in to increase the ability to detect associations. For example, when performing FFB association mapping we may include the fruit size as a cofactor to help increase the ability of the association mapping to pick out markers linked to traits.

10 It is understood that different cut-off values can be used depending on the stringency of the analysis and efficient cut-off values will have to be determined based on the population type, organism, sample size and other consideration. For example, a polymorphic genetic marker with a p-value, empirically derived, of about below 5%, or about below 4%, or about below 3%, or about below 2%, or about below 1%, or about below 0.5%, or about below 0.1% is considered to have a statistically significant association to the phenotype of interest.

15 In some examples, the method is described, wherein for association mapping a polymorphic genetic marker with a p-value below a 5% empirically derived value is considered to have a statistically significant association to the phenotype of interest. In a further example, the method as described herein is a method for association mapping a polymorphic genetic marker with a p-value below a 1% empirically derived value is considered to have a statistically significant association to the phenotype of interest.

20 In some examples, the method as disclosed herein, relates on a method of mapping QTL based on the regression of phenotype on the marker effect. As explained in further detail below under the heading "Validation", the coefficient of determination, (R-squared or R^2) estimates the proportion of the phenotypic variation using markers. That is to say, R^2 may be used to determine or predict the marker or allele effect, which anticipates how much a trait of interest, is affected by the presence of specific alleles. In some examples, the R-square value may be applied to a set of markers having a p-value of less than 0.01. Thus, in some examples, the R-square cut-off value may be about more than about 0.01, or more than about 0.02, or more than about 0.03, or more than about 0.04, or more than about 0.05, or more than about 0.06, or more than about 0.07, or more than about 0.08, or more than about 0.09, or more than about 0.1, or more than about 0.15, or more than about 0.2. In some examples, for a set of markers having a p-value of less than 0.01, there is described an R-square cut-off value of more than about 0.1.

35 The R-square shows how much an allele or combination of alleles can explain the variances in a trait as compared to the total variance of that trait in the sample set. For example, if the R-square is 0.8 for a set of markers, that set of markers can predict 80% of the trait of interest. It also means that there additional unknown or undiscovered markers that effect the

traits by 20%. In other words, the R-square value may be seen as an index of the ability of one, two, three, four or more markers to predict a phenotypic trait accurately. The effect size however relates to the magnitude at which the allele impacts the value of the phenotype. For example, if the effect size for fruit weight is 50, then that allele influences the trait by 50kg. In
5 other words, the effect size may be seen as an index of the quantitative changes that the allele controls.

In some examples, the description further provides a method, wherein the polymorphic markers comprising a coefficient of determination having a value of less than 0.1 and wherein the polymorphic markers with the most significant negative effect are those with an estimate
10 effect size of <-10 or <-15 or <-20 or <-30, when tested *a posteriori*.

In some examples, there is described a method of predicting a plant phenotype of interest of a plant not yet showing the phenotype of interest for screening and selection of plants possessing the phenotype of interest to be planted or bred with other plants of the same species using polymorphic markers obtained in the method described herein. In other words,
15 the method allows anticipating the phenotype of interest without having to wait until the traits is measurable. Advantageously, by being able to predict the phenotype by determining the phenotype, the inventors are able to save time and costs.

In a further example, there is described a method, wherein the phenotype predicted is yield. Other phenotype of interest may be predicted. Some examples have been described
20 herein. In yet another example, there is described a method wherein the plant is oil palm and the phenotype predicted is fresh fruit bunch yield, bunch number or a combination thereof. In a further example, there is described a method, wherein the plant phenotype of interest is a yield component selected from the group consisting of fresh fruit bunch weight, bunch size, fruit size, oil extraction rate, bunch numbers and combinations thereof.

25 In another example, the phenotype of interest is associated with fresh fruit bunch weight or bunch number yield.

The present disclosure may also provide other yield component comprising, but not limited to, increased biomass (weight) of one or more parts of a plant, particularly above-ground (harvestable) parts, increased root biomass or increased biomass of any other
30 harvestable part; increased total seed yield, which includes an increase in seed biomass (seed weight) and which may be an increase in the seed weight per plant or on an individual seed basis; increased number of flowers ("florets") per panicle; increased number of (filled) seeds; increased seed size, which may also influence the composition of seeds; increased seed volume, which may also influence the composition of seeds (including oil, protein and carbohydrate
35 total content and composition); increased individual seed area; increased individual seed length and/or width; increased harvest index, which is expressed as a ratio of the yield of harvestable

parts, such as seeds, over the total biomass; and increased thousand kernel weight (TKW), which is extrapolated from the number of filled seeds counted and their total weight. An increased TKW may result from an increased seed size and/or seed weight.

In one example, there is described that the presence of one or more polymorphic markers is associated with a fresh fruit bunch yield (FFB) phenotype, wherein the markers include, but are not limited to, SEQ ID NOs: 1 to 79. In another example the method comprises detecting the presence or absence of two or more, three or more, four or more, five or more, six or more, seven or more, eight or more, nine or more, ten or more, or eleven or more or all polymorphic genetic marker(s) including, but not limited to SEQ ID NOs. 1 to 79. In another example, there is described that the presence of one or more polymorphic markers is associated with a bunch number yield (BNO) phenotype, wherein the markers include, but are not limited to, SEQ ID NO: 8, SEQ ID NO: 27, SEQ ID NO: 48, SEQ ID NO: 53, SEQ ID NO: 56, SEQ ID NO: 57, SEQ ID NO: 58, SEQ ID NO: 62, SEQ ID NO: 80, SEQ ID NO: 81, SEQ ID NO: 82, SEQ ID NO: 83, SEQ ID NO: 84, SEQ ID NO: 85, SEQ ID NO: 86, SEQ ID NO: 87, SEQ ID NO: 88, SEQ ID NO: 89, SEQ ID NO: 90, SEQ ID NO: 91, SEQ ID NO: 92, SEQ ID NO: 93, SEQ ID NO: 94, SEQ ID NO: 95, SEQ ID NO: 96, SEQ ID NO: 97, SEQ ID NO: 98, SEQ ID NO: 99, SEQ ID NO: 100, SEQ ID NO: 101, SEQ ID NO: 102, SEQ ID NO: 103, SEQ ID NO: 104, SEQ ID NO: 105, SEQ ID NO: 106, SEQ ID NO: 107, SEQ ID NO: 108, SEQ ID NO: 109, SEQ ID NO: 110, SEQ ID NO: 111, SEQ ID NO: 112, SEQ ID NO: 113, SEQ ID NO: 114, SEQ ID NO: 115. The polymorphic markers were selected according to methods as described herein. Exemplary combinations of markers and their uses in determining plant phenotypes can be found in the present specification, for example in Tables 6, 7 and 12.

In one example, the method as described herein provides SNPs associated with an increase/decrease in fresh fruit bunch yield, whereby the SNPs are one, or two, or more genetic markers including, but not limited to, SEQ ID NO: 1, SEQ ID NO: 3, SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 8, SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 13, SEQ ID NO: 14, SEQ ID NO: 18, SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21, SEQ ID NO: 23, SEQ ID NO: 26, SEQ ID NO: 27, SEQ ID NO: 30, SEQ ID NO: 33, SEQ ID NO: 34, SEQ ID NO: 35, SEQ ID NO: 36, SEQ ID NO: 37, SEQ ID NO: 39, SEQ ID NO: 40, SEQ ID NO: 41, SEQ ID NO: 43, SEQ ID NO: 44, SEQ ID NO: 45, SEQ ID NO: 46, SEQ ID NO: 47, SEQ ID NO: 48, SEQ ID NO: 51, SEQ ID NO: 52, SEQ ID NO: 53, SEQ ID NO: 54, SEQ ID NO: 56, SEQ ID NO: 57, SEQ ID NO: 58, SEQ ID NO: 60, SEQ ID NO: 62, SEQ ID NO: 63, SEQ ID NO: 64, SEQ ID NO: 65, SEQ ID NO: 66, SEQ ID NO: 67, SEQ ID NO: 70, SEQ ID NO: 71, SEQ ID NO: 72, SEQ ID NO: 73, SEQ ID NO: 75, SEQ ID NO: 76, SEQ ID NO: 77 and SEQ ID NO: 79.

In one example, the method as described herein provides SNPs, wherein the presence of said SNPs is associated with an increase or decrease in fresh fruit bunch yield and wherein the SNPs include, but are not limited to, Marker Nos. 1 (SEQ ID NO: 10), 2 (SEQ ID NO: 20), 3 (SEQ ID NO: 13), 4 (SEQ ID NO: 8), 5 (SEQ ID NO: 19), 6 (SEQ ID NO: 34), and combinations thereof. In a further example, the presence of at least one, at least two, at least three, at least four, at least five, or all markers including, but not limited to, Marker no. 1 (SEQ ID NO: 10), Marker no. 2 (SEQ ID NO: 20), Marker no. 3 (SEQ ID NO: 13), Marker no. 4 (SEQ ID NO: 8), Marker no. 5 (SEQ ID NO: 19) and Marker no. 6 (SEQ ID NO: 34) is associated with an increase in fresh fruit bunch weight yield. In another example, the presence of at least one, two, three or all markers including, but not limited to Marker no. 1 (SEQ ID NO: 10), Marker no. 3 (SEQ ID NO: 13), Marker no. 5 (SEQ ID NO: 19) and Marker no. 6 (SEQ ID NO: 34), is associated with a decrease in fresh fruit bunch weight yield. Examples of marker usage can be seen in Tables 4, 6 and 7 of the present description.

In one example, the SNPs as previously mentioned are described, wherein the SNPs associated with an increase or decrease in bunch number (BNO) yield are Marker Nos. 7 (SEQ ID NO: 90), 8 (SEQ ID NO: 105), 9 (SEQ ID NO: 106), 10 (SEQ ID NO: 99), 11 (SEQ ID NO: 80) and combinations thereof. In a further example, the SNPs are one or two or more genetic markers including, but not limited to, SEQ ID NO: 48, SEQ ID NO: 53, SEQ ID NO: 58, SEQ ID NO: 80, SEQ ID NO: 82, SEQ ID NO: 83, SEQ ID NO: 84, SEQ ID NO: 86, SEQ ID NO: 90, SEQ ID NO: 92, SEQ ID NO: 94, SEQ ID NO: 96, SEQ ID NO: 99, SEQ ID NO: 100, SEQ ID NO: 105, SEQ ID NO: 106 and SEQ ID NO: 113. In a further example, the SNPs are one, or two, or more genetic markers including, but not limited to SEQ ID NO: 80, SEQ ID NO: 90, SEQ ID NO: 99, SEQ ID NO: 105 and SEQ ID NO: 106.

In one example, the method as described herein, provides that the presence of a combination of Markers no. 8 (SEQ ID NO: 105) and 11 (SEQ ID NO: 80) is associated with a change in bunch number yield. In another example, the presence of a combination of Markers no. 7 (SEQ ID NO: 90), 9 (SEQ ID NO: 106) and 10 (SEQ ID NO: 99) is associated with a change in bunch number yield. In yet another example, the presence of Marker no. 11 (SEQ ID NO: 80) is associated with a change in bunch number yield. Examples of marker usage can be seen in Tables 10 and 12 of the present description.

Also described herein is use of one or more polymorphic genetic markers as criteria for the selection of a plant, wherein the selection is based on the presence or absence of one or more polymorphic genetic markers.

EXAMPLES

Non-limiting examples of the invention, including the best mode, and a comparative example will be further described in greater detail by reference to specific Examples, which should not be construed as in any way limiting the scope of the invention.

5

Materials and Methods

The plant materials for use in the 'fresh fruit bunch yield' and 'bunch number' marker identification methods were derived from three populations of oil palms, each population containing crosses of Deli, Nigerian and AVROS backgrounds and therefore having a mixture of different genetic backgrounds. The combination of alleles found in these three populations is a representation of the alleles present in commercial oil palm genetic materials. For example, the commercial materials sold and planted in Malaysia are a combination of different crosses of breeding populations of restricted origins (BPROs). The total parental BPROs used to generate the populations for the association mapping covers all the commonly found genetic backgrounds that are introgressed by plant breeders to create the commercial oil palm that is planted in Malaysia and Indonesia. The palms were individually yield recorded over a specified period of time.

DNA Extraction

Samples of spear leaves, which are the first unopened young leaves from the highest frond, were obtained from the oil palm trees. The leaves were cleaned and frozen in liquid nitrogen before being placed in a freezer (-80°C) for storage. When a DNA sample was required, a leaf sample weighing approximately 3-5 grams was removed from storage, treated with liquid nitrogen and ground to a powder in a mortar and pestle. The ground leaves were then transferred into a tube, into which 15 ml of modified cetyl trimethyl ammonium bromide (CTAB) buffer was added. The CTAB buffer contained 2% CTAB (weight/volume), 20 mM EDTA (pH 8.0), 1.4 M NaCl, 100 mM Tris-HCL (pH 8.0), 5 mM ascorbic acid, 4 mM diethyldithiocarbamic acid sodium salt and 2% polyvinylpyrrolidone-40 and 100 µl of β-mercaptoethanol. The mixture was incubated at 60°C for 30 minutes, after which an equal volume of chloroform: isoamyl alcohol was added before mixing thoroughly. The mixture was centrifuged at 10,000 rpm for 15 minutes and the upper aqueous phase was transferred into a fresh tube. DNA was then precipitated with the addition of 0.6x volume of ice-cold isopropanol. The solution was then stored in -20°C for a few hours and then centrifuged at 12,000 rpm for 15 minutes. The DNA pellet, seen at the bottom of the tube, was washed in 5 ml wash buffer containing 70% ethanol and 10 mM ammonium acetate. The sample was then centrifuged to precipitate the DNA, dried and dissolved in Tris-EDTA (TE) buffer containing

10 mM Tris-HCl (pH 8.0) and 1 mM EDTA (pH 8.0). The DNA solution was then treated with RNase to remove RNA.

Marker Identification

5 The extracted DNA was then sequenced using the Illumina Genome Analyser as per manufacturer's instruction. The data was analysed using bioinformatics tools such as but not limited to Burrows-Wheeler Aligner and SAMTOOLS to detect possible SNPs. A list of SNPs was sent to Illumina, Inc., and 1,208 SNP markers were arrayed onto a GoldenGate™ assay format. Illumina's method was used to perform the GoldenGate™ analysis using the DNA of
10 the 269 palms, from three populations as defined above with 96 palms from a first population, 96 from a second population and 77 from a third population.

 The results of the analysis were then interpreted using Illumina's Genome Studio™ - Genotyping Module. The results, which showed the clustering of the alleles, were transferred onto a spread sheet and further analysed as described below.

15

Association Mapping

 The SNP markers identified above were then statistically tested to identify if they are linked to phenotype data and by extension, able to predict phenotype performance and differentiate genotypes that influence traits and those that do not. Statistical linear models were used to assess if there is a significant link between the genotype and the phenotype tested. The
20 genotype data were obtained from the GoldenGate™ analysis described above and the phenotype data were collected over a period of 54 months to increase the accuracy of the data. The phenotypes used in this analysis were fresh fruit bunch yield and bunch number. The analysis was conducted using the general linear model (GLM) to identify SNP genetic
25 markers that were linked to two phenotypes, fresh fruit bunch yield data and bunch number. To improve accuracy, the mixed linear model (MLM) was used together with the GLM model. SNP markers significantly associated with yield traits were shortlisted using general linear model (GLM) and mixed model association (MLM) analysis with TASSEL (Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. *Bioinformatics*; 2007; 23:2633-
30 2635.) program that combines genotype, phenotypic data containing fresh fruit bunch yield data and bunch number, and both pairwise kinship coefficients (K-matrix) derived from SPAGeDi (Hardy OJ, Vekemans X; 2002); SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels (Molecular Ecology Notes 2; 618-620) and structural information (Q-value) from the STRUCTURE software (Pritchard, J. K., Stephens, M., Rosenberg, N. A., Donnelly, *Am. J. Hum. Genet.*; 2000; 67: 170-181) to
35 decrease the possibility of spurious associations due to population stratification. The parameter

used to run STRUCTURE was 100,000 repetitions for both the BURN-IN and MCMC parameters. The K value was determined using the analysis proposed by Evanno *et al.* (Evanno, G., Regnaut, S., Goudet, J.; *Molecular Ecology*; 2005; 14: 2611-2620). P-values for the markers were obtained and studied. A lower the p-value represents a higher chance that the marker is linked to the traits of interest. A polymorphic site with a p-value below the 5% or 1% empirically derived value is considered to have a statistically significant association to the trait (Greenhalgh T; *BMJ* 315; 1997; 540-543.; Tommasini L., Schnurbusch T., Fossati D., Mascher F., Keller B. , *Theor. Appl. Genet.*;2007; 115: 697–708.). As a result, the cut-off point was determined to be 0.01 and the best few markers with the acceptable p-value were selected for further analysis.

For the selected markers, the genotype of the plant was compared with the trait of interest. Effect estimate was calculated to determine the effect of that allele either positively or negatively associated to the trait of interest.

Validation

The shortlisted markers from performing the association mapping step were validated on oil palms from commercially available planting materials available at oil palm plantations. The samples were genotyped using the “OpenArray” platform and analysed using the association mapping methods described above. The results of the analysis provided a list of markers, the different alleles that the marker can detect, and a measure of how accurately the marker can be used to predict a trait of interest (p-value), how strongly they affect the trait that they are linked to (allele effect), and how strongly the marker fits into a mathematical model to explain the variation that the marker causes to the overall trait that is tested (r-squared).

Markers were shortlisted based on a cut-off p-value of <0.01, R-squared (R^2) cut-off value of <0.1 and their predicted allele effect which predicts how much the oil palm is affected by the presence of specific alleles. The r-squared value can be derived using standard statistical methods and the closer the value is to 1, the stronger influence the marker has on the overall total effect of the trait tested. The allele effect predicts whether that particular combination of allele will give a positive or negative change to the tested plant as compared to an average oil palm. Larger positive and larger negative values will be more advantageous for screening and selecting oil palms that have the desired allele and lack the unwanted alleles.

Ideally the selected oil palms should only contain alleles predicted to give positive effects. Care should be taken to avoid oil palms carrying alleles predicted to have a negative impact on the trait tested, but in practice, this may be difficult as such an ideal genotype combination will be very rare. For the purpose of selection of alleles, allelic combinations that gave the best predicted value for allele effect were selected, and those that were predicted to

give the strongest negative influence on the trait tested were removed. The data from the genotyping, combined with the prediction of the association mapping together with the phenotype data were analysed to check if the predictions of the shortlisted markers were accurate and if oil palm phenotype performance can be predicted based on the genotype.

5 Based on the validation trial performed, the palms that were identified to be excellent for the trait being tested were analysed, and their trait data were compared to the average value for that trait to show that selection with these markers will allow for accurate prediction of oil palm that have the potential for increased performance for the trait being evaluated. The method described above was used to identify genetic markers linked to two phenotypic traits
10 (FFB yield and bunch number). The results are described in Example 1 and Example 2 below.

Results

Example 1: FFB yield associated markers

Identification of markers to predict plant performance in relation to FFB yield was
15 carried out to select high yielding plants for planting as well as breeding.

Genotyping of the 269 palms of the three populations as mentioned previously, was carried out as described above. Phenotyping of the FFB yield was carried out by measuring the weight of the fruit bunches produced by individual oil palms. Association mapping using both GLM and MLM analysis was also performed.

20 To determine the marker(s) which are positively associated to FFB yield, a cut-off value was required so that the number of selected markers is manageable. If the cut-off value is too high, it will be difficult to find enough plants that satisfy the criteria but the effect on the phenotype will be higher. Conversely, if the cut-off value is too low, it will be easier to find plants that fit the criteria but there will be too many markers, some of which will only have a
25 small improvement (increase) in the trait being tested. The cut-off value used in this example was a p-value <0.01, in order to produce a list containing 79 shortlisted markers that are putatively related to fresh fruit bunch (FFB) yield trait. The 79 shortlisted markers predicted to be associated to the FFB trait together with the GLM and MLM predicted value for the p-value are listed in Table 2 below.

30

Table 2: GLM (general statistical linear model) and MLM (mixed statistical linear model) values of markers tested for association with the fresh fruit bunch (FFB) yield trait

SEQ ID No.	Association Mapping Analysis	
	GLM p-value	MLM p-value
1	7.08E-04	1.42E-04

2	3.26E-03	2.83E-04
3	1.27E-06	2.76E-03
4		5.97E-03
5		7.63E-03
6	4.04E-06	8.83E-03
7	5.74E-03	9.72E-03
8	2.96E-03	
9	8.22E-03	
10	3.17E-06	
11	1.39E-03	
12	1.19E-05	
13	1.71E-03	
14	2.85E-04	
15	6.65E-03	
16	6.65E-03	
17	6.65E-03	
18	9.25E-03	
19	5.21E-03	
20	8.04E-04	
21	7.27E-06	
22	3.48E-03	
23	1.47E-03	
24	8.42E-03	
25	3.44E-04	
26	5.37E-03	
27	9.38E-04	
28	3.39E-03	
29	1.01E-03	
30	8.87E-04	
31	3.27E-03	
32	4.55E-04	
33	4.12E-03	
34	5.61E-04	
35	3.92E-03	
36	1.66E-03	
37	2.37E-03	
38	8.31E-03	
39	1.28E-03	
40	4.19E-03	
41	2.68E-03	
42	6.34E-03	
43	4.09E-03	
44	5.15E-03	
45	4.73E-05	

46	2.15E-03	
47	1.09E-04	
48	7.28E-05	
49	8.69E-03	
50	9.41E-03	
51	3.82E-05	
52	5.30E-03	
53	3.45E-03	
54	1.91E-05	
55	4.25E-03	
56	1.27E-03	
57	4.06E-03	
58	2.45E-03	
59	3.45E-03	
60	8.20E-03	
61	3.12E-03	
62	7.57E-03	
63	1.40E-03	
64	1.47E-03	
65	2.82E-03	
66	5.23E-03	
67	5.57E-04	
68	2.05E-05	
69	6.48E-03	
70	1.51E-03	
71	1.61E-03	
72	3.56E-04	
73	8.04E-03	
74	6.54E-03	
75	1.97E-03	
76	1.82E-03	
77	6.68E-03	
78	9.85E-03	
79	9.68E-03	

The 79 markers were validated on a population of 225 palms which were planted as three separate experimental blocks consisting of 75 oil palms per block. As these three blocks of planting materials were from different origins, these three blocks would not be considered as a replication, but as a representative of the diversity of different oil palm populations that are indicative of the diversity found in commercial materials in the oil palm industry.

Genotyping was conducted so that the allelic content of the individual oil palms were identified. Phenotype data for these palms were collected. The genotyping was carried out

using the 79 markers that have been identified as listed in Table 2 above, and the various alleles that are linked to the markers were identified on an individual tree basis. As multiple alleles may be possible for each of the 79 markers, a cut-off was applied based on the R-squared (R^2) value. All alleles that gave an R^2 value of <0.1 were removed in order to reduce the complexity of the data and simplify the analysis. The R^2 value is a description of how much a particular allele explains how well the data points fits the statistical model used to predict the variations in the trait or phenotype being determined. Therefore, an allele which has a low R^2 value will not be particularly accurate and thus, not useful for the purpose stated herein. For this example, we used a value of 0.1 as a cut-off point and all alleles which were below 0.1 were not considered useful for further selection. However, it is known that different values of the R^2 value can be accepted, depending on the how many markers is feasible for the selection of the trait or how much accuracy is required for the prediction.

The R^2 value was determined and applied on an individual population basis, where the markers will result in different R^2 values when tested in the different populations. This is because some markers are better predictors for traits in certain populations and some markers functions well across populations. By determining the R^2 value of markers based on population markers for population-specific traits, markers for universal traits can be identified, allowing for greater freedom in determining which combination of markers should be used to predict traits in a given population.

After applying the p-value cut-off of less than 0.01, 53 of the 79 markers were still retained as being good predictors of the FFB trait.

Table 3: R^2 Value for markers that have a p-value of <0.01 in the 3 tested populations

SEQ ID No.	R^2 of Population A	R^2 of Population B	R^2 of Population C
58	0.010	0.223	0.020
67	0.006	0.151	0.025
45	0.028	0.090	0.235
60	0.006	0.175	0.004
35	0.014	0.056	0.136
76	0.026	0.133	0.067
71	0.005	0.142	0.051
30	0.022	0.082	0.169
75	0.008	0.108	0.200
10	0.103	0.155	0.000
44	0.033	0.122	0.008
20	0.133	0.076	0.187
1	0.061	0.119	0.015

39	0.006	0.132	0.001
72	0.013	0.101	0.019
66	0.023	0.101	0.130
13	0.125	0.062	0.001
12	0.055	0.188	0.006
53	0.004	0.116	0.040
63	0.023	0.076	0.235
77	0.011	0.134	0.004
73	0.017	0.109	0.153
36	0.007	0.195	0.004
47	0.001	0.163	0.016
54	0.026	0.163	0.136
56	0.010	0.111	0.134
3	0.075	0.182	0.015
51	0.032	0.130	0.001
26	0.004	0.140	0.000
65	0.011	0.116	0.039
79	0.065	0.170	0.036
21	0.003	0.159	0.155
37	0.011	0.113	0.121
52	0.059	0.126	0.017
7	0.098	0.047	0.235
62	0.002	0.122	0.002
14	0.040	0.185	0.007
8	0.110	0.042	0.000
6	0.094	0.141	0.018
48	0.008	0.147	0.118
27	0.006	0.146	0.064
40	0.018	0.047	0.125
33	0.019	0.210	0.015
19	0.132	0.047	0.029
18	0.046	0.103	0.046
23	0.007	0.117	0.173
57	0.035	0.081	0.118
46	0.018	0.151	0.004
70	0.013	0.128	0.024
41	0.029	0.169	0.002
64	0.004	0.122	0.022
34	0.104	0.053	0.006
43	0.005	0.169	0.003

The genotype of the plants from the association mapping sample set can be either homozygous positive, homozygous negative or heterozygous alleles. These data were obtained

from analysing the different alleles present in the population. The data were used to predict FFB yield performance.

Unlike qualitative traits where a phenotype trait is influenced by a single gene, yield is a quantitative trait that is affected by the interaction and integration of many component traits (multiple genes) that can interact with variations in environmental conditions such as rainfall, terrain, soil type, temperature and solar radiation. Hence, to increase the accuracy of selecting high yielding oil palms in order to increase overall average yield of a planting area, multiple markers were used. Markers with allele effects that are considered minor were also included as a combination of these 'minor' variants can be used to predict a significant change to the expected value.

The identified markers were then used to screen oil palm populations for plants that are predicted to be high yielding. At least one, at least two, at least three or more markers were selected that indicated a positive/negative allele effect. The effects are more enhanced with more markers and the number of markers used will depend on the sum total of the R-squared value. The cut-off point may be selected by taking into consideration the cost linked to the number of markers used as opposed to the accuracy of using a larger number of markers. In other words, in case a larger number of markers, the more accurate the prediction, but the higher the cost of running the experiment.

The sum value of the R-square indicate how much of the variation can be explained by the allele combination in a regression model. For example, if the combination of markers gives a sum R-square of 1 then the combination of markers will accurately predict the phenotype change. If the sum R-square is 0.3 for example, that means the marker can predict with 30% accuracy based on the regression model for that trait.

Ideally the markers with the highest R-square value, such a R-square value of 1, or a R-square value of 0.9, or a R-square value of 0.8, or a R-square value of 0.7, or a R-square value of 0.6, or a R-square value of 0.5, or a R-square value of 0.4, or a R-square value of 0.3, or a R-square value of 0.2, or a R-square value of 0.1 may be selected with the highest marker effects and recursively additional markers may be selected until the sum R-squared reaches an accuracy of prediction that is reasonable and reproducible. This however is subjective as it depends on the practicality of the measurement.

If the range of effect size is very small, such as for example, the difference between 100kg and 110kg (in other words 10kg is an effect size of effect size), then it would not be desirable to use many markers as accuracy would not enable the selection of plants, such as palms that have significant increase/decrease to the trait of interest. If the range of variation is very large (e.g. the difference between 50kg – 150kg), then it would be desirable to select more markers so that the plants with the most significant trait increase or decrease will be selected.

At least one, at least two, at least three, at least four or more markers that indicated a positive allele effect were selected. Markers were recursively added in to balance the highest possible R-squared value with the lowest possible markers. The population tested was screened for plants that carry the allele linked to good allele effects, and plants that carry that allele were selected. In this example, six markers selected, as shown in Table 4, which also shows the different alleles associated with the marker and the estimated GLM value for each of the allele observed.

Table 4: Shortlisted markers that were selected for screening a population of palms

Marker No.	SEQ ID No.	Observed palms with specific genotype	Allele Combination	GLM
1	10	50	C:C	13.68
	10	196	T:T	-22.22
	10	23	C:T	0
2	20	56	A:T	25.8187037
	20	137	A:A	20.3456752
	20	74	T:T	0
3	13	219	T:T	-36.82
	13	32	A:T	-34.72
	13	18	A:A	0
4	8	76	A:G	19.8369966
	8	193	G:G	0
5	19	50	A:A	19.741822
	19	180	G:G	-2.4188407
	19	26	A:G	0
6	34	189	C:T	14.5823867
	34	58	C:C	-10.50
	34	22	T:T	0

For validation of the markers, six markers were selected as best examples and were used to genotype a total of 225 oil palms of various backgrounds comprising of mixes of Deli, AVROS and Nigerian pedigree. These plants were individually yield recorded for 54 months and the average FFB yield of each plant was calculated. The allele in question for each of the six markers is listed in Table 5.

Table 5: Genotyping results for the six markers for the validation population of 225 palms.

Oil Palm #	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6	Average Yield/year
1	T:T	T:T	T:T	G:G	A:G	C:T	375.04
2	C:C	A:A	A:A	A:G	A:A	C:T	368.07
3	C:C	A:T	T:T	A:G	A:A	C:T	343.38
4	C:C	A:T	T:T	G:G	A:A	C:T	343.04
5	T:T	A:A	T:T	A:G	A:G	C:T	335.98
6	T:T	A:T	T:T	A:G	G:G	C:T	332.16
7	C:C	A:A	A:A	A:G	A:A	C:T	323.56
8	C:T	T:T	T:T	G:G	A:A	T:T	319.58
9	C:C	A:A	T:T	A:G	G:G	C:T	309.78
10	C:C	A:T	T:T	A:G	G:G	C:T	306.64
11	C:C	A:A	A:A	G:G	G:G	T:T	305.22
12	T:T	T:T	T:T	A:G	G:G	C:T	305.08
13	C:T	A:A	T:T	A:G	G:G	C:T	303.33
14	T:T	T:T	A:A	A:G	G:G	C:T	301.71
15	T:T	A:T	T:T	G:G	G:G	C:C	301.18
16	T:T	A:A	A:A	A:G	A:A	C:T	299.64
17	T:T	A:A	T:T	G:G	G:G	C:T	295.98
18	C:C	A:T	A:A	G:G	A:A	C:T	291.07
19	T:T	T:T	T:T	A:G	A:A	C:T	290.38
20	C:C	A:A	A:T	A:G	G:G	T:T	288.31
21	T:T	A:A	T:T	G:G	G:G	C:T	286.64
22	T:T	A:A	T:T	A:G	A:A	C:T	283.18
23	C:C	A:A	A:A	A:G	A:A	C:T	281.38
24	T:T	A:A	T:T	G:G	G:G	C:T	278.84
25	T:T	A:A	A:A	A:G	A:A	C:T	278.29
26	T:T	A:T	T:T	G:G	G:G	C:T	278.02
27	T:T	A:A	T:T	G:G	G:G	T:T	277.91
28	T:T	A:A	T:T	G:G	G:G	C:T	277.67
29	T:T	A:A	A:A	A:G	A:A	C:T	277.13
30	C:C	A:T	T:T	G:G	A:G	C:T	276.8
31	T:T	A:A	T:T	G:G	G:G	C:T	276.18
32	C:C	T:T	A:T	A:G	A:A	C:T	274.29
33	C:C	A:T	A:T	G:G	N:N	T:T	273.87
34	C:C	T:T	A:T	A:G	A:A	C:T	273.64
35	C:C	A:A	A:A	G:G	N:N	C:T	271.78
36	C:C	A:T	A:T	A:G	A:A	T:T	270.18
37	C:C	A:A	A:T	A:G	A:A	C:T	269.81
38	T:T	T:T	T:T	G:G	A:G	C:T	269.67
39	C:C	A:T	A:A	G:G	A:A	C:T	269.22
40	T:T	A:A	T:T	G:G	G:G	C:T	268.49
41	T:T	T:T	T:T	G:G	G:G	C:C	268.22

42	T:T	A:A	T:T	G:G	G:G	C:T	268.16
43	T:T	A:A	T:T	G:G	G:G	C:T	267.32
44	C:C	T:T	T:T	G:G	N:N	C:T	266.93
45	T:T	A:A	T:T	A:G	G:G	C:T	266.91
46	T:T	A:T	A:T	A:G	G:G	C:T	266.72
47	T:T	A:A	T:T	G:G	G:G	C:T	266.4
48	T:T	A:A	T:T	A:G	A:A	C:C	266.27
49	T:T	A:A	A:T	A:G	A:A	C:T	266.02
50	T:T	A:T	T:T	G:G	G:G	C:C	263.84
51	T:T	A:A	T:T	A:G	G:G	C:T	263.81
52	T:T	A:A	T:T	G:G	G:G	C:T	262.67
53	C:C	A:T	T:T	A:G	G:G	C:T	262.54
54	T:T	A:A	T:T	G:G	G:G	C:C	259.72
55	T:T	A:A	T:T	G:G	G:G	C:T	259.56
56	T:T	T:T	T:T	G:G	G:G	C:C	259.22
57	C:C	A:T	T:T	G:G	A:A	C:T	258.6
58	C:T	A:T	T:T	G:G	A:A	C:C	258.22
59	T:T	A:A	T:T	G:G	G:G	C:T	257.93
60	C:C	T:T	A:A	G:G	G:G	C:T	257.62
61	C:C	A:A	T:T	A:G	G:G	C:T	257.09
62	T:T	A:T	T:T	A:G	G:G	C:T	256.88
63	T:T	A:T	T:T	G:G	G:G	C:T	256.36
64	C:C	T:T	A:T	G:G	G:G	C:T	255.56
65	T:T	N:N	T:T	G:G	G:G	C:T	255.47
66	T:T	T:T	T:T	G:G	G:G	C:C	254.36
67	T:T	A:A	T:T	G:G	G:G	C:T	254
68	T:T	A:A	T:T	G:G	G:G	C:T	253.78
69	C:T	T:T	T:T	G:G	A:A	C:T	253.04
70	T:T	A:A	T:T	G:G	G:G	C:T	252.93
71	T:T	A:A	T:T	G:G	G:G	C:T	252.78
72	C:C	A:T	T:T	G:G	G:G	T:T	252.67
73	C:T	A:T	A:T	A:G	N:N	C:T	251.76
74	T:T	A:A	T:T	G:G	G:G	C:T	251.17
75	C:C	T:T	T:T	A:G	N:N	C:T	250.09
76	C:T	T:T	T:T	G:G	A:A	C:T	249.38
77	T:T	A:T	T:T	A:G	A:A	C:T	249.36
78	T:T	A:T	T:T	G:G	G:G	C:C	249.28
79	C:T	A:T	A:T	G:G	G:G	C:T	249.09
80	T:T	A:A	A:T	G:G	A:G	C:T	247.63
81	C:C	A:T	A:T	G:G	G:G	C:T	247.16
82	T:T	A:A	T:T	G:G	G:G	C:T	247.07
83	T:T	T:T	A:A	A:G	A:A	C:T	244.71
84	T:T	A:A	T:T	G:G	G:G	C:T	243.8
85	C:C	A:A	T:T	G:G	N:N	C:T	243.6

86	T:T	A:A	T:T	G:G	G:G	C:T	243.47
87	T:T	A:A	A:T	G:G	G:G	C:C	242.36
88	T:T	A:A	T:T	G:G	G:G	C:C	241.98
89	C:C	A:A	A:T	G:G	A:A	C:T	241.96
90	T:T	A:T	T:T	A:G	G:G	C:T	241.84
91	T:T	T:T	A:T	A:G	G:G	C:C	241.22
92	T:T	A:T	T:T	A:G	G:G	C:C	240.93
93	T:T	A:A	A:T	A:G	A:A	C:T	240.11
94	T:T	A:T	T:T	G:G	G:G	C:T	239.91
95	C:C	A:T	T:T	G:G	A:A	C:T	239.07
96	T:T	A:T	T:T	G:G	G:G	C:T	238.18
97	T:T	A:A	T:T	A:G	G:G	C:T	237.91
98	T:T	A:A	T:T	G:G	G:G	C:T	237.6
99	C:C	T:T	T:T	G:G	A:A	C:C	237.47
100	T:T	A:A	T:T	G:G	G:G	C:T	237.24
101	T:T	A:T	T:T	G:G	G:G	C:T	237.16
102	T:T	A:T	T:T	G:G	G:G	C:T	236.63
103	T:T	T:T	T:T	G:G	A:G	C:C	236.07
104	T:T	A:A	T:T	A:G	A:G	C:C	234.91
105	T:T	A:A	T:T	G:G	G:G	C:T	234.56
106	T:T	A:A	T:T	G:G	G:G	C:T	234.16
107	T:T	T:T	T:T	G:G	G:G	C:T	233.56
108	T:T	A:A	T:T	G:G	G:G	C:T	232.92
109	T:T	A:T	T:T	A:G	A:G	C:T	232.56
110	T:T	A:T	T:T	G:G	A:A	C:C	232.42
111	T:T	A:T	T:T	G:G	G:G	C:T	232.4
112	C:C	T:T	T:T	G:G	A:G	T:T	232.04
113	T:T	N:N	T:T	G:G	A:G	C:T	232.02
114	T:T	A:T	A:T	G:G	G:G	T:T	231.73
115	T:T	A:A	T:T	G:G	G:G	T:T	231.69
116	T:T	A:A	T:T	G:G	G:G	C:T	231.42
117	C:T	T:T	T:T	G:G	N:N	C:C	231.2
118	T:T	T:T	T:T	G:G	A:A	C:C	230.93
119	T:T	A:A	T:T	G:G	G:G	C:T	230.41
120	T:T	A:A	T:T	G:G	G:G	C:T	230.2
121	T:T	A:T	T:T	A:G	G:G	C:T	229.87
122	C:C	A:A	T:T	A:G	G:G	C:T	229.52
123	T:T	A:A	T:T	A:G	A:G	C:T	229.38
124	T:T	A:T	T:T	A:G	G:G	C:T	229.15
125	T:T	A:A	T:T	G:G	G:G	C:T	229.04
126	T:T	T:T	T:T	G:G	A:A	C:C	228.96
127	T:T	A:A	T:T	G:G	G:G	C:T	228.22
128	C:C	A:A	A:A	G:G	A:A	C:T	228.2
129	T:T	A:A	T:T	G:G	G:G	C:T	227.69

130	T:T	T:T	T:T	A:G	G:G	C:T	227.44
131	T:T	A:A	T:T	G:G	G:G	C:T	227.11
132	T:T	T:T	T:T	A:G	N:N	C:T	226.18
133	C:C	A:T	T:T	G:G	G:G	T:T	225.98
134	T:T	A:T	T:T	G:G	A:G	C:C	225.27
135	T:T	T:T	T:T	G:G	A:A	C:T	225.05
136	T:T	A:A	T:T	G:G	G:G	C:T	224.89
137	T:T	A:A	T:T	G:G	G:G	C:T	224.42
138	C:T	A:T	T:T	G:G	G:G	C:T	223.82
139	T:T	T:T	T:T	A:G	G:G	C:T	223.49
140	T:T	A:A	T:T	G:G	G:G	C:T	222.42
141	T:T	T:T	A:T	A:G	G:G	C:C	222.31
142	T:T	A:A	T:T	G:G	G:G	C:T	221.38
143	C:T	A:A	T:T	G:G	G:G	T:T	220.73
144	T:T	A:A	T:T	G:G	G:G	C:C	219.78
145	C:T	A:T	T:T	A:G	G:G	C:C	219.53
146	T:T	T:T	T:T	A:G	A:A	C:T	218.91
147	T:T	A:A	T:T	G:G	A:G	C:T	218.6
148	C:C	T:T	T:T	G:G	A:G	C:T	217.44
149	C:C	A:A	T:T	G:G	A:A	C:T	217.31
150	T:T	A:A	T:T	G:G	G:G	C:T	217.28
151	T:T	A:A	T:T	G:G	G:G	C:T	216.6
152	T:T	A:A	T:T	G:G	G:G	C:T	216.58
153	C:C	A:A	A:A	G:G	G:G	C:T	216.13
154	T:T	A:A	T:T	A:G	A:G	C:C	215.49
155	T:T	A:A	A:T	G:G	G:G	C:T	215.47
156	C:C	T:T	T:T	A:G	G:G	C:T	214.91
157	T:T	A:A	T:T	G:G	G:G	T:T	214.47
158	T:T	A:A	A:A	A:G	A:A	T:T	214.34
159	C:T	T:T	T:T	G:G	A:A	C:C	214.2
160	T:T	A:A	T:T	G:G	G:G	C:T	214
161	T:T	A:T	T:T	G:G	A:G	C:T	213.74
162	C:C	A:T	T:T	G:G	A:G	C:T	213.09
163	T:T	T:T	T:T	A:G	A:A	C:T	212.87
164	C:C	A:A	A:T	G:G	G:G	C:T	212.64
165	T:T	A:A	T:T	A:G	G:G	C:T	212.36
166	T:T	A:A	T:T	G:G	G:G	C:T	211.56
167	T:T	T:T	T:T	A:G	G:G	C:T	211.29
168	T:T	T:T	T:T	G:G	A:A	C:C	210.76
169	T:T	A:A	T:T	G:G	G:G	C:T	210.33
170	T:T	A:A	T:T	G:G	G:G	C:T	209.86
171	C:C	A:T	A:T	A:G	A:G	C:T	208.11
172	T:T	A:A	T:T	G:G	G:G	C:T	207.3
173	T:T	A:T	T:T	G:G	G:G	C:C	205.82

174	T:T	A:A	T:T	G:G	G:G	C:T	205.78
175	C:T	T:T	T:T	A:G	A:A	C:T	205.49
176	C:C	A:T	T:T	G:G	G:G	C:C	203.76
177	T:T	T:T	T:T	A:G	G:G	C:T	203.69
178	T:T	T:T	T:T	A:G	G:G	C:T	203.62
179	T:T	T:T	A:T	A:G	G:G	C:T	203.51
180	T:T	A:A	T:T	G:G	G:G	T:T	203.13
181	T:T	T:T	T:T	G:G	G:G	C:C	201.62
182	T:T	T:T	T:T	G:G	G:G	T:T	200.18
183	T:T	T:T	T:T	G:G	G:G	C:C	199.62
184	T:T	A:A	T:T	G:G	G:G	C:T	199.04
185	T:T	A:A	T:T	G:G	A:G	C:T	198.96
186	T:T	A:A	A:T	G:G	A:G	C:T	198.76
187	T:T	A:A	T:T	G:G	A:G	C:T	198.38
188	T:T	A:A	T:T	G:G	G:G	C:T	197.25
189	C:C	A:A	T:T	G:G	A:A	C:C	197.2
190	T:T	A:A	T:T	G:G	G:G	C:T	195.64
191	T:T	A:T	A:T	A:G	A:G	C:T	195.24
192	C:T	T:T	T:T	G:G	G:G	C:T	194.78
193	T:T	A:A	T:T	G:G	G:G	C:T	194.5
194	T:T	T:T	T:T	A:G	G:G	C:C	192.8
195	T:T	A:A	A:T	G:G	A:G	C:C	192.69
196	T:T	A:A	T:T	G:G	G:G	C:T	192.2
197	T:T	A:A	T:T	G:G	A:A	C:T	191.95
198	T:T	T:T	T:T	G:G	N:N	C:C	190.62
199	C:T	T:T	T:T	G:G	G:G	C:C	189.53
200	T:T	T:T	A:A	A:G	G:G	C:T	189.45
201	T:T	A:A	T:T	G:G	G:G	C:C	188.82
202	T:T	A:A	T:T	G:G	G:G	C:T	187.54
203	T:T	A:T	T:T	G:G	G:G	C:C	185.13
204	T:T	A:A	T:T	G:G	G:G	C:T	184.98
205	T:T	T:T	T:T	G:G	G:G	C:C	184.13
206	T:T	A:A	T:T	G:G	G:G	C:T	181.96
207	T:T	T:T	T:T	G:G	G:G	T:T	180.53
208	C:T	A:A	A:T	G:G	N:N	C:T	179.33
209	T:T	A:A	T:T	G:G	G:G	C:T	178.31
210	T:T	A:A	T:T	A:G	G:G	C:C	177.78
211	T:T	A:A	T:T	G:G	G:G	C:T	175.31
212	T:T	T:T	T:T	G:G	G:G	C:T	171.64
213	T:T	A:A	T:T	G:G	G:G	C:T	171.18
214	C:T	T:T	T:T	G:G	G:G	C:T	168.54
215	T:T	A:A	T:T	G:G	A:A	C:C	166.64
216	T:T	T:T	T:T	A:G	G:G	C:T	165.89
217	T:T	T:T	A:T	G:G	G:G	C:C	163.42

218	T:T	A:A	T:T	A:G	N:N	T:T	160.22
219	T:T	A:T	T:T	G:G	G:G	C:T	156.62
220	T:T	T:T	T:T	G:G	N:N	C:C	151.09
221	T:T	T:T	T:T	G:G	A:A	C:C	125.8
222	T:T	T:T	T:T	G:G	G:G	C:C	121.51
223	T:T	T:T	A:T	G:G	G:G	C:T	98.64
224	C:C	T:T	T:T	G:G	N:N	T:T	82.76
225	T:T	T:T	T:T	G:G	G:G	C:T	77.56

Plants were selected based on the allele criteria of the markers in Table 4 above, and the FFB average of the selected plants were recalculated to determine if there is an improvement over the average FFB for the total unselected population. The average FFB of the 225 palms that were unselected was 233.8 kg/palm/year.

If palms were selected based on Marker 1 by choosing palms that contain the positive effect allele combination C:C and discarding palms that have negative allelic effect T:T and zero allelic effect combination C:T then 45 palms will be selected from the 225 palms and the average yield would be 255.77 kg/palm/year which is an improvement of FFB by 21.97 kg/palm/year.

If palms were selected based on Marker 2 by choosing palms that contain the positive effect allele combination A:T or A:A then 161 palms will be selected from the 225 palms and the average yield will be 239.66 kg/palm/year which is an improvement by 5.86 kg/palm/year.

For selection of palms using markers that only have negative effect allele combinations, palms containing these allele combinations will be rejected for selection. An example is the use of Marker 3 which contains negative effect allele combination T:T and A:T where 208 palms will be rejected and 17 palms will be selected from the 225 palms and the average yield will be 271.58 kg/palm/year which is an improvement by 37.7 kg/palm/year.

If palms were selected based on Marker 4 which contains the positive effect allele combination A:G and discarding palms that contains the negative effect allele combination G:G then 67 palms will be selected from the 225 palms and the average yield will be 249.71 kg/palm/year which is an improvement by 15.91 kg/palm/year.

If palms were selected based on Marker 5 which contains the positive effect allele combination A:A and discarding palms that contains the negative effect allele combination G:G and the neutral effect allele combination A:G then 45 palms will be selected from the 225 palms and the average yield will be 252.91 kg/palm/year which is an improvement by 19.11 kg/palm/year.

If palms were selected based on Marker 6 which contains the positive effect alleles combination C:T and discarding palms that contains the negative effect allele combination C:C

and neutral allele combination T:T then 161 palms will be selected from the 225 palms and the average yield will be 238.97 kg/palm/year which is an improvement by 5.1 kg/palm/year.

The selection of palms can be done using either a single marker or a plurality of markers. Table 6 shows the yield gain for FFB if the validation population of 225 palms were
5 selected using a plurality of markers.

Table 6: Validation of using a plurality of markers for selection

Marker Combinations Used	Palms selected from 225 palms	Average FFB yield (kg/palm/year) of selected subpopulation	Improvement over mean FFB yield (kg/palm/year) in unselected population
All 6	3	324.34	90.54
1,2,4,5 and 6	5	317.24	83.44
1,2,3,4 and 5	3	324.34	90.54
1,3 and 5	6	293.58	59.79
2,4 and 6	34	268.81	35.01
2 and 6	124	242.14	8.35
1 and 3	10	281.23	47.43
1 and 5	18	273.75	39.95
1 and 4	17	278.31	44.52

To further determine if the markers obtained can predict for alleles that should be
10 avoided in the selection of oil palms, the same validation test was done on the same validation population, where instead of choosing the positive allele effect combinations, the negative allele effect was chosen. This is similar as the allele selection that was done on Marker 3 as in the example above. The results are shown in Table 7, where it can be observed that selecting
15 negative allele effect does have a detrimental effect on the trait tested for.

Table 7: Validation of using a plurality of markers for selection

Markers	Allele Combination Selected and allele effect	Average FFB yield (kg/palm/year) of selected subpopulation	Reduction over mean FFB yield (kg/palm/year) in unselected population
1	T:T (-22.22)	227.99	5.80
3	T:T (-36.82) and A:T (-34.72)	230.7	3.09
5	G:G (-2.52)	229.48	4.31
6	C:C (-10.50)	216.54	17.26
Combination of 1,3,5 and 6	As above	221.09	12.71

The above examples clearly show that oil palm trait prediction can be made by searching for plants that carry alleles that are predicted to be beneficial to the trait (positive allele effect) being tested, and avoiding alleles with detrimental effect on the traits tested (negative allele effect).

5

Example 2: Bunch number associated markers

Identification of markers for predicting bunch number in oil palm was carried out using the same method as described in Example 1.

Using the methods described above, 44 markers were shortlisted (p-value <0.01) to be putatively related to bunch number of oil palms using the same plant materials as in Example 1. These are listed in Table 8 below.

10

Table 8: GLM and MLM values of markers tested for association with the bunch number (BNO) yield trait

SEQ ID No.	GLM	MLM
80	2.52E-04	2.42E-03
81	3.56E-04	3.29E-03
56	3.71E-04	
82	5.73E-04	1.28E-03
83	9.50E-04	2.04E-03
84	1.30E-03	4.72E-03
85	1.45E-03	
86	1.57E-03	
87	1.60E-03	6.55E-03
88	2.66E-03	1.23E-03
89	2.70E-03	
8	3.00E-03	6.23E-03
62	3.03E-03	
90	3.03E-03	
91	3.08E-03	8.09E-03
92	3.33E-03	1.77E-03
53	3.51E-03	
93	3.92E-03	
94	3.96E-03	
95	3.98E-03	2.20E-03
96	4.37E-03	
97	4.57E-03	
98	4.98E-03	
99	5.70E-03	
100	5.84E-03	
48	5.87E-03	
101	5.91E-03	9.04E-03

102	6.07E-03	5.58E-03
103	6.46E-03	
57	6.79E-03	
104	6.85E-03	
105	7.14E-03	
27	7.30E-03	
106	7.74E-03	
58	8.14E-03	
107	8.33E-03	
108	8.33E-03	9.97E-03
109	8.58E-03	
110	9.49E-03	
111	9.71E-03	
112	9.73E-03	5.41E-03
113	9.78E-03	
114		7.37E-03
115		9.11E-03

The markers were then tested to identify markers that have an $R^2 > 0.1$ value as shown in Table 8. The validation was performed using the steps described above, with the exception that the phenotype tested for was bunch number (BNO), and the results are set out in Table 9 below.

Table 9: R^2 values for markers that have a p-value of < 0.01 in the 3 tested populations

SEQ ID No.	Population A	Population B	Population C
58	0.005	0.189	0.040
84	0.010	0.175	0.071
100	0.026	0.104	0.011
90	0.109	0.009	0.039
105	0.141	0.044	0.036
53	0.030	0.156	0.014
83	0.018	0.116	0.040
106	0.110	0.029	0.000
99	0.186	0.039	0.008
94	0.019	0.120	0.031
113	0.010	0.044	0.101
92	0.037	0.002	0.123
48	0.016	0.101	0.006
86	0.075	0.145	0.032
82	0.030	0.077	0.124
80	0.102	0.081	0.016
96	0.006	0.132	0.109

The markers identified were tested in the same manner as set out above in Example 1. The markers were shown to be able to predict the bunch traits.

5 **Table 10:** Shortlisted markers with that were selected for screening a population of palms

Marker No.	SEQ ID No.	Observed palms with specific genotype	Genotype	GLM
7	90	35	A:A	-0.963
	90	151	A:G	-1.564
	90	83	G:G	0
8	105	76	C:C	1.623
	105	74	C:T	0.253
	105	119	T:T	0
9	106	154	C:C	-1.329
	106	8	T:T	-1.125
	106	107	C:T	0
10	99	54	A:A	-0.614
	99	194	G:G	-1.891
	99	21	A:G	0
11	80	93	C:T	1.439
	80	176	T:T	0

The five markers, as in Table 10, lists allele combinations that have an effect for the bunch number trait and can be used by themselves or in combination for selection of plants with predicted bunch number that is higher than average or lower than average. The markers are named Marker 7, 8, 9, 10 and 11. The genotyping results of the 225 palms for Markers 7, 8, 9, 10 and 11 are as in Table 11 and validation of the ability to select are as in Table 12.

Table 11: Genotyping results for the six markers for the validation population of 225 palms.

Oil Palm #	Marker 7	Marker 8	Marker 9	Marker 10	Marker 11	Average Yield/year
1	A:A	C:C	C:C	A:A	C:T	19.33
2	A:A	C:T	C:C	A:A	T:T	10.89
3	A:A	T:T	C:C	G:G	T:T	8.22
4	A:A	T:T	C:T	G:G	C:T	14.44
5	A:A	C:C	T:T	G:G	C:T	12.67
6	A:G	C:C	C:T	G:G	T:T	8.44

7	A:G	C:T	C:T	G:G	C:T	13.33
8	A:G	C:C	C:T	A:A	C:T	12
9	A:G	T:T	C:C	A:G	T:T	18.67
10	A:G	T:T	C:C	G:G	C:T	11.33
11	A:G	C:C	C:T	A:G	C:T	15.78
12	A:G	C:T	C:T	G:G	C:T	14.44
13	A:G	C:T	C:C	A:A	T:T	10
14	G:G	C:C	C:T	G:G	C:T	19.78
15	A:G	C:C	C:C	A:G	C:T	19.56
16	A:G	C:C	C:C	A:A	C:T	11.11
17	A:A	T:T	C:T	G:G	T:T	10.89
18	A:G	C:T	C:T	G:G	T:T	11.11
19	A:G	C:T	C:T	A:G	T:T	16.67
20	A:G	C:T	C:T	A:A	T:T	13.56
21	A:G	C:C	C:T	G:G	T:T	16.67
22	A:G	C:T	C:C	A:A	T:T	9.33
23	A:A	C:C	C:T	G:G	T:T	14.89
24	A:G	C:C	C:T	A:A	C:T	12.67
25	A:G	C:C	C:C	G:G	C:T	14.89
26	G:G	C:T	C:C	A:G	C:T	12.44
27	A:G	C:T	C:C	A:G	C:T	11.33
28	A:G	C:T	C:C	G:G	C:T	14.44
29	A:G	C:T	C:T	G:G	C:T	20.44
30	A:G	C:T	C:C	A:A	C:T	14
31	A:G	T:T	C:C	G:G	T:T	12.22
32	A:G	C:T	C:T	G:G	T:T	14.22
33	A:G	T:T	C:C	G:G	C:T	17.78
34	A:A	C:C	C:T	A:A	T:T	16.22
35	A:G	C:T	C:C	A:A	C:T	12
36	A:G	C:T	C:C	A:A	T:T	14
37	A:G	C:T	C:C	A:A	T:T	14
38	A:G	C:C	C:T	A:A	T:T	13.11
39	A:G	C:T	C:C	A:G	T:T	10.89
40	A:G	T:T	C:C	A:A	T:T	9.78
41	A:G	C:C	C:C	G:G	T:T	11.56
42	A:G	C:C	C:C	A:A	T:T	10.44
43	G:G	C:C	C:T	A:A	T:T	9.11
44	G:G	C:C	C:T	A:A	T:T	16.89
45	A:G	T:T	C:C	A:A	T:T	13.11
46	A:G	T:T	C:T	G:G	T:T	8
47	A:G	C:C	C:T	A:A	C:T	12
48	A:G	C:T	C:T	A:A	C:T	16.44
49	A:G	T:T	C:T	A:G	T:T	12
50	A:G	T:T	C:T	G:G	T:T	16.89

51	A:G	C:T	C:T	A:A	C:T	19.11
52	A:G	C:C	C:C	A:A	C:T	16.89
53	A:A	C:T	C:T	G:G	T:T	6
54	A:G	C:T	C:C	G:G	C:T	8.22
55	A:G	C:T	C:T	G:G	C:T	14.44
56	A:G	C:T	C:T	A:A	T:T	13.56
57	A:G	C:T	C:C	A:A	C:T	11.11
58	A:G	T:T	C:T	A:A	T:T	12
59	A:G	C:C	C:T	G:G	T:T	15.78
60	A:G	C:C	C:C	G:G	T:T	13.56
61	A:G	C:C	C:T	A:G	C:T	17.78
62	A:G	C:T	C:C	G:G	T:T	12.22
63	A:G	C:C	C:C	G:G	T:T	8
64	A:G	C:T	C:T	G:G	C:T	13.33
65	A:G	C:C	C:T	A:G	T:T	15.11
66	A:G	C:T	C:T	A:A	T:T	15.78
67	A:G	C:C	C:C	A:A	T:T	12.22
68	A:G	C:C	C:T	G:G	T:T	13.11
69	A:G	C:T	C:T	A:A	C:T	13.33
70	A:G	C:C	C:C	G:G	C:T	17.11
71	A:G	C:T	C:C	A:A	C:T	14.89
72	A:G	C:T	C:T	G:G	T:T	11.11
73	A:G	T:T	C:C	A:A	T:T	13.11
74	A:G	C:T	C:C	A:A	C:T	8.22
75	A:G	C:C	C:C	G:G	T:T	17.11
76	A:G	C:C	C:C	G:G	C:T	18.22
77	G:G	C:C	C:C	G:G	C:T	11.11
78	A:A	C:C	C:C	G:G	T:T	14.22
79	G:G	C:T	C:T	A:A	T:T	18.22
80	A:A	C:C	C:T	G:G	T:T	12.22
81	G:G	C:T	C:C	G:G	T:T	9.78
82	A:G	C:C	C:T	G:G	T:T	14.44
83	A:A	C:T	C:T	G:G	T:T	12.67
84	G:G	C:T	C:T	A:G	T:T	14.44
85	A:G	C:T	C:T	A:A	T:T	10.67
86	A:G	C:T	C:T	G:G	T:T	7.78
87	G:G	C:C	C:C	G:G	C:T	18
88	A:G	C:T	T:T	G:G	T:T	9.33
89	A:G	C:T	T:T	G:G	T:T	10.89
90	G:G	C:C	C:T	A:A	C:T	18.22
91	A:A	C:T	C:C	G:G	C:T	12
92	A:G	T:T	C:T	G:G	T:T	12
93	A:G	C:T	T:T	G:G	C:T	10.67
94	A:G	T:T	C:C	G:G	T:T	14.22

95	A:G	C:T	C:T	G:G	T:T	14
96	A:G	C:T	C:T	A:G	T:T	17.11
97	A:G	C:C	C:T	G:G	T:T	12
98	A:G	C:T	C:C	G:G	T:T	11.56
99	A:G	C:T	C:T	G:G	T:T	11.78
100	A:A	C:T	C:C	G:G	C:T	13.56
101	A:G	C:C	C:T	G:G	C:T	17.11
102	A:A	C:T	C:C	G:G	T:T	11.11
103	A:G	C:C	C:C	A:A	C:T	12
104	A:G	C:T	C:C	G:G	T:T	9.78
105	A:G	C:T	C:T	G:G	T:T	11.56
106	A:G	C:T	C:T	G:G	C:T	14.67
107	G:G	C:C	C:T	A:A	T:T	16.44
108	A:G	C:C	C:T	A:A	T:T	14
109	A:G	C:T	C:T	G:G	T:T	12.44
110	A:G	C:T	T:T	G:G	T:T	15.33
111	A:G	C:T	C:C	G:G	T:T	9.11
112	A:G	T:T	C:C	G:G	T:T	8.89
113	G:G	C:C	C:T	G:G	C:T	16.44
114	A:G	C:C	C:T	G:G	T:T	14.22
115	A:G	C:C	C:T	A:G	C:T	21.78
116	A:G	C:C	C:T	A:G	T:T	18.22
117	G:G	C:C	C:T	A:A	C:T	22.44
118	G:G	C:T	C:T	G:G	C:T	12.22
119	A:G	C:C	C:T	G:G	T:T	10.67
120	A:A	C:C	C:T	G:G	T:T	14.22
121	A:G	T:T	C:T	A:A	C:T	14.44
122	A:G	C:C	C:T	G:G	C:T	13.56
123	G:G	C:T	C:T	G:G	T:T	14.44
124	A:G	C:T	C:C	A:A	T:T	9.78
125	A:A	C:T	C:C	G:G	T:T	14.67
126	A:G	T:T	C:T	G:G	C:T	12.44
127	G:G	C:C	C:T	G:G	T:T	13.11
128	G:G	C:T	C:T	G:G	T:T	10.67
129	G:G	C:C	T:T	G:G	T:T	12.67
130	A:G	C:T	C:C	G:G	T:T	12.67
131	G:G	C:C	C:C	A:G	T:T	11.56
132	A:G	C:C	C:T	G:G	T:T	12.89
133	G:G	C:C	C:C	A:A	T:T	17.78
134	G:G	C:C	C:T	G:G	C:T	18
135	A:G	C:C	C:T	G:G	C:T	12
136	A:G	C:C	C:T	G:G	T:T	12.67
137	A:G	T:T	C:T	G:G	T:T	9.78
138	G:G	C:C	C:T	G:G	T:T	14

139	A:G	T:T	C:C	G:G	T:T	11.11
140	A:G	T:T	C:C	G:G	T:T	10.22
141	G:G	C:C	C:C	G:G	T:T	10
142	G:G	C:C	C:T	G:G	T:T	10.89
143	G:G	C:C	C:T	A:A	T:T	20
144	G:G	C:C	C:T	G:G	T:T	16.44
145	A:G	C:C	C:C	G:G	T:T	7.33
146	G:G	C:C	C:T	G:G	C:T	13.33
147	G:G	C:T	C:T	G:G	T:T	16
148	A:G	C:T	C:T	G:G	T:T	13.11
149	G:G	C:C	C:T	A:A	T:T	14.44
150	A:G	C:C	C:T	A:A	C:T	12
151	G:G	T:T	C:C	A:G	C:T	12.44
152	A:G	T:T	C:C	G:G	T:T	12.22
153	A:A	T:T	C:C	G:G	T:T	12.44
154	A:A	T:T	C:C	G:G	C:T	13.33
155	G:G	T:T	C:C	G:G	C:T	20
156	G:G	T:T	C:C	G:G	T:T	15.78
157	G:G	T:T	C:C	G:G	C:T	14.67
158	G:G	T:T	C:C	G:G	C:T	16.89
159	A:A	T:T	C:C	G:G	T:T	12.44
160	A:G	T:T	C:C	G:G	C:T	12
161	A:G	T:T	C:C	G:G	T:T	12.44
162	A:A	T:T	C:C	G:G	T:T	20
163	A:G	T:T	C:C	G:G	T:T	18
164	G:G	T:T	C:C	G:G	T:T	15.33
165	G:G	T:T	C:C	G:G	T:T	11.56
166	A:G	T:T	C:C	G:G	T:T	13.78
167	G:G	T:T	C:T	A:A	C:T	16.89
168	A:G	T:T	C:C	G:G	T:T	12.89
169	G:G	T:T	C:T	G:G	C:T	10.44
170	G:G	C:T	C:T	G:G	C:T	19.78
171	A:G	T:T	C:C	G:G	T:T	14.44
172	A:G	T:T	C:C	G:G	T:T	10.44
173	G:G	T:T	C:C	G:G	T:T	17.33
174	G:G	T:T	C:C	G:G	C:T	15.11
175	G:G	T:T	C:C	G:G	C:T	11.78
176	G:G	T:T	C:C	G:G	C:T	14.67
177	G:G	T:T	C:C	G:G	C:T	20.44
178	A:G	T:T	C:C	G:G	T:T	11.78
179	G:G	T:T	C:T	G:G	C:T	15.78
180	A:G	T:T	C:C	G:G	T:T	15.33
181	G:G	T:T	C:C	G:G	T:T	11.33
182	A:G	T:T	C:C	G:G	T:T	8.67

183	A:G	C:T	C:C	G:G	T:T	16.22
184	G:G	T:T	C:C	G:G	T:T	16
185	G:G	T:T	C:C	G:G	C:T	10
186	G:G	T:T	C:C	G:G	T:T	14.44
187	G:G	T:T	C:C	G:G	T:T	12.22
188	G:G	T:T	C:C	G:G	T:T	14.22
189	A:A	T:T	C:C	G:G	C:T	15.56
190	A:G	T:T	C:C	G:G	C:T	21.56
191	G:G	T:T	C:C	G:G	T:T	13.33
192	A:A	T:T	C:C	G:G	T:T	15.33
193	G:G	T:T	C:C	G:G	T:T	14
194	A:G	T:T	C:C	G:G	T:T	11.11
195	G:G	T:T	C:C	G:G	T:T	15.78
196	A:G	T:T	C:C	G:G	T:T	14.22
197	G:G	T:T	C:C	G:G	T:T	11.33
198	G:G	T:T	C:C	G:G	T:T	24.89
199	A:A	T:T	C:C	G:G	T:T	14.89
200	A:G	T:T	C:C	G:G	T:T	14.67
201	G:G	T:T	C:C	G:G	T:T	14.44
202	A:A	T:T	C:C	G:G	T:T	12.22
203	G:G	T:T	C:C	G:G	T:T	16.67
204	A:G	T:T	C:C	G:G	T:T	14.89
205	G:G	T:T	C:C	G:G	T:T	16.67
206	A:G	T:T	C:C	G:G	T:T	13.11
207	G:G	T:T	C:C	G:G	T:T	12.89
208	G:G	T:T	C:T	G:G	T:T	11.78
209	A:A	T:T	C:C	G:G	T:T	13.78
210	A:G	T:T	C:C	G:G	T:T	20.89
211	G:G	T:T	C:C	G:G	T:T	15.78
212	G:G	T:T	C:C	G:G	C:T	14.44
213	G:G	T:T	C:C	G:G	T:T	17.56
214	A:A	T:T	C:C	G:G	T:T	14.89
215	G:G	C:T	C:C	G:G	T:T	16
216	A:A	T:T	C:C	G:G	T:T	17.56
217	A:G	T:T	C:C	G:G	T:T	11.78
218	G:G	T:T	C:C	G:G	C:T	17.78
219	G:G	T:T	C:C	G:G	T:T	12.89
220	A:G	T:T	C:C	G:G	T:T	11.56
221	A:G	T:T	C:T	G:G	T:T	11.33
222	G:G	T:T	C:C	G:G	T:T	10.89
223	A:A	T:T	C:T	G:G	T:T	16.44
224	G:G	T:T	C:T	G:G	C:T	13.78
225	A:G	T:T	C:C	G:G	C:T	10.67

Table 12: Validation of using a plurality of markers for selection

Markers	Allele combination selected and allele effect	Average bunch yield (bunch/palm/ year) of selected subpopulation	Change over mean number of bunch yield (bunch/palm/ year) in unselected population
8 and 11	C:C (+1.623) and C:T (+0.253) for Marker 8 & C:T (+1.439) for Marker 11	14.8	+ 1.02
7,9 and 10	A:A (-0.963) and A:G (-1.564) for Marker 7 & C:C (-1.329) and T:T (-1.125) for Marker 9 & A:A (-0.614) and G:G (-1.8913) for Marker 10	12.97	-0.81
11	C:T (1.439)	14.78	+1.00

Table 12 shows that the markers identified to be linked to bunch number can be used to select plants containing the best allelic combination statistically linked to either higher bunch number or lower bunch number, as compared to the mean of the unselected population. This example serves to demonstrate that the method disclosed here can be used to associate different phenotypes to genotypes and therefore detect beneficial or desired allelic combinations.

	GCTCTTATGGTGTAGGGGTGCCATGTGATTTGATCC
16	AGCTCGATTGGACCTAAATCAATTAGATTGAAATCTAATCAAAATTTGATCAAACTTATATT[T/C]GGACCTACTGCCAGTTAGATAT GGAACCCAAATGAGACACACACATATAGAAATIGGCCAA
17	GCTCGATTGGACCTAAATCGATTAGATTGAAATCTAATCAAAATTTGATCAAACTTATATCT[A/G]GACCTATTGCCAACTAGATGTG GAAACCCAAATGGTCACACACATATAGAAATTTGGCCAAAG
18	TAAAGGGGCTATTGTATTTTGAATAAGATTCAAAACCTTTTGGCCTATTTAAAGGGGATCTTCT[T/C]TAAAGGCTCTAAGCATCACATTTT CAACAGCCCCACGCCCTCCCTCTCTCTCTCCCTTGGT
19	TAGCTTATATAGGAAGAGAGACTATTTTTCATGTGAGGTGTCTTTGAAGAGCAAAACTGT[A/G]AGGCTCTAAATGACAAAAGGCC ACTGAAGTCCAAAGATGGTCATGTGCTTTTAAAGAGC
20	TGATCCATCAAGTTGTGTTTATGGAAAGGTAAACGTAACTTTGGCGATGGCTTTAAACCA[A/T]TGAATATATATATTTTTCATT TCTCTTCTCTCTTCTCCCTCTCTGTTTCTTCTTC
21	CTAAAACTGAGCCATCAACCTCTTCTCTGATACCGATTGAAAGAAAAAAATATCTTTT[T/C]CTTTGAGGATCTTTTCAGATTTTCA TCAAATCTGAGGTGAAATAATTTTAAAAAAATTTAT
22	TATATAGTTGTAGCATTTGTTGTAAGTAATTTAGTAATTACCATAGTAGCTAATCACATTCT[T/C]ATCTTTGTTCAGGAGCTTTTAAAGT TTGGCGGGGAAGTTTGTCTCACTTGGTGTCCCAATGG
23	CTTGAGCTACCCAAAGATTTGGTTTCTTACCATTGTCATAAACTCATATAGTGTGGTAAAAAA[A/T]GATTTAAAAAAATCTGATTCA AAAAATGAATCACAAATAAATAATATATATCCCTAGAGA
24	GATGCCGATATTAAATATTCTCAGTCATTATTAAAAATTCACCTGAATAGATATTAGTAATGA[T/C]AGTATCAATGCACCTCGACCATC AATAAAGTCATTATTAAAAATGCACCAAAAAATTTGGCT
25	CCAAACTTCTCTATTTTAAAAAACAGCGTCATGCTGGATGATGTAATCATGTCCCTTCTT[T/C]GTTTTTCGACGAACATAGGTTTCC ATGCAGCCGGAAAAATTATTCAGAAAAATTTGATCATTTAT
26	GGGAGAGAGAACAGTCTCCCCCTTTTAAAAATAGGTCAGAGGGGAGGAGTTTGACTCCTCCT[T/C]AGGGTGTTTTTTGACTCCGATT AGGAGTCGTCGGGAAGAAGAAAGAC'TCTCAGGAGT
27	AAATTAAGGCCATACAGGAGATGACTCCCTCAAGAACGATCAAAAGAAAGTGC AACATCATAT[C/C]GGGAAGACTTGTGTGCTTTGAA CCGATTTGTCTTGAGATCGACAGAATGGTGTCTTACCCCTT
28	CTCAACCATTTTCATGATGTTTCAGTCAATTTTAAAGACAAACCCCATGCAGTGATAGTAAAA[A/T/G]TAAATGTGTGTATAGAATAT AAAAAGGACTACTCATGTTGATATGCAACAAGTTAGACGA
29	GAATGATGGATATCGAGCAGATGGTCTGCCGATGGCTGGAGGGTTGGCAGGCCCGTGCAT[T/C]TGTCTATGATGGATCGGATCA CCTTGATGAGGTCTGTCTCAGCTTTATTTTCAGTGTAAC
30	TCTTCAATTTATGACCACCTCTCCTTTTCTACCATAGTATTAATGCTTTTAAAAAATAGAAA[C/G]TGTAGCAGGCTTTTCTTGAAAGGA AATAAACTCATAGTTTATAGTGAATTATCAATGAA
31	CAATGAATGGACAGGTGCTGTTACATATCTAAACACTGAAAAATTTGATCCATTATACTTC[T/C]TTATCAGCTGAAGTTTAAAAATT

	ATATGCAATACCTCTAGATTGTAGAGTGAAAAAAT
32	CATTCTTATTTAAAAAGAAAAAACCTTTATGAGACTCTTTGATACAACACTGACCTAT[T/G]GTAAAGCTATGAAGGCATTGCAG TGTTTTAGGTGAGTCTAGATAGCTTTATTTCTTGTTT
33	ATAGTAGTGAGTAAAAATTAAATTTTGGTACTAGATATGTATCAAAATTTTGGTACCATT[A/G]TCGGAGATCTTAGCACGATTGT TTTAAAAAAATCAAAAAAAATTTTATAAAAAATAATAAA
34	CCTTGGTCAGAAGGCTATCAGTTTCAACTCCTAGTCGACATACAACACTCTCTGATTTT[T/C]CTCAGAAAGCATATCGAAAAAGCT CAAGTTCAAAAGTCAGGCACCAACTGGCCACAGCTCTC
35	ACTGCATGGTTAACAAAGGAACTTTTCTATTGCTCAACTTTGAACTAATCTTATGTAGGTT[A/G]CTAGCCTACCGAGAGAGACACT ATCATGGTCTTTTTCATGAACCTTTTAAAGACATGTT
36	TATTGTTTTTAAAAATACAGTTACTAAATTTGTCACTAGTCTCTCCAATATTTTGTGTTGA[A/G]AATACATATCCATTGATAATCGAT CGCTATTCACTCGCTAAAAAAAACCTATTTCGGGACA
37	AAAAACAGACGGATGAACAATTTATTAGAAAGGCTATAATTCGAGGAATTCATGGTGGTTAT[A/C]TAAAAGGCCCTTTGTTTGCAGTTT TCACATTCTATTTTCTGGCCATTTTCTCTCTGTTATA
38	TTTAAACATCACTATAAGCATAAACATCATATCTTATCTTAACCAGTCACCGAGCTACTC[A/G]TGTGAATATTTGGATTGGACCA ATTGAGCTTCTTCTCATATTTTGAATACTTCATAGAAA
39	CATGTCATATAACAAATACCCACGACTAACAACTTAGTGAGTGGTGGATTTCCTATCCT[A/T]GCTACCTATAGATTTCAGACATA ACAAACTCTGGTACTTCTCTATATAAAGGTAACAAGAG
40	TTTTTTTTTTGGGTAAAACGTGTACTAGTCTTGTGTTTGTCTTGCAGAAAGGCATC[A/T]ATTTGTGGATCTTTTAAATAAAA GTTTGATGAAAGAGTAGACCAATTATAATGGGTCAAAAT
41	CTCAGCTTTTAAAAAATCAACCTTCTATCTTTTGTGCTGCCCTTTGAGACATAGTCGGCTAA[T/C]TTTCTGACTTAGTCGACTAAGTTT GCGTGCCAAAGCTACGAAGGGTCTTTGAAATGTTT
42	ATCCGTTTAAACATTGGCCAAAAACATCACATTCACATCCTGGAGGAAAAAGACGATAGCT[A/G]ATTTGATGAAGGCTCTATCTA ACATGTATAAAAAACCTTCGACTTCCAAACAAGATATACC
43	GATGTATGCTGTGTGTGTGTAGATTACATCTTGGCCATGCAGTGAGTCAGATCAGCAG[A/G]TTTATGGTATGGCCTAGCAAGA AGCATTGACGAGCACTGAAGGAGATCTTCAGATATATG
44	TTATGCTTGAAGGATATAGAAGATGTTGAAGAAGATTATTTTGAAGACITTCCTGAAAAAT[T/C]TGTGCATAAACTTTCCTCTCTCTT TATTTTGGGATATTCATGTTTTCACAAAAATAAA
45	AACGAGAAAGTTCGGGCCCAAGCTCGAACACCGAGAGAAAGCACAAAATTTGAAGGGTCTC[A/G]AAAGGACCTTCGGTCAATGAA CAAAAAGGATGAATCAATGTGGCGATGACCAAGAAACCTTC
46	TTGGAGCCCTTTGTCATCTACTTTTTCATGTGTACATAGAGCCCTAGGGGATGCCCATGCT[A/G]GTCCCACATGTCCCTCTTTAAAAAT GGGCATGTCCAACTTGATCAAAATCAAGTGACGCCCT
47	TTTAAAAAAAGCTGTGCATTACAGATGACTGCATCTAAAACCATCATTTTGGTAAACAAAA[A/C]AGTCACATTATTTCCGAACCAT

	GAACAGCAATGGTGCAGGATAACTTTCGTTAGATTGC
48	AATATTTATTATTTAAAAATTAATTTTGCAGCATATTTACAGAAAGCTCTATGCCTCGTGT[T/G]ATTACCGAGATCATCTAACGATT GATGTCGGGCCAGTGAAACGAGTTCTCTACTTGACCA
49	TTATAGATAAATCTCATGATCTAATAAGTATAGAGATTCTAAACCCAAAATCTAAATTTAAAAAT[T/C]AGTCCAATACTTATAAAAAAAA AATTTCTACATGAGATAGAAATTGCCATCACATAATGACAA
50	ATATTTTCATCAAAATTTTAAATCTTTCGGATAGTCATGACTTATTGTCTAAACATCAATCT[T/G]GACTTATGGATTTCATCAAAATTA AAAAAATTTAATTCAGAAATTAATTAGGAAGAATCTT
51	AGTAAATGTAAAAATGCAGTTTGGAAGTTACATAATTTTAAACACAAAAAGAAAACT[T/C]GATATGGTCTTCTTCTGGTTTGT GACATAAACACGAAATATGTACCCAAAGTTTATTAGT
52	TTTTCTGACTTTCCCAATGCTGTAGTTTGGCTGGTCTGCTCAGCTTCTCTTGTGCTGCC[A/G]TCGGCTGAATTTTTCAGGTCTT CTTGGTTTCACACTTCCGCTTGGCTCCCCGGTGC
53	AAGGGTCTAGCCACIGTAGCGGACCCCTCATCACCCCTACTCCACAAAACAGTCCGATTGGAT[T/C]GAGAAATCTAACTGCATGCCTAA TGATCAAAAACCTCCCTCCTCATCAACAAGGATGGATC
54	ATATGICTAACGATCCTCCTGGATCAAAAGGATGATAAATACTCATACACTAGTGTTTTTC[T/C]AGTGACCCCTCATGTCTCCATCCA ATAATTTTAATTATCATGTAAATTTAAATATTTTTT
55	TCAGAAAAATAGCAAAAATTTAAAGCTTAATAGTAAACATTTTAAATGGAAATCAAAACCCCTC[A/G]ACCAAAACCTGAAGGCATTTAA ATATAGGGGACACCAATATGGTAAGAAATTTAAGCCCATG
56	TCCTTCTCCTCAACTGACTGTCTGCTGCACTATCAGGAATATCTTTGTCTTGTAGAGGCTT[T/C]TTTCATTGAAGATCATGTCCTTGC TTCTGATGAATTTCTATCCTGCTCATTCAGAACT
57	CATGGCTGGTTTGGAGGTCGAGGACGCTGCGGCCAAAGAAAGCGAGAGAGAAACAAAGGA[A/G]CTTCTCCGAGATGGGGCGGG GGGATAGCGGATCGATCGAGATTTTCTGGAATGGAATTG
58	TAATCTTAATTAGTTTACGATTGTCAATATATCCTAATTTTAAACCTTATTTTAAAGAAC[A/G]AAAGGGATTCTAATCCCTCCCAC TACTCTATACCCACTCTCCTAACATAGATTGAAGTTT
59	AAGGGCTCACTTGGTTGAGGCTCTACATTGGCCGACCTCGATGGCTTTCCTCAA[A/T]GAACCTTGGGCAGTCGACGCTA CCCTCTCTTAGGGATATGTTGCGCCAAAGATCCAATAC
60	AAAGCTATGATGATTGTTGAGCAATTGTCTATTAATGGAAATAATATCATAGTCACCGAATA[A/T]GAGAGTGCAACTAATTGTCTGGC GTGATATATTGAAGCTGGCCTAACAAACCATGAATGTGA
61	TTTTCATTAATTCATCATGAAAAAGATCGTGACGACCCCTACATACCATAGGAGATC[T/C]CATCGAATTGGTAAATAAAGATT ATGAAATCTTGATTTCTTAGGATCGGATACTACGA
62	CCCTATGTTGGGTAAATGTAATCTGGGGTGGCATAAGTGTAGGCTTGAATCTGACTTAAGCCT[A/G]TGATAGGAGGGCCTATTAGGAT TAGTTTAAATGGCTAGATCCCTCTCCAAATTTCTATA
63	ACCAACCAAAATCTTACAAAGTCCAAATCTAGGTTGGATGGCCTAATCCAAATTTCTTT[T/C]ATCCCCCAAGAACTTGAAAAAGA

	AAACAAGAAATATAAGATGCAGAAATTCAGAATAAATTA
64	TTGTTCCAAATCATGTATGGATTTATCTCAATGATTTTTTTGACAGTTTGAGAAAGCTGCAA[A/G]AAAAATAGGTATCCAATGGTAA GACTAATGTTTTAAACCAGATCAATTGACAAAACTGGAA
65	AAGAGTACGCCAGCTTACTGAACCATGGCCGATCAGCACAAAGATCCAATTCGGATACGGC[A/G]AACCTTACTTCTAGTTTCATCA AAGTAATAAGTAACAAAGCAAGGAGGGTAGAGCACTGATG
66	TATCCCTTTAAAGTACTATGATATCTCTCACTAACTAATAGACCTTAAAGGCTATTAT[A/G]TCATCAAAAACTACATAGCATAT GACTTGGTATGCTTCAAAATAGGTGTGATGGTTTT
67	TTAAACTTATCAAAAGACTTTGGATTATGTCTTCAATTAATCCATATATCCAAAACCTTGAT[A/G]AATCATTATAAAAAGTGATGAA ATAGGAGCATCCCTCTCTGATTGAGTTGTCAATTGACC
68	TGTCATTAAAGTAACACTAACTATGGCTTACCAACCCACTATCTTATCTAATTAAAAGTC[T/C]TTTTTCATCACAAATTACCAATGTG CCGTAAACATAAAACGCTGCAAAATGATCTTTTAATGC
69	ATAACGACTGGATCATCATTTGAGTCACAAAAGGGTGCTCTTCGATTCAATTGATTAAATA[A/T]TTTTGTAAAGCCACCATATAGT CTTTGTATTTTAAAAAGTTTCCCTCCCTATTTTCTCAA
70	TCAAGGTTTCAAGCTATCAAGCTAGCTTGACAAAGTTGGTGGAAAGATGTTGATGGATAGAG[A/G]GACTTAAATCTCGGGCAAAAATA ACAAC TAGGATTTGGGCGGCTCCACTTTAAAAATACTTT
71	AGTATCATATTAGATAAGTGTCAATTAATGTCTCTTGTGATGATGAATGTCATTTCTACA[A/G]GTTTTTTTGGCCAAAATTAAATTTTA AATTTTTAAATAAAAGATAAATTTTGTGATATCATTA
72	TCTTCAAAATATCTCTCAATATTCAAAAAAATTTAAAGATAAAAAACGTCATGATTATGAT[A/T]TTTAAATCCCTAAAAATTAGTTTC AAGGAGTGAACATTGAACACACTACCGAGCACATCGCC
73	ATACATGGCCATTAGGACACACTAAAGTAGTAAACACTAAATAAATAAATTTAAAAAATG[A/G]TCAAAAAGTCCTGAAGGGTAAC AATAAAAGGTTTATGAACCTCAAAAGGCTCACATAGTGA
74	TACATTTCCAGTCAGTGTAGGCTTCATAACAATGATTCAGATTGTGTGAGGCACGGCATC[A/C]TGGTTTTCCCATGAAACAAGAC ACGATGCCATCTCGGCCCATCTCTGACATAATGCCTTTG
75	CCAAAAGCTCTTGCTTCCATTTTATTGCACTGCAACTGATTAAACCATATGAGTTCTATT[A/G]GACTCTTCTTCTCTCTTTTCTCT TTTTTATTTAAAAAATGTAGTCCACATGAGGAG
76	ACCCCTTTTAAAGGCCAGTTGTTGTATAGCAATTTTGGATGGTTTAAAAAGAAAAATGATT[T/C]TTAGTGAAAGAAAGGCTGTATAC CCTCTTTCTCTTTCTTATGCTCCCTCTCTTTCTCTTT
77	GCAAAAAGAGTAGTGGTACAGACTGTGAATTAACCATATATCAAGTTATTTAAAGATGTCA[T/C]GTCTTCTTCTATGAACATAAAGC GAAGATGTTCCCTTGTGTACAAAGTATTATTTTTCAT
78	AGTAAAAAACCTCTGCTCCAGTTTGTGCTGACGATACTATTGTTTTTTTGAGAAAGCGAAAAAG[A/G]AAGTTGCTGGCCATCTCAAGTT CATATTATACATCATGAATGGTATCTGATCTCAAAA
79	TATTTCTACTATCTCATAAAGCTGATGCTATGGCACATCCTATTGGTTGATCACTCCTAT[A/G]GATAGCATGCTGAAAAAATATGCT

	TGATAAAATTCTCAAAAAGCTGACAAAGTTGGTTTAAAAAT
--	---

Table 14: Sequence ID No. and SNP for BNO

SEQ ID No.	SEQ ID No.
80	GTCCATGAAGCTTAGAAGTCAGTATCCCAAAGTTGCATCAACTAGTTGGTCGATCTTTGA[T/C]AAAGAAAAAACTGTCTTCGGATG AGCTTCATTAAATTGTATAGTCGATGTAATTCTC
81	CTAAGTGTGGGGCGTGGGGATGCCCAAAGAGAGGGAGAAAGAGAGAGGGGC[A/G]TGGGAAGAGCTTTTGGGGCAT AAGGGCTGCTAGGTTTATGCCCTAAGGATCTTAAAGGA
56	TCTTTCTCCTCAACTGACTGTCTCTGCACTATCAGGAATATCTTTGTCTCTTGTAGAGGCTT[T/C]TTTCATTGAAGATCATGGCCCTTGCT TCTGATGAATTTCCTATCCTGCTCATTCACAGAACT
82	CTGCTGAACTGTATGGTTTAAATTCTCCAAATCATTTGAATATGGGGCTCACAACTAAG[T/G]GTTCATGTATGTTATATCCTTTTCGA AAAGTCTTTTAGAGGTTTCATGTAGGGCCAAATTIG
83	TATCCCAATAAATGCGGTCCAGGTAAGTTTATCAGAGGCAAGAGTCAACCAAGAAAAAG[A/G]TAGAGAGAAACAATAAGCCCAAG GAAATTGTGAGCACTTTTTCATTAGAGAGAGGAT
84	TATCAACATGTCAGAGTCAACTGAAATCTCTTGAAGTTGACTTTAAGGCTTTTTCACCTT[A/G]ATTTTATGAAAAATCTTTATCTA ATCATCCTTTTACTTTCAAAAATCTTTAAATAC
85	CTGTGGCCAGTATGGCCAGTCCATTAAATTGAAAAGCTGTGGAATCCATACACCTCATAA[T/G]TTAAATTTTGGAGAAATAGAAAT CACACTTTTAGATGAATTCATGTGTCTTCAAGAAAGTC
86	ACCTCCTCTCCTGGACCTCCGAGGAGGTACTTTGATGTTAGTCTTGTTCGCCGTCACAGCT[A/G]TATCATTTGCTGGGCCCTCGTCTCTC CGTTTACCATGGAGGAGAAATTGGCGTCCGACCCGG
87	TAGATTGATTAATTTATTGGCCATCCGAATCCGAATCTAACTCATTTAAATAGAAATG[A/G]ACTATATTTTTCGATCCCAATGGAT GGGTTCGAGCCTAAATTCAAATCCCATTTGGTGAAG
88	AGATGAAATTTACTAGTAGTCAAAGTCGGGACGAAAGTCTGACTCCCGTAGGAGTCCAGA[T/C]GGAGTTTCCCCCATGGCCAGAGT CAGAGATGATGTCCGGCTCCCGTAGGAGTCCGACGGA
89	TGGGTTATTTGGTTGGCCATGACTTGTTTTAAACAGGTTAAACAGTTGGGAAAGGATCA[C/G]CTCACCTGGGCAATAAACACAGGTT AGTTAGGGTTTGACTGTCTAACCTACTTAGGGTCCGT
8	ATCCAGGTTATGACTTTGATGATGATGAACACCTCTCTGATGAGATTAAA[A/G]TACTTGCTCAAGGGGCCAAACCAT ATTGCAAGAAGGTTTAAAGCTTTTGCTATGGATAATG
62	CCTTATGTTGGGTAAATGTACTTGGGGTGGCATAAGTGTAGGCTTGAATCTGACTTAAGCCT[A/G]TGATAGGAGGGCCTATTAGGATT

	AGTTTAAATGGCTAGATCCCTCTCCAAATTTCTATA
90	ATGAAAGGAAACATTCCTTATGCAGTTCACATAAAGATGAGTAAATAGACCTGAGGCCGTG[A/G]GCAACAGTGATGCTCACATTAA GAAGTCAAAATTGTGCTCATAGTTACCAAGTACTAT
91	CATATCTGTACTTTTTTGCCTGAATAGCGAAGACATGAAATTTGATGTTACTATTCCATAT[A/T]TGTTTAAAAATGAAGATGAATATGA GCTGGATATAAATACTTCTTTTTTAATTGAATGTGA
92	ACGGCTCCACCCCTCTCCAGATTGCCTGTCTCTGGGCTGGGTTGGGACTTGGATACA[A/G]AAACATGTCTTATCAATGCTATA GCAGTCTGTGAACCTTAAAGGTTTCATGCGGGACGGTAC
53	AAGGGTCTAGCCACTGTAGCGGACCCTCATCACCTACTCCACAAAACAGTCCGATTGGAT[T/C]GAGAATCTAACTGCATGCCTAAT GATCAAAACCTCCCTCCCTCATCAACAAGGAGTGGATC
93	CAACTCGTGGCTAGGCATGGCCTGACATGGTCCAGGCCCTAGTCCAAAGCCCTGCTTGATC[T/C]TTTTTCTGAATTTAAAAAATATA TTTTTTTATAAATTATAAAGTGTAAAGGAGTAAAA
94	TACATAATAGTGAGTAGGTTTAAATTTTGGTATGCTACGACACATATAAAATTTTGACATC[A/G]TTATCGGAGATCTTGGTATGATTG TTTTAAAAAATCAAAAAATATATATTTACTACTT
95	AAAAGCGGGATGAATTGAATTTCTTTAAAAATTTTACTAAATCCAAATCTGAACAATTAT[A/G]TGCTTCAATTTAAGCTAAAATTGA GTGTTGTGAAATTTTGATATGTGTAGCAGCAAAAT
96	GTAAATGCAAAATTTAATAAATAAATTTAAATCCATGGTTCAAAAAAGTCAAAAGATA[A/T]TTTTTGTATTTTAAAGACCGATGC TAGAAATCGATGCAACAATTTGGCAAAATATGAATTA
97	CATCTCCATGCTATCTTTCTAATTCAAAATTGAACCTATGATTCCTAGCAAAAACAGTAAAT[A/C]GCATTTTGCTTACACCATTGATGAA TTATGTTTGTTTTAAACTCAAAAGAGCATTCCTCGT
98	GTGGTGCAGCTCCAAAGGGCTAATAATGAGAAATGCTGGATCTCCTTGATGTTTGTCCGTA[T/C]GAAAGTTATGAACCGCAACAGG GGCCTTAAAGTCAITTAACATGCTGGGTAGGGTAAAGG
99	AGTGTCTGCACATCAGGCACGCCCTTAAATGCCCTTTAACAAACATGATTCCAGCAACA[A/G]AGAGAAAGCAGAAAAATCACGAAA ACCATGTAATAATTACTCCAATTCTTGATCAACCTCCCAT
100	CCCTCTCTTGGACTATCTCAAGAAAAATTCATTTCTCTCACCACAAAGAAATGTTTCTCTTTT[A/C]TTCCCTCTTGGATCACTTCCCCTAAT TGATTTTATATCTTAACTTAGTTGAAAGTATACTGG
48	AATATTTATTATTAAAAATTAATTTTGCAGCATATTTACAGAAAGCTCTATGCCTCGTGT[T/G]ATTACCGAGATCATCTAACGATTG ATGTCGGGCCAGTGAACGAGTTCTCTACTTGACCA
101	TTAATCGGCCACATACACACGTACCATGAGGATTGAAAACTGTAAATGGCTGAGA[T/C]AACTTTTATATGAAAAATGAAAGCC GCTGTAGTTATGATCCTAATCCTACTGATGATGAAGA
102	TTTAAATCGCGATTTTCAGGACACATAGTACTGTTGTGTGGCCATCTTTTTTGACTTTTGT[T/C]GATGCAATTATACAGGGTAAATAATT AAAAATAGGGTAATAATTAAACAGAGTCTACAAAAT
103	ATCATTAATTGTCACTCCTAACCAAAACACTTTCTTGATCAGTTCTCTTTCTTGGTTTCTGA[T/C]ATCCAAAATCTTCCCATGACACAATTA

	AAGAAATCAAGAAATGAACAGATAATCTCATGTTT
57	CATGGCTGGGTTTGGAGTCGAGGACGCCCTGCGGCCAAAGAAAGCGAGAGAGAAACAAAGGA[A/G]CTTCTCCGAGATGGGGCGGGG GGATAGCGGATCGATCGAGATTTTCTGGAATGGAATTG
104	ATGGGTCATTTCTTCTAAGATCCCTTTTCTTCTCAACTACTCCATTTTGTGTGGTATTTT[T/G]ATGCTAAAAAGTTGGGTTGAATCCC ATTTTCATTTATTGAAATTTTCAAAAATTTTATTTT
105	TAACAGTTAATGAGTCCCTCAATCCACCTGATTGGATTCTGTTTTCCCAAGCTACCGGG[T/C]CTTAAATTTGGGAIGCCTTCTATTA AAGTTGTCAATAATCTGAGATTAAATTAGTATCT
27	AAATTAAGGCCATACAGGAGATGACTCCCTCAAGAACGATCAAGAAGTGTCAACATCATA[T/C]GGGAAGACTTGTGTGCTTTGAAC CGATTGTCTTGAGATCGACAGAAATGGTGCTTACCCTT
106	AAATGTGTCTGACTTATCCAAAACATCTCTTAGCTGCATAATGAGCTCCAGATGGCCAG[T/C]TGA AAAA AACTCATTCACCGTTCCC GACATCAAGTATTGGATGATGTCAGTAACACATGAG
58	TAATCTTAATTAGTTTACGATTGTCATATATCCTAATTTTAAAACTTATTTTAAAGAAC[A/G]AAAAGGATTCTAATCCCTCCCCT ACTCTATACCCACTCTCCTAACATAGATTGAAGTTT
107	TTTCATCGATATATATATATGGTTGGATGATTCATGCTTGTGATTTTCTGATATATAC[T/C]GCAAGTTTCTTTGGTATATACTCAG TAGCAATTCAATACACACCTCGGATAAATCGATAA
108	GAGGCCTAAACGATTTCTGTTTCCATTTGATGATGACTTACATCATGATGAAAAGGAGATG[A/T]ATAATGGTTTAAAAATATATATT GAGTGTGTCAATAATCAGGTACCTTCTAGAGGT
109	CCACTCCTTAGTTAAGCTTGACCCGCCGTTACAGGCCAGGTGTGAAGCCGTCAATTGAGTC[T/C]CTCAGACACACACATTCGGGCCAA CCTCATCTTTACCCCTGGAGCAACTCTTTGCTAGTAA
110	CTGATCTGATCAGTCTTCTCGGGAGTGTAGTTTGCATGAAGACCTTCTTCTTAACTGAT[T/C]TGGATCTTTTCTCATAAATTTTAA ATTTTCTTAGAGATTGAAGATAGACTTCTAGAGGA
111	GAAAGAAAAGAAAGAACCCAGATGAAAATCTGAAGGGTACTAAGCATCTGAAGAGGTCAAT[A/G]AATAGTGTCTCGCTTGCGA TGAACATTTCCCACTCTGACTCATTTTAAATACTTATCT
112	CAATTTCATGCGCCAGCTTGTAAATCCCCCAAAGTCGAAGCCCTCAAGCTTCGGGCCAGTTGA[T/C]GCAATCCCCAAAAGCACCAGCTGT CTCGACTGATGCCCTCTCTAGGGAAAGGTCGACAACT
113	CTATGATGAGCTTCAATTGATTTATCTTTATTTCTTTGGGACATGTCTCACTTATCGTAA[T/G]GCATATTTTAAATAGGATTCCCTTCT AAATCCATTCTTACTATACCATATGAGATATGGCA
114	CTTTGTCACTGAGGCCAGGATCTCTTGCTCATACTTCAGCAGTTCTAAAAACATCCACAG[T/C]GGAGACACACCACCTAGTGTAGTG CTGCAATAGCTGGTGACGGCTGCAGCTGCACAGTT
115	CGTTCTATTTTAAATAAAATAAAATAAATAATTTAGGGAGGCAAGACTACGCTGAGTCG[T/C]GGTGTGCGTTTGGTAGATTACGT GCGGTATACCGCAGGCAAAACTGGAACCAAGCGGAGCT

CLAIMS

1. A method of predicting or determining a plant phenotype of interest, wherein the method comprises detecting the presence or absence of one or more polymorphic genetic markers
5 selected from the group consisting of SEQ ID NO: 1 to SEQ ID NO: 115.
2. The method of claim 1, wherein the plant is from the Arecaceae family
3. The method of claim 1, wherein the plant is an oil palm.
4. The method of claim 2, wherein the oil palm is from the genus of *Elaeis*.
10
5. The method of claim 3, wherein the genus of the oil palm may consist of the form of *tenera*.
6. The method of claim 3, wherein the genus of the oil palm may consist of the variations
15 selected from a group consisting of *pisifera* and *dura*.
7. The method according to any one of the preceding claims, wherein the oil palm is selected from a subpopulation, wherein the subpopulation consists of Deli, Nigerian and AVROS breeding backgrounds.
20
8. The method according to any one of the preceding claims, wherein the oil palm has genetic background from least one, at least two or three subpopulations selected from a group consisting of Deli, Nigerian and AVROS breeding backgrounds.
- 25 9. The method of claim 1, wherein presence or absence of one or more polymorphic genetic markers is determined in a plant sample containing genetic material, wherein the sample material is selected from a group consisting of leaves, spear leaves, stem, saplings, roots, buds, flowers, seed, fruit and pieces thereof.
10. The method of claim 1, wherein the plant phenotype of interest is a yield component
30 selected from the group consisting of fresh fruit bunch weight, bunch size, fruit size, oil extraction rate, bunch numbers and combinations thereof.
11. The method of claim 10, wherein the plant phenotype of interest is associated with fresh fruit bunch weight yield or bunch number yield.

12. The method of any one of the preceding claims, wherein polymorphic genetic markers are SNPs (single nucleotide polymorphisms), SSRs (simple sequence repeats), AFLPs (amplified fragment length polymorphisms), RAPDs (random amplification of polymorphic DNAs) and combinations thereof.
- 5
13. The method according to claim 9, wherein the one or more polymorphic genetic markers are single nucleotide polymorphisms (SNPs).
- 10
14. The method according to any one of the preceding claims, wherein the SNPs associated with an increase in fresh fruit bunch yield are one, or two, or more genetic markers selected from the group consisting of SEQ ID NO: 1 to SEQ ID NO 79.
- 15
15. The method of claim 14, wherein the SNPs associated with an increase in fresh fruit bunch yield are one or two or more genetic markers selected from the group consisting of SEQ ID NO: 1, SEQ ID NO: 3, SEQ ID NO: 6, SEQ ID NO: 7, SEQ ID NO: 8, SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 13, SEQ ID NO: 14, SEQ ID NO: 18, SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21, SEQ ID NO: 23, SEQ ID NO: 26, SEQ ID NO: 27, SEQ ID NO: 30, SEQ ID NO: 33, SEQ ID NO: 34, SEQ ID NO: 35, SEQ ID NO: 36, SEQ ID NO: 37, SEQ ID NO: 39, SEQ ID NO: 40, SEQ ID NO: 41, SEQ ID NO: 43, SEQ ID NO: 44, SEQ ID NO: 45, SEQ ID NO: 46, SEQ ID NO: 47, SEQ ID NO: 48, SEQ ID NO: 51, SEQ ID NO: 52, SEQ ID NO: 53, SEQ ID NO: 54, SEQ ID NO: 56, SEQ ID NO: 57, SEQ ID NO: 58, SEQ ID NO: 60, SEQ ID NO: 62, SEQ ID NO: 63, SEQ ID NO: 64, SEQ ID NO: 65, SEQ ID NO: 66, SEQ ID NO: 67, SEQ ID NO: 70, SEQ ID NO: 71, SEQ ID NO: 72, SEQ ID NO: 73, SEQ ID NO: 75, SEQ ID NO: 76, SEQ ID NO: 77 and SEQ ID NO: 79.
- 20
- 25
16. The method of claim 15 wherein the SNPs associated with an increase in fresh fruit bunch yield are one or two or more genetic markers selected from the group consisting of SEQ ID NO: 8, SEQ ID NO: 10, SEQ ID NO:13, SEQ ID NO:19, SEQ ID NO: 20 and SEQ ID NO: 34.
- 30
17. The method according to any one of the preceding claims, wherein the SNPs associated with an increase in bunch number yield are one or two or more genetic markers selected from the group consisting of SEQ ID NO: 8, SEQ ID NO: 27, SEQ ID NO: 48, SEQ ID

NO: 53, SEQ ID NO: 56, SEQ ID NO: 57, SEQ ID NO: 58, SEQ ID NO: 62 and SEQ ID NO: 80 to SEQ ID NO: 115.

- 5 18. The method according to claim 17, wherein the SNPs associated with an increase in bunch number yield are one or two or more genetic markers selected from the group consisting of SEQ ID NO: 48, SEQ ID NO: 53, SEQ ID NO: 58, SEQ ID NO: 80, SEQ ID NO: 82, SEQ ID NO: 83, SEQ ID NO: 84, SEQ ID NO: 86, SEQ ID NO: 90, SEQ ID NO: 92, SEQ ID NO: 94, SEQ ID NO: 96, SEQ ID NO: 99, SEQ ID NO: 100, SEQ ID NO: 105, SEQ ID NO: 106 and SEQ ID NO: 113.
- 10 19. The method according to claim 18, wherein the SNPs associated with an increase in bunch number yield are one or two or more genetic markers selected from the group consisting of SEQ ID NO: 80, SEQ ID NO: 90, SEQ ID NO: 99, SEQ ID NO: 105 and SEQ ID NO: 106.
- 15 20. The method according to any one of the preceding claims, wherein the presence of at least one, or all markers, selected from a group consisting of Marker no. 1 (SEQ ID NO: 10), Marker no. 2 (SEQ ID NO: 20), Marker no. 3 (SEQ ID NO: 13), Marker no. 4 (SEQ ID NO: 8), Marker no. 5 (SEQ ID NO: 19) and Marker no. 6 (SEQ ID NO: 34), is associated with an increase in fresh fruit bunch weight yield.
- 20 21. The method according to any one of the preceding claims, wherein the presence of at least one, or all markers, selected from a group consisting of Marker no. 1 (SEQ ID NO: 10), Marker no. 3 (SEQ ID NO: 13), Marker no. 5 (SEQ ID NO: 19) and Marker no. 6 (SEQ ID NO: 34), is associated with a decrease in fresh fruit bunch weight yield.
- 25 22. The method according to any one of the preceding claims, wherein the presence of a combination of Markers no. 8 (SEQ ID NO: 105) and 11 (SEQ ID NO: 80) is associated with a change in bunch number yield.
- 30 23. The method according to any one of the preceding claims, wherein the presence of a combination of Markers no. 7 (SEQ ID NO: 90), 9 (SEQ ID NO: 106) and 10 (SEQ ID NO: 99) is associated with a change in bunch number yield.
- 35 24. The method according to any one of the preceding claims, wherein the presence of Marker no. 11 (SEQ ID NO: 80) is associated with a change in bunch number yield.

25. Use of the one or more polymorphic genetic markers, according to any one of the preceding claims, as criteria for the selection of a plant, wherein the selection is based on the presence or absence of one or more polymorphic genetic markers.

INTERNATIONAL SEARCH REPORT

International application No
PCT/MY2015/050034

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68 A01H1/04
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q A01H

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, BIOSIS, Sequence Search, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	NGOOT-CHIN TING ET AL: "High density SNP and SSR-based genetic maps of two independent oil palm hybrids", BMC GENOMICS, BIOMED CENTRAL LTD, LONDON, UK, vol. 15, no. 1, 27 April 2014 (2014-04-27), page 309, XP021193979, ISSN: 1471-2164, DOI: 10.1186/1471-2164-15-309 ----- -/--	1-25



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

6 October 2015

Date of mailing of the international search report

16/10/2015

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Kania, Thomas

INTERNATIONAL SEARCH REPORT

International application No

PCT/MY2015/050034

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WIRULDA POOTAKHAM ET AL: "Development and characterization of single-nucleotide polymorphism markers from 454 transcriptome sequences in oil palm (<i>Elaeis guineensis</i>)", PLANT BREEDING., vol. 132, no. 6, 20 December 2013 (2013-12-20), pages 711-717, XP055218401, DE ISSN: 0179-9541, DOI: 10.1111/pbr.12095 -----	1-25
A	JEENNOR S ET AL: "Mapping of quantitative trait loci (QTLs) for oil yield using SSRs and gene-based markers in African oil palm (<i>Elaeis guineensis</i> Jacq.)", TREE GENETICS & GENOMES, SPRINGER BERLIN HEIDELBERG, BERLIN/HEIDELBERG, vol. 10, no. 1, 11 October 2013 (2013-10-11), pages 1-14, XP035365520, ISSN: 1614-2942, DOI: 10.1007/S11295-013-0655-3 [retrieved on 2013-10-11] -----	1-25
A	WO 2013/012308 A1 (ACGT INTELLECTUAL LTD; CHEAH SUAN CHOO [MY]; YONG SOCK HWA [MY]; LEE W) 24 January 2013 (2013-01-24) -----	1-25
A	WO 2011/084047 A2 (MALAYSIAN PALM OIL BOARD [MY]; HARMINDER SINGH RAJINDER SINGH [MY]; CH) 14 July 2011 (2011-07-14) -----	1-25
A	WO 2010/146357 A1 (SUMATRA BIOSCIENCE PRIVATE LTD [SG]; FORSTER BRIAN PETER [GB]; WILKINS) 23 December 2010 (2010-12-23) -----	1-25
A	WO 2010/056107 A2 (MALAYSIAN PALM OIL BOARD [MY]; HARMINDER SINGH RAJINDER SINGH [MY]; AB) 20 May 2010 (2010-05-20) -----	1-25
A	RAJINDER SINGH ET AL: "Development of simple sequence repeat (SSR) markers for oil palm and their application in genetic mapping and fingerprinting of tissue culture clones", ASIA - PACIFIC JOURNAL OF MOLECULAR BIOLOGY & BIOTECHNOLOGY, UNIVERSITI PUTRA MALAYSIA, vol. 15, no. 3, 1 January 2007 (2007-01-01), pages 121-131, XP007914167, ISSN: 0128-7451 ----- -/--	1-25

INTERNATIONAL SEARCH REPORT

International application No

PCT/MY2015/050034

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	RAJINDER SINGH ET AL: "Identification of cDNA-RFLP Markers and Their Use for Molecular Mapping in Oil Palm (<i>Elaeis guineensis</i>)", ASPAC J. MOL. BIOL. BIOTECHNOL. JOURNAL OF MOLECULAR BIOLOGY AND BIOTECHNOLOGY, vol. 16, no. 3, 1 January 2008 (2008-01-01), pages 53-63, XP055160122, -----	1-25
A	TING NGOOT-CHIN ET AL: "Identification of QTLs Associated with Callogenesis and Embryogenesis in Oil Palm Using Genetic Linkage Maps Improved with SSR Markers", PLOS ONE, vol. 8, no. 1, January 2013 (2013-01), XP002745431, ISSN: 1932-6203 -----	1-25

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/MY2015/050034

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2013012308 A1	24-01-2013	CN 103764842 A HK 1197270 A1 US 2014249046 A1 WO 2013012308 A1	30-04-2014 09-01-2015 04-09-2014 24-01-2013
WO 2011084047 A2	14-07-2011	CO 6592066 A2 WO 2011084047 A2	02-01-2013 14-07-2011
WO 2010146357 A1	23-12-2010	NONE	
WO 2010056107 A2	20-05-2010	CO 6390010 A2 WO 2010056107 A2	29-02-2012 20-05-2010

INTERNATIONAL SEARCH REPORT

International application No.
PCT/MY2015/050034

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☒ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-25

A method for predicting or determining a plant phenotype of interest, wherein the method comprises detecting the presence or absence of one or more polymorphic genetic markers selected from the group of markers as claimed as well as subject-matter related thereto.

Use of said markers as criteria for the selection of a plant, wherein the selection is based on the presence or absence of one or more polymorphic genetic markers.

1.1. claims: 1-25(partially)

The method as use, wherein the plant phenotype of interest is fresh fruit bunch weight yield in oil palm and the polymorphic genetic markers are selected from the group consisting of SEQ ID NO:1-79.

1.2. claims: 1-25(partially)

The method as use, wherein the plant phenotype of interest is bunch number yield in oil palm and the polymorphic genetic markers are selected from the group consisting of SEQ ID NO:8, 27, 48, 53, 56-58, 62, and 80-115.
