The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

DEVELOPING NEW APPROACHES FOR TRANSCRIPTOMICS AND GENOMICS –

USING MAJOR RESOURCES DEVELOPED IN MODEL SPECIES FOR RESEARCH IN CROP SPECIES

Chai Hui Hui, BSc.

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

Jan 2014

School of Biosciences

Abstract

   With the estimated increase in global demand for food and over-reliance on staple food crops, the exploitation of agricultural biodiversity is important to address food security challenges. The aim of this study is to develop approaches to transfer major informational and physical resources developed in model plant and major crop species to resources poor crop species, using oil palm and Bambara groundnut as two exemplar crops. XSpecies (cross-species) approach, the core approach of the study, is described as the approach which uses microarrays developed for a given species to analyse another related species.

   The use of the XSpecies approach (here the cross-hybridisation of DNA from oil palm onto heterologous Affymetrix microarrays for *Arabidopsis* and rice), is the first experiment reported in oil palm and focused on a bulked segregant analysis of different shell-thicknesses for oil palm fruit. Primers design involved screening candidate probe-pairs filtered using PIGEONS software against oil palm transcriptome sequences generated using 454 sequencing technology. The results provided an insight into the effects of sequence divergence between oil palm and the reference species (*Arabidopsis* and rice) onto the power of detecting single feature polymorphism (SFPs) in oil palm, implying the importance of close association between studied and model plant/crop in XSpecies approach.

   The XSpecies approach coupled with genetical genomics was also tested within legumes, with Bambara groundnut as the query species compared to soybean as the resource rich species (20 Mya). A mild drought experiment, conducted in a controlled environment glasshouse, used an F_5 segregating population derived from a controlled cross between DipC and Tiga Nicuru in Bambara groundnut. The cross-hybridisation of Bambara groundnut leaf RNA to the soybean GeneChip individual oligonucleotide probes resulted in a total of 1,531 of good quality gene expression markers (GEMs) on the basis of the differences in the hybridisation signal strength. The first 'expression-based'

genetic map (GEM map) was constructed using 165 GEMs spanning 920.3 cM of Bambara groundnut genome. The first high density DNA-marker genetic map of 1,341.3 cM combining dominant DArT and co-dominant SNPs, developed using the DArT Seq approach, with additional pre-existing microarray-based DArT and SSR markers, was also developed in the F_3 segregating population. Both maps were combined to form the first integrated map of 1,250.7 cM with 212 markers.

Morphological differences and the rapid reduction in *stomatal conductance* observed within the F_5 segregating population in the drought experiment provided trait data for a QTL analysis. The comprehensive QTL analysis in Bambara groundnut detected significant QTLs for morphological traits using GEM map, including *internode length, peduncle length, pod number per plant, pod weight per plant, seed number per plant, seed weight per plant, 100-seed weight, shoot dry weight* and *harvest index* across four linkage groups: LG1, LG2B, LG8B and LG11A. The loci controlling *internode length* and *peduncle length* were also consistently mapped to single marker on LG1 in DArTseq map using F_3 segregating population, suggesting that these two traits are probably controlled by single gene or two closely linked genes. Despite significant genotypes effects on *stomatal conductance* tested in ANOVA analysis, no major QTLs were detected, suggesting the contributions of a number of small genetic effects to *stomatal conductance*. A preliminary homology search using the LG1 linkage group markers and associated gene models showed the ability to develop a framework for identification of candidate genes in Bambara groundnut relative to soybean. The present study also developed the resources for an eQTL analysis in a cross-species context.

Translation from major and model plant species to underutilised and resource poor crops is critical to be able to develop many crop species with potential for future agriculture. This study examines some of the approaches which might be adopted and replicated in various underutilised crop species.

Acknowledgements

Table of contents

List of Tables

List of Figures

List of Appendices

List of abbreviations

ABA	Abscisic acid
ABI	Applied bio system
AFLP	Amplified Fragment Length Polymorphism
AMOVA	Analysis of Molecular Variation
ANOVA	Analysis of Variance
BC	Back cross
BLAST	Basic local alignment search tool
bp	Base pair
CID	Carbon isotope discrimination
CIM	Composite interval mapping
cM	centiMorgan
CV	Coefficient of variation
DArT	Diversity Arrays Technology
DAS	Days after sowing
DE	Days to emergence
DF	Days to flowering
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide Triphosphates (usually mix of dATP/dTTP/dCTP/dGTP)
EDP	Estimated days to podding
EDTA	Ethylene diamine Tetracetic Acid
ELP	Expression level polymorphism
GA	Gibberellin
GEM	Gene expression marker
HI	Harvest index
HSW	100-seed weight
IITA	International Institute of Tropical Agriculture
IM	Interval mapping
IN	Internode length
LA	Leaf area
LG	Linkage group
LOD	Logarithm of odds
MAS	Marker Assisted Selection
mM	Millimole, equal to 10^{-3} mole
MM	Mismatch
MQM	Multiple QTL mapping
NN	Number of nodes per plant

PAR	Photosynthetically active radiation
PCA	Principal Component Analysis
PCR	Polymerase chain reaction
PEL	Peduncle length
PL	Petiole length
PM	Perfect match
PN	Pod number per plant
PS	Plant spread
PT	Permutation test
PVE	Phenotypic variation explained
PW	Pod weight/plant
QTL	Quantitative trait loci
RAPD	Randomly Amplified Polymorphic DNA
RFLP	Restriction Fragment Length Polymorphism
RIL	Recombinant inbreed line
RNA	Ribonucleic Acid
RWC	Relative water content
SDW	Shoot dry weight
SFP	Single feature polymorphism
SMA	Single marker analysis
SN	Seed number per plant
SNP	Single Nucleotide Polymorphism
SS	Size standard
SSR	Simple Sequence Repeat
SSR	Simple sequence repeat
STN	Stem number per plant
SW	Seed weight/plant
TBE	Tris/Borate/EDTA
TDM	Transcript derived marker
TLL	Terminal leaflet length
TLW	Terminal leaflet width
UPGMA	Unweighted pair group method with arithmetic means
UV	Ultraviolet

Chapter 1: INTRODUCTION

The present study aims to use major resources and approaches developed in model plant and major crop species for research and development of less researched crop species. In this study, two crop species, oil palm and Bambara groundnut, were used as exemplar crops. The nucleic acids from oil palm and Bambara groundnut were cross-hybridised separately onto heterologous Affymetrix microarrays (*Arabidopsis* and rice, and soybean). This approach is used to attempt to develop potential molecular markers that are linked to the gene(s) controlling shell thickness in oil palm, as an example. In Bambara groundnut, a combination of XSpecies and genetical genomics were employed to evaluate Bambara groundnut at both genetics and transcriptomics levels. Chapter 1 introduces the two exemplar crop species and provides a detailed review of XSpecies and advanced genetical genomics approaches and their potential application in crop improvement programmes. This is followed by a description of the project overview and objectives.

1.1   CROP SPECIES

1.1.1   Oil palm

1.1.1.1 *Introduction*

Oil palm (*Elaeis guineensis* Jacq.) is a tropical perennial crop belonging to the family Arecaceae, or commonly referred to as the palm family, tribe *Cocoseae* and subtribe *Elaeidinae* (Mayes *et al.,* 2008). Oil palm is a monocotyledon and it is believed to have originated from Central and Western Africa as supported by fossil, historical and pollen sedimentation evidence (Corley and Tinker, 2003). Currently, oil palm is grown across the equatorial tropic region of South-East Asia (SEA), Africa, southern and northern parts of America. Malaysia and Indonesia are the two largest palm oil producing nations followed by Thailand and Nigeria (Hazir *et al*, 2012). The total area of oil palm cultivation in Malaysia alone comprises of about 35% of the global oil palm

cultivated area (Hazir *et al*, 2012). Major plantation groups and the government account for 60% of the oil palm plantation ownership while the rest belongs to private smallholders (Hazir *et al*, 2012).

Oil palm is naturally out-crossing and has 16 pairs of chromosomes ($2n = 2x = 32$) with an estimated haploid genome size of about 1.8 billion base pair (Jouannic *et al.,* 2005). The plant is monoecious which is characterised by the successive production of male and female inflorescence in a single palm, allowing out-crossing to occur (Mayes *et al.,* 2008). The production of fresh fruit bunches (FFB) in oil palm varies according to genotypes and the environment. The FFB usually appears in an oval shape consisting around 1500 fruit/bunch (Mayes *et al.,* 2008). At the matured stage, the fruit is red-brown and consists of mesocarp, shell and kernel. The mesocarp produces edible and orange-red oil (palm oil) whereas kernel yields clear yellowish oil (kernel oil) with the former being the major product (Mayes *et al.,* 2008).

Oil palm seeds need around 100-120 days to germinate (after heat-treatment), followed by 10-12 months in the nursery (Mayes *et al.,* 2008). When the young seedlings are ready for field planting, the seedlings are transplanted to the field and fruiting will only commence from the third year onwards. Oil palm reaches maturity after 10 years of planting but harvesting can be done up to 20-30 years, depending on local planting conditions (Corley and Tinker, 2003).

Oil palm planting materials are grouped into three different fruit types based on the shell thickness trait, controlled by two alleles of the gene, *Sh* (Corley and Tinker, 2003): the thick-shelled '*dura'* fruit type (homozygous; D), the '*pisifera'* fruit type (homozygous; P) which has no shell and is often female sterile, and the *tenera* hybrids (heterozygotes; T), with a thin shell and fibre ring around the shell, derived from a cross between D x P (Corley and Tinker, 2003). Shell thickness is the most important trait in oil palm breeding and research, as the thickness of the fruit shell influences the thickness of the oil bearing

mesocarp. Compared to the *tenera* fruit, the thick shell observed in the *dura* fruit typically generates a 30% lower oil extraction rate. As a result, most of the oil palm plantations in Malaysia and Indonesia have adopted *tenera* as the major planting material due to its fertility and high palm oil yield.

1.1.1.2 *Importance of oil palm*

Two SEA countries, Malaysia and Indonesia, contributed approximately 90% of the world palm oil export trade in 2010 (Rupani *et al*. 2010). In Malaysia, the total export of palm oil products such as palm oil, palm kernel oil, palm kernel cake, oleochemicals, biodiesel and other palm products amounted to RM 71.4 billion and constituted close to 10% of the country total export in 2012 (MPOB, 2012). In the same year, 18.8 million tonnes of crude palm oil was produced with 93.6% of the total production being exported to major countries like China, India and the United States (MPOB, 2012).

Oil palm is an economically important crop due to its high oil-yielding capacity, producing 9.8, 7.8 and 5.6 times more oil yield on average per hectare than soybean (*Glycine max*), sunflower (*Helianthus annuus*) and rapeseed (*Brassica napus*), respectively (Oil World, 2007). In recent years palm oil has overtaken soybean oil to become the largest source of edible vegetable oil constituting 33% of the global vegetable oil production (Saeed, *et al*, 2012). Palm oil production cost is much lower compared to that of soybean and with the higher oil extraction rate, the demand for palm oil will continue to increase. Figure 1.1 shows the predicted global palm oil, soybean oil and rapeseed oil production for the next 6 years (Iowa State University, 2011).

Major Oil Production
Historic and Projected Growth in Thousands of Metric Tons
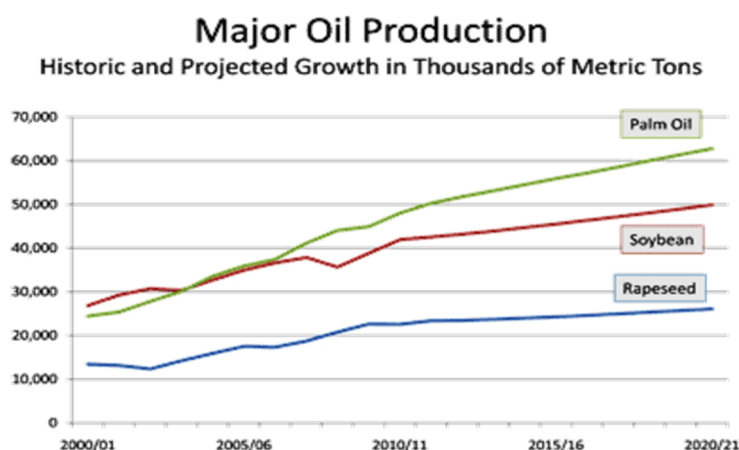
Figure 1.1 The prediction of global production of palm oil, soybean and rapeseed (Iowa State University, 2011).

Palm oil is a versatile commodity and has been used in various sectors ranging from food, pharmaceutical, cosmetic, lubricants to many other industries. Sambanthamurthi *et al.* (2000) stated that 90% of the world's palm oil is used for direct or indirect consumption. Although palm oil has saturated and unsaturated fatty acids ratio of 1:1, research has shown that a diet with a high proportion of palm oil did not promote atherosclerosis and/or arterial thrombosis (Oguntibeju *et al.,* 2009). Palm oil is preferred in producing margarines due to its semi-solid feature at room temperature. Palm oil also offers the advantage of being excluded from the catalyse-based hydrogenation process seen in other temperate vegetable oil which promotes production of trans-fatty acids, leading to cardiovascular diseases (Mayes *et al.,* 2008). In addition, low content of polyunsaturated linoleic acid and a higher level of saturated fatty acids allow palm oil to be used for deep frying purposes (Sambanthamurthi *et al.,* 2000). From the analysis, palm oil has been found to contain high concentration of antioxidants, for example, tocotrienols, beta-carotene, tocopherols and vitamin E (Oguntibeju, *et al.,* 2009). The authors also revealed that the consumption of palm oil can inhibit endogenous cholesterol biosynthesis, reduce blood pressure, reduce oxidative stress, facilitate the

harmoglobinisation of red blood cells and improve immune system (Oguntibeju, *et al.,* 2009).

In addition to edible oil, palm oil also serves as the raw material for biofuel production. The use of biofuels is expected to increase as a consequence of a high demand from developed nations like the US and European countries to fulfil climate change targets and increased energy supply security (Boons and Mendoza, 2010). In Malaysia, the launching of "Envo Diesel" (palm olein blend with diesel) has offered a new opportunity to the local biofuels industry to improve the country's oil palm sector (Jusoff, 2009). In addition, the remaining palm oil mill effluent (POME) is also suggested to be converted into nutraceutical product by Malaysian Palm Oil Board (MPOB) as POME was shown to have phenolics and flavonoids that possess antioxidant properties (Sundram *et al.,* 2003).

Palm kernel oil is widely used in the cosmetic industry to produce luxury soaps or act as a substitute to coconut oil for the production of coffee whiteners, ice cream and confectionary fats (Mayes *et al.,* 2008). Soh *et al.* (2003) also reported the use of palm kernel meal, a by-product of kernel oil extraction, for livestock feed.

1.1.2 Bambara groundnut

1.1.2.1 *Introduction*

Bambara groundnut (*Vigna subterranea* (L.) Verdc) is an indigenous legume that is widely grown by subsistence and small-scale farmers in sub-Saharan Africa. This underutilised crop belongs to the family Fabaceae, subfamily Papilionoideae, and it is the third most important legume after groundnut (*Arachis hypogaea*) and cowpea (*Vigna unguiculata*) in semi-arid Africa (Howell, 1994). It bears protein-rich and nutritious seeds, capable of growing in poor soils and tolerant to drought stress (Heller *et al.,* 1997), allowing

Bambara groundnut to become a potential crop in easing future global food security issues.

The centre of origin of Bambara groundnut has been suggested to be the region between north eastern Nigeria and northern Cameroon, where the wild form of Bambara groundnut were found (Begemann, 1988). The domestication is believed to have occurred within Jos plateau and Yola regions, towards Garoua in Cameroon and probably even Central African Republic (Hepper, 1963; Begemann, 1988). Bambara groundnut has been widely cultivated in tropical regions since the 17th century. In addition to Nigeria, Ghana, Haute Volta as well as Eastern Africa and Madagascar (Benedict, 2010), Bambara groundnut is also grown in South America, Oceania and Asia such as Indonesia, Malaysia, Philippines, India and Sri Lanka (Linnemann and Azam-Ali, 1993; Baudoin and Mergeai, 2001).

There are no improved varieties of Bambara groundnut, all genotypes are mainly landraces that have evolved directly from their wild forms. Doku and Karikari (1971) reported that Bambara groundnut consists of two botanical forms: wild forms (var. *spontanea*) and domesticated forms (var. *subterranea*). Wild forms of Bambara groundnut can be found in the region of Nigeria to Sudan and Cameroon, while domesticated forms are dominant in most of the tropical areas, especially in sub-Saharan Africa (Doku and Karikari, 1971; Basu *et al.,* 2007a). Upon the discovery of high genetic resemblance between wild and domesticated forms by Pasquet *et al.* (1999), the domesticated Bambara groundnut is believed to be derived directly from the wild forms. In addition to further confirming the origin of Bambara groundnut, the higher genetic diversity in var. *spontanea* than var. *subterranea* also allows wild forms of Bambara groundnut to serve as potential sources of advantageous genes for Bambara groundnut breeding programme (Pasquet *et al.,* 1999).

Like most of the underutilised crops, Bambara groundnut has been deprived of extensive research and only limited genomics resources currently

exist.   However, Bambara groundnut possesses highly desirable traits, such as high protein content and tolerance to various biotic and abiotic stresses, enabling this crop to be potentially explored as an alternative crop for food production.

Bambara groundnut is a predominantly self-pollinating crop (cleistogamous) and has 11 pairs of chromosomes, 2n=2x=22 (Forni-Martins, 1986). Bambara groundnut plant has a life cycle of between 110 to 150 days, although some landraces, for example Zebra coloured variety in Ghana takes only 90 days to mature (Berchie *et al.,* 2010). The germination of Bambara groundnut seeds takes 7–15 days under optimal temperature of between 28.5°C and 32.5 °C (Makanda *et al.,* 2009). Flowering starts from 30 to 35 days after sowing and may continue until the end of the crop life cycle. Bambara groundnut requires 30 to 40 days to form pods after fertilisation and reaches maturity under a photoperiod of 12 hours (Basu *et al.*, 2007a).

Bambara groundnut is an annual, herbaceous, intermediate legume of up to 30 cm-35 cm in height with well-developed tap root and lateral roots under the soil (Heller *et al.,* 1997). The roots form nodules in association with *Rhizobia* for nitrogen fixation (Heller *et al.,* 1997). General appearance of the crop, as shown in Figure 1.2, is trifoliate leaves with erect petiole grown from short, creeping, multi-branched lateral stems on the ground level (Heller *et al.,* 1997). Each lateral stem has numerous nodes and the distance (or the length of branch) from the base of the plant to the nearest node is always shorter than the more distant ones (Heller *et al.,* 1997). Due to the length of internodes, Bambara groundnut landraces differ from each other in terms of growth habit, ranging from spreading, semi-bunched to bunch types (Benedict, 2010). The petioles that are borne from the nodes are long, stiff and grooved, with a base of a range of colour such as green, purple and brown (Swanevelder, 1998). In contrast, wild forms of Bambara groundnut exhibit a slightly different appearance in which they have a spreading growth habit, limited numbers of

elongated lateral stems and no distinct tap root with pentafoliate leaves (Swanevelder, 1998; Basu *et al.,* 2007a).
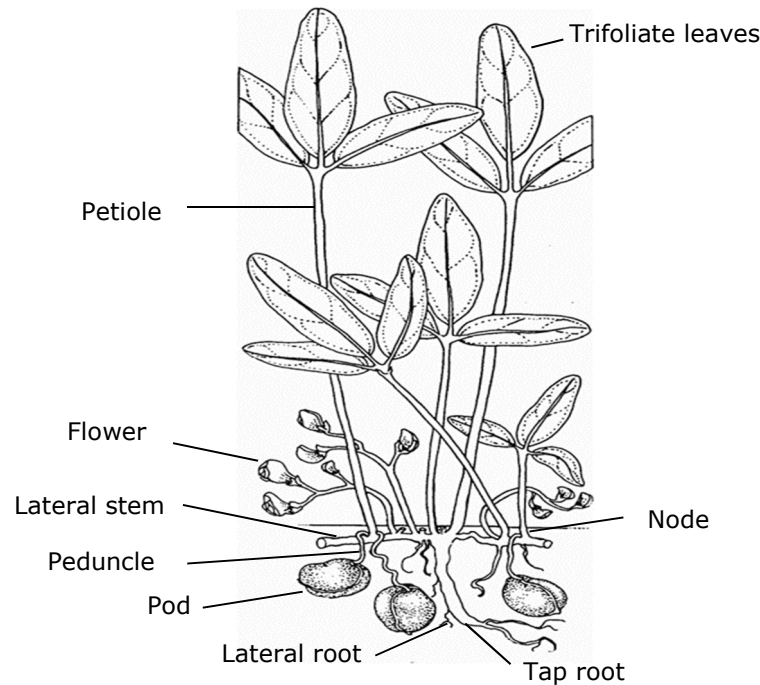
Figure 1.2 The morphology of Bambara groundnut (National Research Council, 2006).

The flowers of Bambara groundnut are typically papilionaceous and produced on long and hairy peduncles which elongates from nodes on the lateral stem (Swanevelder, 1998). The opening of the flowers on the same peduncle does not exceed 24 hours (Benedict, 2010). In addition, the colour of the flowers changes from yellow-whitish in the early morning to pale yellow or even light brown in the evening (Heller *et al.,* 1997). After pollination and fertilisation the peduncles elongate until their maximum length and bring the fertilised ovary into the soil or just above the ground level for pod formation (Heller *et al.,* 1997; Basu *et al.,* 2007a).

The size of the pods ranges from 1.5 cm to 2.5 cm in diameter (Swanevelder, 1998), although some reports show a pod size of 3.7 cm in diameter, depending on the number of seeds inside the pod (Heller *et al.,* 1997). The pods are generally yellow-greenish colour when they are young but when

approaching maturity stage, they are cream yellow and green colour or change to dark brown and red colour in some landraces (Massawe *et al.*, 2003). The pods are round, oval or spherical in shape and many of them contain only one seed. However, Pasquet and Fotso (1997) reported that some landraces produce pods with two or more seeds. The mature pods are indehiscent and contain seeds of various colours, ranging from cream, yellow, brown, red and black, to seeds with or without hilum colouration depending on landraces (Swanevelder, 1998). In addition to growth habit, the pod size is another major difference between wild and domesticated forms of Bambara groundnut. The domesticated material is reported to have larger seeds (1.1-1.5 cm in diameter) which do not wrinkle as compared to wild forms (0.9-1.1 cm in diameter; Basu *et al.,* 2007a).

1.1.2.2 *Importance of Bambara groundnut*

Bambara groundnut is mainly grown for human consumption as it contains sufficient protein, carbohydrates and oil. On average, the seed contains 18%-26% protein with high concentration of essential amino acids such as lysine (6.8%) and methionine (1.3%) (Brough and Azam-Ali, 1992; Borough *et al.,* 1993; Heller *et al.,* 1997). Bambara groundnut therefore, provides an alternative and a cheaper source of protein compared to protein derived from other sources such as animals and fish. Furthermore, the seed contains 51-70% carbohydrates, 3.0-5.0% ash, 5.0-12.0% fibre and 6-12% oil (Rowland, 1993). The crop is not considered as an oil seed legume because the oil content is relatively low compared to oil seed legume, such as groundnut which contains 45.3-47.7% oil (Brough and Azam-Ali, 1992).

Table 1.1 shows a complete composition of micronutrients in Bambara groundnut seed (Amarteifio *et al.,* 2006). Amarteifio *et al.* (2002) and Kemo (2000) suggested that the nutrient content may vary depending on the environment and the landrace. The micronutrients values in Bambara groundnut

are comparable or even higher than some other legumes such as soybean (*Glycine max*), which contains 1,730 mg potassium, 250 mg magnesium and 15.7 mg iron per 100 g of soybean (Holland *et al.,* 1995). Bambara groundnut also has an advantage over the other common pulses such as cowpea (*Vigna unguiculata*), pigeonpea (*Cajanus cajan*) and lentil (*Lens culinaris*) for the high gross energy value in the seed (FAO, 1982).

Table 1.1 The composition of micronutrients in Bambara groundnut seed (mg 100 g^{-1}; Amarteifio *et al.,* 2006).

Micronutrient	Content (mg 100 g^{-1})
phosphorus	313 - 561
iron	23 - 132
calcium	37 - 128
potassium	1,545 - 2,200
magnesium	159 - 332
sodium	16 - 25

Bambara groundnut is largely consumed by the local community in several ways. The fresh pods are boiled with salt and pepper and eaten as a snack in many West African countries (Heller *et al.,* 1997). Linnemann (1990) proposed that Bambara groundnut seed could be pounded into flour for baking purpose or making into a stiff porridge. An experiment was conducted to compare the flavour and composition of milk derived from Bambara groundnut, cowpea, pigeonpea and soybean (Brough *et al.,* 1993). The authors showed that milk produced from Bambara groundnut tends to be more mildly flavoured than other similar competitor such as soybean. Although anti-nutritional factors, tannins and trypsin inhibitor, are reported in Bambara groundnut seed, removing seed coat where tannins are located and pasteurising the milk to denature heat-labile trypsin inhibitors would possibly minimise the issues of the presence of anti-nutritional factors in the milk (Brough *et al.,* 1993). Moreover, Bambara groundnut seed and haulm are found to be a source of animal feed and the

leaves are suitable for animal grazing as they contains high levels of nitrogen and phosphorus (Heller *et al.,* 1997).

In addition to high nutritional value, Bambara groundnut is adapted to harsh and unfavourable environments and could play an important role in increasing food production in Africa. Traits such as drought tolerance, adaptation to poor soils, resistance to pests (Obagwu, 2003) and nitrogen-fixing ability allow Bambara groundnut to contribute to sustainable cropping systems and could potentially play a big role in reducing food insecurity and malnutrition (Basu *et al.,* 2007a).

1.2   FROM MODEL PLANTS TO CROP SPECIES

Genomics tools including sequencing, functional genomic analysis and high throughput gene characterisation, are now been used to complement conventional methods for genetic improvement of crop species (Salentijn *et al.,* 2007). The application of major resources developed in model plants to study crop species is essential and has been reported in many species such as wheat (*Triticum aestivum*; Peng *et al.,* 1999), *Brassica* (Hammond *et al.,* 2005), cowpea (Das *et al.,* 2008), and blueberry (*Vaccinium corymbosum*; Die and Rowland, 2013). Most of the major traits in crop species for breeding purpose, such as high-yielding characters, abiotic and biotic tolerance, involved complex interactions between genetics and environment and polyploid nature in some crop species like banana (*Musa*), wheat, cotton (*Gossypium*) and peanut (*Arachis hypogaea*) makes breeding for these traits difficult and time consuming. The reason behind transferring information from biological models to other crops is that, the knowledge on gene function, structures and molecular pathways of model species is widely studied. In addition, whole genome sequences of model species are publicly available e.g. *Arabidopsis* (The Arabidopsis Genome Initiative, 2000)*,* rice (*Oryza*; Goff *et al.,* 2002), and *Nicotiana benthamiana* (Bombarely, *et al.,* 2012).

In order to translate the gene functions from a model plant species to a crop species, the candidate genes in a model species is first identified, either using functional genomics or genomic mapping approaches, followed by the extraction of orthologs from the target crop species through comparative genomics or genome-wide sequencing (Salentijn *et al.,* 2007). For example, the dwarfing gene 'Rht' of wheat during 'Green Revolution' is an orthologue of *Dwarf 8* in maize and GAI in *Arabidopsis* (Peng *et al.,* 1999)*.* Finally, the candidate genes are validated in the target crop using several methods, depending on the complexity of the plants and traits of interest (Salentijn *et al.,* 2007).

The candidate genes in model species are identified based on the assumptions that genes with a proven or predicted function in the model species or co-localized with a trait-locus could also control the similar function or traits in the crop species (Salentijn *et al.,* 2007), such as salt tolerance from *Arabidopsis* (Quesada *et al.,* 2002). Krutovsky *et al.* (2004) stated that different genomes within plant families could have collinearity which allows the identification of candidate genes to be conducted on the basis of genomic synteny and also functional genomics. However genomic synteny does not always reflect colinearity as rearrangements and duplications could occur during evolutionary process. This is known to occur in the maize genome, and hence minimise the accuracy and efficiency of using comparative mapping (Lai *et al.,* 2006).

Candidate genes identified from model species could be validated in crop species through alignment of nucleotide sequences or amino acid sequence of genes using BLAST database (Salentijn *et al.,* 2007). In blueberry, candidate genes were extracted from *Arabidopsis*, based on transcriptome data that is publicly available, for identification of genes that play important roles in fruit ripening in blueberry (Die and Rowland, 2013). In addition, genetic linkage and comparative maps also serve as genomic tools to validate the candidate genes in crop species. For example, a gene which is homologous to APETALA1 (AP1) was identified through the use of genetic and physical maps of diploid wheat,

combined with comparative mapping of VRN1 and VRN2 regions in rice, hexaploid wheat and *Sorghum* (Yan *et al.,* 2004).

Recently, cross-hybridisation of heterologous nucleic acids from crop species with microarrays that is derived from model species has been reported. For instance, more than 1,000 single feature polymorphisms (SFPs) in cowpea were detected and validated using a soybean genome array (Das *et al.,* 2008). Furthermore, banana leaf transcriptome subjected to drought stress has been investigated through cross-hybridisation with the Rice GeneChip Genome Array (Davey *et al.,* 2009). The result indicated that approximately 33,700 genes are homologous to rice genes and fifty two of the transcripts were identified to be involved in drought and cold tolerance in rice (Davey *et al.,* 2009). The use of cross-species microarray has extended the application of genomic resources from well-researched crop species to minor and less studied crop species. This approach offers the potential for gene discovery as well as the understanding of complex biological responses, such as regulatory networks in response to phosphorus in *Brassica oleracea* (Hammond *et al.,* 2005). Furthermore, relevant genes of interest can also be identified and developed into markers for crop improvement in the future.

1.3 MODERN TECHNIQUES FOR CROP IMPROVEMENT

Food security faces further challenges such as global climate change, water availability, limitation of arable land and sustainable crop production. It is important to exploit the potential of other crop plants and make improvement in yield, increase abiotic and biotic stress tolerance and improve nutritional quality. Several modern techniques, including linkage mapping, molecular markers, genome sequencing, microarray transcriptome analysis and functional genomics have been used to support crop improvement programmes.

1.3.1 XSpecies microarray

1.3.1.1 *Microarray platforms*

Microarrays have become a powerful and popular tool to analyse gene expression on a large scale and improve the understanding of biological systems and gene regulation at the transcriptional level (Pariset *et al.,* 2009). The transcriptome is the total set of transcripts produced from an individual or particular cell type. Unlike the genome, the transcriptome can differ with external environmental conditions, reflecting the gene expression at any given specific time and conditions for a particular tissue (Pevsner, 2009). Microarrays are commonly used to determine the expression level of transcripts because of their rapid production of data, complete coverage of entire transcriptome on a single array for many species and their flexibility (Pevsner, 2009).

In terms of fabrication, several types of microarrays have been established. For example, the spotted array, produced by depositing and spotting the probes (cDNA, PCR products and oligonucleotides) onto the array surface, and the oligonucleotide *in situ* array (*in situ* synthesised array) which is generated by synthesising the probes onto the arrays directly instead of depositing sequences, such as Affymetrix GeneChip array with short oligonucleotide sequences (25-mer probes; Pariset *et al.,* 2009). The Affymetrix GeneChip array is generated through photolithography, using standard oligonucleotide synthesis protocols associated with photolabile nucleotides that allow specific oligonucleotides to be immobilised onto the chip in order to synthesise those oligonucleotides *in situ* onto a silica substrate (Pevsner, 2009). The Affymetrix GeneChip array is usually known as a single-channel array because the array is capable of providing datasets generated from hybridisation of only one labelled RNA/DNA sample onto the array (Pariset *et al.,* 2009). This is in contrast to two-channel array, for instance the Agilent Dual-Mode platform, which allows cDNA from two samples labelled with two fluorescent dyes like Cy3

and Cy5 that differ in their fluorescence emission wavelengths to hybridise simultaneously onto the same microarray (Pevsner, 2009; Pariset *et al.,* 2009).

Microarray technology has been used for massive gene expression profiling in order to explore the transcriptional responses of plants when they are exposed to different conditions, i.e. diseases, abnormal flowering, fruit production and embryogenesis. Microarray can also be employed for comparative genome studies, microbial detection, identification of SNPs, mutant studies and miRNA detection (Pariset *et al.,* 2009). Currently, there are sixteen Affymetrix GeneChip microarrays available for plant species (Affymetrix, 2011). They can provide reproducible and accurate data which can be stored and compared across experiments. However, due to extensive sequence information required in advance and high manufacturing costs for a microarray, this technology is still limited to several species such as *A. thaliana* (L.) Heynh., barley (*Hordeum vulgare*), rice, and wheat (Affymetrix, 2011).

1.3.1.2 *Principles of XSpecies microarray analysis*

One approach recently developed for the Affymetrix GeneChip platform, which is known as XSpecies (cross-species) microarray approach (http://affy.arabidopsis.info/xspecies/), offers a new prospective to exploit the crop species without a species specific microarray. The XSpecies microarray approach is described as a useful approach to explore oligonucleotide targets of a second species by hybridising nucleic acids onto the Affymetrix oligonucleotide-based microarray of a reference species, also known as first species, such as Affymetrix *Arabidopsis* ATH1-121501 (ATH1) GeneChip (Hammond *et al.,* 2005; 2006). The underlying principle of XSpecies microarray is to take advantage of pre-existing homologous sequences that are conserved within related phylogenetic groups and use this information to determine the putative sequences and identities of an unknown species by comparing overlapping sequences derived from reference species.

GeneChip arrays consist of probe sets with up to 16 probe pairs in each probe set. This is in contrast to most other arrays that use single cDNA or long oligonucleotides to assay a gene as each probe set in GeneChip array is specific to a gene transcript. For example, *Arabidopsis* ATH1 GeneChip contains 11 probe pairs per probe set. Each probe pair consists of a perfect match (PM) and a mismatch (MM) probe, with the PM probe having 25 nucleotides complementary to the design sequence, while the MM probe is the same sequence as the PM probe except for a mismatch at the 13th nucleotide in order to evaluate non-specific hybridisation (Wu *et al.,* 2005). The basic principle for this approach is to extract nucleic acid from target species, followed by hybridisation of fluorescence-labeled or biotin-labeled nucleic acids onto microarrays designed for other species. Examples of proof-of-concept studies reported on XSpecies microarray approaches are summarised in Table 1.2.

Table 1.2 Examples of XSpecies (cross-species) microarray approaches.

	Target species	Affymetrix GeneChip array	Sample descriptions	Comments	Reference
1	Woodchuck (*Marmota monax*)	Human (*Homo sapiens*)	Woodchuck liver.	Gene expression was characterised.	Rinaudo and Gerin (2004)
2	Chinese hamster (*Cricetulus griseus*)	Mouse (*Mus musculus*)	Chinese hamster ovary RNA.	Transcriptomics profiling.	Yee *et al.* (2008)
3	Tomato (*Solanum lycopersicum*) Eggplant (*Solanum melongena*) Pepper (*Capsicum* spp.)	Tomato	Both immature and mature fruit tissues.	Gene expression changes at different stages were observed. Groups of EST as well as genes involved in fruit ripening and development in *Solanaceae* were identified.	Moore *et al.* (2005)
4	Potato (*Solanum tuberosum*)	Tomato	Both control and cold-incubated tubers.	Events in potato tubers cold-induced sweetening were investigated.	Bagnaresi *et al.* (2008)
5	Banana (*Musa* spp.)	Rice	*Musa* cultivar 'Cachaco' pooled-RNA from control and drought stressed leaves.	Transcriptional responses of *Musa* to drought stress were assessed.	Davey *et al.* (2009)
6	Cowpea (*Vigna unguiculata* L. Walp)	Soybean	RNA derived from inbred pure lines CB46 and IT93K-503-1.	Single feature polymorphisms were detected and validated.	Das *et al.* (2008)

After hybridisation, the probe sets that are complementary to the heterologous nucleic acids are chosen through computational analysis involving the creation of a software mask, followed by the analysis of the pattern of hybridisation of samples to selected probes for gene expression studies. Imaging of the resultant signal intensities is carried out in order to examine the transcript abundance when target samples bind to each probe set. Subsequently based on the background-adjusted cell intensities, for example, hybridisation differences between PM and MM probes across a probe set, the signal value is calculated (Wu *et al.,* 2005). When MM values are smaller than PM, the MM signal can be used directly as a measurement of non-specific hybridisation and also act as background signal (Affymetrix, 2002). However, Affymetrix (2002) suggested that the MM signal presented on the array should be excluded if MM values are larger than PM value. An ideal mismatch value is then calculated in order to adjust PM intensity as well as establish log-transformation for robust resulting values (Affymetrix, 2002).

However, due to sequence polymorphisms between two different species when XSpecies microarray approach is applied, the potential issue of inefficient hybridisation of certain transcripts to the probes on the array would probably decrease the detection of transcript abundance (Hammond *et al.,* 2005). In order to minimise the problem of sequence divergence during XSpecies hybridisation, the application of genomic DNA-based (gDNA-based) probe-selection was suggested by Hammond *et al.* (2005). Labelled-genomic DNA from the target species is hybridised onto the array and PM which show high hybridisation values with the heterologous gDNA above a defined threshold would be selected for subsequent transcriptome analysis of such species. For example, gene expression profiles of sheep tissues were analysed with the aid of gDNA-based probe selection after cross hybridisation onto Affymetrix Human U133+2 GeneChip array (Graham *et al.,* 2010). A threshold value is either manually or computationally determined and all probe pairs giving a gDNA signal

above this level are retained for the 'virtual' species chip. Hammond *et al.* (2005) reported a 13-fold increase in the sensitivity of *Arabidopsis* ATH1 GeneChip when detecting the regulation of gene expression of *Brassica oleracea* to phosphorus (P) stress following gDNA-based probe selection approach. For example, 111 genes that significantly differentially regulated when exposed to P stress were estimated at a gDNA hybridisation intensity threshold of 500, compared to eight genes when using no probe-selection (Hammond *et al.,* 2005).

Optimal gDNA hybridisation intensity thresholds are different for every single species used in the XSpecies microarray approach, hence re-optimisation of threshold is important as their gDNA origin and quality will affect whole hybridisation intensities across the probe sets. For instance, the gDNA hybridisation threshold of 500 was applied for *B. oleracea* (Hammond *et al.,* 2005) but a cut off level of 550 was used by Davey *et al.* (2009) for *Musa.* Thus, the genomic DNA-based probe selection approach can be used to select appropriate probes and also enhance the sensitivity required for detecting different transcripts expressed between two species. Similar principles could also be applied when XSpecies microarray approach is conducted at DNA level for comparative genome analysis as well as SNP marker development based on the sequence differences identified between two samples (Das *et al.,* 2008).

As compared to species specific arrays, the capability of the XSpecies microarray approach to produce highly reliable data is also questioned (Bar-Or *et al.,* 2007). In this case, several studies have been conducted to compare the sensitivity and efficiency between XSpecies microarray and species specific array. For instance, potato RNA was hybridised to tomato and potato spotted cDNA microarray, respectively, in order to examine the specificity of data obtained from cross species hybridisation as compared to species specific hybridisation (Bar-Or *et al.,* 2006). The result showed the reduction of signal expressed in tomato array in which only 80 and 52 differentially regulated genes

at day 5 and 10 were expressed while potato array showed 591 and 790 differentially regulated genes from homologous potato RNA (Bar-Or *et al.,* 2006). In addition, custom-made Chinese hamster ovary (CHO) Affymetrix array and mouse array have been selected and compared for gene expression profiling in CHO cells (Yee *et al.,* 2008). The authors indicated that seven to eleven probe pairs in most of the probe sets on CHO array passed the minimum criterion for the specificity and sensitivity for XSpecies microarray approach (PM/MM ratio>1.5 and PM-MM>50), but only five probe pairs achieved that criterion on the mouse array.

Although XSpecies microarray approach shows less specificity in the detection of heterologous transcripts, several studies have reported the improvement of XSpecies microarray approach using different strategies such as the type of microarray platforms, hybridisation conditions, experimental design and data validation in addition to gDNA-based probe selection (Bar-Or *et al.,* 2006). For example, cDNA microarrays which have longer probe sequences (over hundred nucleotides) are preferred for cross species hybridisation as the probes are sufficiently large to minimise chances of sequence analysis getting affected by the small interspecies differences in nucleotide sequences (Bar-Or *et al.,* 2006). However, the presence of chimeric clones and contamination in cDNA-based probes due to differential quality of cDNA libraries construction has to be taken into consideration (Bar-Or *et al.,* 2007). A larger number of biological replicates as well as suitable microarray platforms with minimal sequence divergence were also suggested for a better performance during XSpecies hybridisation (Bar-Or *et al.,* 2007; Buckley 2007).

The XSpecies microarray approach might not produce data as specific as those from species specific array due to the sequence polymorphism between reference species and target species, however it is a powerful tool to analyse nucleotide differences and gene expression of species with no species specific microarray. By hybridising heterologous nucleic acids onto the microarray

derived from a closely related species, coupled with appropriate probe selection and data analysis, the XSpecies microarray approach can be improved.

1.3.1.3 *Bioinformatics*

The XSpecies microarray approach involves an appropriate analysis after the hybridisation in order to generate valid results. A programme, Microarray Analysis Suite (MAS, Affymetrix) is commonly used to generate .CEL files through the scanning of the intensities for each probe, followed by data analysis using software such as GeneSpring (Agilent Technologies, Palo Alto, CA, USA) with Robust Multichip Average (RMA) normalisation algorithm (Graham *et al.,* 2010). As probe pairs within a probe set give various signal intensities due to different physical binding properties of each probe pair to transcripts from target species, it is more complicated to produce a single expression value for a gene and often causes background noises (Graham *et al.,* 2007). Thus, the normalisation algorithm is important to amalgamate and generate a single signal value for each probe set.

In addition, when a gDNA-based probe selection is utilised to increase the sensitivity of XSpecies microarray approach, a parser script written in Perl is developed to generate probe-masking files. The probe-masking files provide masking effect which allows probe pairs with gDNA hybridisation intensity greater than a defined threshold from the gDNA CEL files to be selected and hence organised in a custom Chip Description File (CDF; Hammond *et al.,* 2006). The CDF files can then be used to interpret RNA CEL files that are generated from the target species with defined threshold. For example, *B. oleracea* transcriptomics analysis was established by comparing *B. oleracea* RNA CEL files with both the *A. thaliana* CDF file and *B. oleracea* gDNA CDF file after a gDNA hybridisation intensity threshold is defined (Hammond *et al.,* 2005).

Following the normalisation, further examination of the data can also be carried out using one-way ANOVA, Welch's test and Benjamini-Hochberg False

Discovery Rate (FDR) multiple testing correction (Hammond *et al.,* 2005; Graham *et al.,* 2007). Recently, a program known as 'Photographically InteGrated En-suite for the OligoNucleotides Screen' (PIGEONS), was developed to investigate the individual oligonucleotides underlying genomic cross-species studies (Lai, 2009). PIGEONS is used to analyse the CEL files obtained from the XSpecies experiments in order to generate a candidate list for potential probe sets that gave reasonable signal strength as well as showed differential signals between two samples. PIGEONS contains three main sections: PIGEONS filter, a cut-off analysis to remove poorly hybridised oligonucleotides, PIGEONS Mining & Image that provides Fold Change Analysis and statistical analysis, and PIGEONS Query which provides an interface for searching probe sets from the database (Lai, 2009). The cut-off function in PIGEONS is similar to probe masking function in gDNA-based probe selection approach as it gives threshold boundaries in which potential probe sets and oligoprobes are selected, and thus increase the specificity of cross-hybridisation. PIGEONS Mining & Image could be used to search for differentially expressed transcripts with single variation on the nucleotide (i.e. SNPs) from one probe set at the genomic level. Lai (2009) reported the effectiveness of using PIGEONS for XSpecies analysis as compared to those established in Hammond *et al.* (2005), Hammond *et al.* (2006) and Broadley *et al.* (2008), and concluded that PIGEONS is able to produce reliable and valid results.

Although several ways of data analysis for XSpecies microarray approaches have been reported, modifications have to be made when new species are used in order to generate accurate results with higher efficiency. Following the data analysis, putative functions of target sequences can be annotated after *in silico* alignment of PM probes derived from reference species with target species gene sequences using BLAST algorithm against public databases like GeneBank (Lu *et al.,* 2009).

1.3.2 Next generation sequencing (NGS) technologies

As sequence of DNA reveals heritable genetic information that forms the basis for developmental processes of all living organisms, DNA sequencing is now a necessity in modern molecular biology. Analysis and annotation of the function of genes using bioinformatics tools is the next important step in order to determine the genes that regulate phenotypes.

The recent introduction of high-throughput instruments capable of processing millions of sequence reads in a single run has revolutionised sequencing technologies. The technology is known as Next Generation Sequencing (NGS). Unlike Sanger sequencing which requires insertion of fragmented DNA into vectors followed by amplification prior to sequencing, NGS technology with an *in vitro* construction of sequencing libraries bypass complex vector-based cloning. NGS generates shorter sequence reads as compared to Sanger sequencing and this influences the assembly process after sequencing, causing difficulties in identifying overlapping regions and alignment of sequence reads from some DNA fragments, particularly in repetitive genomes (Kantardjieff *et al.,* 2009; Horner *et al.,* 2009). In combination with the advancing development of bioinformatics tools, NGS will be continuously improved in order to increase the sequence lengths, numbers and therefore reduce the overall experimental cost.

In terms of the features and performance of platforms, there are three commercially available next-generation DNA sequencers: the Roche (454) GS FLX sequencer, the Illumina genome analyzer and the Applied Biosystems SOLiD sequencer. Due to the longer sequence read lengths that can be obtained from 454 Life Sciences pyrosequencing method for subsequent sequence assembly purpose, it is mostly preferred and has been widely used in several studies such as transcriptome analysis in *Arabidopsis* (Weber *et al.,* 2007), HIV clinical isolate sequencing (Mardis, 2008) and detection of SNPs in the highly polyploid plant, sugarcane (*Saccharum*; Bundock *et al.,* 2009). The concept of "polymerase-

based sequencing-by-synthesis" that has similar starting workflow to 454 pyrosequencing is applied in Illumina. Although Illumina has shorter sequence reads compared to 454 pyrosequencing, greater than ten times more reads can be obtained per run (Horner *et al.,* 2009). For SOLiD sequencing, ligase is involved in catalysis of the sequencing process after emPCR amplification (Mardis, 2008). Horner *et al.* (2009) reviewed the similarity between Illumina and SOLiD for production of sequence reads and showed that the unique "2-base encoding", a kind of quality check on sequence reads, enables SOLiD to offer more advantages than the other sequencers. Furthermore, ligase based reactions are also highly specific, compared to some polymerase reactions.

Each sequencing platform is unique for different applications, including mutation detection, re-sequencing, identification of genetic variation (i.e. SNPs) and gene expression studies. 454 pyrosequencing produces longer sequence reads which give fewer difficulties in assembly. Illumina and SOLiD give larger coverage, through greater sequence generation. A combination of NGS technologies with different platforms would improve the production of sequence reads. For example, a draft genome sequences (32.5 Mb) that integrates sequence information from Illumina, 454 and Sanger sequence data for the forest pathogen *Grosmannia clavigera*, an ascomycete fungus, was assembled and reported to have higher data quality (DiGuistini *et al.,* 2009). In addition, the draft assembly of the wild strawberry genome, *Fragaria vesca,* was established using a combination of 454, SOLiD and Illumina sequence data (Michael *et al.,* 2010). The authors reported that the wild strawberry assembly was first created by assembling 454 data, followed by SOLiD pairs to grow scaffolds and finally the gaps were filled by mapping Illumina contigs to the 454/SOLiD assembly for higher accuracy.

Recently, third generation deep sequencing approaches such as Ion Torrent and Pacific Biosystems which offer shorter run times and lower costs (Table 1.3) have accelerated the development of NGS tools (Genome Web,

2010; Arthur, 2010), but they are currently not widely available. Comparisons of the performance and features of different types of deep sequencing tools are shown in Table 1.3.

Table 1.3 Comparisons of the performance and features of different platforms in NGS tools (Mardis, 2008; Horner *et al.,* 2009; Genome Web, 2010; Arthur, 2010; Clenn, 2011).

| | Platform | | | | |
	Roche (454)	**Illumina**	**SOLiD**	**Pacific Biosystems**	**Ion Torrent**
Sequencing principle	Pyrosequencing	Polymerase-based sequencing-by-synthesis	Ligation-based sequencing	Single molecule real time sequencing (SMRT)	Semi-conductor sequencing
Average read length	400 bp	50 - 100 bp	35 bp	1,000 bp	100 - 200 bp
Number of reads per full run	1 Million	100 – 200 Million	700 Million	0.01 Million	1 Million
Run time	7 -10 h	3 - 5 days	8 days	0.5 – 2 h	2 h
Cost per run	$ 6,000	$ 8,000 - $10,000	$ 6,000- $ 10,000	$ 100 – 900	$ 750

After sequencing, massive amount of sequence data generated from the same DNA fragments are assembled into contigs or singletons. Several short sequence assemblers are recommended, such as CAP3 (Huang and Madan, 1999), Newbler assembler (454 Life Sciences, Roche Diagnostics, Switzerland) and stackPACK (Electric Genetics, US) for EST clustering (Weber *et al.,* 2007). For example, CAP3 has been used to assemble pyrosequenced ESTs in *A. thaliana* as it offers advantages over the other analysis tools, for example, the capability of putting more ESTs into contigs and generate longer contigs than stackPACK. However, due to small overlapping regions of adjacent ESTs, it is still difficult to create full length contigs using CAP3 (Weber *et al.,* 2007).

The next steps following the assembly involve comparing sequence reads to reference databases for functional annotation. In this case, several programmes have been developed to evaluate the degree of similarity of those sequences with closely related species, for instance BLAST, ELAND (developed

together with Solexa Illumina), Short Oligonucleotide Alignment Program (SOAP; Li *et al.,* 2008), Mapping and Assembly with Quality (MAQ) and RMAP (Horner *et al.,* 2009). A recent report by Brautigam *et al.* (2011) recommended the use of commercial CLC bio genomics workbench (CLC bio, US) for NGS downstream analysis in terms of the hybrid assemblies, contigs length, error tolerance and redundancy reduction after comparing with other assembly programs, including SOAP, CAP3, Velvet, MIRA and TGICL.

For NGS technologies, different bioinformatics tools are used for different purposes such as assembly, mapping, functions annotation and SNP discovery. Appropriate tools will maximise the use of the data and hence increase the accuracy of the analysis.

1.3.3 Genetic markers

1.3.3.1 *Types of markers*

The phenotypes of the crop species are influenced by the interactions between genetics and the environment. Due to insufficient knowledge about the number of genetic factors and their importance in determining the phenotypes, breeders face difficulties in predicting and maintaining the performance of the crop species. Genetic markers offer the advantages of increased efficient selection of individuals prior to breeding, identification of genetic diversity of genotypes, routine quality controls and rapid improvement of varieties for important traits through marker-assisted selection (MAS). There are several types of genetic markers, including protein-based markers such as isozymes or DNA-based markers such as restriction fragment length polymorphisms (RFLP), randomly amplified polymorphic DNA (RAPD), amplified fragments length polymorphisms (AFLP), microsatellite or simple sequence repeat (SSR) and single nucleotide polymorphisms (SNP).

Sax (1923) demonstrated the use of morphological markers to detect the differences in seed size, seed coat and pigmentation patterns in common bean

(*Phaseolus vulgaris*). However, the use of morphological markers is restricted due to the pleiotropic effect observed in the morphological traits as well as limited number of markers in most of the populations (Park *et al.,* 2009). In addition to morphological markers, isozymes and other proteins have been used as marker systems. However, isozymes have disadvantages in terms of limited numbers of detectable isozymes and proteins being tissue and development stage specific.

The development of DNA-based markers has greatly improved the understanding of the genetic of crop plants. DNA markers are DNA sequences located at specific site of the genome and segregate from one generation to the next based on Mendel's Law's of Inheritance (Semagn *et al.,* 2006). DNA markers offer advantages over the other marker systems: firstly, the number of DNA markers found in the populations is effectively unlimited. Secondly, DNA markers are not restricted to specific tissues or development stages like isozymes and thirdly, DNA markers directly reflect the genotypes without being influenced by the environment.

Some examples of DNA markers and their features are compared in Table 1.4. RFLP markers are generated when genomic DNA is fragmented using restriction enzymes and result in fragments whose number and size is different among the individuals, populations and species (Semagn *et al.,* 2006). The differences between two individuals of the same species could be obtained as a consequence of point mutation, insertions, deletions, inversions and translocations, thus result in different length of fragments when DNA is cut at the restriction enzyme recognition sites. For example, a cross between an aphid resistant cultivated cowpea and sensitive wild cowpea was screened using RFLP markers for linkage mapping, marker segregation pattern and also investigation of aphid resistant phenotype (Myers *et al.,* 1996). The result showed that aphid resistance gene is closely linked with one RFLP marker, *bg*4D9b, giving rise to a potential for map-based cloning.

Table 1.4 Comparison of widely used isozymes and DNA markers in plants (Park *et al.,* 2009).

	Isozyme	RFLP	RAPD	AFLP	SSR	SNP
Abundance	Low	Medium	Very high	Very high	High	Very high
Types of polymorphism	Amino acid change in polypeptide	Single base change, insertion, deletion, inversion	Single base change, insertion, deletion, inversion	Single base change, insertion, deletion, inversion	Repeat length variation	Single base change
DNA quality	-	High	Medium	High	Medium	Medium
DNA sequence information	-	Not required	Not required	Not required	Required	Required
Level of polymorphism	Low	Medium	High	High	High	High
Inheritance	Co-dominance	Co-dominance	Dominance	Dominance	Co-dominance	Co-dominance
Reproducibility	Medium	High	Low	Medium	High	High
Technical complexity	Medium	High	Low	Medium	Low	Medium
Developmental cost	Medium	High	Low	Low	High in start	High
Species Transferribility	High	Medium	High	High	Medium	Low
Automation	Low	Low	Medium	Medium	High	High

Instead of using restriction enzymes, the second generation of DNA markers employs PCR (polymerase chain reaction) technique for genetic analysis such as RAPD, AFLP and SSR. PCR-based markers have several advantages over morphological and protein-based markers in terms of low cost, small amount of DNA needed for the analysis as well as rapid speed making it possible to conduct large scale experiments (Park *et al.,* 2009). RAPD markers are developed using a single arbitrary primer of 10-12 nucleotides in the PCR reaction and amplify the target sequences after binding the complementary sequences derived from genomic DNA. RAPD markers have been reported for their application to study genetic diversity in Bambara groundnut using 25 African accessions (Amadou *et al.,* 2001). The authors discovered two main groups of accessions that are divided on the basis of their geographic origin: cluster that contained both Nigerian and Cameroon accessions and another cluster that consisted of Zambian accessions and those originating in Zimbabwe. In addition, AFLP markers have also been used for genetic diversity study in Bambara groundnut.

Eleven AFLP primer combinations were reported to generate 49 scorable polymorphic products across 100 accessions collected from Tanzania and resulted in two main groups: Southern agro-ecological zone and mixed accessions from Central, Lake Victoria and Western agro-ecological zones (Ntundu *et al.,* 2003). AFLP technique involves the amplification of adaptor-ligated restriction fragments with adaptor complementary primers that consist of selective nucleotides at their 3'-ends (Park *et al.,* 2009). Reproducibility of AFLP is high compared to RAPD but both AFLP and RAPD are dominant markers which possibly limits their application to analyse the heterozygous populations such as F_2 population (Park *et al.,* 2009). SSR, which is also known as microsatellites markers, have repeat motifs as short as 1-6 bases long and are codominant markers with high reproducibility (Park *et al.,* 2009). SSR markers are widely used in many studies, for example, genome analysis and DNA fingerprinting of oil palm tissue culture clones (Singh *et al.,* 2007), diversity study in rice (Chakravarthi and Naravaneni, 2006), maize (Enoki *et al.,* 2002), soybean (Tantasawat *et al.,* 2011) and Bambara groundnut (Molosiwa, 2012) as well as genetic mapping in *Sorghum* (Wu and Huang, 2007) and rice (Lang and Buu, 2008).

Furthermore, SNPs (single nucleotide polymorphisms) is the third generation of molecular markers where the polymorphisms of a single base difference can be examined by non-gel based assays, such as invasive cleavage, oligonucleotide ligation assay and primer extension (Park *et al.,* 2009). Numerous SNPs exist in plant genome and their frequency can vary with species, ranging from one per 30 bp to one per 500 bp (Park *et al.,* 2009). For example, one SNP in every 170 bp was found in rice (Yu *et al.,* 2002), but one polymorphism in every 200 bp in barley (Rostoks *et al.,* 2005) as well as every 31 bp in non-coding regions and every 124 bp in coding regions in maize (Ching *et al.,* 2002). SNP markers are widely used in gene or QTL discovery and the genetic maps generated using SNP markers are shown to have higher resolution

compared to RFLP or SSR markers (Yu *et al.,* 2011). Yu *et al.* (2011) reported that the grain width-related QTL in rice, *GW5/qSW5*, were accurately mapped at 123 kb when SNP map was utilised as compared to 12.4 Mb region based on RFLP or SSR map. However, to design SNP markers prior sequence information is required, limiting the application of SNP to major species with extensive nucleotide sequence information (Park *et al.,* 2009).

Each marker system has advantages and disadvantages, thus careful consideration is required in choosing one or more marker systems for respective applications.

1.3.3.2 *Diversity Array Technology (DArT)*

Diversity Array Technology (DArT), which is a relatively new molecular marker technique has been used in several species including rice (Jaccoud *et al.,* 2001), *Arabidopsis* (Wittenberg *et al.,* 2005), *Eucalyptus* (Petroli *et al.,* 2012), oilseed crop *Lesquerella* (Cruz *et al.,* 2013) and perennial ryegrass (King *et al.,* 2013). DArT, a microarray hybridisation-based technique, offers numerous advantages over the other marker systems. DArT technique is reported to be cost effective, capable of detecting single base changes, and it is also a high throughput technique which allows germplasm to be characterised rapidly in a single experiment (Cruz *et al.,* 2013). In addition, DArT technique is well suited for research in minor species or underutilised species for the exploitation of genetic diversity in populations, gene discovery for molecular breeding and construction of genetic linkage maps as no prior sequence information is needed (Petroli *et al.,* 2012; Cruz *et al.,* 2013).

DArT technique involves the isolation and fragmentation of genomic DNA using restriction enzyme such as PstI/TaqI (Semagn *et al.,* 2006) or PstI/BstNI (Cruz *et al.,* 2013), followed by the ligation of restricted fragments with adaptors. In order to reduce the genome complexity, primers complementary to the adapters of the fragments are used in a PCR reaction. After cloning and

amplifying the resulting fragments, the fragments are purified and spotted onto a microarray in order to generate a 'Diversity Panel' (Jaccoud *et al.,* 2001). In addition, the PCR amplified products are also labelled with fluorescent dye, Cy3 or Cy5, and hybridised to DArT Diversity Panel for genotyping. Based on the hybridisation signal intensities, the DArT markers which show polymorphisms are selected and assembled in a 'genotyping array' for routine genotyping whenever the assay of any new specimen is required (Jaccoud *et al.,* 2001; Semagn *et al.,* 2006). For example, 7,680 clones derived from a wide representation of 64 *Eucalyptus* species were selected to generate a high density DArT genotyping array for the construction of high density linkage map (Petroli *et al.,* 2012).

Furthermore, DArTseq technique, a new DArT platform which utilised NGS technique, was developed recently. DArTseq technique allows a plate of DNA samples to run within a single lane on the next generation sequencer such as Illumina Genome Analyzer IIx by tagging PstI/RE site specific adapters of the fragments with 96 different barcodes (Cruz *et al.,* 2013). Both microarray-based DArT and DArTseq markers are used for phylogenetic or genetic diversity analyses as well as the construction of genetic maps. For example, a linkage map of *Eucalyptus* was constructed with 564 DArT markers integrated with 1,930 DArTseq markers and 29 SSR markers (Sansaloni *et al.,* 2011). Compared to the microarray-based DArT, DArTseq technique is reported to produce more polymorphic markers (dominant DArT and SNPs markers) (Sansaloni *et al.,* 2011). In *Eucalyptus,* the DArTseq genotyping was reported to generate 2,835 polymorphic markers whereas microarray-based DArT only produced 1,088 high quality markers (Sansaloni *et al.,* 2011). In *Lesquerella*, Cruz *et al.* (2013) also reported high number of markers, 27,748 polymorphic markers, using DArTseq as compared to 2,833 polymorphic markers when microarray-based DArT was used. In addition, DArTseq technique is more cost-effective when compared to microarry based-DArT due to its capability of producing larger quantity of polymorphic markers at similar cost. Therefore, DArTseq technique is an

essential tool and provides a potential for molecular breeding, germplasm analyses and MAS.

1.3.4 Genetic linkage map and quantitative trait loci (QTL) analysis

One of the major applications of molecular markers is the construction of genetic linkage maps for a wide range of species such as cowpea (Gowda *et al.,* 2002), *Sorghum* (Wu and Huang, 2007) and rice (Lang and Buu, 2008). Genetic linkage map can be constructed through determining the position of genes or molecular markers on the chromosomes and the relative genetic distances between them, followed by the allocation of the molecular markers into their linkage group on the basis of the recombination frequency (Jones *et al.,* 1997). Construction of genetic linkage map is essential for QTL analysis, map-based cloning, marker-assisted selection and comparative mapping. Using genetic linkage map, putative genes controlling traits of interest, either qualitative or quantitative traits, can be identified and selected for breeding purpose.

1.3.4.1 *Mapping population and polymorphisms detection*

The parental lines selected for crossing will differ for one or more traits of interest in order to generate a segregating population for genetic linkage analysis. By calculating the recombination values between the markers in the segregating population, the genetic map can be constructed. In addition, population size ranging from 50 to 250 individuals was suggested for genetic mapping, although a larger number of individuals up to 1000 individuals would be preferred for a higher density genetic map (Mohan *et al.,* 1997; Schneider, 2005). Thus, the selection of an appropriate mapping population which reveals allelic differences for one or more traits of interest is essential.

Based on the reproductive mode, crop species are generally categorised into cross-pollinating and self-pollinating species. Pollination in cross-pollinating species, also known as outcrossing species, involves the delivery of pollen grains

from the anther of a flower to the stigma of a different flower from a different plant (same species). Examples of outcrossing species include oil palm, maize and potato. Due to high genetic heterozygosity, the nature of polyploidy in plants and inbreeding depression, the generation of pure lines from cross-pollinating species for linkage study is difficult (Collard *et al.,* 2005). Semagn *et al.* (2006) stated the use of two-way pseudo-testcross, half sib and full sib families derived from controlled crosses to generate mapping population for cross-pollinating species. For example, in cross-pollinating species white clover (*Trifolium repens*), a double-pseudo testcross population of 92 F_1 progenies was generated by pair crossing two phenotypically divergent, heterozygous parental plants (Barrett *et al.,* 2004). Using a heterozygous parent and a haploid or homozygous plant, mapping population for white clover was established.

For self-pollinating species, several different types of mapping populations can be generated such as F_2 and backcross populations, recombinant inbred lines (RIL) and doubled haploid populations (Figure 1.3; Collard *et al.,* 2005; Semagn *et al.,* 2006). The simplest form of mapping populations is F_2 populations, created from selfing of F_1 hybrids generated by crossing two homozygous parental lines, or backcross populations, which are derived by crossing the F_1 hybrids with one of the parental line (Schneider, 2005). These two mapping populations offer the advantages of easy construction and short generation period. For instance, a F_2 population, consisting of 186 plants, derived from a single cross between japonica variety Nipponbare and indica variety Kasalath in rice was produced and used to construct a high density genetic map with 2,275 markers (Harushima *et al.,* 1998).
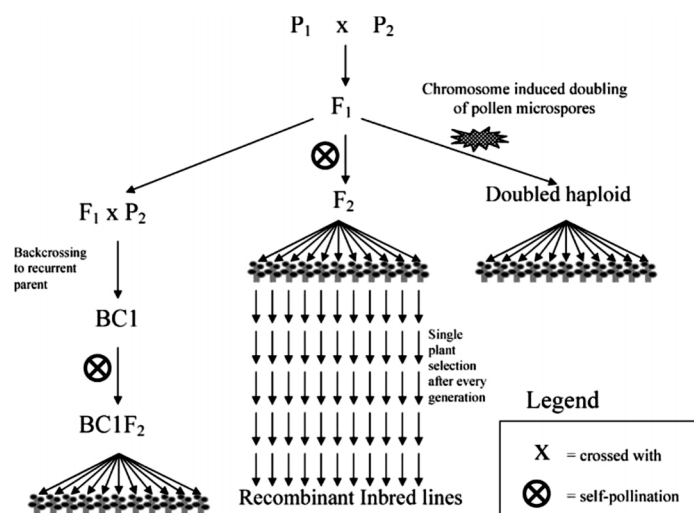
Figure 1.3 The main types of mapping populations for self-pollinating species (Collard *et al.*, 2005).

RILs are developed from inbreeding of individual F_2 plants and also known as single-seed descent lines as one seed of each line in the selfing process is used for next generation (Schneider, 2005). RILs consist of a series of homozygous lines, each containing a different combination of chromosomal segments derived from parental lines. As recombination event can no longer occur in RIL due to complete homozygosity, RIL offers advantages for multiplication and cultivation of plant species without genetic change across different locations and time (Collard *et al.,* 2005). In addition, frequent recombination event that occurred in RIL before reaching homozygosity can result in a higher degree of recombination. As localisation of markers and QTL are highly dependent upon the number of recombination that occurs between genes, RILs are able to produce higher resolution map compared to F_2 populations (Vinod, 2009). The development and use of RILs in several species such as *Arabidopsis thaliana,* rice and oat has been reported. For instance, Meissner *et al.* (2013) reported the use of 250 RILs in *A. thaliana* to construct a linkage map with 391.9 cM as well as to identify QTL for freezing tolerance. The main disadvantage of RILs is the time it takes to generate individual lines. The

generation of RILs requires at least six to eight generations, which is relatively long, in order to obtain nearly 100% of homozygosity in the progenies.

Doubled haploid (DH) plants contain two identical sets of chromosomes in every single cell, resulting from either spontaneous duplication of chromosome number in haploid plants or by the colchicine treatment of haploids (Schneider, 2005). Similar to RIL, DH populations can be considered as permanent resources as they also consists of homozygous plants that can be multiplied and repeatedly planted across different locations and laboratories for further genetic linkage analysis. However, one of the disadvantages is that DH populations are limited to crop species that can be regenerated using tissue culture technique, such as DH wheat lines, which were generated using anther culture technique (Hennawy *et al.,* 2011).

Each population type has its advantages and disadvantages. However, Vinod (2009) proposed that F_2 or RILs are more suitable for use in genetic linkage mapping, followed by QTL analysis as other population types such as backcross population has relatively lower power to detect QTL. Moreover, in order to perform QTL mapping after the construction of genetic linkage map, the same mapping populations have to be phenotypically evaluated and examined (Collard *et al.,* 2005).

Following the selection of mapping population, the next step for constructing the genetic map is to identify the polymorphic markers (Collard *et al.,* 2005). The polymorphic markers employed to screen the whole population, including the parental lines and F_1 hybrids. Polymorphic markers allow an individual to be identified if the individual has inherited phenotypes or traits from maternal or paternal parents. The expected segregation ratios are found to be different according to the types of mapping populations (Table 1.5; Collard *et al.,* 2005). As a result, chi-square analysis is performed for each segregating marker to examine the deviation of the observed segregating patterns from the expected segregation ratios for the mapping population.

Table 1.5 Expected segregation ratios in different types of mapping population (Collard *et al.,* 2005).

Population type	Codominant markers	Dominant markers
F$_2$	1: 2:1 (AA:Aa:aa)	3:1 (B_:bb)
Backcross	1:1 (Cc:cc)	1:1 (Dd:dd)
Recombinant inbred or doubled haploid	1:1 (EE: ee)	1:1 (FF:ff)

1.3.4.2 *Genetic linkage map*

The fundamental principle underlying linkage map construction is the segregation of genes and markers through chromosome recombination during meiosis into the gametes (Semagn *et al.,* 2006). Based on Mendel's second law, which is also known as the law of independent assortment, random assortment of chromosome into gametes during meiosis will result in the alleles of one gene (Aa) to segregate independently with alleles of another gene (Bb), if two genes are unlinked or on different chromosomes (Jones *et al.,* 1997; Semagn *et al.,* 2006). However, genes or markers that are closely linked will segregate together from the parent to the progeny.

The frequency of recombinant genotypes in the segregating population can be used to calculate recombinant value and thus estimate the order and genetic distance between two genes (Collard *et al.,* 2005). The same principle is applied to analyse the segregation of markers for genetic map construction: the lower the frequency of recombination between two markers, the shorter is the genetic distance between two markers on the same chromosome. In contrast, the higher the frequency of recombination between two markers, the further apart two markers located on the same chromosome. As recombination event involves two of the four chromatids at the four-strand stage of meiosis, a recombination frequency of 50% is set as a threshold to determine the linkage of two markers (Jones *et al.,* 1997). Markers with recombinant frequency more

than 50% are considered as unlinked and are assumed to be located on different chromosomes or opposite ends of the same chromosomes where at least one recombination event could occur (Jones *et al.,* 1997; Collard *et al.,* 2005). Conversely, markers that are closely linked will have recombinant frequency of less than 50%.

As a linkage analysis involves large number of markers, it is more feasible to use computer softwares, such as Mapmaker/EXP (Lander *et al.,* 1987), MapManager QTX (Manly *et al.,* 2001) and JoinMap (Van Ooijen, 2006), to calculate the linkages between markers. Among the computer softwares, JoinMap is the most commonly used program to construct the genetic map, for example in rice (Koyama *et al.,* 2001), cotton (Ulloa *et al.,* 2002) and grape (*Vitis*; Wang *et al.,* 2012). Linkages between large numbers of markers are calculated using odds ratios to construct the maps. Odds ratio refers to the ratio of the probability that two markers are linked over the probability that two markers are not linked and it is also often expressed as the logarithm of the ratio, LOD (logarithm of odds) value or LOD score (Risch, 1992). Collard *et al.* (2005) stated that LOD score of 3 and above is always adopted to conduct linkage analysis as a LOD score of 3 indicates that the two markers is 10^3 (1000) times more likely to be linked than unlinked. For example, a RFLP linkage map in rice reported by Xiao *et al.* (1995) showed that RFLP markers were allocated to their respective linkage groups through pairwise analysis with a LOD score of 4.0. Semagn *et al.* (2006) also stated that not all the markers generated for a segregating population were allocated to their respective linkage group. For instance, Petroli *et al.* (2012) reported that the use of 3,198 markers (2,976 DArT and 222 SSR markers) for a F_1 population developed from an inter-specific cross between *Eucalyptus grandis* (clone G38) and *E. urophylla* (clone U15) but only a total of 2,484 markers (2,274 DArT and 210 SSR markers) were mapped.

After assigning markers into respective linkage groups, the genetic distance between markers is calculated prior to map construction.

Recombination frequency is not directly related to the frequency of crossing-over due to the potential of having double or multiple crossovers in the chromosome (Jones *et al.,* 1997; Hart and Jones, 2001). This relationship is likely to happen when the genetic distances between two markers is larger than 10 cM, thus two common mapping functions: Kosambi and Haldane are used to convert recombination fractions into centiMorgans (cM). Kosambi mapping function (Kosambi, 1944) assumes that recombination events can interfere with the adjacent recombination events to a certain extent whereas Haldane mapping function (Haldane, 1931) assumes that there is no interference between crossover in meiosis. Each of the mapping functions has advantages and disadvantages which allow them to be adopted for linkage analysis of different species (Liu *et al.,* 1997).

In addition, it is worth to bear in mind that genetic map is different from a physical map as the genetic distances derived from genetic maps do not directly reflect the physical distances of loci in the chromosomes. Jones *et al.* (1997) reported markers closely linked with genes in a genetic map (1 cM) but the actual distance of the genes in a physical map could be 1 megabase. An example was given by Schmidt *et al.* (1995) who discovered the kilobase pair on the actual chromosome varied from 30 to 550 kb for 1 cM in chromosome 4 of *Arabidopsis*. The relationship between genetic distance and physical distance was suggested to be dependent on the genome size of the plant species (Paterson, 1996) in which the larger the genome size, the larger is the kilobase pair to cM ratio. For example, 120-1,000 kb per cM was found in rice (Kurata *et al.,* 1994) and 118-22,000 kb per cM in wheat (Gill *et al.,* 1996). Thus, a high density genetic map consisting of many markers is required for map-based cloning purpose and also for the integration of genetic and physical maps.

1.3.4.3 *QTL mapping*

Quantitative trait loci (QTL) refer to loci controlling quantitative traits that have measurable phenotypic variation due to several polymorphic genes or environmental factors (Abiola *et al.,* 2003). Most of the quantitative traits are of agronomic importance, such as yield, disease resistance and drought resistance, and can be influenced by one or many QTLs. Compared to qualitative traits, quantitative traits show a normal distribution in the population with phenotypic characteristics that vary in degree among the individuals. The genomic regions controlling quantitative traits can be identified through QTL mapping which involves the process of constructing linkage maps and conducting QTL analysis.

The principle of QTL analysis is to detect the association between phenotypic characteristics and genotype of the markers (Collard *et al.,* 2005). QTL analysis allows the number of genes and their interaction to control the expression of quantitative genes to be identified, and hence provide the tools for crop improvement programs. For example, QTLs that control aluminium tolerance have been analysed using RILs derived from Landsberg *erecta* and Columbia in *Arabidopsis* (Kobayashi and Koyama, 2002). In order to conduct a QTL analysis, linkage map with sufficient polymorphic markers as well as phenotypic data of the same segregating population used for constructing linkage map are essential.

There are at least three widely used methods to detect QTLs: single-marker analysis, simple interval mapping and composite interval mapping. Single-marker analysis employs the use of analysis of variance (ANOVA) and linear regression to detect the association between the QTLs and single markers (Liu, 1998). According to the coefficient of determination (R^2) expressed from markers, the phenotypic variation observed in crop species can be identified if they are regulated by the QTL linked to that marker. For instance, single-marker analysis using simple linear regression was applied in sunflower (*Helianthus*

annuus) to assess the association of markers with nine yield component traits (Anandhan *et al.,* 2010). The authors reported that ORS811 was associated with six traits such as *days to 50% flowering* (R^2=0.33), *days to maturity* (R^2=0.34), *plant height* (R^2=0.46), *volume weight per 100ml* (R^2=0.25), *oil content* (R^2=0.49) and *seed yield* (R^2=0.28). Single-marker system can be easily conducted using statistical software and no complete linkage map is required (Collard *et al.,* 2005). However, single-marker analysis is only applicable when the markers are tightly linked with QTLs. The recombination may occur if the markers are located far from the QTLs and thus minimise the sensitivity and accuracy of detecting QTLs (Tanksley, 1993).

The simple interval mapping (SIM) approach analyses the association of phenotypic variation with the intervals between two adjacent pairs of linked markers along the chromosomes in order to detect the presence of QTL in between two markers (Lander and Botstein, 1989). SIM approach has always been compared with composite interval mapping (CIM) which includes linear regression that examine the association of phenotypic variation with markers in other regions of the genome in addition to an adjacent pair of markers (Jansen, 1993; Basten *et al.,* 2000). For instance, Nagabhushana *et al.* (2006) compared SIM and CIM models to detect QTLs related to growth and yield traits in rice. The result showed that SIM was able to detect five significant QTLs whereas CIM obtained nine significant QTLs that were associated with flowering and maturity. QTLs with higher LOD scores were observed in CIM than in SIM, suggesting that CIM is more accurate and precise in detecting QTLs (Nagabhushana *et al.,* 2006). LOD score is used to identify the position of QTLs located in linkage map (Collard *et al.,* 2005). QTLs with higher LOD score are considered as genuine after comparing LOD score with significant thresholds performed using permutation tests (Churchill and Doerge, 1994).

QTL mapping has been applied in many species in order to identify QTLs controlling agronomic traits that can be employed in crop breeding programme

such as selection and breeding for pest and disease resistance, high-yielding character as well as drought tolerance. For example, the detection of QTLs for flowering and maturity on linkage groups b05 and b06 in common bean under drought stress condition implied the potential of selecting and breeding the genotypes which mature earlier to escape drought (Blair *et al.,* 2012).

1.3.5 Genetical genomics approach

Genetics and gene expression have been studied separately all this while, and these studies have used different technologies, tools and biological materials. In a segregating population, the natural variations observed in the individuals allow the identification of the genomic regions, which are also known as QTL, controlling the phenotypic traits. However, the identification of causal genes within the genomic regions that control phenotypic variation is always a challenge, involving fine mapping or cloning of QTL, which are time consuming and laborious (Joosen *et al.,* 2009).

In addition, gene expression has been largely studied with the increasing availability of genomic sequences and high throughput microarray technology. A typical microarray analysis allows the up- or down- regulation of genes and pathways associated with any specific conditions and developmental stages of a single genotype to be revealed and compared with others. Although transcript abundance and their function can be obtained from gene expression profiling, there is always lack of information regarding the genetic regulation of transcription (Joosen *et al.,* 2009).

Genetical genomics approach, combines gene mapping (genetics) with gene expression analysis to identify loci controlling gene expression and examine the hypothetic regulatory networks (Figure 1.4). Variation in gene expression has been proved to be heritable and shows a quantitative distribution in many studies (Li and Burmeister, 2005). Therefore, linkage map and QTL analyses can be employed for gene expression studies in order to identify the genetic

regulatory loci, or also called expression quantitative trait loci (eQTL), that explain the variation observed in gene expression (Kliebenstein, 2009). The approach was first outlined by Jansen and Nap (2001) and the first proof-of-principle of genetical genomics was performed in *Saccharomyces cerevisiae* (Brem *et al.,* 2002). Following the first report on genetical genomics, the approach has also been applied in crop species such as *Arabidopsis* (Decook *et al.,* 2006; Keurentjes *et al.,* 2007)*,* barley (Potokina *et al.,* 2008) and *Brassica rapa* (Hammond *et al.,* 2011).
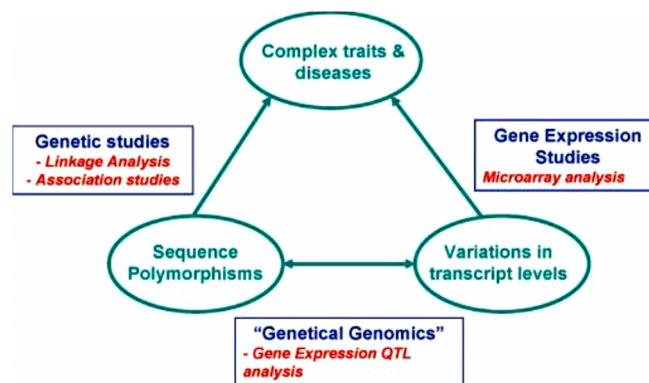
Figure 1.4 Genetical genomics approach combines both genetic studies and gene expression (Li and Burmeister, 2005).

The variation in gene expression generally could be due to many factors, including sequence polymorphisms in target genes, variation in *cis*-regulatory regions in promoter regions or *trans*-regulatory regions, copy number of variation, insertions, deletions and translocation (Joosen *et al.,* 2009). The eQTLs are classified into three types based on the position of variations in DNA structure (Figure 1.5). Local *cis*-eQTLs, as a result of *cis*-regulatory variation in the target gene, can affect the transcription process, transcript stability and also expression of downstream target gene *in trans.* Local *trans*-eQTLs has causal polymorphism near to target gene, within the eQTL confidence interval, but not exactly in the target gene while distant *trans*-eQTLs are located far from target

gene, such as transcription factors. In addition, *trans*-eQTLs are shown to be colocated with variation in the expression level of many genes, ranging from hundreds to thousands of genes (Kliebenstein, 2009). Although the most significant eQTLs are always referred to *cis*-eQTL, identification of hotspots in plants, which are genomic regions with high density of *trans*-eQTL, are thought to represent the major regulatory loci that control the expression of many downstream genes (Kliebenstein, 2009; Joosen *et al.,* 2009).
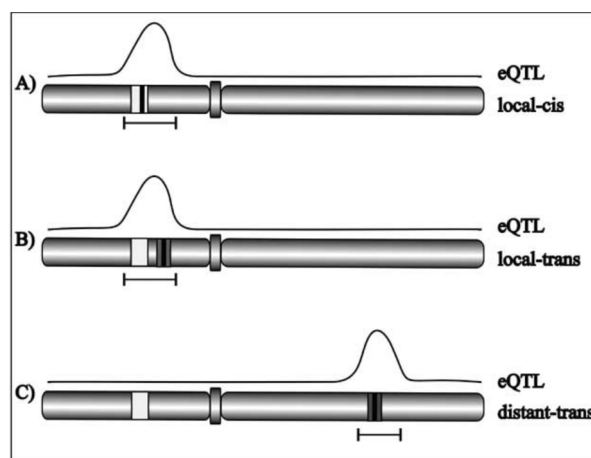
Figure 1.5 Different types of eQTLs (solid line) based on the position of causal polymorphisms (black bar) and the expression of the target gene (light grey box; Joosen *et al.,* 2009).

The use of genetical genomics approach was demonstrated, for example in *Brassica rapa* to examine the regulatory hotspots for phosphorus use efficiency in plants (Hammond *et al.,* 2011). The study reported that using genome sequences that were available, 18,876 eQTL were identified and *trans*-eQTL hotspots occurring on chromosome A06 within *B. rapa* were enriched with phosphorus metabolism-, chloroplast- and photosynthesis-related genes. In addition, Decook *et al.* (2006) also reported the discovery of two eQTL hotspots that were related to shoot generation in *Arabidopsis*. The result showed that most significant eQTLs within the hotspot regions were linked with their corresponding genes but majority were located far apart, suggesting that

heritable *cis*-eQTLs was the causal for the difference in shoot regeneration efficiency whereas *trans*-eQTLs might be involved in downstream effects for the phenotypic traits. Following the identification of eQTLs, the prior information of the selected genes obtained from the gene ontology and biological interactions data can be used to filter the number of potential genes collocated with the phenotypic traits and hence result in the strongest candidate gene for the observed phenotypic variation (Joosen *et al.,* 2009).

An annotated and assembled genome is important to compare the position of genes and the respective eQTLs, but for most of the crop species this is not available. However, several studies in crop species showed that the comprehensive genetical genomics approach can be conducted using genetic maps without the need for annotated genome sequences (Joosen *et al.,* 2009). For instance, Kirst *et al.* (2004) used genetic linkage map to conduct genetical genomic analysis in *Eucalyptus* and discovered that gene expressions of lignin-related genes were regulated by two genetic loci, which were collocated with QTLs associated with stem diameter. Genetic mapping also showed that most of the lignin genes were controlled by *trans*-eQTL hotspots in addition to significant *cis*-eQTL linked to S-adenosylmethionine synthase (Kirst *et al.*, 2004).

Furthermore, the genomic sequence of crop species could also be identified through comparative genomics study with other closely related species. Genes in *VRN2 gene* region in wheat was found to have the same order and orientation in rice and barley, implying that three crop species could potentially use the same genes to control the biological pathway for vernalisation (Yan *et al.,* 2004). Through the use of resources developed from well-established species on the basis of genomic synteny, eQTLs associated with traits of interest could also be identified in crop species that have limited annotated genome sequences. For example, genetical genomics approach was conducted in wheat using a DH population to study eQTLs controlling seed development through synteny analysis (Jordan *et al.,* 2007). Moreover, expressed sequence tag (EST) libraries

(Shi *et al.,* 2007) as well as cDNA-AFLP approach (Vuylsteke *et al.,* 2006) could also be used to generate gene expression profiling when microarray is not available, followed by genetical genomics analysis.

Therefore, by combining gene expression variation to linkage analysis, genetical genomics approach allows the co-localisation of eQTLs, trait QTLs and the actual position of the gene on genetic maps and thus identifies genetic regulatory loci.

1.4   IMPACT OF NEW TECHNOLOGIES IN GENERAL

Due to the potential impact of the global climate change, increasing demand for crop production and limitation of arable land, the development of new technologies to support crop improvement programmes is crucial. In the past, new technologies have been used to understand plant genome organisation, identify regulatory networks in response to abiotic and biotic stress and to establish molecular breeding for the development of new varieties. With the increased knowledge and availability of more powerful technologies, sufficient food supply and poverty alleviation could be achieved in the future.

One of the major applications of the technologies is the exploitation of genetic diversity in crop species using molecular markers. The collection of germplasm resources from different regions is crucial for variety development and improvement of yields. The exploitation of relationships between germplasm allows the understanding of the crop species origin, plant architecture and responses to various abiotic and biotic stresses. With the knowledge, crop improvement could be conducted through continuously breeding with wild strains, domesticated varieties or genetic resources with traits of interest. For example, genetic diversity studies in Bambara groundnut (Amadou *et al.,* 2001; Massawe *et al.,* 2002; Ntundu *et al.,* 2004;) were conducted using molecular markers prior to the selection of parents for mapping studies, leading to effective breeding programme. In addition, the evaluation of genetic diversity

and classification of germplams is essential for the preservation of endangered species and also elimination of the redundant genotypes in gene banks. Park *et al.* (2009) stated that thousands of new accessions are introduced every year as a result of breeding programmes, but due to the limited space and other resources, the redundancy screening using molecular markers is necessary.

As molecular markers can be tightly linked to the genes controlling traits of interest, the use of molecular markers is extended to the generation of high density genetic map in order to locate the QTLs and then predicts the responses and functions of agronomically important genes. For instance, SSR markers have been used to map drought recovery score genes in rice at the position of 0.4 cM from RM201 on chromosome 9, which is related to the length of root and drought tolerance (Lang and Buu, 2008). As a result, the molecular markers allow the breeders to introduce only genes of interest from a related species to cultivated plants, leading to marker-assisted selection (MAS) breeding.

MAS breeding involves screening the population for the absence and presence of the desired traits based on the sequences or band patterns derived from molecular markers associated with genes controlling phenotypic traits (Vinod, 2009). The advantages of MAS breeding include time saving as several characters can be screened simultaneously, selection of desired genotypes at the juvenile stage and the ability to screen the complicated traits such as salt tolerance without phenotypic scoring (Vinod, 2009). Through MAS breeding, new varieties that are environmentally resilient can be produced, such as Bambara groundnut plants with shorter generation time or drought tolerance (Mwale *et al.,* 2007; Mabhaudhi *et al.,* 2013).

In addition, different approaches used in crop breeding research can also be applied in underutilised crops. There are more than 300,000 plant species in the world, but 15 crops (particularly three major crop species; rice, maize and wheat) are used as sources for 90% of human food consumption. By developing new approaches, underutilised crop plants can be explored for medicinal, food,

industrial and renewable energy uses. The XSpecies approach that utilises microarrays derived from major or model plants to evaluate the transcriptomes of crop species provides an alternative way to determine gene expression patterns and also identify nucleotide differences in crop species. For example, single feature polymorphisms related to drought tolerance, brown blotch resistance, photoperiod sensitivity and quality of grain in cowpea were detected using soybean GeneChip (Das *et al.,* 2008). As a result, the polymorphisms can be selected and designed as molecular markers for molecular breeding.

Furthermore, the development of NGS technologies provides opportunities for crop species to be sequenced in a shorter period and at a lower cost (Mardis, 2008; Horner *et al.,* 2009; Genome Web, 2010; Arthur, 2010). Since 2012, genome analyses of 12 crops have been published including that of melon (*Cucumis melo*), chickpea (*Cicer arietinum*), *Citrus* and Cavendish banana (Bevan and Uauy, 2013). Accessing the genetic variation through NGS technologies increases the availability of information for the development of molecular markers and subsequently the genetic mapping of agronomically importance traits. For instance, about 500 SNP markers were obtained in wheat in 2008, however along with the development of NGS technologies the number of the markers have increased from 1,536 to over 90,000 between 2010 to 2012 (Chao *et al.,* 2008; Chao *et al.,* 2010). Through the use of technologies, natural allelic variation can also be discovered and used for improving crop performance.

Global food security is a major concern as an increase of around 70% in crop production is required to fulfil the expected increase in global food demand as the world population rises to 9 billion by 2050 (FAO, 2009). Thus, the development of molecular markers and high throughput technologies would play an important role in meeting future food demand through improved crop production and performance.

1.5   PROJECT OVERVIEW AND OBJECTIVES

This study aims to develop new approaches for genomics and transcriptomics through the use of major resources developed in model species for research in crop species, using oil palm and Bambara groundnut as two exemplar species. In the present study DNA from oil palm was cross-hybridised onto heterologous Affymetrix microarrays (*Arabidopsis* and rice) in order to identify potential SFPs for traits (focusing on the shell thickness genes initially) and thereby generating molecular markers for crop breeding and an understanding of important agronomic traits. The use of XSpecies microarray approach has been demonstrated in many crop species (Hammond *et al.,* 2005; Moore *et al.,* 2005; Bagnaresi *et al.,* 2008; Davey *et al.,* 2009). In Bambara groundnut, a combination of XSpecies and genetical genomics approaches were employed to evaluate Bambara groundnut at both genetics and transcriptomics level. Firstly, a F_5 segregating population derived from the cross between DipC and Tiga Nicuru in Bambara groundnut was subjected to a mild drought condition in a controlled glasshouse, allowing the early responses of Bambara groundnut to drought stress to be studied and also providing the phenotypic traits for QTL mapping. Secondly, RNA from Bambara groundnut was cross hybridised with soybean GeneChip, to develop gene expression markers (GEMs) based on differential hybridisation signals of RNA to individual oligonucleotide probes. These GEMs were used in the construction of a genetic linkage map (GEM map) as well as QTL mapping. In addition, a genetic linkage map (DArTseq map) was also created by combining dominant DArT and SNPs markers (developed using DArT Seq technology) with pre-existing microarray-based DArT and SSR markers using the F_3 segregating population of the the same cross (DipC x Tiga Nicuru), followed by the integration of DArTseq and GEM maps. Thirdly, an attempt was made to overlay Bambara groundnut genetic linkage maps with the 'pseudo physical' map in soybean in order to identify the location of genes on the genetic maps of the two species. The advanced genomic tools provide an insight into

the efficiency of using major resources in model species to study crop species, leading to exploitation of agricultural biodiversity which is potentially important to address food security challenges.

The objectives of the study are:

- To identify potential SFPs, from XSpecies microarray analysis, that are linked to the gene(s) controlling shell thickness in oil palm using a newly developed bioinformatics tool, PIGEONS software.

- To evaluate the effect of drought and changes in gene expression of Bambara groundnut segregating population subjected to mild stress.

- To develop and characterise DArTseq (both dominant DArT and SNPs) markers, and utilise DArTseq markers to construct a high density genetic linkage map using F3 segregating populations.

- To develop GEM markers for use in the construction of an 'expression based' map using F5 segregating population.

- To construct an integrated genetic linkage map using DArTseq and GEM markers derived from two different generational populations of Bambara groundnut, the F_3 and F_5 segregating populations.

- To perform QTL analysis of agronomic and drought-related traits for the mapped populations.

- To provide a framework for identification of candidate genes in Bambara groundnut using soybean 'pseudo physical' map.

Thesis outline:

Chapter 1: The introduction of the oil palm and Bambara groundnut, reviews on modern technologies such as XSpecies microarray approach, molecular markers, genetic linkage map, genetical genomics approach and their impacts are presented. In addition, project overview and the objectives of study are also stated in this chapter.

Chapter 2: Material and methods that generally used throughout the study are described, including list of standard solutions, preparation and quantitation of nucleic acids, polymerase chain reactions (PCR), gel electrophoresis and XSpecies microarrays analysis.

Chapter 3: The use of XSpecies microarray analysis on oil palm using Affymetrix *Arabidopsis* GeneChip and rice GeneChip is reported in this chapter. The development of molecular markers using dataset generated from microarray is focused. In addition, the use of new bioinformatics software, PIGEONS, is also exploited to examine the probe sets and probe pairs that differentially expressed from individual palms with different fruit types.

Chapter 4: A mild drought stress experiment on a F_5 segregating population derived from a cross between DipC and Tiga Nicuru in Bambara groundnut is reported. The distribution of population, morpho-physiological studies and responses of Bambara groundnut plants to early drought stress are focused. Due to the variation between two parental lines, individual plants from segregating population with high-yielding characters and drought tolerance behaviour are reviewed.

Chapter 5: The construction of genetic map of a F_3 segregating population in Bambara groundnut using dominant DArT and SNPs markers, which are developed from DArT Seq technology, is described.

Chapter 6: Cross-hybridisation of Bambara groundnut RNA samples subjected under drought conditions onto Affymetrix soybean GeneChip is described in order to produce GEM markers. Three rounds of analyses for GEM markers development as well as the construction of genetic map using GEM markers for F_5 segregating population are focused. Furthermore, the attempt of integrating DArTseq and GEM map is also reported.

Chapter 7: This chapter focuses on the QTL analysis of agronomically important traits using DArTseq map generated from dominant DArT and SNP markers and GEM map with GEMs, respectively.

Chapter 8: An attempt of identifying the location of genes of the markers represent in Bambara groundnut genetic map using major resources developed in soybean is reported in this chapter.

Chapter 9: General discussion on the study is reported, including potential problem in terms of food security, importance of agricultural biodiversity, review on the application of advanced genomics tools (XSpecies microarray approach combined with genetical genomics approach) in breeding programme, impacts of the findings and also future works.

Chapter 2: GENERAL MATERIALS AND METHODS

In this chapter, materials and methods that were commonly used throughout the study are presented. The protocols and procedures which were specifically conducted in some experiments are explained in respective chapters.

2.6   LIST OF STANDARD SOLUTION

A list of standard solution used for molecular biology experiments are listed as below.

0.5 M EDTA: 186.1 g of EDTA was added into 800 ml H_2O, followed by adding 20 g NaOH pellets while stiring to achieve a pH value of 8.0. After EDTA was dissolved in H_2O, EDTA was filtered using 0.5 micron filter.

5.0 M NaCl: 292.2 g of NaCl was dissolved in 800 ml H_2O, after adjusting final volume up to 1 L NaCl solution was sent for autoclaving.

$T_{10}E_1$ buffer (1X TE): For 50 ml of TE buffer, 500 µl 1 M Tris-HCl (pH 8.0) and 100 µl 0.5 M EDTA (pH 8.0) and H_2O were added together for a final volume of 50 ml, followed by sterilisation with syringe filter.

5X TBE DNA electrophoresis buffer: 54 g Tris base, 27.5 g boric acid and 20 ml 0.5 M EDTA pH 8.0 were added together with 800 ml H_2O. After stiring the final volume was adjusted to 1 L for use.

6X loading buffer (LB): 30% glycerol was prepared using 35 ml molecular grade water. Together with 0.25 g 0.25% (w/v) bromophenol blue, 0.25 g 0.25% (w/v) xylene cyanol and 30 ml 30% glycerol, the solution was topped up with 70 ml H_2O to a final volume of 100 ml.

Lambda DNA (50 ng µl^{-1}): 200 µl uncut Lambda DNA (500 ng µl^{-1}) was mixed with 1,400 µl TE buffer and 400 µl 6X LB for a final volume of 2,000 µl.

2-log DNA ladder (200 ng µl^{-1}): 100 µl 1 µg µl^{-1} 2-log DNA ladder was mixed with 80 µl 6X LB and 320 µl TE buffer.

2.3   QUANTITATION OF NUCLEIC ACID

As each experiment in the study adopted different methods for DNA and RNA extraction, the extraction methods will be explained in respective chapters. The concentration and quality of nucleic acid was examined using spectral absorbance ratios and electrophoretically on an agarose gel.   Spectral absorbance ratios ($A_{260/280}$) of DNA and RNA (ng µl^{-1}) were determined using the Nanodrop ND1000 spectrophotometer (Thermo Scientific, USA) associated with ND-1000 V 3.7.0 software. The pedestal of Nanodrop was first cleaned with 2 µl sterile water, followed by loading 2 µl samples onto the pedestal for measurement. A ratio of ~1.8 was generally accepted for DNA of good quality whereas ratio of ~2.0 was required for RNA.   In addition, Agilent 2100 Bioanalyzer (Agilent Technologies, California, US) was also utilised specifically to determine the integrity of RNA samples.   RNA samples with 2 µl each were loaded into the PCR tubes and sent to Plant Sciences, The University of Nottingham, Sutton Bonington Campus, UK for Agilent analysis. The size of the 18S peak and 28S peaks were then calculated, a ratio of 2 is ideal as 28S/18S ratio is one of the key indicators of RNA quality (Figure 2.1).
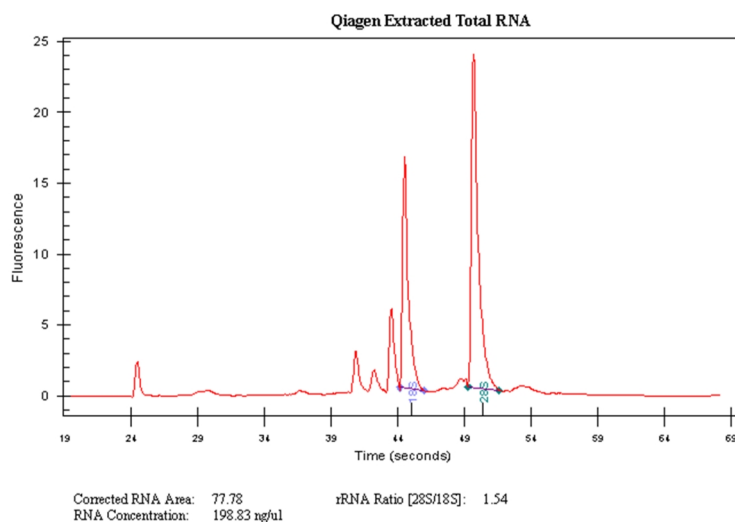
Figure 2.1   Agilent analysis of high quality RNA using Qiagen commercial kit was presented. X-axis: Runtime (s); Y-axis: Fluorescence units.

Another method used to quantitate nucleic acid involved running nucleic acid samples on an agarose gel stained with Ethidium bromide and the intensity of UV-induced fluorescence emitted from the samples was compared with DNA and RNA standards of know concentrations. This method is described in details in Chapter 2.4.

2.4   POLYMERASE CHAIN REACTION (PCR)

The PCR involved *in vitro* amplification of DNA through a series of cycles: DNA denaturation, primers annealing and DNA extension that initiated by thermostable DNA polymerase, such as Taq polymerase.   There are several factors that influence PCR and one of them is annealing temperature. Gradient PCR was used to determine the optimal annealing temperature for the primers.

Annealing temperatures between 50°C to 65°C for each pair of primer were optimised using the Thermo Hybaid PCR Express (Thermo Electron Corporation, US) and the optimal temperature range was used for subsequent reactions performed in the GeneAmp PCR system 9700 (Applied Biosystem, US)

to amplify products of interest. 20 µl PCR master mixes as well as PCR cycle parameters were set up as below:

Table 2.1 PCR mix for 20 µl reactions for each pairs of primers. Larger mastermixes were used in practice to ensure consistency.

PCR components	Volume (µl)	Final concentration
10x Standard PCR buffer (inc. MgCl$_2$ to 1.5 mM)	2	1x
100 mM dNTP mix (25 mM each)	0.16	0.8 mM
NEB Taq	0.1	
DNA (10 ng ul^{-1})	2	
2 µM (10x) reverse primer	2	0.2 µM (1x)
2 µM (10x) forward primer	2	0.2 µM (1x)
MQ-SDW	11.74	
Total	20	

Table 2.2 PCR reaction performed in GeneAmp PCR system 9700 (Applied Biosystem, US).

Pre-denaturation	94 °C	3 min	1 cycle
Denaturation	94 °C	1 min	
*Annealing	50 °C-65 °C	1 min	35 cycles
Extension	72 °C	2 min	
Final Extension	72 °C	10 min	1 cycle
Hold	4 °C/ 20°C	∞	

* Optimal temperature was chosen based on annealing temperature gradient.

For the experiment of oil palm XSpecies microarray analysis, a list of the primers (5′ to 3′ sequences), expected sizes and optimal annealing temperature, resulting from XSpecies analysis, are given in Appendix 2.

2.5   GEL ELECTROPHORESIS

To make a gel, agarose (Bioline, UK) was dissolved in 0.5X TBE buffer and heated in microwave with occasional swirling until a clear solution was observed. After cooling, either SYBR® Safe or Ethidium bromide (0.5 µl; 10 mg ml^{-1} stock; per 50 ml gel) was added and the gel was poured into an appropriate

gel cast tray. DNA, RNA and PCR products were quantitated and/or checked by running them respectively on a 1% (w/v), 1.5% (w/v) and 2% (w/v) stained agarose gel at 80 V for 60 min alongside lambda DNA with two concentrations, 50 ng µl^{-1} and 10 ng µl^{-1}. When PCR products were subjected for analysis, 2-log ladder was also loaded alongside with the samples in order to identify the band size. The gel was then visualised under UV light using the Gel Doc 2000 Gel Documentation System and associated Quantify One 1-D Analysis Software (Biorad, California, US).

2.6   XSPECIES HYBRIDISATION

XSpecies hybridisation involved cross-hybridisation of nucleic acids of target species onto the microarray derived from closely related species. For DNA samples, a minimum volume of 10 µl of 50 ng µl^{-1} DNA sample was prepared to cross-hybridise onto a microarray whereas RNA required a higher concentration which was 100 ng µl^{-1}. Prior to XSpecies analysis, a preliminary quality check was carried out for the samples using Agilent 2100 Bioanalyzer, followed by construction of cDNA or cRNA libraries before hybridisation. XSpecies hybridisation was conducted in The Nottingham Arabidopsis Stock Centre (NASC) International Affymetrix service, The University of Nottingham, Sutton Bonington Campus, UK.

Chapter 3: OIL PALM XSPECIES MICROARRAY ANALYSIS

3.1   INTRODUCTION

3.1.1   Oil palm breeding and selection

Oil palm breeding and selection started formally after 1925 (Hartley, 1967) with the aim of maximising the yield of palm oil and achieving reduced height, disease resistance and high oleic acid oil (Soh *et al.,* 2009). In order to make improvements, establishing a population of palms with substantial genetic variation is important. However, one of the major bottlenecks oil palm breeders have been facing is the relatively narrow genetic base of the *Deli dura* material with only a few ancestral palms contributing to the population (Corley and Tinker, 2003). Thus, there is a need to look for new material to increase the genetic diversity of the base populations, followed by the selection of traits that are genetically variable in order to make genetic improvements in oil palm.

Two main constraints exist when making genetic improvements through oil palm breeding programmes: shell thickness and long selection cycles. In Indonesia and Malaysia, *Deli dura* palms are the main maternal parents which when crossed to *pisifera* palms produce the *tenera* shell-types with 30% more oil per bunch than the thick-shelled *dura* (Corley and Tinker, 2003). *Tenera* has become the preferred commercial planting material that is used today (Figure 3.1; Soh *et al.,* 2009).
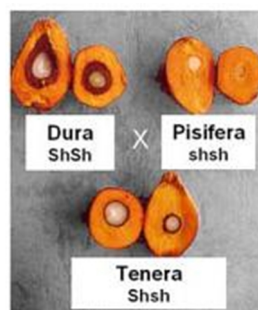
Figure 3.1 The generation of *tenera* by controlled pollination crossing between *dura* and *pisifera* (Soh *et al.,* 2010)*.*

In order to keep *dura* and *pisifera* population separate, a complex breeding scheme is carried out, usually through combined Reciprocal Recurrent Selection (RRS) and Family and Individual Selection (FIS) (Corley and Tinker, 2003; Soh *et al.,* 2009). For the RRS approach, the major disadvantage is that it requires a large population of 500 crosses and 180 selfs to be evaluated over 15-25 years on a land requirement of around 600 ha (Soh *et al.,* 2009). Furthermore, *dura* have a long selection cycle of 10-12 years, with 16 years for *pisifera*, due to the difficult to selection within the often female sterile pisifera pollen parents*.* This requires sib breeding rather than by direct selection of the next generation of *pisifera* (Mayes *et al.,* 2008). As a result, early selection for traits of interest is required to speed up the selection process.

Biotechnology tools could potentially be used to improve and accelerate the selection of individuals in oil palm breeding programmes. Transformation, marker assisted selection (MAS) and tissue culture approaches have been tested to improve selection efficiency for early trait identification and also to propagate selected genotypes (Soh *et al.,* 2009). Mayes *et al.* (2008) reviewed the establishment of bacterial artificial chromosome (BAC) libraries as well as expressed sequence tags (ESTs), combining genetic linkage mapping (as was first reported in oil palm by Mayes *et al.* (1997) in order to begin to construct a physical map. A 3,806 clone array spotted on a slide has also be reported in order to evaluate and compare the expression patterns of more than a thousand genes cloned from normal and abnormal tissue culture material (Low *et al.,* 2006). This is a useful tool for transcriptome analysis, as no microarray has yet been designed and reported for oil palm. However, in the longer term the low number of features of the slide-based microarray would limit the exploitation of hybridisation-based transcriptomic information in oil palm. The first genome sequence to be released for oil palm *Elaies guineensis* and *E. oleifera* was announced in 2013 by a consortium led by Malaysian Palm Oil Board (MPOB), Orion Genomics and the Advanced Biotechnology and Breeding Centre. The

genome had been supplemented with 454-based transcriptome sequences derived from more than 30 tissues, allowed *Sh* gene for fruit types of oil palm and genes affecting other important agronomical and quantitative traits to be identified (Singh *et al.,* 2013). This work represents a major step forward for oil palm, but the physical map is not complete and the process of annotation and functional testing of genes within the physical map is only just beginning.

3.1.2 Application of the XSpecies microarray approach

For species with no available microarray platform, one approach known as the XSpecies microarray could be an alternative option to study oil palm at both genomic and transcriptomic levels.

Microarrays that are designed from *Arabidopsis* have been widely used for cross-species hybridisation (Hammond *et al.,* 2005; Hammond *et al.,* 2006; Graham *et al.,* 2007; Broadley *et al.,* 2008; Davey *et al.,* 2009). As *Arabidopsis* is a model plant, it is believed that a reasonable amount of gene information could be retrieved from *Arabidopsis* in order to understand and study oil palm gene expression patterns after the cross-hybridisation. Willis *et al.* (2008) reported the use of genes from rice to assign thousands of ESTs generated from oil palm into 25 functional clusters of orthologous gene families using the COGsensus software. They also suggested that rice is the most closely related species of monocot for which a complete genomic sequence exists, providing a good source of genetic information to study oil palm.

The aim of the XSpecies microarray approach is to use the oligonucleotides or probes on the microarray developed from a reference species to identify and analyse the corresponding nucleotide sequences from the target species. Using genomic DNA from the target species to select probes allows those probes showing good cross-hybridisation to be identified and a software mask developed to only report signal from those features. Following the genomic DNA hybridisation, gene transcript expression levels under different

conditions, in different stages of development and in different tissues could be extracted and cross-hybridised onto the reference species chip in order to study the pattern of gene expression. The probes identified from reference species could potentially be also used to analyse differences or changes between two different tissues at the transcriptome level such as insertions, deletions, chromosome rearrangements or polymorphisms such as single nucleotide polymorphisms (SNPs), which can then be used as molecular markers.

Mayes *et al.* (1997) reported the development of the first genetic map in oil palm, followed by recent microsatellite-based high density genetic maps reported in Billotte *et al.* (2005) and Seng *et al.* (2011). Using the information or the molecular markers that are derived from genomic hybridisation of oil palm onto high density array chips could improve new high density genetic maps by adding markers based on functional sequences. As molecular markers directly reflect plant genotypes, the development of markers that are closely linked with the shell thickness gene is important in order to accelerate the breeding progress by identifying fruit type at an early stage and before field planting. Oil palm has a long selection cycle, the existence of molecular markers could allow breeders to introduce and introgress only the gene(s) of interest from related species or wide sources of germplasm into their cultivated material.

The XSpecies microarray approach is a promising additional tool until the complete and fully annotated genome of the crop species becomes available and comprehensive oil palm microarrays are created. The differences in hybridisation signal observed in oligonucleotides generated from the XSpecies microarray approach can be validated using quantitative PCR (qPCR) as has been suggested in Hammond *et al.* (2006) in order to confirm the differential expression between different samples. Annotation of functions for different classes of genes can also be carried out using the appropriate bioinformatics tools, such as The Dual Organellar GenoMe Annotator (DOGMA). The use of DOGMA combining BLAST searches against databases developed from tobacco,

rice, and date palm facilitated the chloroplast genome sequence of oil palm to be annotated (Uthaipaisanwong *et al.,* 2012). Therefore, the relevant genes of interest, those homologues to the shell thickness gene, or a series of closely linked molecular markers developed which can distinguish the different haplotypes for shell-thickness, might be identified and their allelic variation in shell thickness can then be distinguished.

In the present study, cross-hybridisation of oil palm genomic DNA onto heterologous microarrays, *Arabidopsis* and rice Affymetrix GeneChip, followed by the identification of differential signal hybridisation between *dura* and *pisifera* across studied populations will be reported. In addition, the alignment of probe-sets and probe-pairs, which are generated from the XSpecies microarray, with oil palm mesocarp transcriptome sequences produced using the 454 next generation sequencing technology is discussed. An attempt using different approaches to design primers is also described in order to generate potential markers for oil palm shell thickness genes, as an example. This work was carried out before the release of the oil palm sequence in 2013.

3.2 MATERIALS AND METHODS

3.2.1 Genomic DNA extraction

3.2.1.1 Minipreparation

Leaf samples from three families of oil palm (751, 768 and 896) were used for these experiments. Each population was derived from the self-fertilisation of a *tenera* palm and palms were identified as either *dura* or *pisifera* and collected individually from the Paloh Estate, Johor, Malaysia (Appendix 1). Leaf samples were kept in a -80°C freezer after surface sterilisation with 3% bleach, followed by three washes with distilled water. DNA was extracted in Malaysia using a modified cetyltrimethylammonium bromide (CTAB) method developed by Applied Agricultural Resources Sdn Bhd (AAR), Malaysia.

Following the DNA extraction from these leaf samples, DNA samples derived from *tenera* self-pollinated (769) family and the four *tenera* parents of the crosses were also supplied directly by AAR, Malaysia (Appendix 1). All DNA samples were then shipped to University of Nottingham (UoN), Sutton Bonington Campus, UK.

3.2.1.2 DNA purification

The DNA samples were purified using a phenol/chloroform extraction method modified from Sambrook and Russell (2001). A total of 50 µl DNA solution was added with 250 µl of SDW to make up 300 µl total DNA solution. An equal volume (300 µl) of phenol: chloroform: isoamyl alcohol (25:24:1) was added, followed by vigorous vortexing for 1 min and centrifugation at 13,000 rpm for 2 min. The aqueous phase was transferred to new tubes without disturbing any protein at the phase interface. The step was repeated by adding an equal volume of chloroform: isoamyl alcohol (24:1) into the tube. The tubes were then centrifuged at 13,000 rpm for 2 min after vigorous vortexing. After cleaning the DNA twice, the clean aqueous phase was transferred to new tubes for additional RNase treatment and ethanol precipitation.

A 1/100 volume of 100 mg ml^{-1} RNAse (NEB, UK) was added to the DNA solution, followed by incubation at 37°C for 30 min to allow RNA digestion. Subsequently, a 1/10 volume of 3 M sodium acetate (pH 5.2) and 2 volumes of ice-cold 100% ethanol were added. The tubes were briefly vortexed, followed by incubation on ice for 30 min. After centrifugation (14,000 rpm, 4°C) for 10 min, a pellet was obtained and washed with 500 μl of 70% ethanol after the supernatant was discarded. The tubes were then centrifuged again at 14,000 rpm for 2 min at 4°C, after removing the supernatant the tubes were left on the bench at room temperature for 15 min to allow residual ethanol to evaporate. Final DNA pellets were dissolved in 50 μl T$_{10}$E$_{0.1}$ buffer at room temperature or incubated at 37°C for 30 min to 1 hour. Quantitation of DNA was carried out using Nanodrop ND1000 spectrophotometer (Thermo Scientific, USA) associated with ND-1000 V 3.7.0 software and by running gel electrophoresis (Chapter 2).

3.2.2 Restriction endonuclease digestion

Two restriction endonucleases (RE), namely *Hin*dIII (Promega, UK) and *Pst*I (NEB, UK), were used for the RE digestion test, to determine whether the DNA was of good quality for further work. Three types of reactions were set up for each RE: DNA with RE, DNA with RE buffer only and DNA with SDW (negative control). Two sets of RE digestion were prepared as below:

<table>
<tr><td colspan="2" align="center">*Pst*I</td><td colspan="2" align="center">*Hin*dIII</td></tr>
<tr><td>**RE component**</td><td>**Volume**</td><td>**RE component**</td><td>**Volume**</td></tr>
<tr><td>DNA</td><td>1 μl</td><td>DNA</td><td>1 μl</td></tr>
<tr><td>10X RE buffer</td><td>2 μl</td><td>10X RE buffer</td><td>2 μl</td></tr>
<tr><td>100X BSA</td><td>0.2 μl</td><td>100X BSA</td><td>0.2 μl</td></tr>
<tr><td>*Pst*I enzyme</td><td>0.2 μl</td><td>*Hin*dIII enzyme</td><td>0.4 μl</td></tr>
<tr><td>SDW</td><td>16.6 μl</td><td>SDW</td><td>16.4 μl</td></tr>
</table>

For the subsequent two reactions, RE enzyme was replaced with SDW. They were then incubated at 37°C for approximately 1 hour before running on a gel.

3.2.3 DNA fingerprinting

The fingerprinting of oil palm DNA samples (Appendix 2) was done at the UoN, Sutton Bonington Campus, using 12 oil palm SSR primers generated by CIRAD (Appendix 3), in order to eliminate illegitimate samples prior to XSpecies analysis. The resulting M13-labelled PCR products were analysed on the Beckman CEQ 8000 DNA sequencer at the Genomics Services Lab, Plant and Crop Sciences, Sutton Bonington Campus, UK.

3.2.4 Bulked segregant analysis

Using a bulked segregant analysis (BSA) approach, equal amounts of DNA from 10 *dura* and 10 *pisifera* palms derived from the same segregating population were pooled into *dura* and *pisifera* bulks, respectively. For example, '768 *dura*' bulk was developed by pooling equal amounts of DNA from 10 *dura* individual palms derived from oil palm 768 family. In total, eight bulks of DNA samples were prepared for cross hybridisation with the *Arabidopsis* Genome ATH1 Array (Affymetrix, US) and five for the Rice Genome Array (Affymetrix, US) (Table 3.1). In addition to the oil palm 768 family and 769 family, 'Superbulk *dura*' and 'Superbulk *pisifera*' that consisted of *dura* and *pisifera* bulked DNA from all four families: 751, 768, 769, and 896 were also prepared with each family contributing equal amounts of DNA for the 'Superbulk'. All samples were then sent to the NASC Affymetrix Service, UoN, Sutton Bonington Campus, UK.

Table 3.1 Bulked DNA samples sent for XSpecies analysis on ATH1 GeneChip and Rice GeneChip.

ATH1	228/05 (*tenera*; parent 768)	Rice	228/05 (*tenera*; parent 768)
	768 *dura* (D)		768 *dura* (D)
	768 *pisifera* (P)		768 *pisifera* (P)
	228/06 (*tenera*; parent 769)		Superbulk *dura* (D)
	769 *dura* (D)		Superbulk *pisifera* (P)
	769 *pisifera* (P)		-
	Superbulk *dura* (D)		-
	Superbulk *pisifera* (P)		-

3.2.5 RNA extraction

The extraction and purification of RNA derived from oil palm mesocarp tissues was conducted prior to transcriptome sequencing through the Roche 454 Pyrosequencing sequencing technology.

3.2.5.1 Minipreparation

Three bunches of oil palm fruits from a single *tenera* oil palm 150/07, each at different developmental stages (F13, F16 and F24) were supplied by AAR, Malaysia. Tissues were kept in a -80ºC freezer after sample collection. A modified TRIzol Reagent protocol (Manufacturer's instruction manual) was followed by AAR-UNMC Biotechnology Research Centre, Selangor, Malaysia to extract RNA from mesocarp tissue. Approximately 50 mg of ground tissue were transferred into a 1.5 ml tube with cold 1 ml TRIzol Reagent (Invitrogen, USA) after grinding under liquid nitrogen. Slight modifications involved an overnight incubation of tubes at -20°C after adding 500 μl isopropanol, instead of incubation at 15°C for 10 min, before precipitation of the nucleic acids.

RNA samples were shipped to the UK in two forms, one was resuspended in SDW whereas the other was as a precipitated pellet under 25 ml 70% ethanol. The latter RNA samples were centrifuged at 3000 rpm for 15 min and the supernatant was discarded. 5 ml of 75% ethanol was added onto the pellet

followed by centrifugation at 3000 rpm for 15 min. After removing the supernatant, they were spun down at 3000 rpm for 1-2 min and the pellets were air-dried. For final resuspension, 100 µl of RNase-free water was added to dissolve the pellets and subsequently transferred to 1.5 ml new tubes.

3.2.5.2 RNA purification

50 µl of RNA samples were first adjusted to a final volume of 100 µl with 50 µl of RNase-free water before purification. A modified RNA cleanup protocol that involved a DNase treatment and the use of RNeasy Mini spin column from RNeasy Qiagen handbook (Qiagen, UK) was then followed. The final RNA products were recovered in 30 µl of RNase-free water. After running RNA samples on the Agilent 2100 Bioanalyzer (Agilent Technologies, US) and using gel electrophoresis for quality control, they were kept at -80°C freezer prior to 454 transcriptome sequencing.

3.2.6 Transcriptome sequencing

RNA were sent to Deep Seq, Centre of Genetics and Genomics, School of Biology, University of Nottingham, UK for transcriptome analysis via Roche 454 sequencing using a full plate of 454 with Titanium reagents (1/3 plate for each developmental mesocarp stage). Subsequently, the transcriptome assembly and analysis was carried out using CLC Genomics Workbench 4th edition (CLC bio, US).

3.2.7 Data analysis using PIGEONS software

Three sets of analyses were done on three different families: 768, 769 and Superbulk using PIGEONS (V1.2) following the guidelines contained in PIGEONS Quick User Guide 2010-2011. Firstly, CDF files derived from the *Arabidopsis* and rice GeneChip were loaded into the software, followed by various CEL files generated from oil palm DNA cross-hybridised onto *Arabidopsis*

and rice, respectively. The use of DNA hybridisation onto the Affymetrix array allows the strength of cross-hybridisation between the subject species (oil palm) and the target species (*Arabidopsis* and rice) to be tested with most gene sequence expected to be at 1 copy per haploid genome. PIGEONS was used to identify the threshold boundary of signal strength, below which probe-pairs were excluded. All features which were above the threshold were then included in a custom CDF file, which were used for subsequent analysis. Secondly, to interpret the experiment, the option 'Mask by single chip' used with the parental *tenera* CEL file to provide the masking data. Thirdly, CEL files of parental *tenera* from each family were also chosen as a 'reference chip' as well as 'parent' while CEL files generated from relevant F_2 offspring *dura* and *pisifera* bulks were selected as 'F_2' in respective analyses. For example, CEL file 228/05 (*tenera*), the parent of the 768 family, was selected as 'reference chip' as well as both 'parent 1' and 'parent 2' whereas CEL files of the offspring 768 D and 768 P bulks were chosen as 'first F_2' and 'second F_2', respectively. In the case of the Superbulk samples, either the CEL file of 228/05 or 228/06 (parent of the 769 family) can be chosen as 'reference chip' and 'parent'.

For Pigeon Filter, cluster validity index, Fukuyama-Sugeno Index with the fuzziness value of two was chosen. The maximum threshold value was set as 1000 with increment of 10. Based on the suggested threshold generated from Pigeon Filter, Dual-fold-change Analysis (DFC) from Pigeon Mining and Image was carried out. As the same parent was used for 'parent 1' and 'parent 2', fold-change value for the 'Parent' was selected as 1 while 'F_2' fold-change value could go from 1.5 to 5, depending on the number of candidate probe-sets as well as oligonucleotide probes that were desired. A signal intensity of 500 was selected as the cut-off for including features in the custom CDF, two categories of signal intensity: 1. 500 and above; 2. 500 and below were implied. After several rounds of filtering probe-sets for one family (i.e. 768) in one analysis, the potential candidates that fit the criteria each time were recorded as a list. They

were then entered into Pigeons Query to cross-check with the other two families (i.e. 769 and Superbulk) to increase the confidence of getting potentially useful probe-sets with consistent hybridisation differences across the comparisons.

3.2.8   Primer design

From the candidate list, PCR primers were designed in four different ways (Appendix 4).   Probe-pairs that flanked the target sequences, where a signal intensity difference between *dura* and *pisifera* were detected in PIGEONS, could be chosen and used to design primers.   Firstly, primers were designed directly from Affymetrix array probe sequences (http://www.affymetrix.com).   Secondly, the Primer 3 software (http://www.bioinformatics.nl/primer3plus), a widely used program for designing PCR primers, was utilised to produce primers with appropriate primer size, GC content, melting temperature (Tm) and also product size on the basis of hybridisation signals observed in PIGEONS software and based on the chip design sequence.   Thirdly, target probe sequences obtained from Array oligonucleotides were entered and searched throughout TblastX database for sequences that align to such sequences from monocot families, such as rice and the palm family.   Degenerate primers were designed from regions that flanked the target sequences after protein to DNA reverse translation.   Fourthly, candidate probe-sets and probe-pairs generated from PIGEONS were overlaid onto the 454 sequence gene models generated from oil palm mesocarp tissue assembly.   Chosen probe-pairs were blasted against the oil palm transcriptome and where they were consistently associated with a single gene model, PCR primers were designed directly from the isotig to amplify the region of potential polymorphism to test if the differences observed *in silico* are genuine differences.   All primers were then tested via PCR on six bulks of oil palm DNA samples: 228/05, 768 D, 768 P, 228/06, 769 D and 769 P.

3.2.9 PCR product clean up and DNA sequencing

Once the PCR amplification products were were purified using GenElute^{TM} PCR Clean Up Kit (Sigma-Aldrich, US) and GenElute^{TM} Gel Extraction Kit (Sigma-Aldrich, US) based on the manufacturer's instructions.   Subsequently, PCR products were sent to Source BioScience LifeSciences, Nottingham, UK for Sanger Sequencing.

3.3 RESULTS

3.3.1 Quantitation of DNA

Quantitation of purified DNA using the Nanodrop indicated high yields of DNA (Table 3.2). No smearing was observed from all genomic DNA on agarose gel but high concentration and an intact band moving at limited mobility on the gel, suggesting that the purified DNA is of good size and quality (Figure 3.2).
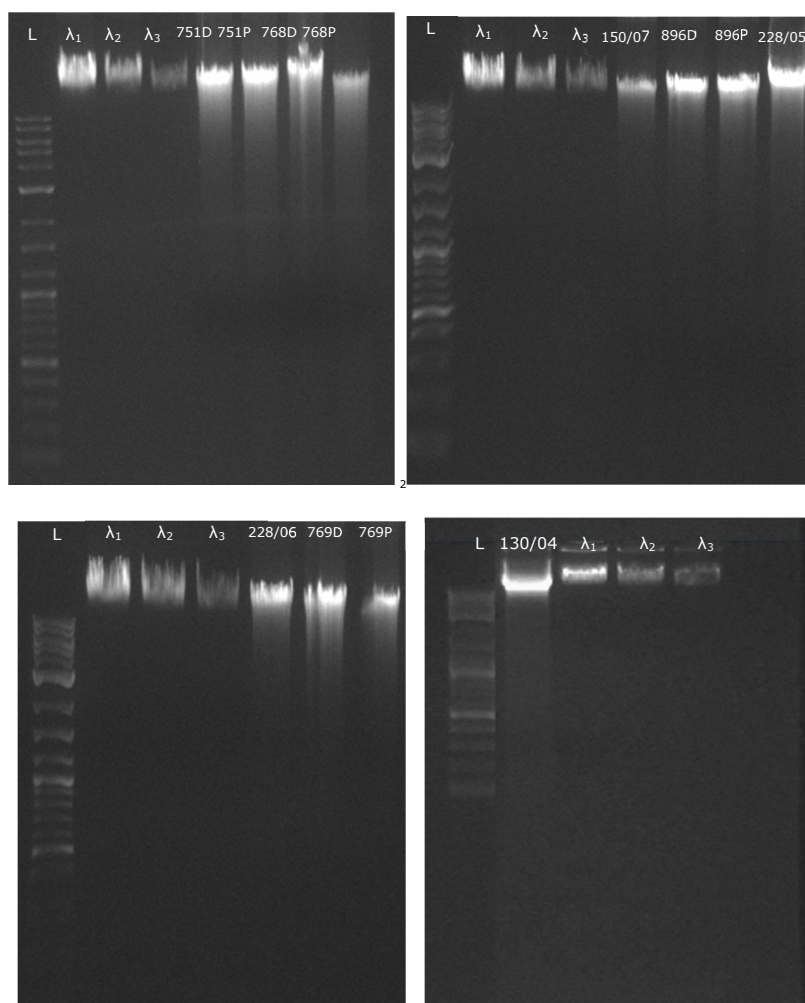
Figure 3.2 Quantitation of DNA used for XSpecies analysis after DNA purification. Lane L= New England Biolabs 2-log DNA Ladder (0.1-10.0 kb). λ is Lambda DNA: λ_1= 500 ng μl^{-1}, λ_2= 250 ng μl^{-1} and λ_3= 125 ng μl^{-1}. D: *dura*; P: *pisifera.*

Table 3.2 Results of DNA quantitation using the Nanodrop for *dura* and *pisifera* bulks as well as *tenera* (parental palm) after DNA purification.

Samples	ng µl^{-1}	260/280
150/07 (*tenera*; parent 896)	644.0	1.74
896 *dura* (D)	675.0	1.78
896 *pisifera* (P)	827.9	1.79
228/05 (*tenera*; parent 768)	728.5	1.78
768 *dura* (D)	721.7	1.66
768 *pisifera* (P)	989.0	1.73
228/06 (*tenera*; parent 769)	725.9	1.76
769 *dura* (D)	292.2	1.73
769 *pisifera* (P)	1102.2	1.67
130/04 (*tenera*; parent 751)	660.2	1.72
751 *dura* (D)	502.2	1.76
751 *pisifera* (P)	677.0	1.72

To test the suitability of the DNA to be digested and to detect the presence of contamination that could inhibit RE digestion and other enzyme activity, a digestion was carried out. All genomic DNA showed smearing when digested with both enzymes but lack of digestion with RE buffer only or SDW (Figure not shown). This confirmed that the quality of the DNA was good. There was neither obvious phenol nor protein contamination and the samples were free from nucleases.

3.3.2 DNA fingerprinting

Oil palm DNA fingerprinting data produced using the CEQ8000 software showed that the 12 SSR primers gave clear and consistent signals (Appendix 2). The markers were polymorphic in the samples and identified that all oil palm materials are derived from the self-pollination of the appropriate parental *Tenera*, except for three incorrect samples, namely; 768/28(D)-D18, 769/A/36(D)–D109, 751/48(P)–P25. These were excluded from the bulks for *dura* and *pisifera* before XSpecies analysis. Table 3.3 shows the DNA fingerprinting data of the *dura* 768 population using 12 SSR markers and the identification of the illegitimate samples which is 768/28(D)-D18. Most of the

individuals are correctly derived from their parental palm using this number of SSRs markers. For the *dura* 768/28(D)-D18 sample it can be seen that among the 12 SSRs used, eight primer sets (namely OP 1, OP5, OP13, OP 11, OP 2, OP 20, OP 18, OP 29) resulted in fragment sizes that are not compatible with the palm being a true descendant of the expected parental palm F1 228/05. OP 1 amplifies two alleles from F1 228/05, with sizes 213 bp or 219 bp, but 768/28(D)-D18 produces a 203 bp allele, indicating that the material is probably resulted derived from a different cross. Overall, the fingerprinting results were good and samples which passed the QC test were pooled into bulks for subsequent analysis and study.

Table 3.3 DNA fingerprinting of *dura* 768, as an example, using 12 SSR primers (A1-C2).

		SAMPLE NAME	Overall	OP1		OP5		OP13		OP12		OP11		OP24/6	
A1	*F1 228/05*	*D2B2par*		*213/219*		*236*		*314*		*192/200/206*		*240/253*		*177/181/190*	
B1	768/49(D)	D9	✓	219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181/190	✓
C1	768/44(D)	D10	✓	219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181/190	✓
D1	768/35(D)	D11	✓	213/219	✓	236	✓	314	✓	192/(200)/206	✓	(240)/253	✓	177/181/190	✓
E1	768/42(D)	D12	✓	213/219	✓	236	✓	314	✓	192/(200)/206	✓	240/253	✓	177/181	✓
F1	768/57(D)	D13	✓	219	✓	236	✓	314	✓	192/(200)/206	✓	253	✓	177/181	✓
G1	768/41(D)	D14	✓	213/219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181/190	✓
H1	768/56(D)	D15	✓	219	✓	236	✓	314	✓	192/(200)/206	✓	253	✓	181/190	✓
A2	768/60(D)	D16	✓	213	✓	236	✓	314	✓	192/(200)/206	✓	(240)/253	✓	177/181	✓
B2	768/31(D)	D17	✓	219	✓	236	✓	314	✓	192/(200)/206	✓	240/253	✓	177/181/190	✓
C2	768/28(D)	D18	✗ (8)	203	✗	236/257	✗	322	✗	192	✓	253/258	✗	181/190	✓

		SAMPLE NAME		OP2		OP7		OP20		OP18		OP21		OP29	
A1	*F1 228/05*	*D2B2par*		*164*		*240*		*225/246*		*292/303*		*225*		*116/122*	
B1	768/49(D)	D9		164	✓	240	✓	225/246	✓	303	✓	225	✓	116	✓
C1	768/44(D)	D10		164	✓	240	✓	246	✓	292	✓	225	✓	116/122	✓
D1	768/35(D)	D11		164	✓	240	✓	246	✓	303	✓	225	✓	122	✓
E1	768/42(D)	D12		164	✓	240	✓	NA*		292	✓	225	✓	116/122	✓
F1	768/57(D)	D13		164	✓	240	✓	NA		292	✓	225	✓	122	✓
G1	768/41(D)	D14		164	✓	240	✓	NA		292	✓	225	✓	116/122	✓
H1	768/56(D)	D15		164	✓	NA		NA		292	✓	225	✓	122	✓
A2	768/60(D)	D16		164	✓	240	✓	225	✓	292/303	✓	225	✓	122	✓
B2	768/31(D)	D17		164	✓	240	✓			292	✓	225	✓	116/122	✓
C2	768/28(D)	D18		158/166	✗	NA		222	✗	294	✗	225	✓	133	✗

3.3.3 RNA quantitation

Quantitation for both sets of RNA was done after RNA purification. There was no significant difference in quality of the RNA resuspended in SDW compared to the resuspended ethanol precipitate pellet. Both preservation methods appear to show clear and intact ribosomal bands on agarose gel, indicating that the yield of RNA was high, no RNA degradation was observed and absence of a band at limiting mobility suggested that the RNA was free from DNA contamination (Figure 3.3). The concentration of the RNA pellet resuspended after ethanol precipitation was slightly lower than the RNA in SDW, perhaps with small amount of RNA is lost during recovery from ethanol solution.
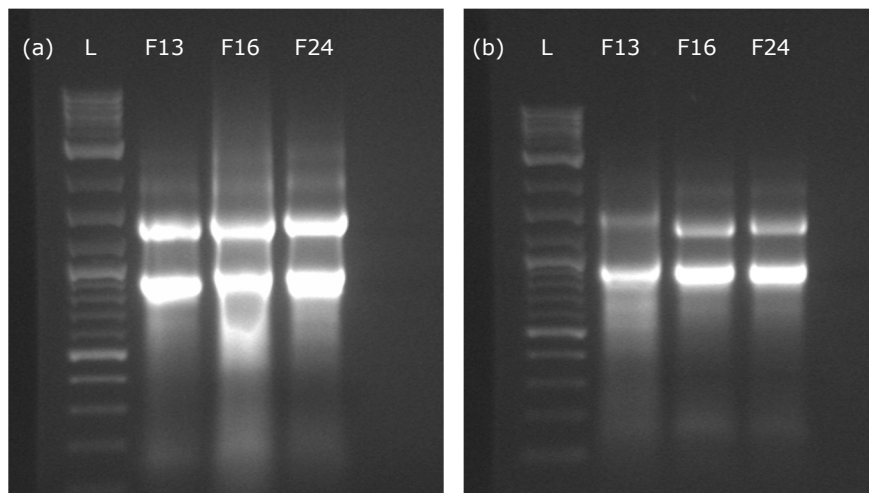
Figure 3.3   Quantitation of RNA after purification and resuspension prior to 454 transcriptome sequencing. (a) RNA samples sent in SDW and (b) RNA samples sent in 70% ethanol. Lane L: New England Biolabs 2-log DNA Ladder (0.1-10.0 kb).

Ribosomal RNA accounts for more than 80% of total RNA with the majority contributed by 18S and 28S rRNA (Bruns *et al.,* 2007). Therefore 28S:18S ratio is one of the key indicators of good RNA quality. However, RNA analysis using the Agilent 2100 Bioanalyzer (Table 3.4) revealed that the RNA probably was not ideal as their 28S:18S ratio was less than 1, whereas a ratio of 2 is preferable (NASC'S International Affymetrix Service, 2011). In addition,

F16_W appears to have substantial degradation from the two peak profile (Figure 3.4). Normally two peaks can be observed in the profiles, with the first peak (18S) giving lower signal than the higher peak (28S).

The three RNA samples, F13_W, F16_E and F24_W which gave the better results based on an Agilent analysis were sent to Deep Seq for initial testing and cDNA library construction. The results confirmed that the RNA was good enough to proceed with 454 transcriptome analysis (result not shown).

Table 3.4 The Concentration and 28S:18S ratio of RNA extracted using Trizol.

RNA samples	ng µl^{-1}	rRNA Ratio [28s/18s]
F13_W	1,726	0.56
F16_W	4,307	-
F24_W	1,739	0.63
F13_E	1,613	0.21
F16_E	1,221	0.40
F24_E	1,045	0.35
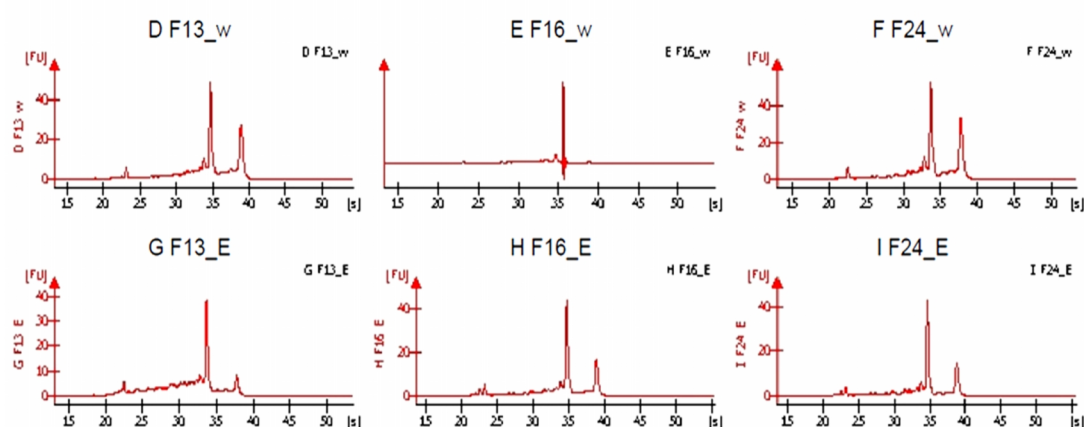
Figure 3.4 The profiles produced by the Agilent 2100 Bioanalyzer for Trizol extracted total RNA. The first peak is 18S while the second peak is 28S.

3.3.4 Generation of potential probes using PIGEONS software

In total 13 CEL files were generated from XSpecies analysis of bulked DNA samples, eight from Affymetrix *Arabidopsis* ATH1 GeneChip and five from Affymetrix Rice GeneChip. As mentioned, PIGEONS software was used to analyse the CEL files, to check for fold-change differences in hybridisation signals of the same oligonucleotide probes on the chips against different *dura* and *pisifera* DNA pools in order to identify potentially polymorphic markers at the DNA level. In this case, two major steps were applied in completing the analysis; threshold selection as well as potential probe-set identification.

3.3.4.1 *Threshold selection*

Pigeon Filter provides a range of threshold boundaries for selection. Based on the threshold value calculated by Pigeon Filter, the number of probe-sets and probe-pairs resulted from cross hybridisation of each oil palm family onto *Arabidopsis* and rice, respectively, is presented in Table 3.5.

For example, CEL files generated from cross-hybridisation of *dura* and *pisifera* bulks derived from the oil palm 768 family on the *Arabidopsis* GeneChip has a suggested threshold of 100, with a target level between 90 and 110 and tolerance interval from 70 to 130 (Figure 3.5). The number of probe-sets and probe-pairs retained varied from 21,777 to 18,619 and 75,387 to 41,743 when the threshold was set between 70 and 130. When a threshold value of 100 is chosen, a relatively high probe-set retention rate was obtained of 89.68% with the ratio of average probe-pairs retained per probe-set being 2.7. However, compared to the *Arabidopsis* microarray*,* cross-hybridisation of the oil palm 768 family-derived bulks onto the rice GeneChip gave stronger hybridisation. At a threshold value of 100, 93.58% of probe-sets are retained, while the ratio of average probe-pairs retained per probe-set was 4.1. The probe-pair retention decreased at a rate of 0.74% per threshold value when the threshold value was changed from 70 to 130 after oil palm 768 family cross-hybridised onto

Arabidopsis GeneChip whereas a 0.59% decrement is seen using the rice GeneChip. As more probe-sets and oligonucleotide probes are retained using the rice GeneChip, the identification of probes was done only on the candidate list generated from rice GeneChip.

Table 3.5 The summary of threshold selection using Pigeon Filter after cross-hybridisation of oil palm to *Arabidopsis* and rice respectively.

Microarray GeneChip	Oil palm family	Threshold		Probe-set	Probe-pairs	Avg Probes /Set	probe-set ret. rate (%)	probe-pairs ret. rate (%)
ATH1	768	L-tolerance interval	70	21,777	75,397	3.5	95.74	30.15
		L-target interval	90	20,911	60,266	2.9	91.93	24.10
		Suggested threshold	100	20,399	54,570	2.7	89.68	21.82
		H-target interval	110	19,815	49,683	2.5	87.11	19.87
		H-tolerance interval	130	18,619	41,743	2.2	81.86	16.69
ATH1	769	L-tolerance interval	30	22,609	121,656	5.4	99.40	48.64
		Suggested threshold	50	22,238	88,420	4.0	97.77	35.35
		H-tolerance interval	60	21,854	76,617	3.5	96.08	30.63
ATH1	Bulk	L-tolerance interval	80	21,969	80,781	3.7	96.58	32.30
		L-target interval	120	20,558	55,936	2.7	90.38	22.37
		Suggested threshold	130	20,067	51,637	2.6	88.22	20.65
		H-target interval	140	19,563	47,844	2.4	86.01	19.13
		H-tolerance interval	170	18,021	38,575	2.1	79.23	15.42
Rice	768	L-tolerance interval	70	55,520	276,965	5.0	96.96	44.00
		L-target interval	100	53,580	217,419	4.1	93.58	34.54
		Suggested threshold	110	52,754	200,876	3.8	92.13	31.91
		H-target interval	120	51,963	188,102	3.6	90.75	29.88
		H-tolerance interval	140	49,907	163,441	3.3	87.16	25.97
Rice	Bulk	L-tolerance interval	90	54,812	249,009	4.5	95.73	39.56
		L-target interval	120	52,686	199,751	3.8	92.02	31.74
		Suggested threshold	135	51,342	180,335	3.5	89.67	28.65
		H-target interval	150	49,979	164,299	3.3	87.29	26.10
		H-tolerance interval	180	47,028	136,930	2.9	82.13	21.75

* Suggested threshold: cut-off point to remove the poorly hybridising oligonucleotides; L-target interval, H-target interval: the lowest and highest value for potential cut-off; L-tolerance interval, H-tolerance interval: the lowest and highest value for feasible cut-off where probe-sets and probe-pairs could be retained.
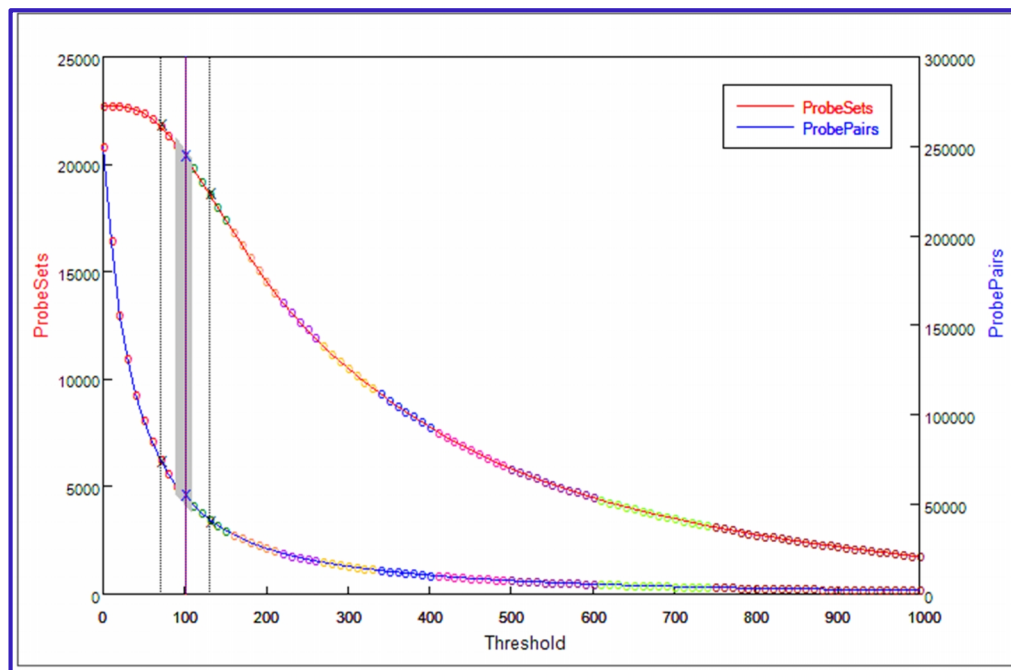
Figure 3.5 Threshold boundaries for the XSpecies analysis obtained from the hybridisation of DNA from oil palm 768 family on Affymetrix *Arabidopsis* ATH1 GeneChip. Red solid line: suggested exclusion threshold; Grey shaded block: target interval; Black dotted line: tolerance interval.

3.3.4.2 *Potential probe set identification*

After cross hybridising on the rice GeneChip, the oil palm 768 family showed a significant decline of the number of probes at the threshold value of 110 when the fold-change value between *dura* and *pisifera* increases from 2.0 to 5.0 (Figure 3.6). For example, 1,533 probe-sets and 1,653 probe-pairs were retained at a fold-change value of 2.0 whereas only four probe-sets as well as probe-pairs were retained at the value of 5.0.

The influence of fold change (FC) value on probe set and probe pair retention
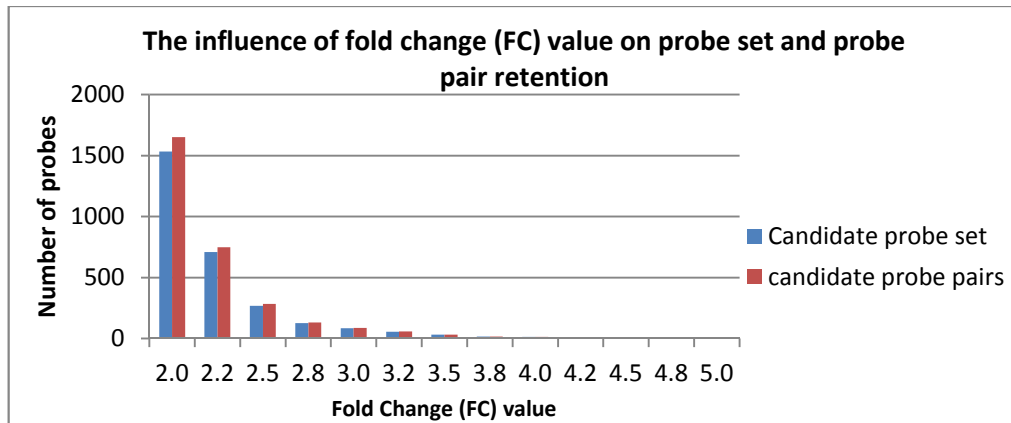
Figure 3.6 The impact of fold-change (FC) value on the number of probe-sets (red) and probe-pairs (blue) retained in rice GeneChip. The number of probes on the rice GeneChip at the threshold value of 110 decreases when the fold-change value between *dura* and *pisifera* increases from 2.0 to 5.0.

Dual-fold-change Analysis (DFC) analysis from Pigeons Mining & Image allows the generation of a candidate list when the probe-sets and probe-pairs are analysed at each threshold level and fold-change value. After cross screening with all families, the potential probe-sets with reasonable fold-change values between two samples at all threshold levels, which are identified from Pigeon Filter, are listed in Appendix 5. From the *Arabidopsis* and rice GeneChips, 31 and 60 probe-sets, respectively, were identified using an initial signal intensity of 500 and above. When the signal intensity falls below 500, only 14 probe-sets from the rice GeneChip were identified.

For example, probe-set 245050_at from the Affymetrix *Arabidopsis* ATH1 GeneChip produces stronger hybridisation signals in the *dura* than *pisifera* at a cut-off value of 100 across bulks of oil palm 768 family, 769 family and superbulks (Figure 3.7). The oil palm 768 family showed a signal intensity of 945 and 353 for *dura* and *pisifera* respectively. In comparison, oil palm 769 family showed signal intensity of 1,567 and 784 while Superbulk family recorded intensity value of 2,022 and 1,722, for the *dura* and *pisifera* bulks. In addition, the oil palm 768 family gives a fold-change difference of 6.3 for probe-pair 10

and 2.7 for probe pair 7 between the *dura* and *pisifera* bulks (Figure 3.7; Appendix 5a)*.* A similar pattern of hybridisation observed across all oil palm samples is hypothesised, if the same allele of the shell-thickness determining gene is present and the marker close enough to the gene. Expected results are shown in 769 and Superbulks, but with a lower fold-change value in probe 10 (769: 1.7; Superbulks: 1.2) and probe 7 (769: 2; Superbulk: 1.2). A lower fold-change value observed in Superbulks would be expected to be due to average signal strength obtained in *dura* and *pisifera*, respectively, as a result of pooling of four families of oil palm DNA into a bulk. However, the region that surrounds probe 10 is more consistent - with low signal strength and small fold-change - so could be used to design primers to amplify the probe 10 region for sequence confirmation. If a difference in a single nucleotide between the two bulks is discovered, it could lead to the development of a marker, for instance, testing the difference observed in the high fold-change value (6.3) at probe 10.

Similar principles were applied to probe-sets generated from the rice GeneChip with two different categories of signal intensity. As the oil palm 769 family is not included in the cross-hybridisation experiments using the rice GeneChip, cross-screening of probe-sets were done only within oil palm 768 family and Superbulk. Although only 14 probe-sets are obtained when restricted to a signal intensity of 500 and below, on average higher fold-change difference was observed between two samples in both oil palm 768 family and Superbulk (Appendix 5c). For instance, probe-pair 4 from probe-set Os.53248.1.A1_at gives the highest fold-change value of 8.1 between *dura* and *pisifera* for the oil palm 768 family and a value of 3.5 from Superbulk.

(a) 768 analysis
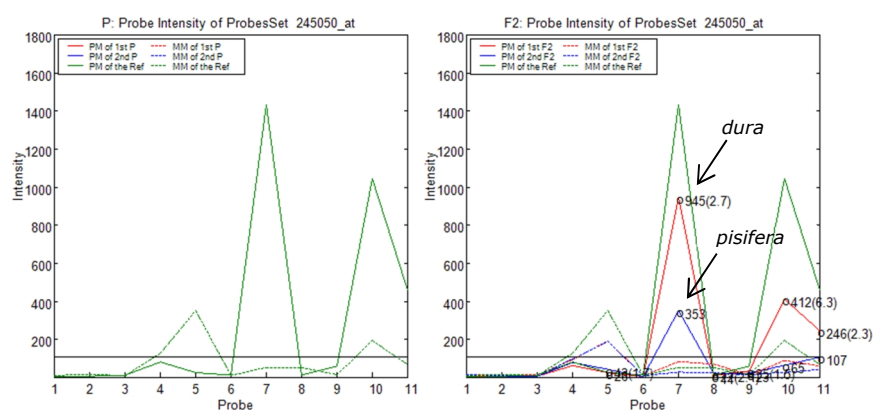
(b) 769 analysis
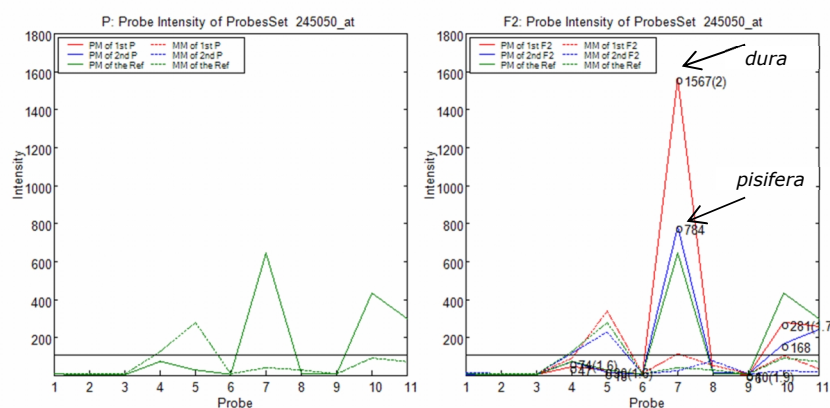
(c) Superbulk
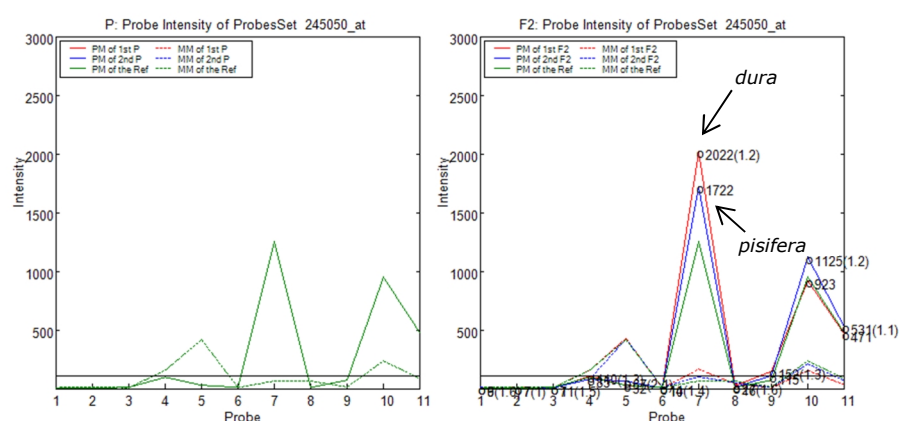
Figure 3.7   Analysis of probe-set 245050_at from the Affymetrix *Arabidopsis* ATH1 GeneChip in (a) 768, (b) 769 and (c) Superbulk using PIGEONS at a threshold of 100. The left panel gives the parental *tenera* DNA and the right-hand panel gives the bulk analysis. Red: *dura*; Blue: *pisifera*; Green: *tenera.*

3.3.5 Primer design and selection

Appendix 4 shows a list of primers from all four approaches described in Section 3.2.8. Most of the primers that were designed directly from Affymetrix array probe sequences using the first approach (as might be expected) failed to amplify oil palm DNA of the expected size. A total of 24 primers, seven from the *Arabidopsis* Affymetrix GeneChip and 17 from the rice Affymetrix GeneChip, were designed.   Of seven primers designed from the *Arabidopsis* Affymetrix GeneChip, only Af_2 and Af_3 primers, both with annealing temperature of 50°C, generate bright PCR bands with the expected size, 99 bp and 260 bp respectively (Figure 3.8).   For primers designed from the rice Affymetrix GeneChip, three primers Os_4, Os_6 and Os_11 out of 17 amplified clear PCR bands with sizes of 700 bp, 800 bp and 280 bp (result not shown).   Optimum annealing temperature for primers were chosen based on the results of the annealing temperature gradient.   For the rest of the primers, either no amplification was observed or primer-dimers and non-specific products were obtained.   For example, Os_1 failed to amplify PCR products whereas multiple bands (non-specific products) were observed when oil palm DNA was amplified using Os_3 at 50°C annealing temperature (result not shown).
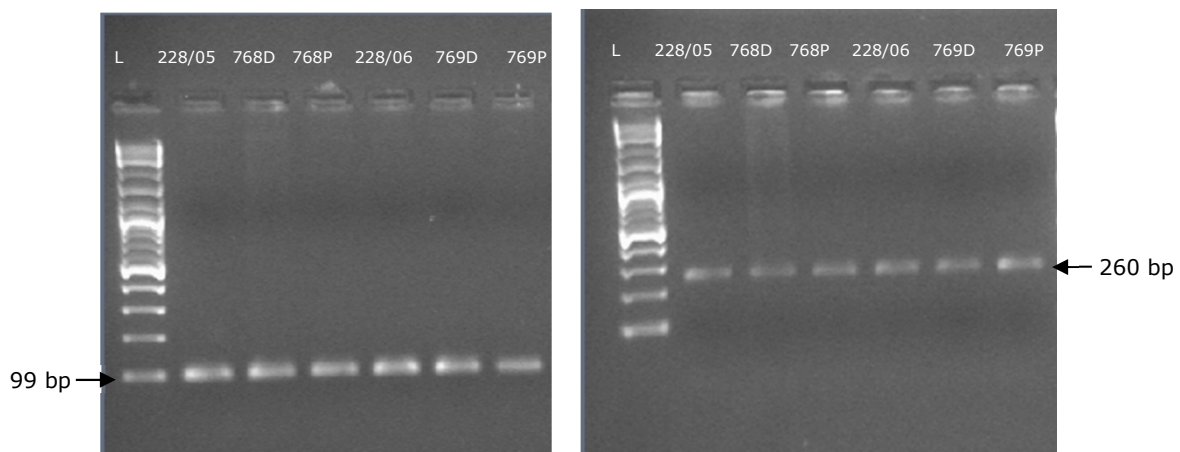
Figure 3.8   Analysis of PCR products from six oil palm DNA samples amplified using primer pairs Af_2 (left) and primer Af_3 (right) on agarose gel. Lane L: New England Biolabs 2-log DNA Ladder (0.1-10.0 kb). D: *dura*; P: *pisifera.*

Using the second approach, only probe-sets identified from the Affymetrix *Arabidopsis* ATH1 GeneChip were used to design primers using the Primer 3 software. Similar to primers designed from the Affymetrix array probe sequences directly, although the parameters were optimised using Primer 3 software, most of the primers failed to amplify a correctly sized product. There were eight primers and only Pr_5 successfully amplified oil palm DNA with the expected band size of 237 bp (Figure 3.9).
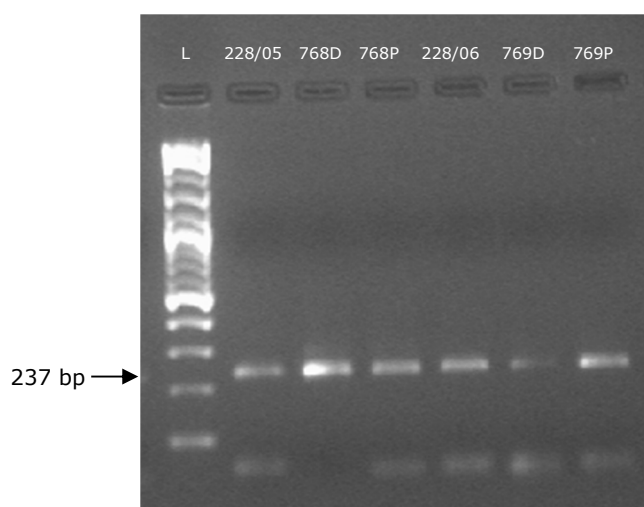
Figure 3.9 Analysis of PCR products from six oil palm DNA samples amplified using primer pair Pr_5 on agarose gel. Lane L: New England Biolabs 2-log DNA Ladder (0.1-10.0 kb). D: *dura*; P: *pisifera.*

Unfortunately primers designed using the third approach, based on protein sequences, were also unable to amplify oil palm DNA successfully. Degenerate primers were generated from the surrounding sequences after aligning target design sequences against monocot plant species. Of eight primers, Tbx_4, Tbx_5 and Tbx_6 were amplified at the annealing temperature of 40°C, 55°C and 40°C, respectively, giving multiple bands on agarose gel (results not shown). Touchdown PCR, a PCR technique that involves an initially high annealing temperature which reducing as cycles proceed, was utilised to minimise the production of non-specific products. However, the results were

negative, indicating that the third approach was not successful as degenerate primers did not amplify oil palm DNA specifically.

In order to confirm the failure of PCR amplification which is most likely to be the problem of distantly related species rather than technical error, *Arabidopsis* and rice DNA were used in PCR amplification with primers Af_1-Af_7; Pr_1-Pr_8 and Os_1-Os_17, respectively. Figures 3.10 and 3.11 showed that primer design was good and most of the products with expected sizes were amplified. However, there are some PCR products showing larger sizes than expected after amplification; for example, products amplified using Pr_7 are expected to have a band of 240 bp from the design sequence but the actual size seen on the gel is 430 bp, implying the presence of introns in the products. Optimisation of the PCR process is necessary for those that failed to amplify, such as Os_5. After sending the PCR products for Sanger sequencing to confirm the nature of the products, it was discovered that all the products appear to have homologous sequences that matched the target probe-pairs from the Affymetrix *Arabidopsis* and rice GeneChips, confirming that the primers designed from *Arabidopsis* and rice are able to amplify *Arabidopsis* and rice DNA accurately but not the heterologous oil palm DNA.
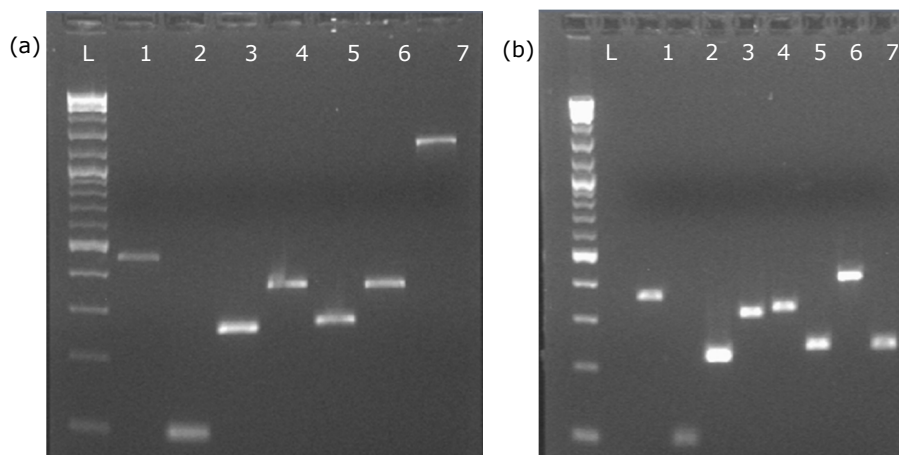
Figure 3.10   Analysis of PCR products from *Arabidopsis* DNA samples amplified using primer pairs Af_1-Af_7 (a) and primer pairs Pr_1-Pr_8 (b) on agarose gel. Lane L: New England Biolabs 2-log DNA Ladder (0.1-10.0 kb).
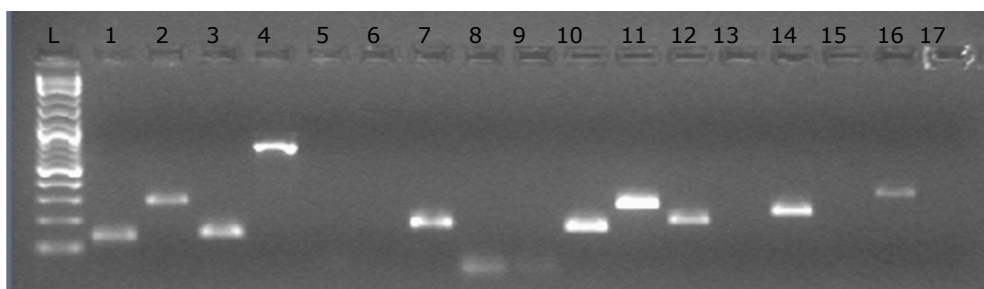
Figure 3.11 Analysis of PCR products from rice DNA samples amplified using primer pairs Os_1-Os_17 on an agarose gel. Lane L: New England Biolabs 2-log DNA Ladder (0.1-10.0 kb).

The fourth approach of designing primers involved the use of oil palm transcriptome sequences that were generated via 454 next generation sequencing technology using RNA samples derived from oil palm mesocarp. The candidate probe-pairs filtered using PIGEONS on the GeneChips for *Arabidopsis* and rice (two categories of signal intensity) were overlaid onto the 454 transcriptome. Subsequently, primers were designed directly from the isotigs to amplify the region of potential polymorphism to test if the differences observed *in silico* are genuine differences. Three groups of primers namely OP_AT_1-OP_AT_10 (*Arabidopsis*), OP_OS_1-OP_OS_9 (rice; high signal intensity) and OS_L_1b-OS-L_14b (rice; low signal intensity) were generated. Table 3.6 shows the performance of these primer sets when oil palm DNA is subjected to PCR amplification and also their function annotation based on overall homology.

Table 3.6 The summary of primers designed from oil palm isotigs and their behaviour in PCR amplification after overlaying candidate probe-pairs derived from GeneChips, *Arabidopsis* and rice onto the oil palm 454 transcriptome.

	Primer Name	Probe set	Descriptions	Product size	*PCR amplification
1	OP_AT_1	245050_at	Photosystem II protein K, Chloroplast	438 bp	√
2	OP_AT_2	245024_at	ATPase alpha subunit, Chloroplast	500 bp	√
3	OP_AT_3	245001_at	Photosystem II protein M, Chloroplast; NADH dehydrogenase subunit 7	434 bp	√
4	OP_AT_4	245002_at	Photosystem II protein D2, Chloroplast	383 bp	√
5	OP_AT_5	265228_s_at	Nucleotide binding;ATP binding;poly(U) RNA binding;zinc ion binding	363 bp	X
6	OP_AT_6	252041_at	NRPB11; DNA binding / DNA-directed RNA polymerase	350 bp	X
7	OP_AT_7	265090_at	Calcium ion binding	207 bp	√
8	OP_AT_8	258484_at	STE1 (STEROL 1); C-5 sterol desaturase;STE1 Fatty acid biosynthetic process;steroid biosynthetic process	388 bp	√+
9	OP_AT_9	245270_at	TUA6; structural constituent of cytoskeleton; microtubule cytoskeleton organisation	447 bp	√
10	OP_AT_10	256293_at	AGO7 (ARGONAUTE7); nucleic acid binding;AGO7 Vegetative phase change;production of ta-siRNAs involved in RNA interference	369 bp	√+
11	OP_OS_1	OsAffx.32330.1.S1_x_at	Cytochrome b6/f complex subunit IV, Chloroplast	371 bp	√+
12	OP_OS_2	Os.38100.1.S1_at	Cellular component organisation; actin cytoskeleton organisation AFH1 (FORMIN HOMOLOGY 1); actin binding / actin filament binding / protein binding	419 bp	√
13	OP_OS_3	Os.23127.1.S1_s_at	S-adenosylmethionine:2-demethylmenaquinone methyltransferase-like	409 bp	√
14	OP_OS_4	OsAffx.32237.1.A1_at	NADH dehydrogenase ND4L, chloroplast	353 bp	√
15	OP_OS_5	Os.28037.1.A1_at	-	386 bp	X
16	OP_OS_6	OsAffx.32279.1.S1_at	NADH-ubiquinone oxidoreductase chain 4, putative	438 bp	√
17	OP_OS_7	Os.57569.1.S1_at	rRNA processing;ribosome biogenesis;nucleotide binding;helicase activity	430 bp	X
18	OP_OS_8	Os.12924.1.S1_s_at	Putative clathrin coat assembly protein AP17; Protein transporter activity	381 bp	X
19	OP_OS_9	Os.33607.2.S1_x_at	Translation;aminoacyl-tRNA hydrolase activity	406 bp	X

*PCR amplification: √ means primers amplify DNA successfully; √+ means primers are working well but further testing is ongoing; X indicates no amplification.

Table 3.6 (cont.) The summary of primers designed from oil palm isotigs and their behaviour in PCR amplification after overlaying candidate probe-pairs derived from GeneChips, *Arabidopsis* and rice onto the oil palm 454 transcriptome.

	Primer Name	Probe set	Description	Product size	PCR amplification
20	OS_L_1b	OsAffx.13276.1.S1_at	hydrolase activity; Protein of unknown function DUF620 family protein	317 bp	√
21	OS_L_3b	Os.9523.1.S1_at	transcription; regulation of transcription, DNA-dependent; sequence-specific DNA binding transcription factor activity; nucleus	350 bp	√
22	OS_L_4b	Os.49922.1.S1_at	-	354 bp	X
23	OS_L_5b	Os.51235.1.S1_at	Plastid	351 bp	X
24	OS_L_6b	OsAffx.18742.1.S1_at	DEFL32 - Defensin and Defensin-like DEFL family; hypothetical protein	440 bp	√
25	OS_L_9b	Os.54523.1.S1_at	-	358 bp	√
26	OS_L_12b	Os.53248.1.A1_at	coiled-coil domain-containing protein 12, putative, expressed similar to *Arabidopsis* TAIR8: At3g05070.1, contains InterPro domainmRNA splicing factor, Cwf18 family protein	371 bp	√
27	OS_L_13b	Os.12010.1.S1_x_at	ATP biosynthetic process; cation transport; ATP binding; ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism plastid; membrane; integral to membrane	449 bp	X
28	OS_L_14b	Os.54503.1.A1_at	expressed protein	430 bp	X

*PCR amplification: V means primers amplify DNA successfully; V+ means primers are working well but further testing is ongoing; X indicates no amplification.
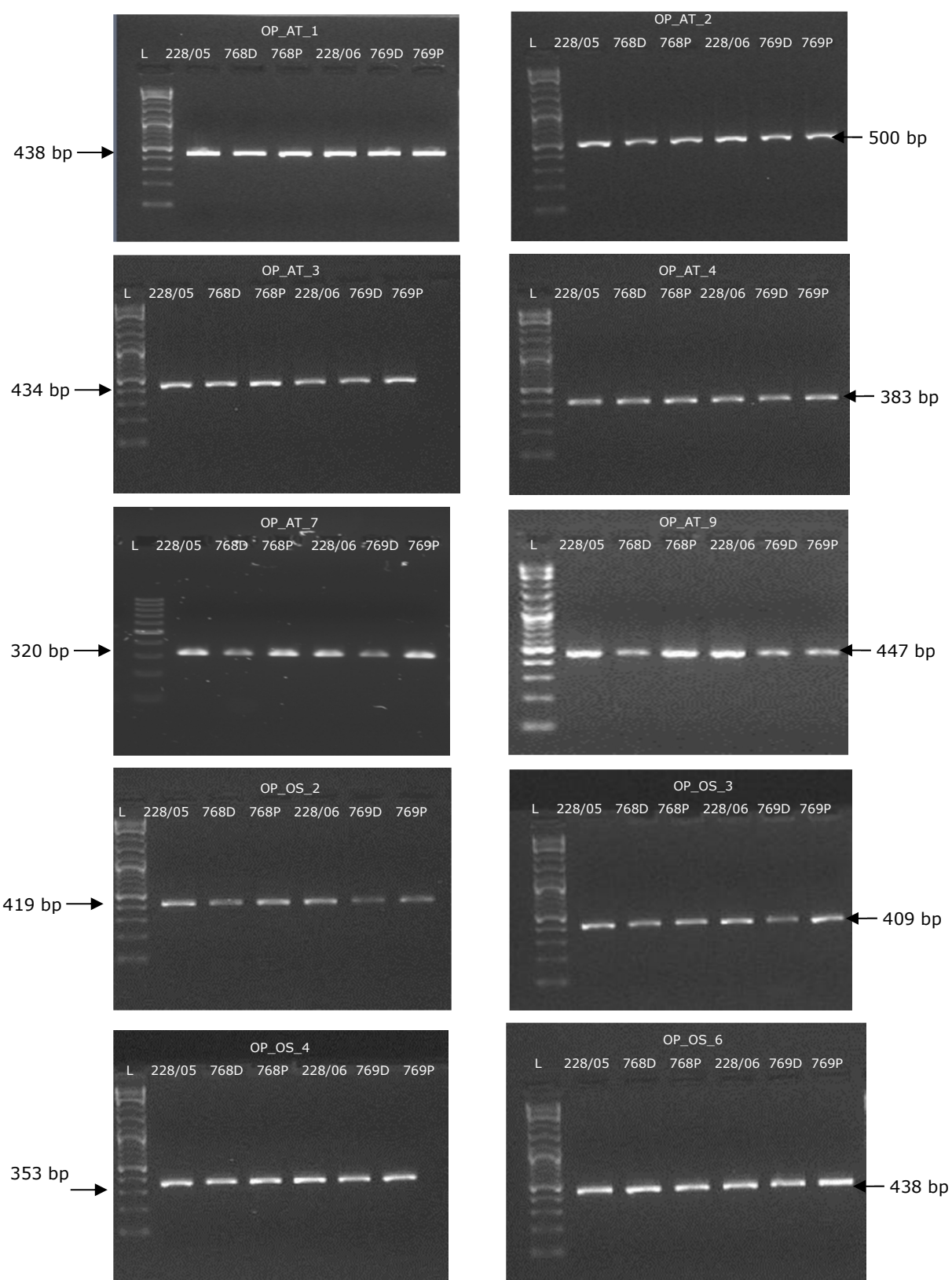
Figure 3.12   Gel image of PCR products generated from oil palm DNA samples amplified using primer pairs OP_AT and OP_OS designed from oil palm transcriptome gene models. Lane L: New England Biolabs 2-log DNA Ladder (0.1-10.0 kb). D: *dura*; P: *pisifera.*
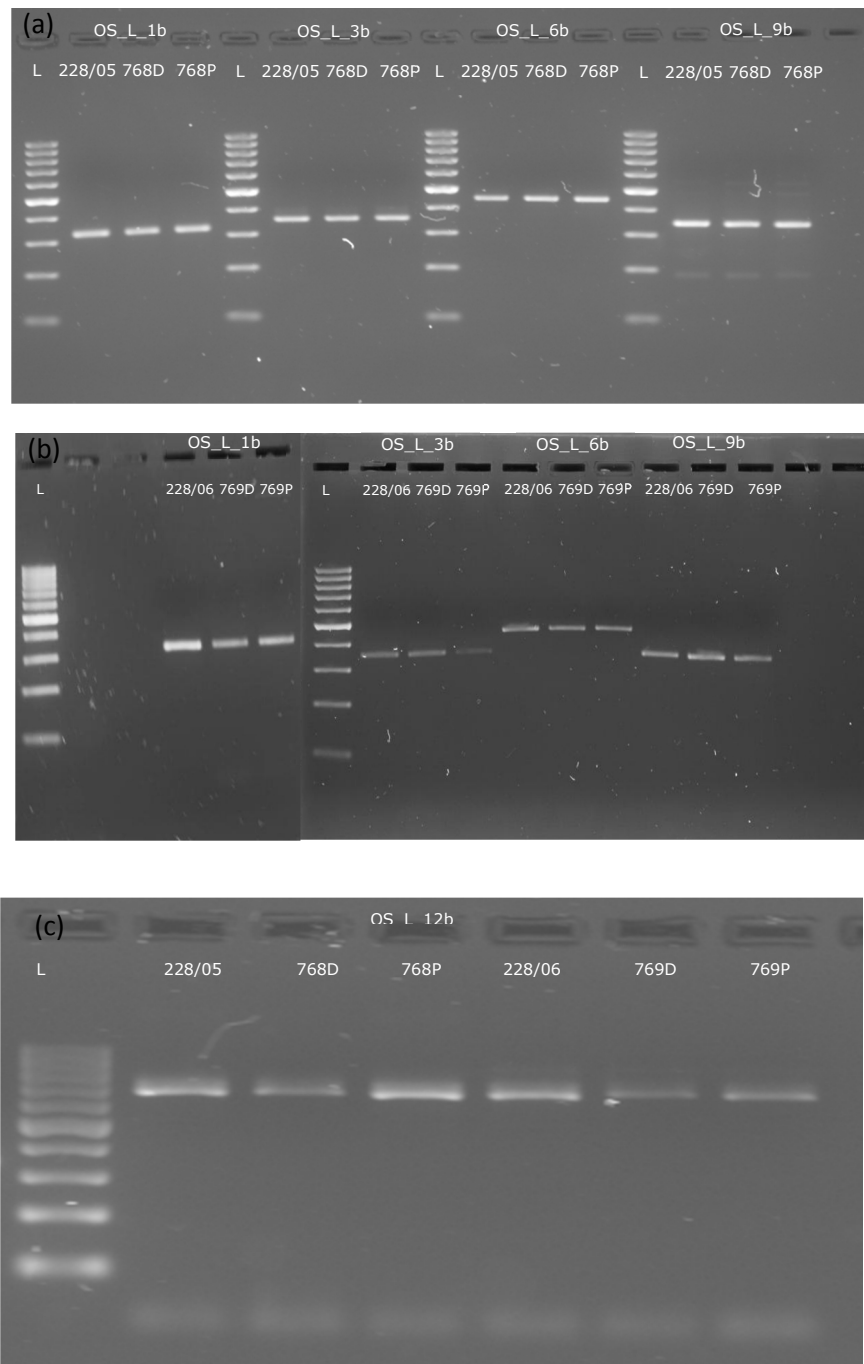
Figure 3.13 Gel image of PCR products from oil palm DNA samples amplified using OS_L primers designed from oil palm trancriptome gene models. Primers pairs OS_L_1b, OS_L_3b, OS_L_6b and OS_L9b were tested on (a) oil palm 768 family (b) oil palm 769 family while (c) primer pairs OS_L_12b tested on both oil palm 768 and 769 family. Lane L: New England Biolabs 2-log DNA Ladder (0.1-10.0 kb). D: *dura*; P: *pisifera.*

This batch of primers worked better for PCR amplification, in most cases single bands were generated from each set of primers. In total 15 sets of primers out of 28 amplified the six different oil palm DNA bulk samples with the expected band size (Figures 3.12 and Figure 3.13). OPAT 7 had a slightly larger band size (320 bp) than expected, which was 207 bp. For those that did not amplify, redesign of primers was required.

Following the PCR amplication, the purification of PCR products was conducted. As 1 ng μl^{-1} per 100 bp PCR product was required, quantitation of purified PCR products was done before sending for sequencing. Nearly 95% of PCR products were recovered using a commercial purification kit and some examples are given in Figure 3.14.
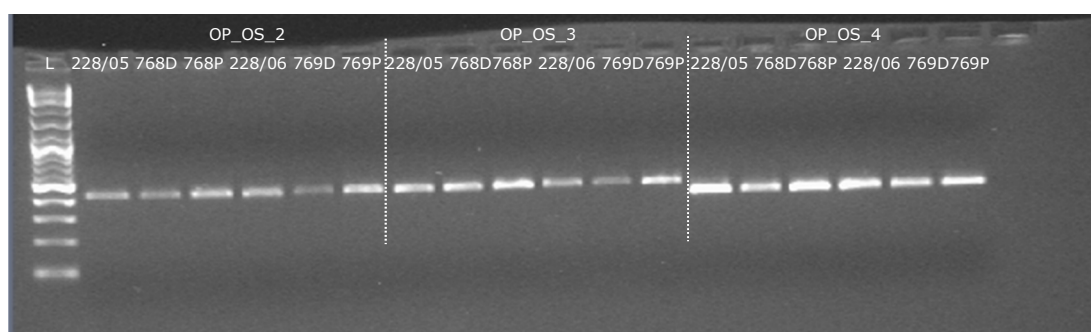
Figure 3.14 Gel image of purified PCR products derived from six oil palm DNA samples amplified using three sets of primers, OP_OS_2, OP_OS_3 and OP_OS_4. Lane L: New England Biolabs 2-log DNA Ladder (0.1-10.0 kb). D: *dura*; P: *pisifera*.
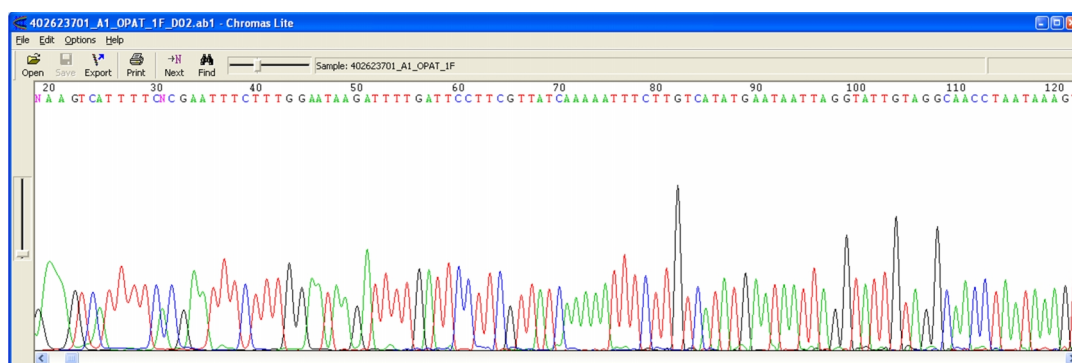
Figure 3.15 DNA sequencing trace of oil palm genomic DNA 228/05 amplified using primers OP_AT_1.

Based on the results obtained from Sanger Sequencing, all the primers generated monomorphic products across all the samples. Most of the sequencing results give a relatively good sequencing signal, for example oil palm 228/05 amplified using primer OP_AT_1 (Figure 3.15). Gene sequences from Sanger sequencing, generated from the *Arabidopsis* and rice GeneChip with high signal intensity, were BLAST searched. The putative functions, as stated in Table 3.6, revealed that most of the target genes are chloroplast-related genes or belong to structural and regulatory gene families. They are all highly conserved between and/or within species, thus no differences in sequences among the six oil palm DNA genotypes were detected.

The sequences were also aligned against each other through ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/) to confirm the consistency of the product mplified by the primers. One example is shown in Figure 3.16 with the sequences generated from six oil palm DNA samples amplified using primer OP_AT_1 being nearly similar to each other. The base highlighted in blue probably resulted from poor sequence at the end of chromatogram, rather than a genuine difference among the samples.

For sequences that were amplified using the OS_L primers, which are generated from the rice GeneChip feature which cross hybridised with lower signal intensity, the putative functions of target genes are mainly related to transcription, structural regulation, transport, enzymes and some domain-containing proteins (Table 3.6). Gene sequences were aligned using ClustalW as well, however there was no significant difference observed in bases scores across all six samples. Although mixed bases can be seen from the chromatogram, there are no really consistent differences between samples with different shell types.

```
228/05   -NNNNNNNNNNTTTTTGAANAAGTCATTTTTCNCGAATTTCTTTGGAATAAGATTTTGATT 59
768D     NNNNNNNNNNTTTTTTGAANAAGTCATTTTTCNCGAATTTCTTTGGAATAAGATTTTGATT 60
768P     --NNNNANNNNTTTTTGAANAAGTCATTTTTCNCGAATTTCTTTGGAATAAGATTTTGATT 58
228/06   NNNNNNNNNNGNTTTTTGAANAAGTCATTTTTCNCGAATTTCTTTGGAATAAGATTTTGATT 60
769D     --NNNNNNNNNTTTTTGAANAAGTCATTTTTCNCGAATTTCTTTGGAATAAGATTTTGATT 58
769P     NNNNNNNNNNTTTTTTGAANAAGTCATTTTTCNCGAATTTCTTTGGAATAAGATTTTGATT 60
         ****  **..******************************************************

228/05   CCTTCGTTATCAAAAATTTCTTGTCATATGAATAATTAGGTATTGTAGGCAACCTAATAA 119
768D     CCTTCGTTATCAAAAATTTCTTGTCATATGAATAATTAGGTATTGTAGGCAACCTAATAA 120
768P     CCTTCGTTATCAAAAATTTCTTGTCATATGAATAATTAGGTATTGTAGGCAACCTAATAA 118
228/06   CCTTCGTTATCAAAAATTTCTTGTCATATGAATAATTAGGTATTGTAGGCAACCTAATAA 120
769D     CCTTCGTTATCAAAAATTTCTTGTCATATGAATAATTAGGTATTGTAGGCAACCTAATAA 118
769P     CCTTCGTTATCAAAAATTTCTTGTCATATGAATAATTAGGTATTGTAGGCAACCTAATAA 120
         ************************************************************

228/05   AGTCTTTGCTCACTGTAAGGTCAGAACGAGGAAATAAGTTGATCAAAATTCATCGCCGTG 179
768D     AGTCTTTGCTCACTGTAAGGTCAGAACGAGGAAATAAGTTGATCAAAATTCATCGCCGTG 180
768P     AGTCTTTGCTCACTGTAAGGTCAGAACGAGGAAATAAGTTGATCAAAATTCATCGCCGTG 178
228/06   AGTCTTTGCTCACTGTAAGGTCAGAACGAGGAAATAAGTTGATCAAAATTCATCGCCGTG 180
769D     AGTCTTTGCTCACTGTAAGGTCAGAACGAGGAAATAAGTTGATCAAAATTCATCGCCGTG 178
769P     AGTCTTTGCTCACTGTAAGGTCAGAACGAGGAAATAAGTTGATCAAAATTCATCGCCGTG 180
         ************************************************************

228/05   GTTATTCAATATACAAGAATTTCGATTTTTGAATCGAGGGTTCATAATGTAAGACTTATC 239
768D     GTTATTCAATATACAAGAATTTCGATTTTTGAATCGAGGGTTCATAATGTAAGACTTATC 240
768P     GTTATTCAATATACAAGAATTTCGATTTTTGAATCGAGGGTTCATAATGTAAGACTTATC 238
228/06   GTTATTCAATATACAAGAATTTCGATTTTTGAATCGAGGGTTCATAATGTAAGACTTATC 240
769D     GTTATTCAATATACAAGAATTTCGATTTTTGAATCGAGGGTTCATAATGTAAGACTTATC 238
769P     GTTATTCAATATACAAGAATTTCGATTTTTGAATCGAGGGTTCATAATGTAAGACTTATC 240
         ************************************************************

228/05   TGGTCTTATCAATTTTTCGAATTTTGATTTATCGAATAAATCATGAATTTAGCAGAGTAT 299
768D     TGGTCTTATCAATTTTTCGAATTTTGATTTATCGAATAAATCATGAATTTAGCAGAGTAT 300
768P     TGGTCTTATCAATTTTTCGAATTTTGATTTATCGAATAAATCATGAATTTAGCAGAGTAT 298
228/06   TGGTCTTATCAATTTTTCGAATTTTGATTTATCGAATAAATCATGAATTTAGCAGAGTAT 300
769D     TGGTCTTATCAATTTTTCGAATTTTGATTTATCGAATAAATCATGAATTTAGCAGAGTAT 298
769P     TGGTCTTATCAATTTTTCGAATTTTGATTTATCGAATAAATCATGAATTTAGCAGAGTAT 300
         ************************************************************

228/05   TAAATCATCGAAAACTTACAGCAGCTTGCCAAACAAAGGCTAAGAGAAAAAAAAGTACAG 359
768D     TAAATCATCGAAAACTTACAGCAGCTTGCCAAACAAAGGCTAAGAGAAAAAAAAGTACAG 360
768P     TAAATCATCGAAAACTTACAGCAGCTTGCCAAACAAAGGCTAAGAGAAAAAAAAGTACAG 358
228/06   TAAATCATCGAAAACTTACAGCAGCTTGCCAAACAAAGGCTAAGAGAAAAAAAAGTACAG 360
769D     TAAATCATCGAAAACTTACAGCAGCTTGCCAAACAAAGGCTAAGAGAAAAAAAAGTACAG 358
769P     TAAATCATCGAAAACTTACAGCAGCTTGCCAAACAAAGGCTAAGAGAAAAAAAAGTACAG 360
         ************************************************************

228/05   GTATGACAGGCATAACATCTACGATTGGATTTAAAAANGG-CATAAAA 406
768D     GTATGACAGGCATAACATCTACGATTGGATTTAAAAAGGG-CATAAA- 406
768P     GTATGACAGGCATAACATCTACGATTGGATTTAAAAANGGGCATAAA- 405
228/06   GTATGACAGGCATAACATCTACGATTGGATTTAAAAAAGG-CATAAA- 406
769D     GTATGACAGGCATAACATCTACGATTGGATTTAAAAAGGG-CATAAA- 404
769P     GTATGACAGGCATAACATCTACGATTGGATTTAAAAAGGG-CATAAA- 406
         ********************************** ** ******
```

Figure 3.16 Alignment of sequences that were generated from six oil palm DNA samples amplified

using the OP_AT_1 primer pairs analysed with ClustalW.

3.3.6 Transcriptome profiling

An initial analysis of the oil palm transcriptome using the CLC Genomics Workbench software generate a total number of 1,087,824 of reads with an average length of 358.25 and total base number of 389, 715, 590 (Table 3.7). A total of 989,298 reads were incorporated into the contig assembly, accounting for 90.94% of the total reads and 92.24% of total bases with an average read length of 363.38. The fragment size of matched read length varies from 50 bp to 550 bp, with the highest number of reads falling between 440 bp and 550 bp (Figure 3.17). After the alignment of reads, 46,770 contigs were produced with an average length of 536 bp whereas the N50 was 554 bp.

Table 3.7 Summary of oil palm transcriptome analysis using CLC Genomics Workbench after 454 pyrosequencing.

	Count	Average length	Total bases	Count %	Bases %
Reads	1,087,824	358.25	389,715,590		
Matched	989,298	363.38	359,489,817	90.94	92.24
Not matched	98,526	306.78	30,225,773	9.06	7.76
Contigs	46,770	536	25,080,424		

Quality measurement	
N75	445
N50	554
N25	826
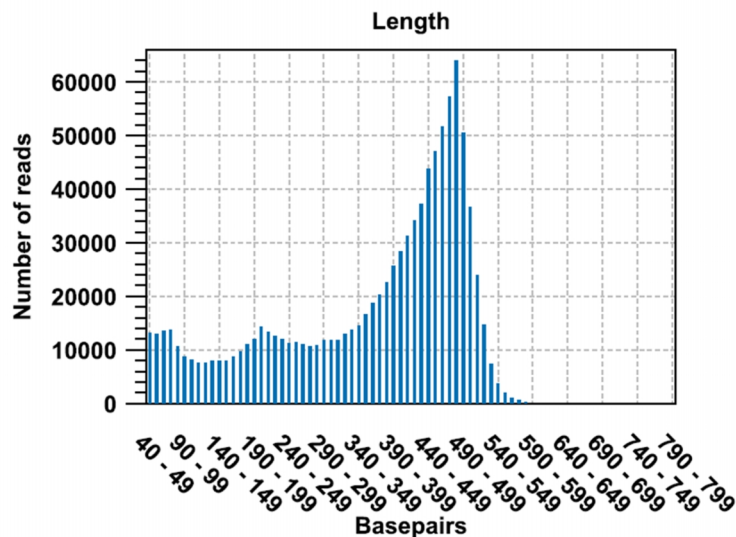
Figure 3.17 The fragment size of matched reads (bp) in relative to number of reads in oil palm.

All of the 454 contigs were overlaid on a partial reference genome generated from date palm, which is also a member of *Aracaceae* family. Table 3.8 showed that less than 50% of contigs have similarity with date palm. Only 20,842 out of 46,770 of total contigs from oil palm match with the total contig number of 57,277 from the reference genome. The read length of matched contigs ranges from 50bp to 1500 bp and more than 3,000 matched reads are of 500 bp to 550 bp (Figure 3.18).

Table 3.8 The assembly data obtained from assembled oil palm transcriptome overlaid on the date palm genome sequence.

	Count	Average length	Total bases	%
Contigs	46,770	536.25	25,080,424	
Matched	20,842	482.51	10,056,540	44.56
Not matched	25,928	579.45	15,023,884	55.44
References	57,277	6,661	381,563,256	
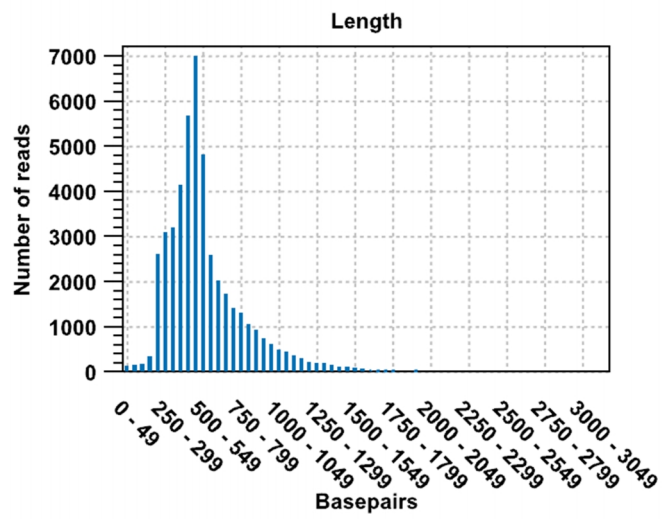
Figure 3.18 The fragment size of oil palm matched reads against the date palm reference genome.

3.4 DISCUSSION

3.4.1 Examination of the segregating population

Ways of using resources developed in model plants, such as *Arabidopsis* and rice, for research in oil palm was investigated in order to develop potential molecular markers that are linked to gene(s) controlling shell thickness, as an example. The oil palm families, 768, 769, 751 and 869, were first examined for the presence of illegitimate samples prior to XSpecies analysis. The assessment of segregation in a population derived from a controlled cross is crucial as the traits to be studied need to be polymorphic between two parental lines and heritable across all the progenies. DNA fingerprinting is always carried out to screen the populations prior to experimental analysis to ensure that the palm identity is correct and this is achieved through the use of molecular markers (Mayes *et al.,* 2008). SSR markers are among the most commonly used molecular markers and they are publicly available in a range of plant species. They can reveal genetic relationships and ensure effective quality control in plants such as rice (Chakravarthi and Naraveneni, 2006) and oil palm (Billotte *et al.,* 2001). SSR markers differentiate cultivars based on differences in the length of the SSR repeat units present in particular alleles. SSR is a co-dominant marker system, thus it is chosen in many cases for fingerprinting (Billotte *et al.,* 2001) and also provides information on heterozygosity and/or inbreeding. SSRs can be detected by PCR, using two flanking primers designed from genomic or coding sequences containing SSR repeats. After screening the oil palm population, those individual palms that were suspected to have resulted from out-breeding or mis-labelling, namely; 768/28(D)-D18, 769/A/36(D)–D109, 751/48(P)–P25, were excluded from the population in order to ensure that all individuals are segregating from a single controlled cross and hence the polymorphism detected among the population in subsequent analysis is genuine.

Subsequently, an approach known as bulked segregant analysis (BSA) was used to construct the DNA pools that should differ for the trait of interest, for instance, shell thickness in this study. This approach was first developed by Michelmore *et al.* (1991) who screened the bulks with random amplified polymorphic DNAs (RAPDs) in a screen for disease resistance gene loci. BSA allows rapid identification of markers linked to a trait controlling gene by comparing two bulked DNA samples derived from individuals generated from a segregating population of a single controlled cross. Each bulk contains individuals that have similar phenotypes for a particular trait and random genotypes for genes or loci that are not linked with the target gene, if the control of the trait is mono (or oligogenic). In this case, markers showing polymorphism between *dura* and *pisifera* bulks (constructed for 'thick' and 'no shell' types in oil palm) should be genetically linked to the locus that determines shell thickness, the trait used to construct the pools. The process of genotyping the plants using a microarray approach for contrasting phenotype bulked samples is reduced to only two samples, *dura* and *pisifera*, instead of screening of all the individuals, XSpecies analysis thus potentially becomes relatively simpler and cheaper.

3.4.2 Selection of potential probe-sets

PIGEONS software was used to select the potential probe-sets and probe-pairs that might be used as molecular marker for single nucleotide polymorphism (SNPs) detection. PIGEONS (Lai et al., 2014) was developed to overcome the issue of human-dependant cut-off selection of poorly hybridising oligonucleotide probes within a probe-set through genomic DNA-based approach that is driven from previous script parser (Hammond *et al.,* 2005; Davey *et al,* 2009). Cut-off selection, based on threshold boundaries, is important as it gives an idea of which threshold should be chosen for the analysis, as well as the number of probe-pairs, probe-sets and ratio of average probe-sets to probe-pairs in order to generate a feasible result.

By using a heuristic method known as Automated Threshold Mapping (ATM), PIGEONS software is able to provide three types of cut-off choices: suggested threshold value, target interval and tolerance interval (Lai *et al*, 2014). Suggested threshold values can directly be taken as a cut-off point to remove the poorly hybridising oligonucleotides while any values in the target interval could serve as potential cut-offs in which more probe-sets and probe-pairs could be retained. Those values that fall in the tolerance interval are still considered as a feasible threshold value, however, probes with signal intensity that falls outside the tolerance interval are considered as poorly hybridising oligonucleotides and should be excluded (Lai *et al.,* 2014). As the threshold is increased from 0 to 1000, probe-pairs retention decreases rapidly although entire probe-sets which represent transcripts are lost relatively slowly as only a minimum of one PM probe is required to retain a probe-set (Hammond *et al.,* 2005; Davey *et al.,* 2009). As a result there will be a significant loss of probe-sets as well if the threshold value chosen is beyond the upper limit.

Lai *et al.* (2014) also recommended the use of the Fukuyama-Sugeno Index when the ATM approach is established, as it can improve the fuzzy boundaries as well as serve as the best approach for studies where no particular interest in expressed genes is required, for instance, when seeking for SNPs markers between two samples.

Dual-fold-change Analysis (DFC) in PIGEONS Mining & Image was carried out at several threshold values within the boundary area in order to identify potential probe-sets for primer design. Although the selection of threshold values no longer needs to be completely dependent on human judgement, the filtering of probe-sets and probe-pairs through DFC is still defined by the user, from a ratio of 1.5 to 5, in order to obtain as many potential probe-sets as possible. Dual-fold-change is defined as the ratio of signals from differentially hybridised oligonucleotides between two samples. The number of potential probe-sets and probe-pairs decreases simultaneously as the threshold value as

well as fold-change value increases. The higher the threshold value or fold-change difference between samples, the smaller the number of probe-sets obtained, as more poor probes which fail to achieve the defined value will be removed and the increased fold-change stringency removes less differentially hybridised probe sets. Thus several threshold values within the boundary, combined with different fold-change values, were tried in order to minimise the loss of any potentially informative probe-sets.

The XSpecies analysis used two cross-species high density microarrays; the *Arabidopsis* and rice Affymetrix GeneChips. As discussed earlier, the retention rate of probe-sets and probe-pairs after cross-hybridising onto the rice GeneChip (92.13%; 31.91%) is higher than in *Arabidopsis* (87.11%; 19.87%) at threshold value of 110, suggesting a taxonomically closer relationship between oil palm and rice. Oil palm and *Arabidopsis* are believed to have diverged between 145-208 Mya (Sanderson *et al.,* 2004) while oil palm and rice split at the level of clade *Commelinids* in the early Cretaceous, in the range of 91-99 Mya (Wikstrom *et al.,* 2001), confirming that oil palm is taxonomically closer to rice rather than to *Arabidopsis.* Lai *et al.* (2014) also pointed out the importance of using closely related species for cross-species hybridisation as only a few SNPs were identified in the study after cross-hybridising Bambara groundnut onto the *Arabidopsis* GeneChip.

3.4.3 Potential markers for the oil palm shell thickness locus

Several approaches were used to design primers from candidate probe-sets and probe-pairs for validation purposes. The first three approaches, as reported earlier, encountered poor success. The poor PCR amplification in the first approach could be the result of designing primer sequences based purely on the probe-pairs sequences that flank target probe-pairs showing differential hybridisation directly in each probe-set. These are derived from heterologous sequence from a model species. *Arabidopsis* is a dicotyledonous plant species,

as a result of sequence divergence there is a higher chance that target sequences observed in model plant no longer reliably detect an orthologue in oil palm. Similar PCR amplification results were obtained in the second approach using primers designed using the Primer 3 software, based on the original model species design sequence. For the third approach in which primers were designed based on protein sequences, it is suspected that the introduction of ambiguity into the primers has reduced the efficiency of binding and subsequent amplification.

The large scale changes in the genome over time since the divergence from the common ancestor of the pair of species could also be a factor causing problems during amplification of DNA using primers designed from the heterologous species. *Arabidopsis* is reported to have a relatively small genome size of 120 Mbp with only approximately 10% repetitive sequences whereas the rice genome is reported to be three times larger (389 Mbp) and contains at least 35% repetitive DNA (Wicker and Keller, 2007). In comparison, the oil palm repetitive DNA content was estimated to account for 75% of 1.8 Gbp oil palm genome (Singh *et al.,* 2013). Of these repeat sequences, nearly 57% show no sequence similarity to previously identified repeat elements. Repeat elements could have a number of effects on evolution of genome, for instance, recombination events that lead to genome rearrangements (Brown, 2002).

In addition, Singh *et al.* (2013) also observed a large number of segmental duplications (homologues duplicated sequences) in oil palm genome, suggesting that oil palm is a paleotetraploid which is the result of genome duplications during the evolutionary history of plants. Genome duplications can result from either duplication of the genome of a single species or the combination of the chromosome sets from different species during plant evolutions (Edger and Pires, 2009). However, most of the paleopolyploids have experienced extensive chromosome restructuring (as is the case in maize and oil palm) and often the concommitant silencing of genes. Once meiosis is

stabilised, there can be a rapid effective loss of the polyploidy status. The resultant genomes behave like diploids in meiosis, such as maize, cotton and soybean (Blanc and Wolfe, 2004). Other species, such as wheat, are allopolyploids and despite being composed of the A, B and D ancestral genome, essentially behaves as a diploid ($2n = 6x = 42$; Blanc and Wolfe, 2004). Genome evolution involves rearrangement of existing gene or acquisition of new genes by gene duplication or by polyploidisation, therefore, it is possible that primers designed from *Arabidopsis* and rice sequences may show evolutionary divergence, may detect multiple genes complicating the technical complexity of the PCR or may even cross-hybridise to repetitive sequences, as the oligonucleotide probes and most PCR oligonucleotide primers are relatively short.

The use of XSpecies approach combined with 454 next generation sequencing technology provided a test platform into the development of potential markers that are linked to gene(s) controlling traits of interest. A preliminary analysis of the oil palm transcriptome using CLC Genomics Workbench was undertaken in order to allow the alignment of candidate probe-sets and probe-pairs from XSpecies microarray against the 454 transcriptome prior to primer design. The first step of the analysis was to determine if the assembly of sequences is of good quality and ready for annotation. One of the useful statistics to examine the completeness of the genome assembly is the N50 value. N50 is calculated by first ranking the contigs according to size, followed by totalling up the lengths of each contig until the sum equals to 50% of the total length of all contigs in the assembly (Yandell and Ence, 2012). The N50 is identified as the length of the shortest contigs in this list (Yandell and Ence, 2012). The larger the N50 size is, the better the assembly is. A contig N50 of 554 bp obtained from the oil palm transcriptome analysis through 454 pyrosequencing indicated that the assembly is of reasonable quality, therefore alignment with the probe-sets and probe-pairs from the XSpecies microarray to design primers seems a reasonable approach. Although longer reads are

generated from 454 pyrosequencing (~400 bp) which are more amenable to *de novo* assembly and alignment, ideally the initial assembly should be followed by the production of shorter sequences (~35 bp) using SOLiD or Illumina platforms in order to generate a far more comprehensive depth of sequence coverage (Mardis, 2008; Kumar and Blaxter, 2010).

The putative functions of the products are compared and annotated against the databases from other plant species. Several primer pairs are reported to be able to amplify products homologous to photosystem II protein K, ATPase alpha subunits and NADH dehydrogenase that are located in the chloroplast (Table 3.6). This could explain one of the possibilities for the lack of polymorphism observed among the samples. The chloroplast haplotype is maternally inherited, highly conserved, has a copy number which can vary in different tissues and under different conditions compared with the nuclear DNA content (Palmer and Zamir, 1982). In addition, it is believed that the probe-sets and probe-pairs with high signal intensity after cross hybridising with *Arabidopsis* and rice sequences are resulted from repeating units of highly conserved genes in oil palm, such as chloroplast-related genes, and could be an indicator of non-nuclear genes. One of the examples given is probe-set 245050_at. The high hybridisation signal strength and fold-change difference between *dura* and *pisifera* observed in PIGEONS across oil palm bulks from the 768 family, 769 family and Superbulk at probe-set 245050_at could be due to the number of copies of chloroplast DNA present in the original leaf samples (Figure 3.7).

Thus, targeting genes which appear to show variation between the bulks, but have lower signal intensity was attempted with probe-sets and probe-pairs filtered using PIGEONS after cross-hybridising oil palm DNA on the rice GeneChip in order to avoid selecting multiple copy genes, such as chloroplast or mitochondrially coded sequences and target putatively genuine polymorphisms derived from single copy gene across the samples. Although fewer candidate probe-sets and probe-pairs were selected due to more stringent parameters,

probe-sets that have sequence homology to transcription regulators, hydrolase activity, defensin-like genes and domain-containing protein were successfully identified. However, lack of consistent differences between samples with different shell type after sequencing the PCR products (monomorphic products) means that this approach requires further investigation to determine why the apparent hybridisation signal differences from the Affymetrix analysis are not clearly reflected in base pair differences in the PCR products.

The XSpecies approach that has been investigated so far seem to hold some promise, with primers that are designed based on available within target species transcriptomic sequences potentially serving as putative markers for the trait of shell thickness in oil palm. The development of a more comprehensive genetic map in oil palm for this cross would be helpful to identify the distribution and location of any putative markers on the oil palm genome in relation to the known position of the shell-thickness gene itself. As the cost of sequencing becomes cheaper, further sequencing analysis using ABI SOLiD or Illumina sequencing platforms could be used to generate more complete isotigs. Furthermore, as the current transcriptome is only composed of three stages of mesocarp development, a much broader sampling of different transcriptomes would be useful, thus more primers could be designed from the oil palm transcriptome which allow polymorphism detected using an XSpecies microarray to be evaluated.

3.4.4   Challenges of the XSpecies study in oil palm

Although the bulked segregant analysis (BSA) approach is utilised, the number of oil palm plants for each bulk is relatively small (10 individual plants from one fruit type in each family). A small bulk size has been suggested to introduce difficulties to determine if the polymorphisms observed between bulks are genetically linked to the gene or loci that control the trait used to construct the pools. Quarrie *et al*. (1999) suggested the use of DNA from at least 50

individuals to construct the bulks for improving drought resistance in maize when using codominant markers (such as RFLPs, SSR markers and SNPs) for analysis in order to ensure that the allele is represented in the bulks at the same frequency as in the population. This is important to reduce the background noise in a cross-species study.

In addition, another major issue with cross hybridisation is the sequence divergence between the target species and the species that was used to design the microarray. Hybridisation efficiency is suggested to be influenced by the evolutionary relationships, with the lowest efficiency obtained in comparisons between diverged species (Buckley, 2007). Rise *et al.* (2004) reported the cross hybridisation of cDNAs from lake white fish (*Coreogonus clupeaformis)* and smelt (*Osmerus mordax*) to a 7356-feature cDNA microarray, derived from ESTs from Atlantic salmon and rainbow trout. As expected, the lowest number of features on cDNA microarray (38%) was detected in smelt, which is the most diverged species, as compared to Atlantic salmon targets (70%) whose ESTs are used to construct microarray. Inefficient hybridisation of certain transcripts to the probes on the array could result in background noise and lack of clear signal. This could affect the ability to differentiate variation observed in intensity due to differential gene expression between samples mismatches of sequences design and target sequence. In terms of the microarrays used in this study, *Arabidopsis* has been proven to be less sensitive and less efficient to detect the probe-set targets from oil palm when compared to rice. However, rice is still not an ideal reference species to cross hybridise with.

Date palm, which is also a member of *Aracaceae* family, has been sequenced recently and ~380 Mb of the sequence assembled (gene-rich region) covering and estimated 90% genes and 60% of the genome (Al-Dous *et al.,* 2011). Although the alignment of the oil palm transciptome sequences with date palm genome sequences was carried out using CLC Genomics Workbench 4th edition, less than 50% of the oil palm sequences were assembled to the

reference genome. The result suggest that the genome information derived from the oil palm mesocarp is still not sufficient and is at least partly a reflection of the limited proportion of the full complement of genes expressed in oil palm mesocarp across the three stages of development studied.

Due to reasons that have been mentioned, two species which are more closely related to oil palm than *Arabidopsis* or rice are recommended in order to further study the application of XSpecies approach in oil palm. A better pair of subject species and model/crop would be within legumes, with Bambara groundnut compared to *Medicago truncatula* as well as soybean (Schmutz *et al.,* 2010; Young *et al.,* 2011). While the sequences available for Medicago and soybean are not as comprehensive and are poorly annotated compared to both rice and Arabidopsis, the genetic distance to the target species is far smaller (54 Mya; Cannon *et al.,* 2009). In addition, the fact that Bambara groundnut has a relatively small genome size, ~882 Mb, which is approximately twice the size of rice genome (http://data.kew.org/cvalues/introduction.html) with diploid genetics ($2n = 2x = 22$) allows testing and development of molecular genetic tools through these approaches in Bambara groundnut, an important contrast to oil palm.

Chapter 4: EFFECT OF MILD DROUGHT STRESS IN BAMBARA GROUNDNUT

4.1   INTRODUCTION

4.1.1   Bambara groundnut landraces: DipC and Tiga Nicuru

The genetic resources of Bambara groundnut are widely conserved by indigenous farmers across sub-Saharan Africa. In addition, there are also approximately 2000 and 972 accessions in gene banks held by the International Institute of Tropical Agriculture (IITA) in Nigeria and Southern Africa Development Community (SADC), respectively (Massawe *et al.,* 2005). Nonetheless, Bambara groundnut germplasm has not been fully exploited yet. Most of the Bambara groundnut accessions exist in the form of landraces, which have evolved directly from their wild relatives (Massawe *et al.,* 2005). High genetic variation in Bambara groundnut provides breeders with genetic sources to improve yield, biotic and abiotic resistance and adaptability of crops to various environments.

In the present study, a segregating population generated from a narrow cross between two landraces, DipC from Botswana and Tiga Nicuru from Mali, was used to study the effect of a mild drought stress in Bambara groundnut. Botswana and Mali are semi-arid, landlocked countries in the centre of southern Africa and West Africa, respectively. Botswana has a mean annual rainfall of about 450mm ranging from 250mm in the extreme southwest to 650mm in the extreme northeast (Burgess, 2006; Kgathi *et al.,* 2012). The temperature in Botswana ranges from 12°C-15°C during the early morning, to 30°C-40°C by late afternoon in the dry season (April to October), but the maximum temperature is 25°C-30°C during the rainy season   (November to March; Burgess, 2006). Nevertheless, Botswana experiences extremely low humidity with average annual evaporation of about 2000mm (Burgess, 2006). For Mali, the annual precipitation varies across the country and can be divided into three climatic zones. With the average annual rainfall 440mm across the country, the highest

mean rainfall of between 700-1000mm can be obtained in Sudanic in the South, followed by 200-400mm rainfall in the Sahelian in the central and West and little or zero rainfall in the Saharan in the North (Pedercini *et al.,* 2012). The temperature in Mali ranges from 16°C to 39°C with 4-5 months of rainy season from April to October (Pedercini *et al.,* 2012). As a result, both Botswana and Mali face the challenges of drought and desertification as most of the areas receive limited to negligible rainfall. The climatic conditions in Botswana and Mali suggested that both DipC and Tiga Nicuru are likely to be more tolerant to drought than many landraces, but potentially with some variation between them for climatic adaptation.

DipC and Tiga Nicuru have significant differences in terms of average seed yield as well as growth habits. Differences have been observed in yield production between DipC and Tiga Nicuru with DipC producing greater *seed number* and *seed weight per plant* than Tiga Nicuru (Ahmad, 2012). In addition, Ahmad (2012) recorded DipC having different plant architecture with greater *petiole length*, *leaf area* and *plant height* than Tiga Nicuru. However, as DipC has shorter *internode length* and *peduncle length* than Tiga Nicuru, DipC is classified as bunched type while Tiga Nicuru is categorised as a semi-spreading type morphology (IPGRI, 2000; Figure 4.1).

Figure 4.1 The comparison of the DipC (left) 'bunched type' and Tiga Nicuru (right) 'semi-spreading growth habit' (Ahmad, 2012).

In a genetic diversity UPGMA analysis based on DArT markers, DipC from Botswana was found to be allocated to a different cluster to Tiga Nicuru from Mali (Figure 4.2; Standler, 2009). While the absence of branch confidence scores makes this harder to interpret, it seems likely that this is a real genetic differentiation between the two parental lines. As a result, DipC which is the maternal parent and Tiga Nicuru the paternal parent were selected for crossing in order to achieve both good drought tolerance and relatively high seed weight in a single line.
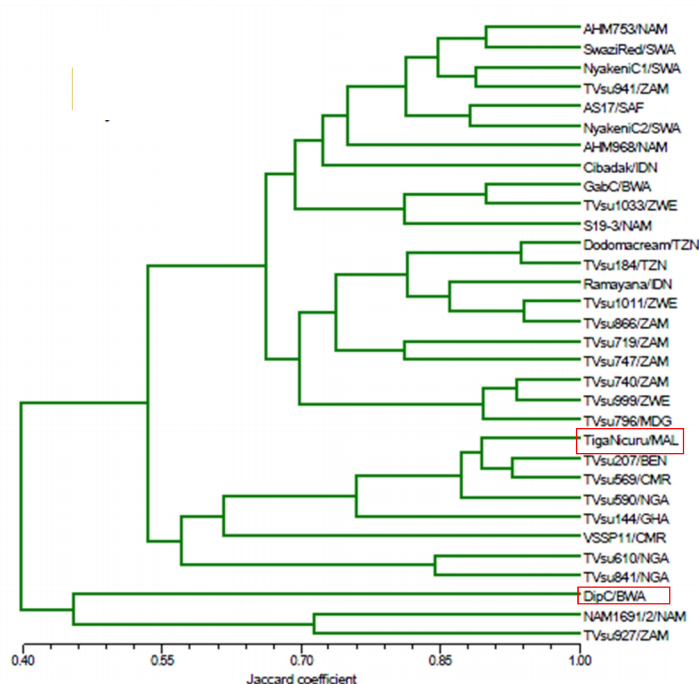
Figure 4.2   UPGMA dendrograms representing Bambara groundnut landraces collected from different regions based on the similarity matrix of DArT markers (Standler, 2009). The analysis is based on 201 DArT markers and single seed genotypes form each landrace or line.

4.1.2   Drought stress in crop plants

Abiotic stresses such as salinity, drought, temperature and flooding are the major limiting factors to plant growth and crop productivity. Drought stress, one of the most important constraints for agriculture, is defined as stress that is caused by inadequate soil moisture to meet the needs of a particular crop at a

particular time (National Drought Mitigation, 2003). Drought stress influences several plant processes and can cause a change in growth parameters, for instance, a reduction in leaf area and dry matter production in groundnut (Collino *et al.,* 2001), cowpea (Anyia and Herzog, 2004) and chickpea (Singh, 1991). In pea, germination and early seedling growth were reported to be influenced by drought (Okcu *et al.,* 2005). In wheat, a decrease of the number of grains, grain yield, shoot dry weight and harvest index were observed when wheat was subjected to drought stress (Gupta *et al.,* 2001). Drought stress can affect crop growth at any developmental stage including the vegetative stage, reproductive stage and during grain filling (Blair *et al.,* 2012). In soybean, the loss of seed yield was reported to be maximal when drought appeared during anthesis and the early reproductive stages (Liu *et al.,* 2003; Eslami *et al.,* 2010).

As water available for agriculture continues to decline worldwide, the development of drought-tolerant plants or the improvement of the tolerance level to drought in plants is important. For example, advanced lines BAT477 and SEA5 that are highly tolerant to drought have been identified in common bean (Teran and Singh, 2002; Singh *et al.,* 2001). Furthermore, Budak *et al.* (2013) also reported the introgression of wild emmer wheat (*T. dicoccoides*) which is highly tolerant to drought, into modern wheat cultivars in order to obtain drought related candidate genes for breeding purpose. The authors reported that the investigation of the morphological and physiological characteristics of developed cultivars in field trials was conducted to access their performance and their contributions to yield under drought condition.

Bambara groundnut has long been recognised as a drought-tolerant crop as it can survive and produce higher seed yield than other legume crops under drought conditions (Colinson *et al.,* 1996), although a comprehensive set of comparisons between legume species is still needed. Landrace differences in Bambara groundnut in response to drought have been reported (Berchie *et al.,* 2012; Mwale *et al.,* 2007), providing the potential to select and breed higher

yielding landraces and/or cultivars under water stress. In order to investigate the species' genetic diversity for drought tolerance, an exploration of the mechanisms underlying the response of Bambara groundnut to drought is essential.

4.1.3   Plant response to drought stress

Drought stress in crop plants triggers various responses and these can be categorised into three groups: escape, avoidance and tolerance (Turner, 1979). Drought escape is described as the capability of plants to complete their growth cycle and reach maturity before drought-stress develops (Collinson *et al.,* 1997). Drought avoidance is demonstrated by crop species which are able to maintain high water potential in the plant by minimising water loss and maximising water uptake under drought conditions, as seen in *Siratro*, the tropical legume (Ludlow, 1989) and Chickpea (Gaur *et al.,* 2008). Mechanisms of avoidance include improved root traits, for greater extraction of soil moisture, decreased *stomatal conductance*, decreased radiation absorption and decreased leaf area for minimal water loss (Harb *et al.,* 2010). Drought tolerance allows plants to survive the drought period despite stresses. Such mechanisms are seen in a range of species, including mung bean (Ocampo *et al.,* 2000) and pigeonpea (Subbarao *et al.,* 2000). Plants with drought tolerance mechanisms are able to maintain their cell turgor through osmotic adjustment, which in turn will contribute to maintaining stomatal opening, leaf expansion and photosynthesis throughout the drought period (Collinson *et al.,* 1997).

For Bambara groundnut, several studies have been carried out to investigate the response to drought stress. For instance, the change in leaf orientation, which is known as paraheliotropic movement, was observed in drought-stressed Bambara groundnut landrace AS-17 (Stadler, 2009). The author stated that in drought stressed plants leaflet angles were shown to be parallel to the incident radiation, leading to less transpiratory water loss due to

the lower leaf temperature that resulted from decreased light interception. In addition, a higher root dry weight was reported when Bambara groundnut landrace, Burkina, was subjected to drought (Berchie *et al.,* 2012). The allocation of assimilates to root growth rather than shoots would have allowed Bambara groundnut plants to exploit greater soil moisture when the plants were drought-stressed, probably through deeper root growth.

As mentioned, crop plants could have shorter life cycles in order to escape from drought stress. Bambara groundnut was shown to have a shortened vegetative growth period, earlier flowering, have a reduced reproductive stage and mature earlier in response to water stress, at the expense of yield (Mabhaudhi *et al.,* 2013). Landraces from Jozini, South Africa, such as 'Red' and 'Brown' landraces matured early (mean: 122.75 DAP, $p<0.01$) when the plants were stressed at 30% of the crop water requirement (ETa) as compared to 100% ETa (mean: 128 DAP, $p<0.01$; Mabhaudhi *et al.,* 2013). The findings are also comparable with a previous study which identified S19-3 from Namibia to have faster rates of development, resulting in a shorter phenology (mean: 110 DAS; Mwale *et al.,* 2007) under drought stress.

Stomatal closure plays an important role in regulating transpiration and hence improving plant water status over the drought stress period. Stomatal closure has been recognised as a universal response to drought stress in many species, such as rice (Huang *et al.,* 2009), maize (Benesova *et al.,* 2012) and Bambara groundnut (Collinson *et al.,* 1997; Vurayai *et al.,* 2011). The reduction of *stomatal conductance* in Bambara groundnut could reach 90% when drought stress is imposed during the pod-filling stage (Vurayai *et al.,* 2011). Drought-tolerance species regulate stomatal function to allow some carbon fixation during the drought period and hence to improve photosynthetic efficiency (Yordanov *et al.,* 2003).

In addition to stomatal regulation of water loss, Collinson *et al.* (1997) suggested that Bambara groundnut maintains plant water status over the

drought period through osmotic adjustment and reduced leaf area. Osmotic adjustment, which involves accumulation of osmolytes, has been proposed to occur either through passive movement where water is withdrawn from the cell due to drought or the active accumulation of solutes such as proline (Collinson *et al.*, 1997). Drought-induced accumulation of soluble sugars and proline has been observed in other species. For example, free proline levels in maize increased by 1.56 to 3.13 times when the plants were subjected to drought stress (Mohammadkhani and Heidari, 2008).

Furthermore, Vurayai *et al.* (2011) stated that reduced leaf area in drought-stressed Bambara groundnut plants due to turgor reduction within expanding cells is common and is one of the earliest physiological responses to water stress. The decline in leaf expansion which in turn causes decreased total leaf area has also been observed in crop species like cowpea and common bean. For example, Akyeampong (1986) reported that drought stress reduced total leaf area by 58% in cowpea cultivar TVu 4552 as compared to control plants. In addition, common bean also showed a 22% of reduction in leaf area when water stress was imposed (Ghanbari *et al.,* 2013).

Water serves as the medium and substrate for photosynthesis, transportation of nutrients and minerals, cell expansion, biochemical and enzymatic reactions in plants (Hsiao, 1973). Drought stress could easily effect plant growth and physiological responses, as the water content in plants ranges from 70%-90% of the plant fresh mass (Gardner *et al.,* 1984). However, the nature and degree of drought damage in Bambara groundnut due to drought is also dependent upon the developmental stage affected (Vurayai *et al.,* 2011; Jorgensen *et al.,* 2011). Bambara groundnut is more vulnerable to drought during the pod filling stage, then the flowering stage and then the vegetative stage, as plants stressed at the pod filling stage failed to fully recover their *relative water content* and chlorophyll fluorescence after irrigation was resumed (Vurayai *et al.,* 2011).

In the current study, mild drought stress was applied to a Bambara groundnut F_5 segregating population at the early flowering stage in order to investigate the immediate response of Bambara groundnut to water stress and the effects of mild drought on final yield. The study investigated how the crop deals with the early stages of drought stress, when the changes in gene expression are likely to reflect initial protective mechanisms, rather than extreme stress. Gene expression in situations of extreme stress may represent plants in a terminal state beyond full recovery. As the segregating population consists of lines which may show genetic variation for a number of characters (the parental genotypes being derived from landraces derived from Botswana (DipC) and Mali (Tiga Nicuru)), potential lines that have higher yielding characteristics and also greater tolerance under drought stress could be selected for future breeding programmes.

4.2   MATERIALS AND METHODS

4.2.1   Experimental site and plant material

The drought stress experiment was conducted in controlled-environment glasshouses at the FutureCrop Glasshouses, Sutton Bonington Campus, the University of Nottingham, UK. The dimension of glasshouse was approximately 9 m x 12 m and had an automatic drip irrigation system, automated blackouts as well as automated recording of temperature, humidity, CO_2 levels and light flux. Two soil pits of 1.2m deep containing a sandy loam soil in the glasshouse were used as the droughted plot (left) and the irrigated plot (right), with each plot having a dimension of 5m x 5m. The F_5 segregating population derived from the cross between single genotypes derived from the DipC (maternal) and Tiga Nicuru (paternal) landraces were evaluated in this drought stress experiment Plant material, consisting of two parental lines and 65 F_5 individual lines were planted in both droughted and irrigated plots with the experimental design described below.

Figure 4.3 The FutureCrop Glasshouses at Sutton Bonington Campus, The University of Nottingham, UK. (a) Three FutureCrop glasshouses located at the Sutton Bonington campus. (b) The soil plots inside one glasshouse can be seen.

4.2.2 Experimental design and crop management

The soil pits were first irrigated to encourage the germination and the growth of weeds as well as residual of Bambara groundnut seeds from last planting season. Then the soil pits were prepared by digging, raking and levelling in order to remove unwanted plants and also for a uniform soil structure, followed by the application of 50 kg/ha of Ammonium Nitrate fertilizer. The soil pits were covered with black plastic for two weeks prior to planting to kill any germinating weeds.

The experiment was conducted between late June 2012 and late November 2012. The experiment was arranged in a randomised block design (RBD) with three blocks for each soil pit. Each line had three replicates, each replicate was represented by a single plant in each block after thinning. Three seeds were sowed per replicate in each soil pit, a total of 9 seeds per line, at a depth of 3-4 cm and spacing of 25 cm x 25 cm between each individual, giving 20 plants per row. On 20 DAS the plants were thinned to one plant per hole. A spacing of 12.5 cm from the border of the plots was used with the wood plank to provide a physical barrier to spread of the edge plants. The photoperiod was set at 12 hours using an automated blackout system (Cambridge Glasshouses, Newport, UK) with a set 'day' temperature of 28°C and 23°C during the 'night'. Trickle tape irrigation consisted of PVC micro-porous tubing placed beside each row set to irrigate the plants at 0600 hrs and 1800 hrs for 20 minutes, twice per day with a measured flow rate of 1L/hr per tube. Four PVC micro-porous tubes were used for each soil pit. For the droughted plot, after 100% flowering was observed across all the lines at 50 DAS, the irrigation system was terminated for six weeks until 92 DAS when a 50% reduction in *stomatal conductance* was observed and irrigation resumed.

Throughout the growing season, *Phytoseilus persimilis*, a biological control agent, was used against red spider mite (*Tetranychus urticae*) and applied every two weeks. In addition, chemicals such as Savona (soap) against

114

Aphids and Thiovit (sulphur) against mildew or fungal infections were applied as needed.

4.2.3 Environmental factor measurements

To maintain a consistent environment for the growing of Bambara groundnut in the glasshouse, environmental factors within the glasshouse were monitored using an automated record system (Cambridge Glasshouses, Newport, UK) placed in the glasshouse. The conditions, such as photosynthetically active radiation (PAR), humidity and temperature were recorded every eight minutes throughout the experiment. For soil moisture measurements, three PR2 profile tubes (Delta-T devices, UK) were inserted into each soil pit across the diagonal from the irrigation source towards the end of the trickle tape, at least 1m apart from each other. Three PR2 readings, which are displayed in the unit of %Vol (volumetric water content as a percentage), were taken twice a week at 1000 hrs starting from 16 DAS until 133 DAS at soil depths of 300mm, 400mm, 600mm and 1000mm.

4.2.4 Morpho-physiological traits and drought-related trait measurement

A range of morphological and physiological traits were measured on both droughted and irrigated plots based on the Bambara groundnut descriptor list (IPGRI, IITA, BAMNET, 2000). The measurements were done during vegetative growth, flowering, podding and after harvesting. Table 4.1 states all the traits that were measured throughout the growing period.

In addition, drought-related traits including *stomatal conductance*, *relative water content*, *leaf carbon (Delta C^{13}) isotope analysis (CID)* and *stomatal density* were measured. Methods for measuring *stomatal conductance* and *relative water content* were modified from Vurayai *et al.* (2011). Due to time constraints, seven measurements were carried out on the droughted plot and only four on the irrigated plot during the course of the experiment.

Table 4.1 The morpho-physiolgical traits that were examined and their brief description based on Bambara groundnut descriptors list (IPGRI, 2000; Mwale *et al.,* 2007).

Morpho-physiological traits	Character and description
Days to emergence (DE)	Recorded as a number of days from sowing to discovering the first fully expanded leaf on the soil surface.
Days to flowering (DF)	Recorded from the emergence date to the appearance of the first flower(s).
Estimated days to podding (EDP)	Recorded from emergence to the day of first pod(s) discovery.
Internode length (IN)	Measured during harvest as the average length of the fourth internodes of the five longest stems/plant.
Peduncle length (PEL)	Measured during harvest as the average length of five peduncles per plant.
Pod. No/plant (PN)	Counted during harvest.
Pod weight/plant (PW)	Weight of pods per plant after incubating for 3 weeks at 37ºC.
Seed. No/plant (SN)	Counted after removing the shell of all pods.
Seed weight/plant (SW)	Weight of seeds per plant after incubating for 3 weeks at 37ºC.
100-seed weight (HSW)	Average weight of 100 seeds after incubating for 3 weeks at 37ºC.
Shoot dry weight (SDW)	Weight of above ground material after drying for 48 hours at 80ºC.
Harvest index (HI)	Fraction of pod weight to above ground plant weight.

Stomatal conductance (mmol $m^{-2} s^{-1}$): The reading of *stomatal conductance* (g_s) on only the abaxial side of the leaf was undertaken using an AP4 leaf porometer (Delta-T devices, UK) as initial readings of g_s on the adaxial side of the leaf were very low, in agreement with Jorgensen *et al.,* (2011). The middle leaflet of three fully expanded leaves, per plant, per replicate, were measured between the hours of 0800 hrs and 1200 hrs. Measurements were taken weekly and started from 49 DAS, before the drought treatment was applied, until two weeks after drought recovery (107 DAS). Throughout the measurement, the artificial lights were switched off manually to minimise the stress from the environment and the calibration of the porometer was done whenever there was a change of cup temperature registered on the porometer between 0800 hrs and 1200 hrs.

Relative water content (%): Relative water content (*RWC*) was determined from 48 DAS, before application of the drought treatment, until two weeks after water recovery (104 DAS). Every week one middle leaflet of three fully expanded leaves was chosen randomly and harvested from each plant per replicate. Three leaf discs (13 mm diameter) were punched from the leaflet and then placed on a pre-weighed weighing boat to obtain the fresh weight (Fw). The leaf discs were placed in a petri dish containing distilled water and left overnight under a light source to allow discs to fully hydrate to their turgid weight (Tw). Next morning the leaf discs were dried with tissue paper and Tw was obtained. The leaf discs were placed in an oven at 80°C for 48 hours to allow dry weight (Dw) to be measured. Their *RWC* was calculated as:

$$RWC = [(Fw-Dw)/(Tw-Dw)] \times 100$$

Leaf carbon (Delta C^{13}) isotope analysis: Seed samples collected from both parental lines (three replicates) and 65 individual line (one replicate) were freeze-dried using a Benchtop Freeze Dryer LSBC50 (MechaTech Systems, UK) for a week. These samples were then milled into a fine powder using an Ultra Centrifugal Mill ZM200 (Retsch, Germany). The leaf carbon (Delta C^{13}) isotope analysis was performed at the Mylnefield Research Services Ltd, Invergowrie, Dundee, DD2 5DA, Scotland. Based on their recommended protocol, approximately 0.2-0.3 mg of milled samples was encapsulated in the tin capsules that were provided. Care was taken to avoid contamination from the surroundings and the plates containing tin capsules were sealed prior to sample delivery. $^{13}C/^{12}C$ ratio values were expressed as carbon isotope composition ($\Delta 13C$) values which were calculated with reference to the Vienna Pee Dee belemnite (VPDB) scale, using laboratory standards calibrated against international standards (IAEA). The reported precision of the analysis was 0.07‰. Therefore,

$$\delta 13C(‰) = [(R \text{ sample}/R \text{ standard}) -1] \times 1000$$

where R is the $^{13}C/^{12}C$ ratio. The value of the discrimination (Δ) for ^{13}C was calculated from $δ_a$ and $δ_p$, where **a** refers to air and **p** refers to plant (Farquhar *et al.,* 1989):

$$Δ = (δ_a - δ_p \, Δ/(1+δ_p)$$

As on the VPDB scale, free atmospheric CO_2 has a current deviation of approximately -8.0 ‰ (Farquhar *et al.,* 1989), thus the final equation was:

$$Δ = 1000 \ X \ (-0.008 - δ13C(‰)/1000)/(1+ δ13C(‰)/1000)$$

Stomatal density: One leaf from each replicate for both parental lines and 65 individual lines was harvested. The abaxial side of the leaf was painted using nail polish and a thin film was mounted on a glass slide after peeling from the leaf. A drop of water was then added on top of the thin film. The counts of stomata were performed after the images were captured using a Leica BF200 compound microscope with Leica LAS EZ software (Leica Microsystems, Switzerland) at a magnification of 40X. Three counts per impression were done with a square area of 0.8071 mm^2 per impression. Therefore,

$$stomatal \ density = count \ of \ stomata/0.8071 \ mm^2 \ leaf \ area$$

4.2.5 Statistical analysis

Data for all the traits were subjected to analysis of variance (ANOVA) using Genstat 15th edition (VSN International, 2012) to determine whether statistical differences existed between lines for a given trait and to investigate the population distributions through descriptive statistics. Non-normally distributed traits were also transformed using a square root function after failing the Anderson-Darling normality test. Genstat was also used to examine the correlation relationships between the traits and also the characters that contributed the greatest variance observed among the individual lines using Principal Component Analysis (PCA).

4.3 RESULTS

4.3.1 Environmental factors

Throughout the Bambara groundnut growing season, environmental factors including temperature, humidity and photosynthetically active radiation (PAR) were recorded and were largely consistent in the fully controlled glasshouses. For example, Figure 4.4 shows the measurement of environmental factors over a day in September. The plants in the glasshouse received PAR, ranging from 160 W/m^2 to 255 W/m^2, for 12 hours from 0700 hrs to 1900 hrs. Both temperature and humidity were shown to be correlated with PAR. During the 12-hour exposure to PAR in the glasshouse, the temperature increased from 23°C to 31°C, which is the maximal temperature of that day in in the glasshouses in September, while humidity value decreased from 62% to 40% due to evaporation in the glasshouse. It is important to bear in mind that both soil pits are present in the same controlled environment glasshouse, so the overall humidity recorded is shared by the soil pits.
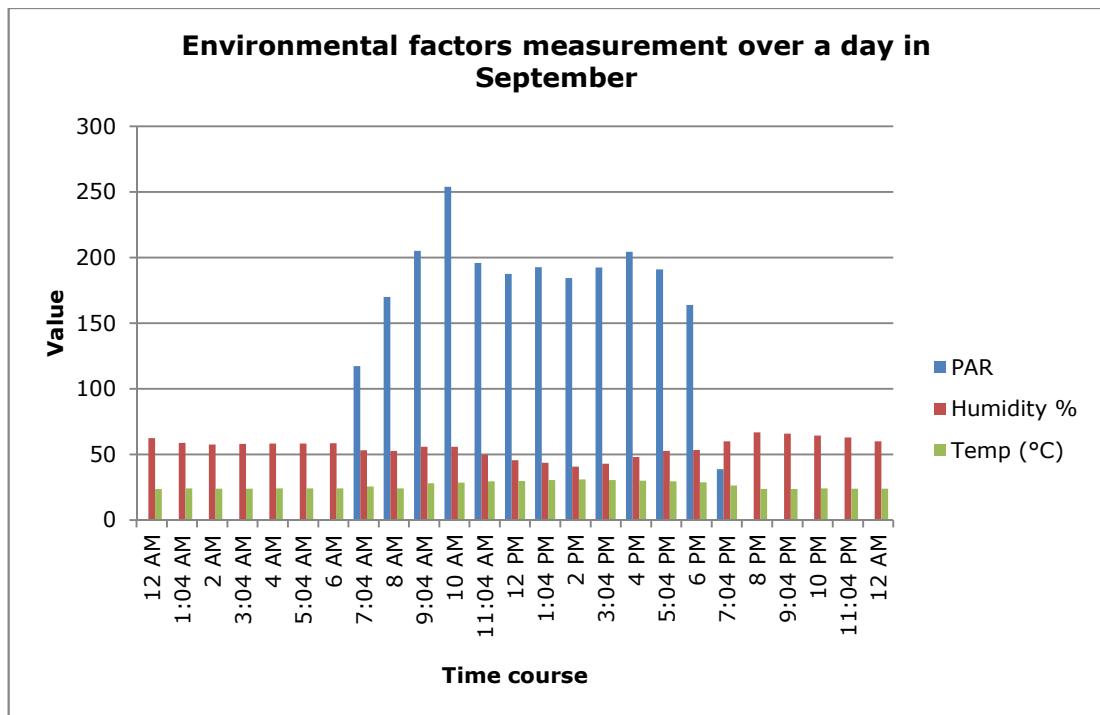
Figure 4.4 The measurement of environmental factors in September 2012 over a day (16 September 2012).

In addition, the average temperature (°C) per day was compared over growing season (Figure 4.5). At the same time point the temperature increased to a maximum (31.5°C) in August while in October the day time temperature was maintained around 28°C. As the growing season fell into the UK summer season (June to August), a slightly higher temperature than target was sometimes recorded in the glasshouse. During night time, the temperature in the glasshouse starts to drop and was maintained at average temperatures between 22°C - 24°C. The control and determination of temperature in the glasshouse is crucial as it could easily affect the growth and development of Bambara groundnut.
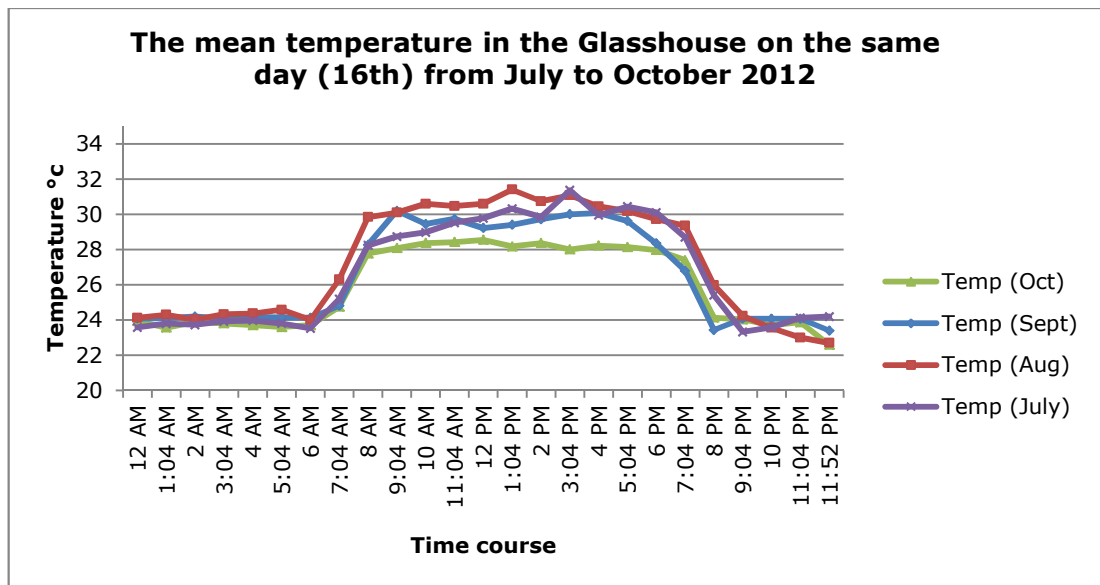
Figure 4.5   The mean temperature in glasshouse on the same date (16th) for four months from July to October 2012.

On average, the total reduction in soil moisture content during the drought treatment based on the PR2 reading was 52.7% for all depths. The irrigated plot reduced by 9.5%, from 50 DAS to 92 DAS. Soil moisture was lost rapidly at a rate of 1.95% per day at a soil depth of 400 mm, followed by 1.65% per day at a soil depth of 600 mm, from 50 DAS to 92 DAS (Figure 4.6). At 1000 mm droughted plots showed relatively constant soil moisture content and losses only became apparent at 86 DAS. In contrast, no significant changes occurred in irrigated plot from 50 DAS to 92 DAS for all depths (Figure 4.7). As a result the droughted plot has consistently lower soil moisture content as compared to fully irrigated plot from 58 DAS for all depths after imposing the drought treatment and until after the recovery treatment.
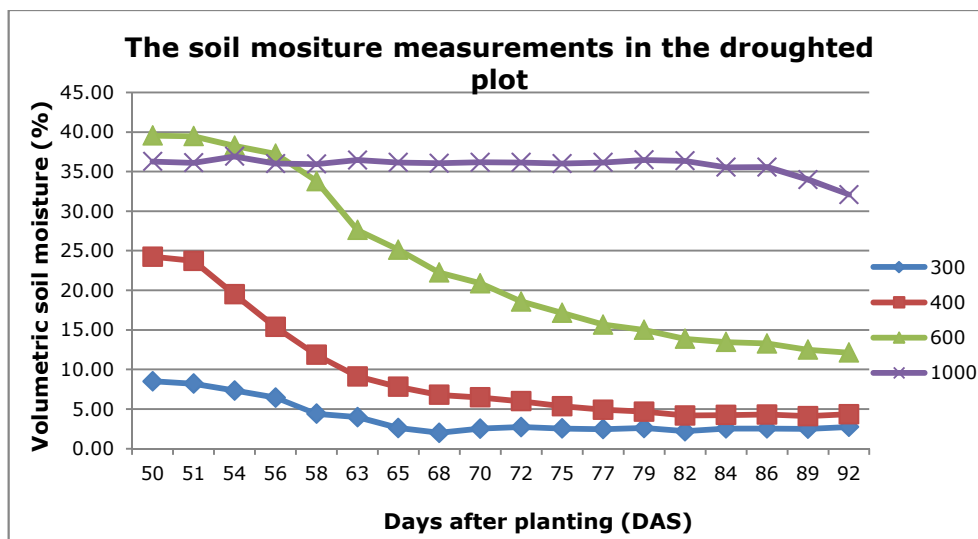
Figure 4.6 Soil moisture content based on a PR2 reading (%vol) in the drought treatment plot throughout the treatment from 50 DAS to 92 DAS.
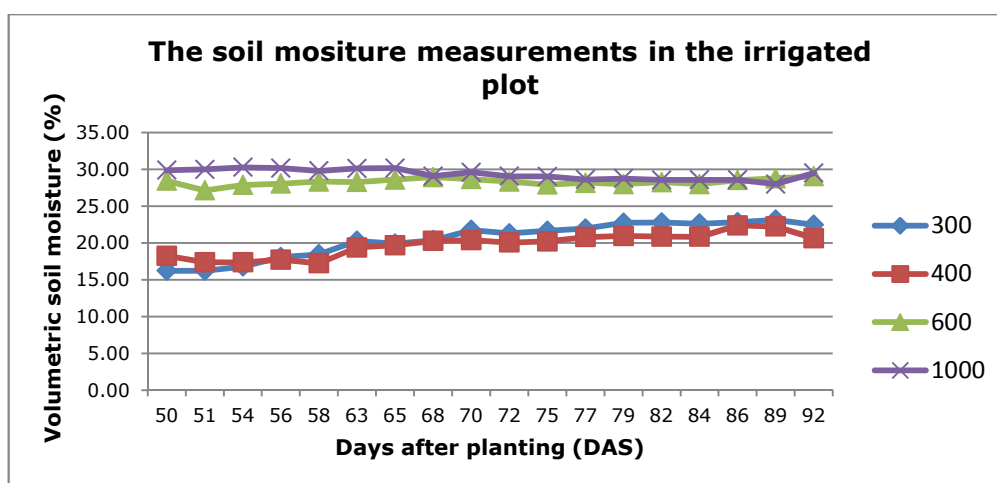
Figure 4.7 Soil moisture content based on a PR2 reading (%vol) in the fully irrigated plot throughout the treatment from 50 DAS to 92 DAS.

4.3.2 Morpho-physiological traits

4.3.2.1 *Population distributions*

Table 4.2 presents the results obtained from the analysis of morpho-physiological traits of the two parental lines and the F_5 segregating population. These are generated from single genotypes under drought and irrigated conditions, except for parental samples (n=3). The results showed that most of

the traits were normally distributed, except for *days to emergence*, *internode length* (irrigation), *pod weight per plant* (irrigation) and *seed weight per plant* (irrigation). A standard normal distribution has a kurtosis value and skewness value of zero. Non-normal distributed traits, for example, *days to emergence,* exhibited a right-skewed (1.30) and a leptokurtotic distribution (2.80) while *internode length*, *pod weight* as well as *seed weight* in the irrigation plot showed a platykurtic distribution. Nevertheless, after transformation of the data using the square root function the data showed a normal distribution (Figure 4.8).

Internode length (irrigation)
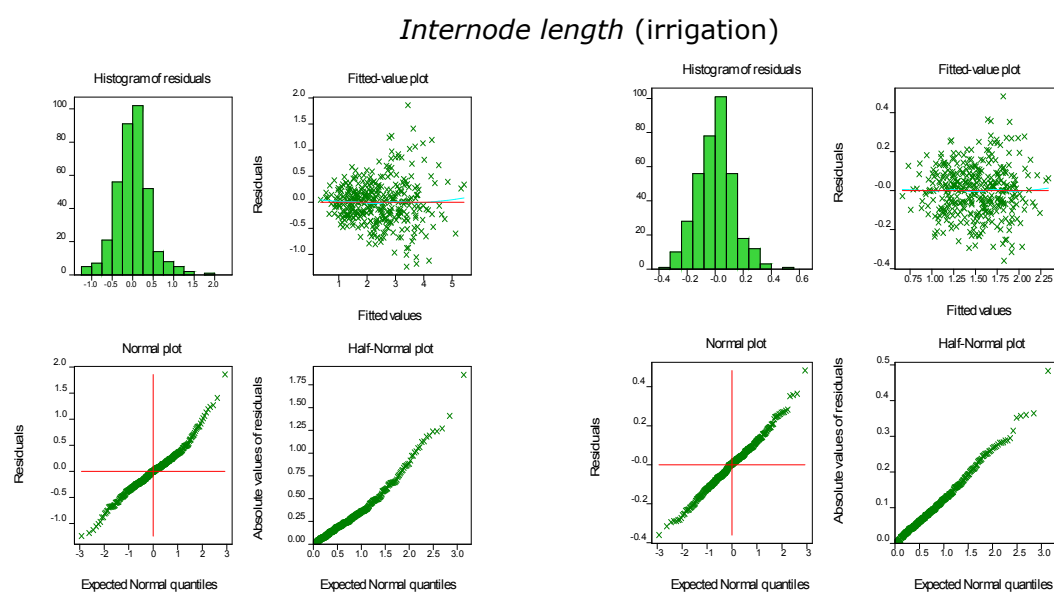
Figure 4.8 The histogram, fitted-value plot, normal plot and half-normal plot of normal distribution for *internode length* (irrigation) before (left) and after (right) transformation using square root function.

Table 4.2 Descriptive statistics for morphological and physiological traits measured in two parental lines and the F_5 segregation population under both drought and irrigated conditions.

Traits	Treatment	Mean	SD	Min	Max	Skewness	Kurtosis	Normality	DipC		Tiga Nicuru	
									Min	Max	Min	Max
Days to emergence	-	7.38	0.61	6.46	9.83	1.30	2.80	1.27^{**}	7.00	8.00	6.00	6.50
Days to flowering	-	31.84	2.53	27.33	41.17	0.80	1.39	0.49^{ns}	28.00	33.00	28.50	35.00
Estimated days to Podding	Drought	57.35	3.45	49.67	64.33	-0.16	-0.41	0.28^{ns}	55.00	61.00	50.00	56.00
	Irrigation	57.31	3.24	50.33	63.67	-0.39	-0.32	0.73^{ns}	53.00	58.00	51.00	54.00
Internode length (cm)	Drought	2.48	1.00	0.71	5.29	0.46	-0.13	0.35^{ns}	1.74	2.22	2.54	3.04
	Irrigation	2.21	0.92	0.52	4.15	0.32	-0.91	0.85^{*}	1.54	2.04	1.57	2.82
Peduncle length (cm)	Drought	3.50	1.48	0.60	7.28	0.12	-0.65	0.37^{ns}	2.54	3.06	3.54	4.60
	Irrigation	3.12	1.48	0.57	6.15	0.17	-0.98	0.59^{ns}	1.65	2.38	1.94	3.56
Pod. No/plant	Drought	53.40	25.45	7.50	126.70	0.50	0.08	0.46^{ns}	59.00	73.00	20.00	32.00
	Irrigation	46.79	23.76	3.00	105.70	0.41	-0.24	0.38^{ns}	44.00	106.00	21.00	23.00
Pod weight (g/plant)	Drought	36.01	19.12	4.36	83.09	0.47	-0.12	0.48^{ns}	39.21	49.64	11.32	14.36
	Irrigation	38.25	22.65	1.98	85.51	0.33	-1.05	1.13^{**}	28.41	76.83	10.77	11.26
Seed. No/plant	Drought	53.47	26.60	6.50	129.30	0.50	-0.06	0.44^{ns}	58.00	72.00	26.00	28.00
	Irrigation	48.28	26.35	3.00	116.70	0.49	-0.46	0.68^{ns}	38.00	105.00	15.00	16.00
Seed weight (g/plant)	Drought	26.47	13.96	1.95	62.36	0.45	-0.12	0.46^{ns}	28.00	39.40	8.57	8.59
	Irrigation	27.12	16.24	1.28	57.72	0.33	-1.14	1.36^{**}	23.14	61.79	6.81	8.24
100-seed weight (g)	Drought	49.24	12.02	24.48	81.89	0.42	0.10	0.66^{ns}	52.83	58.80	37.34	44.15
	Irrigation	53.55	12.53	26.67	89.42	0.36	-0.16	0.68^{ns}	58.85	60.89	45.40	51.50
Shoot dry weight (g/plant)	Drought	50.62	16.83	17.03	100.20	0.61	0.59	0.59^{ns}	44.75	51.36	26.23	32.96
	Irrigation	45.88	17.41	14.93	92.30	0.58	0.16	0.61^{ns}	48.47	105.26	27.31	29.39
HI index	Drought	0.65	0.23	0.19	1.23	-0.10	-0.31	0.51^{ns}	0.81	1.11	0.43	0.44
	Irrigation	0.77	0.31	0.10	1.65	0.02	-0.15	0.39^{ns}	0.73	1.00	0.38	0.39

ns not significant.

[a] Standard error for average in preceding column; [b] Level of significance * $p<0.05$; ** $p<0.01$

124

The key traits that differentiate the two parental lines, DipC and Tiga Nicuru, irrespective of drought treatment are *internode length*, *pod number per plant*, *pod weight per plant*, *seed weight per plant* and *100-seed weight* at a significance level of $p<0.05$ as well as *peduncle length* and *harvest index* at a significance level of $p<0.01$.

As the F_5 is a segregating population derived from two genetically dissimilar parental genotypes, genetic variability between lines for key traits would be expected. For example, *internode length* in the population had a minimum range of between 0.5 cm and 0.7 cm and a maximum range of between 4.2 cm and 5.3 cm, *pod weight per plant* minimum range of between 3.0 g and 7.5 g and maximum range of between 126.7 g and 105.7 g, *seed weight per plant* minimum range of between 1.3 g and 2.0 g and a maximum range of 57.7 g and 62.4 g (Table 4.2). The distribution and segregation of each trait from parental lines to the progenies are described as below:

Days to emergence: *Days to emergence* varied among the lines in the F_5 segregating population. The minimum and maximum number of days for Bambara groundnut seedlings to emerge was 6.5 and 9.8, respectively. As examined in the ANOVA, both plots showed that on average the parental line Tiga Nicuru (6.3 days) germinated earlier than DipC (7.5 days) at a significance level of $p<0.01$.

Days to flowering: Bambara groundnut plants in the F_5 segregating population required a minimum of 27.3 days to start flowering although some of the lines took 41.2 days to flower under a day-length of 12 hours. Although Tiga Nicuru emerges earlier than DipC, both of them started flowering on the same day which is on average 30.5 days after emergence, suggesting transgressive segregation in the offspring.

Estimated days to podding: The data was recorded when the first visible pod appeared on the surface of the soil with a diameter of 5mm or greater. However, Tiga Nicuru usually bears seed under the soil and is likely to be more developed before a pod breaking the soil is first observed. Similar seed-bearing characters would be expected within some of the individual lines, thus *estimated days to podding* was measured for the current segregating population, but the caveat should be noted. ANOVA analysis showed significant differences between the two parental lines ($F_{(1,6)}$=15.96, $p<0.01$) in which Tiga Nicuru (mean: $52.3_{d;i}$) had an earlier *estimated day to podding* than DipC (mean: 58.0_d; mean: 55.7_i). As the trait segregates in the F_5 population, the minimum *estimated days to podding* of 49.7 and maximum 64.3 was recorded in the droughted plot while the minimum value of 50.3 and maximum 63.7 was recorded in irrigated plot, based on single plant measurements

Shoot dry weight: Reflects the amount of energy stored in plant biomass during growth. DipC (44.8 g-105.3 g) generally produced more *shoot dry weight* than Tiga Nicuru (26.3 g-33.0 g) and a normal and continuous distribution for *shoot dry weight* was observed in the F_5 segregating population. However, no significant differences were observed between the two parental lines as well as between the treatments.

Internode length: High genetic variability was observed for this trait because DipC is well known as a bunched morphology type and is expected to have shorter *internode length* than Tiga Nicuru, which has a spreading growth habit. ANOVA results showed that DipC has a shorter *internode length* than Tiga Nicuru at a significance level of $p<0.05$. In the F_5 segregating population *internode length* ranged from 0.7 cm to 5.3 cm in the droughted plot whereas *internode length* varied from 0.5 cm to 4.2 cm in the irrigated plot.

Peduncle length: A significant difference between the two parental lines was obtained *(p<0.01)*. The trait segregates in the offspring lines. A wide range of *peduncle length* was also observed in both the droughted plot (0.6 cm–7.3 cm) and the irrigated plot (0.6 cm–6.2 cm). Lines that exhibited a short peduncle are believed to have inherited the trait from DipC (mean: 2.1_i) while the long *peduncle length* was inherited from Tiga Nicuru (mean: 2.9_i), although there is also evidence for transgressive segregation in the F_5 for the trait.

Pod number per plant: *Pod number* was measured in order to estimate the yield of Bambara groundnut. Higher *pod number* was obtained in DipC, compared to Tiga Nicuru ($F_{(1,6)}=8.33$, $p<0.05$). For the segregating population, the plants produced as few as 7 (droughted) and 3 *pods per plant* (irrigated) and or as high as 126.7 (droughted) and 105.7 *pods per plant* (irrigated).

Pod weight per plant: The *pod weight per plant* ranged from 4 to 83 g and 2 to 85 g under droughted and irrigated conditions, respectively. As a result of having low numbers of pods, a two-fold reduction of *pod weight per plant* was observed in Tiga Nicuru compared to DipC under both irrigated and droughted conditions ($p<0.05$).

Seed number per plant: Similar to *pod number per plant*, *seed number per plant* of the current population was highly variable among individual lines. After the removal of the pod shell, 6.5 to 129.3 *seeds per plant* in the droughted plot were observed while 3 to 116.7 *seeds per plant* in the irrigated plot were obtained. Although there was no significant difference, DipC showed two-fold increase in the number of seeds produced by Tiga Nicuru in both droughted and irrigated plots. Some of lines also contain double-seeded pods, thus higher *seed number per plant* could be due to the presence of double-seeded pods. Nevertheless, some of the seeds also possibly abort inside the pods, thus fewer seed were obtained in some of the lines, compared to *pod number per plant*.

Seed weight per plant: Bambara groundnut plants produced seed with a minimum and maximum weight of 2.0 g and 62.4 g per plant in the droughted plot and 1.3 g and 57.7 g in the irrigated plot. For the parental lines, Tiga Nicuru has a lower mean value for *seed weight per plant* (mean: 8.6_d; mean: 7.5_i) which is significantly different from DipC (mean: 33.0_d; mean: 42.5_i) at a significance level of $p<0.05$.

100-seed weight: Overall, significantly lower *100-seed weight* ($F_{(1,258)}=19.4$, $p<0.01$) was obtained when the F_5 population was placed under drought stress. When plants were stressed, a minimum value of 24.5 g and maximum value of 81.9 g for *100-seed weight* was recorded with a mean value of 49.2 g. Higher *100-seed weight* was obtained in the irrigated plot, ranging from 26.7 g to 89.4 g with a mean value of 53.6 g. Although no significant difference was observed between the drought and irrigated treatments for the two parental lines, significantly higher *100-seed weight* was observed in DipC than in Tiga Nicuru ($F_{(1,6)}=12.22$, $p<0.05$).

Harvest index: No significant difference was observed between the two treatments for the parental lines. However, the *harvest index* in DipC was significantly higher than in Tiga Nicuru ($F_{(1,6)}=14.98$, $p<0.01$). For the population a significantly higher mean value of 0.77 was obtained in the irrigated plot as compared to the droughted plot which had a *harvest index* mean value of 0.65 ($F_{(1,258)}=12.87$, $p<0.01$).

Overall, broad trait variation between the two parental lines allowed different characters to segregate in the F_5 population. For each trait several lines were better or worse than the parental lines in the drought treatment, suggesting possible transgressive segregation in the population, although some traits are more likely to have large environmental components to the observed variation than others, particularly complex yield traits. For example, plants in the population had a maximum *internode length* of 4.5 cm-5.8 cm (L64; mean: 5.3; s.d.: 0.7; n=3) and minimum 0.7 cm-0.8 cm (L103; mean: 0.8; s.d.: 0.09; n=3) while DipC has

internode length ranged from 1.7 cm-2.2 cm (parental mean: 2.0; parental s.d.: 0.2; n=3) and Tiga Nicuru 1.6 cm-3.0 cm (parental mean: 2.5; parental s.d.: 1.0; n=3; Table 4.2).

Although ANOVA analysis shows significant differences ($p<0.01$) among the lines for all traits, drought stress did not significantly influence plant phenology as measured by *estimated days to podding*, or morphology and growth parameters, including *pod weight per plant*, *seed number per plant* and *seed weight per plant*. Nevertheless, a significant increase of *internode length* ($F_{(1,258)}= 27.45$, $p<0.01$), *peduncle length* ($F_{(1,258)}= 33.09$, $p<0.01$) and *shoot dry weight* ($F_{(1,258)}= 8.56$, $p<0.01$) as measured at final harvest is observed between lines in droughted plot, which is suspected could be the result of rapid plant growth when the water stress is relieved. Although *pod number per plant* was higher in the droughted plot ($p<0.05$), a significant reduction of *100-seed weight* and *harvest index* ($F_{(1,258)}= 19.4$, $p<0.01$; $F_{(1,258)}= 12.87$, $p<0.01$) by 8% and 15.6%, respectively, in the drought treatment occurred between lines, implying that mild drought may negatively influence yield accumulating processes in Bambara groundnut, particularly partitioning to seed (although perhaps not altering sink number). Given that the segregating population has high trait variability, there were lines that produced high *100-seed weight* such as L89 (D: 81.9 g; IR: 89.4 g), L5 (D: 72.5 g; IR: 70.8 g) and L101 (D: 69.4 g; IR: 64.2 g) and lines obtained low *100-seed weight* such as L41 (D: 24.7 g; IR: 37.3 g), L45 (D: 27.2 g; IR: 28.8 g) and L37 (D: 33.5 g; IR: 26.7 g) in the droughted plot (population mean and s.d.: 49.2 and 12.0) and irrigated plot (population mean and s.d.: 53.6 and 12.5), respectively, suggesting an intrinsic rather than a treatment related difference in trait.

4.3.2.2 *Correlation between the traits*

In addition to descriptive statistics, investigation of any associations among the traits is important. The correlations among different morphological and physiological traits in the F_5 segregating population under drought and irrigation

treatment were investigated and are presented in Table 4.3. Correlations could potentially be used to assist selection and breeding, if one early trait is strongly correlated within final production traits.

A negative correlation between *estimated days to podding* and *100-seed weight* was obtained in both the droughted plot ($p<0.01$) and the irrigated plot ($p<0.05$). The droughted and irrigated plot showed a significant coefficient correlation of r=-0.42 and r=-0.28 between *estimated days to podding* and *100-seed weight*, respectively. A possible reason for the negative correlation could be due to the underestimation of scoring first pod dates in lines with traits inherited from Tiga Nicuru, which buries the pods in the soil. In addition, these lines would probably produce lower *100-seed weight* due to the seed traits inherited from Tiga Nicuru, thus producing a potential negative correlation of *estimated days to podding* could be related to lower *100-seed weight*. However, this is quite speculative as a possible explanation and would require some degree of linkage between the loci determining *estimated days to podding* and yield component genes.

Internode length was positively correlated with several traits such as *peduncle length* (r=0.80_d; r=0.82_i), *shoot dry weight* (r=0.60_d; r=0.62_i), *pod number per plant* (r=0.53_d; r=0.66_i), *pod weight per plant* (r=0.60_d; r=0.66_i), *seed number per plant* (r=0.58_d; r=0.65_i), *seed weight per plant* (r=0.56_d; r=0.59_i) and *harvest index* (r=0.50_d; r=0.58_i) in the irrigated plot. *Internode length* was shown to be closely correlated to *peduncle length* while a moderate correlation was observed between *internode length* and other yield-related traits.

Shoot dry weight was found to be highly correlated with *pod number per plant* (r=0.87_d; r=0.79_i), *pod weight per plant* (r=0.89_d; r=0.86_i), *seed number per plant* (r=0.86_d; r=0.78_i) and *seed weight per plant* (r=0.85_d; r=0.84_i). In addition, *shoot dry weight* was also shown to have a moderate correlation with *100-seed weight* in the irrigated plot (r=0.48) but not in the droughted plot.

A strong correlation between *pod number per plant* and yield traits in both plots such as *pod weight per plant*, *seed number per plant*, *seed weight per plant*

130

and *harvest index* was shown at a significance level of *p*<0.01. Thus *pod number per plant* is suggested as an early indicator for the yield of Bambara groundnut plant. The higher the number of pods produced in a plant, the larger is the *pod weight per plant* as well as *seed number per plant* and *seed weight per plant*. There was also a positive relationship between *harvest index* and *pod weight per plant*, *seed number per plant*, *seed weight per plant* and *shoot dry weight*. Furthermore, a positive impact of *pod weight per plant* and *seed weight per plant* on *100-seed weight* was observed in both the droughted plot (r=0.43; r=0.48) and the irrigated plot (r=0.49; r=0.54) respectively.

Table 4.3 Pearson's Correlation Coefficients between different morphological and physiological traits measured in the F_5 segregating population derived from the cross between DipC and Tiga Necaru, under drought condition and irrigation condition.

		1	2	3	4	5	6	7	8	9	10
Estimated days to podding_D	1	-									
Estimated days to podding_IR	1	-									
Internode length_D	2	0.11	-								
Internode length_IR	2	0.03	-								
Peduncle length_D	3	0.21	**0.80****	-							
Peduncle length_IR	3	-0.02	**0.82****	-							
Shoot dry weight_D	4	0.11	**0.60****	**0.59****	-						
Shoot dry weight_IR	4	-0.06	**0.62****	**0.54****	-						
Pod. No/plant_D	5	0.10	**0.53****	**0.47****	**0.87****	-					
Pod. No/plant_IR	5	0.01	**0.66****	**0.57****	**0.79****	-					
Pod weight/plant_D	6	-0.01	**0.60****	**0.58****	**0.89****	**0.87****	-				
Pod weight/plant_IR	6	-0.08	**0.66****	**0.63****	**0.86****	**0.92****	-				
Seed. No/plant_D	7	0.11	**0.58****	**0.53****	**0.86****	**0.97****	**0.88****	-			
Seed. No/plant_IR	7	0.01	**0.65****	**0.58****	**0.78****	**0.97****	**0.94****	-			
Seed weight/plant_D	8	-0.05	**0.56****	**0.53****	**0.85****	**0.85****	**0.98****	**0.86****	-		
Seed weight/plant_IR	8	-0.11	**0.59****	**0.58****	**0.84****	**0.89****	**0.98****	**0.92****	-		
100-seed weight_D	9	**-0.42****	0.17	0.21	0.23	0.11	**0.43****	0.07	**0.48****	-	
100-seed weight_IR	9	**-0.28***	0.18	**0.28***	**0.48****	0.21	**0.49****	**0.25***	**0.54****	-	
Harvest index_D	10	-0.18	**0.50****	**0.48****	**0.60****	**0.67****	**0.86****	**0.70****	**0.88****	**0.61****	-
Harvest index_IR	10	-0.09	**0.58****	**0.60****	**0.52****	**0.81****	**0.85****	**0.83****	**0.83****	**0.41****	-
		1	2	3	4	5	6	7	8	9	10

* Significant level of $p<0.05$; ** Significant level of $p<0.01$

In addition to the relationships between the traits, the pattern of variation in F_5 segregating population for ten morpho-physiological traits was examined through Principal Component Analysis (PCA). In this study, three principal components (PC) having eigenvalues more than one were extracted. Table 4.4 showed that three PCs contributed 78.78% and 78.33% of the total variability among the segregating lines for droughted plot and irrigated plot, respectively. The first principal component (PC 1) contributed 51.60% and 54.07% of the variation in droughted and irrigated plots respectively, and the characters that gave higher values were *shoot dry weight*, *internode length*, *peduncle length*, *pod number per plant*, *pod weight per plant*, *seed number per plant* and *seed weight per plant* and *harvest index*. The second principal component (PC 2) accounted for 17.67% and 14.68% in the droughted plot and the irrigated plot, respectively, and the characters with high loadings were *estimated days to podding* (drought) and *100-seed weight* (irrigation). *Estimated days to podding* and *100-seed weight* accounted for most of the 9.58% identified at the third principal component (PC 3) in irrigated plot. None of the characters showed a significant contribution to the 9.51% of variation observed in the droughted plot. PCA analysis summarised the amount of diversity for the characters among the segregating lines, despite the application of drought treatment, into three components with *shoot dry weight*, *internode length*, *peduncle length*, *pod number per plant*, *pod weight per plant*, *seed number per plant*, *seed weight per plant* and *harvest index* being the main contributors. The utilisation of genetic variability for various morpho-physiological traits could be exploited to conduct breeding programmes in Bambara groundnut as it is assumed that maximum variability observed within the population produces maximum heterosis (Ali *et al.,* 2011).

Table 4.4 Principal component analysis for ten characters measured in the F_5 segregating population of Bambara groundnut cross between DipC and Tiga Nicuru under drought and irrigation conditions.

	Drought			Irrigation		
	PC 1	PC 2	PC 3	PC 1	PC 2	PC 3
Eigenvalues Variance	6.19	2.12	1.14	6.48	1.76	1.14
% variation	51.60	17.67	9.51	54.07	14.68	9.58
Estimated days to podding	0.02	**0.54**	0.06	-0.02	-0.59	**0.31**
Shoot dry weight	**0.36**	0.09	0.04	**0.33**	0.01	-0.03
Internode length	**0.28**	0.10	-0.48	**0.29**	-0.18	-0.29
Peduncle length	**0.27**	0.11	-0.43	**0.28**	-0.12	-0.15
Pod. No/plant	**0.36**	0.12	0.18	**0.36**	-0.10	-0.02
Pod weight/plant	**0.39**	-0.05	0.11	**0.38**	0.02	0.08
Seed. No/plant	**0.36**	0.13	0.13	**0.37**	-0.09	0.01
Seed weight/plant	**0.38**	-0.09	0.15	**0.37**	0.07	0.11
100-seed weight	0.15	-0.50	-0.02	0.18	**0.39**	**0.33**
Harvest index	**0.33**	-0.23	0.11	**0.33**	0.02	0.17

4.3.3 Responses of Bambara groundnut to mild drought

In addition to morphological and physiological traits, drought-related variables such as *stomatal conductance*, *relative water content (RWC)*, *leaf carbon (Delta C¹³) isotope analysis* and *stomatal density* were examined to understand the immediate responses of Bambara groundnut plants subjected to a mild drought conditions.

4.3.3.1 *Stomatal conductance*

Throughout the drought stress period, grand mean values for *stomatal conductance* (g_s) declined gradually in the droughted plot from 540 mmol m^{-2} s^{-1} to 220 mmol m^{-2} s^{-1} (Figure 4.9). Drought treatment was applied at 50 DAS, g_s before treatment (49 DAS) was measured and served as a baseline for g_s over the drought period. Although there are some missing sampling dates due to the priority given to the droughted plot, consistently high values were observed in the irrigated plot (500 mmol m^{-2} s^{-1} – 600 mmol m^{-2} s^{-1}). The sudden increase in g_s at 107 DAS in the droughted plot was a result of the water recovery treatment at 92 DAS. Rewatering Bambara groundnut after the drought stress resulted in a significant increase of g_s ($p<0.01$). The analysis of data using ANOVA showed significant differences among the lines ($F_{(64,130)}=16.27$, $p<0.01$), as well as between the treatments ($F_{(1,130)}=2259.59$, $p<0.01$). Some lines are shown to have high g_s under both drought and irrigation conditions, for example, L101 (D: 274.1 mmol m^{-2} s^{-1}; IR: 581.1 mmol m^{-2} s^{-1}), L89 (D: 269.3 mmol m^{-2} s^{-1}; IR: 584.4 mmol m^{-2} s^{-1}) and L94 (D: 261.8 mmol m^{-2} s^{-1}; IR: 617.8 mmol m^{-2} s^{-1}) at 84 DAS. However, L5 (D: 166.1 mmol m^{-2} s^{-1}; IR: 432.8 mmol m^{-2} s^{-1}), L7 (D: 185.9 mmol m^{-2} s^{-1}; IR: 519.4 mmol m^{-2} s^{-1}) and L37 (D: 193.6 mmol m^{-2} s^{-1}; IR: 524.2 mmol m^{-2} s^{-1} showed lower g_s at 84 DAS.
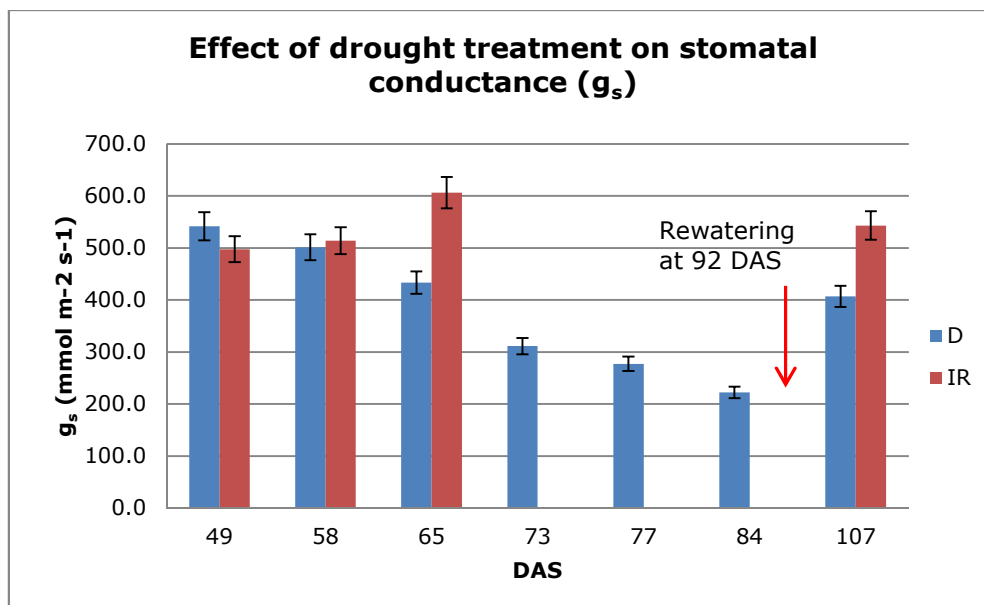
Figure 4.9   The effect of mild drought treatment on *stomatal conductance* (g_s) in the droughted and irrigated plot between 49 DAS to 107 DAS. Data points represent mean value ± standard error, n=65. Arrow: re-watering of plants at 92 DAS.

4.3.3.2 *Relative water content (RWC)*

As shown in Figure 4.10, Bambara groundnut plants in the droughted plot appear to have higher *RWC* (although not significantly) compared to the irrigated plot at the beginning of drought stress. One possible reason could be that soil moisture content was higher in the 'droughted' plot than the 'irrigated' plot before drought stress was applied. Although leaf *RWC* in the droughted plot starts to decrease (albeit, erratically) after 65 DAS and consistently remains lower than the irrigated plot, the reduction in *RWC* is not significantly different between the treatments.
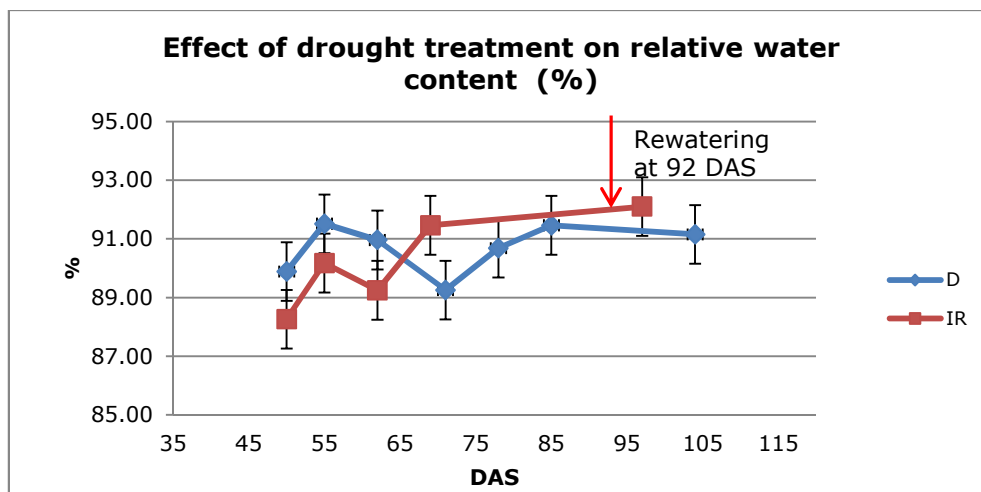
Figure 4.10 The effect of drought treatment on *relative water content* (%) in the droughted and irrigated plots between 48 DAS to 104 DAS. Data points represent mean value ± standard error, n=65. Arrow: rewatering of plants at 92 DAS.

4.3.3.3 *Leaf carbon (Delta C^{13}) isotope analysis*

Significant differences between the two parental lines for leaf carbon isotope analysis (δC^{13}; $F_{(1,6)}$=21.33, p<0.01) were found. Table 4.5 showed that lower δC^{13} was associated with higher yield as observed in DipC, compared to Tiga Nicuru. However, there was no obvious effect of drought treatment on δC^{13} as no significant difference was observed between treatments for two parental lines. In the F_5 segregating population, δC^{13} values ranging from 17.2 to 21.1 in the droughted plot and 15.5 to 21.3 in the irrigated plot were obtained. Although no ANOVA analysis was carried out in the segregating population due to the lack of replicates, the population exhibited variation for δC^{13} and, based on the use of this surrogate measure, water use efficiency was expected to show variation due to genotypic differences derived from two parental lines.

137

Table 4.5 The δC^{13} value of DipC and Tiga Nicuru under drought and irrigation conditions.

Sample	Treatment	Average δC^{13}	Average yield (g/plant)
DipC	Drought	17.85	33.0
DipC	Irrigation	17.77	31.6
Tiga Nicuru	Drought	19.65	8.6
Tiga Nicuru	Irrigation	19.73	7.5

4.3.3.4 *Stomatal density*

Stomatal density was significantly different among the individual lines ($F_{(64,258)}$= 4.08, $p<0.01$) and also between the treatments ($F_{(1,258)}$=22.55, $p<0.01$). Higher *stomatal density* was observed in the droughted plot compared to the irrigated plot as the plants that were stressed had a mean value of 11.64 pores cm^{-2} for *stomatal density* while plants that were fully irrigated had a mean value of 10.07 pores cm^{-2}. Among the segregating population, some lines showed high *stomatal density* such as L37 (D: 13.9 pores cm^{-2}; IR: 12.3 pores cm^{-2}), L94 (D: 12.7 pores cm^{-2}; IR: 11.2 pores cm^{-2}) and L7 (D: 11.1 pores cm^{-2}; IR: 12.1 pores cm^{-2}) whereas there were lines that showed low *stomatal density*, L112 (D: 6.3 pores cm^{-2}; IR: 7.6 pores cm^{-2}), L101 (D: 7.0 pores cm^{-2}; IR: 6.9 pores cm^{-2}) and L5 (D: 7.4 pores cm^{-2}; IR: 9.0 pores cm^{-2}).

In addition, the leaf area of the same leaf that was used for the stomatal count, total three leaves per line, was also analysed. ANOVA showed that smaller leaf areas were obtained in the droughted plot (mean: 18.92 cm^2) than in irrigated plot (mean: 22.25 cm^2) at a significance level of $p<0.01$. Stomatal density was also discovered to have a moderate and negative relationship with *100-seed weight* and *harvest index* (r=-0.40, $p<0.01$; r=-0.42, p <0.01). However, a low negative Pearson's Correlation Coefficients (r=-0.28, $p<0.05$) is observed between *stomatal density* and *stomatal conductance*.

4.4 DISCUSSION

4.4.1 Effect of mild drought on Bambara groundnut

A rapid reduction in g_s when mild drought is applied is consistent with observations reported by Collinson *et al.* (1997) and Vurayai *et al.* (2011) in Bambara groundnut, implying that the regulation of stomata closure for water loss is one of the early events to occur in Bambara groundnut in response to drought. Stomatal regulation is known to be closely linked to soil moisture content as stomata are sensitive and respond towards chemical signals such as ABA produced by dehydrating roots (Davies and Zang, 1991). The present study also showed a strong and positive relationship (R^2=0.96) between g_s and soil moisture content, for example, at the depth of 600 cm (Figure 4.11). Given that fewer data points were obtained in the irrigated plot, g_s for the plants under the irrigation treatment remain consistently higher than droughted plot, with no significant changes occurring for soil moisture content in the irrigated plot.
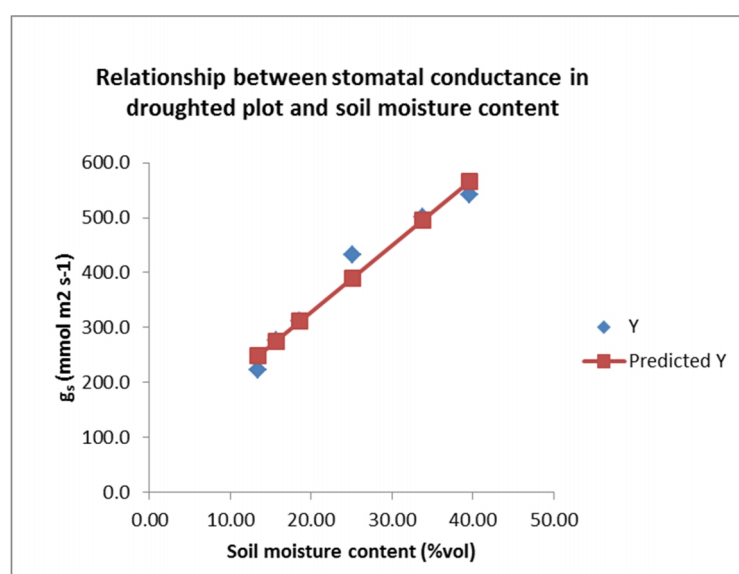
Figure 4.11 The relationship between the observed *stomatal conductantce* g_s (mmol m^{-2} s^{-1}, Y) and the observed soil moisture content (%vol) and the predicted *stomatal conductance* and observed soil moisture content based on the soil moisture at a depth of 600 cm in droughted plot (R^2=0.96, *p*<0.01).

It is worth noting that the experimental conditions measure soil water deficit alone. As both soil pits were in the same glasshouse, it is likely that vapour pressure deficit reflects the combined effects of droughted and irrigated plots and the humidity within the glasshouse did not decrease below 30%. This is also likely to have mitigated the effects of the drought treatment. The observation is supported by Franks *et al.* (1997) who reported the stomatal respond to air humidity and water status in addition to soil moisture content.

A rapid decline in g_s between 65 DAS to 72 DAS (15.23 mmol m^{-2} s^{-1} per day), followed by a relatively slow and steady decline between 72 DAS and 84 DAS (8.07 mmol m^{-2} s^{-1} per day) was observed in the droughted plot (Figure 4.9). Collinson *et al.* (1997) stated that stress-induced stomata closure is believed to be accompanied by osmotic adjustment. Once the decline of g_s reaches a threshold value due to drought stress, g_s shows little or no change as the plants are speculated to keep the stomata opened for carbon uptake while maintaining their plant water status by osmotic adjustment. Collinson *et al.* (1997) also observed a relatively unchanged g_s value (0.13 cm s^{-1}-0.25 cm s^{-1}) at lower leaf water potentials in Bambara groundnut and thus suggested that this is a common response to drought using osmotic adjustment to maintain turgor in the plant. Osmotic adjustment could be attributed to various osmolytes, such as proline. Although the relationship between g_s and proline was not directly proven, during the experiment 10-fold reduction of g_s in stressed Bambara groundnut plants together with 4 times higher in the concentration of free proline accumulation in Bambara groundnut leaves was obtained, compared to control plants (Vurayai *et al.,* 2011). Changes in proline content, as a result of drought stress, have been observed in other crops as well, for instance, cowpea (Somal *et al.,* 1998), maize (Mohammadkhan and Heidari, 2008) and wheat (Cattivelli *et al.,* 2000). Thus, osmotic adjustment is believed to maintain plant water status along with stomatal regulation for water loss.

When mild drought stress is imposed slowly under field conditions along with decreased *stomatal conductance* a reduction in carbon assimilation and utilisation

may also occur (Yordanov *et al.,* 2003). In the present study the reduction of *100-seed weight* in the segregating population by 8% in the drought treatment suggests that the rate of CO_2 assimilation may be depressed by restricted gaseous diffusion due to drought stress resulting in lower intercellular net CO_2 as well as lower chloroplastic CO_2 (Vurayai *et al.,* 2011). Insufficient CO_2 in the plant will thus produce a negative impact on plant growth and yield as a result of decreased photosynthetic efficiency and dry matter production (Vurayai *et al.,* 2011). This speculation is also supported by Maroco *et al.* (2002) who reported a decrease in the activity of enzymes of the Calvin cycle, such as Rubisco, G3PHD, Ru5PKin and FruBPase, from modelled responses of net photosynthetic to internal CO_2 when field-grown grapevine were subjected to drought conditions. The authors concluded that limitation of CO_2 assimilation due to stomatal closure in grapevines, followed by reduced photosynthetic activities is one of the major responses of plants to drought stress.

Drought stress reduced *stomatal conductance* varied significantly in the segregating population but did not show significant differences between lines for *RWC* or δC^{13} analysis. An unstable decline of *RWC* (Figure 4.10; non-significant) was also observed between 36 DAS and 60 DAS in a previous study that was conducted in controlled environment glasshouses, followed by a gradual decline of *RWC* from 93% to 83% between 60 DAS and 137 DAS (Collinson *et al.,* 1997). As a mild drought (in total 42 days) was imposed in the present study, *RWC* appeared not to be significantly influenced by the stress. The maintenance of relatively high *RWC* despite the drought stress in Bambara groundnut appears to be a common trait in drought-resistant species (Collinson *et al.,* 1997).

For δC^{13} analysis, the lower the value of δC^{13}, the higher is the water use efficiency. In drought prone environments, this may feed through into higher yields (Ebdon and Kopp, 2004), although the direction of the relationship with respect to yield can be influenced by the severity of the drought. For example, a positive correlation between δC^{13} and yield was identified in barley and wheat in

Mediterranean irrigated conditions, whereas in Australian environments where crop growth is reliant on stored soil water, a negative correlation is associated with higher grain yield (Araus *et al.,* 2007). No significant difference was observed between the treatments for parental lines, indicating that mild drought did not significantly bias carbon fixation during the drought period. δC^{13} implies that there is no significant impact of the drought on water use efficiency.

In addition, *stomatal density* was found to be significantly influenced by the drought stress. However, no direct relationship can be determined as the total leaf area per plant was not determined in this study. Instead of stomatal effects, reduced leaf area seems to be the main factor that causes a higher *stomatal density* observed in plants that are stressed. Drought stress reduces leaf area index as well as the size of the canopy in Bambara groundnut (Collinson *et al.,* 1997; Mwale *et al.,* 2007). In the present study, although total leaf area per plant was not determined, an analysis of the leaf area of the same leaves used for stomatal count showed that the smaller leaf areas were obtained in the droughted plot than in irrigated plot. The observation of high *stomatal density* and reduced leaf area in stressed Bambara groundnut plant is consistent with previous studies that report a negative correlation between leaf area and *stomatal density* in *Leymus chinensis* under moderate drought (Xu and Zhou, 2008). The moderate and negative relationship between *stomatal density* and *100-seed weight* as well as *harvest index* observed in the present study is also comparable with the result presented by Meng *et al.* (1999) in which the net photosynthetic rate is significantly negatively correlated with *stomatal density* in rice. Thus, in addition to stomatal closure, a reduction in leaf area is also an early response to drought stress in Bambara groundnut, allowing plants to reduce water loss, although with an inevitable decrease in carbon uptake, leading to limitations in photosynthetic assimilation (Xu and Zhou, 2008).

The effect of the drought treatment could only be observed after 1-2 weeks after application, which was at the pod filling stage. Therefore, the plants in the droughted plot are believed to have been well-established before the drought stress

took effect, resulting in a better crop performance overall and one that was not significantly different from plants in irrigated plot. However, it is also possible that rapid plant growth recovery occurred when the plants were relieved from the mild drought stress, thus resulting in a significant increase of *internode length, peduncle length* and *shoot dry weight* in the segregating population in the droughted plot. There are two possible reasons for the rapid growth: Otieno *et al.* (2005) stated that cell wall elasticity was improved and better-adapted water-conducting vessels were developed during prolonged moderate drought stress in *Acacia xanthophloea*, thus result in rapid growth and allowing plants to recover rapidly after water stress is relieved. Another possible reason for rapid plant growth is associated with a decrease in sugar and proline content after rewatering and these solutes are likely to be utilised in growth after the stress is alleviated (Kameli and Losel, 1993). Hare and Cress (1997) also stated that a decrease in proline content after rewatering could serve as a sources for recovering tissues, generation of ATP for recovery from stress as well as repair of stress-induced damage.

The pod filling stage was affected by drought stress, despite more *pods per plant* being observed in the droughted plot, *100-seed weight* and *harvest index* are reduced significantly in the segregating population. Reduction in *100-seed weight* under drought conditions in the present study agrees with previous reports in Bambara groundnut (Vurayai *et al.,* 2011; Mwale *et al.,* 2007), common bean (Szilagyi, 2003) and soybean (Liu 2004). In addition, ANOVA analysis showed significant differences ($F_{(64,258)} = 7.66$, $p<0.01$) among the lines for *100-seed weight* as well as the interaction between the lines and the drought treatment ($F_{(64,258)} = 1.93$, $p<0.01$). The significant interaction indicates that the line in the segregating population are responding differently to drought in relation to *100-seed weight.* The finding is in consistent with Mwale *et al.* (2007) who reported that seed weight as a result of drought stress may vary across different genotypes. For example, in pea, seed weight of one cultivar was increased by drought while decreased seed weight was observed in another cultivar (Baigorri *et al.,* 1999).

The mean *harvest index* of 0.77 under irrigation conditions in the F_5 segregating population is relatively higher than those reported in Bambara groundnut landraces, for instance, a *harvest index* of 0.46 by Mwale *et al.* (2007) and between 0.30 and 0.37 by Collinson *et al.* (1999). High values of *harvest index* implied that large portion of dry matter is allocated to pods in Bambara groundnut under irrigated conditions (Mwale *et al.,* 2007). Same authors also reported that the *harvest index* would be increased in varieties, which are bred specifically for their yield, rather than in landraces. Selection for high harvest index has driven a lot of the historical yield increases in cereals. In addition, *harvest index* remained as high as 0.65 in the F_5 segregating population in the droughted plot and was distinctly different from Collinson *et al.* (1999) who reported a *harvest index* of 0-0.08 in drought-stressed Bambara groundnut plants. The large variation observed suggests that reduction in *harvest index* is dependent on the timing and severity of drought (Mwanamwenge *et al.,* 1999). Although no significant difference between treatments for *100-seed weight* and *harvest index* was observed in the parental lines, variation in the segregating populations allows high-yielding lines to become potential candidates for future improvement programmes for maintenance of yield under mild drought.

The reduction in final yield was possibly the combined result of stomatal closure and reduced leaf area which could reduce water loss, but also limit photosynthesis capacity, and hence carbon deposition in the seeds. *Stomatal conductance* is able to provide some indication of transpiration rates, nevertheless the relationship is not direct as transpiration in the current experiment will be effected by both soil pits in the glasshouse with vapour pressure deficit experienced by the droughted plot likely to be quite low, compared to a true field drought situation (Collinson *et al.,* 1997). In addition, transpiration rate is also suggested to be associated with the development of leaf area. For instance, high transpiration rates were shown to be the result of greater g_s in S19-3 but for Uniswa bigger leaf area was most likely to be the more important factor for higher overall rates of

transcription (Jorgensen *et al.,* 2011). However, for low rates of transpiration in LunT, Jorgensen *et al.* (2011) suggested that it was possibly due to the combined result of a moderate g_s and a small leaf area. Unlike other legumes, such as pea, chickpea and mungbean, Mwale *et al.* (2007) found that Bambara groundnut did not carry out a redistribution of dry matter during the pod filling stage as Bambara groundnut probably lacks important vegetative structures to store carbohydrates before redistributing assimilates to the pods. A decrease in seed yield in Bambara groundnut plants is believed to be contributed to by lower photosynthetic levels of plants during the pod filling stage due to mild drought stress (Mwale *et al.,* 2007).

4.4.2 Potential candidates for future programmes

Combining the responses of Bambara groundnut plants to mild drought stress, there are some lines in the segregating population that performed better in terms of both the ability to tolerate drought and also produce high seed weight per plant, which could potentially be selected as candidates for future breeding programmes. Using eight lines in the segregating population as examples, Table 4.6 shows a comparison of the lines for *100-seed weight, stomatal conductance* and *stomatal density.* Among the individual lines, i.e. L89, L5 and L101 produce higher yield under the current drought conditions while L41, L45 and L37 produced the lowest yield. The result shows that *100-seed weight* could be possibly affected by g_s and *stomatal density*, which is related to *leaf area*. Under the current drought conditions, L89 and L94 showed high g_s and moderate *stomatal density* (moderate leaf area) while L5 showed low g_s and low *stomatal density* (large leaf area), but both generated reasonably high *100-seed weight.* However, different responses were shown by L101 and L112 in which both showed high g_s and low *stomatal density* (large leaf area), but gave reasonably high *100-seed weight*. For L41 and L37, low *100-seed weight* is associated with both low g_s and high *stomatal density* (small leaf area) while L45 showed both moderate g_s and *stomatal density* (moderate leaf area).

Table 4.6 Comparison of potential candidates in the segregating population for the *100-seed weight, stomatal conductance* and *stomatal density* traits under drought (D) and irrigation (IR) conditions.

Line	100-seed weight (g)		g_s (mmol m^{-2} s^{-1})		Stomatal density (pores cm^{-2})	
	Drought	Irrigation	Drought	Irrigation	Drought	Irrigation
L89	81.89	89.42	269.3	584.4	8.6	8.9
L5	72.46	70.79	166.1	432.8	7.4	9.0
L101	69.42	64.19	274.1	581.1	7.0	6.9
L112	67.08	63.25	262.0	578.3	6.3	7.6
L94	63.07	71.19	261.8	617.8	12.7	11.2
L41	24.69	37.33	162.7	555.6	15.1	8.9
L45	27.22	28.77	224.9	530.8	11.0	13.7
L37	33.53	26.67	193.6	524.2	13.9	12.3
Population mean	49.24	53.55	220.5	541.8	11.6	10.1
Population s.d.	12.02	12.53	34.8	50.9	2.5	3.5

The final yield is relatively complex to determine due to the combined effects of g_s and *stomatal density*, both of which are related to leaf area. It is worth noting that apart from a genotypic effect, the yield is strongly affected by the environment and repeating the experiment elsewhere would probably give different results as well. However, based on the result, *100-seed weight* seems to be attributed to the *stomatal density* rather than g_s. The observation is also supported by the moderate and negative relationship between *stomatal density* and *100-seed weight* (r=-0.40, *p*<0.01), but not g_s, in the present study. In short, lines with high g_s accompanied with moderate or low *stomatal density* (moderate or large leaf area) could potentially result in higher yield in Bambara groundnut plants even under similar mild drought conditions.

The differences among the landraces in their response to drought stress are suggested to be related to their climatic and agro-ecological origins. For example, S19-3 from Namibia with a mean annual rainfall of 365 mm has faster rates of development which allows S19-3 to avoid terminal drought (Mwale *et*

al., 2007) and is an example of drought escape. Both of the two parental lines used for the population, DipC and Tiga Nicuru, are most likely to be more tolerant to drought than many landraces as both of them are derived from water limited countries, Botswana and Mali, whose average rainfall is 450 mm and 440 mm per year, respectively (Burgess, 2006; Pedercini *et al.,* 2012). While the headline figures for rainfall give an initial indication, a far more extensive analysis of patterns of rainfall, temperature and daylength during the planting season is needed. DipC showed significantly higher *100-seed weight* and larger leaf area, based on the measurement of the same leaves used for the stomatal count, than Tiga Nicuru ($p<0.05$). The differences between two parental lines in the drought treatment for a number of traits and their origins may suggest that some of their mechanisms for adaption to drought could be non-identical in the different landraces. Therefore, it could be possible to map and select for the best in the offspring for further research and breeding work.

Under glasshouse conditions, the responses of a Bambara groundnut F_5 segregating population to mild drought imposed at the early flowering stage were studied. *Stomatal conductance*, *100-seed weight*, *harvest index* and *stomatal density* could be potential criteria for breeding selection for drought tolerance. However, the relationship between the impact of the drought and final yield is not straightforward. Several measurements such as total leaf area, number of leaves, transpiration rate and photosynthetic level in plants would need to be carried out in order to establish a clearer relationship. As DipC is different from Tiga Nicuru in terms of plant morpho-physiological traits and possibly in adaptation to drought, variation is expected to be observed among the segregating population. Potential candidates that have higher yield characteristics and perform better than parental lines under drought stress could be selected for future breeding programmes.

Chapter 5: CONSTRUCTION OF A DArTseq GENETIC MAP IN BAMBARA GROUNDNUT

5.1   INTRODUCTION

5.1.1 DArTseq overview

In the early 2000s a relatively new molecular technique, known as Diversity Array Technology (DArT), was developed. DArT markers are widely used in construction of genetic linkage map, genetic diversity analysis and assessment of genetic structure of collections of germplasm in various crop species (DArT, 2012). Sohail *et al.* (2012) reported the development and utilisation of DArT markers for assessment of population structure and genetic diversity in *Aegilops tauschii.* Cruz *et al.* (2013) determined genetic diversity of the *Physaria* collection with DArT markers. In addition, Oliver *et al.* (2011) reported the first complete oat genetic linkage map and markers linked to domestication genes in tetraploid cultivated oat (*Avena sativa*) using DArT markers.

Two types of platforms are described in developing DArT markers, which are microarray-based DArT and the DArTseq platform (Cruz *et al.,* 2013). The details of DArT marker development which involves the use of a combination of restriction endonucleases for genome complexity reduction is described in Chapter 1.3.2.2. The DArTseq platform can generate two types of data; presence/absence dominant markers (0/1) and SNPs (DArT, 2013). The DArTseq platform is associated with the use of Next Generation Sequencing (NGS) for sequencing of the genomic representations, providing advantages over the microarray-based DArT which involves fluorescent labelling of representations and hybridisation to DArT microarrays (DArT, 2013). Both microarray-based DArT and DArTseq platforms have approximately the same development and application costs, however, the higher number of markers produced from the DArTseq platform (up to 10-fold) gives a lower cost per datapoint than microarray-based DArT (Cruz *et al.,* 2013). Thus, DArTseq is

suggested for high-throughput work, including high resolution mapping and detailed genetic dissection of traits (Cruz *et al.,* 2013; DArT, 2013). In addition, as most of the DArTseq platform uses the methylation sensitive restriction enzymes (*Pst*I), the distribution of DArTseq markers could reflect genomic methylation patterns and allow epigenetic variation to be detected (DArT, 2013).

The construction of genetic maps and QTL mapping using DNA markers can assist in marker-assisted selection based breeding. When markers are closely linked with the genes or QTLs controlling traits of interest and they are inherited together in the segregating offspring, the breeder can use the DNA markers to screen through the population at the seedling stage for plants carrying the desirable allele of genes or QTLs prior to cultivation, especially on a large scale (Collard and Mackill, 2008). Thus, instead of screening the plants based on the phenotype which may develop late in the plant life cycle and be difficult or expensive to measure, the selection of plants with favourable traits could be based on the genotype of a marker or flanking markers. This could improve the efficiency of a breeding program. In addition, the localisation of genes of interest on the genetic map could also lead to a better understanding of the genes controlling desired traits and hence provides information to breeders about which new genes could be introduced into cultivated materials for an improved genotype or enhanced landrace (Basu *et al.,* 2007a).

5.1.2 Regression mapping and maximum likelihood mapping

The computer software, JoinMap v4.1 (Ooijen *et al.,* 2006), was utilised in this study for the construction of a genetic map in a controlled cross F_3 Bambara groundnut population. The software allows two mapping approaches as calculation options: regression mapping and maximum likelihood mapping. The two mapping approaches adopt slightly different techniques and principles for mapping (Ooijen *et al.,* 2006). One fundamental problem with genetic mapping (particularly now with new high density genetic markers) is that the

theoretically possible arrangements of the map cannot be tested computationally. For this reason, algorithms have been developed which allow the number of tested combinations to be significantly reduced, while still tending to produce the more parsimonious solutions.

Regression mapping was first proposed by Stam (1993) in which the underlying principle involves the addition of markers sequentially into the map by using the most informative pair of markers as the starting point. This is defined as the pair of markers for which the greatest evidence exists (highest pairwise LOD score). A weighted least squares procedure (linear regression) was used in regression mapping to estimate the recombination fractions and, hence, map distances (Stam, 1993). There are three rounds of analysis for regression mapping. In the first round, the best position for each added marker is determined by calculating the goodness-of-fit for each possible map position for the new marker. 'Jump' refers to the measurement of normalised difference in the goodness-of-fit value before and after adding a marker. A large jump indicates that the added marker has a poor fit in the map and thus needs to be removed. When a rapid reduction of goodness-of-fit for all possible positions of the additional marker (a large 'jump') or negative genetic distances between markers are obtained, the marker is removed from the map. Local order of added markers in the map is tested after the addition of each marker by 'ripple'. This calculates the likelihoods based upon testing of the best position of the added marker and the surrounding two markers. Ripple helps to avoid the map developing local minima in the overall likelihood which are actually not the best solution. If the ripple produces a more likely order, it is adopted. The mapping procedure is repeated and continued until all markers in round 1 have been tested.

Following the first round of mapping, the order of the accepted markers is fixed and the removed markers are re-tested in the second round. The jump threshold is unchanged and high jump markers are removed again. However,

sometimes the development of the complete Round 1 map can allow the mapping of some of the originally rejected markers under the same stringency of conditions. As such, Round 2 is the map in which there is limited conflicting data The third round of analysis incorporates all markers within a grouping node into the map, regardless the thresholds for reduction of goodness-of-fit or negative distances. Therefore, Ooijen *et al.* (2006) suggested that the map generated from third round is not the preferred map as there are questions about the quality due to conflicting marker data.

The speed of regression mapping progressively slows as the number of potential marker in a linkage group increases, with 50 markers being near the limit of what can be handled within a reasonable time. As a result, a new approach for mapping, the maximum likelihood mapping approach, was introduced by Jansen *et al.* (2001). The maximum likelihood mapping approach employs three techniques to locate the markers and calculate their distances: simulated annealing, Gibbs sampling and spatial sampling (Jansen *et al.,* 2001). Simulated annealing is used to estimate the best position for the markers based on the maximum likelihood or the recombination frequencies. However, one linkage group may be divided into two or more groups when simulated annealing is used, especially for dense maps where markers contain typing errors. Thus, through the use of spatial sampling to obtain a framework map with a few of the selected markers in the first stage, the problem in simulated annealing can be overcome as the framework map can be adopted as the basis for the construction of the map for all the markers. In addition, Gibbs sampling is used to estimate the recombination frequencies that are used for likelihoods calculations, given the map order. As the expected numbers of recombinants obtained whenever Gibbs sampling is used will vary, a new round of simulated annealing is applied to construct a new map with a (hopefully) improved map order for all markers based on the new recombination frequencies. These two techniques work in sequential order one after another in a repetitive cycle until

no further progress is achieved. Jansen *et al.* (2001) suggested that three to four repeated cycles were sufficient to construct the final map.

5.1.3 Genetic linkage mapping in Bambara groundnut

Bambara groundnut genotypes with desirable traits such as high yield, large seed size, early maturity and bunched morphology types have been selected throughout the period of cultivation of the crop by farmers. With the development of artificial hybridisation, an improved cultivar with a combination of traits of interest that cannot be found in one landrace or single pure line could be developed (Massawe *et al.,* 2005). In addition, by crossing two accessions with contrasting desirable traits, individuals with variation in the desirable traits inherited from the parents could be obtained in the segregating F_2 population, allowing a genetic map to be constructed and the identification of molecular markers that are closely linked to genes controlling both qualitative and quantitative traits for marker-assisted breeding.

In 2007, with the objective of constructing an initial genetic map, a successful controlled cross was reported between an ancestral wild type (VSSP11) and the domesticated form (DipC) of Bambara groundnut on the basis of differences observed in growth habit, maturity and yield performance (Basu *et al.,* 2007a; Basu *et al.,* 2007b). The genetic map was constructed based on an F_2 population which has the advantage of having heterozygous individuals that provide the opportunity to evaluate the effects of additive and dominant gene action at a specific locus (Collard and Mackill, 2008). As a population size of 50 to 250 individuals is usually suggested (Ferraira *et al.,* 2006), a set of 98 individuals from the segregating F_2 population were used to construct the map as the population size contributes to the resolution of the genetic map and the ability to determine marker order (Basu *et al.,* 2007b). Extensive polymorphism was observed between the individuals in the segregating population facilitating the construction of the genetic map. According to Basu *et al.* (2007b), 20

linkage groups were identified using 67 AFLP markers and one SSR marker in a total length of 516 cM, with the inter-markers distance varying from 4.7 cM to 32 cM.

In addition to the generation of a genetic linkage map, the same population was studied for the inheritance of a number of plant morphological traits, such as *internode length*, *number of stems per plant* and *days to flowering*. The major difference observed between VSSP11 and DipC was reported to be growth habit, in which VSSP11 has a spreading habit (long internode length with low stem number) while the opposite characteristic is obtained in DipC leading to a 'bunched' morphology (Basu *et al.,* 2007a). The F_1 hybrid is a spreading type, quite similar to the wild parent, VSSP11, but with more leaves and pods than VSSP11. A spectrum of variation was observed within the F_2 population (Basu *et al.,* 2007a). With the existence of a genetic map, the relevant molecular markers could be used to assist in the identification of quantitative trait loci (QTL) controlling the differences observed between the plants. Basu *et al.* (2007a) reported the localisation of four QTLs that contribute to *seed weight*, *specific leaf area*, *number of stems per plant* and ΔC^{13} through the use of the interspecific genetic map.

Ahmad (2012) also reported the construction of the first intraspecific genetic map using a F_3 segregating population derived from two domesticated landraces DipC x Tiga Nicuru. The intraspecific map covered 608.6 cM in 21 linkage groups using 29 SSR and 209 microarray-based DArT markers, with marker-marker distances ranging between 0 cM and 10.1 cM. QTL mapping for the phenotypic variation observed within the controlled cross of DipC and Tiga Nicuru was also conducted by Ahmad (2012). A major QTL contributing to *internode length* was mapped on linkage group 4 (LG4) with LOD values of 7.9 at a distance of 3.0 cM from marker bgPabg-596988. Another significant QTL contributing to *peduncle length* was found to map on LG4 as well with LOD values of 9.7 with the nearest marker being bgPt-423527 at 2.4 cM on LG4.

In the present study, an improved genetic linkage map in the F_3 segregating population derived from a cross between DipC and Tiga Nicuru is attempted by adding DArT dominant markers and SNPs markers onto pre-existing genetic map (Ahmad, 2012). Given the slower speeds expected from the regression mapping approach when using large marker numbers, a combination of regression mapping and maximum likelihood mapping are used and compared in this study to obtain the optimal position of markers in the genetic map.

5.2   METHODS AND MATERIALS

5.2.1 List of molecular markers

The construction of an initial genetic linkage map was reported by Ahmad (2012).   For the generation of markers, the genomic representations were prepared using 73 individuals from the F_3 segregating population derived from a cross between DipC and Tiga Nicuru as described in Ahmad (2012).   A total of 3,670 classical dominant DArT (presence/absence) markers and an additional 2993 bi-allelic SNP markers were developed by Diversity Arrays Technology Pty. Ltd (Yarralumla, Austrialia) in the current study using the DArT seq platform.   In addition, 210 microarray-based DArT and 33 SSR markers used by Ahmad (2012) were also included into the present study for construction of a higher density genetic map.

5.2.2 Coding and selection of markers

Dominant DArT markers for each individual in the F_3 segregating population were genotype coded either as (a,c) or (d,b) based on presence or absence of hybridisation in the two parental lines: DipC and Tiga Nicuru (Table 5.1).   When presence, or absence, of hybridisation for both parental lines were observed ('1':'1'; '0':'0'), they were considered as monomorphic markers or unreliable and were excluded from the analysis.

Table 5.1 Conversion of genotype code for dominant DArT markers.

DipC	Tiga Nicuru	Genotype code	Conversion	
absence (0)	presence (1)	(a,c)	0 -> a	1-> c
presence (1)	absence (0)	(d,b)	0 -> b	1-> d

Each SNP marker was reported as two lines in the Excel sheet: 'variant' line and 'reference' line. Based on the scoring pattern in both parental lines, co-dominant SNP markers were assigned as 'a', 'h' and 'b' as appropriate in each individual with data (Table 5.2). SNP markers that had identical scores in the two parental lines were eliminated, irrespective of whether there was evidence for segregation in the offspring.

Table 5.2 Conversion of genotype code in SNP markers.

	DipC	Tiga Nicuru	Conversion		
			progeny 1	progeny 2	progeny 3
'Variant'	0	1	1	0	1
'Reference'	1	0	1	1	0
Genotype code			**h**	**a**	**b**
'Variant'	1	0	1	0	1
'Reference'	0	1	1	1	0
Genotype code			**h**	**b**	**a**

Following the conversion of markers, polymorphic markers with ≤5 missing values across the individuals in F_3 segregating population were selected for linkage analysis. Secondly, based on the ratio of alleles (presence:absence), SNPs markers with ratio less than 25% and more than 75% and dominant DArT markers with ratio less than 37.5% and more than 62.5% were excluded from the analysis based on the expected segregation patterns of 3:2:3 and 5:3 ratio in order to remove poorly scored markers.

5.2.3 Linkage analysis

A total of 1,361 markers were used for construction of the initial genetic linkage map using JoinMap v4.1 (Ooijen *et al.,* 2006). As per the JoinMap v4.1 instruction manual, the data was arranged in an Excel (.xlsx) file, copied and pasted into the JoinMap4.1 software spreadsheet to conduct the linkage analysis. Of 73 individuals in the F_3 segregating population, 71 individuals were subjected to linkage analysis as two individuals: L19 and L54 contained more than 5%

missing data and were excluded from the analysis. The population type was entered as 'Rlx; x:3' for the F_3 segregating population. The grouping of markers was set between LOD 2.0 and 10.0 with a step of 1.0 and the Independence LOD option adopted. Once the grouping trees were generated, the grouping and ordering of the markers for each linkage group were established using the maximum likelihood (ML) mapping approach of JoinMap4.1 with grouping at LOD>6.0. After creation of grouping nodes based on the initial splitting of markers into linkage groups, the initial ML maps were generated. Markers were manually removed when two adjacent markers were too closely located (1-3 cM) through the use of the information in the 'plausible positions' tab. In addition, the markers that showed double crossover events between two neighbouring markers within a distance of between 1 and 3 cM were also removed. When the number of markers in a linkage group reached approximately 80 or below, the regression mapping approach with a recombination fraction ≤4.0, ripple value=1, jump in goodness-of-fit threshold=5 under a Haldane mapping function was applied. Through the alternate use of the maximum likelihood mapping approach and then the regression mapping approach, a framework map consisting of dominant DArT markers, SNP markers, microarray-based DArT markers and SSR markers, which were spaced at approximately 5 cM, was obtained. This reiterative process of removal based on graphical genotyping and stress and fit testing allow a high quality framework map to be generated for QTL analysis and further development.

5.3   RESULTS

5.3.1 The selection of polymorphic markers

Of the 3,670 dominant DArT markers developed, 1,859 dominant DArT markers (50.7%) were polymorphic and able to distinguish between the parental alleles across the individuals in the F_3 segregating population.   Following a stringent selection of markers, 282 dominant DArT markers which represent the best of those available were selected based on defined missing value (≤5).   SNP markers were also filtered for missing values in addition to being polymorphic, as a result, 1,014 out of 2,993 markers (33.9%) were identified to be polymorphic and of good quality for linkage analysis.   Furthermore, 32 SSR markers and 33 microarray-based DArT markers out of 210 (15.7%) were also selected for linkage analysis.   As a result, a total of 1,361 markers were pasted into JoinMap v4.1 for construction of the genetic map.

5.3.2   The segregation distortion of markers

Markers were analysed using a Chi-square test in JoinMap4.1 against the expected segregation patterns for their segregation pattern and also potential segregation distortion at a significance level of $p < 0.05$.   The result showed that 1,043 markers did not deviate significantly from the expected segregation ratio of 3:2:3 for co-dominant markers and 3:5 for dominant markers in F_3 segregating population.   However, 318 markers (23.4%) tested significant for segregation distortion ($p > 0.05$).   The highest segregation distortion rates were found in SSR markers (28.1%), followed by SNP markers (25.5%), dominant DArT markers (16.3%) and microarray-based DArT (12.1%).

5.3.3   Linkage group and markers distribution

A group of 1,361 markers (282 dominant DArT, 1014 SNP, 32 SSR and 33 microarray-based DArT) were subjected to linkage analysis and only 18 markers could not be mapped.   Grouping analysis at LOD>6.0 resulted into 11

linkage groups (LG) with 171 spaced markers covering 1,341.3 cM of Bambara groundnut genome (in a final mapping interaction based on regression mapping and the Haldane mapping function). The markers were distributed evenly over 11 LGs with an average of 15.5 markers in each LG. The highest number of markers was observed in LG5 (19) whereas the lowest number of markers was observed in LG9 (11). In addition, as the map was developed as a framework map, an average distance of 7.8 cM between two adjacent markers across all LG was achieved. The shortest distance between adjacent markers of 2.7 cM was found in LG8 whereas the longest distance of 33.0 cM were found on LG11.

Among the linkage groups, LG5 with 19 markers (15 SNP, 3 dominant DArT and 1 SSR) was the longest group covering 176.8 cM, followed by LG2 and LG1 with sizes of 173.2 cM and 149.4 cM, respectively. The shortest group, LG9, was 76.4 cM with 11 markers (9 SNP, 1 dominant DArT, 1 microarray-based DArT). The number of markers, marker distance and corresponding LG are presented graphically in Figure 5.1.

Figure 5.1 Genetic linkage map of Bambara groundnut F₃ segregating population constructed using dominant DArT, SNPs, SSR and microarray-based DArT markers. Left: name of the markers. Right: positions of markers (cM).

Furthermore, the distribution of each type of marker across each LG, marker density and average marker interval are also summarised in Table 5.3. Of 171 total markers present in the framework map, the results showed that 124 SNP markers (72.5%), followed by dominant DArT (17.5%), microarray-based DArT (7%) and SSR (2.92%).

Table 5.3 The distribution of dominant DArT, SNPs, SSR and microarray-based DArT markers across each LG for the framework genetic map in the F_3 segregating population of Bambara groundnut.

Linkage group (LG)	Length of LG (cM)	Dominant DArT	SNPs	SSR	Microarray-based DArT	Total number of markers	Average marker interval (cM)
1	149.4	3	8	1	5	17	8.8
2	173.2	0	13	1	4	18	9.6
3	90.6	1	15	0	0	16	5.7
4	101.2	3	14	0	0	17	6.0
5	176.8	3	15	1	0	19	9.3
6	93.4	6	8	0	0	14	6.7
7	103.4	2	9	1	0	12	8.6
8	117.9	5	10	0	0	15	7.9
9	76.4	1	9	0	1	11	6.9
10	134.2	3	13	1	1	18	7.5
11	124.8	3	10	0	1	14	8.9
Grand total	1341.3	30	124	5	12	171	85.8
Mean	121.9	2.7	11.3	0.5	1.1	15.5	7.8

In addition to regression mapping, maximum likelihood mapping was also employed to construct the genetic map. For the maximum likelihood mapping, the total map size was 1,723.9 cM, with an average spacing of 9.95 cM between adjacent markers for all the LG. Moreover, marker location and order were generally similar in all the LGs except for LG 1, 3 and 5 which showed one to two inverted orders of the markers compared to regression mapping.

5.4   DISCUSSION

5.4.1 Selection of molecular markers for genetic linkage mapping

The construction of the genetic linkage map in an F_3 segregating population of Bambara groundnut is following up on a study by Ahmad (2012) who utilised both SSR and microarray-based DArT markers. In the present study, more marker types were introduced, namely dominant DArT markers and SNPs markers which are produced using DArTseq technology, in order to generate LGs that have complete coverage of the genome. The use of various marker types, both dominant and co-dominant markers, are believed to be complementary to each other and thus produce a genetic map with good genome coverage.

The deviation of the observed segregation ratio in the segregating population from the expected Mendelian segregation ratio is known as segregation of distortion (Semagn *et al.,* 2006). Marker segregation distortion is common in mapping studies (Causse *et al.,* 1994; Yu *et al.,* 2011; Yang *et al.,* 2011). For example, 182 out of 466 polymorphic DArT markers (39.1%) were reported to be distorted when DArT Array markers were utilised for constructing the first map of pigeon pea (Yang *et al.,* 2011). A lower distortion rate was reported in cowpea in which 410 out of 1,375 SNPs (29.8%) deviating from the expected ratio (Muchero *et al.,* 2009). In the present study, the distortion rate of the various markers types obtained for Bambara groundnut was 28.1% (SSR), 25.5% (SNP), 16.3% (dominant DArT) and 12.1% (microarray-based DArT). This finding is in agreement with Ahmad (2012) who used both microarray-based DArT and SSR in generating the first genetic linkage analysis for Bambara groundnut. That author reported that 69 out of 210 microarray-based DArT (33%) exhibited segregation distortion whereas SSR markers had a distortion rate of 24% when the same F_3 segregating population was used in linkage analysis.

The occurrence of segregation distortion can be due to technical issues such as sample size, genotyping and missing data (Boopathi, 2012) or biological factors including chromosome rearrangements, incompatible genes (Semagn *et al.,* 2006), alleles inducing gamete or zygotic selection (Lu *et al.,* 2000) parental reproductive differences (Blanco *et al.,* 1998) and also possibly the use of wild relatives as parental lines (Yang *et al.,* 2011). Although segregation distortion is always an issue when conducting genetic linkage analysis, the effects of including the distorted markers in the final genetic maps seem to be contradictory between different studies.

Segregation distortion is reported to have impacts on linkage distances in several linkage maps (Wu *et al.,* 2010). Two genetic linkage maps were produced using F_2 segregating populations derived from rice inter-subspecific crosses, TNG67/TCS10 and TNG67/TCS17, respectively (Wu *et al.,* 2010). The authors reported that a longer linkage length of 1,481.6 cM for TNG67/TCS10 than 1,267.4 cM for TNG67/TCS17 in rice was most likely to be related to distribution of more severe distorted markers at more chromosome regions in TNG67/TCS10 (Wu *et al.,* 2010). This is in agreement with Knox and Ellis (2002) who also reported an increased linkage map length in an F_2 population in pea as a result of segregation distortion due to excess heterozygosity. However, some authors argued that the effect of segregation distortion on both marker order and map length by simulation was minimal (Hackett and Broadfoot, 2003). The generation of a genetic linkage map is commonly associated with QTL mapping. The introduction of distorted markers in a genetic linkage analysis was suggested to increase the marker coverage of the genome and lead to identification of more QTLs in such regions (Wang *et al.,* 2005). In addition, Zhang *et al.* (2010) also reported that the effect of distorted markers on the genetic map prior to QTL analysis was associated with the distances between distorted markers and QTL. If the distorted marker is not closely linked to the QTL, it will have no significance impacts on QTL analysis (Zhang *et al.,* 2010).

Thus, it is suggested that markers exhibiting segregation distortion at a significance level of 5% should be included in the construction of the genetic linkage map to reduce the frequency of false positives (Doucleff *et al.,* 2004). In the present study, distorted markers were included in the linkage analysis in order to avoid losing a number of markers which might be linked to the traits of interest. In addition, in terms of marker type, calculations of recombinant values using co-dominant markers such as SSR markers are suggested to be less affected by segregation distortion than dominant markers (Lorieux *et al.,* 1995).

A total of 3,670 dominant DArT markers, 2,993 SNP markers, 210 microarray-based DArT and 33 SSR were generated in the present study prior to selection. However, like any other analysis, genetic linkage analysis is susceptible to errors as well. In addition to segregation distortion, missing data and genotyping errors are also the major concerns for the construction of genetic maps, with the first priority to generate a genetic map with good genome coverage based on the best quality data available. Individuals with many missing data points are unable to contribute to mapping calculations and should be eliminated from the analysis (Ooijen *et al.,* 2006). Furthermore, missing data is reported to be a causal agent for incorrect marker order during dense genetic map construction where a single recombination event could determine the relative order of two closely linked markers (Semagn *et al.,* 2006). Thus, in the present study individuals such as L19 and L54 containing more than 5% of missing value were eliminated from the analysis as they were likely to introduce significant noise into the mapping process. In addition, markers with missing data of ≥ 5 across the individuals in the F_3 segregating population (<66 data points out of 71 individuals) were also removed in order to minimise the effect of missing data on the final genetic mapping.

Genotyping errors can result in inaccurate estimations of the map distance and also produce incorrect marker orders (Cheema and Dicks, 2009). In order to minimise the impact of errors on the genetic map, the detection of

genotyping errors by searching for double recombinants over a short distance (up to 5 cM) in the genetic map is commonly practised (Cheema and Dicks, 2009). In the present study, markers with more than one to two double recombinants over 5 cM were removed from the genetic map to enhance the accuracy of mapping markers in the correct order. The problem of having genotyping errors increases when the marker density is greater as the errors can lead to incorrect ordering (Cheema and Dicks, 2009).

The molecular markers used to conduct genetic linkage analysis were filtered through several criteria, such as level of missing data and genotyping errors, in order to increase the data quality for mapping. It is worth noting that the stringent criteria used in this study for marker selection could have deleted some genuine markers. As a result, 1,361 out of 6,906 possible markers were selected to construct the genetic map for the F_3 segregating population in Bambara groundnut.

5.4.2 Framework linkage mapping

A framework map consisting of 171 markers with an average spacing of 7.8 cM between neighbouring markers for the F_3 segregating population in Bambara groundnut was constructed. The marker intervals obtained in the present study was slightly higher than previous reports for Bambara groundnut, for example, a mean value of 3 cM between two consecutive markers in Ahmad (2012) was obtained, although genome coverage was not comprehensive. However, a framework map consisting of evenly spaced markers (i.e. 10 cM) and potentially 100-200 markers is generally recommended for use in QTL analysis (Boopathi, 2012). This is supported by Darvasi *et al.* (1993) who subjected a backcross population to a simulation study and found that irrespective of genetic effect and population size a marker spacing of 10 cM is sufficiently precise QTL detection. In addition, 97% of the RNA-based markers used to construct a

framework genetic map for *Miscanthus sinensis* showed a marker interval of 10 cM (Swaminathan *et al.,* 2012).

It has been reported that a genetic map with higher marker density could improve the precision of localisation of the QTL as the chances of having QTL tightly linked with markers are slightly higher (Stange *et al.,* 2013). However for Bambara groundnut which has a limited established genetic map, the production of a high quality and robust framework map is essential to provide accurate marker order using the best quality markers which can then be fixed and serve as the backbone for other individual maps in Bambara groundnut and for saturation mapping using large numbers of dominant DArT markers.

As the sequence of the Bambara groundnut genome is not yet available, marker order dictated by the physical map length is not known and may also vary between individuals, if translocations and other rearrangements between individuals exist within the species. However, by employing two mapping approaches maps of the expected 11 linkage groups in Bambara groundnut were generated and the marker ordering of each linkage group, based on the best marker data, can be compared and used for further work.

Maximum likelihood mapping was first applied for each linkage group analysis as regression mapping is very time consuming when more than 50 markers are subjected to analysis (Ooijen *et al.,* 2006). Maximum likelihood mapping as implemented in JoinMap4.1 provides a function tab 'plausible positions' which reveals other potential locations where markers might be acceptable and also gives a good indication of the amount of uncertainty in the map concerning the positions of the markers in the map. The markers which are close to each other will often appear to be interchangeable as the amount of evidence present in the dataset for the adopted order may be limited. However when the markers are located further apart, they are likely to be 100% fixed at their estimated position (Ooijen *et al.,* 2006). Thus, the higher the value obtained in plausible positions, the higher confidence that the marker is located

in the chosen position and the more confidence in the relatively order of markers in the genetic map (Ooijen *et al.,* 2006). Theoretically, maximum likelihood mapping has a number of advantages, such as greater robustness to missing data and the ability to find the most likely group wide marker order (De Keyser *et al.,* 2010). This mapping approach uses the Haldane mapping function that assumes that adjacent chromosomal recombination events exhibit no crossover interference in the F_3 segregating population (Ooijen *et al.,* 2006). Following genetic linkage analysis using maximum likelihood mapping, each linkage group is alternately mapped with regression mapping when the number of markers on the linkage group reaches 80 or below. During the construction of genetic linkage map using regression mapping approach in JoinMap4.1, it was observed that the genetic linkage maps were not able to be produced when insufficient linkage to other marker groups was detected. The observation is in agreement with De Keyser *et al.* (2010), suggesting that regression mapping approach allows the markers which are poorly fit to be removed.

In the present study, two mapping approaches for genetic linkage analysis were alternately applied until the marker location and order produced was similar between maps constructed with the two approaches. The practice is supported by Doligez *et al.* (2006) which revealed that only well-conserved marker order irrespective of mapping algorithms can be considered as genuine marker orders. In addition, the use of maximum likelihood followed by regression mapping was also adopted in loblolly pine (Martinez-Garcia *et al.,* 2013). A high density consensus linkage map was produced with the expected 12 linkage group for loblolly pine, covering 1,475.9 cM with 2,466 markers (Martinez-Garcia *et al.,* 2013).

As discussed earlier, the construction of a framework map with a marker interval of approximately 10 cM could minimise the impacts resulting from missing data, genotyping errors and segregation distortion, concentrating on developing a map based on the best available data. The marker order was

observed to be minimally effect by missing data and genotyping errors when a genetic map with 10 cM marker intervals was constructed (Hackett and Broadfoot, 2003). In addition, by applying two mapping approaches alternately a final map with a well-conserved marker order is established. The stable and consistent order of markers in the framework map is important for use in QTL analysis and also for the integration of other marker-types and the development of saturated genetic map. For example, to compare the genetic order in the wide cross map using SSR, DArT and AFLP markers for an F_2 segregating population derived from DipC and VSSP11 (non-domesticated accession) in Bambara groundnut (Basu *et al.,* 2007a; Ahmad, 2012).

Chapter 6: DEVELOPMENT OF A LINKAGE MAP FOR BAMBARA GROUNDNUT USING MAJOR RESOURCES DEVELOPED IN SOYBEAN

6.1   INTRODUCTION

6.1.1   Gene expression markers (GEMs)

The advanced state of development of gene expression microarrays allows the analysis of the transcriptome in a wide range of organisms. The differences in apparent gene expression between individuals could be due either to sequence polymorphism affecting hybridisation of the labelled probe to the target sequence or actual variation in the mRNA abundance of the gene of interest (or both). Variation in gene expression has been reported to be heritable and present often as a quantitatively distributed trait (Li and Burmeister, 2005). The differences in hybridisation signal strength for individual features on a microarray chip can be used to identify so-called 'expression quantitative trait loci' (eQTL) through treating the variation in signal hybridisation across the population as a quantitative trait for analysis. By determining the position of eQTL loci, trait QTL loci and the location of known or putative candidate genes associated with traits of interest within the controlled cross population, existing candidates can be evaluated and new candidates identified. The identification of candidate genes can be done either through direct mapping of markers to a genetic or physical location of a gene within a genome or through conserved synteny relationships across closely related species. For example, *Hls* for leaf shape in cowpea was determined through identification of syntenic loci in other legume species, *Medicago trunculata* and soybean (Pottorff *et al.,* 2012).

The heritable pattern of differences in hybridisation strength across the population can be affected by eQTL loci being either 'cis' or 'trans'. *Cis* effects refer to variation in gene expression across the population where the causal agent appears in or close to the structural gene represented by the chip feature. In contrast, *trans* effects reflect a pattern of gene expression variation across the

169

population resulting from the causal agent being located away from structural gene. An example of a *trans*-acting element would be a transcription factor which controls the expression of many downstream genes (Kliebenstein, 2009; Joosen *et al.,* 2009). The hybridisation pattern observed across the population is regulated by allelic variation resulting from sequence polymorphism in transcription factor gene, not due to differences between expression patterns of the structural genes themselves, whose expression is altered in level by inheritance of the different alleles of the transcription factor that acts upon the structural gene and alters expression levels. Therefore, *trans*-eQTL will not be mapped with the physical position of affected structural genes, but with the location of the transcription factor gene polymorphism. This potentially allows *trans*-eQTL representing master regulatory loci to be located for the traits. Such loci will have a number of structural genes co-located to the genetic position of the regulatory genes.

Variation of gene expression can be examined using several technologies such as reverse transcription polymerase chain reaction (RT-PCR), gene-based microarrays and next generation sequencing (NGS; Druka *et al.,* 2010). Recently, gene-based microarrays have been widely adopted to exploit novel marker information for comprehensive QTL and eQTL studies. Winzeler *et al.* (1998) first proposed the hybridisation of genomic DNA to oligonucleotide microarrays to identify DNA sequence polymorphisms in haploid yeast. The approach was then extended by hybridising cRNA instead of genomic DNA to microarrays in order to obtain both phenotypic (gene expression) and genotypic (marker) data for linkage mapping, simultaneously (Ronald *et al.,* 2005). The identification of genetic polymorphism across a population from gene expression microarrays enables the production of reliable genetic markers to construct a framework map from the same dataset that is used for both map construction and eQTL mapping (Druka *et al.,* 2010).

The markers produced from gene expression microarrays have been classified into several groups based on the different principles underlying selection of the markers (Table 6.1). Gene expression markers (GEMs) refer to sequence polymorphisms which lead to a difference in hybridisation signal strength when cRNA is hybridised to GeneChip arrays. For instance, the majority of GEMs that were developed in a *Brassica* eQTL experiment had been selected based on differences in hybridisation signal intensity in the parental plants. These were likely to have resulted from (but were not unequivocally proven to be) sequence polymorphisms which effected binding of the test RNA samples to the microarray targets (Hammond *et al.,* 2011). In contrast, Gupta *et al.* (2013) defined the resulting DNA polymorphisms as single feature polymorphisms (SFPs) when DNA instead of cRNA is used for hybridisation.

Table 6.1 The definitions of different categories of markers produced using microarrays designed for analysing gene expression.

Markers	Definition	Reference
Gene expression markers (GEMs)	Sequence polymorphisms that represent the difference in hybridisation signal strength obtained from hybridising RNA to microarrays.	Hammond et al. (2011)
Expression level polymorphisms (ELPs)	The identification of differences in expression level (transcript abundance) between samples.	West *et al.* (2006); Calvino *et al.* (2009)
Single feature polymorphisms (SFPs)	Sequence polymorphisms obtained from hybridising genomic DNA to microarrays.	Gupta *et al.* (2013)
Transcript derived markers (TDMs)	Represent both GEMs and ELPs.	Potokina *et al.* (2007)

If two parents of a segregating population that exhibit bimodal distribution of hybridisation signal in the segregating progeny differ in the expression level (transcript abundance), the difference is considered an expression level polymorphism (ELPs; West *et al.,* 2006; Calvino *et al.,* 2009). Using the Affymetrix microarray GeneChip system, the expression level was calculated as the average value of 11 probe-pairs (West *et al.,* 2006). Using

Arabidopsis as the study organism, West *et al.* (2006) demonstrated the development of ELPs based on gene expression measurements from the Affymetrix ATH1 GeneChips. A total of 1,431 genes with a two-fold or higher differential expression ratio between the two parental genotypes, Bay-0 and Sha were identified. Subsequently, a 'gap' value was calculated for each gene by dividing the minimum expression value of the higher expression allele with the maximum expression value of the lower expression allele. Of 1431 genes, a subset of 324 genes detected in an *Arabidopsis* RIL population were identified as potential ELPs as they showed a gap of ≥1.0 (no overlapping distributions) among segregating progenies. Following the development of ELPs, a genetic linkage map covering 393 cM with ELPs and microsatellite markers was constructed and the map order was shown to be consistent with the gene order predicted from the genomic sequence of the *Col*-0 accession (West *et al.*, 2006). In addition, the authors also discovered that most of the ELPs were the result of *cis*-regulatory polymorphisms.

GEMs are considered as robust genetic markers that are able to identify eQTLs. For instance, Hammond *et al.* (2011) showed the identification of *cis*-elements and *trans*-eQTL regulatory hotspots that regulated low phosphorus availability in *Brassica rapa* using a genetic map constructed from 125 GEMs. In addition, a total of 1,596 transcript derived markers (TDMs), with no separation of GEMs and ELPs, derived from two commercial varieties of barley (Steptoe x Morex) were identified in barley for use in genetic mapping and eQTL analysis of 16,000 genes (Potokina *et al.*, 2007). The authors reported that 23,738 significant eQTLs representing 12,987 genes were identified and more than 50% of them were *cis*-eQTLs. However, GEMs and ELPs are identified at the transcription level so they could be highly affected by environmental factors, resulting in irreproducibility and dissimilarity of performance under different conditions and when different sets of tissues are used (West *et al.*, 2006).

The potential for analysing less intensively studied species using GEMs was suggested when Calvino *et al.* (2009) reported the hybridisation of RNA derived from stems of grain and sweet sorghum onto the sugarcane Affymetrix GeneChip, followed by identification of both GEMs and ELPs linked to high sugar content. As a result, 154 genes differentially expressed between grain and sweet sorghum were reported to be related to sugar and cell wall metabolism (Calvino *et al.,* 2009). The combination of cross-species hybridisation and genetical genomics approaches, combining gene mapping with gene expression analysis, is initially applied here in Bambara groundnut in order to produce a genetic map for use in QTL and eQTL studies.

6.1.2 Integration of linkage maps in crops

Genetic linkage maps serve as a major resource to study genome organisation, gene space, position of coding regions and genome evolutionary relationships. A complete genetic linkage map can also be used to identify quantitative trait loci (QTLs) for subsequent use, such as in marker-assisted selection (MAS) breeding or gene cloning (Wu *et al.,* 2000). The advanced development of molecular genetics, by which large numbers of molecular markers such as RFLP, SSR and SNPs can be produced, has resulted in the development of various versions of genetic maps in the same crop that are often developed by different research groups using a range of mapping populations. As a result, a representative integrated map for a single species is of interest to support genome studies, provide tools for high resolution mapping as well as to assist the correct assignment and orientation of sequences to the respective chromosome locations (Stam, 1993; Wang *et al.,* 2011). The development of integrated linkage maps have been reported in numerous crop species, such as *Brassica rapa* (Wang *et al.,* 2011), sugarcane (Garcia *et al.,* 2006), *Populus deltoides* (Wu *et al.,* 2000), maize (Cone *et al.,* 2002), soybean (Choi *et al.,*

2007), barley (Wenzl *et al.,* 2006), cowpea (Muchero *et al.,* 2009) and peanut (Hong *et al.,* 2010).

Several studies have shown the application of 'two-way pseudo testcross' mapping approach to align two parental genetic maps such as *Populus deltoides* and *Calluna vulgaris.* In *P. deltoids*, the authors reported the use of heteroduplex markers (intercross markers) that were heterozygous in both parents to combine two parent-specific genetic maps, resulting in an integrated map that covered 2,927 cM with 19 linkage groups (Wu *et al.,* 2000). Wu *et al.* (2000) proposed that the first step of integrating genetic maps was to create a framework map for each parent line. For example, testcross markers segregating in parent I-63 and heteroduplex markers were used to construct the first framework map for I-63, followed by a second framework map for parent C-135 using the testcross markers segregating in parent C-135 and the same heteroduplex markers. The two genetic linkage maps were then combined based on the relative positions of the heteroduplex markers in the framework map (Wu *et al.,* 2000). A similar approach is also presented in Behrend *et al.* (2013) in which separate maps were constructed in paternal and maternal line of *C. vulgaris* respectively, followed by integration of the two maps using biparental markers in both parental maps. As a result, an integrated map spanning 601.1 cM total map distance across nine linkage groups was reported in *C. vulgaris*, which is a perennial shrub (Behrend *et al.,* 2013)*.*

In addition, early attempts to integrate genetic maps by pooling genetic information from different mapping populations, followed by log-likelihood statistical mapping algorithms have also been reported (Beavis *et al.,* 1991). However, potential problems arise due to the use of different types of mapping populations, missing data and limited numbers of linking loci between maps. Stam (1993) suggested that common markers shared between individual genetics maps that are derived from different mapping populations or using different marker systems are the key problem for map integration. For instance,

in *Brassica* an effort was made to integrate linkage maps derived from different mapping populations based on shared markers, however, a low resolution integrated map was obtained due to the low number of shared markers (Hu *et al.,* 1998). Recently, the first genome wide integration of *Brassica* genetic maps using three extensively studied *B. napus* doubled haploid mapping populations was established as they shared a high number of common markers (Wang *et al.,* 2011). The approach used to integrate *B. napus* genetic maps involved the development of population-specific consolidated maps from each mapping population, followed by development of a skeleton map which consisted of only representative markers and common markers for use in subsequent map integration (Figure 6.1). Three skeleton maps were combined using JoinMap v4.0, as a result, an integrated map with 5,162 genetic markers representing 2,196 loci and a total genetic map length of 1,792 cM was produced. Wang *et al.* (2011) also showed that the marker density of the integrated map in *B. napus* increased at least three-fold compared to the original maps with one locus every 0.82 cM, corresponding to 515 kbp being obtained. Thus, a high-density and high-resolution integrated genetic map potentially provides 'bridges' connecting different mapping populations and also serves as a resources to study closely-related but less researched crop species.
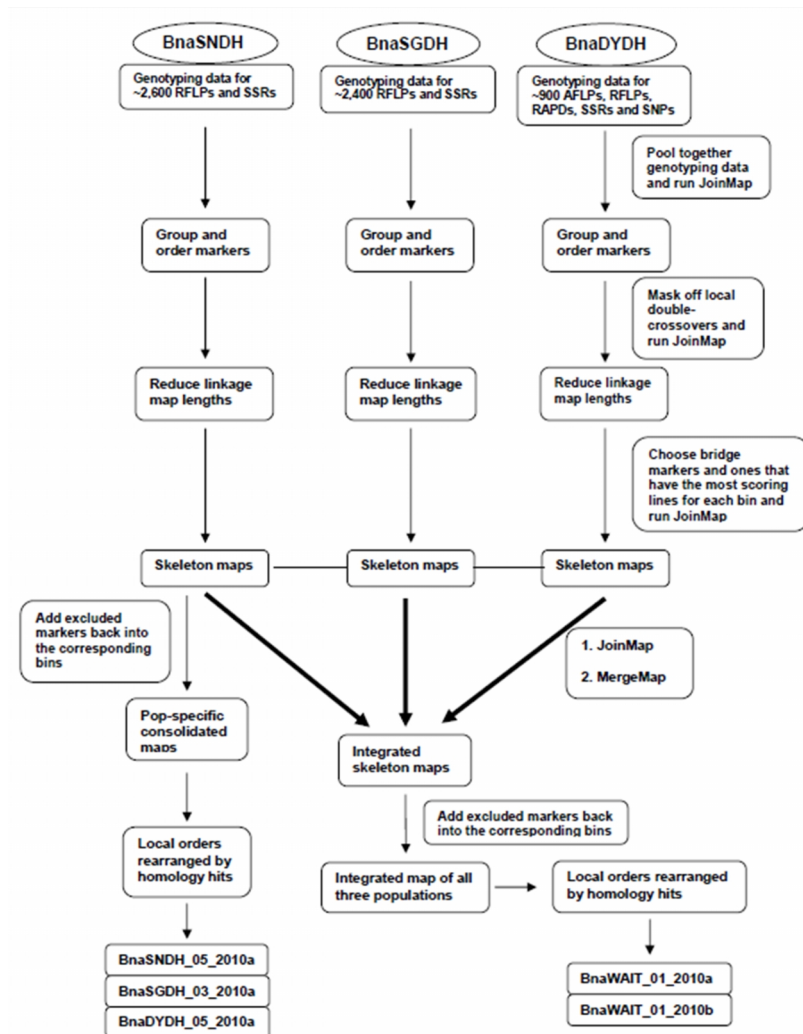
Figure 6.1 The automated pipeline indicating the process of the integration of the genetic linkage maps of *B. napus* using doubled haploid populations (Wang *et al.*, 2011).

For legume crop species, a new integrated linkage map of soybean was first reported by Song *et al.* (2004). The authors showed that by combining five genetic linkage maps using JoinMap v3.0, an integrated linkage map that covered 2,523.6 cM (Kosambi) across 20 linkage groups with a total of 1,849 markers (including 1,015 SSRs, 709 RFLPs, 73 RAPDs, 24 classical traits, six AFLPs, ten isozymes and 12 others) was produced. The present integrated map of soybean was then improved by adding SNPs that were discovered through resequencing of sequence-tagged sites (STSs) developed from ESTs sequences (Choi *et al.* 2007). The integration of two maps allowed additional SNP markers

to effectively fill in the 'gaps' (marker intervals) between pre-existing markers in the SSR-based map developed by Song *et al.* (2004). For examples, 291 genes were mapped into 72 'gaps' with gap interval distances between 5 cM to 10 cM and a further 111 genes filled up 19 'gaps' that had more than 10 cM gap distances between adjacent markers (Choi *et al.* 2007). The integration of different genetic linkage maps potentially increases map resolution and density, providing an important resource for QTL studies and map-based cloning.

Computing software, such as JoinMap and MergeMap, are widely used to extract all available information from individual datasets, assign markers into their respective linkage groups and to estimate the marker order as well as the genetic distances. JoinMap v4.0 which employs regression mapping or maximum likelihood mapping approaches, allows the search for the optimal position of markers in the genetic map. MergeMap uses directed acyclic graphs (DAGs) to represent maps from populations (Wang *et al.,* 2011). Using *Brassica* as an example, Wang *et al.* (2011) also compared the use of JoinMap v4.0 and MergeMap to construct the integrated map. A higher density integrated map developed from *B. napus* was produced by JoinMap with one marker every 515 kbp as compared to one marker every 630 kbp when MergeMap was applied. However, MergeMap had advantages in terms of run time, as regression mapping in JoinMap proved to be time-consuming (Ooijen *et al.,* 2006; Wang *et al.,* 2011). Based on calculations of Spearman's rank correlation in the marker order for the integrated maps, MergeMap was shown to generate an integrated map with higher marker order consistency (Spearman's correlation coefficient r>0.90) when the marker order of each linkage group from the *B. napus* integrated map was compared with the population-specific maps developed from three doubled haploid mapping populations. Although JoinMap obtained lower marker order consistency, it was proven to be able to produce a greater number of unique loci and more accurate estimates of genetic distance as it made use of all pairwise recombination frequencies and LOD scores. For example, the

integrated map in *B. napus* produced by JoinMap generated 2,196 unique loci covering 1,791.9 cM across 19 linkage groups, as compared to 1,796 unique loci and total map length of 5,547.4 cM generated by MergeMap (Wang *et al.,* 2011). The authors concluded that JoinMap performs well and produces an integrated map with reliable genetic distances.

The first genetic map for a narrow cross derived from two landraces DipC x Tiga Nicuru in Bambara groundnut using an F_3 segregating population was reported in Ahmad (2012). The addition of dominant DArT markers and SNP markers derived from the same segregating population into the first genetic map, as described in Chapter 5, has improved the resolution of the present framework map in Bambara groundnut. In this chapter, the generation and use of GEMs produced by hybridisation of Bambara groundnut leaf RNA to the Soybean GeneChip to construct a framework map using the F_5 segregating population derived from two landraces DipC x Tiga Nicuru will be reported. This map was constructed using an F_5 population, compared to the F_3 population which was used for construction of the DArTseq genetic map, so individual lines have had a further two generations of inbreeding. In addition, the attempt to integrate two genetics maps, the DArT-seq map and the GEM map, is also described in order to facilitate the identification of eQTL (West *et al.,* 2006).

6.2 MATERIALS AND METHODS

6.2.1 Leaf harvest and RNA preparation

After Bambara groundnut parental lines and the F_5 segregating population had received six weeks of drought treatment, two leaves from each of the individual plants were harvested. Leaf samples from all parental lines in the irrigated plot were also collected as experimental controls. Following the harvest, each piece of leaf measuring approximately 5 cm x 4 cm, was wrapped in labelled aluminium foil packets and flash frozen in a large dewar of liquid nitrogen, where it remained to prevent thawing of the leaf samples. The leaf samples were then transferred to a -80°C freezer for longer term storage.

RNA was extracted from leaf samples from one replicate of each line (n=65). In addition, three replicates of each parental line (DipC and Tiga Nicuru) under drought and irrigation conditions (12 samples in total) were also extracted. All extraction were carried out using the QIAGEN RNeasy Plant Mini Kit (Qiagen, UK) according to the manufacturer's instructions. The final RNA product was resuspended using 30 µl RNase-free water. The total RNA were checked for integrity and quality using both the Agilent 2100 Bioanalyzer (Agilent Technologies, US) and gel electrophoresis. As a result, 10 µl of RNA samples (100 ng µl^{-1}) derived from the 12 parental samples and 60 individual lines were sent to NASC Affymetrix Service, UoN, Sutton Bonington Campus, UK for cross-hybridisation analysis onto the Soybean GeneChip.

6.2.2 Generation of GEMs

A total of 72 data files were generated and sent to Plant Sciences, UoN, Sutton Bonington Campus, UK for initial data analysis using GeneSpring GX (version 11.0.2; Agilent Technologies). The analysis approach adopted for Bambara groundnut data was based on, but modified from, Hammond *et al.* (2011) which used *Brassica rapa* as the experimental organism. Three sets of normalised data were produced at three different levels: probe-sets, CDF

179

masked probe-sets and unmasked probe-pairs (oligonucleotide). A new custom .CDF file was created using PIGEONSv1.2 software by filtering the original DipC .CDF file and Tiga Nicuru .CDF file at threshold 141.00. The custom .CDF file was then used to mask the signals derived from each probe-set/probe-pair in order to generate a custom masked probe-sets/pairs data set. To generate potential GEMs, a series of analyses were conducted using the three sets of normalised chip data.

Firstly, the mean and standard deviation (s.d.) of each $\log_2$-normalised hybridisation signal was calculated for each of the parent from the drought plot (DipC [n=3] and Tiga Nicuru [n=3]), followed by the segregating population (n=60) for each putative marker.

Secondly, each individual line for each putative marker was provisionally assigned into parental 'DipC' and 'Tiga Nicuru' scores based on the mean of signal value of the population (n=60). Conventionally, the female parent is represented as the first parent in a cross. Here the female parent is DipC and the male parent Tiga Nicaru. An 'a' allele score was given when the signal value of individual line was on the same side of the mean population signal as the DipC parent. A 'b' score was given when the hybridisation signal for an individual line was on the same side of the mean as the parental value 'Tiga Nicaru'.

Thirdly, the mean and s.d. of signal value was computed for individual lines scored as 'a' and 'b', respectively. The s.d. values from 'a' and 'b' for each marker were averaged. By dividing the s.d. of the hybridisation signal of the entire population by the average s.d. of the hybridisation signal derived from 'a' and 'b', a 'distinctness' score that indicated the likely degree of separation between group 'a' and group 'b' was calculated (Figure 6.2).

After the threshold value was defined, the probe-sets or probe-pairs in their respective normalised data sets with distinctness score of equal or higher than a selected threshold value were selected as potential GEMs. The threshold values of markers used in map construction were retrospectively checked

through visual inspection of the graphical distribution of group 'a' and group 'b'. A good separation of 'a' and 'b' allele scores within the individual lines would allow the production of polymorphic GEMs of good quality.
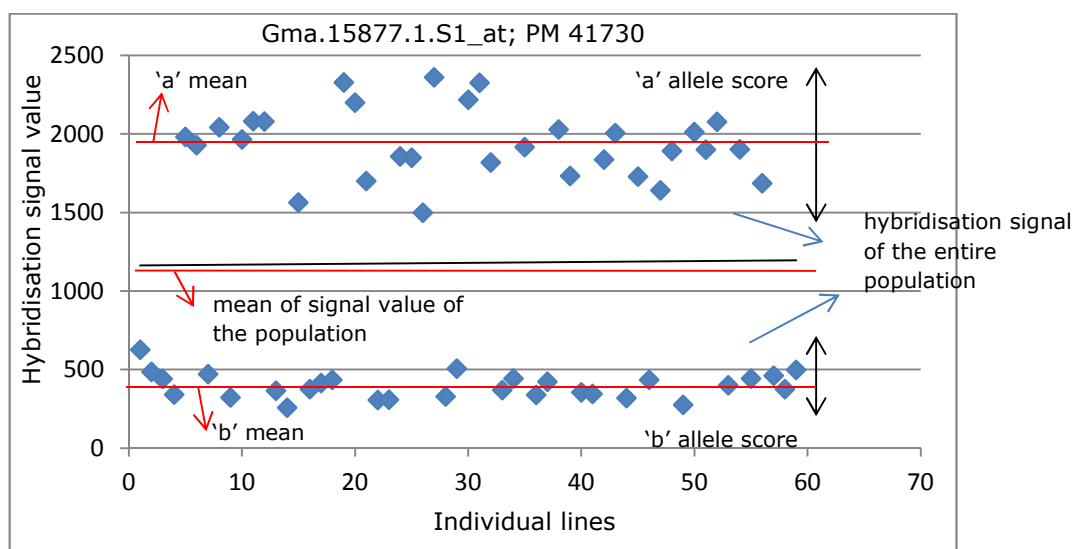
Figure 6.2 An illustration of the estimates generated to develop the 'distinctness' score for potential GEMs.

6.2.3 Examination of markers

The hybridisation patterns of mapped GEMs generated by cross-hybridising DNA and RNA samples onto the soybean GeneChip, respectively, were examined using PIGEONS (V1.2.1) following the guidelines contained in PIGEONS Quick User Guide 2010-2011. Firstly, CDF files derived from the soybean GeneChip were loaded into the software, followed by various CEL files generated from Bambara groundnut DNA and RNA cross-hybridised onto soybean, respectively. In this case, CEL files from two parental lines, DipC and Tiga Nicuru, were selected for a preliminary comparison at the DNA and RNA level. Secondly, the suggested threshold of 120 was used based on the graphical results shown in 'Pigeon Filter'. Thirdly, dual-fold-change Analysis (DFC) from Pigeon Mining and Image was carried out using a fold-change value,

for instance, 2 and 1.5 for the 'Parent' and 'F$_2$', respectively. The potential candidates that fit the criteria were exported as a list.

6.2.4 Conversion and selection of markers for map construction

GEMs were scored as dominant markers using a similar approach to that adopted in Chapter 5. Based on the scoring of 'a' and 'b' in two parental lines (DipC and Tiga Nicuru), GEMs for each individual were given genotype codes either (a,c) or (d,b). When DipC scored 'a' and Tiga Nicuru 'b', a genotype code of (a,c) was given to individuals. Conversely, when DipC scored 'b' and Tiga Nicuru 'a', a genotype code of (d,b) was assigned. Table 6.2 shows the scoring based on the genotype code derived from parental lines. Potential GEMs that showed no polymorphisms across the individual lines or between parental samples were removed.

Table 6.2 The scoring and conversion of GEMs as dominant markers.

DipC	Tiga Nicuru	Genotype code	Conversion	
a	b	(a,c)	a > hh > a	b > k- > c
b	a	(d,b)	a > hh > b	b > k- > d

6.2.5 The construction of the GEM map

The GEM map was constructed using JoinMap v4.1 (Ooijen *et al.,* 2006). As per the JoinMap v4.1 instruction manual, the data was arranged in an Excel (.xlsx) file type, copied and pasted into the JoinMap software spreadsheet to conduct linkage analysis (Chapter 5). The population type was entered as 'Rlx; x:5' for the F$_5$ segregating population. Once the grouping trees were generated, the maximum likelihood mapping approach of JoinMap with grouping at LOD 3.0 and above was first applied to obtain the GEM order for each linkage group. The

GEMs were ordered through a reiterative process of removing markers by examining 'plausible positions', focusing on close genetic distances of two adjacent markers (1-3 cM). The higher the value obtained in plausible positions, the higher confidence that the marker is located in the chosen position and the more confidence in the relatively order of markers in the genetic map. In addition, markers that showed double crossover events in individuals within distances between 1 and 5 cM were also removed. When the number of markers reached 80 and below, the regression mapping approach with a recombination fraction ≤4.0, ripple value=1, jump in goodness-of-fit threshold=5 under a Haldane mapping function was then introduced. Through the alternate use of the maximum likelihood mapping approach and the regression mapping approach, a framework map consisting of GEMs was then generated.

6.2.6 Integration of the DArTseq map and GEM map

The integration of the DArTseq framework map (as described in Chapter 5) and GEMs framework map was attempted using JoinMap v4.1 (Ooijen *et al.,* 2006). All genotype information derived from both framework maps was pooled together prior to linkage analysis. Using the DArTseq framework as the backbone, the population type was entered as 'Rlx; x:3' as the DArTseq map are derived from an F_3 segregating population. The grouping of markers was set between LOD 2.0 and 10 with a step of LOD 0.5. In addition, the marker order derived from the DArTseq framework map was listed as the 'fixed order'. The regression mapping approach was used under default conditions in the linkage analysis as presented in Ahmad (2012), and the final order of markers in the integrated map was computed through a repeated process of removing markers by examining the 'Mean Chi-square Contributions' tabsheet. The markers with the large contribution to the chi-square goodness-of-fit increase of the map as well as high values of neighbour fit were removed as they did not fit well at the proposed map location (Ooijen *et al.,* 2006). Results from the regression

mapping approach were used to attempt to create an initial integrated map derived from DArT Seq (dominant DArT markers and SNPs markers) and GEMs.

6.3   RESULTS

6.3.1 The development of GEMs from the Soybean GeneChip for mapping

Three rounds of separate analyses on three sets of normalised data at different levels: probe-sets, CDF masked probe-sets and unmasked probe-pairs (oligonucleotide) were conducted (Table 6.3).

A data matrix with 61,035 probe-sets generated with normalised data at the probe-sets level was analysed. After a series of post-analyses, a distinctness score as high as 4.03 was obtained, followed by 3.44, 3.14 and 3.05, with the lowest distinctness score being 1.18. A threshold value of 2.50 was set in order to obtain a relatively good separation between 'a' and 'b' alleles across the individual lines. For example, the probe set 'GmaAffx.92555.1.S1_s_at' with a distinctness score of 4.03 is presented in Figure 6.3a. When the distinctness score fell below 2.50, a more scattered graph was usually observed, such as the probe set 'GmaAffx.57563.1.S1_at' with distinctness score of 2.15 (Figure 6.3b). As a result, 15 potential GEMs with distinctness score of 2.50 and above were retrieved from the data matrix based on unmasked probe-sets.

Table 6.3 The summary of GEMs development at three different levels: probe-sets, CDF masked probe-sets and unmasked oligonucleotides.

Level of analysis	Highest distinctness score	Cut-off point	Total GEMs	Number of potential GEMs
Probe-sets	4.03	2.5	61,035	15
CDF masked probe-sets	6.09	2.6	53,651	48
Unmasked probe-pairs (Oligonucleotide)				
(a)Perfect-match probes (PM)	7.99	2.34	669,982	1,030
(b)Mis-match probes (MM)	8.58E+13	2.3	669,982	501

(a)

(b)
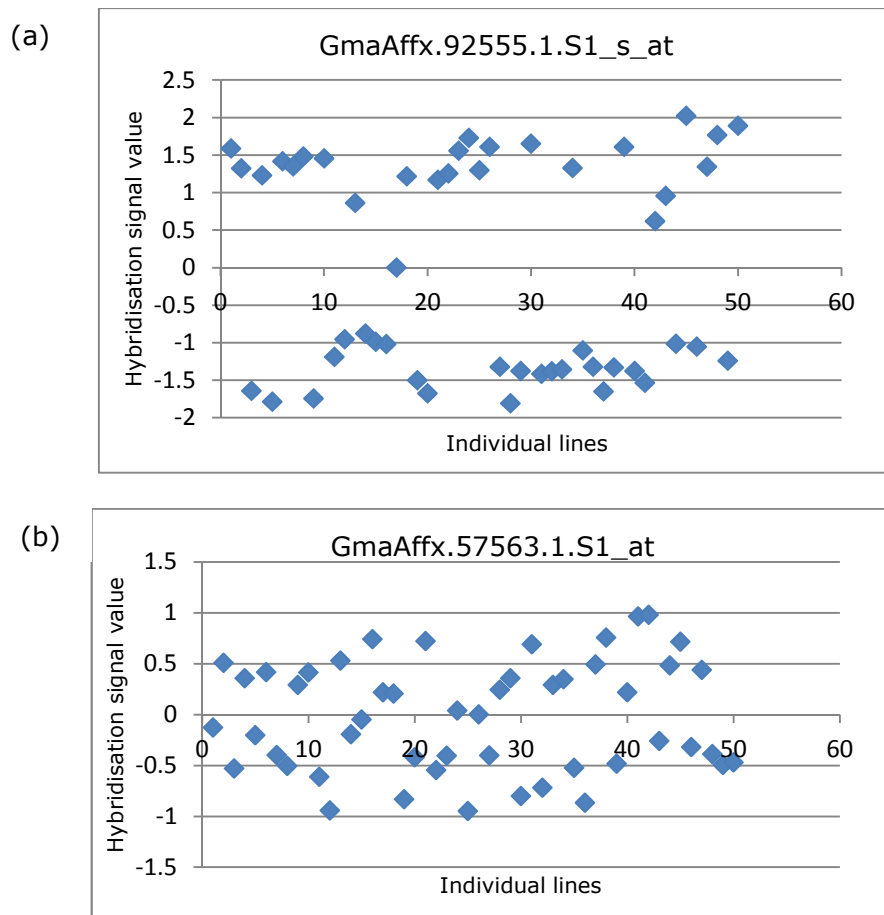
Figure 6.3 A visual inspection of the trait distribution of 'a' and 'b' allele scores across the individual lines at the unmasked probe-set level.   (a) The probe set 'GmaAffx.92555.1.S1_s_at' with distinctness score of 4.03 and (b) probe set 'GmaAffx.57563.1.S1_at' with distinctness score of 2.15.

A total number of 53,651 CDF masked probe-sets were obtained in the second round of analysis after filtering CDF masked probe-sets that were differentially expressed in the parental lines and all the individual lines. The post-analysis result showed that a higher distinctness score up to 6.09, followed by 5.23 and 5.21 were obtained for CDF masked probe-sets as compared to distinctness score of 4.03 in the previous analysis. Based on the visual inspection of graphical distribution of 'a' and 'b' allele scores across the individual lines, a minimum distinctness score of 2.60 for the CDF masked dataset was suggested in order to obtain a clear separation, for example PsAffx.C32000037_at, Gma.7135.3.S1_a_at and GmaAffx.88141.1.S1_at (Figure

6.4). In total there were 48 potential GEMs with distinctness score of 2.60 and above extracted from CDF masked data set.
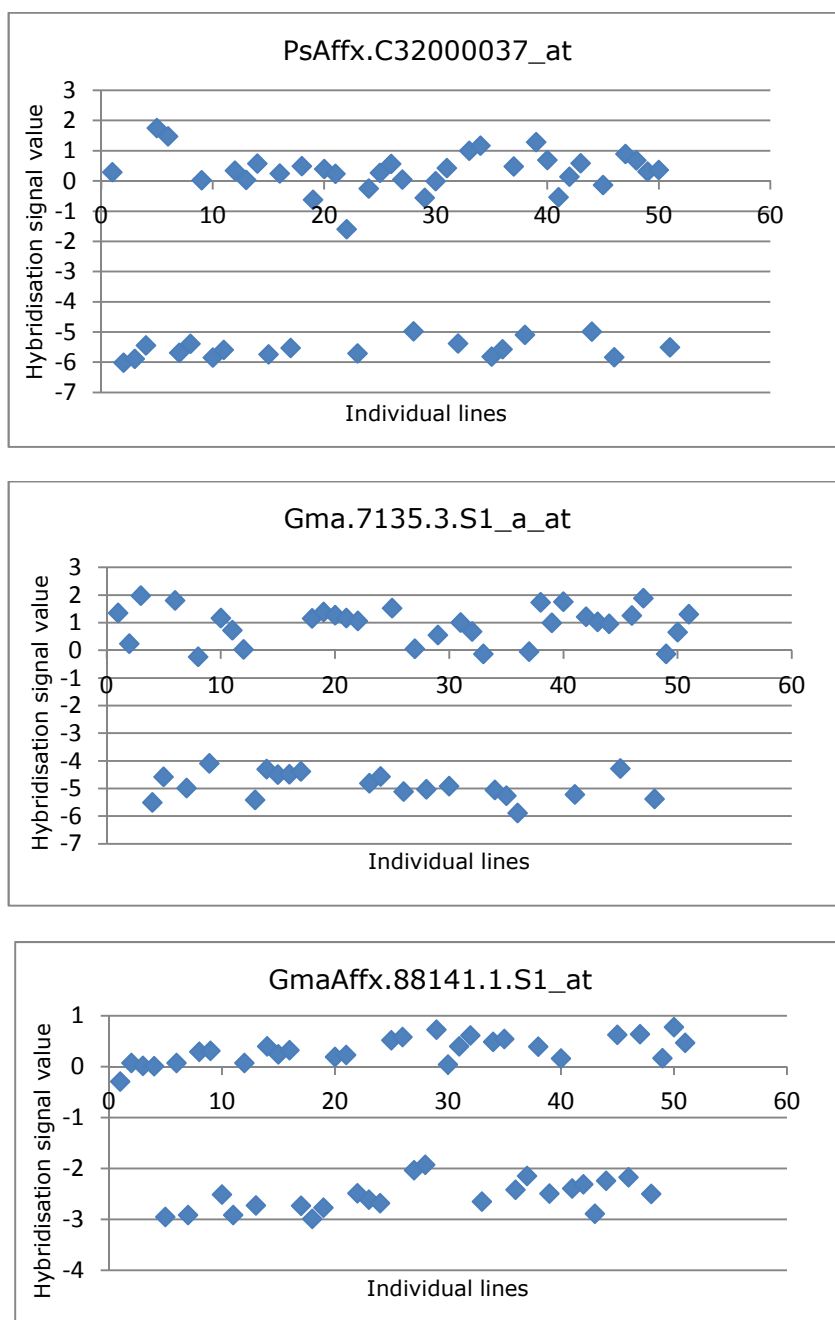
Figure 6.4 A graphical distribution of 'a' and 'b' alleles scores across the individual lines at the CDF masked probe-sets level. A clear separation was observed in PsAffx.C32000037_at (6.09), Gma.7135.3.S1_a_at (5.23) and GmaAffx.88141.1.S1_at (5.21).

Although the number of potential GEMs generated from CDF masked probe-sets had been improved, the relatively low number was still insufficient for mapping studies. Therefore, normalised data containing unmasked probe-pairs was used in the third round analysis. The hybridisation signal values of the individual unmasked probe-pair for each probe-set was calculated, resulting in 669,982 perfect-match probes (PM) and 669,982 mis-match probes (MM) respectively. For PM probes, the result showed that the highest distinctness score was 7.99 (Gma.3025.1.S1_at; PM-933459), followed by 7.30 (GmaAffx.69054.1.S1_s_at; PM-460879) and 5.31 (Gma.15877.1.S1_at; PM-41730). The distinctness graphs of the top 100 PM probes, ranking from the highest to lowest distinctness score, are presented in Appendix 7. A distinctness score of 2.34 was set as a cut-off point for good separation between 'a' and 'b' alleles across individual lines, resulting in a total number of 1,030 potential GEMs.

A similar result was also observed in MM probes in which a distinctness score of up to 7.10 was obtained by Gma.289.1.S1_s_at; MM-1048658, followed by 4.73 (Gma.17784.1.S1_at; MM-177722) and 4.47 (Gma.15877.1.S1_at; MM-42894). The lowest distinctness score for MM probes was 0.37 from 'PsAffx.CL2153Contig1_at; MM- 193965'. As a result, 501 potential GEMs were generated from MM probes where a threshold value of 2.30 was selected based on the graphical distribution of group 'a' and 'b' across all individual lines.

6.3.2 The comparison of hybridisation patterns in GEMs

Following the identification of GEMs at three different levels, an initial examination of the hybridisation pattern of GEMs derived from cross-hybridisation with the soybean GeneChip was conducted using PIGEONS. The results showed that similar hybridisation patterns were observed at the DNA and RNA level on the same set of probe-pairs, but as might be expected, the hybridisation signal strength varied between the DNA and RNA levels. For

example, probe-pair 4 from Gma.12977.1.S1_at had hybridisation signal differences of up to 3.6-fold between DipC and Tiga Nicuru at the RNA level and 2.6-fold at DNA level (Figure 6.5).

In addition, there were also some examples where hybridisation signal differences were not observed at the same probe-pairs of a single probe set. For instance, Gma.12147.1.S1_at and GmaAffx.23289.1.S1_at showed a high signal value and fold-change differences between two parental lines on different sets of probe-pairs in RNA samples compared to the DNA samples (Figure 6.6). The variation of signal values at the DNA and RNA level between DipC and Tiga Nicuru was presented using the PIGEONS software. By comparing the hybridisation patterns of two parental lines, DipC and Tiga Nicuru, at both DNA and RNA level, a preliminary insight into *cis-* or *trans-* marker variation observed could be provided.
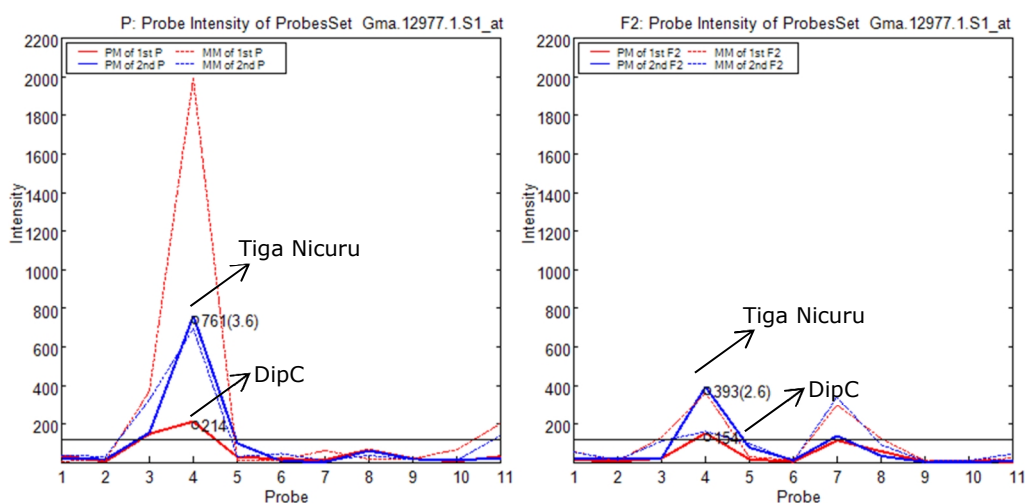
Figure 6.5   An initial examination of the hybridisation patterns of GEMs derived from cross-hybridisation with the soybean GeneChip. A comparison of hybridisation signals derived from RNA (left) and DNA (right) samples was presented. Red: DipC; Blue: Tiga Nicuru. Solid line: PM; Dotted line: MM.
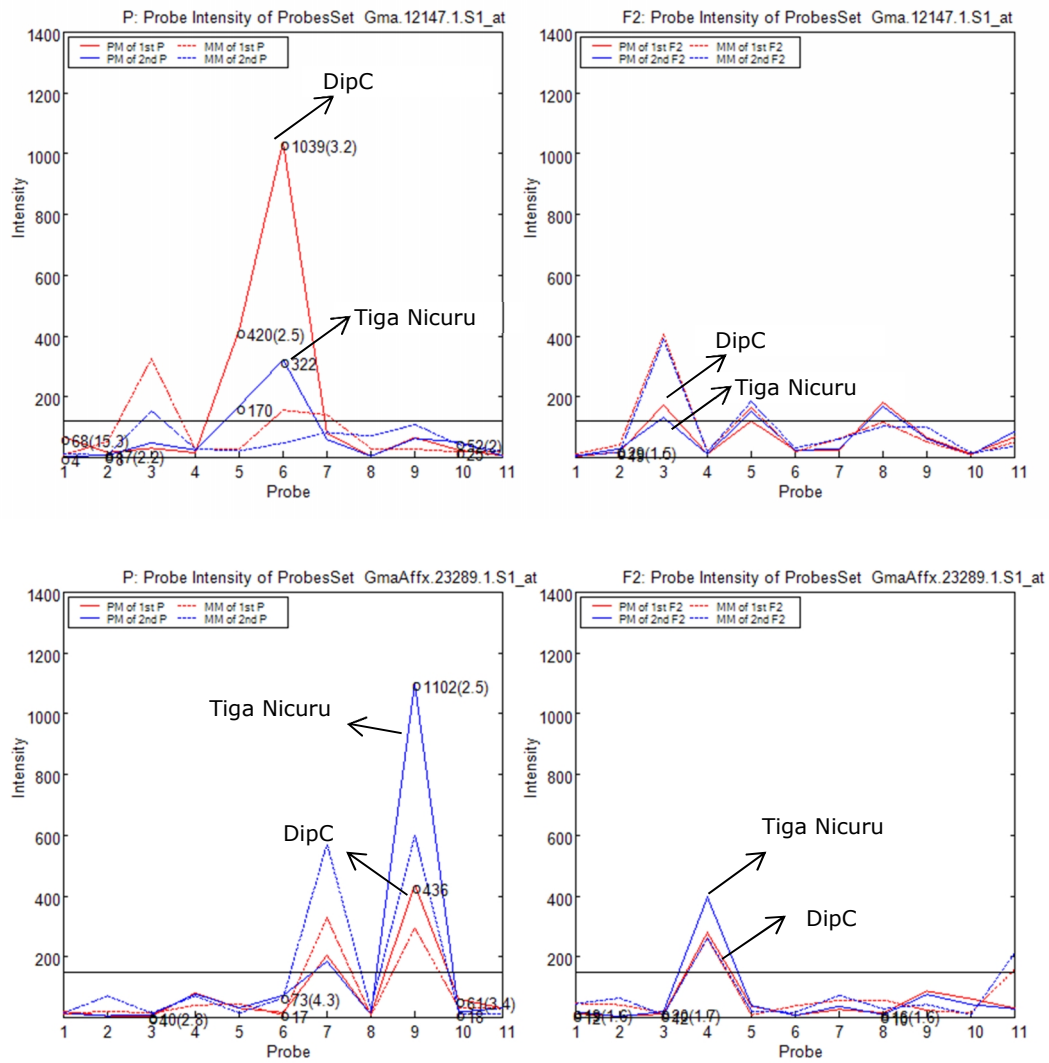
Figure 6.6 Presentation of different hybridisation patterns of GEMs derived from cross-hybridisation with the soybean GeneChip. The comparison of hybridisation signals derived from DNA (right) and RNA (left) samples using A) Gma.12147.1.S1_at and B) GmaAffx.23289.1.S1_at as exemplar probe sets. Red: DipC; Blue: Tiga Nicuru. Solid line: PM; Dotted line: MM.

6.3.3 Linkage groups and marker distribution in the GEM map

Potential GEMs that showed no polymorphism or a low distinctness score were eliminated from the list. Of 1,531 markers (1,030 from PM probes and 501 from MM probes), 753 potential GEMs were identified and subjected to linkage analysis. The segregating pattern of the GEMs was also examined using a Chi-square test against the predicted patterns. The result showed that only 55 markers (7.3%) presented significant segregation distortion whereas 698 GEMs segregated in a way consistent with the expected Mendelian ratio of 17:15 for an F_5 segregating population using dominant markers.

Of 753 GEMs, 527 markers were provisionally mapped before being removed during the construction of framework map. An initial linkage analysis at LOD>3 generated 19 linkage groups with 165 GEMs (120 PM probes and 45 MM probes), spanning 920.3 cM of the Bambara groundnut genome based on the regression mapping approach implemented in JoinMap4.1. The distribution of GEMs across all LGs and their corresponding map lengths is summarised in Table 6.4. There was an average of 8.7 markers per LG with the highest number of 23 observed on LG1, followed by 12 markers on LG5A and LG10A. The lowest number of marker was seen in LG10B which contained only 3 markers. As the GEM map was intended to produce a framework map, a mean distance of 5.1 cM between two neighbouring markers across all LG was obtained, which is in the target for QTL analysis of 5-10 cM between markers. The closest distance between two adjacent markers was 0.1 cM, for example, PM100 (6.4 cM) and PM184 (6.5 cM) in LG8B. However, a spacing distance of 23.1 cM in LG9 between PM193 (20.6 cM) and MM238 (43.7 cM) is reported as the largest distance between two markers across 19 LGs. Further work to fill gaps identified through reintroduction of excluded markers in regions with low marker density would improve coverage in future.

Table 6.4 The distribution of GEMs across 19 LGs for genetic linkage analysis in the F_5 segregating population of Bambara groundnut.

Linkage group (LG)	Length of LG (cM)	GEMs	Average marker interval (cM)
1	113.9	23	5.0
2A	63.7	10	6.4
2B	27.4	7	3.9
3A	61.3	9	6.8
3B	29.2	6	4.9
4A	22.0	7	3.1
4B	21.4	4	5.4
5A	92.1	12	7.7
5B	25.6	6	4.3
6A	69.3	10	6.9
6B	18.4	5	3.7
7	70.6	10	7.1
8A	58.5	9	6.5
8B	7.5	6	1.3
9	89.4	10	8.9
10A	73.7	12	6.1
10B	2.2	3	0.7
11A	62.1	10	6.2
11B	12.0	6	2.0
Grand total	920.3	165.0	96.8
Mean	48.4	8.7	5.1

In addition, the number of markers, marker distances and corresponding LGs were also presented graphically in Figure 6.7. An average map length of 48.4 cM was calculated across all 19 LGs. LG1 appeared to be the longest linkage group and covered 113.9 cM with 23 GEMs, followed by LG5A and LG9 which had a map length of 92.1 cM and 89.4 cM, respectively. The shortest map length was reported to be LG10B, covering 2.2 cM with 3 markers.
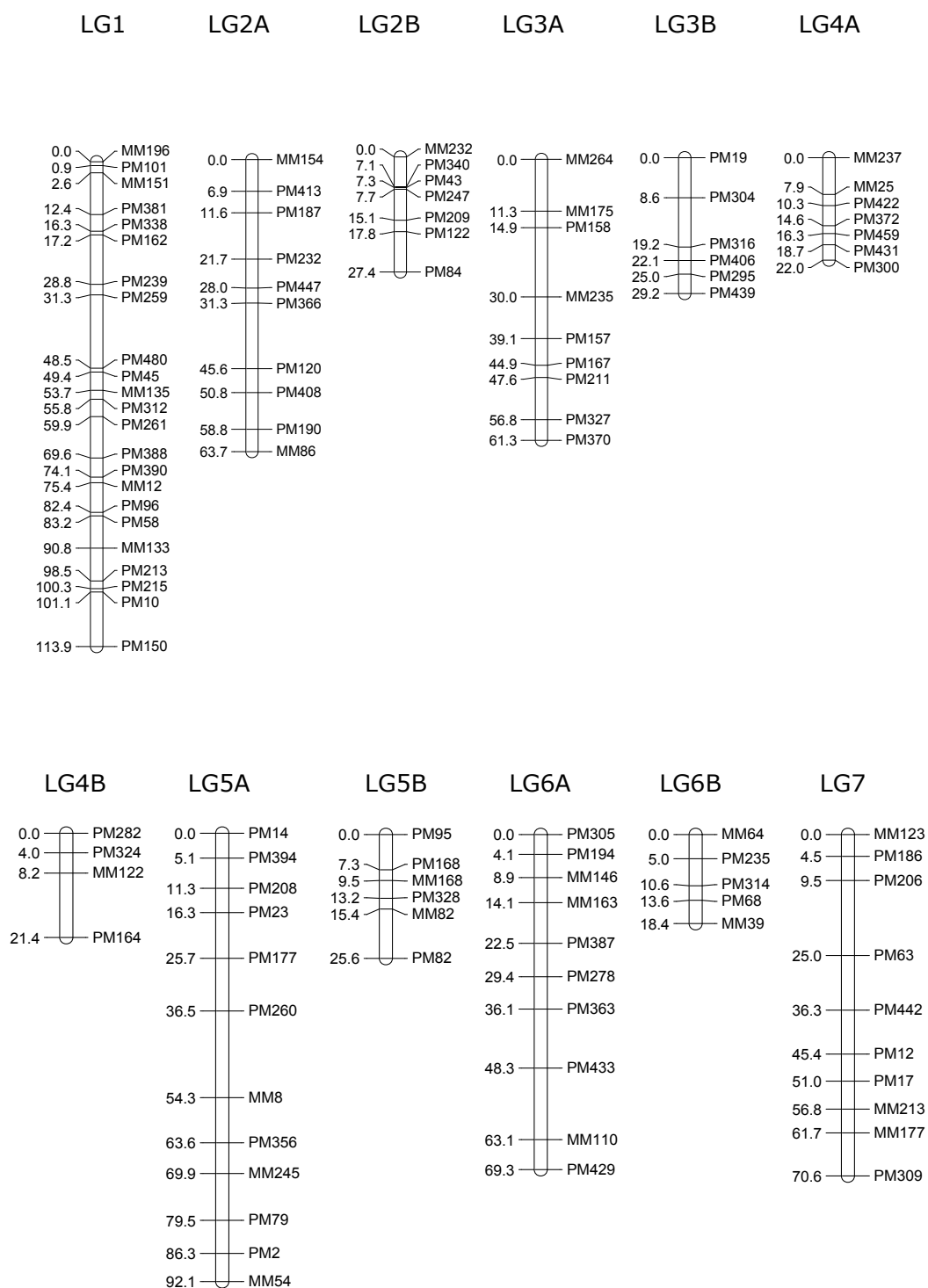
Figure 6.7 Genetic linkage map of the F_5 segregating population in Bambara groundnut constructed by GEMs. Right: positions of markers (cM); left: name of the markers.

LG8A LG8B LG9 LG10A LG10B LG11A LG11B

LG8A
0.0 — PM234
5.3 — PM25
13.3 — PM319
28.2 — PM93
34.4 — PM86
40.1 — PM224
48.6 — PM35
54.9 — MM69
58.5 — PM382

LG8B
0.0 — PM271
4.9 — PM367
6.4 — PM100
6.5 — PM184
6.7 — PM54
7.5 — MM262

LG9
0.0 — PM402
10.4 — PM342
20.6 — PM193
43.7 — MM238
50.7 — PM88
60.9 — MM227
63.1 — PM376
72.4 — PM398
80.7 — MM206
89.4 — PM443

LG10A
0.0 — PM276
6.2 — PM191
13.3 — PM288
14.7 — PM230
23.5 — MM131
32.5 — MM214
35.6 — MM223
40.5 — PM179
45.8 — PM130
52.6 — PM308
58.9 — PM183
73.7 — MM194

LG10B
0.0 — MM125
1.3 — PM242
2.2 — PM135

LG11A
0.0 — PM73
2.9 — PM417
21.3 — PM269
24.5 — PM160
26.4 — PM393
39.7 — MM37
45.9 — MM236
50.0 — PM301
55.0 — MM149
62.1 — PM115

LG11B
0.0 — PM449
2.8 — MM161
6.2 — MM121
7.8 — MM85
10.1 — MM199
12.0 — MM74

Figure 6.7 (cont.)   Genetic linkage map of F_5 segregating population in Bambara groundnut constructed by GEMs. Right: positions of markers (cM); left: name of the markers.

Compared to the regression mapping approach, the maximum likelihood mapping approach generated a longer map at 1125.5 cM with an average spacing of 6.8 cM between neighbouring markers across all the LGs. However, the marker locations were similar and the marker order was the same in all the LGs except for LG5A, LG8B and LG11A where one to two inverted markers were obtained when linkage maps generated from the two mapping approaches were compared (Appendix 8).

6.3.3 Integration of the genetic linkage maps and comparison with the Ahmad original map

A total of 343 markers (31 dominant DArT, 130 SNPs, 5 SSR, 12 microarray-based DArT, 120 PM probes and 45 MM probes) derived from both DArTseq-based and GEM maps were pooled together for map integration. Using the regression mapping approach, grouping analysis at LOD>3.0 resulted in 11 linkage groups and 11 unmapped markers. An integrated linkage map that covered a total of 1,250.7 cM Haldane map distances across 11 LGs with 212 markers, including 18 dominant DArT, 97 SNPs, 3 SSR, 7 microarray-based DArT, 64 PM probes and 23 MM probes, was derived from the initial groupings. The markers were one every 6 cM across all 11 LGs and a mean number of 19.3 markers was obtained for each LG. The highest number of markers was obtained in LG 5 (26) whereas the lowest number of markers was observed in LG7 (13). In addition, the shortest marker spacing of 1.3 cM between two neighbouring markers was found on LG6, with SNP100005109|0-5_6 (27.0 cM) and MM146 (28.3 cM), whereas the longest distance was 16.7 cM between SNP100012935|0-32_8 (70.5 cM) and PM100 (87.2 cM) in LG8.

Among the linkage groups, LG5 was reported to be the longest group as it was mapped with 26 markers (2 dominant DArT, 10 SNPs, 1 SSR, 10 PM probes, 3 MM probes) covering a map length of 143.4 cM, followed by LG11 and LG10 with map lengths of 128.6 cM and 127.3 cM, respectively. The shortest LG had map length of 88.3 cM mapped with 17 markers in LG6. The distribution of each type of markers across each LG, number of markers, markers distance and corresponding LGs map length are summarised in Table 6.5. The results showed that SNPs contributed 45.75% (97 out of 212 markers) to the integration of map, with the highest number of 10 markers observed in LG2, LG5 and LG10. 87 GEMs (41.04%; 64 PM probes and 23 MM probes), 18 dominant DArT (8.49%), 7 microarray-based DArT (3.30%) and 3 SSR (1.42%) were also represented.

Table 6.5 The distribution of dominant DArT, SNPs, SSR, microarray-based DArT, GEMs (PM probes and MM probes) across each LG for map integration in Bambara groundnut.

Linkage group (LG)	Length of LG (cM)	dominant DArT	SNPs	SSR	Microarray-based DArT	PM	MM	Total number of markers	Average marker interval (cM)
1	110.7	1	5	1	3	6	1	17	6.5
2	121.8	0	10	1	2	6	2	21	5.8
3	111.7	1	9	0	0	6	2	18	6.2
4	102.8	3	13	0	0	7	2	25	4.1
5	143.3	2	10	1	0	10	3	26	5.5
6	88.3	2	9	0	0	3	3	17	5.2
7	88.7	0	6	0	0	5	2	13	6.8
8	118.9	4	8	0	0	8	0	20	5.9
9	108.6	1	9	0	1	5	2	18	6.0
10	127.3	2	10	0	0	3	2	17	7.5
11	128.6	2	8	0	1	5	4	20	6.4
Grand total	1250.7	18	97	3	7	64	23	212	66.1
Average	113.7	1.6	8.8	0.3	0.6	5.8	2.1	19.3	6.0

The integrated map of each LG was compared graphically with respective LGs derived from original DArTseq map, GEM map and also the first genetic linkage map derived from the cross between DipC and Tiga Nicuru in Bambara groundnut (Ahmad, 2012; Figure 6.8a-6.8k). For unknown reasons, the two original maps appeared to be poorly integrated. The graphical presentation showed that a relatively large number of markers were missing in the integrated maps, resulting in the loss of genotypic marker information for subsequent QTL and/or eQTL analysis. However, despite the loss of marker information the marker order from both original DArTseq map and GEM map were adequately conserved in the integrated map, except for LG2 and LG6 which present an inverted marker order in the combined map when compared to the DArTseq map. A larger number of linkage groups (n=19) were obtained from the GEM map, however, two LGs could be aligned into a single LG when additional marker information from the DArTseq map was presented in integrated map. For

instance, LG2A and LG2B from the GEM map were aligned to LG2 in the integrated map. There were also some LGs with short map length that remained unmapped, such as LG2, LG6, LG9, LG13, LG20 and LG21 derived from pre-existing genetic map (Ahmad, 2012; Figure 6.9).
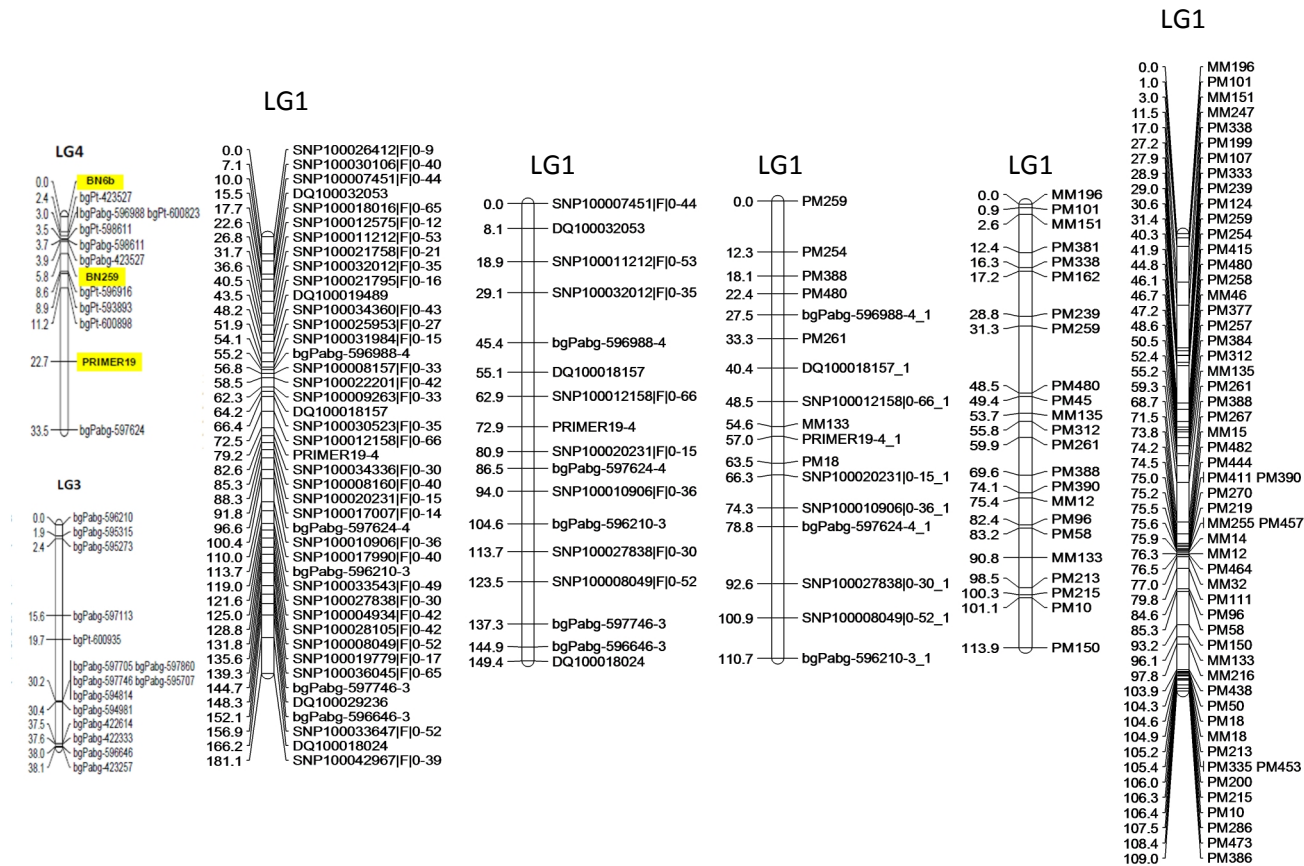
Figure 6.8(a) The graphical comparison of the integrated map with original maps for LG1. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
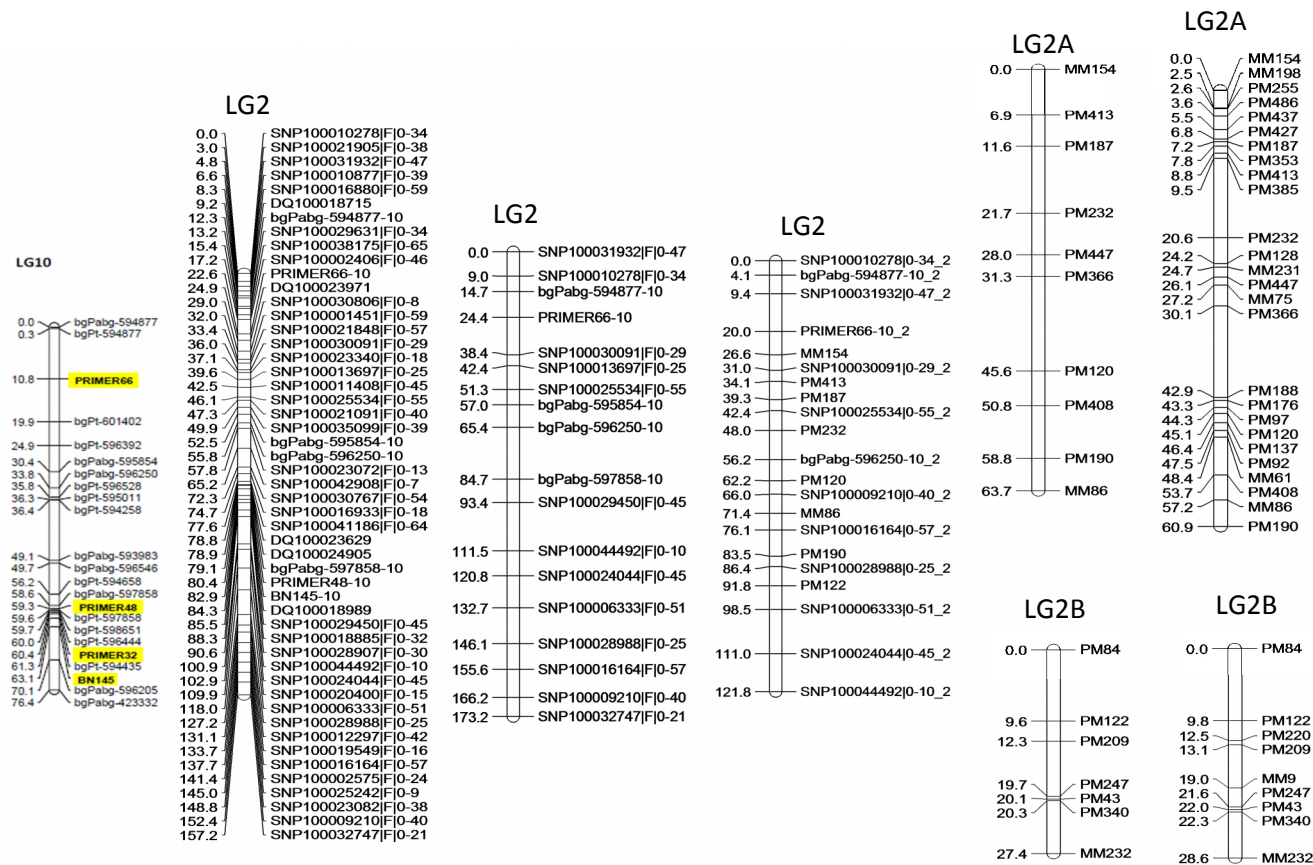
Figure 6.8(b) The graphical comparison of integrated map with original map for LG2. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
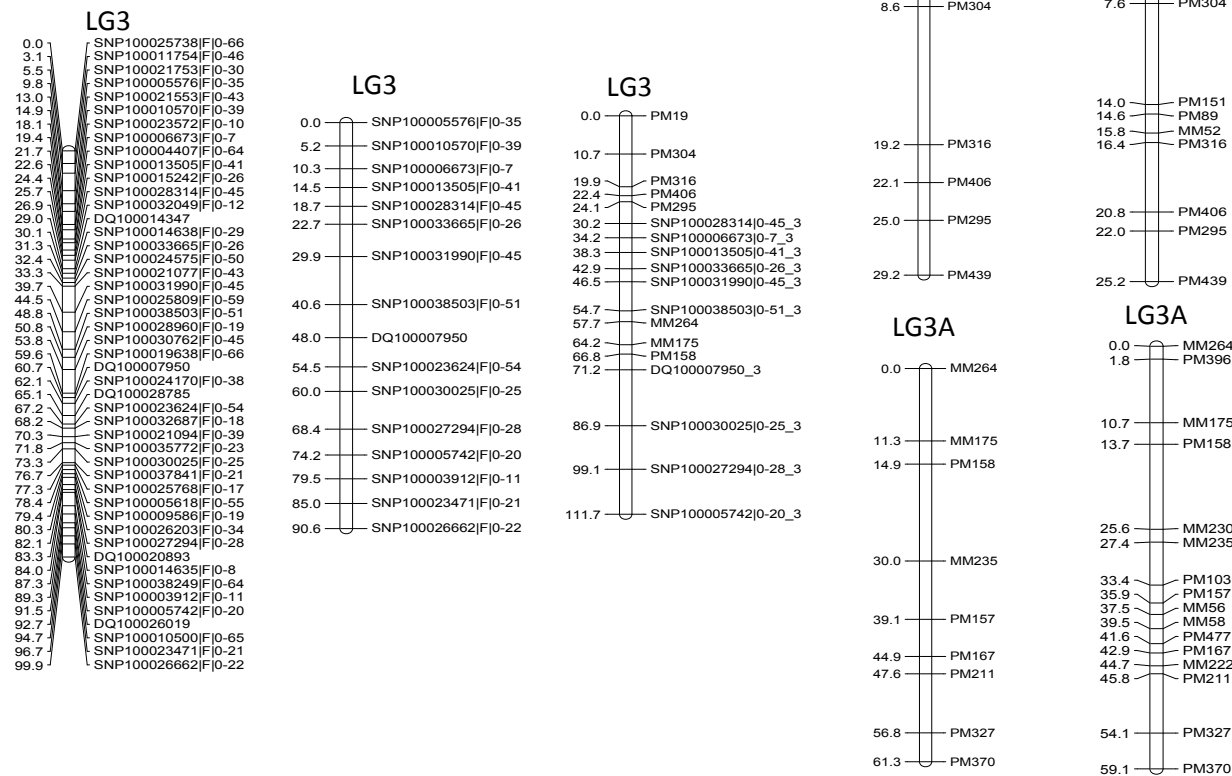
Figure 6.8(c) The graphical comparison of integrated map with original map for LG3. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
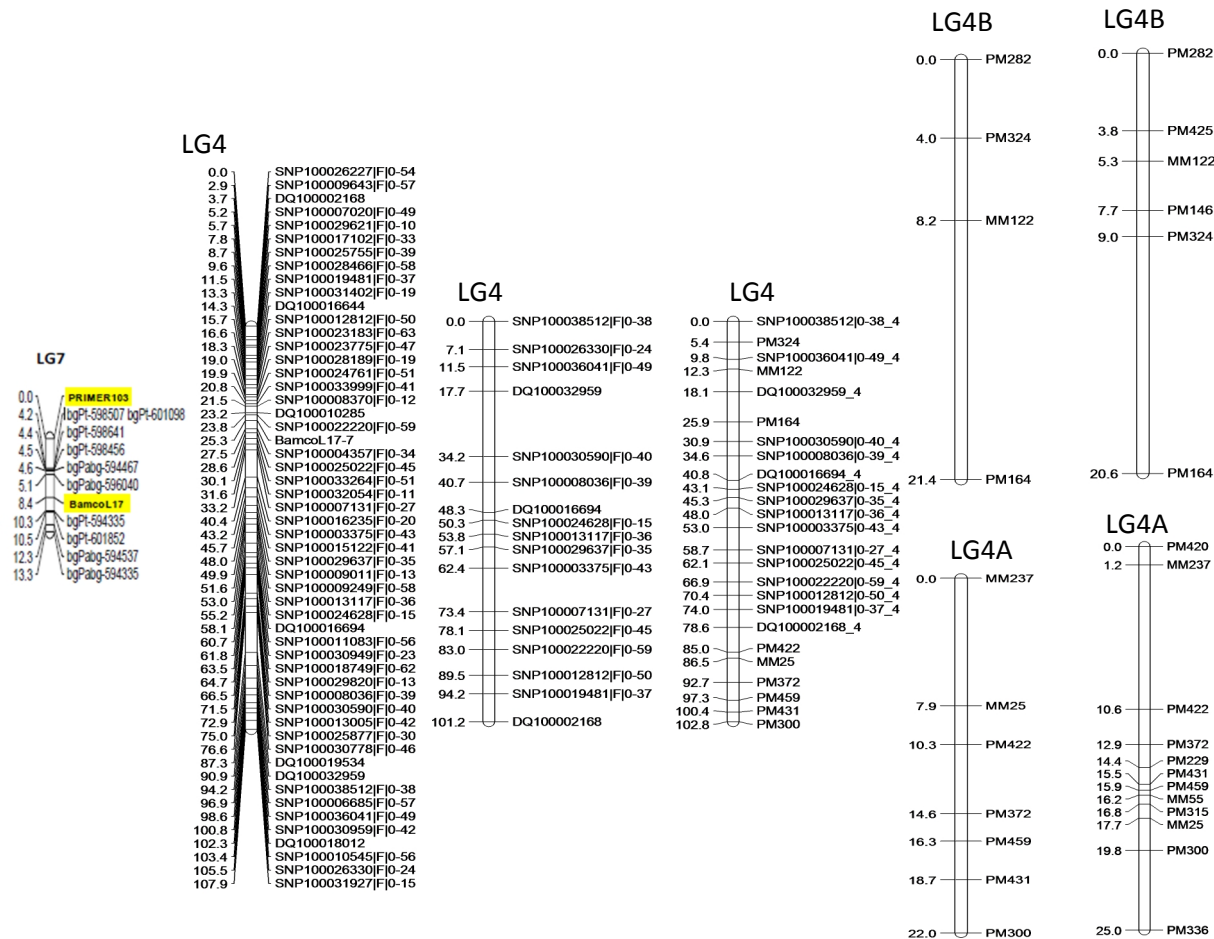
Figure 6.8(d) The graphical comparison of integrated map with original map for LG4. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
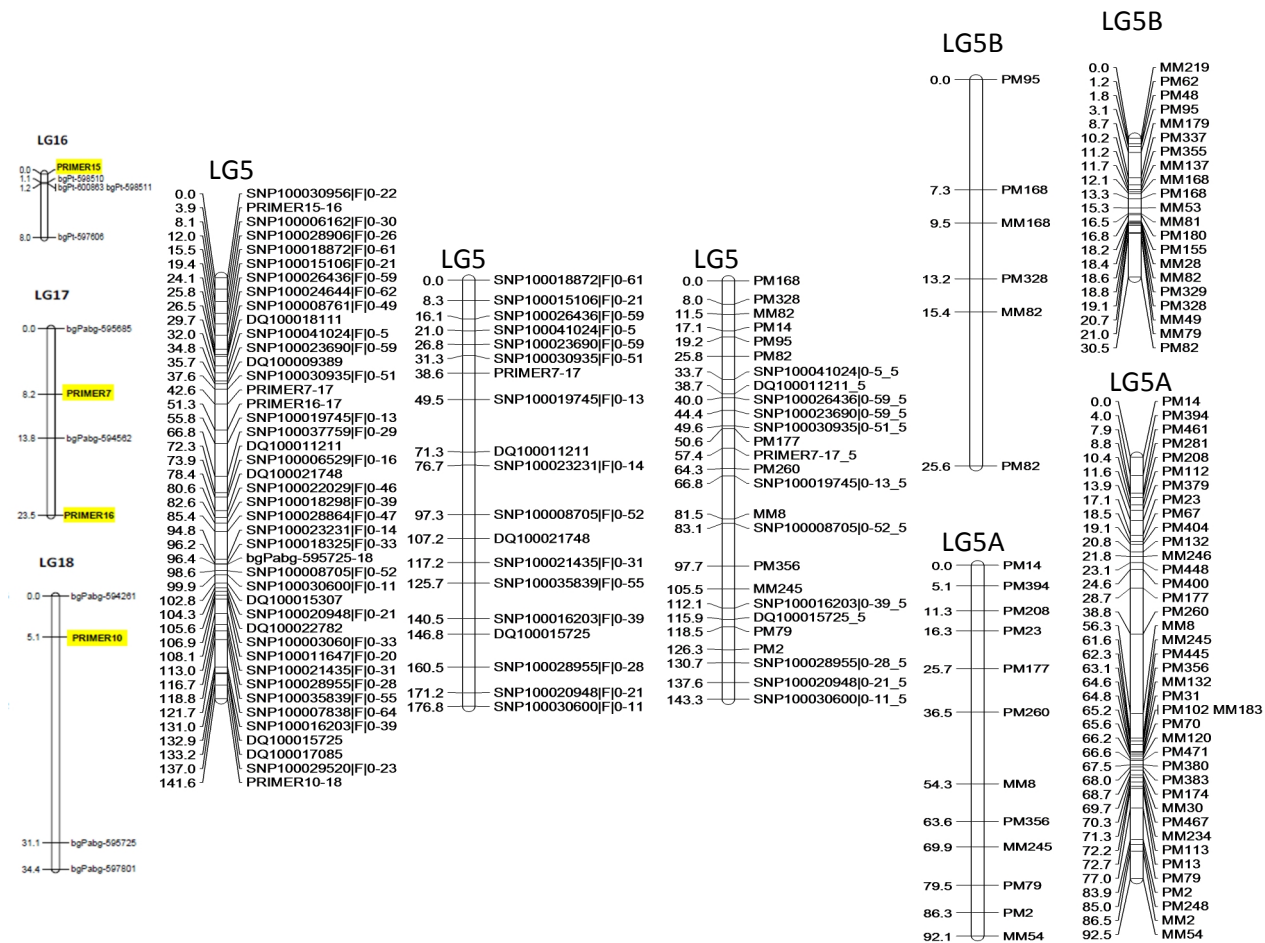
Figure 6.8(e) The graphical comparison of integrated map with original map for LG5. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
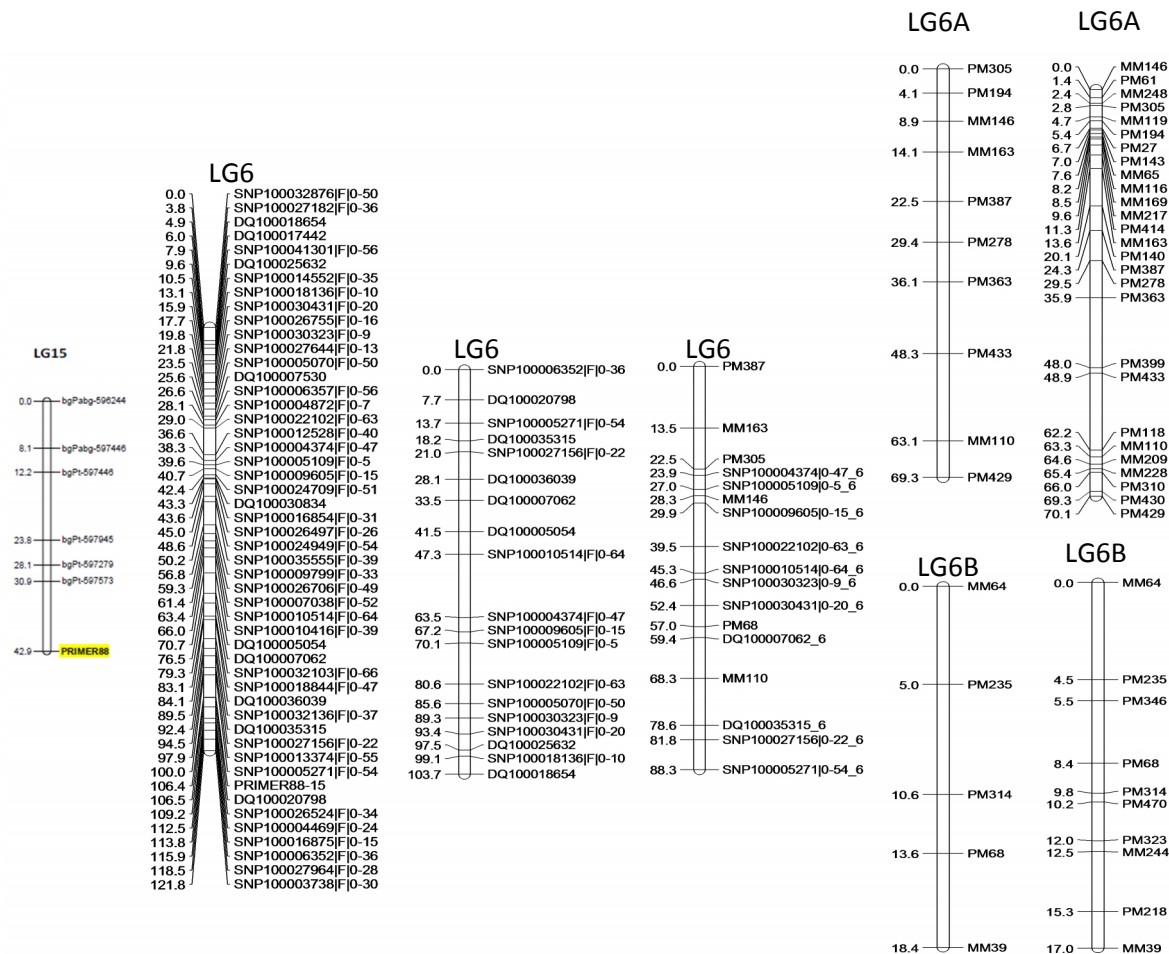
Figure 6.8(f) The graphical comparison of integrated map with original map for LG6. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
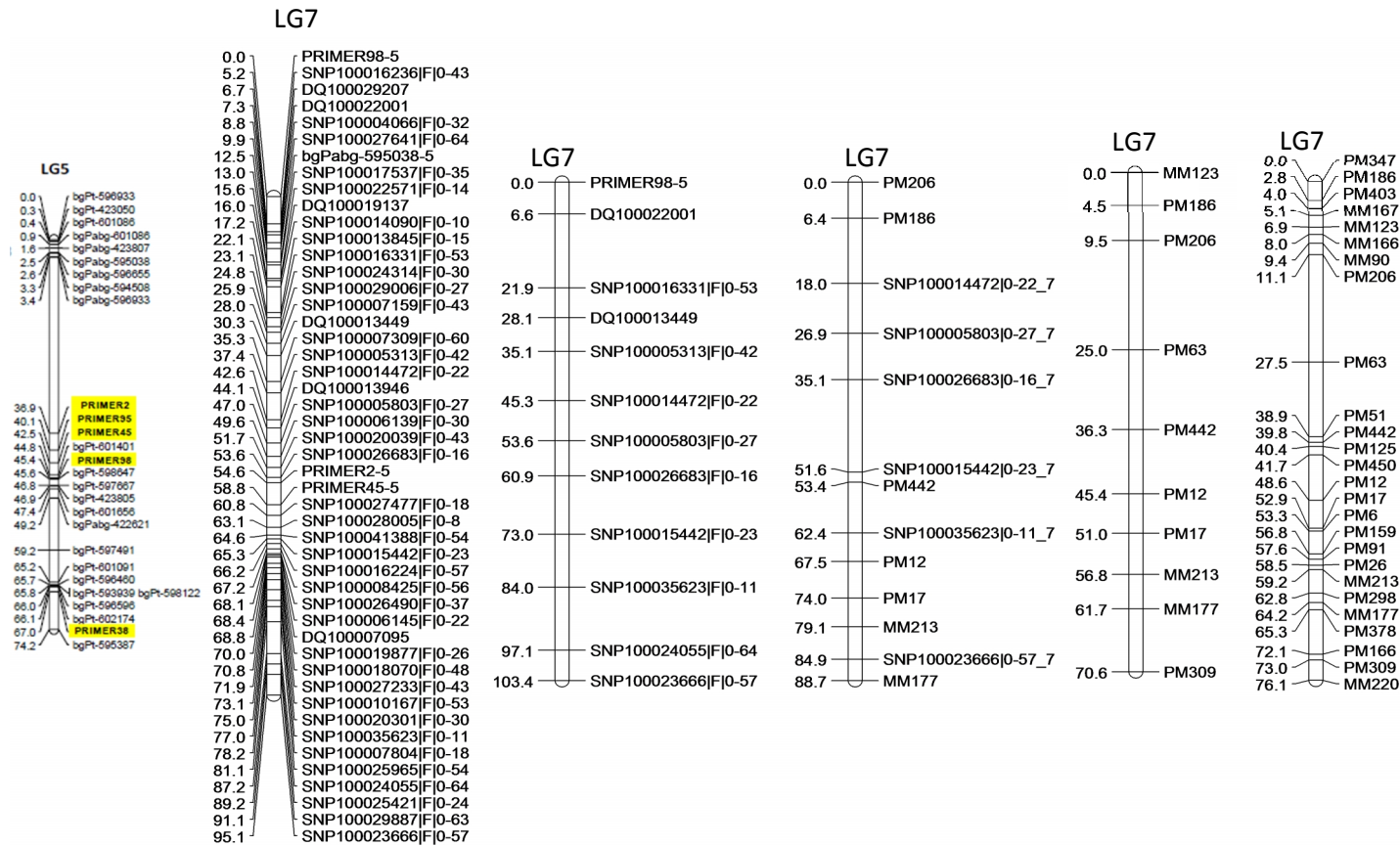
Figure 6.8(g) The graphical comparison of integrated map with original map for LG7. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
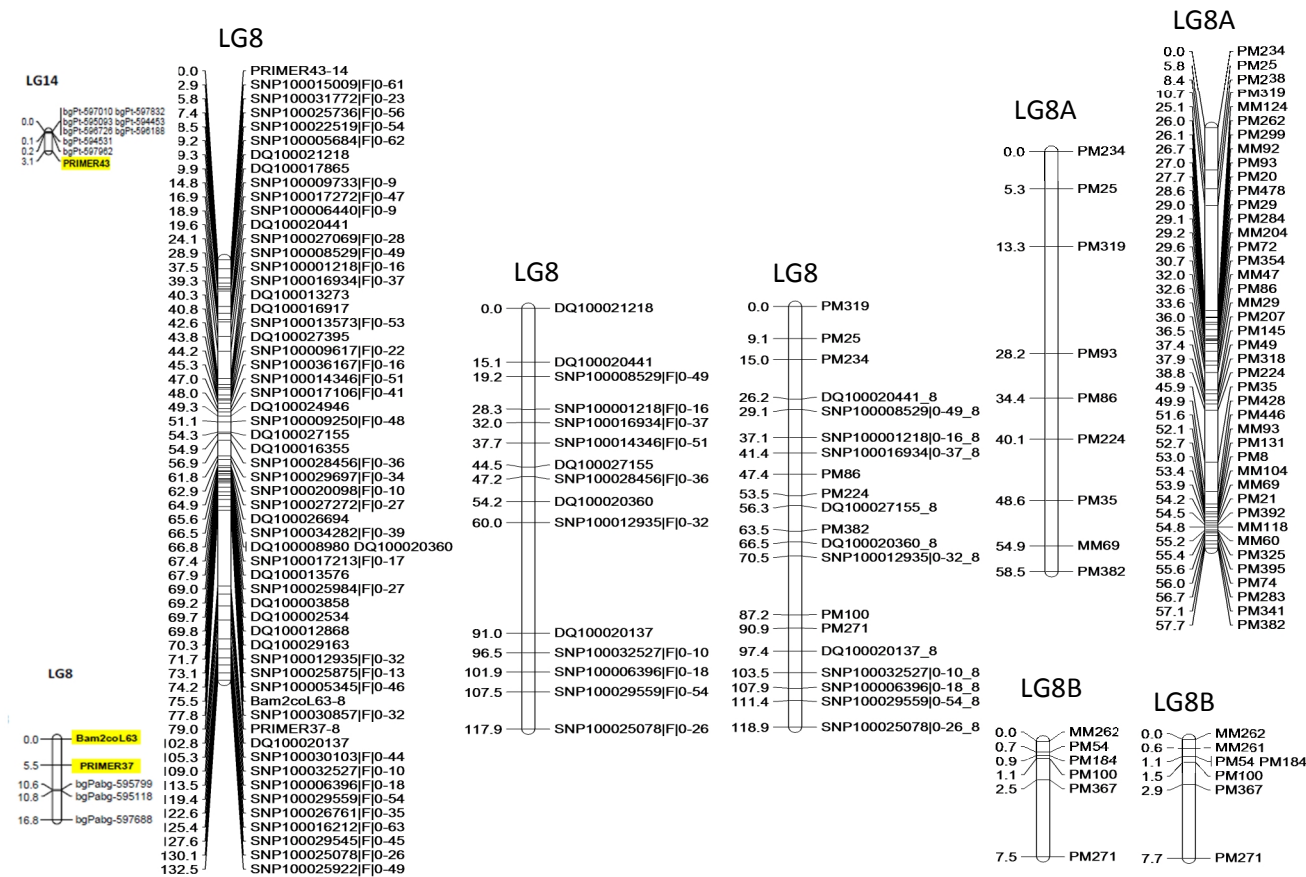
Figure 6.8(h) The graphical comparison of integrated map with original map for LG8. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
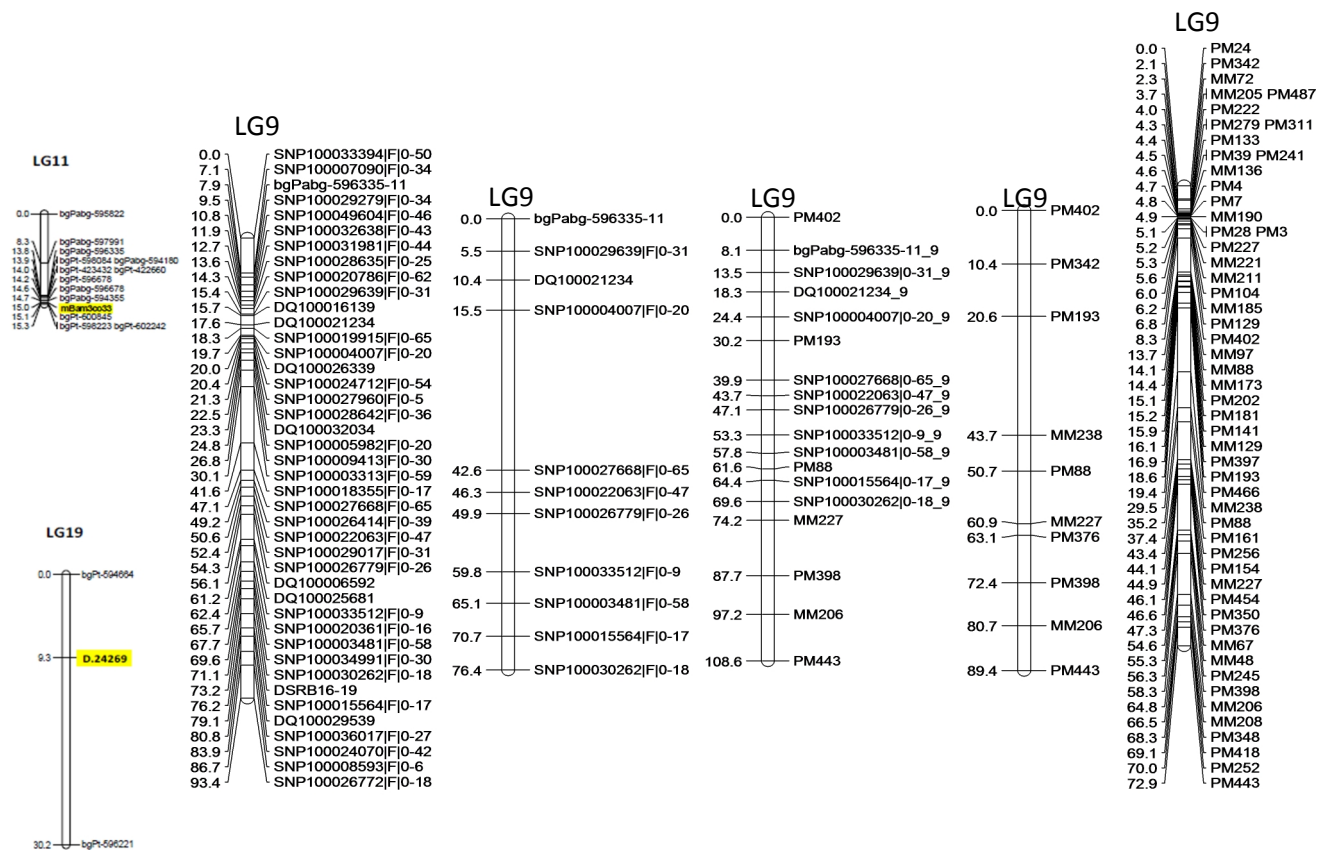
Figure 6.8(i) The graphical comparison of integrated map with original map for LG9. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
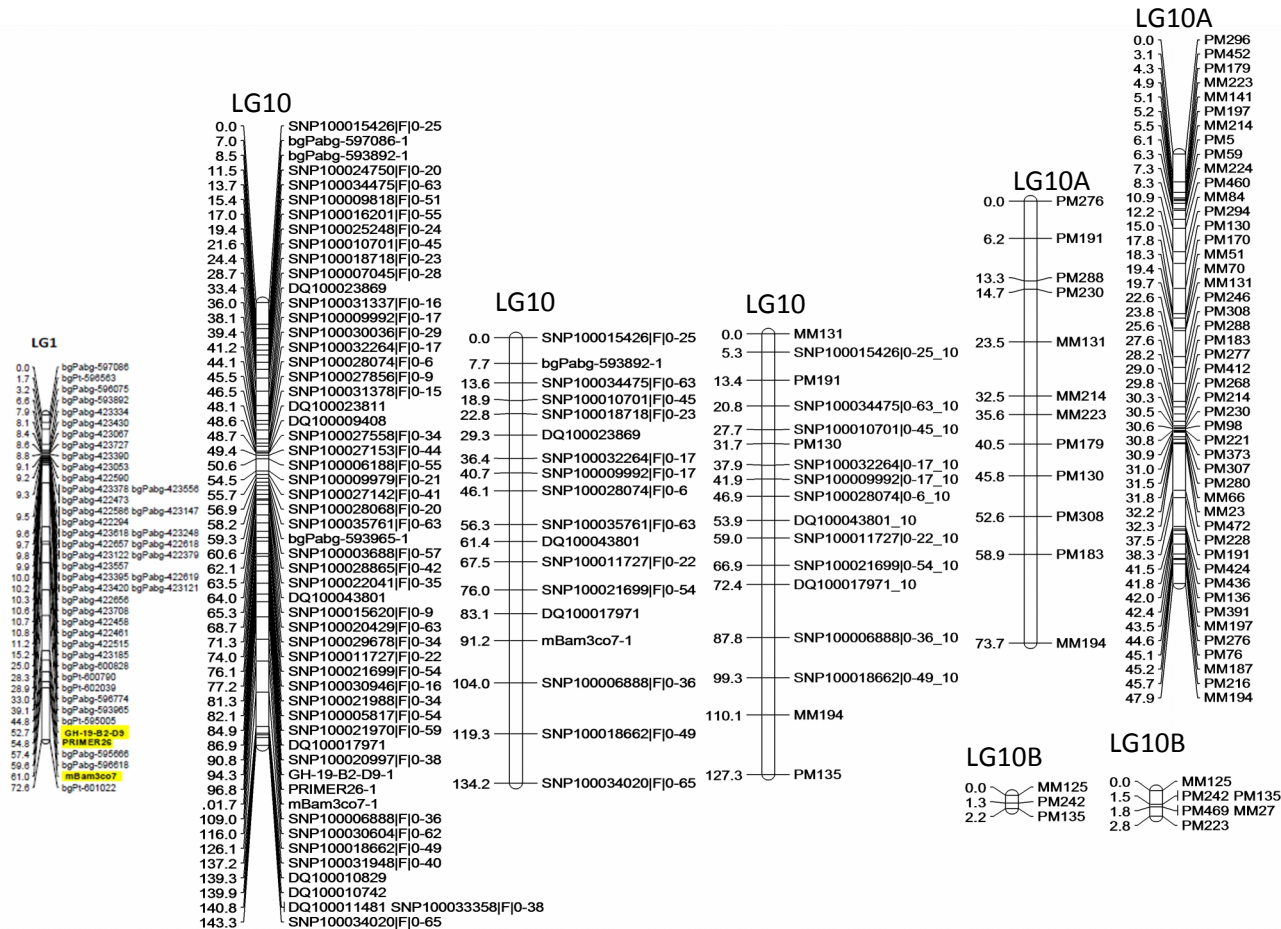
Figure 6.8(j) The graphical comparison of integrated map with original map for LG10. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
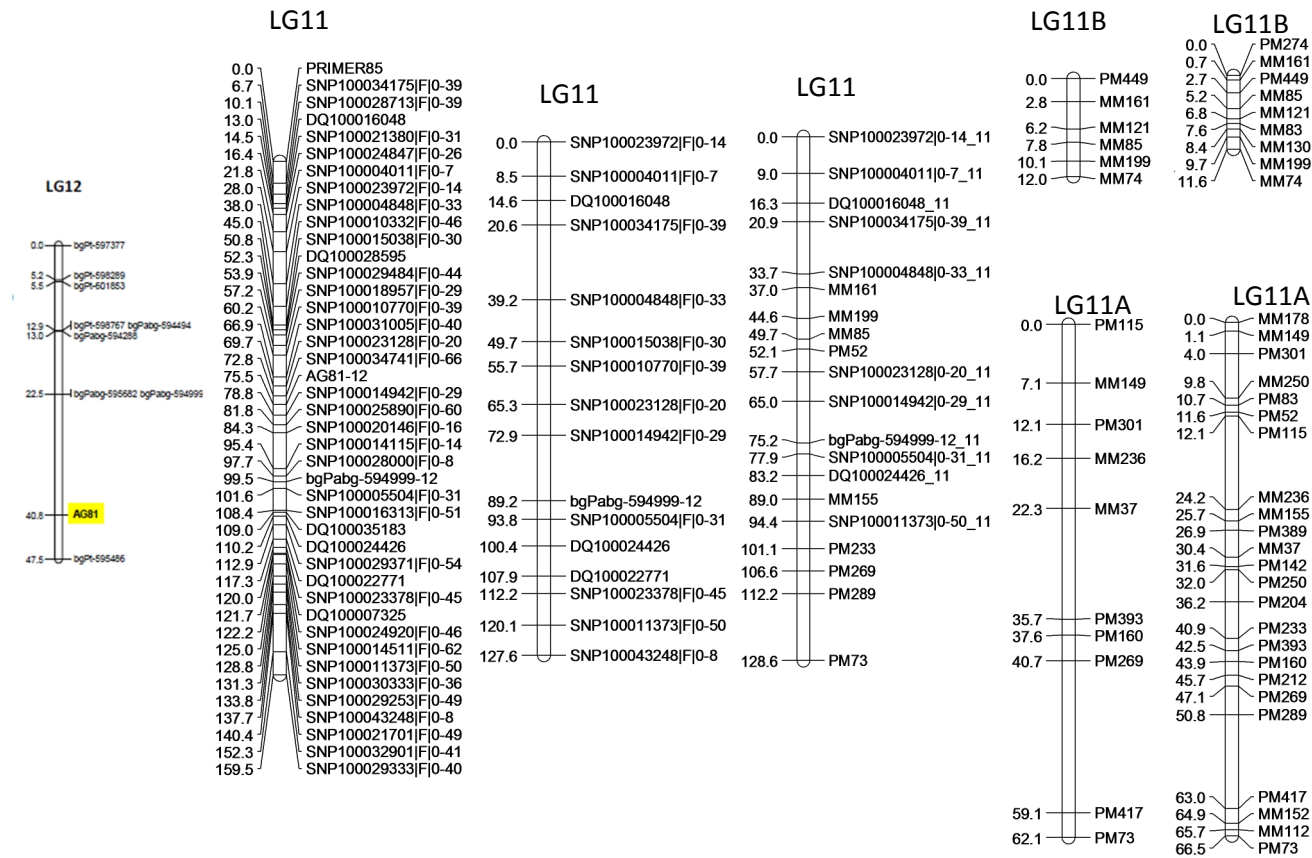
Figure 6.8(k) The graphical comparison of integrated map with original map for LG11. Left to right: Respective LG derived from Ahmad (2012), DArTseq high density map, DArTseq framework map, integrated map, GEMs framework map and GEMs high density map.

*high density map: map before removing markers; framework map: map with limited number of markers; integrated map: combining DArTseq and GEMs framework maps.
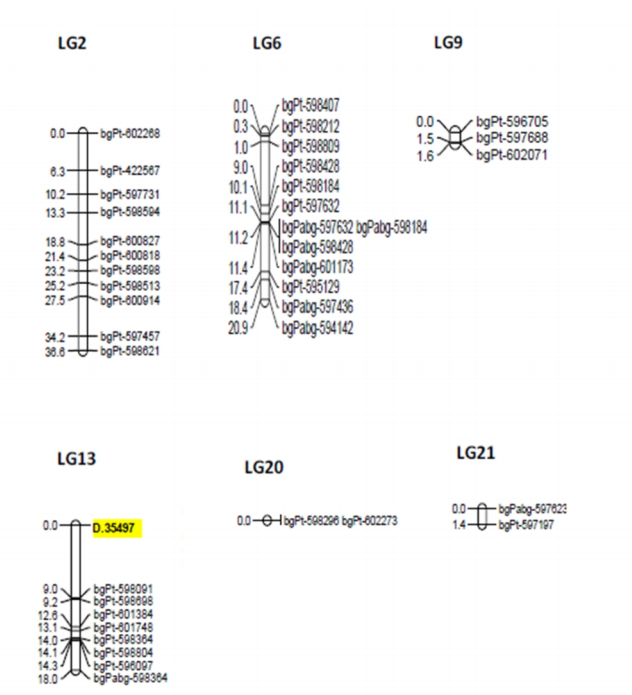
Figure 6.9 The remaining unmapped linkage groups from Ahmad (2012).

6.4   DISCUSSION

6.4.1 Novel GEMs generated using the soybean Affymetrix GeneChip

The use of a microarray offers the potential to obtain both gene expression variation (pseudo phenotypic data) and genotypic markers for construction of a genetic linkage map simultaneously, allowing the identification of thousands of eQTL in a single experiment. When integrated with trait QTLs, the causal loci within the genomic regions and the hypothetic regulatory networks controlling phenotypic variation could be analysed. As reported in Calvino *et al.* (2009), a cross-species hybridisation approach was adopted for less studied crop species, such as sorghum, to exploit markers that were differentially expressed between two parents as well as to identify the candidate genes that have functions related to sugar and cell wall metabolism. Despite the transcriptome sequence information for Bambara groundnut not being fully available and the lack of a genome sequence, the development of a novel marker system through cross-hybridising Bambara groundnut RNA onto the Soybean GeneChip contributed to the construction of a high density genetic map and will lead to subsequent trait QTL and eQTL analyses studies for Bambara groundnut.

In general the development of GEMs is based on the average hybridisation signal produced from a single probe-set, which is represented by 11 probe-pairs when the Affymetrix GeneChip is used or by a number of other features when an Agilent chip is used (usually 1-3 60-mer probes). The first approach in which hybridisation signal was measured at the level of the soybean probe-sets gave limited results as only 15 potential GEMs were identified out of 61,035 genes. The fact that only 0.02% markers were found to be suitable candidates for GEMs may be due to the hybridisation of RNA samples from Bambara groundnut onto a heterologous Soybean genome microarray (evolutionary separation of the two species being around 20 million years; Cannon *et al.,* 2009) leading to underestimation of hybridisation signals at the

probe-set level. Despite reasonably high signal strength that might be generated by one probe-pair in probe set, the hybridisation signal is averaged across all probe-pairs that represent that probe set. Poor hybridisation to other probe-pairs could reduce the overall mean and as a result, a relatively low distinctness score (used to differentiate between the 'a' and 'b' alleles in the segregating population) could be produced, resulting few GEMs being identified. The results improved when probe-sets were masked with a custom made .CDF file to remove probe-pairs with poor hybridisation signal. Forty-eight genes out of 53,651 (0.09%) were selected as GEMs.

However, the number of GEMs generated at probe-set (15) and CDF masked probe-set level (48) was insufficient for use in genetic linkage analysis as a single marker type. Therefore, the development of GEMs at the probe-pair level was established to overcome the likely signal damping effect resulting from averaging of signal across all probe-pairs in each probe-set. From the list of probe-pairs with different distinctness scores, differential expression of the probe-pairs from the same probe set was also discovered, for example, Gma.12360.1.S1_at; PM-394638, Gma.12360.1.S1_at; PM-1346833 and Gma.12360.1.S1_at; PM-432117 which showed distinctness scores of 3.72, 2.08 and 1.73. The analysis of the hybridisation signal data at the probe-pair level offers an advantage in terms of retrieving potential probe-pairs with a high distinctness score and to remove poorly hybridised probe-pairs from each probe-set in order to obtain as much information as possible for GEMs development. When the hybridisation signal is analysed at probe-set level, there is a possibility of not using probe-sets containing probe-pairs with high distinctness scores but where there is far lower distinctness across the entire probe-set. Such markers could have the potentially to be selected as GEMs.

Of 669,982 probes, 1,030 PM probes and 501 MM probes (0.23%) were chosen as potential GEMs when the analysis was conducted at the probe-pair level. The result is in agreement with Hammond *et al.* (2011) as 838 putative

GEMs out of 92k transcripts (<1%) from the Agilent Brassica 60-mer array were selected. Although the approach adopted was slightly different, West *et al.* (2006) also reported the selection of <1% of the genes from Affymetrix ATH1 GeneChip for ELPs. In addition, as gene expression is highly dependent on the environment and condition of the tissues samples, it is believed that not all potential differentially expressed genes are expressed in a single experiment (West *et al.,* 2006). The cross-species analysis has not only to contend with the small number of genes actually showing a DNA (sequence) or RNA (expression level) based difference during a homologous species-chip eQTL analysis, but must also contend with lower signal strength due to evolutionary distance between target species and microarray used.

A distinctness score is used to enrich for the separation of 'a' and 'b' allele scores across the individual lines, allowing the probe-pairs to be selected as a potential GEM. As shown in Appendix 7, a high distinctness score could distinguish between 'a' and 'b' allele across individual lines and assign them into two distinct groups. For example, Gma.3025.1.S1_at; PM-933459 with a distinctness score of 7.99 obtained two distinct groups that showed hybridisation signal as high as 3200 and as low as 6, respectively. One of the possible causes for hybridisation signal differences observed in each probe-pair on a microarray chip between individual lines could be due to the binding of the detecting oligonucleotide to a repetitive sequence. A transposon that inserts itself into a functional gene can cause the modification of the gene sequences and prevent expression or effect the function of the gene, could be one of may causes of gene inactivation (Belancio *et al.,* 2008). For decreased values of the distinctness score, the distribution of two distinct groups becomes more scattered and there is the potential of having hybridisation signal from some individual lines falling in between the two distinct groups, for instance, GmaAffx.71175.1.S1_s_at; PM-1195578 with a distinctness score of 3.82. This noise could be due to technical errors such as the strength of hybridisation of

nucleic acids onto individual GeneChip (although all chips were normalised before analysis). Thus, a series of cut-off points are set during data analysis in order to remove probe-pairs with poor performance, very similar signal or with high scatter across lines.

The initial examination of hybridisation patterns using PIGEONS also provides a preliminary insight into *cis* or *trans* variation. Differences in hybridisation signal observed at the same probe-pairs observed between parental lines at both DNA and RNA levels provisionally suggests that a polymorphism in the detecting probe-pair could be affecting hybridisation signal strength between the two parental lines and the segregation in the offsprings could be due to sequence differences (Figure 6.5). In this case, the observation might be more likely to indicate the detection of *cis* polymorphism. *Cis* variation in the structural gene or nearby could influence transcript stability, the transcription process and also expression of downstream structural gene *in trans* (Kliebenstein, 2009). For *trans*-acting element, such as a transcription factor, the causal agent of any sequence polymorphism detected by the probe-pair would not co-locate with the structural gene. Therefore, differential hybridisation signal would be observed to map at a different location to the designed sequences in single probe set at the DNA and expression level. Compared to *cis*-, *trans*-acting elements are believed to often be associated with variation in the expression level of many downstream genes, with the mapping location representing the major regulator, leading to a clustering of *trans* effects away from the structural location of the genes. Such major regulators could also have pleiotropic effects (Kliebenstein, 2009).

However, the expectation here is that the parental DNA hybridisation signals will be the same with the levels varying in the RNA of the offspring, as the difference in transcript abundance is not a function of a sequence polymorphism within the oligonucleotide target site for a structural gene whose population variation is actually mapping at a *trans* location. A similar effect

would be expected for *trans* markers, as the polymorphism relates to transcript abundance differences in the population, rather than sequence differences in the detecting oligonucleotides. It is always possible that both mechanisms could be active, particularly if the sequence polymorphism also effects the stability or processing of the transcript detected by the sequence.

Using the PIGEONS software is not sufficient to provisionally identify *cis-/trans*-acting elements, although profiles which show consistent (and similar magnitude) differences between the DNA and RNA levels are likely to be sequence polymorphisms between the parental genomes and to be *cis* effects. The distinction between *cis* and *trans* can only be determined through location mapping of the differential hybridisation of the transcript. The idea of aligning gene expression profiles from the microarray analysis in Bambara groundnut onto a soybean 'pseudo physical' map could potentially determine the correlation between hybridisation signals observed and variation in gene expression, through location of *cis* and *trans* effect on the pseudogenome.

The identification of GEMs derived from RNA samples based on hybridisation signal differences could be influenced by the environment. As each experimental sample is represented by a single plant (one replicate) from each individual line in the drought treatment, it is worth bearing in mind that the use of RNA samples for cross-hybridisation with the soybean GeneChip could possibly reflect an interaction between Bambara groundnut plant and the environment of that particular plant. Indeed, in many ways this is the main point behind the analysis, to identify differential expression of genes under a mild drought stress, in this cross. For the F_5 segregating population which is a fairly stable population (6% heterozygotes expected to remain) there is a possibility that GEMs under one set of experimental conditions may not perform similarly under different environmental conditions or when different sets of tissue samples are used. Therefore, the integration of GEM markers into a framework map containing DNA-based markers which are not dependent on the conditions

under which experiment is carried out is recommended (West *et al.,* 2006). Although there is no replicate for the gene expression analysis, the use of 60 segregating lines should be sufficient to represent the population derived from DipC and Tiga Nicuru for GEMs development and provide robust segregation data within the specific experiment, at least.

6.4.2   Use of GEMs for genetic mapping

In order to evaluate marker quality and to increase the genotyping efficiency for mapping, a marker filtration process was required.   Before constructing a genetic map, the filtration and selection of GEMs was carried out to minimise missing data, remove distorted markers. During the mapping process, removal of markers showing double cross-over events in small genetic distances was also carried out.   In terms of quantity, the use of both PM probes and MM probes increases the number of markers available for genetic mapping. Of 753 GEMs (487 PM probes and 266 MM probes), 527 markers were grouped to construct the first GEMs-based genetic linkage map. As a result, a framework genetic map of 920.3 cM containing a final 165 GEMs (120 PM probes and 45 MM probes) with a spacing of 5.1 cM between adjacent markers for the F_5 segregating population in Bambara groundnut was produced.   The PM probes and MM probes in each probe-pair could have different hybridisation signals due to the single nucleotide difference present at the 13th nucleotide between the PM and MM probe of a probe-pair. This could results in a variation of the distinctness score and might give some indication of the basis of the polymorphism mapped. The combination of PM probes and MM probes in genetic mapping is then said to maximise the potential of GEMs to be mapped and markers mapped to the same location could represent PM and MM versions of the same probe-set.

The method that was used to generate GEMs provided sufficient markers to create a genetic map which was expected to have reasonable coverage for the F_5 segregating population of 60 individual lines.   The first priority in map

construction must be to use the best quality data to produce the most accurate map, then additional putative markers could be introduced using less stringent criteria, but fixing the order of the framework map. If the data quality is good, greater marker information will allow denser maps to be constructed. If the additional marker quality is poor, approximate positions can still be assigned which could be useful in any conserved synteny comparisons or subsequent fine mapping.

As GEMs are potentially 'transient' markers, the integration of GEMs into a stable DNA sequence-based framework map is recommended (West *et al.,* 2006). There are several potential advantages for integration of maps. For example, integration of maps allows the potential alignment of GEMs and thus also the DNA sequence-based framework map to the soybean physical map. In addition, the integration also facilitates some evaluation of *cis* and *trans* regulatory elements during subsequent eQTL analysis and the identification of potential *trans* hot spots.

6.4.3 Integration of the genetic map using resources at DNA and RNA level

In spite of the presence of common markers, several issues still arise during the integration of two framework maps (Stam, 1993). First, the accuracy of the estimates of recombination frequency vary between each of the data sets. Recombination is a stochastic process, it is believed to be distributed roughly according to a Poisson distribution. Second, the mapping populations for each map could be of different types, for instance an F_2 population, a backcross and doubled haploid populations. The issues also occur when all the genotype information from individual maps is pooled together prior to integration of map, as applied in the present study. Stam (1993) reported the use of 'fixed' sequences to produce an ordering which will not conflict with any sequences when additional raw data is introduced. The present study showed that no

significant differences were obtained using a 'fixed' marker order based on the DArTseq map during the integration of the GEM map using JoinMap4.1.

GEMs were scored as dominant markers in accord with the study reported in West *et al.* (2006). DArTseq map and GEM map were generated from F_3 and F_5 segregating populations, respectively, with expected residual heterozygosity of 25% and 6% respectively. The main concern of the integration of two maps at different levels of inbreeding is most likely to be the regions where individuals are still heterozygous in the F_3 (25%) and F_5 (6%) segregating populations. Therefore, it is expected that in around 19% of the map, there could be some conflict of the markers. However, by treating the F_5 population markers as dominant markers, the approach should be reasonably accurate (as only 6% of individuals are expected to be present which could be scored as co-dominants, in theory at least) and avoids the danger of mis-scoring GEMs through making assumptions about how GEM markers, heterozygous or homozygous in an individual, might present themselves in a hybridisation dataset from a particular probe-pair.

There are also some concerns about the loss of marker information from original maps in constructing an integrated map, although marker orders of integrated map in the present study were in reasonable agreement with the original maps (DArTseq map and GEM map). The resulting effects of losing markers could potentially influence the QTL analysis, particularly the accuracy of QTL position. Furthermore, insufficient marker information could also result in a gap/break in the middle of the original maps, leading to inverted orientation of parts or all of a chromosome arm when constructing the integrated map, particularly where marker data appear contradictory. For instance, an inverted marker order was observed in LG2 and LG6 between the integrated map and DArTseq map. The marker distance between bgPabg-596250-10 (65.4 cM) and bgPabg-597858-10 (84.7 cM) in LG2 from the DArTseq map was 19.3 cM,

resulting in inaccurate marker position in integrated map when JoinMap was used.

The first attempt to integrate all GEMs developed at probe-pair level into the DArTseq map, instead of into the representative markers from the framework map, was also made (result not shown). However, the clustering of markers from the DArTseq map and GEM map, respectively, was observed in the integrated map. The reason for such clustering in the integrated map is possibly due to noise introduced by the use of different datasets produced from two mapping populations, which are F_3 and F_5 segregating populations. It is also possible, that some degree of clustering of the markers could be a genuine reflection of the different distribution of markers in the Bambara groundnut genome or the mechanism used to detect the markers. For example, GEMs are based on expression patterns, while the DArT Seq method includes the use of a methylation sensitive enzyme (*Pst*I) to create a genomic representation of Bambara groundnut which is then translated into dominant DArT markers and SNPs markers. Therefore, the distribution of expressed genes within the genome (GEMs) and unexpressed, but also unmethylated, genomic fragments could be different, particularly around the centromeres.

Although the integration of the two original framework maps in this study is not complete, the construction of framework maps prior to map integration for Bambara groundnut was considered a reliable approach based on established publications (Wang *et al.,* 2006). The authors used skeleton maps derived from three mapping populations in *B. napus* to develop an integrated map. In addition, Wang *et al.* (2006) also identified the conserved collinearity blocks relative to *Arabidopsis* in the *B. napus* integrated map. The finding provides an insight into the use of the Bambara groundnut integrated map for comparative studies with Soybean, a reasonably closely related species in the legume family in order to facilitate an understanding and annotations of genes controlling the traits of interest.

Despite the incomplete integration of the two original maps, the integrated map was used to assign groupings of markers from the GEM map into LGs which correspond to the DArTseq map. The alignment of LGs in both original maps allows the two aligned maps to be used separately in the subsequent QTL analysis, with a comparative analysis of the detected effects made possible. When the identification of the positions of the QTLs using both the DArTseq map and GEM map is made, the presence of a number of probably common QTLs for important traits could be identified. A detailed integrated map is required in future works to fill in the gaps between two or more linkage groups. In addition, as dominant DArT markers and co-dominant SNPs markers are generated based on DNA sequence with the 6-base staggered *Pst*I sequence (CTGCAG) while GEMs are developed on the basis of hybridisation signal differences at transcript level, genuine effects of each mapping population at the DNA and RNA level could be revealed together in the detailed integrated map.

Chapter 7: QUANTITATIVE TRAIT LOCI (QTL) ANALYSIS

7.1   INTRODUCTION

The majority of agronomically important crop traits, such as yield, disease tolerance and drought resistance, are quantitative traits, also known as polygenic, continuous or complex traits.   Quantitative traits are usually controlled by multiple genes in which the majority of these genes have minor effects on the traits while a few genes have major effects. The identification of gene loci controlling quantitative traits (QTLs) provides an insight into potential molecular mechanisms underlying the traits as well as the genetic effect of the QTLs on the traits. This can lead to more informed classical breeding (i.e. better selection approaches or concentration on specific trait components) or the application of marker assisted selection, leading to efficient crop breeding. Using a genetic linkage map consisting of polymorphic markers and accurate phenotypic data in the segregating population, QTL analysis to map regions of genome containing genes that regulate quantitative trait can be conducted.

QTL analysis and its applications in crop plants have been widely studied. The identification and mapping of QTLs can provide a fundamental understanding of mechanisms controlling traits of interest. For instance, *days to flowering* in *Vicia faba* was first identified to be controlled by five QTLs located at chromosomes IA, III and V (Cruz-Izquierdo *et al.,* 2012). The identification of genes controlling flowering time is useful in *V. faba* to counter the effects of late frost damage and providing adequate water supply for grain filling at middle and lower latitudes (Nelson *et al.,* 2010). In soybean, Diers (1992) first reported eight protein QTLs using the $F_{2:3}$ segregating population from a cross between population A81-356022 and PI-468916. In *Lotus japonica*, the first QTL analysis using RILs developed from *Miyakojima MG-20* x *Gifu B-129* identified a total of 40 QTLs that explained some of the variation observed for thirteen agronomic traits (Gondo *et al.,* 2007). These findings could provide a genetic

understanding of the study traits of interest and also provide markers for marker-assisted breeding of important legume crops.

Furthermore, QTL mapping can improve the understanding of the domestication process in crop species, allowing putative useful genes from wild relatives to be introgressed into cultivated crops by marker assisted selection (MAS) for crop improvement. For example, domestication-related traits in soybean were discovered to be contributed to by one or two major QTLs and a number of minor QTLs when a population of 96 RILs derived from a cross between cultivated (ssp. *max*) and wild (ssp. *soja*) was subjected into QTL analysis (Liu *et al.,* 2007). One of the major QTLs (*qPD-J*) identified as accounting for variation in pod dehiscence and seed hardness was also reported to be a possible key factor leading to larger seed size during the domestication of soybean (Liu *et al.,* 2007). This has also been investigated in Bambara groundnut, where a single F_2 population derived from a cross between a domesticated landrace (DipC) from Botswana and a wild accession (VSSP11) from Cameroon was used to study domestication-related traits (Basu *et al.,* 2007a). The domestication of Bambara groundnut involved the alteration of plant morphology and agronomic traits. A number of genes are suggested to control the morphological changes from extreme spreading growth habit (wild type VSSP11) to compact growth habit (DipC) with variation observed in *internode length* in particular (Basu *et al.,* 2007a). In addition, the authors also identified *leaf area, specific leaf area, 100 seed weight* and *carbon discrimination isotope (CID)* as quantitative traits with significant variation observed domesticated by wild type offspring population. Using QTL mapping, it was possible to examine the QTLs along with their positions on the map for traits like *specific leaf area, seed weight* and *CID* (Basu *et al.,* 2007b). Once the associations between markers and QTLs have been identified, flanking markers around the QTLs could serve as the tools to improve quantitative and qualitative traits through MAS breeding.

In addition to the use of single-marker analysis and interval mapping (Chapter 1) to conduct QTL analysis, the application of multiple QTL mapping (MQM) is also suggested to identify multiple QTLs controlling components of the same trait of interest, such as southern corn rust in tropical sweet corn (Wanlayaporn *et al.,* 2013). MQM mapping was first developed by Jansen (1993) based on the multiple QTL model and offers a number of advantages over conventional interval mapping, including greater power and accuracy in detecting QTLs. However, the computational work involved in the MQM mapping is not feasible if the number of QTLs is large. Jansen (1993) proposed to select one QTL at a time and use selected markers close to the detected QTLs as cofactors, allowing them to account for the variation associated with the QTL assigned to the marker, simplifying them to 'mendelian-like' markers. The proposed approach is able to reduce the residual variance (by accounting for genetic effects at different positions in the genome) and increase the power of searching for other segregating QTLs, when the marker closely linked to a QTL that explains a large component of the genetic variation of the traits is selected as cofactor (Ooijen, 2009). When MQM mapping was used, the explained genotypic variance showed an increase of up to 6-fold compared to conventional interval mapping, indicating that part of the residual variance was accounted for by marker cofactors (Jansen, 1993). The approach of MQM mapping is reported to be similar to composite interval mapping (CIM), however MQM has advantages over CIM in terms of reducing type I error (a QTL is detected at a location when a QTL is actually absent) and type II error (a QTL is not indicated, despite one being present) during QTLs detection (Jansen *et al.,* 2010).

The example of using MQM mapping to identify QTLs and their association with markers was reported in tropical sweet corn (Wanlayaporn *et al.,* 2013). Eighty nine tropical sweet corn RILs derived from a cross between *hA9104* and *hA9035* inbred lines were subjected to QTL analysis with the MQM mapping algorithm to identify the QTLs related to southern corn rust resistance. The

authors discovered that phenotypic variation for rust resistance was explained by one major QTL, which was flanked by markers umc2025 and umc1919 on chromosome 1, as well as two minor QTLs detected on chromosome 6 and 10. Based on the example given, adopting the MQM mapping approach should be beneficial for identifying QTLs that regulate agronomic traits and also drought-related traits in Bambara groundnut.

The first QTL analysis in Bambara groundnut using a F_3 segregating population developed from the cross between DipC and Tiga Nicuru was conducted by Ahmad (2012). The author reported that a total of 37 QTLs were mapped on the DipC x Tiga Nicuru genetic linkage map, which consisted of 209 microarray-based DArT markers and 29 SSR markers, for 23 morphological and agronomical traits in Bambara groundnut. Among the traits, *internode length* was shown to be controlled by a major QTL detected on linkage group (LG) 4 with a peak LOD score of 7.9. When the LOD score is higher than the predefined value from a permutation test, the QTL is concluded to be significant (Ooijen, 2009). This significant QTL was mapped close to marker bgPabg-596988 at 3.0 cM on LG4 and explained 43.5% of the total phenotypic variation. On the same LG4, a major QTL related to *peduncle length* with a LOD score of 9.7 at 1.0 cM was also discovered and the closely linked marker was bgPt-423527 located at 2.4 cM. The morphological trait *internode length* is of importance for breeding programs. For example, Bambara groundnut landraces with bunched type (short internode length) offer easier management when Bambara groundnut plants are planted in a mixed cropping system (Ahmad, 2012). The identification of QTLs for *internode length* allows the development of planting material through MAS for use in different planting systems.

As described in Chapter 5, the construction of a new genetic linkage map using dominant DArT and co-dominant SNPs markers in addition to pre-existing microarray-based DArT and SSR markers was completed. The QTL analysis in the Bambara groundnut F_3 segregating population is expected to be improved

with a higher density genetic linkage map with better genome coverage. In addition, the generation of a gene expression marker (GEM) map (Chapter 6) also allowed QTL mapping in a Bambara groundnut F_5 segregating population to be completed. By comparing the two QTL analyses, the position and magnitude of putative QTLs for common agronomic and morphological traits will be evaluated and hence may facilitate the understanding of the genetic and molecular mechanisms underlying the traits.

7.2 MATERIALS AND METHODS

7.2.1 Plant materials

Two different generational populations of Bambara groundnut, an F_3 and an F_5 segregating population derived from the same cross between DipC and Tiga Nicuru were grown and mapped to evaluate QTLs involved in agronomic, morphological traits and drought-related traits. The phenotypic data for both F_3 and F_5 segregating populations were adopted from Ahmad (2012) and Chapter 4, respectively. Based on Ahmad (2012), 13 agronomical traits were used for QTL analysis in F_3 segregating populations whereas 16 traits recorded in Chapter 4 were adopted to study associations between markers and traits in the F_5 segregating populations. A total of 71 individual lines from the F_3 segregating population (Ahmad, 2012) were subjected to QTL analysis using an improved genetic linkage map spanning 1354.4 cM across 11 linkage groups with dominant DArT, SNPs, microarray-based DArT and SSR markers. In addition, another genetic linkage map with map length of 872.2 cM across 21 linkage groups was constructed using GEMs (Chapter 6). A total of 59 individual lines from the F_5 segregating population (Chapter 4) were used for mapping and QTL analysis. Identification of QTLs was conducted using MapQTL$^®$ v6.0 (Ooijen, 2009) with interval mapping (IM) and multiple-QTL mapping (MQM) model, where appropriate. The pre-testing and transformation of trait data (where necessary) have been discussed in Chapter 4)

7.2.2 Preparation of data files

Three types of data files were prepared prior to QTL analysis according to manual of MapQTL (Ooijen, 2009):

1. *Locus genotype file*: The file (*loc file*) contained the information of all the loci for a single segregating population. The header of the file defined four instructions: name of the population, the type of the population (F_2, RIx, BCpxFy and IMxFy), the number of loci and the number of individuals (Figure 7.1).

Figure 7.1 An example of *.loc* file used for QTL mapping.

2. *Map file*: The file contained the positions of all the loci. The grouping and order of markers were the same as the *map file* resulting from JoinMap v4.1. The *map* file had no header but is line-structured (Figure 7.2).
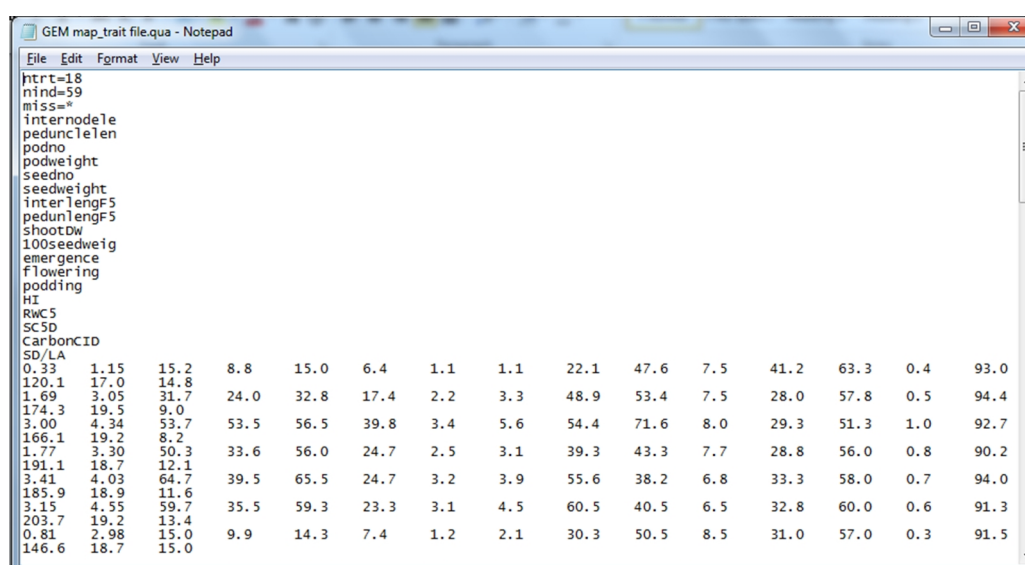
Figure 7.2 An example of *.map* file used for QTL mapping.

3. *Quantitative data file*: The file (*qua file*) consists of data of all quantitative traits for each individual. The header of the file defined three instructions, followed by the names and numerical values of the traits (Figure 7.3). The three instructions were: number of traits, number of individuals and the symbol that indicates a missing value (*). The name of the traits could not be longer than 20 characters and could not contain spaces. The number and order of the individuals should correspond to the *.loc file.* For non-normalised trait data, transformation was carried out in Chapter 4 and the transformed data was used for QTL mapping.

Figure 7.3 An example of *.qua* file used for QTL mapping

7.2.3 QTL mapping approach

Three data files were loaded into MapQTL® v6.1 followed by analysis using two mapping approaches, IM and MQM mapping, to detect and identify the QTLs. The analysis options were set by default, including using the regression algorithm for IM and MQM mapping and fitting *dominance* for the population types. The permutation test using 10,000 reiterations was first conducted in order to determine the significance threshold of the LOD score. Following the permutation test, IM mapping was carried out. The LOD score obtained from IM

mapping was compared with the Genome Wide (GW) threshold at $p \leq 0.05$ from the permutation test. Significant QTLs were identified if the LOD score was equivalent or higher than GW threshold. However, QTLs were considered as 'putative' when the LOD score was lower than GW threshold by up to a one LOD interval. Once QTLs with significant LOD scores were identified from IM mapping model, the closest linked marker was selected as a cofactor prior to MQM mapping. The positions of QTLs picked up by marker cofactors were verified through visual inspection of LOD profile and LOD table produced by MapQTL v6.0.

7.3 RESULTS

7.3.1 Detection of QTLs in the F_5 segregating population using the GEM map

A total of 16 traits relating to agronomy, morphology and drought response were analysed for QTL based on the GEMs genetic linkage map. The MQM mapping results produced a total of 10 QTLs, 6 significant QTLs and 4 putative QTLs, associated with 10 studied traits distributed over 4 linkage groups including LG1, LG2B, LG8B and LG11A (Figure 7.4).
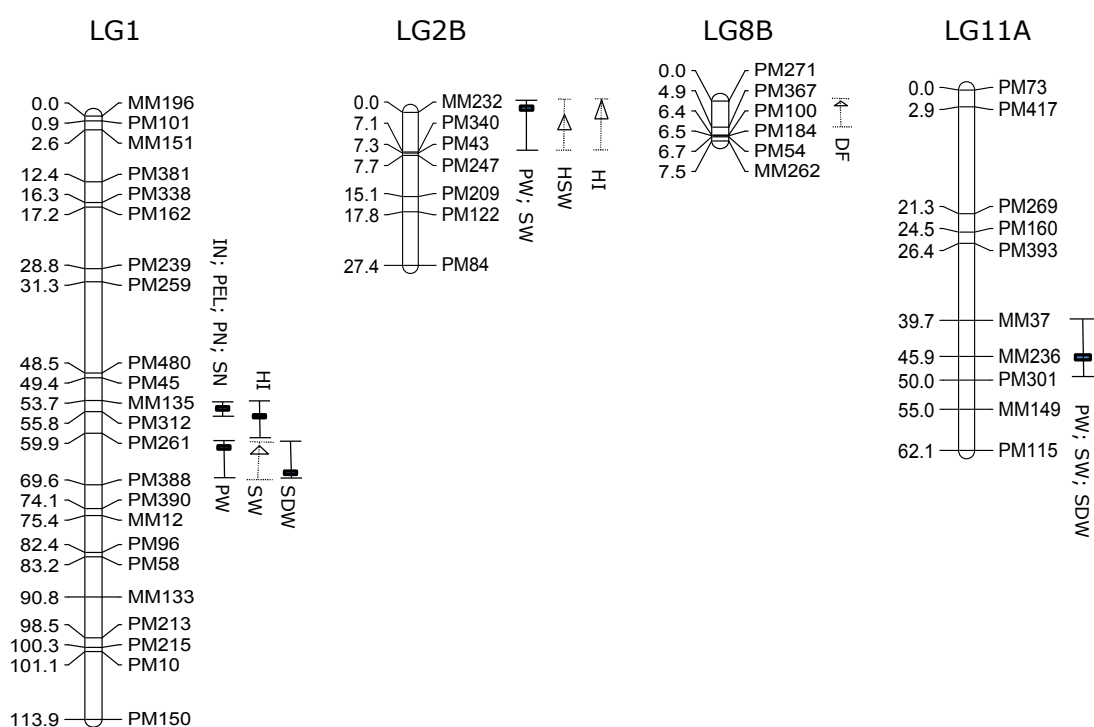
Figure 7.4 Map positions of the QTLs across four linkage groups in the F_5 segregating population developed from a cross between DipC and Tiga Nicuru. GEMs identity is described on the right and map positions (cM) on the left. The rectangular box (▬) with the solid confidence intervals indicated the location of significant QTL and their flanking markers whereas triangular boxes (Δ) with dotted confidence intervals represent putative QTLs and their neighbouring markers. *DF, days to flowering; IN, internode length; PEL, peduncle length; PN, pod number per plant; PW, pod weight per plant; SN, seed number per plant; SW, seed weight per plant; HSW, 100-seed weight; SDW: shoot dry weight; HI, harvest index.*

Most of the QTLs were clustered, especially on LG1 and LG2B. Of 10 QTLs, 5 QTLs (4 significant QTLs and 1 putative QTLs) were located on LG1 whereas the other 5 QTLs were identified on LG2B (1 significant QTLs and 2 putative QTLs), LG8B (1 putative QTL) and LG11A (1 significant QTL). Some of the QTLs had overlapping confidence intervals, opening the possibility that they are being influenced by the same underlying gene. For example, QTLs controlling four traits *internode length, peduncle length, pod number per plant* and *seed number per plant* were detected at loci closely linked with MM135 (53.7 cM) on LG1. In addition, MM236 (45.9 cM) on LG11A was also found to be linked to loci that mapped with QTLs controlling *pod weight per plant*, *seed weight per plant* and *shoot dry weight*. Some of the traits were shown to be controlled by multiple loci across different linkage groups. For instance, *pod weight per plant* was mapped at three loci with closely linked markers PM261 (59.9 cM), MM232 (0.0 cM) and MM236 (45.9 cM) on LG1, LG2B and LG11A respectively.

In addition to graphical presentation of the QTL location, a summary of QTLs associated with 16 studied traits, LOD score, position of QTLs, location of nearest markers, phenotypic variation explain (PVE) and additive effect is also presented in Table 7.1. Based on the result, the distribution of QTLs and their effects for each trait are described as below:

Days to flowering (DF): A putative QTL for *days to flowering* located at 0.0 cM on LG8B was identified. The QTL had a LOD score (3.8) lower than GW threshold (4.8) by 1 LOD interval and explained 25.1% of the total phenotypic variation. Marker PM271 at 0.0 cM was the nearest marker to this locus.

Internode length (IN): A single significant QTL for *internode length* was identified at 54.7 cM on LG1 with a LOD score of 7.28 and PVE of 43.4%. The closest marker to the locus was reported as MM135 (53.7 cM).

Peduncle length (PEL): A QTL for *peduncle length* was also mapped at 54.7 cM on LG1 with a LOD score of 9.52 and PVE of 52.4%. The closest marker to the locus was observed as MM135.

Pod number per plant (PN): A single significant QTL for *pod number per plant* was mapped at locus 53.7 cM on LG1 with PVE of 26.5%. In addition, MM135, a closely linked marker to the significant QTL, was found to be associated with *internode length* and *peduncle length.*

Pod weight per plant (PW): Multiple loci mapped on LG1 (62.9 cM), LG2B (2.0 cM) and LG11A (46.9 cM) were detected to control *pod weight per plant*. All three significant QTLs with LOD scores of 4.04, 3.89 and 4.16 accounted for 17.5%, 17.2% and 17.9% of total phenotypic variation respectively. The nearest markers were shown to be PM261 (59.9 cM) on LG1, MM232 (0.0 cM) on LG2B and MM236 (45.9 cM) on LG11A.

Seed number per plant (SN): A significant QTL for *seed number per plant* was discovered at 53.7 cM on LG1, which was also in the confidence intervals of the loci that control *internode length, peduncle length* and *pod number per plant.* The QTL had a maximum LOD score (4.82) which is higher than the GW threshold (3.70) and contributed 31.4% of total phenotypic variability.

Seed weight per plant (SW): One putative QTL and two significant QTLs were detected for *seed weight per plant* on LG1 (62.9 cM), LG2B (2.0 cM) and LG11A (46.9 cM), overlapping with the confidence interval of QTLs detected for *pod weight per plant.* The QTL located on LG1 was considered as putative with a LOD score within the one LOD drop interval (2.61) as compared to the GW threshold (3.60). Three QTLs accounted for 11.5%, 21.9% and 15.9% of total phenotypic variation accordingly.

100-seed weight (HSW): A putative QTL for *100-seed weight* at 4.0 cM on LG2B was identified with PVE of 21.7%. The nearest marker to the QTL was MM232 at 0.0 cM.

Shoot dry weight (SDW): The trait QTL was mapped at 66.9 cM on LG1 and 46.9 cM on LG11A, overlapping with the confidence interval of QTLs that controlled *pod weight per plant* and *seed weight per plant.* The significant QTLs were linked to markers PM338 (69.6 cM) and MM236 (45.9 cM) and accounted for 24.0% and 23.0% of the trait variation, respectively.

Harvest index (HI): One significant QTL and one putative QTL were identified for *harvest index* on LG1 (55.7 cM) and LG2B (2.0 cM) individually. The significant QTL was linked to PM312 (55.8 cM) and accounted for 22.5% of the phenotypic variation of the trait. The putative QTL showed a lower LOD score of 2.86 compared to GW threshold (3.70) and PVE of 13.5%.

However, for *days of emergence (DE), estimated days of podding (EDP), relative water content (RWC), stomatal conductance (SC), carbon isotope discrimination analysis (CID)* and *stomatal density (SD),* as the LOD score was lower than GW threshold generated from permutation test at $p \leq 0.05$ by more than 1 LOD interval, so no putative or significant QTLs were identified.

Table 7.1 QTLs for 16 traits involved in agronomic, morphology and drought traits detected in a F_5 segregating population derived from a cross between DipC and Tiga Nicuru.

*Traits	QTL-LG	Position (cM)	Nearest marker	LOD	PT	PVE%	Additive effect
DE	7	8.5	PM403 (7.5 cM)	2.20^{ns}	5.00	15.80	-0.05
DF	8B	0.0	PM271 (0.0 cM)	3.83^{p}	4.80	25.10	1.31
EDP	8B	6.4	PM100 (6.4 cM)	1.95^{ns}	3.70	12.10	1.16
	11B	4.2	MM130 (4.2 cM)	1.94^{ns}	3.70	14.00	1.22
IN	1	54.7	MM135 (53.7 cM)	7.28	3.70	43.40	-0.66
PEL	1	54.7	MM135 (53.7 cM)	9.52	3.70	52.40	-1.13
PN	1	53.7	MM135 (53.7 cM)	3.94	3.80	26.50	-12.27
PW	1	62.9	PM261 (59.9 cM)	4.04	3.80	17.50	-10.00
	2B	2.0	MM232 (0.0 cM)	3.89	3.80	17.20	9.09
	11A	46.9	MM236 (45.9 cM)	4.16	3.80	17.90	10.46
SN	1	53.7	MM135 (53.7 cM)	4.82	3.70	31.40	-14.33
SW	1	62.9	PM261 (59.9 cM)	2.61^{p}	3.60	11.50	-5.81
	2B	2.0	MM232 (0.0 cM)	4.59	3.60	21.90	7.36
	11A	46.9	MM236 (45.9 cM)	3.60	3.60	15.90	7.07
HSW	2B	4.0	MM232 (0.0 cM)	3.14^{p}	3.90	21.70	5.79
SDW	1	66.9	PM338 (69.6 cM)	4.20	3.70	24.00	-8.90
	11A	46.9	MM236 (45.9 cM)	4.14	3.70	23.00	9.50
HI	1	55.7	PM312 (55.8 cM)	4.41	3.70	22.50	-0.12
	2B	2.0	MM232 (0.0 cM)	2.86^{p}	3.70	13.50	0.09
RWC	4A	18.2	PM431 (18.2 cM)	2.10^{ns}	3.80	15.10	0.10
SC	2A	39.9	PM97 (39.9 cM)	2.10^{ns}	3.80	15.20	14.14
CID	2B	0.0	MM232 (0.0 cM)	2.49^{ns}	3.80	18.20	-0.39
SD	9	47.7	MM238 (42.7 cM)	2.16^{ns}	3.70	15.50	0.71

ns: non-significance at p≤0.05 by permutation test using 10,000 reiterations.

p: putative QTLs where LOD score is lower than the GW threshold by up to a 1 LOD interval.

PT: permutation test threshold using 10,000 reiterations at p≤0.05.

*DF, days to flowering; IN, internode length; PEL, peduncle length; PN, pod number per plant; PW, pod weight per plant; SN, seed number per plant; SW, seed weight per plant; HSW, 100-seed weight; SDW: shoot dry weight; HI, harvest index.

7.3.2 Comparison of the QTL analyses between the F_3 and F_5 segregating populations

The QTL analysis in the F_3 segregating population was conducted using an improved genetic linkage map constructed with dominant DArT and co-dominant SNPs markers in addition to pre-existing markers such as microarray-based DArT and SSR. Of 13 traits subjected into QTL analysis, MQM mapping identified QTLs for four of the traits, *terminal leaf length* (one significant QTL)*, internode length* (one significant QTL)*, peduncle length* (one significant QTL) and *stem number per plant* (one putative QTL) across two linkage groups including LG1 and LG8. The QTLs associated with 13 studied traits, LOD score, closely linked marker, position of QTLs, PVE and additive effect are summarised in Table 7.2. According to the result, the distribution of QTLs and their effects for the four traits mapped with QTLs are described as below:

Terminal leaf length (TLL): A significant QTL for *terminal leaf length* with PVE of 25.5% was identified at 53.2 cM on LG8. The nearest marker to the loci controlling *terminal leaf length* was reported as DQ100020360 at 54.3 cM.

Stem number per plant (STN): A putative QTL located at 24.9 cM on LG1 was identified for *stem number per plant*. The QTL had a LOD score lower than GW threshold by 0.55 and explained 22.7% of the total phenotypic variation. Marker SNP100032012|F|0-35 at 29.1 cM position was the nearest marker to this locus.

Internode length (IN): The trait was mapped with one significant QTL located at 44.1 cM on LG1. The significant QTL was linked with marker bgPabg-596988-4, had a LOD score of 6.05 and accounted for 37.1% of phenotypic variation.

Peduncle length (PEL): QTL for *peduncle length* was mapped at 50.4 cM on LG1 and linked with the same marker used for *internode length,* bgPabg-596988-4 at 45.4 cM. The significant QTL had a LOD score of 6.6, explaining 39.8% of total phenotypic variation.

For the remaining nine traits, *petiole length, terminal leaf width, plant spread, number of nodes per plant, pod number, pod weight, seed weight, shoot dry weight* and *leaf area,* the distribution of QTLs and their effects were considered as not significant. The non-significant QTLs obtained for the traits are determined based on their LOD scores, which are lower than the respective GW threshold that was generated from permutation test at $p \leq 0.05$ by more than 1 LOD interval.

Table 7.2 QTLs for 13 traits involved in agronomic and morphology detected in a F_3 segregating population derived from a cross between DipC and Tiga Nicuru.

*Traits	QTL-LG	Position (cM)	Nearest marker	LOD	PT	PVE%	Additive effect
PL	3	18.69	SNP100028314|F|0-45 (18.7 cM)	1.82[ns]	3.90	12.0	0.66
TLL	8	53.19	DQ100020360 (54.3 cM)	3.83	3.80	25.5	-0.52
TLW	3	18.69	SNP100028314|F|0-45 (18.7 cM)	1.79[ns]	3.90	11.3	0.18
PS	1	51.39	DQ100018157 (55.1 cM)	2.70[ns]	4.00	18.7	-3.45
STN	1	24.91	SNP100032012|F|0-35 (29.1 cM)	3.35[p]	3.90	22.7	1.14
NN	4	74.39	SNP100007131|F|0-27 (73.4 cM)	2.17[ns]	3.80	13.1	-1.07
	1	91.10	mBam3co7-1 (91.2 cM)	2.09[ns]	4.00	14.8	1.04
IN	1	44.06	bgPabg-596988-4 (45.4 cM)	6.05	3.90	37.1	-0.73
PN	10	91.18	mBam3co7-1 (91.2 cM)	2.41[ns]	3.80	16.9	8.04
PEL	1	50.39	bgPabg-596988-4 (45.4 cM); DQ100018157 (55.1 cM)	6.60	3.80	39.8	-0.86
PW	10	43.68	SNP100028074|F|0-6 (46.1 cM)	2.61[ns]	4.00	18.1	8.43
SW	10	44.68	SNP100028074|F|0-6 (46.1 cM)	2.43[ns]	3.80	17.0	0.62
SDW	10	42.68	SNP100009992|F|0-17 (40.7 cM)	2.86[ns]	4.00	19.7	0.84
LA	11	112.25	SNP100023378|F|0-45 (112.3 cM)	1.91[ns]	3.80	12.4	4.81
	1	119.72	SNP100008049|F|0-52 (123.6 cM)	1.53[ns]	4.00	11.1	-4.64

ns: non-significance at p≤0.05 by permutation test using 10,000 reiterations.

p: putative QTLs whereby LOD score was lower than GW threshold by 0.1 to 1 interval.

PT: permutation test using 10,000 reiterations at p≤0.05.

*PL, petiole length; TLL, terminal leaf length; TLW, terminal leaf width; PS, plant spread; STN, stem number per plant; NN, number of node per plant; IN, internode length; PN, pod number; PEL, peduncle length; PW, pod weight; SW, seed weight; SDW, shoot dry weight; LA, leaf area.

The segregating populations derived from the same cross between DipC and Tiga Nicuru with two different generations were compared to verify the location of QTLs for common traits that were recorded and used in both QTL analysis: *internode length, peduncle length, pod number, pod weight, seed weight* and *shoot dry weight* (Table 7.3)*.* Of six common traits used, *internode length* and *peduncle length* were mapped with strong QTLs located on LG1 across two segregating populations. For *internode length,* the significant QTL (LOD= 7.28) was identified at 54.7 cM on LG1 in the F_5 segregating population whereas the significant QTL (LOD=6.05) was mapped at position of 44.1 cM on LG1 in the F_3 segregating population. The mapping position differed by over 10.0 cM for *internode length* between two segregating populations, although the confidence intervals of these two positions overlapped, suggesting that they are likely to represent the same effect. In addition, the QTL mapped for *peduncle length* on LG1 was consistent across two segregating populations, with a 4.0 cM difference in the maximum LOD and within the confidence interval overlap for both crosses. A significant QTL (LOD=9.52) for *peduncle length* was mapped at the same position as *internode length* (54.7 cM) in the F_5 segregating population whereas a QTL for this trait (LOD=6.60) was detected at 50.4 cM on LG1, 6.0 cM away from QTL associated with *internode length* (44.1 cM), in the F_3 segregating population. Despite the difference in mapping distance, the markers bgPabg-596988-4 and MM135 derived from the F_3 and F_5 segregating populations, respectively, were closely linked to both QTLs associated with *internode length* and *peduncle length.*

Table 7.3 The comparison of QTL analysis between F_3 and F_5 segregating population derived from the same cross between DipC and Tiga Nicuru.

F3 segregating population							*Traits	F5 segregating population						
Additive effect	PVE%	PT	LOD	Nearest marker	Position (cM)	QTL-LG		QTL-LG	Position (cM)	Nearest marker	LOD	PT	PVE%	Additive effect
-0.73	37.10	3.90	6.05	bgPabg-596988-4 (45.4 cM)	44.1	1	IN	1	54.7	MM135 (53.7 cM)	7.28	3.70	43.40	-0.66
-0.86	39.80	3.80	6.60	bgPabg-596988-4 (45.4 cM); DQ100018157 (55.1 cM)	50.4	1	PEL	1	54.7	MM135 (53.7 cM)	9.52	3.70	52.40	-1.13
8.04	16.90	3.80	2.41[ns]	mBam3co7-1 (91.2 cM)	91.1	10	PN	1	64.9	PM261 (59.9 cM);PM388 (69.6 cM)	3.50[p]	3.80	23.90	-12.94
								1	53.7	MM135 (53.7 cM)	3.94	3.80	26.50	-12.27
8.43	18.10	4.00	2.61[ns]	SNP100028074|F|0-6 (46.1 cM)	43.7	10	PW	1	62.9	PM261 (59.9 cM)	4.04	3.80	17.50	-10.00
								2B	2.0	MM232 (0.0 cM)	3.89	3.80	17.20	9.09
								11A	46.9	MM236 (45.9 cM)	4.16	3.80	17.90	10.46
0.62	17.00	3.80	2.43[ns]	SNP100028074|F|0-6 (46.1 cM)	44.7	10	SW	1	62.9	PM261 (59.9 cM)	2.61[p]	3.60	11.50	-5.81
								2B	2.0	MM232 (0.0 cM)	4.59	3.60	21.90	7.36
								11A	46.9	MM236 (45.9 cM)	3.60	3.60	15.90	7.07
0.84	19.70	4.00	2.86[ns]	SNP100009992|F|0-17 (40.7 cM)	42.7	10	SDW	1	66.9	PM338 (69.6 cM)	4.20	3.70	24.00	-8.90
								11A	46.9	MM236 (45.9 cM)	4.14	3.70	23.00	9.50

ns: non-significance at p≤0.05 by permutation test using 10,000 reiterations; p: putative QTLs whereby LOD score was lower than GW threshold by 0.1 to 1 interval; PT: permutation test using 10,000 reiterations at p≤0.05.

*IN, internode length; PEL, peduncle length; PN, pod number; PW, pod weight; SW, seed weight; SDW, shoot dry weight.

In addition, the comparison of QTLs controlling *internode length* and *peduncle length* across the two segregating populations is presented graphically (Figure 7.5). The comparison also included an initial QTL analysis reported by Ahmad (2012) on the F_3 segregating population using a genetic linkage map constructed with microarray-based DArT and SSR markers. The result showed that *internode length* and *peduncle length* associated QTLs in the DArT map were located at 3 cM and 2.4 cM on LG4 (Ahmad, 2012), which corresponded to LG1 in the improved DArTseq map (Chapter 5). Despite variation in map position of the markers and QTL between the DArT map and DArTseq map in the F_3 segregating population, marker bgPabg-596988 was a common marker that linked to QTL controlling *internode length* as it was found across two genetic linkage maps in the F_3 segregating population. The differences in absolute position could result from the degree of marker density in each genetic map. The present DArTseq map (Chapter 5) is an improved genetic linkage map with higher marker density after adding dominant DArT and SNP marker into the pre-existing DArT map (Ahmad, 2012). Comparing the location of marker bgPabg-596988 and MM135 from the improved DArTseq map and the GEM map, respectively, the loci controlling *internode length* and *peduncle length* are consistent due to the overlap of the confidence intervals observed for these two positions. In this case, the integration of two maps derived from two different generations of segregating populations is important in identifying potential positions of QTLs controlling traits of interest for detailed comparison in Bambara groundnut.
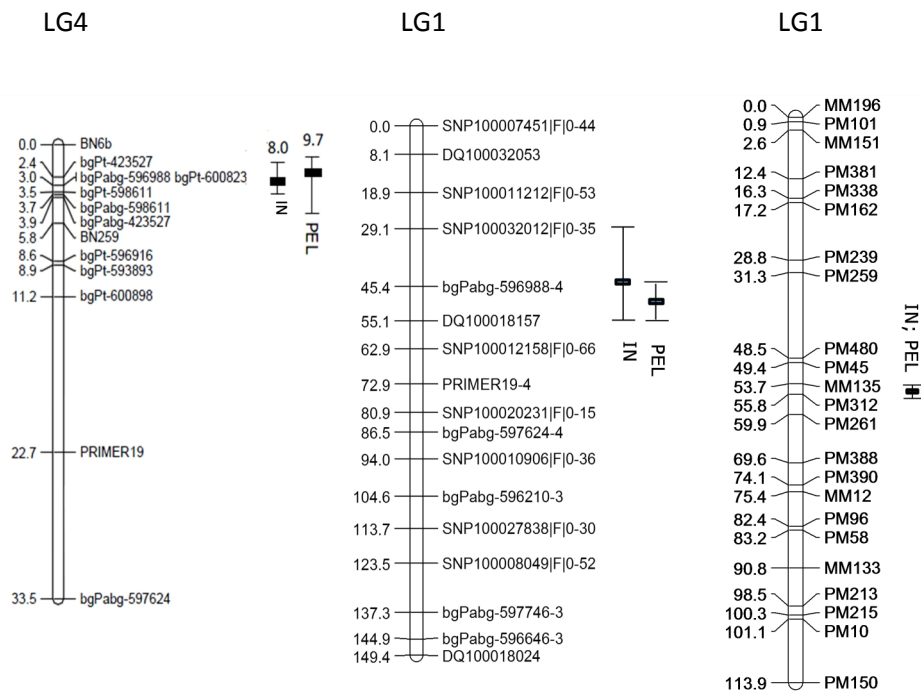
Figure 7.5 Map positions of the QTLs for *internode length* and *peduncle length* across the three genetic linkage maps in the F_3 and F_5 segregating populations derived from a cross between DipC and Tiga Nicuru. Left to right: genetic linkage group in the F_3 segregating population constructed using microarray-based DArT and SSR marker (Ahmad, 2012), the improved genetic linkage map (DArTseq map) in the F_3 segregating population with the addition of DArTseq and SNPs marker (Chapter 5) and the GEM map from F_5 segregating population (Chapter 6). *IN, internode length; PEL, peduncle length*.

7.4   DISCUSSION

7.4.1   The MQM mapping algorithm

The MQM mapping model, which uses marker cofactors to absorb the detected variance explained by a located QTL, allows a genome-wide search for additional effects who residual variance is partly masked by the detected QTL. If the genetic variance from the current QTL at the detected position can be removed from the remaining phenotypic variance, then the residual phenotypic variation is reduced, resulting in a more powerful analysis with decreased error or unexplained residuals. The application of the MQM mapping model potentially allows multiple QTLs to be identified for a given trait and mapped more accurately as compared to conventional IM mapping. Using *internode length* and *peduncle length* as examples, the power of MQM mapping in detecting QTLs is graphically presented in Figure 7.6.   The results show that MQM mapping produced a smaller confidence interval for the position of the detected QTL (53.0 cM to 55.0 cM) compared to IM mapping (45.0 cM to 55.0 cM) after residual variance was absorbed by cofactors. However, it is important to bear in mind that MQM is a model and if any of the assumptions underlying the model are incorrect, the location identified by MQM could be quite misleading.

GW thresholds generated from permutation tests at $P \leq 0.05$ appear to be high, such as a GW threshold of 5.00 for *days to emergence*, when the two segregating populations were subjected to QTL analysis.   The reason is suspected to be the result of having small population sizes in the F_3 ($n=71$) and F_5 segregating populations ($n=59$).   Thus, 'putative' QTL were also included in the QTL analysis when the LOD score was within a 1 LOD drop from the expected GW threshold in order to reduce the possibility of losing potential QTL. However, the consistency of QTLs mapped for traits of interest would need to be further examined using a larger sample size.

Internode length
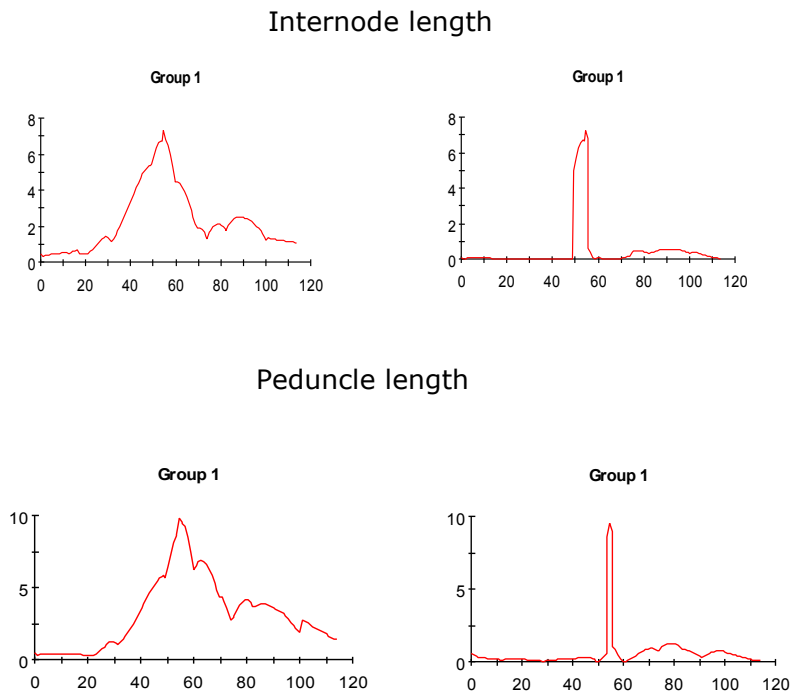
Peduncle length

Figure 7.6 The comparison of LOD profiles between IM mapping (left) and MQM mapping (right) for *internode length* and *peduncle length.* X-axis: Haldane mapping position (cM); Y-axis: LOD score.

There have been some concerns about QTL mapping in terms of map distances, number of markers, missing or incomplete (dominant) marker genotypes and combining over populations. Theoretically, the map is used to calculate the likelihood of the QTL position based on marker genotypes flanking the estimated position of QTL. The correct map distance is important in order to enhance the power of detecting multiple QTLs on the map. However in practice there may be missing observations and even mapping errors, resulting in incorrect map distances. The impacts of these deviations on the resulting LOD scores would lead to underestimation or overestimation of QTL positions, depending on the accuracy of trait data (Ooijen, 2009).

The number of markers could be increased in the genomic regions where a segregating QTL is identified with significant LOD score in order to determine

the effect and location of QTL more precisely within the region. However, the adding of markers is argued to be not necessary when the distance between the flanking markers of identified QTL is less than 5cM as the likelihood of QTL positions would probably remain the same within a short mapping distance (Ooijen, 2009). The precision of QTL locations depends more on the sample size than on the number of markers and the quality of the trait data. In addition, the addition of dominant markers could also probably increase the memory requirements of computers (Ooijen, 2009) as well as lead to a loss of information where the population is expected to contain heterozygotes. For example, when a dominant cofactor marker is selected in MQM mapping, twice of the normal space in the design matrix used in computations is needed (Ooijen, 2009). Subsequently, when two dominant cofactor markers are used, the calculation will take up four times of the normal space, and so on.

QTL can be mapped using different population types but RILs are the most effective population types. However, the construction of RILs is time consuming as it requires at least six generations of self-pollination in order to obtain a level of confidence that loci will be homozygous (Seymour et $al.,$ 2012). The present QTL analysis in the F_5 segregating population is genotyped by GEMs that have been scored as dominant markers, in the absence of clear evidence for each locus to avoid active mis-calling of individual values in lines. An alternative approach whereby translating the marker genotypes from RI_5 to doubled haploid (DH) population type, which has no heterozygotes, is suggested when dominant markers are used in QTL analysis. This could be used for the GEMs map, but would involve a significant loss of data for the other maps. The advantage of RILs is that there are more recombination events than for DH populations.

The present QTL analysis using MQM mapping for traits of interest in Bambara groundnut was conducted individually in two generations of segregating populations, F_3 and F_5, derived from a cross between DipC and Tiga Nicuru. The possibility that QTL positions could be mapped more precisely if an

243

integrated map combined over populations, such as the combination of the DArTseq map and GEM map derived from F_3 and F_5 segregating population, respectively, is constructed, could be investigated. An integrated map would probably represent recombinant events over all populations, thus it can be used to identify QTLs that correspond between different populations for traits of interest (Ooijen, 2009). Although an initial construction of an integrated map was attempted, the result shows incomplete integration between the DArTseq map and GEM map (Chapter 6). Therefore, a detailed integration for these two maps would be needed in future prior to QTL analysis in order to detect and identify QTLs controlling traits of interest in Bambara groundnut with higher accuracy. The reasons for the partial integration of the map need further investigation.

7.4.2 Association between markers and traits in Bambara groundnut

Broad trait variation between the two parental lines of Bambara groundnut, DipC and Tiga Nicuru, allows different traits to segregate in the offspring. The use of QTL analysis in the segregating populations allows the identification of the loci controlling the traits of interest, potentially leading to molecular breeding and MAS for crop improvement.

Based on the MQM mapping result, QTL associated with *internode length* and *peduncle length* consistently mapped to LG1 across two generations of segregating population. The same single marker linked to a single locus suggested that these two traits are probably controlled by single gene or two closely linked genes. The hypothesis is supported by Basu *et al.* (2007) who reported that the segregation pattern of *internode length* was consistent with primarily monogenic inheritance in a domesticated (DipC) by *V. subterranea spontenea* (VSSP11) cross created to evaluate the domestication process in Bambara groundnut. The regulation of *internode length* by a single gene has also reported in pea (Reinecke *et al.,* 2013). *Internode length* in pea was

discovered to be controlled by single gene *Le* which encodes a gibberellin 3 β-hydoxylase that catalyses the conversion of GA_{20} to biologically active GA, an important regulator of plant growth and development (Lester et al., 1997; Reinecke et al., 2013). The transgenic pea plants with increased expression of GA_1 exhibited longer internode length, larger stipules, altered vascular development and displayed delayed flowering as compared to wild type (Reinecke *et al.,* 2013). The findings in pea suggested that morphological changes such as *internode length* and *peduncle length* in Bambara groundnut could be related to cell proliferation and expansion controlled by a single gene that is involved in gibberellin regulation. Therefore, future work involving the application of exogenous gibberellin to Bambara groundnut could confirm whether the observed morphological differences are gibberellin-sensitive or insensitive

Complex yield traits, such as *pod weight per plant, seed weight per plant, pod number per plant* and *100-seed weight* are more likely to have a larger environmental component in their phenotypic variation. In addition, the discovery of a number of QTLs (rather than a single major locus) explaining more limited phenotypic variation for yield traits, suggested that these traits could probably be controlled by many genes with minor effects and also affected by the environment. For instance, *pod weight per plant* and *seed weight per plant* were contributed to by multiple loci located across LG1, LG2B and LG11A in the F_5 segregating population. Similar observations were also reported by Zhang *et al.* (2004) who discovered four QTL located on three linkage groups (A2, B1 and D2) for seed weight in RILs derived from soybean vars. *Kefeng No.1* X *Nannong 1138-2*.

Although the QTLs identified on LG10 for *pod weight per plant, seed weight per plant* and *shoot dry weight* in the F_3 segregating population were not significant based on LOD scores that were lower than GW threshold, it is interesting to observe a similar distribution pattern of QTLs across the F_3 and F_5

245

segregating population.   In F_3 segregating population, QTLs controlling *pod weight per plant, seed weight per plant* and *shoot dry weight* were located particularly close to each other at 43.7 cM, 44.7 cM and 42.7 cM on LG10. A similar distribution pattern was observed in the F_5 segregating population in which the three studied traits *pod weight per plant, seed weight per plant* and *shoot dry weight* were also mapped at the same location at 45.89cM on LG11A. The observation indicates a close relationship among these three traits and a possibility that they are controlled by the same QTL. Pleiotropism was reported to be common in many QTL studies. For instance, a soybean locus was shown to affect five traits, including *days to flowering, plant height, lodging, nodes on the main stem* and *pods per node* (Zhang *et al.,* 2004). In order to further identify and detect QTLs related to *pod weight, seed weight* and *shoot dry weight*, the production of a good integrated map from the F_3 and F_5 segregating populations is important for a detailed comparison.

Most of the QTLs mapped in the cluster on LG1 in the F_5 segregating population are related to plant morphology, as well as yield traits. The clustered QTLs could correspond to single genes controlling plant architecture which has pleiotropic effects on different traits, including seed and plant growth-related traits. In pea, QTL detected for seed traits were found to be located in the genomic regions regulating traits such as plant morphology, phenology and plant biomass (Burstin *et al.,* 2007). The authors showed that *Le* allele which is related to internode length has pleiotropic effects on other traits such as plant height, vegetative biomass and plant nitrogen content. In wheat, the dwarfism gene (*Rht-1*) is associated with many QTLs including grain yield and root development QTLs (Laperche *et al.,* 2006). In rice, a single gene controlling erect leaf development is associated with higher grain yield (Sakamoto *et al.,* 2006). In Bambara groundnut, the present study showed that QTLs controlling *internode length*, *peduncle length, pod number per plant* and *seed number per plant* were linked with the same marker MM135 at 53.7 cM on LG1. On the basis

of pleiotropism, a speculation could be made that by altering the morphology of the plants, the genes may contribute to pod number.

For drought-related traits, the non-significant QTLs observed could probably be explained by the effect of the mild drought introduced to the F_5 segregating population, resulting in a relatively weak association between loci and traits. Although no significant QTL was found for *relative water content, stomatal conductance, carbon isotope discrimination analysis* and *stomatal density* using MQM mapping, a putative QTL of 3.04 (GW= 3.8) on LG2A was identified for *stomatal conductance* during the analysis using IM (result not shown). The result is correlated with phenotypic traits reported in Chapter 4, whereby the F_5 segregating population showed significant differences among the lines ($F_{(64,130)}$=16.27, $p<0.01$) for *stomatal conductance* using an ANOVA analysis. *Stomatal conductance* is probably controlled by multiple genes with minor effects, therefore could not be detected using MQM mapping which uses cofactors to eliminate the effects of additional QTLs (Jansen *et al.,* 1993).

The application of QTL analysis can be extended to the identification of candidate genes that control these respective traits. For instance, a QTL for *beginning of flowering* in pea was mapped onto linkage group LGV at 49 cM (Burstin *et al.,* 2007) which harbors the gene, *Det*, that is involved in the regulation of flowering time and of inflorescence architecture (Foucher *et al.,* 2003). In addition, the identification of a seed weight QTL on LGIII at 189 cM was close to the location of candidate gene *PepC* that encodes a phosphoenolpyruvate carboxylase (181 cM), and was also reported by Burstin *et al.* (2007). The identification of candidate genes could be done by aligning and comparing the map of QTLs with a genetic map with functional markers. For instance, a pea genetic map containing a total of 111 gene-anchored markers was developed by Aubert *et al.* (2006). This genetic map was used to identify the candidate genes in RILs derived from the cross between Terese and K586 in pea (Burstin *et al.,* 2007). However, like other underutilised crop species, the

genetic map with genes of known functions and/or physical map of Bambara groundnut is not yet available.

An example of using cross-species approaches to identify candidate genes is reported in cowpea. Through the syntenic relationship between cowpea with *Medicago trunculata* and soybean, the syntenic locus for *Hls* (hastate leaf shape) was discovered and led to the identification of a candidate gene controlling leaf morphology in cowpea (Pottorff *et al.,* 2012). The cross-species approach presented in cowpea provides an alternative option to identify candidate genes in underutilised crop species. Following QTL analysis, syntenic loci and candidate genes controlling traits of interest in Bambara groundnut could be identified by projecting the map of QTLs onto physical map or genetic map with functional markers derived from closely related species such as soybean and *Medicago trunculata*.

Chapter 8: PROVIDING A FRAMEWORK FOR IDENTIFICATION OF CANDIDATE GENES IN BAMBARA GROUNDNUT

8.1   INTRODUCTION

Legumes are generally categorised into three subfamilies: Papilionoideae, Mimosoidaeae and Caesalpinoideae, which accounted for approximately 70%, 15% and 15% of the legume species (Doyle and Luckow, 2003). These authors also reported a separation of Papilionoideae subfamily into four large divisions at approximately 50 Mya, which are galegoid (*Medicago truncatula, Lotus japonicus*, chickpea and pea), millettioid (soybean, cowpea and Bambara groundnut), dalbergioid (*Arachis*) and genistoid clades (*Lupinas*; Figure 8.1). The completion of genome sequences of three major legume crops from different clades, soybean, *Medicago* and *Lotus* has been reported, facilitating these assembled and annotated genomes to be compared and transferred from model plants to other crop species (Cannon *et al.,* 2009).
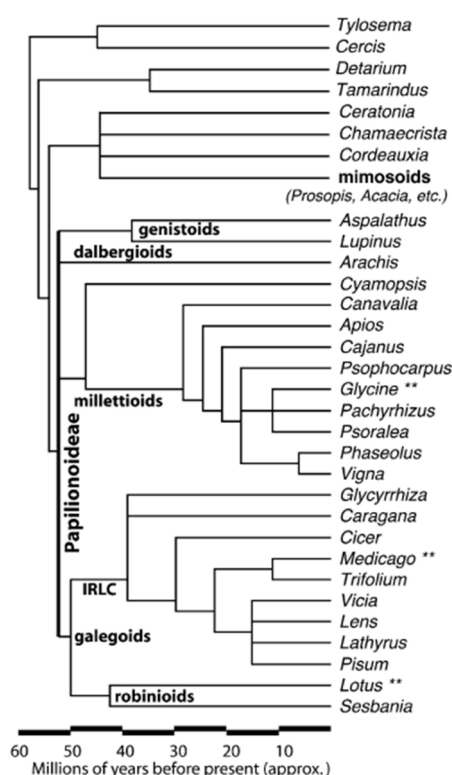
Figure 8.1 Taxonomic relationships among legume species (Cannon *et al.,* 2009).

The genome size of soybean ($2n = 2x = 40$) is estimated to be 1,115 Mb and the current assembled sequences are reported to contain 950 Mb across 20 chromosomes sequences and 23 Mb in smaller, unanchored scaffold sequences (Glyma1.01, http://www.phytozome.net). Compared to *Medicago* and *Lotus* which have been widely used for studies of mycorrhization, nodulation and plant-symbiont signalling (Oldroyd and Downie, 2008), soybean has mainly served as the model legume to study seed development (Vodkin *et al.,* 2008), mineral uptake, protein, oil biosynthesis and root hair development. As a result of having a narrow genetic distance across the Papilionoideae subfamily, most of the genes examined seem to be located within syntenic regions shared among the papilionoideae species (Cannon *et al.* 2009). The finding suggested that the position of an orthologous gene could probably be identified in one legume species if another closely related legume shows an association between a gene and phenotype. Cannon *et al.* (2009) reported the divergence of soybean and Bambara groundnut at approximately 20 Mya. The major genomic resources developed in soybean are thus believed to provide opportunities to study Bambara groundnut, an underutilised crop species which is tolerant to drought and serves as a source of useful dietary protein in many developing countries.

Several studies have reported the translation of genomics information from model plants to taxonomically related crop species. Pottorff *et al.* (2012) reported that an orthologous gene, EZA1/SWINGER, was found in the *Hls* region which controlled hastate leaf shape in cowpea using synteny with *M. truncatula* and soybean. Yang *et al.* (2008) reported the identification of the *RCT1* gene that is responsible for anthracnose resistance in alfafa (*M. sativa*) through the syntenic relationship with *M. truncatula*. In common bean, the genetic linkage map anchored with corresponding syntenic regions of the soybean was identified, allowing the specific genomic regions to be targeted for the discovery of genes and loci that affect phenotypic expression in both species (McClean *et al.,* 2010). Based on the studies given, it is believed that the location of

candidate genes controlling traits of interest in Bambara groundnut could be determined using the conserved synteny relationship with soybean, due to the relatively close taxonomic relationship with soybean (20 Mya).

Syntenic relationships between model species and crop species could be determined through BLAST search and positional alignment of sequences that show strong homology. For instance, the identification of the *Hls* region for hastate leaf shape in cowpea was conducted by subjecting an EST-derived SNP marker to a BLAST search and then aligning markers which are closely linked with the trait of interest, to other legume species such as soybean, *Medicago* and *Arabidopsis* (Pottorff *et al.*, 2012). The result showed that the *Hls* region was highly correlated with *Medicago* chromosome 7 and two soybean chromosomes, 3 and 19. From the three syntenic loci, an ortholog for EZA1/SWINGER was annotated as a candidate gene for the *Hls* region (Pottorff *et al.*, 2012). In a nodulation study in pea, a series of gene markers were mapped onto the pea genetic map and their homologues were BLAST searched against *M. truncatulata, L. japonicas*, soybean and poplar pseudomolecules (Bordat *et al.,* 2011). Based on the map position, a promising candidate gene in pea was identified to a homologue of Pub1, a gene which negatively regulates nodulation in *M. truncatulata.* As the homologues of Pub1 are located on the top of pea LG1 in the region of a hypernodulation mutant, *nod3* (Gualtieri *et al.,* 2002), the pea ortholog of Pub1 is predicted to be a candidate gene for Nod3.
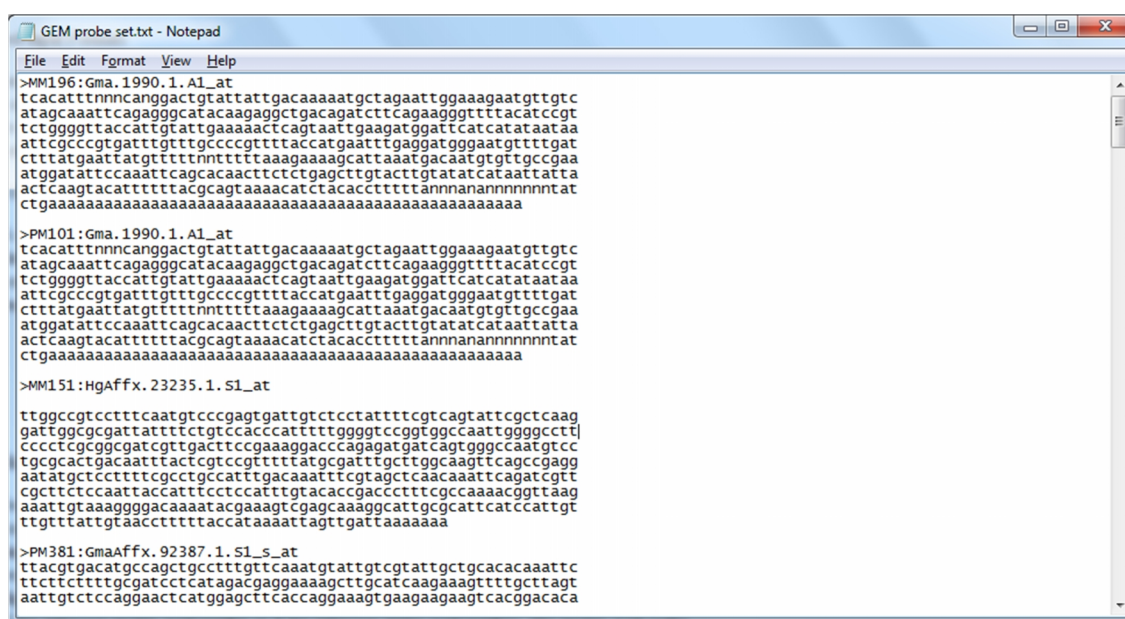
A preliminary evaluation of this approach for the creation of a conserved synteny framework for identification of candidate genes in Bambara groundnut was attempted, through evaluating LG1 as an example linkage group for alignment as most of the QTLs were clustered in LG1 and it is one of the longer groups. A series of markers derived from the DArTseq map and GEM map (Chapter 5 and 6) were subjected to a homology search, respectively, against soybean transcripts and assembled genome in order to localise the position of promising candidate gene.

8.2 MATERIALS AND METHODS

8.2.1 Preparation of FASTA files

A total of 78 markers (DArTseq map) and 28 markers (GEM map) on LG1 were subjected to a BLAST search, respectively. The sequences of DArTseq-based markers were derived from the tag sequence associated with each DArTseq marker generated by Diversity Arrays Technology Pty. Ltd (Yarralumla, Australia). In contrast, as GEMs were developed through cross-hybridisation of Bambara groundnut RNA samples onto the soybean Affymetrix GeneChip, the consensus sequences used to design the probe-set in soybean Affymetrix (http://www.affymetrix.com) were extracted for BLAST search.

Marker sequences arranged in FASTA format were required prior to BLAST search. The FASTA file was started with a single-line description of the sequence, followed by sequence data. The single-line description was distinguished from the sequence data by placing a symbol ">" in front of the description. In addition, a text of description was recommended shorter than 80 characters in length. An example of a FASTA file based on Affymetrix design sequences is shown below:

Figure 8.2 An example of a FASTA file based on Affymetrix design sequences.

8.2.2 BLAST search

A BLAST search was conducted using CLC Genomics Workbench v6.5.1 (http://www.clcbio.com) against a local BLAST databases constructed according to the instructions in CLC user manual. A total of three files, Bambara groundnut leaf transcripts, soybean transcripts (Gmax_189_transcript; Schmutz *et al.,* 2010) and soybean assembled genome (Gmax_189; Schmutz *et al.,* 2010) were imported into the CLC Genomics Workbench interface for the creation of local BLAST databases using the option 'Create BLAST Database'. Subsequently, FASTA files containing marker sequences were also imported and two types of BLAST searches were conducted (Figure 8.3).

First, the marker tag sequences (FASTA format) derived from the DArTseq map were searched using the BLAST program called 'blastn: DNA sequence and database' against the Bambara groundnut leaf transcripts under default settings. When a single good hit was collected, the gene model identified was searched against the soybean transcripts, to identify the most complete soybean homologues. This soybean homologue was then BLAST searched using BLAST program 'tblastx' against the soybean assembled genome using default settings, to identify the location of the transcript. In order to maximise the sensitivity when comparing coding sequences between two species, translated searches are preferred as they convert nucleotide sequences to a more conserved protein translation before the comparisons are made (NCBI news, 2002).

Second, the markers sequences (FASTA format; derived from the original design sequences for the Affymetrix soybean GeneChip) from the GEM map were directly searched against the soybean transcripts. The consensus sequence of soybean transcript was also extracted and subjected to a similar 'tblastx' search against the soybean assembled genome. The best hit was then selected based on E-value and %identity. The E-value served as a measure of the quality of the match with a smaller E-value indicates greater homology. The %identity showed

the percentage of identical residues between query sequences and hit sequences from a database, with longer stretches of homology more likely to indicate a genuine match.

Markers from
GEM map

↓

64bp marker sequence   Identify design sequence
tag from DArTseq map   from Gene Chip

↓ BLAST   ↓ BLAST
                                        tblastx
Bambara groundnut ──────→   Soybean   ──────→   Soybean
leaf transcripts   transcripts   assembled genome
                   Consensus   Consensus
                   sequence   sequence
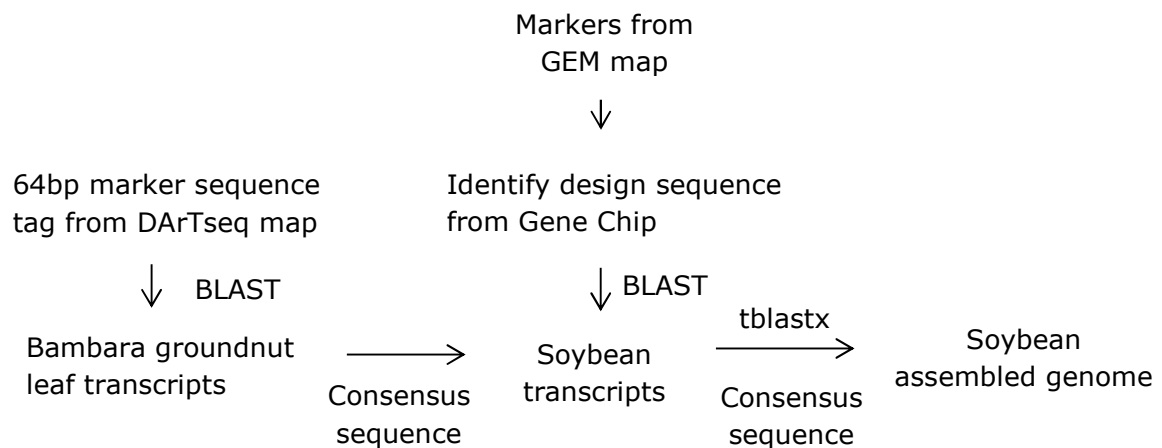
Figure 8.3 A flow chart of BLAST searches conducted in CLC Genomics Workbench v6.5.1 using markers derived from the DArTseq map and GEM maps, respectively, against three local BLAST databases: Bambara groundnut leaf transcripts, soybean transcripts (Gmax_189_transcript; Schmutz *et al.,* 2010) and soybean assembled genome (Gmax_189; Schmutz *et al.,* 2010).

8.3   RESULTS

A preliminary test of the construction of a framework for the identification of candidate genes for Bambara groundnut was conducted using LG1 derived from the DArTseq and GEM maps (Chapter 6). Sequences of each marker on LG1 were subjected to a BLAST search against three local BLAST databases: Bambara groundnut leaf transcripts, soybean transcripts (Gmax_189_transcript; Schmutz *et al.,* 2010) and soybean assembled genome (Gmax_189; Schmutz *et al.,* 2010) in an effort to compare the homology and identify the location of the gene of interest from Bambara groundnut within the soybean genome. The result of BLAST searches for markers derived from the DArTseq map and GEM map are presented in Figure 8.4 and Figure 8.5, respectively.

Of 78 markers (dominant DArT and SNPs markers) derived from DArTseq map, 12 markers (15%) with unique best hits were identified on the soybean assembled genome. The twelve markers showed locations across five chromosomes in soybean, which were Gm2 (2 SNPs), Gm5 (2 SNPs), Gm12 (2 SNPs), Gm13 (4 SNPs and 1 dominant DArT) and Gm15 (1 SNP). From a Bambara groundnut genetic perspective, the longest syntenic region was 50.6 cM in length, which corresponded to Gm5 with a physical position between 2.2 Mb and 7.7 Mb in the soybean chromosome. In contrast, 19 GEMs with a single best hit, out of 28 GEMs (68%) derived from GEM map, were mapped onto the soybean assembled genome. The identified regions between the two species were located across eight soybean chromosomes, including Gm2, Gm3, Gm5, Gm11, Gm12, Gm13, Gm14 and Gm17. The longest syntenic region was identified between marker MM196 and PM58 (83.9 cM) which showed coherence with Gm17 between 1.1 Mb and 11.6 Mb.
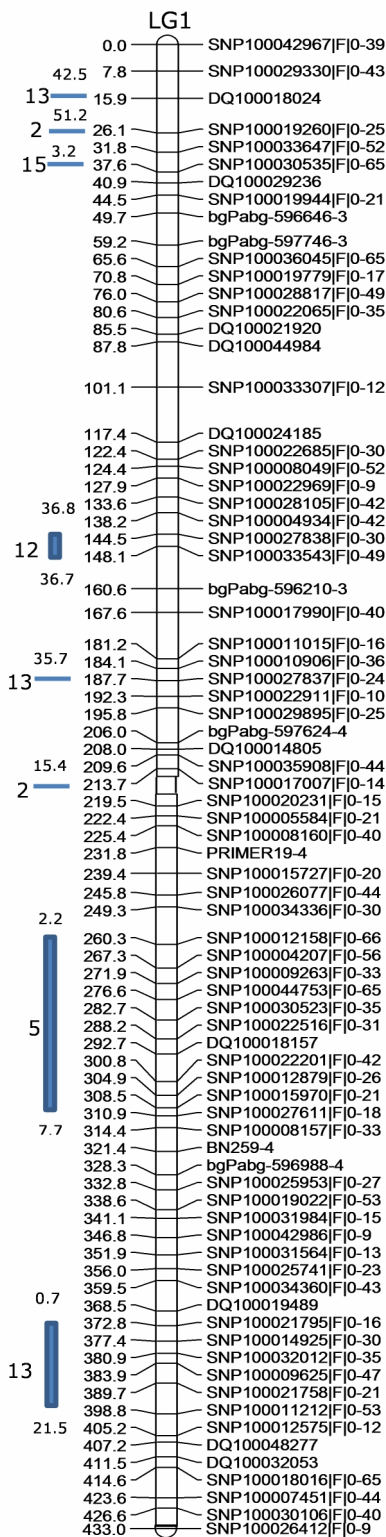
Figure 8.4 Syntenic relationship of LG1 derived from Bambara groundnut full density DArTseq map with soybean. Corresponding syntenic regions of soybean (Gmax 189; Schmutz *et al.,* 2010) were anchored in the Bambara groundnut DArTseq map. The soybean fragments are highlighted in blue with their respective chromosome number and their locations (in megabase pairs).
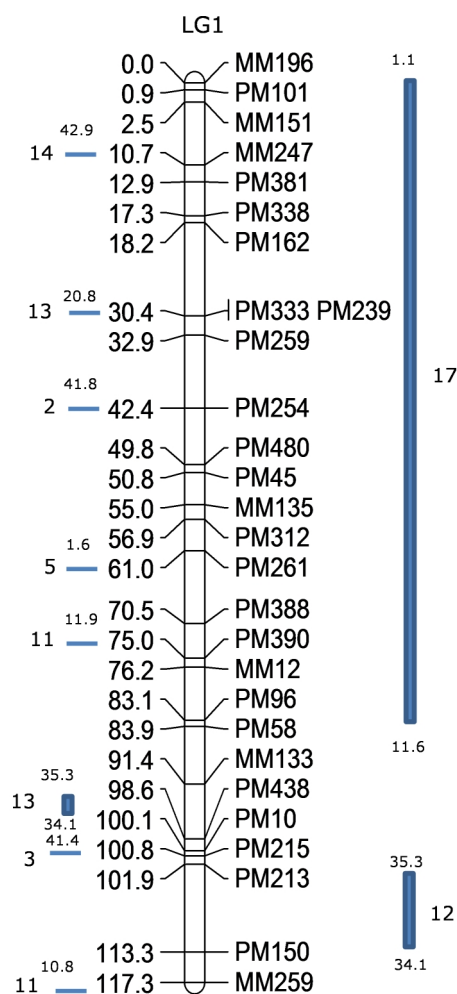
Figure 8.5 Syntenic relationship of LG1 derived from the Bambara groundnut GEM map with soybean. The corresponding syntenic regions of soybean (Gmax 189; Schmutz *et al.,* 2010) were anchored in Bambara groundnut GEM map. The soybean fragments are highlighted in blue with their respective chromosome number and their locations (in megabase pairs).

In addition, the identification of syntenic regions in both Bambara groundnut genetic maps relative to soybean were compared with the syntenic relationship between a common bean genetic map and soybean (McClean *et al.,* 2010). Figure 8.6 suggests that LG1 of Bambara groundnut corresponds to Pv3 in the common bean through the comparison with soybean physical locations (Mb). Bambara groundnut and common bean shared several syntenic regions relative to soybean, especially Gm2, Gm5 and Gm17. For example, the region between g417 (86 cM) and g665 (150 cM) from common bean Pv3 was mapped with Gm17 at physical locations between 2.9 Mb and 18.7 Mb. A similar syntenic region was also observed in Bambara groundnut at a genetic distance between 0.0 cM and 83.9 cM. Furthermore, based on the syntenic relationship of common bean Pv3 relative to soybean, it was observed that Gm17 in soybean corresponded with Gm5 and Gm2. This observation is in agreement with Bambara groundnut as seen in LG1 from the GEM map. Syntenic regions in Bambara groundnut relative to Gm17 between 1.1 Mb and 11.6 Mb (0.0 cM-83.9 cM) corresponded to Gm2 at physical location of 41.8 Mb (42.4 cM) and Gm5 at 1.6 Mb (61.0 cM). A high level of co-linearity between markers in both Bambara groundnut genetic maps and soybean might allow the determination of the syntenic relationships between Bambara groundnut linkage groups and soybean linkage groups, providing a framework for overlaying the QTL detected in Bambara groundnut onto the soybean genome.
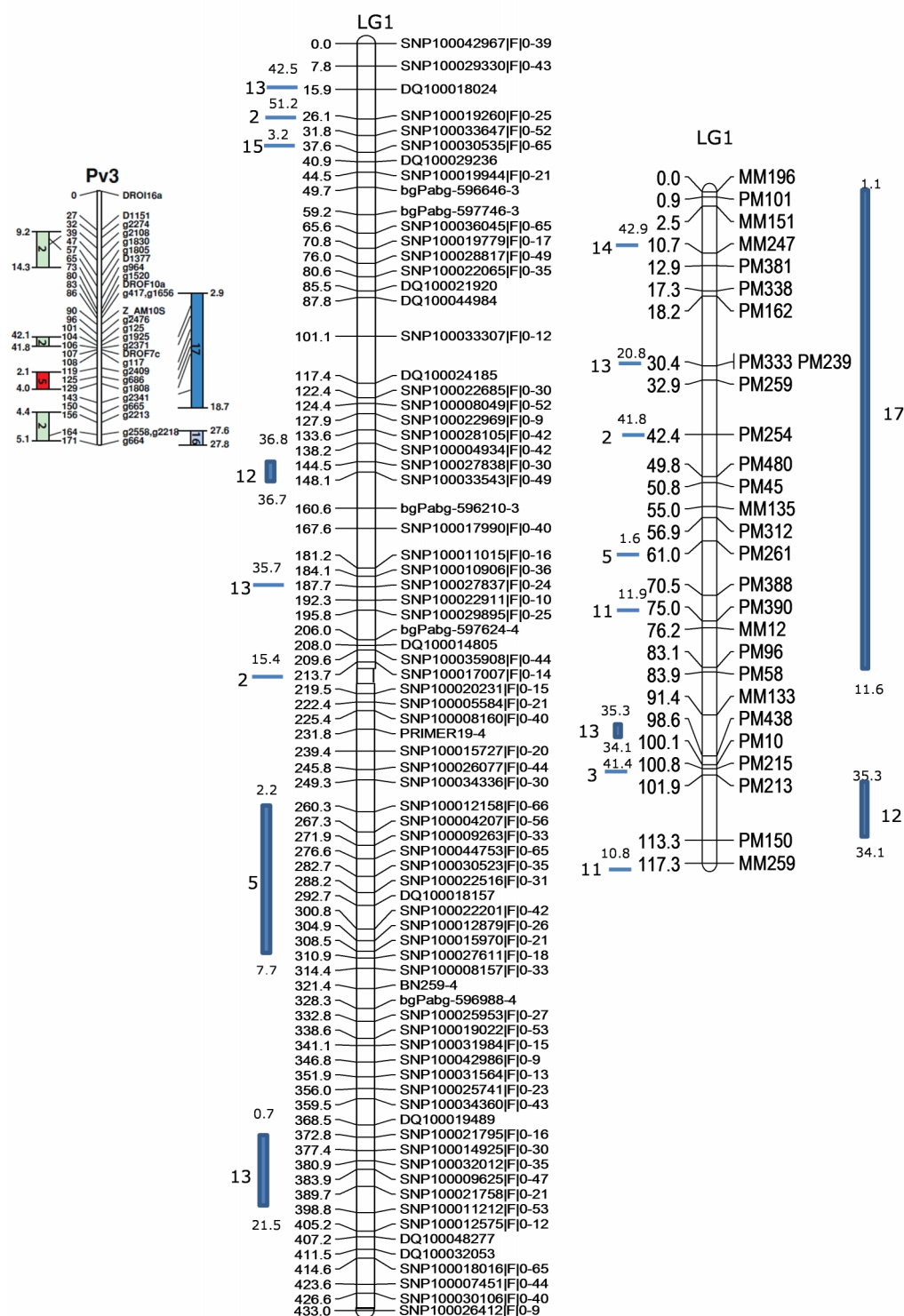
Figure 8.6 Comparison of syntenic regions of LG1 in both Bambara groundnut genetic maps relative to soybean with syntenic relationships between a common bean genetic map and soybean (McClean *et al.,* 2010). The soybean fragments are highlighted in blue with their respective chromosome number and their locations (in megabase pairs).

8.4   DISCUSSION

The development of a genetic framework for the discovery of candidate genes for Bambara groundnut using the conserved syntenic relationship with soybean was tested. The short marker tag sequences (64 bp) associated with dominant DArT and SNP markers are unlikely to align directly to the soybean genome sequence at high frequency, given 20 Mya of evolutionary divergence. However, an initial alignment of the mapped marker 64 bp tag with the Bambara groundnut leaf transcriptome produced through 454 sequencing technology (unpublished) in order to obtain longer gene models prior to BLAST search would increase the rate at which orthologues in the soybean genome could be detected by cross-species BLAST searches. The longer gene model sequences are anticipated to give a more accurate match with homologous sequences from the database and provide a clearer association between Bambara groundnut and soybean.

Due to limited number of Bambara groundnut gene models identified by the 64 bp marker tags, LG1 from original map grouped using maximum likelihood approach instead of a DArTseq framework map was used to increase the number of good gene model matches. However, the disadvantage of using this full density map is some potential inaccuracy of marker location and genetic distance between adjacent markers. The inflation of the total genetic distance from 149.4 cM to 433.0 cM in LG1 suggests some conflict between marker genotypes. An initial test showed that 15% of 64 bp marker tags aligned uniquely with the Bambara groundnut transcriptome, allowing a good candidate to be extracted and blasted against the assembled soybean genome. Given that the DArT Seq analysis returned large numbers of potential mapping markers (dominant DArT: 3,670; SNPs: 2,993), a framework map derived from the best quality data could be supplemented with those markers which show a clear blast match between the marker tag and the Bambara groundnut transcriptome. Once the genetic linkage map associated with supplementary markers is

developed, they could be used to identify orthologues in soybean, allowing more effective comparison of gene order in the two species.

Based on marker and cytogenetic information and targeted sequencing, the soybean genome is believed to have undergone polyploidy at approximately 13 Mya (Shoemaker *et al.,* 2006). This means that any given legume genome, such as *Medicago* and *Lotus*, could correspond to two soybean regions (Cannon *et al.,* 2009). The finding is in agreement with the observation in the present study, whereby Gm17 in soybean appears to correspond with Gm5 and Gm2 as seen in LG1 derived from Bambara groundnut (Figure 8.6). The markers obtained two good hits corresponded to two soybean chromosomes, respectively, after undergoing BLAST search against soybean assembled genome (result not shown). For example, SNPs markers SNP100012158|F|0—66 showed a clear blast match with both Gm5 (E-value: 0.0; %identity: 94.8%) and Gm17 (E-value: 0.0; %identity: 93.2%) at physical locations of 7.7 Mb and 10.0 Mb, respectively. GEMs from the GEM map in Bambara groundnut, such as PM58, also showed a corresponding location with the soybean assembled genome Gm5 at 3.6 Mb and Gm17 at 11.6 Mb. The results were supported by McClean *et al.* (2010) who reported the sharing of duplication blocks between Gm5 and three other soybean chromosomes, including Gm8, Gm17 and Gm19, based on reference ordering of common bean sequences. The present study is a preliminary test using LG1 in Bambara groundnut, the full set of relationships between Bambara groundnut and soybean can be reviewed when all linkage groups are subjected to a syntenic search.

The matching of markers in LG1 from the DArTseq map on the soybean chromosomes, such as Gm2, Gm5, Gm12, Gm13 and Gm15, indicated the possibility of having rearrangement and reshuffling of genomic regions in Bambara groundnut compared to soybean (Bordat *et al.,* 2011). The reshuffling of synteny blocks across pea, *M. truncatula, L. japonica* and soybean in legume families was also observed, with pea and *M. truncatula* have the most conserved

synteny blocks due to a reasonably close taxonomic distance, which is approximately 24 Mya (Cannon *et al.,* 2009). In contrast, for GEMs, which represents the expression patterns of genes in Bambara groundnut, the identification of synteny blocks using GEMs from the GEM map on other chromosomes in soybean (i.e. Gm3, Gm11 and Gm14) could at least partly be due the mapping of *trans* effects. However, there are also likely to be markers potentially mapping with *cis* effects, such as markers which have coherent positions with Gm17, as conserved synteny regions in Gm17 were also observed in common bean genetic map relative to soybean. A detailed integrated map comprising of dominant DArT, SNPs and GEMs markers is important to have an in depth comparison for the syntenic regions in Bambara groundnut relative to soybean as well as the identification of *cis* and *trans* effects in Bambara groundnut.

There is a concern for the 64 bp dominant DArTs and SNPs, which are aligned uniquely with the Bambara groundnut leaf transcriptome, not detecting Gm17 in soybean as it is the major region of synteny between Bambara groundnut and soybean. The finding is suspected to be resulted from the criteria of choosing single best hit of markers to the soybean assembled genome. The blast match of the Bambara groundnut gene models identified by the 64bp marker tags with Gm17 of soybean was observed when the marker derived gene model hit on a fragment of soybean with lower E-value. For example, SNP100011212|F|0—53 obtained two hits with Gm13 and Gm17, whereby the second hit on Gm17 showed a matching of 39 amino acids out of 47 (E-value: 1.2E-88), compared to the first hit (best hit) on Gm13, which obtained 116 out of 122 match amino acids (E-value: 0.0). This may or may not represent a genuine match and needs further investigation. The present study aimed to use the data of the highest quality to construct a framework for identification of candidate genes between Bambara groundnut and soybean. The additional markers with slightly lower E-value hits and %identity that hit other

chromosomes in soybean could be included in the future work for a more detailed comparison between the two species.

While the identification of gene locations in Bambara groundnut are incomplete as only LG1 was subjected to the test in the present study, the determination of syntenic regions in LG1 corresponded to common bean Pv3 relative to soybean suggested that the principle underlying this test is valid. In future work, more linkage groups from genetic linkage maps in Bambara groundnut should be included in the BLAST search to provide a complete framework for assisting the discovery of candidate genes for Bambara groundnut using the conserved syntenic relationship with soybean.

Genome resources in model and major crop species are important to improve crop species that have limited genetic and genomic tools. Molecular markers can be developed from references genomes and then applied in closely related species (Cannon *et al.,* 2009). The present study shows that the cross-hybridisation of Bambara groundnut RNA samples onto the soybean GeneChip and the development of an 'expression-based' genetic map (GEM map) can allow the identification of positions in the major crop species which are likely to correspond to the location of QTL in the minor crop species. This potentially allows the identification of a cross-species candidate gene list which corresponds to the candidate gene underlying these region of QTL in the species of interest. A structured bioinformatics pipeline will be necessary in order to translate biological information from model species to crop species. For example, Bordat *et al.* (2011) reported the use of the 'Pea Medicago translational tool kit' that is hosted on an Apache web-server to search for the putative position of a gene on pea map as well as putative candidate genes in closely related species such as *Medicago*.

Chapter 9: GENERAL DISCUSSION

9.1   ISSUES AND CHALLENGES

Of the 270,000 species of higher plants, about 7,000 species are used for food, fibre, medicine and other purposes, but only 15 crops (including three major crop species, which are rice, maize and wheat) contribute 90% of global food production (Cromwell *et al.,* 1997; Padulosi, 1999). Although global food production relies on a few crop species, there are actually other crop species that provide major sources of energy intake at the regional level. For example, cassava, beans, groundnuts, cowpeas and yams in Central Africa are reported to serve as the dietary staples of millions of people (Cromwell *et al.,* 1997).

The loss of crop diversity is often related to the intensification of agriculture and the growing of cash (commodity) crops (FAO, 2011). The United Nations Food and Agriculture Organisation (FAO) estimate that about 75% of the original varieties of crop species have been lost since 1900 and the trend has increased in the last 50 years (FAO, 1993). Since the launch of the Green Revolution in the 1960s, farmers have adopted a preference for cultivation of single, high-yielding varieties in place of traditional landraces. This agricultural practice which is highly dependent on a few crop varieties has narrowed the genetic base, causing cultivation to be at risk from pest and diseases. One of the examples is the Irish Potato Famine in the 1840s which resulted in the deaths of millions of people from starvation and disease (FAO, 1993). The Irish working population relied primarily on potato with a very narrow genetic base which proved susceptible to late blight disease. Genetic uniformity renders potatoes vulnerable to attack when virulent pathogen strains spread rapidly throughout the population (FAO, 1993).

With the estimated increase in world population from 6.6 billion to 9 billion by 2050 (FAO, 2009), an increase in crop production by around 70% is predicted to be necessary to fulfil the increased global demand for food. Agricultural biodiversity is a potential tool for improving food and nutritional

security (Hunter and Fanzo, 2013). There is considerable interest generated about the potential use of the under-exploited species to provide basic resources for crop improvement in order to adapt to variable environments, climates and to overcome issues of pests and diseases. The lack of agro-biodiversity is a crucial issue especially in regions where diet depends solely on starchy staples with limited access to high micronutrient containing foods. FAO (2012) reported that there are now approximately 868 million people suffering from hunger and malnutrition. In addition, 35% of all children are at risk of malnutrition (Black *et al.,* 2008) and over 2 billion people experience micronutrient deficiencies (Micronutrient Initiative, 2009).

However, the barriers to the exploitation of new crop species are often related to factors such as competition with commodity crops, cultivation practices, inefficiencies in processing and value addition, insufficient market demand and the politics of agriculture. In addition, financial support received from national governments, international and local breeding companies for research and breeding in new crop species over decades is often limited.

9.2   A POTENTIAL APPROCH FOR FOOD SECURITY

The importance of exploiting existing plant biodiversity and developing underutilised crop species for use in future agriculture (particularly those with advantageous traits) in order to tackle these global challenges is clear. The application of conventional and molecular breeding using biotechnology is important to select plants which may have a role in agriculture of the future, particularly in response to drought, disease and pests, waterlogging and eroded soils. A structured breeding program is required for the development and improvement of underutilised crop species, but also an understanding of where knowledge is missing across the whole of the research value chain.

Genomics is a study of an organism's entire genome. The development of genomics tools enables the genes to be identified, providing a foundation to

understand gene expression and biological responses. Given that genetic control of agronomic traits in major and model species, such as *Arabidopsis* and rice are well studied, major resources or knowledge developed in biological models can be transferred to underutilised crop species using genomics tools. With the increased knowledge and advanced development of new technologies, a fundamental understanding of plant genome organisation and regulatory network responses to stress and the molecular mechanisms underlying crop traits can be developed prior to using this knowledge for molecular breeding and production of new varieties with desired traits.

The present study aimed to develop approaches to study and evaluate genomes and transcriptomes of crop species by utilising data and resources derived from major crops and model plants. Although further studies and validation of preliminary results are required, the proposed approaches, including the XSpecies approach coupled with genetical genomics – either through microarrays or through next generation sequencing once the prices drop further - seems to be promising and potentially effective for use in research on underutilised crop species. If extensive genomics studies and breeding for a crop species with exceptional traits can be developed, some of the current issues could be resolved, such as over-reliance on staple food crops and the development of equivalent species without the long development cycles which major crops have undergone.

9.3   THE XSPECIES APPROACH IN CROP SPECIES

The close association between studied crop species and major and/or model plant species is of importance for the XSpecies approach. The cross hybridisation of oil palm onto the dicotyledonous plant, *Arabidopsis*, and the monocotyledonous Poaceae family member, rice, is first experiment reported in this study. The evolutionary distance between oil palm and *Arabidopsis* (145-208 Mya; Sanderson *et al.,* 2004) compared to oil palm and rice (91 to 99 Mya;

266

Wikstrom *et al.,* 2001) provides an insight into the effects of sequence divergence between the target species and the reference species onto the power of detecting SFPs in oil palm. Although oil palm is taxonomically closer to rice than *Arabidopsis*, rice is still not an ideal reference species to compare with oil palm in XSpecies approach and is informationally poorer compared to *Arabidopsis*. Inefficient hybridisation of certain transcripts to the probes as a result of sequence divergence would lead to the production of background noise which could be an obstacle in data analysis or even lead to the complete loss of signal. For the Affymetrix technology, where cross-hybridisation is dependent upon a set of 11 oligonucleotides which constitute a probe-set and each probe being only 25 nucleotides in length, this could be a particular problem. However, even for other microarray technologies, such as Agilent where the detecting probe is a 60-mer, evolutionary distance is still expected to be a confounding factor.

The application of the XSpecies approach was extended to legume family, with Bambara groundnut compared to soybean. This is the first XSpecies study reported in Bambara groundnut. Despite the sequences available for soybean not being as comprehensive or extensive annotated as *Arabidopsis* and rice, the phylogenetic distance between the soybean and Bambara groundnut is smaller (20 Mya; Cannon *et al.,* 2009). Although *Medicago* and *Lotus* are also well studied legume plants with assembled and annotated genomes, their phylogenetic distance from Bambara groundnut is reported to be 54 Mya (Cannon *et al.,* 2009). A complication in the use of soybean for work in Bambara groundnut is the duplication of the soybean genome since evolutionary divergence of the two species ($2n = 2x = 22$ for Bambara groundnut compared with $2n = 2x = 40$ for soybean). From the two exemplar crops used in the present study, the phylogenetic relationship between model plants and crop species provides an insight into the cut-off point for use in the XSpecies approach to translate information from model plants to other crop species.

Comparing the genetic distance of oil palm with *Arabidopsis* and rice, the XSpecies approach applied in Bambara groundnut with soybean was expected to be more effective and sensitive.

The principle of the XSpecies approach is worth exploring to develop genomic resources in non-model crop species based on publications (e.g. Graham *et al.,* 2007; Broadley *et al.,* 2008; Davey *et al.,* 2009). The application of the XSpecies approach in oil palm was conducted in 2011, when the oil palm genome sequence was still not available yet. Given that oil palm sequences have been released in 2013 (Singh *et al.,* 2013), it would be easier to map the potential genes and determine the functions that differ between two fruit types based on the hybridisation signal differences revealed from PIGEONS, although no publically available high density microarray currently exists for oil palm. Future work focusing on developing and using bioinformatics tools to exploit oil palm genomic and transcriptomic information from an XSpecies microarray analysis would be needed and within species approaches based on Next Generation Sequencing may soon be cost-effective. An *in silico* analysis of why the approach failed to identify markers to shell-thickness (based on the released oil palm genome sequence) would most likely provide an insight into the principles underlying the XSpecies approach in oil palm. The choice of date palm, which is also a member of the *Aracaceae* family, would be a better candidate for oil palm in an XSpecies approach, although resources available in this species are also relatively limited. However, the XSpecies approach which uses major resources from model plants to identify gene sequences of target crop species could become less useful, especially when the reported sequences of oil palm genome have become more comprehensive and fully publicly available.

By cross-hybridising target species onto microarrays derived from major plants, the XSpecies approach serves as an alternative pipeline to develop genomics sequences in crop species. While the cost of sequencing is declining,

the most challenging part of sequencing is the analysis of sequencing data which involves complex assembly and annotation work that may not be straight forward, especially given the absence of a viable reference species. In some cases, there may be also limited public access to sequence resources which renders the XSpecies approach as a valid alternative. There are often large existing data and plant resources available in the model species which could allow a first evaluation of the effects of the candidates in model species. The XSpecies approach offers advantages by allowing pre-existing resources to be used in identifying candidate genes for traits in the related crop species where genes of interest can be identified through the use of model species microarrays.

9.4   APPLICATION OF THE XSPECIES APPROACH COMBINED WITH THE GENETICAL GENOMIC APPROACH

The preliminary results obtained from the XSpecies study in oil palm also suggest the importance of having a genetic linkage map to localise genes that control traits of interest in addition to gene expression profiling generated from the microarray analysis itself. The combination of the XSpecies approach with a genetical genomics approach provides an insight into the evaluation of crop species at both genetics and transcriptomics level.

DArT Seq was first applied in Bambara groundnut for the generation of dominant DArT and co-dominant SNPs markers prior to the construction of a genetic linkage map. Combined with pre-existing microarray-based DArT and SSR markers, the first high density map and also the first framework map using DArTseq for Bambara groundnut was produced. DArTseq map could serve as the backbone for QTL analysis and also the integration of other marker-types in the future.

In this study, gene expression markers (GEMs) were produced at the (unmasked) probe-pairs (oligonucleotide) level after cross-hybridising leaf RNA from a segregating Bambara groundnut cross under a mild drought treatment

with the soybean GeneChip. This is the first development of GEMs in Bambara groundnut and they are expected to represent differences in hybridisation signal of RNA to individual oligonucleotide probes. A first spaced GEM map was then developed and this is also the first 'expression based' map in Bambara groundnut. The construction of two genetic linkage maps in this study provides an initial look into the use of markers that are anchored by a short 64 bp tag sequence and markers that show hybridisation signal difference derived from RNA samples for which the original soybean design sequences are available. Following the construction of a framework map, the initial integration of the DNA and RNA marker maps was conducted. This is the first attempt to develop a consensus map for Bambara groundnut. However, while composite chromosomes could be reliably identified, the final integration of markers was uncertain with a clustering of marker types. Although this could be a genuine effect, it needs further exploration and a possible change in approach. In future, a detailed integrated map would probably offer greater potential to map QTLs with traits of interest more accurately.

There are some concerns when using segregating populations with different generations (even of the same cross) of Bambara groundnut and also with a wide range of marker systems. Two segregating populations may possess different genetic effects and interactions in different environments. Moreover, the relative balance between dominance and additive effects will change as the population undergoes further inbreeding. For the F_3 population, the expectation of the proportion of heterozygotes is 25%, while for F_5 population it is 6%, so the relative effect of any dominance in the QTL will decrease between the two populations.

The first drought treatment in a Bambara groundnut controlled cross (F_5 segregating population) derived from a cross between DipC and Tiga Nicuru was conducted to explore the mechanisms underlying any segregation of drought response in this population to drought, prior to selection and breeding of high

270

yielding lines under drought stress. The phenotyping of an F_5 segregating RIL population could provide fundamental information to determine the location of QTLs as well as eQTLs (the genetic regions which are associated with variation in gene expression). In accordance with West *et al.* (2006), GEMs were scored as dominant markers for use in the genetic linkage map. The conversion of GEMs, which reflect the variation underlying hybridisation signals (regardless of the cause), into dominant markers gives novel markers for Bambara groundnut and enhances the availability of markers to conduct comprehensive QTL and eQTL analysis. The first comprehensive QTL analysis with good genome coverage on the GEM map was conducted and it showed the usefulness of the GEM map in potentially mapping QTLs in the F_5 segregating population. The present study showed that no significant QTLs were mapped for drought-related traits. However, the identification of QTLs controlling plant morphology and yield traits under drought, which are also a concern for farmers, in the segregating population, gives a first piece of information for a number of fundamental questions about genetic control of quantitative traits.

Due to time constraints and limited bioinformatics support during the study, the more advanced analysis for eQTL could not be performed. GEM markers rely on hybridisation signal differences at the oligonucleotide level and correspond to variation in gene expression levels, hybridisation strength or both, that tends to be quantitative in distribution. By converting microarray hybridisation signals into quantitative data and treating this as a trait in itself, the application of eQTL analysis using the GEM map developed here will be done in a future study. In addition, the development of the resources for an eQTL analysis in the present study will also provide a new channel for future work involving the identification of eQTLs related to morphological features and even drought-related traits in Bambara groundnut.

The Affymetix GeneChip$^{®}$Soybean Genome Array (2006) was designed based on 37,500 soybean transcript (www.affymetrix.com). In addition, the

genome of soybean has also been assembled and annotated (Cannon *et al.,* 2009). On the basis of the relatively close relationship between Bambara groundnut and soybean, a first attempt to test overlaying the genetic linkage maps developed in Bambara groundnut with the 'pseudo physical' map in soybean was made. Based on the current transcriptome, 15% of dominant DArT and SNPs markers are demonstrated to hit the current leaf Bambara groundnut transcriptome uniquely. Therefore addition of other tissue and stage transcriptomes or the sequencing of the gene space of Bambara groundnut should improve this figure. While the proposed approach is able to reveal genetic information in Bambara groundnut, it is considered as a preliminary attempt and probably not practical when a large number of markers are studied without bioinformatics support. By focusing on generating genetic linkage maps, perhaps an integrated map, using dominant DArT, SNPs and GEMs with linkages between orthologous genes would be a sensible way to allow comparison with soybean. In addition, the potential positions of eQTL detected in the minor crop can also be compared to the locations in the major crop, allowing the translation of information and possible identification of candidate genes.

9.5   IMPLICATIONS OF THE STUDY AND FUTURE RESEARCH OPPORTUNITIES

Oil palm is a high oil-yielding crop species used for global vegetable oil production and initial analysis on this species allowed the XSpecies approach to be refined in this study. Bambara groundnut is the third most important legume after groundnut (*Arachis hypogaea*) and cowpea (*Vigna unguiculata*) in semi-arid Africa (Howell, 1994). This underutilised crop is a potential crop for the future due to its good nutritional content and its drought tolerance. The development of varieties of Bambara groundnut with traits of interest is essential for different environments, especially in water-scarce areas. Understanding the basis of plant architecture, morphology, physiology and its interactions with the environment

offers breeders the potential to develop new material and appropriate agronomic practice for the future.

The current study, which used oil palm and Bambara groundnut as exemplar crop species, aimed to develop new approaches and understanding for transcriptomics and genomics by using major resources developed from model and major crop species for studies in less researched crop species. The results obtained in the present study would provide a platform for use in the experimental analysis of landraces and breeding for varieties with desired traits, especially for Bambara groundnut. In addition, the research can be expanded to the use of segregating populations derived from other landraces in order to examine the flexibility and effectiveness of this combined approach. For example, crosses between different landraces in Bambara groundnut could produce potential hybrids with enhanced characters, such as decreased photoperiod requirement for pod filling and enhanced protein content in seeds. The existence of the high density genetic maps also provides a reference for further study. The development and application of the DArT Seq technology provides a tool which will allow comparison of results from genetic analysis in future crosses to the current work. When genetic linkage maps across different segregating populations are integrated, the genetic location of traits observed in multiple populations can be analysed. In any follow up research, a bioinformatics pipeline is required in order to determine potential candidate genes in crop species using resources developed from major and/or model plants.

The identification of gene location in Bambara groundnut which corresponds to positions in the soybean genome would allow a better understanding of legume evolution and domestication. In recent years, with the establishment of complete genome sequences in legumes, such as *Medicago*, *Lotus* and soybean, the genomic architecture of domestication has been better understood. Given the advanced studies done on model plants, the information on what genes and/or traits are commonly selected during domestication can be

translated for use in research of a few potential underutilised crop species. This will lead to production of new varieties with desired traits in a much shorter time frame as compared to major crops.

The application of the XSpecies approach may not necessary provide a better alternative to next generation sequencing as both methods are applicable perhaps in different situations. Each strategy possesses advantages and disadvantages, but the present study provides additional information and shows that the combined approach is a sensible and valid alternative that could allow molecular mechanisms underlying traits of interest to be studied at DNA and RNA level simultaneously. Translation from model plants and major crop species to underutilised crop species is critical to develop various underutilised crop species with potential for future agriculture. This study is a small contribution to the exploitation of agricultural biodiversity which is potentially important to address food security challenges.

References

454 Life Sciences. 2010. *Sequence Capture/ Targeted Resequencing.* http://www.454.com/applications/sequence-capture-targeted-region.asp (accessed 27/05/2010)

Abiola, O., Angel, J.M., Avner, P., Bachmanov, A.A., Belknap, J.K., Bennett, B., Blankenhorn, E.P., Blizard, D.A., Bolivar, V., Brockmann, G.A., Buck, K.J., Bureau, J.F., Casley, W.L., Chesler, E.J., Cheverud, J.M., Churchill, G.A., Cook, M., Crabbe, J.C., Crusio, W.E., Darvasi, A., de Haan, G., Dermant, P., Doerge, R.W., Elliot, R.W., Farber, C.R., Flaherty, L., Flint, J., Gershenfeld, H., Gibson, J.P., Gu, J., Gu, W., Himmelbauer, H., Hitzemann, R., Hsu, H.C., Hunter, K., Iraqi, F.F., Jansen, R.C., Johnson, T.E., Jones, B.C., Kempermann, G., Lammert, F., Lu, L., Manly, K.F., Matthews, D.B., Medrano, J.F., Mehrabian, M., Mittlemann, G., Mock, B.A., Mogil, J.S., Montagutelli, X., Morahan, G., Mountz, J.D., Nagase, H., Nowakowski, R.S., O'Hara, B.F., Osadchuk, A.V., Paigen, B., Palmer, A.A., Peirce, J.L., Pomp, D., Rosemann, M., Rosen, G.D., Schalkwyk, L.C., Seltzer, Z., Settle, S., Shimomura, K., Shou, S., Sikela, J.M., Siracusa, L.D., Spearow, J.L., Teuscher, C., Threadgill, D.W., Toth, L.A., Toye, A.A., Vadasz, C., Van Zant, G., Wakeland, E., Williams, R.W., Zhang, H.G., Zou F. 2003. The nature and identification of quantitative trait loci: A community's view. *Natural Reviews Genetics* **4:** 911-916

Affymetrix. 2002. *Statistical Algorithms Description Document.* http://media.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf (accessed 18/01/2010)

Affymetrix. 2005. *GeneChip® Expression Analysis Technical Manual*.http://www.medsci.uu.se/klinfarm/arrayplatform/One%20cycle_two%20cycle_expression_analysis_technical_manual_PN%20702232%20Rev2.pdf (accessed 7/02/2010)

Affymetrix. 2011. http://www.affymetrix.com/estore/ (accessed 17/07/2011)

Ahmad, N.S. 2012. Genetic analysis of plant morphology in Bambara groundnut (*Vigna subterranea* (L.) *Verdc*.). PhD Dissertation. The University of Nottingham.

Akyeampong, E. 1986. Some responses of cowpea to drought stress. In: *Potentials of forage legumes in farming systems of sub-Saharan Africa* (Ed. by Haque, I., Jutzi, S., Neate, P.J.H.). ILCA, Addis Ababa.

Al-Dous, E.K., George, B., Al-Mahmoud, M.E., Al-Jaber, M.Y., Wang, H., Salameh, Y.M., Al-Azwani, E.K., Chaluvadi, S., Pontaroli, A.C., DeBarry, J., Arondel, V., Ohlrogge, J., Saie, I.J., Suliman-Elmeer, K.M., Bennetzen, J.L., Kruegger, R.R, Malek, J.A. 2011. *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotechnology* **29:** 521-528

Ali, M.A., Jabran, K., Awan, S.I., Abbas, A., Ehsanullah, Z.M., Acet, T., Farooq, J., Rehman, A. 2011. Morpho-physiological diversity and its implications for improving drought tolerance in grain sorghum at different growth stages. *Australian Journal of Crop Science* **5:** 308-317

Alqudah, A.M., Samarah, N.H., Mullen, R.E. 2011. Drought stress effect on crop pollination, seed set, yield and quality. In: *Alternative farming systems, biotechnology, drought stress and ecological fertilisation* (Ed. by Lichtfouse, E.) New York: Springer, pp. 193-196

Amadou, H.I., Bebeli, P.J., Kaltsikes, P.J. 2001. Genetic diversity in Bambara groundnut (*Vigna subterranea* L.) germplasm revealed by RADP markers. *Genome* **44:** 995-999

Anandhan, T., Manivannan, N., Vindhiyavarman, P., Jeyakumar, P. 2010. Single marker analysis in sunflower (*Helianthus annuus* L.). *Electronic Journal of Plant Breeding* **1:** 1227-1234

Anyia, A.O. and Herzog, H. 2004. Water-use efficiency, leaf area and leaf gas exchange of cowpeas under mid-season drought. *European Journal of Agronomy* **20:** 327–339

Arthur, D.M. 2010. *Pacific Biosciences introduces new third-generation sequencing instrument at AGBT. http://scienceblogs.com/geneticfuture/2010/02/pacific_biosciences_session_at.php* (accessed 25/05/2010)

Athar, H.R. and Ashraf, M. 2009. Strategies for crop improvement against salinity and drought stress an overview. In: *Salinity and water stress* (Ed. by Ashraf, M., Ozturk, M., Athar, H.R.)*.* New York: Springer, pp.1-16.

Aubert, G., Morin, J., Jacquin, F., Loridon, K., Quillet, M.C., Petit, A., Rameau, C., Lejeune-He´naut, I., Huguet, T., Burstin, J. 2006. Functional mapping in pea, as an aid to the candidate gene approach and for investigating the synteny with the model species Medicago truncatula. *Theoretical and Applied Genetics* **112:** 1024–1041

Bagnaresi, P., Moschella, A., Beretta, O., Vitulli, F., Ranalli, P., Perata, P. 2008. Heterologous microarray experiments allow the identification of the early events associated with potato tuber cold sweetening. *BMC Genomics* **9:** 1-23

Baigorri, H., Antolin, M.C., Sanchez-Diaz, M. 1999. Reproductive responses of two morphologically different pea cultivars to drought. *European Journal of Agronomy* **10:** 119-128

Bar-Or, C., Bar-Eyal, M., Gal, T.Z., Kapulnik, Y., Czosnek, H., Koltai, H. 2006. Derivation of species-specific hybridisation-like knowledge out of cross-species hybridisation results. *BMC Genomics* **7:** 110-122

Bar-Or, C., Czosnek, H., Koltai, H. 2007. Cross-species microarray hybridisations: a developing tool for studying species diversity. *Trends in Genetics* **23:** 200-207

Barrett, B., Griffiths, A., Schreiber, M., Ellison, N., Mercer, C., Bouton, J., Ong, B., Forster, J., Sawbridge, T., Spangenburg, G., Bryan, G., Woodfield, D. 2004. A microsatellite map of white clover. *Theoretical and Applied Genetics* **109:** 596–608.

Basten C.J., Weir B.S. and Zeng Z.B. 2000. *QTL Cartographer*. North Carolina State University, NC.

Basu, S., Mayes, S., Davey, M., Roberts, J.A., Azam-Ali, S.N., Mithen, R., Pasquet, R.S. 2007a. Inheritance of 'domestication' traits in bambara groundnut (*Vigna subterranea* (L.: *Verdc*.). *Euphytica* **157:** 59-68

Basu, S., Roberts, J.A., Azam-Ali, S.N., Mayes, S. 2007b. Bambara groundnut. *In: Genome mapping and molecular breeding in plants – pulses, sugar and tuber* (Ed. by Kole, C.M.). New York: Springer, pp. 159-173

Beavis, W.D. and Grant, D. 1991. A Linkage Map Based on Information from 4 F2 Populations of Maize (*Zea Mays* L). *Theoretical and Applied Genetics* **82:** 636-644.

Begemann, F. 1988. Ecogeographic differentiation of Bambara groundnut (*Vigna subterranea*) in the collection of the International Institute of Tropical Agriculture (IITA). *PhD thesis.* Giessen, Germany.

Behrend, A., Borchert, T., Spiller, M., Hohe, A. 2013. AFLP-based genetic mapping of 'bud-flowering' trait in heather (*Calluna vulgaris*). *BMC Genetics* **14:** 64–74

Belancio, V.P., Hedges, D.J., Deininger, P. 2008. Mammalian non-LTR transposons: For better or worse, in sickness and in health. *Genome Research* **18:** 343-358

Benešová, M., Holá, D., Fischer, L., Jedelský, P.L., Hnilička, F., Wilhelmová, N., Rothová, O., Kočová, M., Procházková, J., Fridrichová, L., Hniličková, H. 2012. The physiology and proteomics of drought tolerance in maize: early stomatal closure as a cause of lower tolerance to short-term dehydration? *PLoS ONE* **7:** e38017. doi:10.1371/journal.pone.0038017

Berchie, J.N., Opoku, M., Adu-Dapaah, H., Agyemang, A., Sarkodie-Addo, J., Asare, E., Addo, J., Akuffo, H. 2012. Evaluation of five Bambara groundnut (*Vigna subterranea* (L.) *Verdc*.) landraces to heat and drought stress at Tono-Navrongo, Upper East Region of Ghana. *African Journal of Agricultural Research* **7:** 250-256

Billotte, N., Risterucci, A.M., Barcelos, E., Noyer, J.L., Amblard, P., Baurens, F.C. 2001. Development, characterisation, and across-taxa utility of oil palm (*Elaeis guineensis* Jacq.) microsatellite markers. *Genome* **44:** 413-425

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

Black, C.R., Tang, D.Y., Ong, C.K., Solon, A., Simmonds, L.P. 1985. Effects of soil moisture stress on the water relations and water use of groundnut stands. *New Phytologist* **100:** 313-328

Black, R.E., Allen, L.H., Bhutta, Z.A., Caulfield, L.E., de Onis, M., Ezzati, M., Mathers, C. and Rivera, J. 2008. Maternal and child undernutrition: global and regional exposures and health consequences. *The Lancet* **371:** 243–260.

Blair, M.W., Galeano, C.H., Tovar, E., Torres, M.C.M., Castrillon, A.V., Beebe, S.E., Rao, I.M. 2012. Development of a Mesoamerican intra-genepool genetic map for quantitative trait loci detection in a drought tolerant x susceptible common bean (*Phaseolus vulgaris* L.) cross. *Molecular Breeding* **29:** 71-88

Blanco, A., Bellomo, M.P., Cenci, A., De Giovanni, C., D'Ovidio, R., Iacono, E., Laddomada, B., Pagnotta, M.A., Porceddu, E., Sciancalepore, A., Simeone, R., Tanzarella, O.A. 1998. A genetic linkage map of durum wheat. *Theoretical and Applied Genetics* **97:** 721–728.

Bombarely, A., Rosli, H.G., Vrebalov, J., Moffett, P., Mueller, L.A., Martin, G.B. 2012. A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. *Molecular plant-microbe interactions* **25:** 1523-1530

Boons, F. and Mendoza, A. 2010. Constructing sustainable palm oil: How actors define sustainability. *Journal of Cleaner Production* **18:** 1686 – 1695.

Boopathi, N.M. 2012. *Genetic mapping and marker assisted selection - basics, practice and benefits.* Bucher: Springer Link, pp. 154-156

Bordat, A., Savois, V., Nicholas, M., Salse, J., Chauveau, A., Bourgeois, M., Potier, J., Houtin, H., Rond, C., Murat, F., Marget, P. 2011. Translational genomics in legumes allowed placing *in sillico 5460* unigenes on the pea functional map and identified candidate genes in *Pisum sativum* L. *G3: Genes, Genomes, Genetics* **1:** 93-103

Brautigam, A., Mullick, T., Schliesky, S., Weber, A.P.M. 2011. Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C_3 and C_4 species. *Journal of Experimental Botany* **62:** 3093-3012

Brem, R.B., Yvert, G., Clinton, R., Kruglyak, L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296(5568):** 752-755

Broadley, M.R., White, P.J., Hammond, J.P., Graham, N.S., Bowen, H.C., Emmerson, Z.F., Fray, R.G., Iannetta, P.P.M., McNicol, J.W., May, S.T. 2008. Evidence of neutral transcriptome evolution in plants. *New Phytologist* **180:** 587-593

Brough, S.H. and Azam-Ali, S.N. 1992. The effect of soil moisture on the proximate composition of Bambara groundnut (*Vigna subterranea* (L.) *Verdc*). *Journal of Science of Food and Agriculture.* **60:** 197-203

Brown, T.A. 2002. How genome evolve. *Genomes 2nd edition*. Oxford: Wiley-Liss.

Bruns, D.E., Ashwood, E.R., Burtis, C.A. 2007. *Fundamentals of Molecular Diagnostics.* US: Saunders, Elsevier Health Sciences, pp. 40-43

Buckley, B.A. 2007. Comparative environmental genomics in non-model species: using heterologous hybridization to DNA-based microarrays. *The Journal of Experimental Biology* **209:** 1602-1606

Budak, H., Kantar, M., Kurtoglu, K.Y. 2013. Drought tolerance in modern and wild wild wheat. *The Science World Journal.* http://dx.doi.org./10.1155/2013/548246 (accessed on 29/12/13).

Bundock, P.C., Eliott, F.G., Ablett, G., Benson, A.D., Casu, R.E., Aitken, K.S., Henry, R.J. 2009. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploidy plant species using 454 sequencing. *Plant Biotechnology Journal* **7:** 347-354

Burgess, J. 2006. Country Pasture/ Forage Resource Profiles – Botswana. FAO. Rome.

Burstin, J., Marget, P., Huart, M., Moessner, A., Mangin, B., Duchene, C., Desprez, B., Munier-Jolain, N., Duc, G. 2007. Developmental genes have pleiotropic effects on plant morphology and source capacity, eventually impacting on seed protein content and productivity in pea. *Plant Physiology* **144:** 768-781

Calvino, M., Miclaus, M., Bruggmann, R., Messing, J. 2009. Molecular markers for sweet sorghum based on microarray expression data. *Rice* **2:** 129-142

Cannon, S.B., Gregory, D.M., Jackson, S.A. 2009. Three sequenced legume genomes and many crop species: rich opportunities for translational genomics. *Plant Physiology* **151:** 970-977

Cattivelli, L., Di Fonzo, N., Mastrangelo, A.M., Mazzucco, L., Rascio, A., Russo M. 2000. Molecular aspects of abiotic stress resistance in durum wheat. In: *Durum wheat improvement in the Mediterranean region: New challenges* (Ed. by Royo, C., Nachit. M., Di Fonzo, N., Araus, J.L.) Zaragoza: CIHEAM, pp. 207-213

Causse, M. A., Fulton, T.M., Cho, Y.G., Ahn, S.N., Chunwongse, J., Wu, K., Xiao, J., Yu, Z., Ronald, P.C., Harrington, S.E. 1994. Saturated molecular map of the rice genome based on an interspecific backcross population. *Genetics* **138:** 1251–1274

Chakravarthi, B.K. and Naravaneni, R. 2006. SSR marker based DNA fingerprinting and diversity study in rice (*Oryza sativa.* L). *African Journal of Biotechnology* **5:** 684-688

Chao, S., Zhang, W., Akhunov, E., Sherman, J., Ma, Y., Luo, M.C., Dubcovsky, J. 2008. Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. *Molecular Breeding* **23:** 23-33

Chao, S., Dubcovsky, J., Dvorak, J., Luo, M.C., Baenziger, S.P., Matnyazov, R., Clark, D.R., Talbert, L.E., Anderson, J.A., Dreisigacker, S., Glover, K., Chen, J., Campbell, K., Bruckner, P.L., Rudd, J.C., Haley, S., Carver, B.F., Perry, S., Sorrells, M.E., Akhunov, E.D. 2010. Population and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L.). *BMC Genomics* **11:** 727 – 744

Chee, P.W., Rong, J., Williams-Coplin, D., Schulze, S.R., Paterson, A.H. 2004. EST-derived PCR-based markers for functional gene homologous in cotton. *Genome* **47:** 449-462

Cheema, J. and Dicks, J. 2009. Computational approaches and software tools for genetic linkage map estimation in plants. *Briefing in Bioinformatics* **6:** 595-608

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S.H., Tingey, S., Morgante, M., Rafalski, A.J. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics.* **3:** 19-40

Choi, I.Y., Hyten, D.L., Matukumalli, L.K., Song, Q.J., Chaky, J.M., Quigley, C.V., Chase, K., Lark, K.G., Reiter, R.S., Yoon, M.S., Hwang,, E.Y., Yi, S.I., Young, N.D., Shoemaker, R.C., van Tassell, C.P., Specht, J.E., Cregan, P.B. 2007. A soybean transcript map: Gene distribution, haplotype and single-nucleotide polymorphism analysis. *Genetics* **176:** 685-696.

Churchill, G.A. and Doerge, R.W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* **138:** 963-971

Clenn, T.C. 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11:** 759-769

Collard, B.C.Y., Jahufer, M.Z.Z., Brouwer, J.B., Pang, E.C.K. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* **142:** 169-196

Collino, D.J., Dardanelli, J.L., Sereno, R., Racca, R.W., 2001. Physiological responses of argentine peanut varieties to water stress - light interception, radiation use efficiency and partitioning of assimilates. *Field Crops Research* **70:** 177–184

Collinson, S.T., Clawson, E.J., Azam-Ali, S.N., Black, C.R. 1997. Effects of soil moisture deficits on water relations of bambara groundnut (*Vigna subterranea* L. *Verdc.*). *Journal of Experimental Botany* **48:** 877-884

Cone, K.C., McMullen, M.D., Bi, I.V., Davis, G.L., Yim, Y.S., Gardiner, J.M., Polacco, M.L., Sanchez-Villeda, H., Fang, Z., Schroeder, S.G., Havermann, S.A., Bowers, J.E., Paterson, A.H., Soderlund, C.A., Engler, F.W., Wing, R.A., Coe, E.H. 2002. Genetic, physical, and informatics resources for maize. On the road to an integrated map. *Plant Physiology* **130:**1598-1605.

Corley, R.H.V. and Tinker, P.B. 2003. *The Oil Palm 4th edition.* Malden, Massachusetts: Blackwell, pp. 163-173

Cromwell, E., Cooper, D and Mulvany, P. 1997. Agriculture, biodiversity and livelihoods: issues and entry points for development agencies. http://www.ukabc.org/odi_agbiod.pdf

Cruz, V.M.V., Kilian, A., Dierig, D.A. 2013. Development of DArT Marker Platforms and Genetic Diversity Assessment of the U.S. Collection of the New Oilseed Crop Lesquerella and Related Species. *PLoS ONE* **8(5):** e64062. doi:10.1371/journal.pone.0064062

Cruz-Izquierdo, S., Avila, C.M., Satovic, Z., Palomino, C., Gutierrez, N., Ellwood, S.R., Phan, H.T.T., Cubero, J.I., Torres, A.M. 2012. Comparative genomics to bridge *Vicia faba* with model and closely-related legume species: stability of QTLs for flowering and yield-related traits. *Theoretical and Applied Genetics* **125:**1767–1782

DArT. 2012. Papers about DArt. Diversity Arrays Technology, Pty Ltd. Available: http://www.diversityarrays.com/publications.html accessed 15/1/14

DArT. 2013. DArTseq overview. Available: http://www.diversityarrays.com/dart-application-dartseq (accessed on 12/1/14)

Das, S., Bhat, P.R., Sudhakar, C., Ehlers, J.D., Wanamaker, S., Roberts, P.Q., Cui, X.P., Close, T.J. 2008. Detection and validation of single feature polymorphisms in cowpea (*Vigna unguiculata* L. Walp) using a soybean genome array. *BMC Genomics* **9:** 1-12

Davey, M.W., Graham, N.S., Vanholme, B., Swennen, R., May, S.T., Keulemans, J. 2009. Heterologous oligonucleotide microarrays for transcriptomics in a non-model species; a proof-of-concept study of drought stress in *Musa*. *BMC Genomics* **10:** 1-19

Davies, W. and Zhang, J.J. 1991. Root signals and the regulation of growth and development of plant in drying soil. *Annual Review of Plant Physiological and Plant Molecular Biology* **42:** 55-76

De Keyser, E., Shu, Q., Van Bockstaele, E., De Riek, J. 2010. Multipoint-likelihood maximization mapping on 4 segregating populations to achieve an integrated framework map for QTL analysis in pot azalea (*Rhododendron simsii* hybrids). *BMC Molecular Biology* **11:** 1–20

DeCook, R., Lall, S., Nettleton, D., Howell, S.H. 2006. Genetic regulation of gene expression during shoot development in Arabidopsis. *Genetics* **172(2):** 1155-1164

Die, J.V. and Rowland, L.J. 2013. Superior Cross-Species Reference Genes: A Blueberry Case Study. *PLoS ONE* **8:** e73354. doi:10.1371/journal.pone.0073354

Diers, B.W. 1992. RFLP analysis of soybean seed protein and oil content. *Theoretical and Applied Genetics* **83:** 608- 612

Doligez, A., Adam-Blondon, A.F., Cipriani, G., Di Gaspero, G., Laucou, V., Merdinoglu, D., Meredith, C.P., Riaz, S., Roux, C., This, P. 2006. An integrated SSR map of grapevine based on five mapping populations. *Theoretical and Applied Genetics* **113:** 369-382.

Doyle, J.J. and Luckow, M.A. 2003. The rest of the iceberg: legume diversity and evolution in a phylogenetic context. *Plant Physiology* **131:** 900–910

Druka, A., Potokina, E., Luo, Z., Jiang, N., Chen, X., Kearsey, M., Waugh, R. 2010. Expression quantitative trait loci analysis in plants. *Plant Biotechnology Journal* **8:** 10-27

Ebdon, J.S. and Kopp, K.L. 2004. Relationships between water use efficiency, carbon isotope discrimination, and turf performance in genotypes of Kentucky bluegrass during drought. *Crop Science* **44:** 1754-1762

Eckardt, N.A. 2009. Deep sequencing maps the maize epigenomic landscape. *The Plant Cell* **21:** 1024-1026

Edger, P.P. and Pires, C. J. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research* **17**: 699–717

El-Hennawy, M.A., Adballa, A.F., Shafey, S.A., Al-Ashkar, I.M. 2011. Production of doubled haploid wheat lines (*Triticum aestivum* L.) using anther culture technique. *Annals of Agricultural Sciences* **56:** 63-72

Ellwood, S.R., D'Souza, N.K., Kamphuis, L.G., Burgess, T.I., Nair, R.M., Oliver, R.P. 2006. SSR analysis of the *Medicago truncatula* SARDI core collection reveals substantial diversity and unusual genotype dispersal throughout the Mediterranean basin. *Theoretical and Applied Genetics* **112:** 977–983.

Enoki, H., Sato, H., Koinuma, K. 2002. SSR analysis of genetic diversity among maize inbred lines adapted to cold regions of Japan. *Theoretical and Applied Genetics* **104:** 1270-1277

Eslami, S.V., Gill, G.S., McDonald, G., 2010. Effect of water stress during seed development on morphometric characteristics and dormancy of wild radish (*Raphanus raphanistrum* L.) seeds. *International Journal of Plant Production* **4:** 159-168

Farquhar, G.D., Ehleringer, J.R., Hubick, T. 1989. Carbon isotope discrimination and photosynthesis. *Annual Review of Plant Physiology and Plant Molecular Biology* **40:** 503-537

Food and Agriculture Organization of the United Nations (FAO). 1993. *Harvesting nature's diversity*. FAO, Rome, Italy (accessed on 13/01/14)

Food and Agriculture Organization of the United Nations (FAO). 2009. *How to feed the world in 2050.* FAO, Rome, Italy (accessed on 17/01/14)

Food and Agriculture Organization of the United Nations (FAO). 2011. *Biodiversity for Food and Agriculture – Contributing to food security and sustainable in a changing world*, FAO, Rome, Italy (accessed on 17/01/14)

Food and Agriculture Organization of the United Nations (FAO). 2012. *The Commission on Genetic Resources for Food and Agriculture (CGRFA) Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture (PGRFA)*, FAO, Rome, Italy (accessed on 17/01/14)

Foucher, F., Morin, J., Courtiade, J., Cadioux, S., Ellis, N., Banfield, M.J., Rameau, C. 2003. Determine and late flowering is two Terminal FLOWER1/CENTRORADIALIS homologs that control two distinct phases of flowering initiation and development in pea. *Plant Cell* **15:** 2742–2754

Franks, F., Hatley, R.H.M., Mathias, S.F. 1991. Materials science and the production of self-stable biological. *BioPharm* **4:** 38-42

Fuste, B. 2009. *Present and future of DNA sequencing.* http://www.docstoc.com/docs/24567906/PRESENT-AND-FUTURE-OF-DNA-SEQUENCING (accessed 11/02/2010)

Garcia, A.A.F, Kido, E.A., Meza, A.N., Souza, H.M.B., Pinto, L.R., Pastina, M.M., Leite, C.S., da Silva, J.A.G., Ulian, E.C., Figueira, A., Souza, A.P. 2006. Development of an integrated genetic map of a sugarcane (*Saccharum* spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. *Theoretical and Applied Genetics* **112:** 298-314

Gardner, W.R. and H.R. Gardner. 1983. Principles of water management under drought conditions. *Agricultural Water Management*. **7:**143-155.

Gaur, P.M., Krishnamurthy, L., Kashiwagi, J. 2008. Improving drought-avoidance root traits in Chickpea (*Cicer arietinum.* L.) – Current status of research at ICRISAT. *Plant Production Science* **11:** 3-11

Genome Web. 2010. *Ion Torrent Systems Presents $50,000 Electronic Sequencer at AGBT.* http://www.genomeweb.com/sequencing/ion-torrent-systems-presents-50000-electronic-sequencer-agbt?page=2 (accessed 25/05/2010)

Ghanbari, A.A., Shakiba, M.R., Toorchi, M., Choukan, R. 2013. Morpho-physiological responses of common bean leaf to water deficit stress. *European Journal of Experimental Biology* **3:** 487-492

Gharizadeh, B., Ghaderi, M., Nyren, P. 2007. Pyrosequencing Technology for Short DNA Sequencing and Whole Genome Sequencing. *Biophysics* **47:** 129-132

Gill, K.S., Gill, B.S., Endo, T.R., Boyko, E.V. 1996. Identification of high density mapping of gene-rich regions in chromosome group 5 of wheat. *Genetic* **143:** 1001-1012

Goff, S.A., Ricke, D., Lan, T.H. *et al.* **2002**. A draft sequence of the rice genome (Oryza sativa L. ssp. Japonica). *Science.* **296:** 92-100

Gondo, T., Sato, S., Okumura, K., Tabata, S., Akashi, R., Isobe, S. 2007. Quantitative trait locus analysis of multiple agronomic traits in the model legume *Lotus japonicas*. *Genome.* **50:** 627-637

Gowda, B.S., Miller, J.L., Rubin, S.S., Sharma, D.R., Timko, M.P. 2002. Isolation, sequence analysis, and linkage mapping of resistance gene analogs in cowpea (*Vigna unguiculata L.Walp.). Euphytica* **126:** 365-377

Graham, N.S., Broadley, M.R., Hammond, J.P., White, P.J., May, S.T. 2007. Optimising the analysis of transcript data using high density oligonucleotide arrays and genomic DNA-based probe selection. *BMC Genomics* **8:** 344-352

Graham, N.S., May, S.T., Daniel, Z.C.T.R., Emmerson, Z.F., Brameld, J.M., Parr, T. 2010. Use of the affymetrix human geneChip array and genomic DNA hybridisation probe selection to study ovine transcriptomes. *Animal* **5:** 861-866

Gualtieri, G., Kulikova, O., Limpens, E., Kim, D. J., Cook, D. R., Bisselin, T., Geurts, R. 2002. Microsynteny between pea and *Medicago truncatula* in the SYM2 region. *Plant Molecular Biology* **50:** 225-235.

Guillaume, B. and Kenneth, H.W. 2004. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *The Plant Cell* **16:** 1667-1678

Gupta, N.K., Gupta, S., Kumar, A. 2001. Effect of water stress on physiological attributes and their relationship with growth and yield in wheat cullivars at different growth stages. *European Journal of Agronomy* **86:** 1437-1439

Gupta, P.K., Rustgi, S., Mir, R.R. 2013. Array-Based High-Throughput DNA Markers and Genotyping Platforms for Cereal Genetics and Genomics. In: *Cereal Genomics II* (Ed. by Gupta, P.K. and Varshney, R.K.). New York: Springer Science+Business Media Dordrecht

Hackett, C.A. and Broadfoot, L.B. 2003. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* **90:** 33-38

Haldane, J.B.S. 1931. The cytological basis of genetical interference. *Cytologia* **3:** 54–65

Hammond, J.P., Broadley, M.R., Craigon, D.J., Higgins, J., Emmerson, Z.F., Townsend, H.J., White, P.J., May, S.T. 2005. Using genomic DNA-based probe-selection to improve the sensitivity of high-density oligonucleotide arrays when applied to heterologous species. *Plant Methods* **1:** 1-10

Hammond, J.P., Bowen, H.C., White, P.J., Mills, V., Pyke, K.A., Baker, A.J.M., Whiting, S.N., May, S.T., Broadley, M.R. 2006. A comparison of the *Thlaspi caerulescens* and *Thlaspi arvense* shoot transciptomes. *New Phytologist* **170:** 239-260

Hammond, J.P., Mayes, S., Bowen, H.C., Graham, N.S., Hayden, R.M., Love, C.G., Spracklen, W.P., Wang, J., Welham, S.J., White, P.J., King, G.J., Broadley, M.R. 2011. Regulatory hotspots are associated with plant gene expression under varying soil phosphorus supply in *Brassica rapa*. *Plant Physiology* **156:** 1230-1241

Harb, A., Krishnan, A., Ambavaram, M.M.R, Pereira, A. 2010. Molecular and physiological analysis of drought stress in Arabidopsis reveals early responses leading to acclimation in plant growth. *Plant Physiology* **154:**1254-1271

Hare, P.D. and Cress, W.A. 1997. Metabolic implications of stress induced proline accumulation in plants. *Plant Growth Regulation* **21:** 79-102

Hart, D. and Jones, E. 2001. *Genetics: Analysis of genes and genomes*. Jones and Bartlett Publishers, Sudbury, MA.

Hartley, C.W.S. 1967. *The Oil Palm.* London: Longmans, pp. 706

Harushima, Y., Yano, M., Shomura, A., Sato, M., Shimano, T., Kuboki, Y., Yamamoto, T., Lin, S.Y., Antonio, B.A., Parco, A., Kajiya, H., Huang, N., Yamamoto, K., Nagamura, Y., Kurata, N., Khush, G.S., Sasaki, T. 1998. A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics.* **148:** 479–494

Hazir, M.H.M., Shariff, A.R.M., Amiruddin, M.D. 2012. Determination of oil palm fresh fruit bunches – Based on flavonoids and anthocyanin content. *Industrial Crops and Products* **36:** 466 – 475.

Heller, J., Begemann, F., Mushonga, J. 1997. Bambara groundnut (*Vigna subterranea* (*L.*) *Verdc.*). In: Promoting the conservation and use of underutilized and neglected crops. 9. Proceedings of the workshop on Conservation and Improvement of Bambara Groundnut (*Vigna subterranea* (L.) *Verdc.*), 14–16 November 1995, Harare, Zimbabwe. International Plant Genetic Resources Institute (IPGRI), Rome, Italy.

Hocking, P.J. 1994. Dry-matter production, mineral nutrient concentrations, and nutrient distribution and redistribution in irrigated spring water. *Journal of Plant Nutrition* **17:** 1289-1308

Horner, D.S., Pavesi, G., Castrignano, T., De Meo, P.D., Liuni, S., Sammeth, M., Picardi, E., Pesole, G. 2009. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefing In Bioinformatics* **11:** 181-197

Hsiao, T.C. 1973. Plant responses to water stress. *Annual Review of Plant Physiology* **24:** 519-570

Hu, J., Sadowski, J., Osborn, T.C., Landry, B.S., Quiros, C.F. 1998. Linkage group alignment from four independent Brassica oleracea RFLP maps. *Genome* **41:** 226-235.

Huang, X.Y., Chao, D.Y., Gao, J.P., Zhu, M.Z., Shi, M., Lin, H.X.. 2009. A previously unknown zinc finger protein, DST, regulates drought and salt tolerance in rice via stomatal aperture control*. Genes and Development.* **23:** 1805-1817

Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* **9:** 868-877

Hunter, D. and Fanzo, J. 2013. Agricultural biodiversity, diverse diets and improving nutrition. In: *Diversifying food and diets: using agricultural biodiversity to improve nutrition and health* (Ed. by Fanzo, H., Hunter, D., Borelli, T. Mattei, F.). London: Routledge, pp. 1-14

Iowa State University. 2011. FAPRI-ISU 2011 World Agricultural Outlook. Food and Agricultural Policy Research Institute, Lowa: Ames

IPGRI, IITA, BAMNET, 2000. *Descriptors for Bambara groundnut (Vigna subterranea)*. International Plant Genetic Resources Institute, Rome, Italy; International Institute of Tropical Agriculture, Ibadan, Nigeria; The International Bambara Groundnut Network, Germany.

Jaccoud, D., Peng, K., Feinstein, D., Kilian, A. 2001. Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Research.* **29:***e25.*

Jansen, J., De Jong, A.G., Van Ooigen, J.W. 2001. Constructing dense genetic linkage maps. *Theoretical and Applied Genetics* **102:** 1113-1122

Jansen, R.C. 1993. Interval mapping of multiple quantitative trait loci. *Genetics* **135:** 205-211

Jansen, R.C. 1994. Controlling the Type I and Type II errors in mapping quantitative trait mapping. *Genetics* **138:** 871-881

Jansen, R.C. and Nap, J.P. 2001. Genetical genomics: The added value from segregation. *Trends in Genetics.* **17:** 388-391

Jefferies, S.P., Pallotta, M.A., Paull, J.G., Karakousis, A., Kretschmer, J.M., Manning, S., Islam, A.K.M.R., Langridge, P., Chalmers, K.J. 2000. Mapping and validation of chromosome regions conferring boron toxicity tolerance in wheat (*Triticum aestevum*). *Theoretical and Applied Genetics* **101:** 767–777

Jones, N., Ougham, H., Thomas, H. 1997. Markers and mapping: We are all geneticists now. *New Phytologist* **137:**165-177

Joosen, R.V.L., Ligterink, W., Hilhorst, H.W.M.M, Keurentjes, J.J.B. 2009. Advances in genetical genomics of plants. *Current Genomics* **10:** 540-549

Jordan, M.C., Somers, D.J., Banks, T.W. 2007. Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. *Plant Biotechnology Journal. 5*: 442-453

Jorgensen, S.T., Ntundu, W.H., Ouedraogo, M., Christiansen, J.L., Liu, F. 2011. Effect of a short and severe intermittent drought on transpiration, seed yield, yield components, and harvest index in four landraces of Bambara groundnut. *International Journal of Plant Production* **5:** 25-36

Jouannic, S., Argout, X., Lechauve, F., Fizames, C., Borgel, A., Morcillo, F., Aberlenc-Bertossi, F., Duval, Y., Tregear, J. 2005. Analysis of expressed sequence tags from oil palm (*Elaeis guineensis*). *Federation of European Biochemical Societies* **579:** 2709-2714

Jusoff, H.K. 2008. Airborne hyperspectral imagery for agricultural businesses in Malaysia. *International Business Research* **3:** 54 – 62.

Kameli, A. and Losel, D.M. 1993. Carbohydrates and water status in wheat plants under water stress. *New phytologist* **125:** 609-614

Kantardjieff, A., Nissom, P.M., Chuah, S.H., Yusufi, F., Jacob, N.M., Mulukutla, B.C., Yap, M., Hu, W.S. 2009. Developing genomic platforms for Chinese hamster ovary cells. *Biotechnology Advances* **27:** 1028-1035

Keurentjes, J.J.B., Fu, J., Terpstra, I.R., Garcia, J.M., van den Ackerveken, G. Snoek, L.B., Peeters, A.J., Vreugdenhil, D., Koornneef, M., Jansen, R.C. 2007. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences* **104**: 1708-1713

Kgathi, D.L., Mazonde, I., Murray-Hudson, M. 2012. Water Implications of biofuel development in semi-arid sub-saharan Africa: Case studies for Four Countries In: *Bioenergy for Sustainable Development in Africa* (Ed. by Janssen, R and Rutz, D.). New York: Springer, pp. 261-280

King, J., Thomas, A., James, C., King, I., Armstead, I. 2013. A DArT marker genetic map of perennial ryegrass (*Lolium perenne* L.) integrated with detailed comparative mapping information; comparison with existing DArT marker genetic maps of *Lolium perenne, L. multiflorum and Festuca pratensis. BMC Genomics* **14:** 437 - 444

Kirst, M., Myburg, A.A., De Leon, J.P., Kirst, M.E., Scott, J., Sederoff, R. 2004. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. *Plant Physiology* **135:** 2368 – 2378

Kliebenstein, D. 2009. Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annual Review of Plant Biology* **60:** 93-114

Kobayashi, Y. and Koyama, H. 2002. QTL analysis of Al tolerance in recombinant inbred lines of *Arabidopsis thaliana. Plant Cell Physiology* **12:** 1526-1533

Kosambi, D.D. 1944. The estimation of map distances from recombination values. *Annals of Eugenics* **12:** 172-175

Koyama, M.L., Levesley, A., Koebner, R.M.D., Flowers, T.J., Yeo, A.R. 2001. Quantitative trait loci for component physiological traits determining salt tolerance in rice. *Plant Physiology* **125:** 406-422

Krutovskii, K.V., Vollmer, S.S., Sorensen, F.C., Adams, W.T., Knapp, S.J., Strauss, S.H. 1998. RAPD genome maps of Douglas-fir. *The Journal of Heredity* **89:** 197-205

Krutovsky, K.V., Troggio, M., Brown, G.R., Jermstad, K.D., Neale, D.B. 2004. Comparative mapping in the Pinaceae. *Genetics* **168:** 447-461

Kubisiak, T.L., Nelson, C.D., Nance, W.L., Stine, M. 1995. RAPD linkage mapping in a longleaf pine 3 slash pine F1 family. *Theoretical and Applied Genetics* **90:** 1119–1127

Kumar, S. and Blaxter, M.L. 2010. Comparing *de novo* assemblers for 454 transcriptome data. *BMC Genomics* **11:** 571- 582

Kurata, N., Nagamura, Y., Yamamoto, K., Harushima, Y., Sue, N., Wu, J. 1994. A 300 kilobase interval genetic map of rice including 883 expressed sequences. *Nature Genetics* **8:** 365-372

Lai, H.M. 2009. *PIGEONS: An innovative bioinformatics software to across species investigation in minor crops based on the XSpecies microarray technologies.* M.Res. Thesis. The University of Nottingham.

Lai, H.M., May, S.T., Mayes, S. 2014. Pigeons: A novel GUI software for analysing and parsing high density heterologous oligonucleotide microarray probe level data. *Microarrays* **3:**1-23

Lai, J., Ma, J., Swigonova, Z., Ramakrishna, W., Linton, E., Llaca, V., Tanyolac, B., Park, Y.J., Jeong, O.Y., Bennetzen, J.L., Messing, J. 2006. Gene loss and movement in the maize genome. *Genome Research* **14:** 1924-1931

Lander, E. and Botstein, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121:** 185-199

Lander, E.S, Grene, J., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E., Newburg, L. 1987. Mapmaker an interative computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1:** 174-181

Lang, N.T. and Buu, B.C. 2008. Fine mapping for drought tolerance in rice (*Oryza sativa* L.). *Omonrice* **16:** 9-15

Laperche, A., Devienne-Barret, F., Maury, O., Le Gouis, J., Ney, B. 2006. A simplified conceptual model of carbon/nitrogen functioning for QTL analysis of winter wheat adaptation to nitrogen deficiency. *Theoretical and Applied Genetics* **113:** 1131–1146

Lester, D.R., Ross, J.J., Davies, P.J., Reid, J.B. 1997. Mendel's stem length gene (Le) encodes a gibberellin 3b-hydroxylase. *Plant Cell* **9:** 1435–1443

Linnemann, A.R. and Azam-Ali, S.N. 1993. Bambara groundnut (*Vigna subterranea* L. Verdc). In: *Under-utilised Crops Series. II. Vegetable and Pulses.* Chapman and Hall, London.

Li, J. and Burmeister, M. 2005. Genetical genomics: Combining genetics with gene expression analysis. *Human Molecular Genetics* **14:** 163 – 169

Li, R., Li, Y., Kristiansen, K., Wang, J. 2008. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24:** 713-714

Liu, B.H. 1997. *Statistical Genomics: Linkage, Mapping, and QTL Analysis.* Florida: CRC Press, pp. 328-329.

Liu, B.H., Fujita, T., Yan, Z.H., Sakamoto, S., Xu, D., Abe, J. 2007. QTL mapping of domestication-related traits in Soybean (*Glyxine max*). *Journals of Botany* **100:** 1027-1038

Liu, F. 2004. Physiological regulation of pod set in soybean (*Glycine max* L. Merr.) during drought at early reproductive stages. Phd Dissertation. The Royal Veterinary and Agicultural University, Copenhagen, Denmark.

Liu, F., Andersen, M.N., Jensen, C.R. 2003. Loss of pot set caused by drought stress is associated with water status and ABA content of reproductive structures in soybean. *Functional Plant Biology.* **30:** 271-280

Lorieux, M., Perrier, X., Goffinet, B., Lanaud, C., Gonzalez, D.L. 1995. Maximum-likelihood models for mapping genetic markers showing segregation distortion. 2. F2 populations. *Theoretical and Applied Genetics.* **90:** 81-89

Low, E.T.L., Tan, J.S., Chan, P.L., Boon, S.H., Wong, Y.L. 2006. Developments toward the application of DNA chip technology in oil palm tissue culture. *Journal of Oil Palm Research* **(Special Issue – April 2006):** 87-96

Lu, H., Romero-Severson, J., Bernardo, R. 2002. Chromosomal regions associated with segregation distortion in maize. *Theoretical Applied Genetics* **105:** 622–628.

Lu, Y., Huggins, P., Bar-Joeseph, Z. 2009. Cross species analysis of microarray expression data. *Gene Expression* **25:** 1476-1483

Ludlow, M.M. 1989. Strategies of response to water stress. In: *Structural and functional responses to environmental stresses: Water shortage* (Ed. by Kreb, K.H., Richter, H., Hinkley, T.M.) The Netherlands: Academic publishing, pp. 269-282

Mabhaudhi, T., Modi, A.T., Beletse, Y.G. 2013. Growth, phenological and yield responses of a bambara groundnut (Vigna subterranea L. *Verdc*) landrace to imposed water stress: II. Rain shelter conditions. *Water SA* **39:** 191-198

Manly, K.F., Cudmore, J.R.H, Meer, J.M. 2001. Map Manager QTX, cross-platform software for genetic mapping. *Mammalian Genome* **12:** 930-932

Mardis, E.R. 2008. The Impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24:** 133-141

Maroco, J.P., Pereira, J.S., Chaves, M. 1997. Stomatal responses to leaf-to-air vapour pressure deficit in Sahelian speices. *Australian Journal of Plant Physiology* **24:** 381-387

Martinez-Garcia, P.J., Stevens, K.A., Wegrzyn, J.L., Liechty, J., Crepeau, M., Langley, C.H., Neale, D.B. 2013. Combination of multipoint maximum likelihood (MML) and regression mapping algorithms to construct a high-density genetic linkage map for loblolly pine (*Pinus taeda* L.). *Tree Genetics and Genomes* **9:** 1529 - 1535

Massawe, F.J., Dickson, M., Roberts, J.A., Azam-Ali, S.N. 2002. Genetic diversity in bambara groundnut (*Vigna subterranea* (L.) *Verdc*) landraces revealved by AFLP markers. *Genome* **45:** 1175-1180

Massawe, F.J., Mwale, S.S., Azam-Ali, S.N., Roberts, J.A. 2005. Breeding in bambara groundnut (*Vigna subterranea* (L) *Verdc.*): Strategic considerations. *African Journal of Biotechnology* **4:** 463-471

Mayes, S., James, C.M., Horner, S.F., Jack, P.L., Corley, R.H.V. 1996. The application of restriction fragment length polymorphism for the genetic fingerprinting of oil palm (*Elaeis guineensis* Jacq). *Molecular Breeding* **2:** 175-180

Mayes, S., Jack, P.L., Marshall, D.F., Corley, R.H.V. 1997. Construction of a RFLP genetic linkage map for oil palm (*Elais guineensis* Jacq). *Genome* **40:** 116-122

Mayes, S., Hafeez, F., Price, Z., MacDonald, D., Billotte, N., Roberts, J. 2008. Molecular Research in Oil Palm, the Key Oil Crop for the Future. In: *Genomics of tropical crop plants* (Ed. by Moore, P.H. and Ming, R.)*. New York: Springer, pp. 371-404

McClean, P.E., Mamidi, S., McConnell, M., Chikara, S., Lee, R. 2010. Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. *BMC Genomics* **11:** 184 http://www.biomedcentral.com/1471-2164/11/184

Meissner, M., Orsini, E., Ruschhaupt, M., Melchinger, A.E., Hincha, D.K., Heyer, A.G. 2013. Mapping quantitative trait loci for freezing tolerance in a recombinant inbred line population of Arabidopsis thaliana accessions Tenela and C24 reveals REVEILLE1 as negative regulator of cold acclimation. *Plant, Cell and Environment* **36:** 1256-1267

Meng, L., Li, L., Chen, W., Xu, Z., Liu, L. 1999. Effect of water stress and temperature on leaf size and number of epidermal cells in grain sorghum. *Crop Science* **14:** 751-755

Michael, C.S., Arthur, L.D., Steven, L.S. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research* **20:** 1165-1173

Michelmore, R.W., Paran, I, Kesseli, R.V. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences* **88:** 9828-9832

Micronutrient Initiative. 2009. *Investing in the Future: A united call to action on vitamin and mineral deficiencies. http://*www.unitedcalltoaction.org (accessed on 16/01/04)

Miflin, B.J. and Habash, D.Z. 2002. The role of glutamine synthetase and glutamate dehydrogenase in nitrogen assimilation and possibilities for improvement in the nitrogen utilization of crops. *Journal of Experimental Botany* **53:** 979-987

Mohammadkhani, N. and Heidari, R. 2008. Drought-induced accumulation of soluble sugar and proline in two maize varieties. *World Applied Sciences Journal* **3:** 448-453

Mohan, M., Nair, S., Bhagwat, A., Krishna, T.G., Yano, M., Bhatia, C.R., Sasaki, T. 1997. Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding* **3:** 87-103

Molosiwa, O.O. 2012. *Genetic diversity and population structure analysis of bambara groundnuts (Vigna subterranea (L.) Verdc.) landraces using morpho-agronomic characters and SSR markers.* PhD thesis, University of Nottingham.

Moore, S., Payton, P., Wright, M., Tanksley, S., Giovannoni, J. 2005. Utilisation of tomato microarrays for comparative gene expression analysis in the Solanaceae. *Journal of Experimental Botany* **56:** 2885-2895

MPOB. 2012. Overview of the Malaysian oil palm industry 2012. http://bepi.mpob.gov.my/images/overview/Overview_of_Industry_2012.pdf (assessed on 2/10/2013).

Muchero, W., Diop, N.N., Bhat, P.R., Fenton, R.D., Wanamaker, S., Pottorff, M., Hearne, S., Cisse, N., Fatokun, C., Ehlers, J.D., Roberts, P.A., Close, T.J. 2009. A consensus genetic map of cowpea [*Vigna unguiculata* (L) Walp.] and synteny based on EST-derived SNPs. *Proceedings of the National Academy of Sciences of the United States of America* **106:** 18159-18164

Mwale, S.S., Azam-Ali, S.N., Massawe, F.J. 2007. Growth and development of bambara groundnut (*Vigna subterranea*) in response to soil moisture 1. Dry matter and yield. *European Journal of Agronomy* **26:** 345-353

Mwanamwenge, J., Loss, S.P., Siddique, K.H.M., Cocks, P.S. 1999. Effect of water stress during floral initiation, flowering and podding on the growth and yield of faba bean (*Vicia faba* L.). *European Journal of Agronomy* **11:** 1-11

Myers, G.O., Fatokun, C.A., Young, N.D. 1996. RFLP mapping of an aphid resistance gene in cowpea (*Vigna unguiculata* L. Walp). *Euphytica* **91:** 181-187

Nagabhushana, L., Mane, S.P., Hittalmani, S. 2006. Comparative studies on QTL mapping by simple interval mapping and composite interval mapping models for selected growth and yield traits in rice (*Oryza sativa* L.). *Indian Journal of Crop Science* **1:** 97-101

NASC's International Affymetrix Service. 2011. *QC guide.* http://affy.arabidopsis.info/qc.html (accessed on 03/07/2011)

National Drought Mitigation Center. 2003. *Defining Drought: Overview.* http://www.drought.unl.edu

NCBI news. 2002. *BLAST Lab – Searching the trace achieve with discontiguous MegaBlast.* http://www.ncbi.nlm.nih.gov/Web/Newsltr/FallWinter02/blastlab.html (access on 26/01/14)

Nelson, M.N., Moolhuijzen, P.M., Boersma, J.G., Chudy, M., Lesniewska, K., Bellgard, M., Oliver, R.P., Swiecicki, W., Wolko, B., Cowling, W.A., Ellwood, S.R. 2010. Aligning a dense genetic map of Lupinus angustifolius with the genome sequence of the model legume *Lotus japonicus*. *DNA Research* **17**:73–83

Ntundu, W.H., Bach, I.C., Christiansen, J.L., Anderson, S.B. 2003. Analysis of genetic diversity in Bambara groundnut (*Vigna subterranea* L. *Verdc*) landraces using amplified fragment length polymorphisms (AFLP) markers. *African Journal of Biotechnology* **3:** 220-225

Obagwu, J. 2003. Evaluation of bambara groundnut (*Vigna subterranea* (L.) *Verdc*) lines for reaction to *Cercospora* spot. *Journal of Sustainable Agriculture* **22:** 93–100

Ocampo, E.T.M. and Robles, R.P. 2000. Drought tolerance in Mungbean. I. Osmotic adjustment in drought stressed Mungbean. The *Philippine Journal* of *Crop Science* **25:** 1-5

Oguntibeju, O.O., Esterhuyse, A.J., Truter, E.J. 2009. Red palm oil: nutritional, physiological and therapeutic roles in improving human wellbeing and quality of life. *British Journal of Biomedical Science* **66:** 216-222

Oil World. 2007. *Palm Oil: Gift to Nature, Gift to Life.* http://www.americanpalmoil.com/publications/Brief%20Palm%20Oil%20Story.pdf (accessed 29/09/2011)

Okcu, G., Kaya, M.D., Atak, M. 2005. Effects of salt and drought stresses on germination and seedling growth of pea (*Pisum sativum* L.). *Turkish Journal of Agriculture and Forestry* **29:** 237–242

Oldroyd, G.E. and Downie, J.A. 2008. Coordinating nodule morphogenesis with rhizobial infection in legumes. *Annual Review of Plant Biology* **59:** 519–546

Oliver, R.E., Jellen, E.N., Ladizinsky, G., Korol, A.B., Kilian, A., Beard, J.L., Dumlupinar, Z., Wisniewski-Morehead, N.H., Svedin, E., Coon, M., Redman, R.R., Maughan, P.J., Obert, D.E., Jackson, E.W. 2011. New Diversity Arrays Technology (DArT) markers for tetraploid oat (Avena magna Murphy et Terrell) provide the first complete oat linkage map and markers linked to domestication genes from hexaploid *A. sativa* L. *Theoretical and Applied Genetics* **123:** 1159-1171

Otieno, D.O., Schmidt, M.W.T., Adiku, S., Tenhunen, J. 2005. Physiological and morphological responses to water stress in two *Acacia* species from contrasting habitats. *Tree Physiology* **25:** 361-371

Padulosi, S. 1999. *Final Report:Conservation and Use of Underutilized Mediterranean Species.* Aleppo, Syria: IPGRI Regional Office for Central and West Asia and North Africa.

Palmer, J.D. and Zamir, D. 1982. Chloroplast DNA evolution and phylogenetic relationships in *Lycopersicon. Proceedings of the National Academy of Sciences* **79:** 5006-5010

Pariset, L., Chillemi, G., Bongiorni, S., Spica, V.R., Valentini, A. 2009. Microarrays and high-throughput transcriptomic analysis in species with incomplete availability of genomic sequences. *New Biotechnology* **25:** 272-279

Park, Y.J., Lee, J.K., Kim, N.S. 2009. Simple Sequence Repeat Polymorphisms (SSRPS) for Evaluation of Molecular Diversity and Germplasm Classifications of Minor Crops. *Molecules* **14:** 4546-4569

Paterson, A.H. 1996. Making genetic maps. In: *Genome mapping in plants* (Ed. by Paterson, A.H.). San Diego, California: R.G. Landes Company Academic Press

Pedercini, M., Kanamaru, H. and Derwisch, S. 2012. Potential impacts of climate change on food security in Mali. Natural Resources Management and Environment Department, FAO, Rome

Peng, J.R., Richards, D.E., Hartley, N.M. Murphy, G.P., Devos, K.M., Flintham, J.E., Beales, J., Fish, L.J., Worland, A.J., Pelica, F., Sudhakar, D., Christou, P., Snape, J.W., Gale, M.D., Harberd, N.P. 1999. 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* **400:** 256-261

Petroli, C.D., Sansaloni, C.P., Carling, J., Steane, D.A., Vaillancourt, R.E., Myburg, A.A., da Silva, O.B.J., Georgios, J.P.J., Kilian, A., Grattapaglia, D. 2012. Genomic Characterization of DArT Markers Based on High-Density Linkage Analysis and Physical Mapping to the Eucalyptus Genome. *PLoS ONE* **7(9):** e44684. doi:10.1371/journal.pone.0044684

Pevsner, J. 2009. *Bioinformatics and functional genomics (2nd edition).* New Jersey: Wiley Blackwell, pp. 312-370

Potokina, E., Druka, A., Luo, Z.W., Wise, R., Waugh, R. and Kearsey, M.J. 2007. eQTL analysis of 16,000 barley genes reveals a complex pattern of genome wide transcriptional regulation. *The Plant Journal* **53:** 90–101.

Potokina, E., Druka, A., Luo, Z., Waugh, R., Kearsey, M.J. 2009. Transcriptome analysis of barley (*Hordeum Vulgare* L.) using the Affymetrix Barley1 Genechip. *Russian Journal of Genetics* **45:** 81–92.

Pottorff, M., Ehlers, J.D., Fatokun, C., Roberts, P.A., Close, T.J. 2012. Leaf morphology in Cowpea [*Vigna unguiculata* (L.) Walp]: QTL analysis, physical mapping and identifying a candidate gene using synteny with model legume species. *Genomics* **13:** 234 – 245

Quarrie, S.A., Lazić-Jančić, V., Kovaćević, D., Steed, A., Pekić, S. 1999. Bulk segregant analysis with molecular markers and its use for improving drought tolerance in maize. *Journal of Experimental Botany* **50:** 1299-1306

Quesada, V., Garcia-Martinez, S., Piqueras, P., Ponce, M.R., Micol, J.L. 2002. Genetic architecture of NaCl tolerance in *Arabidopsis. Plant Physiology* **120:** 951-963

Reinecke, D.M., Wickramarathna, A.D., Ozga, J.A., Kurepin, L.V., Jin, A.L., Good, A.G., Pharis, R.P. 2013. *Gibberellin 3-oxidase* gene expression patterns influence gibberellin biosynthesis, growth and development in pea. *Plant Physiology* **163:** 929-945

Rinaudo, J.A.S. and Gerin, J.L. 2004. Cross-species hybridization: Characterization of gene expression in Woodchuck liver using human membrane arrays. *Journal of Medical Virology* **74:** 300-313

Risch, N. 1992. Genetic linkage: Interpreting LOD scores. *Science* **255:** 803-804

Rise, M.L., von Schalburg, K.R., Brown, G.D., Mawer, M.A., Devlin, R.H., Kuipers, N., Busby, M., Beetz-Sargent, M., Alberto, R., Gibbs, A.R. Hunt, P., Shukin, R., Zeznik, J.A., Nelson, C., Jones, S.R., Smailus, D.E., Jones, S.J., Schein, J.E., Marra, M.A., Butterfield, Y.S., Stott, J.M., Ng, S.H., Davidson, W.S., Koop, B.F. 2004. Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Research*. **14**: 478-490

Ronald, J., Akey, J.M., Whittle, J., Smith, E.N., Yvert, G., Kruglyak, L. 2005. Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Reserach.* **15:** 284–291

Rostoks, N., Mudie, S., Cardle, L., Russell, J., Ramsay, L., Booth, A., Svensson, J.T., Wanamaker, S.I., Walia, H. Rodriguez, E. M. 2005. Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. *Molecular Genetics and Genomics* **274:** 515–527

Rupani, P.F., Singh, R.P., Ibrahim, M.H., Esa, N. 2010. Review of the current palm oil mill effluent (POME) treatment methods: Vermicomposting as a sustainable practice. *World Applied Sciences Journal* **10:** 1190 – 1201

Saeed, O.M.B., Sankaran, S., Shariff, A.R.M., Shafri, H.Z.M., Ehsani, R., Alfani, M.S. and Hazir, M.H.M. 2012. Classification of oil palm fresh fruit bunches on their maturity using portable four-band sensor system. *Computers and Electronics in Agriculture* **82:** 55–60

Sakamoto, T., Morinka, Y., Ohnishi, T., Sunohara, H., Fujioka, S., Ueguchi-Tanaka, M., Mizutani, M., Sakata, K., Takatsuto, S., Yoshida, S., Tanaka, H., Kitano, H., Matsuoko, M. 2006. Erect leaves caused by brassinosteroid deficiency increase biomass production and grain yield in rice. *Nature Biotechnology* **24:** 105–109

Salentijn, E.M.J., Pereira, A., Angenent, G.C., van der Linden, C.G., Krens, F., Smulders, M.J.M., Vosman, B. 2007. Plant translational genome: from model species to crops. *Molecular Breeding* **20:** 1-13

Sambanthamurthi, R., Sundram, K., Tan, Y.A. 2000. Chemistry and biochemistry of palm oil. *Progress in Lipid Research* **39:** 507-558

Sambrook, J. and Russell, D.W. 2001. *Molecular cloning: a laboratory manual (3rd edition)*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press

Sanderon, M.J., Thorne, J.L., Wikstrom, N., Bremer, K. 2004. Molecular evidence on plant divergence times. *American Journal of Botany* **9:** 1656-1665

Sansaloni, C., Petroli, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., Killian, A. 2011. Diversity Array Technology (DArT) and next-generation sequencing combined: Genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus. BMC Proceedings* **5:** 54-55

Sax, K. 1923. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics 8*: 552–560

Schmidt, R., West, J., Love, K., Lenehan, Z., Lister, C., Thompson, H., Bouchez, D., Dean, C. 1995. Physical map and organisation of *Arabidopsis thaliana* chromosome 4. *Science* **270:** 480-483

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, GD., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D,. Stacey, G., Shoemaker, R.C., Jackson, S.A. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463:** 178-183

Schneider, K. 2005. Mapping populations and principles of genetic mapping. In: *The handbook of plant genome mapping- genetic and physical mapping* (Ed. by Meksem, K. and Kahl, G). Weinheim: WILEY

DiGuistini, S., Liao, N.Y., Platt, D., Robertson, G., Seidel, M., Chan, S.K., Docking, T.R., Birol, I., Holt, R.A., Hirst, M., Mardis, E., Marra, M.A., Hamelin, R.C., Bohlmann, J., Breuil, C., Jones, S.J.M. **2009.** *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology* 10: doi:10.1186/gb-2009-10-9-r94

Semagn, K., Bjornstad, A., Ndjiondjop, M.N. 2006. An overview of molecular marker methods for plants. *African Journal of Biotechnology* **5:** 2540-2568

Seng, T.Y., Mohamed Saad, S.H., Chin, C.W., Ting, N.C., Harminder Singh, R.S., Qamaruz Zaman, F., Tan. S.G., Rabiah Syed Alwee, S.S. 2011. Genetic Linkage Map of a High Yielding FELDA Deli×Yangambi Oil Palm Cross. *PLoS ONE* **6**: e26593. doi:10.1371/journal.pone.0026593

Seymour, D.K., Filiault, D.L, Henry, I.M., Monson-Miller, J., Ravi, M., Pang, A., Comai, L., Chan, S.W.L., Maloof, J.N. 2012. Rapid creation of *Arabidopsis* doubled haploid lines for quantitative trait locus mapping. *Genetics* **109:** 4227-4232

Shi, C., Uzarowska, A., Ouzunova, M., Landbeck, M., Wenzel, G., Lübberstedt, T. 2007. Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint x Flint maize recombinant inbred line population. *BMC Genomics 8*, 22.

Shoemaker, R.C., Schlueter, J., Doyle, J.J. 2006. Paleopolyploidy and gene duplication in soybean and other legumes. *Current Opinion in Plant Biology* **9:** 104–109

Singh, P. 1991. Influence of water deficits on phenology, growth and dry matter allocation in chickpea (*Cicer arietinum*). *Field Crops Research* **28:** 1–15

Singh, R., Nagappan, J., Tan, S.G., Panandam, J.M., Cheah, S.C. 2007. 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 and Biotechnology* **15:** 121-131

Singh, R., Ong-Abdullah, M., Low, E.T., Manaf, M.A., Rosli, R., Nookiah, R., Ooi, L.C., Ooi, S.E., Chan, K.L., Halim, M.A., Azizi, N., Nagappan, J., Bacher, B., Lakey, N., Smith, S.W., He, D., Hogan, M., Budiman, M.A., Lee, E.K., DeSalle, R., Kudrna, D., Goicoechea, J.L., Wing, R.A., Wilson, R.K., Fulton, R.S., Ordway, J.M., Martienssen, R.A., Sambanthamurthi, R. 2013. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature* **500:** 335-339

Singh, S.P., Teran, H., Gutierrez, J.A. 2001. Registration of SEA5 and SEA13 drought tolerant dry bean germplasm. *Crop Sciences* **41:** 276-277

Soh, A.C., Wong, C.K., Ho, Y.W., Choong, C.W. 2009. Oil palm. In: *Oil Crops* (Ed. by Vollmann, J. and Rajcan, I.). New York: Springer, pp. 333-368

Soh, A.C., Wong, G., Hor, T.Y., Tan, C.C., Chew, P.S. 2010. Oil palm Genetic Improvement. In: *Plant Breeding Review* (Ed. by Janick, J.). New York: John Wiley & Sons, pp. 165-21

Sohail, Q., Shehza, T., Kilian, A., Eltayeb, A.E., Tanaka, H., Tsujimoto, H. 2012. Development of diversity array technology (DArT) markers for assessment of population structure and diversity in *Aegilops tauschii*. *Breeding Science* **62:** 38–45

Somal, T.L.C. and Yapa, P.A.J. 1998. Accumulation of proline in cowpea under nutrient, drought and saline stresses. *Journal of Plant Nutrition* **21:** 2465 - 2473

Song, Q.J., Marek, L.F., Shoemaker, R.C., Lark, K.G., Concibido, V.C., Delannay, X., Specht, J.E., Cregan, P.B., 2004. A new integrated genetic linkage map of the soybean. *Theoretical and Applied Genetics* **109(1):**122-128.

Stadler, F. 2009. Analysis of differential gene expression under water-deficit stress and genetic diversity in bambara groundnut (*Vigna subterranea* (L.) *Verdc*.) using novel high-throughput technologies. *PhD Thesis.* Technical University of Munich, Germany.

Stam, P. 1993. Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *Plant Journal* **3:** 739-744

Subbarao, G.V., Chauhan, Y.S., Johansen, C. 2000. Patterns of osmotic adjustments in pigeonpea – its importance as a mechanism of drought resistance. *European Journal of Agronomy* **12:** 239-249

Sundram, K., Sambanthamurthi, R., Tan, Y.A. 2003. Palm fruit chemistry and nutrition. *Asia Pacific Journal of Clinical Nutrition* **12:** 355-362

Surzycki, S. 2000. *Basic techniques in molecular biology*. New York: Springer, pp. 374-380

Swaminathan, K., Chae, W.B., Mitros, T., Varala, K., Xie, L., Barling, A., Glowacka, K., Hall, M., Jezowsky, S., Ming, R., Hudson, M., Juvik, J.A., Rokhsar, D.S., Moose, S.P. 2012. A framework genetic map for *Miscanthus sinensis* from RNA-seq-based markers shows recent tetraploidy. *BMC Genomics* **13:** 1-17

Szilagyi, L. 2003. Influence of drought on seed yield components in common bean. *Bulgarian Journal Plant Physiology* **Special Issue**: 320-330

Tambussi, E.A., Bort, J., Araus, J.L. 2007. Water use efficiency in C_3 cereals under Mediterranean conditions: a review of physiological aspects. *Annals of Applied Biology* **150:** 307–321. doi: 10.1111/j.1744-7348.2007.00143.x

Tanksley, S.D. 1993. Mapping polygenes. *Annual Review of Genetics* **27:** 205-233

Tantasawat, P., Trongchuen, J., Prajongjai, T., Jenweerawat, S., Chaowiset, W. 2011. SSR analysis of soybean (*Glycine max* (L.) Merr.) genetic relationship and variety identification in Thailand. *Australian Journal of Crop Science* **5:** 283-290

Teoh, C.H. 2010. Key Sustainability Issues in the Palm Oil Sector – A Discussion Paper for Multi-Stakeholder Consultations. *World Bank Discussion Paper,Washington.*http://www.ifc.org/ifcext/agriconsultation.nsf/AttachmentsByTitle/Discussion+Paper/$FILE/Discussion+Paper_FINAL.pdf(accessed 14/08/2011)

Teran, H. and Singh, S.P. 2002. Comparison of sources and lines selected for drought resistance in common bean. *Crop Sciences* **42:** 64-70

The Arabidopsis genome initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature.* **408:** 796-815

Thung, M. and Rao, I.M. 1999. Integrated management of abiotic stresses. In: Singh SP (ed) Common bean improvement in the twenty-first century. Kluwer Academic Publishers, Dordrecht, pp. 331-370

Turner, N.C. 1979. Drought resistance and adaptations to water deficits in crop plants. In: *Musell H. Staple RC* (Ed. by*Stress physiology in crop plants.* New York: Wiley-Interscience, 343-373

Ulloa, M., Meredith, W.R.J., Shappley, Z.W., Kahler, A.L. 2002. RFLP genetic linkage maps from four F (2.3) populations and a joinmap of *Gossypium hirsutum L*. *Theoretical and Applied Genetics* **104:** 200-208

Uthaipaisanwong, P., Chanprasert, J., Shearman, J.R., Sangsrakru, D., Yoocha, T., Jomchai, N., Jantasuriyarat, C., Tragoonrung, S., Tangphatsornruang, S. 2012. Characterization of the chloroplast genome sequence of oil palm (*Elaeis guineensis* Jacq.). *Gene* **500:** 172-180

Van Ooijen, J.W. 2006. JoinMap ® 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B.V., Wageningen, Netherlands.

Van Ooijen, J.W. 2009. MapQTL®6, Software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma B.V., Wageningen, Netherlands.

Van Tassell, C.P., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., Sonstegard, T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* **5(3):** 247-252

Vinod, K.K. 2009. Genetic mapping of quantitative trait loci and marker-assisted selection in plantation crops. In: *Proceedings of the training programme on 'In vitro Techniques in Plantation Crops':* Central Plantation Crops Research Institute, Kasaragod, India. pp. 111-132

Vision, T.J., Brown, D.G., Shmoys, D.B., Durrett, R.T., Tanksley, S.D. 2000. Selective mapping: A strategy for optimizing the construction of high density linkage maps. *Genetics* **155:** 407 – 420.

Vodkin, L., Jones, S., Gonzales, O.D., Thibaud-Nissen, F., Tutega, Z.G, 2008. Genomics of soybean seed development. In: *Genetics and Genomics of Soybean* (Ed: Stacey,G). New York: Springer, pp. 163–184

VSN International. 2012. *Genstat for Windows 15th Edition*. VSN International, Hemel Hempstead, UK. Web page: Genstat.co.uk

Vurayai, R., Emongor, V., Moseki, B. 2011. Physiological responses of bambara groundnut (*Vigna subterranea* (L.) *Verdc*) to short periods of water stress during different development stages. *Asian Journal of Agricultural Sciences* **3:** 37-43

Vuylsteke, M., Van Den Daele, H., Vercauteren, A., Zabeau, M., Kuiper, M. 2006. Genetic dissection of transcriptional regulation by cDNA-AFLP. *Plant Journal.* **45:** 439-446.

Wang, J., Lydiate, D.J., Parkin, I.A.P., Falentin, C., Delourme, R., Carion, P.W.C., King, G.J. 2011. Integration of linkage maps for the Amphydiploid *Brassica napus* and comparative mapping with *Arabidopsis* and *Brassica rapa. BMC Genomics* **12:** 10 -121

Wang, N., Fang, L., Xin, H., Wang, L., Li, S.H. 2012. Construction of a high-density genetic map for grape using next generation restriction-site associated DNA sequencing. *BMC Plant Biology* **12:** 1-15

Wanlayaporn, K., Authrapun, J., Vanavichit, A., Tragoonrung, S. 2013. QTL mapping for partial resistance to southern corn rust using RILs of tropical sweet corn. *American Journal of Plant Sciences* **4:** 878-889

Weber, A.P.M., Weber, K.L., Wilkerson, C., Ohlrogge, J.B. 2007. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiology* **144:** 32-42

Wenzl, P., Li, H.B., Carling, J., Zhou, M.X., Raman, H., Paul, E., Hearnden, P., Maier, C., Xia, L., Caig, V., Oversna, J., Cakir, M., Poulsen, D., Wang, J.P., Raman, R., Smith, K.P., Muehlbauer, G.J., Chalmes, K.J., Kleinhofs, A., Huttner, E., Killian, A. 2006. A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. *BMC Genomics* **7:** 206 – 228

West, M.A.L., van Leeuwen, H., Kozik, A., Kliebenstein, D.J., Doerge, R.W., St Clair, D.A., Michelmore, R.W. 2006. High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. *Genome Research* **16:** 787–795.

Wicker, T. and Keller, B. 2007. Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Research* **17:** 1072-1081

Wikstrom, N., Savolainen, V., Chase, M.W. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society of London, series B* **268:** 2211-2220

Willis, L.B., Lessard, P.A., Parker, J.A., O'Brien, X.M., Sinskey, A.J. 2008. Functional annotation of oil palm genes using an automated bioinformatics approach. *Journal of Oil Palm Research* **(Special Issue – April 2008):** 35-43

Winzeler, E.A., Richards, D.R., Conway, A.R., Goldstein, A.L., Kalman, S., McCullough, M.J., McCusker, J.H., Stevens, D.A., Wodicka, L., Lockhart, D.J., Davis, R.W. 1998. Direct allelic variation scanning of the yeast genome. *Science* **281:** 1194–1197

Wittenberg, A.H.J., van der Lee, T., Cayla, C., Kilian, A., Visser, R.G.F., Schouten, H.J. 2005. Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*. *Molecular Genetics and Genomics* **274:** 30-39.

Wu, C., Carta, R., Zhang, L. 2005. Sequence dependence of cross-hybridisation on short oligo microarrays. *Nucleic Acids Research* **33:** 1-7

Wu, R.L., Han, Y.F., Hu, J.J., Fang, J.J., Li, L., Li, M.L., Xeng, Z.B. 2000. An integrated genetic map of *Populus deltoides* based on amplified fragment length polymorphisms. *Theorical and Applied Genetics* **100:** 1249-1256

Wu, Y.Q. and Huang, Y. 2007. An SSR genetic map of Sorghum bicolor (L.) Moench and its comparison to a published genetic map. *Genome* **50:** 84-89

Xiao, J., Li, J., Yuan, L., Tanksley, S.D. 1995. Dominance in the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers. *Genetics* **140:** 745-754

Xu, Z.H. and Zhou, G.S. 2008. Reponses of leaf stomatal density to water status and its relationship with photosynthesis in a grass. *Journal of Experimental Botany* **59:** 3317-3325

Yan, L., Loukoianov, A., Blechl, A., Tranquilli, G., Ramakrishna, W., SanMiguel, P., Bennetzen, J.L., Echenique, V., Dubcovsky, J. 2004. The wheat*VRN2* gene is a flowering repressor down-regulated by vernalization. *Science* **303:** 1640–1644

Yandell, M. and Ence, D. 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* **13:** 329-342

Yang, M., Zhang, X., Liu, G., Yin, Y., Chen, K., Yun, Q., Zhao, D., Al-Mssallem, I.S., Yu, J. 2010. The Complete Chloroplast Genome Sequence of Date Palm (*Phoenix dactylifera* L.). *PLoS One* **5:** 1-14

Yang, S. Y., Saxena, R. K., Kulwal, P. L., Ash, G. J., Dubey, A., Harper, J. D. I., Upadhyaya, H. D., Gothalwal, R., Kilian, A., Varshney, R. K. 2011. The first genetic map of pigeon pea based on diversity arrays technology (DArT) markers. *Journal of Genetics.* **90:** 103–109

Yang, S., Gao, M., Xu, C., Gao, J., Deshpande, S., Lin, S., Roe, B.A., Zhu, H. 2008. Alfalfa benefits from *Medicago truncatula*: the RCT1 gene from *M. truncatula* confers broad-spectrum resistance to anthracnose in alfalfa. *Proceedings of the National Academy of Sciences* **105:** 12164–12169

Yee, J.C., Wlaschin, K.F., Chuah, S.H., Nissom, P.M., Hu, W.S. 2008. Quality Assessment of Cross-Species Hybridization of CHO Transcriptome on a Mouse DNA Oligo Microarray. *Biotechnology and Bioengineering* **101:** 1359-136

Yordanov, I., Velikova, V., Tsonev, T. 2003. Plant responses to drought and stress tolerance. *Bulgarian Journal Plant Physiology* **Special Issue**: 187-206

Yu, J., Hu, S., Wang, J., Wong, G.K.S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296:** 79–92

Yu, Y., Yuan, D.J., Liang, S.G., Li, X., Wang, X.Q., Lin, Z.X., Zhang, X.L. 2011. Genome structure of cotton revealed by a genome-wide SSR genetic map constructed from a BC1 population between *Gossypium hirsutum* and *G. barbadense. BMC Genomics* **12:** 15 - 29

Zhang, L., Wang, S., Li, H., Deng, Q., Zheng, A., Li, S., Li, P., Li, Z., Wang, J. 2010. Effects of missing marker and segregation distortion on QTL mapping in F2 populations. *Theoretical and Applied Genetics* **121:** 1071-1082

Zhang, W.K., Wang, Y.J., Luo, G.Z., Zhang, J.S., He, C.Y., Wu, X.L., Gai, J.Y., Chen, S.Y. 2004. QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. *Theoretical and Applied Genetics* **108:** 1131-1139

Zhao, W.G., Chung, J.W., Lee, G.A., Ma, K.H., Kim, H.H., Chung, I.M., Kang, H.K., Kim, T.S., Lee, J.K., Kim, N.S., Park, Y.J. 2011. Molecular genetic diversity and population structure of a selected core set in garlic (*Allium sativum* L.) using novel SSR markers. *Plant breeding* **130:** 46-54

Appendix 1- Lists of oil palm plant materials for DNA fingerprinting and XSpecies analysis.

(A) Oil palm leaf samples from Paloh Estate, Johor, Malaysia derived from *Tenera* self-crossing.

	751 *dura*		751 *pisifera*		768 *dura*		768 *pisifera*		896 *dura*		896 *pisifera*	
						Sample Name						
1	751/7	(D19)	751/26	(P17)	768/49	(D9)	768/46	(P7)	896/27	(D1)	896/14	(P1)
2	751/8	(D20)	751/27	(P18)	768/44	(D10)	768/45	(P8)	896/42	(D2)	896/48	(P2)
3	751/22	(D21)	751/29	(P19)	768/35	(D11)	768/52	(P9)	896/3	(D3)	896/38	(P3)
4	751/25	(D22)	751/30	(P20)	768/42	(D12)	768/50	(P10)	896/25	(D4)	896/51	(P4)
5	751/28	(D23)	751/31	(P21)	768/57	(D13)	768/43	(P11)	896/34	(D5)	896/20	(P5)
6	751/39	(D24)	751/34	(P22)	768/41	(D14)	768/34	(P12)	896/4	(D6)	896/44	(P6)
7	751/40	(D25)	751/43	(P23)	768/56	(D15)	768/59	(P13)	896/10	(D7)	-	
8	751/42	(D26)	751/44	(P24)	768/60	(D16)	768/58	(P14)	892/18	(D8)	-	
9	751/45	(D27)	751/48	(P25)	768/31	(D17)	768/32	(P15)	-		-	
10	751/46	(D28)	751/49	(P26)	768/28	(D18)	768/51	(P16)	-		-	

(B) Oil palm DNA samples provided directly by Applied Agricultural Resources Sdn. Bhd (AAR), Malaysia.

	769 *dura*		769 *pisifera*		Parent *Tenera*	
			Sample Name			
1	769/B/35	(D106)	769/B/40	(P104)	F1 150/07 (PAR 896)	(B1)
2	769/B/36	(D109)	769/B/44	(P103)	F1 228/05 (PAR 768)	(B2)
3	769/B/39	(D107)	769/B/52	(P109)	F1 228/06 (PAR 769)	(B3)
4	769/B/43	(D108)	769/B/53	(P105)	138/04 (PAR 751)	(B4)
5	769/B/49	(D101)	769/B/54	(P107)	-	
6	769/B/55	(D105)	769/B/57	(P106)	-	
7	769/A/8	(D103)	769/A/1	(P102)	-	
8	769/A/12	(D104)	769/A/19	(P108)	-	
9	769/A/24	(D110)	769/A/21	(P110)	-	
10	769/A/23	(D102)	769/A/27	(P101)	-	

Appendix 2- DNA fingerprinting of oil palm using 12 SSR primers.

(a) DNA fingerprinting of *dura* 768 (A1-C2)

		SAMPLE NAME	Overall	OP1		OP5		OP13		OP12		OP11		OP24/6	
A1	**F1 228/05**	**D2B2par**		**213/219**		**236**		**314**		**192/200/206**		**240/253**		**177/181/190**	
B1	**768/49(D)**	D9	✓	219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181/190	✓
C1	**768/44(D)**	D10	✓	219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181/190	✓
D1	**768/35(D)**	D11	✓	213/219	✓	236	✓	314	✓	192/(200)/206	✓	(240)/253	✓	177/181/190	✓
E1	**768/42(D)**	D12	✓	213/219	✓	236	✓	314	✓	192/(200)/206	✓	240/253	✓	177/181	✓
F1	**768/57(D)**	D13	✓	219	✓	236	✓	314	✓	192/(200)/206	✓	253	✓	177/181	✓
G1	**768/41(D)**	D14	✓	213/219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181/190	✓
H1	**768/56(D)**	D15	✓	219	✓	236	✓	314	✓	192/(200)/206	✓	253	✓	181/190	✓
A2	**768/60(D)**	D16	✓	213	✓	236	✓	314	✓	192/(200)/206	✓	(240)/253	✓	177/181	✓
B2	**768/31(D)**	D17	✓	219	✓	236	✓	314	✓	192/(200)/206	✓	240/253	✓	177/181/190	✓
C2	**768/28(D)**	D18	✗ (8)	203	✗	236/257	✗	322	✗	192	✓	253/258	✗	181/190	✓

		SAMPLE NAME		OP2		OP7		OP20		OP18		OP21		OP29	
A1	**F1 228/05**	**D2B2par**		**164**		**240**		**225/246**		**292/303**		**225**		**116/122**	
B1	**768/49(D)**	D9		164	✓	240	✓	225/246	✓	303	✓	225	✓	116	✓
C1	**768/44(D)**	D10		164	✓	240	✓	246	✓	292	✓	225	✓	116/122	✓
D1	**768/35(D)**	D11		164	✓	240	✓	246	✓	303	✓	225	✓	122	✓
E1	**768/42(D)**	D12		164	✓	240	✓	NA*		292	✓	225	✓	116/122	✓
F1	**768/57(D)**	D13		164	✓	240	✓	NA		292	✓	225	✓	122	✓
G1	**768/41(D)**	D14		164	✓	240	✓	NA		292	✓	225	✓	116/122	✓
H1	**768/56(D)**	D15		164	✓	NA		NA		292	✓	225	✓	122	✓
A2	**768/60(D)**	D16		164	✓	240	✓	225	✓	292/303		225	✓	122	✓
B2	**768/31(D)**	D17		164	✓	240	✓			292		225	✓	116/122	✓
C2	**768/28(D)**	D18		158/166	✗	NA		222	✗	294		225	✓	133	✗

(b) DNA fingerprinting of *pisifera* 768 (D2-F3).

		SAMPLE NAME	Overall	OP1		OP5		OP13		OP12		OP11		OP24/6	
D2	F1 228/05	D2B2(2)par		213/219		236		314		192/200/206		240/253		177/181/190	
E2	768/46(P)	P7	✓	219	✓	236	✓	314		192/(200)/206	✓	253	✓	181/190	✓
F2	768/45(D)	P8	✓	219	✓	236	✓	314	✓	192/199	✓	253	✓	177/181	✓
G2	768/52(P)	P9	✓	219	✓	236	✓	314	✓	192/200/206	✓	240/253	✓	177/181	✓
H2	768/50(P)	P10	✓	219	✓	236	✓	314	✓	192/(200)/206	✓	240/253	✓	177/181/190	✓
A3	768/43(P)	P11	✓	213	✓	236	✓	314	✓	192/(200)/206	✓	240/253	✓	177/(181)	✓
B3	768/34(P)	P12	✓	213/219	✓	236	✓	314		192/200	✓	253	✓	181/190	✓
C3	768/59(P)	P13	✓	213/219	✓	236	✓	314	✓	192/(200)/206	✓	253	✓	177/181/190	✓
D3	768/58(P)	P14	✓	213/219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181	✓
E3	768/32(P)	P15	✓	213/219	✓	236	✓	314	✓	192/206	✓	253	✓	181/190	✓
F3	768/51(P)	P16	✓	213/219	✓	236	✓	314	✓	192/200	✓	253	✓	177/181	✓

		SAMPLE NAME		OP2		OP7		OP20		OP18		OP21		OP29	
D2	F1 228/05	D2B2(2)par		164		240		225/246		292/303	✓	225		116/122	
E2	768/46(P)	P7		164	✓	240	✓	NA		292/303	✓	225	✓	122	✓
F2	768/45(D)	P8		164	✓	240	✓	NA		303	✓	225	✓	122	✓
G2	768/52(P)	P9		164	✓	240	✓	NA		303	✓	225	✓	116	✓
H2	768/50(P)	P10		164	✓	NA		NA		303	✓	225	✓	116	✓
A3	768/43(P)	P11		164	✓	240	✓	225	✓	292/303	✓	225	✓	116	✓
B3	768/34(P)	P12		164	✓	240	✓	225/246	✓	292/303	✓	225	✓	116/122	✓
C3	768/59(P)	P13		164	✓	240	✓	246	✓	292/303	✓	225	✓	122	✓
D3	768/58(P)	P14		164	✓	240	✓	225/246	✓	292/303	✓	225	✓	116	✓
E3	768/32(P)	P15		164	✓	240	✓	NA		292/303	✓	225	✓	122	✓
F3	768/51(P)	P16		164	✓	240	✓	NA		292	✓	225	✓	122	✓

(c) DNA fingerprinting of *dura* 769 (G3-A5).

		SAMPLE NAME	Overall	OP1		OP5		OP13		OP12		OP11		OP24/6	
G3	*F1 228/06*	*P101B3Par*		*205/219*		*236/(257?)*		*314/320*		*184/192/206*		*253*		*177/181/190*	
H3	**769/B/49(D)**	D101	✔	205	✓	236	✓	320	✓	192/206	✓	253	✓	177/181	✓
A4	**769/A/23(D)**	D102	✔	205	✓	236	✓	314/320	✓	184/206	✓	253	✓	190	✓
B4	**769/A/8(D)**	D103	✔	205/219	✓	236/257	✓	314/320	✓	184/192/206	✓	253	✓	177/181/190	✓
C4	**769/A/12(D)**	D104	✔	205/219	✓	236	✓	314/320	✓	184/192/206	✓	253	✓	181/190	✓
D4	**769/B/55(D)**	D105	✔	205/219	✓	236/257	✓	320	✓	184/192/206	✓	253	✓	177/181/190	✓
E4	**769/B/35(D)**	D106	✔	205	✓	236	✓	320	✓	184/206	✓	253	✓	177/181	✓
F4	**769/B/39(D)**	D107	✔	205	✓	236	✓	314	✓	184/192/206	✓	253	✓	177/181/190	✓
G4	**769/B/43(D)**	D108	✔	205/219	✓	236	✓	314	✓	192/206	✓	253	✓	177/181/190	✓
H4	**769/A/36(D)**	D109	✘ (3)	205/(217)	✗	236/257	✓	314	✓	192/(200)/206	✗	253	✓	177/181	✓
A5	**769/A/24(D)**	D110	✔	205/219	✓	236	✓	314	✓	184/206	✓	253	✓	177/181/190	✓

		SAMPLE NAME		OP2		OP7		OP20		OP18		OP21		OP29	
G3	*F1 228/06*	*P101B3Par*		*152/154*		*242/246*		*NA*		*292/294*		*225/240*		*122/134*	
H3	**769/B/49(D)**	D101		154	✓	242	✓			294		225/240	✓	122/134	✓
A4	**769/A/23(D)**	D102		152/154	✓	242	✓	225	✓	292	✓	225/240	✓	122/134	✓
B4	**769/A/8(D)**	D103		154	✓	242/246	✓	225	✓	292/294	✓	22/240	✓	134	✓
C4	**769/A/12(D)**	D104		154	✓	246	✓			292	✓	225	✓	122/134	✓
D4	**769/B/55(D)**	D105		154	✓	242/246	✓	225	✓	292/294	✓	225	✓	134	✓
E4	**769/B/35(D)**	D106		152	✓	242/246	✓	NA		294		225/240	✓	122/134	✓
F4	**769/B/39(D)**	D107		152	✓	242	✓	NA		292/294		225/240	✓	134	✓
G4	**769/B/43(D)**	D108		152/154	✓	242	✓	NA		292		225/240	✓	122/134	✓
H4	**769/A/36(D)**	D109		160/164	✗	240/246	✓	NA		292/294	✓	240	✓	122/134	✓
A5	**769/A/24(D)**	D110		152/154	✓	242/246	✓	225	✓	292	✓	225/240	✓	134	✓

(d) DNA fingerprinting of *pisifera* 769 (B5-D6).

		SAMPLE NAME	Overall	OP1		OP5		OP13		OP12		OP11		OP24/6	
B5	*F1 228/06*	*P101B3(2)Par*		*205/219*		*236/(257?)*		*314/320*		*184/192/206*		*253*		*177/181/190*	
C5	769/A/27(P)	P101	✓	205/219	✓	236	✓	320	✓	184/192/206	✓	253	✓	177/181/190	✓
D5	769/A/1(P)	P102	✓	205/219	✓	236	✓	NA		184/192/206	✓	253	✓	177/181/190	✓
E5	769/B/44(P)	P103	✓	205/219	✓	236	✓	314	✓	192/206	✓	253	✓	177/181/190	✓
F5	769/B/40(P)	P104	✓	219	✓	236	✓	314/320	✓	192/206	✓	253	✓	181/190	✓
G5	769/B/53(P)	P105	✓	205/219	✓	236/257	✓	314/320	✓	184/192/206	✓	253	✓	177/181	✓
H5	769/B/57(P)	P106	✓	205/219	✓	236	✓	320	✓	184/206	✓	253	✓	177/181/190	✓
A6	769/B/54(P)	P107	✓	205/219	✓	236	✓	314/320	✓	184/192/206	✓	253	✓	177/181/190	✓
B6	769/A/19(P)	P108	✓	219	✓	236/257	✓	314	✓	184/206	✓	253	✓	181/190	✓
C6	769/B/52(P)	P109	✓	205	✓	236	✓	NA		184/192/206	✓	253	✓	177/181/190	✓
D6	769/A/21(P)	P110	✓	NA		236/257	✓	314	✓	184/192/206	✓	253	✓	177/190	✓

		SAMPLE NAME		OP2		OP7		OP20		OP18		OP21		OP29	
B5	*F1 228/06*	*P101B3(2)Par*		*152/154*		*242/246*		*225*		*292/294*		*225/240*		*122/134*	
C5	769/A/27(P)	P101		?		246	✓	225	✓	292	✓	225/240	✓	122	✓
D5	769/A/1(P)	P102		154	✓	242/246	✓			292/294	✓	225	✓	134	✓
E5	769/B/44(P)	P103		152/154	✓	242/246	✓	NA		292	✓	225/240	✓	122/134	✓
F5	769/B/40(P)	P104		152/154	✓	246	✓	NA		292/294	✓	240	✓	134	✓
G5	769/B/53(P)	P105		152	✓	246	✓	NA		294	✓	225/240	✓	134	✓
H5	769/B/57(P)	P106		152	✓	246	✓	NA		294	✓			134	✓
A6	769/B/54(P)	P107		154	✓	242/246	✓	225	✓	292/294	✓	240	✓	122/134	✓
B6	769/A/19(P)	P108		152	✓	246	✓	225	✓	292/294	✓	NA		122/134	✓
C6	769/B/52(P)	P109		152	✓	242/246	✓	NA		294	✓	225/240	✓	122/134	✓
D6	769/A/21(P)	P110		152	✓	246	✓	225	✓	294	✓	225/240	✓	134	✓

(e) DNA fingerprinting of *dura* 751 (E6-G7).

		SAMPLE NAME	Overall	OP1		OP5		OP13		OP12		OP11		OP24/6	
E6	*138/04*	*P107B4a*		*217*		*257*		*314*		*206*		*253/258*		*183*	
F6	751/7(D)	D19	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
G6	751/8(D)	D20	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
H6	751/22(D)	D21	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
A7	751/25(D)	D22	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
B7	751/28(D)	D23	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
C7	751/39(D)	D24	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
D7	751/40(D)	D25	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
E7	751/42(D)	D26	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
F7	751/45(D)	D27	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
G7	751/46(D)	D28	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓

		SAMPLE NAME		OP2		OP7		OP20		OP18		OP21		OP29	
E6	*138/04*	*P107B4a*		*154/160*		*240*		*NA*		*294*		*225/240*		*122*	
F6	751/7(D)	D19		160	✓	240	✓	NA		294	✓	225/240	✓	122	✓
G6	751/8(D)	D20		154/160	✓	240	✓	NA		294	✓	225/240	✓	122	✓
H6	751/22(D)	D21		160	✓	240	✓	NA		294	✓	240	✓	122	✓
A7	751/25(D)	D22		154	✓	240	✓	240		294	✓	225/240	✓	122	✓
B7	751/28(D)	D23		154	✓	240	✓	240		294	✓	240	✓	122	✓
C7	751/39(D)	D24		154/160	✓	240	✓	240		294	✓	240	✓	122	✓
D7	751/40(D)	D25		160	✓	240	✓	NA		294	✓	225/240	✓	122	✓
E7	751/42(D)	D26		154	✓	240	✓	NA		294	✓	225/240	✓	122	✓
F7	751/45(D)	D27		154/160	✓	240	✓	NA		294	✓	240	✓	122	✓
G7	751/46(D)	D28		154	✓	240	✓	NA		294	✓	225/240	✓	122	✓

(f) DNA fingerprinting of *pisifera* 751 (H7-B9).

		SAMPLE NAME	Overall	OP1		OP5		OP13		OP12		OP11		OP24/6	
H7	*138/04*	*P107B4b*		*217*		*257*		*314*		*206*		*253/258*		*183*	
A8	751/26(P)	P17	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
B8	751/27(P)	P18	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
C8	751/29(P)	P19	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
D8	751/30(P)	P20	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
E8	751/31(P)	P21	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
F8	751/34(P)	P22	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
G8	751/43(P)	P23	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
H8	751/44(P)	P24	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓
A9	751/48(P)	P25	✗ (6)	217	✓	236/257	✗	314	✓	192/206	✗	253		183/190	✗
B9	751/49(P)	P26	✔	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓

		SAMPLE NAME		OP2		OP7		OP20		OP18		OP21		OP29	
H7	*138/04*	*P107B4b*		*154/160*		*240*		*NA*		*294*		*225/240*		*122*	✓
A8	751/26(P)	P17		154/160	✓	240	✓	240		294	✓	225	✓	122	✓
B8	751/27(P)	P18		154	✓	240	✓	240		294	✓	225/240	✓	122	✓
C8	751/29(P)	P19		154/160	✓	240	✓	240		294	✓	240	✓	122	✓
D8	751/30(P)	P20		154	✓	240	✓	240		294	✓	225	✓	122	✓
E8	751/31(P)	P21		160	✓	240	✓	NA		294	✓	240	✓	122	✓
F8	751/34(P)	P22		154/160	✓	240	✓	NA		294	✓	225/240	✓	122	✓
G8	751/43(P)	P23		154	✓	240	✓	NA		294	✓	225/240	✓	122	✓
H8	751/44(P)	P24		154/160	✓	240	✓	NA		294	✓	240	✓	122	✓
A9	751/48(P)	P25		152/165	✗	240/242	✗	225		292/294	✗	225	✓	122	✓
B9	751/49(P)	P26		160	✓	240	✓	240		294	✓	225/240	✓	122	✓

(g) DNA fingerprinting of *pisifera* 896 (C9-A10).

		SAMPLE NAME	Overall	OP1		OP5		OP13		OP12		OP11		OP24/6	
C9	*F1 150/07*	*B1*		*205/219*		*236/257*		*316/322*		*188/192/198/206*		*253/260*		*181/190*	
D9	896/14(P)	P1	✔	205/219	✓	257	✓	322	✓	188/206	✓	253	✓	181/190	✓
E9	896/48(P)	P2	✔	205/219	✓	257	✓	316/322	✓	192/198	✓	253	✓	181/190	✓
F9	896/38(P)	P3	✔	219	✓	236/257	✓	316/322	✓	206	✓	253/260	✓	181/190	✓
G9	896/51(P)	P4	✔	219	✓	257	✓	316/322	✓	206	✓	253/260	✓	181/190	✓
H9	896/51(P)	P5	✔	205/219	✓	NA		316	✓	192/206	✓	253/260	✓	181/190	✓
A10	896/44(P)	P6	✔	205/219	✓	257	✓	316/322	✓	188/206	✓	253/260	✓	181/190	✓

		SAMPLE NAME		OP2		OP7		OP20		OP18		OP21		OP29	
C9	*F1 150/07*	*B1*		*164/166*		*242*		*225/248*		*254/294*		*225/240*		*110/122*	
D9	896/14(P)	P1		164/166	✓	242	✓	248	✓	254/294	✓	240	✓	110	✓
E9	896/48(P)	P2		164	✓	242	✓	NA		254	✓	240	✓	122	✓
F9	896/38(P)	P3		164/166	✓	242	✓	NA		254	✓	225/240	✓	110	✓
G9	896/51(P)	P4		166	✓	242	✓	NA		254/294	✓	225/240	✓	110	✓
H9	896/51(P)	P5		166	✓	242	✓	NA		254	✓	225/240	✓	110/122	✓
A10	896/44(P)	P6		164/166	✓	242	✓	225	✓	254/294	✓	225/240	✓	110/122	✓

(h) DNA fingerprinting of *dura* 896 (B10-B11).

		SAMPLE NAME	Overall	OP1		OP5		OP13		OP12		OP11		OP24/6	
B10	F1 150/07	P6B1(1)		205/219		236/257		316/322		188/192/198/206		253/260		181/190	
C10	896/27(D)	D1	✔	205/219	✓	257	✓	316/322	✓	188/206	✓	253/260	✓	181/190	✓
D10	896/42(D)	D2	✔	219	✓	257	✓	316/322	✓	192/198	✓	253/260	✓	181/190	✓
E10	896/3(D)	D3	✔	NA		257	✓	316	✓	188/192/198/206	✓	253/260	✓	181/190	✓
F10	896/25(D)	D4	✔	219	✓	236/257	✓	322	✓	188/192/198/206	✓	253/260	✓	181/190	✓
G10	896/34(D)	D5	✔	205	✓	257	✓	316/322	✓	188/192/198/206	✓	253	✓	181/190	✓
H10	896/4(D)	D6	✔	205	✓	236/257	✓	316/322	✓	188/206	✓	253	✓	181/190	✓
A11	896/10(D)	D7		205	✓	257	✓	316	✓	n/a		253	✓	181/190	✓
B11	896/18(D)	D8		205	✓	257	✓	316	✓	188/206	✓	253/260	✓	181/190	✓

		SAMPLE NAME		OP2		OP7		OP20		OP18		OP21		OP29	
B10	F1 150/07	P6B1(1)		164/166		242		225/248		254/294		225/240		110/122	
C10	896/27(D)	D1		164/166	✓	NA		225/248	✓	254	✓	225/240	✓	110/122	✓
D10	896/42(D)	D2		166	✓	242	✓	225	✓	254/294	✓	225/240	✓	122	✓
E10	896/3(D)	D3		164/166	✓	242	✓	NA		254/294	✓	225	✓	122	✓
F10	896/25(D)	D4		166	✓	242	✓	NA		294	✓	225	✓	122	✓
G10	896/34(D)	D5		164	✓	242	✓	NA		254/294	✓	225	✓	122	✓
H10	896/4(D)	D6		164/166	✓	242	✓	NA		254/294	✓	225	✓	110	✓
A11	896/10(D)	D7		166	✓	242	✓	225	✓	254/294	✓	225/240	✓	110/122	✓
B11	896/18(D)	D8		164/166	✓	242	✓	225/248	✓	254/294	✓	225	✓	110	✓

Appendix 3 - List of SSR primers developed by CIRAD to amplify oil palm.

Local code	CIRAD locus	Repeat motif	reference size in LM2T	Linkage group
OP1	mEgCIR0146	(GT)2(GA)27	301	10
OP2	mEgCIR0163	(GA)23	143	8
OP5	mEgCIR0779	(CA)11(GA)22	238	14
OP7	mEgCIR0790	(GA)19	215	12
OP11	mEgCIR0874	(CA)11(GA)18	235	1
OP12	mEgCIR0878	(GA)22	185	11
OP13	mEgCIR0894	(GA)18	186	7
OP18	mEgCIR2518	(GT)6(GA)32	277	3
OP20	mEgCIR2670	(GA)20	226	15
OP21	mEgCIR2813	(GT)7(GA)11	210	5
OP24	mEgCIR3328	(GA)22	185	8
OP29	mEgCIR3809	(GA)22	113	1

Appendix 4- Lists of primers designed from candidate probe-sets and probe-pairs using four approaches to amplify oil palm.

(A) Primers designed from oil palm DNA cross-hybridised on Affymetrix *Arabidopsis* ATH1 GeneChip.

		Primer Name	Primer Sequence (5' - 3')	Primer Length	*Ta (º C)	Product Size (bp)	Organism	Designed by
1	Af_1	Affy. 255662_at_F	TATCTCTTACCTATTCGTATCCGAA	25	-	383	*Arabidopsis*	Affymetrix array probe sequences
		Affy. 255662_at_R	GGCCGAGATCAGGTGATTCGTTACC	25			*Arabidopsis*	Affymetrix array probe sequences
2	Af_2	Affy.245050_at_F	TAGTCGCCAAATTGCCAGAGGCCTA	25	50	99	*Arabidopsis*	Affymetrix array probe sequences
		Affy.245050_at_R	GCCAAACAAAGGCTAAGAGAAGAAA	25			*Arabidopsis*	Affymetrix array probe sequences
3	Af_3	Affy.245024_at_F	GAGTATGACTGCCTTACCAATCGTC	25	50	260	*Arabidopsis*	Affymetrix array probe sequences
		Affy.245024_at_R	ATTGGGAAAAGGCTTCTAATTCAGC	25			*Arabidopsis*	Affymetrix array probe sequences
4	Af_4	Affy.262910_at_F	GATTCTCTTGATTTCACACCTGGAT	25	50	296	*Arabidopsis*	Affymetrix array probe sequences
		Affy.262910_at_R	AGACGCAATGGGAAAAGCTTCCCGT	25			*Arabidopsis*	Affymetrix array probe sequences
5	Af_5	Affy.255530_at_F	GAGACGAGCCTAGTCTTTTTCCATC	25	50	176	*Arabidopsis*	Affymetrix array probe sequences
		Affy.255530_at_R	AACGGGAGTAGATTCAAGCTTGTGT	25			*Arabidopsis*	Affymetrix array probe sequences
6	Af_6	Affy.250113_at_F	GAATTTGAGCCAATCCCTGTTTTGA	25	50	357	*Arabidopsis*	Affymetrix array probe sequences
		Affy.250113_at_R	TGGTCTAGAAAGTAGCTGCTGACTC	25			*Arabidopsis*	Affymetrix array probe sequences
7	Af_7	Affy.249135_at_F	GTGGATAGTTCTGTATTGTCCCCAA	25	50	408	*Arabidopsis*	Affymetrix array probe sequences
		Affy.249135_at_R	TGGCAGCAACAAGCATATGGAAGAT	25			*Arabidopsis*	Affymetrix array probe sequences
8	Pr_1	P3.255662_at_F	ttctctcgttaccattcgtcatta	24	-	288	*Arabidopsis*	Primer 3 software
		P3.255662_at_R	cttcaatctctgttcccaaaactt	24			*Arabidopsis*	Primer 3 software
9	Pr_2	P3.245050_at_F	ttagtcgccaaattgccagag	21	50	100	*Arabidopsis*	Primer 3 software
		P3.245050_at_R	gccaaacaaaggctaagagaaga	23			*Arabidopsis*	Primer 3 software
10	Pr_3	P3.245024_at_F	agtatgactgccttaccaatcgtc	24	50	226	*Arabidopsis*	Primer 3 software
		P3.245024_at_R	ccaattccaattttaattttccag	24			*Arabidopsis*	Primer 3 software
11	Pr_4	P3.262910_at_F	gcttcatcattctgattctcttga	24	53	248	*Arabidopsis*	Primer 3 software
		P3.262910_at_R	cattgctctcttctttcaatctca	24			*Arabidopsis*	Primer 3 software
12	Pr_5	P3.255530_at_F	gtcgtcttcatgcgaagagactat	24	50	237	*Arabidopsis*	Primer 3 software
		P3.255530_at_R	caagcttgtgtgaagtatctctgg	24			*Arabidopsis*	Primer 3 software

		Primer Name	Primer Sequence (5' - 3')	Primer Length	*Ta (° C)	Product Size (bp)	Organism	Designed by
13	Pr_6	P3.250113_at_F	aggagaaagttgaggaacgtgtag	24	50	240	*Arabidopsis*	Primer 3 software
		P3.250113_at_R	aggtcataagaccataagggttca	24			*Arabidopsis*	Primer 3 software
14	Pr_7	P3.249135_at_F	gatgggacatctagaagagtggtt	24	55	240	*Arabidopsis*	Primer 3 software
		P3.249135_at_R	agttctgctgccaagctacttatt	24			*Arabidopsis*	Primer 3 software
15	Pr_8	P3.245001_at_F	ttaaatcccgagatattccaaaga	24	50	238	*Arabidopsis*	Primer 3 software
		P3.245001_at_R	gagtctaatgcgttcctttcattt	24			*Arabidopsis*	Primer 3 software
16	Tbx_1	TbX_255662_at_F	GTNCCNTTYGTNTAYGTNACNTAY	24	-	221	*Orysa*	TblastX database
		TbX_255662_at_R	ACCATDATNCKNGGNSWNGCNGTD	24			*Oryza*	TblastX database
17	Tbx_2	TbX_245050_at_F	TTYTTYTTYGCNAARYTNCCNGAR	24	-	104	*Elaies*	TblastX database
		TbX_245050_at_R	TGCCANACRA ANGCNARNAR RAAR	24			*Elaies*	TblastX database
18	Tbx_3	TbX_245024_at_F	WSNATGACNGCNYTNCCNATHGTN	24	-	194	*Elaies*	TblastX database
		TbX_245024_at_R	TTCATNGCYTTDATYTGNGCNGCN	24			*Elaies*	TblastX database
19	Tbx_4	TbX_262910_at_F	TGGGAYGTNGARATHGTNCARGTN	24	40	341	*Oryza*	TblastX database
		TbX_262910_at_R	CCYTTRTCRT CNGCDATRTG RTAR	24			*Oryza*	TblastX database
20	Tbx_5	TbX_255530_at_F	ACNGCNGCNGCNGGNGCNAC	20	55	173	*Sorghum*	TblastX database
		TbX_255530_at_R	GGNGGNGGNCKNCKNCKNGG	20			*Sorghum*	TblastX database
21	Tbx_6	TbX_250113_at_F	AARYTNGGNAARYTNGARAARGAR	24	40	204	*Zea mays*	TblastX database
		TbX_250113_at_R	TTDATRTANC CNGCRTGNGG NGCR	24			*Zea mays*	TblastX database
22	Tbx_7	TbX_249135_at_F	TGGAARGARATHWSNAARYTNMGN	24	-	281	*Oryza*	TblastX database
		TbX_249135_at_R	GCCATNGTRT GCATNACRTC NACN	24			*Oryza*	TblastX database
23	Tbx_8	TbX_245001_at_F	TGGGAYTAYATHCCNWSNTAYTGY	24	-	89	*Potamophila*	TblastX database
		TbX_245001_at_R	ACNGTYTTNA CRTADATDAT NARN	24			*Potamophila*	TblastX database
24	OP_AT_1	245050_at_F	CACTATTTTGTTTTGACATGACACC	25	59	438	Oil Palm	Oil palm 454 Seq via Primer 3
		245050_at_R	TTATGCCTTTTTAAATCCAATCGT	24			Oil Palm	Oil palm 454 Seq via Primer 3
25	OP_AT_2	245024_at_F	GCTACATTACAATACCTCGCTCCT	24	50	500	Oil Palm	Oil palm 454 Seq via Primer 3
		245024_at_R	AATTGTGCAAAGGCTTCTAACTCT	24			Oil Palm	Oil palm 454 Seq via Primer 3
26	OP_AT_3	245001_at_F	GCCAAATCGTTTCATTTAAAACTt	24	50	434	Oil Palm	Oil palm 454 Seq via Primer 3
		245001_at_R	GAATCCCATTTCGGATTTAGTATG	24			Oil Palm	Oil palm 454 Seq via Primer 3

		Primer Name	Primer Sequence (5' - 3')	Primer Length	*Ta (º C)	Product Size (bp)	Organism	Designed by
27	OP_AT_4	245002_at_F	TGGTGTTTTCGGACTAATAGGTTT	24	50	383	Oil Palm	Oil palm 454 Seq via Primer 3
		245002_at_R	AATAAGTCTCTTCGGCTTGAGTTG	24			Oil Palm	Oil palm 454 Seq via Primer 3
28	OP_AT_5	265228_s_at_F	ATTTGTTTTtCAATTGGAAGTGGT	24	-	363	Oil Palm	Oil palm 454 Seq via Primer 3
		265228_s_at_R	TGGCTTTTGATTTATATCGTGCTA	24			Oil Palm	Oil palm 454 Seq via Primer 3
29	OP_AT_6	252041_at_F	ATGAGAGAGACACCAAGATTGTCA	24	-	350	Oil Palm	Oil palm 454 Seq via Primer 3
		252041_at_R	CACACTACATaACAAGCCACATGA	24			Oil Palm	Oil palm 454 Seq via Primer 3
30	OP_AT_7	265090_at_F	GAAtGAGAGTTACTTTACACTACGTGA	27	50	207	Oil Palm	Oil palm 454 Seq via Primer 3
		265090_at_R	GCATCTTCTTCCATagAAAGCCTA	24			Oil Palm	Oil palm 454 Seq via Primer 3
31	OP_AT_8	258484_at_F	AACAAAGGGCTACAGAAGTACACC	24	50	388	Oil Palm	Oil palm 454 Seq via Primer 3
		258484_at_R	CAAATATCTTCATGCAACACATCA	24			Oil Palm	Oil palm 454 Seq via Primer 3
32	OP_AT_9	245270_at_F	CTTtCTGTCGCTGAGATCACTAAC	24	50	447	Oil Palm	Oil palm 454 Seq via Primer 3
		245270_at_R	ATCTTCaTAATCCTTCTCCAGTGC	24			Oil Palm	Oil palm 454 Seq via Primer 3
33	OP_AT_10	256293_at_F	AGCGAAGGACAATTCTATCAAGTC	24	50	369	Oil Palm	Oil palm 454 Seq via Primer 3
		256293_at_R	CCGAGCATATGTGTAGCATAGATT	24			Oil Palm	Oil palm 454 Seq via Primer 3

*Ta (º C) refers to optimal annealing temperature where good amplification was obtained, '-' indicates no amplification.

(B) Primers designed from oil palm DNA cross-hybridised on Affymetrix Rice GeneChip.

		Primer Name	Primer Sequence (5' - 3')	Primer Length	*Ta (° C)	Product Size (bp)	Organism	Designed by
1	Os_1	Os.15514.1.S2_at_F	TGGTGTCCATATGGCCAACAGGTAA	25	-	139	Rice	Affymetrix array probe sequences
		Os.15514.1.S2_at_R	TGAGACCATGAGAAATTGTGCCATC	25			Rice	Affymetrix array probe sequences
2	Os_2	Os.34142.1.S1_at_F	AGAACTGTCACATGCTACCGAGAAG	25	50	286	Rice	Affymetrix array probe sequences
		Os.34142.1.S1_at_R	TCTCAAGTGTGTATCATGGTGTATT	25			Rice	Affymetrix array probe sequences
3	O.s 3	Os.17834.1.S1_at_F	gtcaaggctagttttgggggttag	23	50	144	Rice	Affymetrix array probe sequences
		Os.17834.1.S1_at_R	GCTATGGCGTCGTCAGTGCTACTTC	25			Rice	Affymetrix array probe sequences
4	Os_4	Os.54144.1.S1_at_F	TAAGTTCTAGGCCTTACCTGACAGC	25	50	392	Rice	Affymetrix array probe sequences
		Os.54144.1.S1_at_R	ATGGACTGTCATGGTAAGCTTACTC	25			Rice	Affymetrix array probe sequences
5	Os_5	Os.23612.1.S1_at_F	ATCTCTGGCGCCCTCTCTGTTGTTT	25	50	453	Rice	Affymetrix array probe sequences
		Os.23612.1.S1_at_R	TAACCTTGTAAATCAGGAGGCGTGG	25			Rice	Affymetrix array probe sequences
6	Os_6	Os.42585.1.S1_at_F	aggaggaggaggaggaaagag	21	55	225	Rice	Affymetrix array probe sequences
		Os.42585.1.S1_at_R	CATCGCCTGTAATTCCAAGAAAATA	25			Rice	Affymetrix array probe sequences
7	Os_7	Os.45970.1.S1_at_F	CCGTTAGCCCTATTCATATCCTATA	25	-	181	Rice	Affymetrix array probe sequences
		Os.45970.1.S1_at_R	CAAAAACAGTTTCGGAGAGGCCTAA	25			Rice	Affymetrix array probe sequences
8	Os_8	OsAffx.9410.1.S1_x_at_F	CAAATTTCTCACCAGTCTACTTCAC	25	-	399	Rice	Affymetrix array probe sequences
		OsAffx.9410.1.S1_x_at_R	TCGCCGCTAAAGTTCCCACTACGTG	25			Rice	Affymetrix array probe sequences
9	Os_9	OsAffx.9731.1.S1_at_F	TGCAGGATCCACACTGGATCCGCTG	25	-	419	Rice	Affymetrix array probe sequences
		OsAffx.9731.1.S1_at_R	GCGAGGACGGCATCAACAGAATCAG	25			Rice	Affymetrix array probe sequences
10	Os_10	OsAffx.32196.1.S1_x_at_F	GCATCCACATGTCCGTTTTCAAAGT	25	-	171	Rice	Affymetrix array probe sequences
		OsAffx.32196.1.S1_x_at_R	TGTCGAAATCCCTATAATGAGTAGC	25			Rice	Affymetrix array probe sequences
11	Os_11	Os.46267.1.S1_x_at_F	accagagacttaatattgggatcg	24	50	271	Rice	Affymetrix array probe sequences
		Os.46267.1.S1_x_at_R	CAGCAATAATCAATTTTAGCGCGAA	25			Rice	Affymetrix array probe sequences
12	Os_12	Os.26548.1.S1_at_F	ATTGCGCTATCTTATGTCATTGGTG	25	-	197	Rice	Affymetrix array probe sequences
		Os.26548.1.S1_at_R	GCAAGCGCACCGATAATAGCAGTTT	25			Rice	Affymetrix array probe sequences
13	Os_13	Os.24952.1.S1_at_F	TTCTTGGACATAGTTCTTCTTCTTC	25	-	268	Rice	Affymetrix array probe sequences
		Os.24952.1.S1_at_R	AGTTAAAAAGAACAGATTGATGCTC	25			Rice	Affymetrix array probe sequences

		Primer Name	Primer Sequence (5' - 3')	Primer Length	*Ta (º C)	Product Size (bp)	Organism	Designed by
14	Os_14	Os.36236.1.S1_at_F	CGATCTCATCCAGTCTTATTTGCAG	25	-	239	Rice	Affymetrix array probe sequences
		Os.36236.1.S1_at_R	GAAGGGAAACTTGACTATGAAAACA	25			Rice	Affymetrix array probe sequences
15	Os_15	OsAffx.29383.1.S1_x_at_F	atgcaaccggatggcccgtcagag	24	50	282	Rice	Affymetrix array probe sequences
		OsAffx.29383.1.S1_x_at_R	GGAGACGGCCAGATCTGCTGCCGCC	25			Rice	Affymetrix array probe sequences
16	Os_16	Os.29823.3.S1_at_F	TTGCTGTTAGGCAATAGCCCCTTGT	25	-	334	Rice	Affymetrix array probe sequences
		Os.29823.3.S1_at_R	TCTCAGCAACTCGATTGGGTGTAGT	25			Rice	Affymetrix array probe sequences
17	Os_17	OsAffx.21085.1.S1_at_F	GATGGTTCTCATCGGATACGCCGAC	25	50	177	Rice	Affymetrix array probe sequences
		OsAffx.21085.1.S1_at_R	TCGACGGTGAAGTCTGAGCAACCTC	25			Rice	Affymetrix array probe sequences
18	OP_OS_1	OsAffx.32330.1.S1_x_at_F	GGATGGATTATGGGAGTAACAAAG	24	59	371	Oil Palm	Oil palm 454 Seq via Primer 3
		OsAffx.32330.1.S1_x_at_R	TGGAATTTATTGACATTCTCCAAA	24			Oil Palm	Oil palm 454 Seq via Primer 3
19	OP_OS_2	Os.38100.1.S1_at_F	AGATCATTAAAATTCCAGGCACAT	24	50	419	Oil Palm	Oil palm 454 Seq via Primer 3
		Os.38100.1.S1_at_R	AATAAATAAGTGGCATGTGGATTC	24			Oil Palm	Oil palm 454 Seq via Primer 3
20	OP_OS_3	Os.23127.1.S1_s_at_F	GTTTTTGAGGACAATGTTCTTGTG	24	50	409	Oil Palm	Oil palm 454 Seq via Primer 3
		Os.23127.1.S1_s_at_R	CAACAGTGCTGATACAAAGACAGA	24			Oil Palm	Oil palm 454 Seq via Primer 3
21	OP_OS_4	OsAffx.32237.1.A1_at_F	GAATCGGTTgAATTGTTGTTCATA	24	50	353	Oil Palm	Oil palm 454 Seq via Primer 3
		OsAffx.32237.1.A1_at_R	ACAAATTCGATTGATTGATACGAG	24			Oil Palm	Oil palm 454 Seq via Primer 3
22	OP_OS_5	Os.28037.1.A1_at_F	TCGAGTATAGGTGAGTACGCTTGA	24	-	386	Oil Palm	Oil palm 454 Seq via Primer 3
		Os.28037.1.A1_at_R	ACGTAAAGCGAATGATTAGAGGTC	24			Oil Palm	Oil palm 454 Seq via Primer 3
23	OP_OS_6	OsAffx.32279.1.S1_at_F	AATTTCCAGAAACCACACGATTAT	24	50	438	Oil Palm	Oil palm 454 Seq via Primer 3
		OsAffx.32279.1.S1_at_R	aagaagtggagttagaaagccgta	24			Oil Palm	Oil palm 454 Seq via Primer 3
24	OP_OS_7	Os.57569.1.S1_at_F	TTCCAAACAATCGAGAACTTTACA	24	-	430	Oil Palm	Oil palm 454 Seq via Primer 3
		Os.57569.1.S1_at_R	TGTCGGTAATGAAGTCATCAGTTT	24			Oil Palm	Oil palm 454 Seq via Primer 3
25	OP_OS_8	Os.12924.1.S1_s_at_F	GATATCAAGCTCACACACATTTCC	24	50	381	Oil Palm	Oil palm 454 Seq via Primer 3
		Os.12924.1.S1_s_at_R	GCAGGtaaCAAGaAAgGGaAAAA	23			Oil Palm	Oil palm 454 Seq via Primer 3
26	OP_OS_9	Os.33607.2.S1_x_at_F	GGCAACATACCACTGAATCAAGTA	24	50	406	Oil Palm	Oil palm 454 Seq via Primer 3
		Os.33607.2.S1_x_at_R	CCCTCTGAAACGTAAAGTCAATCT	24			Oil Palm	Oil palm 454 Seq via Primer 3

*Ta (º C) refers to optimal annealing temperature where good amplification was obtained, '-' indicates no amplification.

(C) Primers designed from oil palm DNA cross-hybridised on Affymetrix Rice GeneChip at signal intensity of 500 and below.

		Primer Name	Primer Sequence (5' - 3')	Primer Length	*Ta (º C)	Product Size (bp)	Organism	Designed by
1	OS_L_1b	OsAffx.13276.1.S1_454_F	ACCTCACCAAACctAAAAAGTGTC	24		317	Rice	Oil palm 454 Seq via Primer 3
		OsAffx.13276.1.S1_454_R	catTgGAGAGaAGAAgGTCAATG	23			Rice	Oil palm 454 Seq via Primer 3
2	OS_L_3b	Os.9523.1.S1_at_454_F	TGTTCTTTTATATTTTGCTTGTCAGC	26		350	Rice	Oil palm 454 Seq via Primer 3
		Os.9523.1.S1_at_454_R	CATTTTTCATATTCTTGCACCATT	24			Rice	Oil palm 454 Seq via Primer 3
3	OS_L_4b	Os.49922.1.S1_at_454_F	ATGAGATTTCAATTTGATGCTGTC	24		354	Rice	Oil palm 454 Seq via Primer 3
		Os.49922.1.S1_at_454_R	AAAGAAGTCCAAGATGAAGGTTGA	24			Rice	Oil palm 454 Seq via Primer 3
4	OS_L_5b	Os.51235.1.S1_at_454_F	CTATCATCCCCTGAATCCTTTTTA	24		351	Rice	Oil palm 454 Seq via Primer 3
		Os.51235.1.S1_at_454_R	TTATAGAGGATCCAACTTGCCTTC	24			Rice	Oil palm 454 Seq via Primer 3
5	OS_L_6b	OsAffx.18742.1.S1_at_454_F	TTACATTTACCTGCTGAtCCTGAA	24		440	Rice	Oil palm 454 Seq via Primer 3
		OsAffx.18742.1.S1_at_454_R	CACTTgAAtTGtTGCTTTTCAATC	24			Rice	Oil palm 454 Seq via Primer 3
6	OS_L_9b	Os.54523.1.S1_at_454_F	GTTCTGGCTGCATTGAAGAAG	21		358	Rice	Oil palm 454 Seq via Primer 3
		Os.54523.1.S1_at_454_R	AGACTGAGGATGTGACCTATCTCC	24			Rice	Oil palm 454 Seq via Primer 3
7	OS_L_12b	Os.53248.1.A1_at_454_F	TTGAGGTAGAGCTTAGGAGATTGA	24		371	Rice	Oil palm 454 Seq via Primer 3
		Os.53248.1.A1_at_454_R	TGAAAAATTCAGCTCAAACATCTC	24			Rice	Oil palm 454 Seq via Primer 3
8	OS_L_13b	Os.12010.1.S1_x_at_454_F	TACTTTGCTTtCTCaTGCctCATA	24		449	Rice	Oil palm 454 Seq via Primer 3
		Os.12010.1.S1_x_at_454_R	CAACCAGCACTTaATCAGAGAATG	24			Rice	Oil palm 454 Seq via Primer 3
9	OS_L_14b	Os.54503.1.A1_at_454_F	TTCCAAGGGTCCTGTAAATAGTTC	24		430	Rice	Oil palm 454 Seq via Primer 3
		Os.54503.1.A1_at_454_R	CCTTGGTAAGAAAGAAGAAACCAG	24			Rice	Oil palm 454 Seq via Primer 3

*Ta (º C) refers to optimal annealing temperature where good amplification was obtained, '-' indicates no amplification.

Appendix 5 – List of potential probe-sets with reasonable fold-change value between *dura* and *pisifera* at all threshold level.

(A)   Potential probe-set and probe-pairs that are generated from cross-hybridisation on *Arabidopsis* GeneChip.

	Potential probe-set	Potential probe-pairs	Fold-change Value		
			768	769	Superbulk
1	255662_at	probe 7	3.7	1.3	1.5
2	245050_at	probe 10	6.3	1.7	1.2
		probe 7	2.7	2	1.2
3	245024_at	probe 4	2	1.2	1.3
		probe 3	4	4.2	2.1
4	262910_at	probe 9	1.3	2.4	1.8
5	255530_at	probe 7	1.2	2.2	1.1
6	250113_at	probe 4	1.4	2.4	1.3
7	249135_at	probe 5	2.1	2.3	1.4
8	245001_at	probe 3	2.2	2.1	1.7
9	262702_at	probe 7	1.6	2.5	3.9
10	256913_at	probe 8	1.2	1.5	2.1
11	254929_at	probe 6	1.2	1.7	3.1
12	254144_at	probe 7	1.3	2	4.5
13	252750_at	probe 8	1.2	2	2.1
14	247792_at	probe 6	1.5	1.3	2.7
15	247241_at	probe 9	1.1	1.4	2
16	246168_at	probe 4	1.2	1.3	2.3
17	245983_at	probe 9	2.4	1.3	2.4
18	245025_at	probe 3	2.9	1.3	1.1
19	245026_at	probe 3	3.2	1.5	1.1
20	245001_at	probe 4	3.2	1.8	1.4
21	245002_at	probe 2	1.7	1.4	1.8
22	245017_at	probe 3	3.3	1.7	1.1
23	244974_at	probe 9	3.5	1.1	1.1
24	244982_at	probe 6	4	1.1	1.1
25	244961_at	probe 9	3.2	1.2	1.1
26	252041_at	probe 8	3.4	-	1.3
27	265090_at	probe 6	1.8	1.3	1.3
28	258484_at	probe 8	1.7	1.2	-
		probe 9	3.1	-	-
29	244968_at	probe 2	3.2	1.3	1.2
30	245270_at	probe 8	1.4	1.1	1.1
		probe 9	1.4	1.3	1.1
31	256293_at	probe 5	1.6	1.2	1.1

(B)   Potential probe-set and probe-pairs that are generated from cross-hybridisation on rice GeneChip at signal intensity of 500 and above.

	Potential probe-set	Potential probe-pairs	Fold-change Value 768	Fold-change Value Superbulk
1	Os.15514.1.S2_at	probe 8	3.5	1.4
2	Os.34142.1.S1_at	probe 3	3.4	1.3
3	Os.17834.1.S1_at	probe 3	4.2	1.4
4	Os.54144.1.S1_at	probe 3	3.6	1.7
5	Os.23612.1.S1_at	probe 4	3.5	1.2
6	Os.42585.1.S1_at	probe 2	3.4	1.3
7	Os.45970.1.S1_at	probe 4	3.4	1.2
8	OsAffx.9410.1.S1_x_at	probe 3	3.4	1.9
9	OsAffx.9731.1.S1_at	probe 5	5.5	4.1
10	OsAffx.26469.2.S1_at	probe 5	5.1	1.4
11	OsAffx.13460.1.S1_at	probe 6	3.6	1.2
12	OsAffx.32196.1.S1_x_at	probe 5	2.6	1.6
13	Os.46267.1.S1_x_at	probe 7	2.3	1.6
14	Os.26548.1.S1_at	probe 6	1.7	2.6
15	Os.24952.1.S1_at	probe 7	1.5	2.9
16	Os.36236.1.S1_at	probe 7	1.4	3.1
17	OsAffx.6968.1.S1_x_at	probe 10	1.8	2.6
18	OsAffx.29383.1.S1_x_at	probe 3	1.4	3.2
19	OsAffx.29383.1.S1_x_at	probe 7	1.5	3.1
20	OsAffx.21085.1.S1_at	probe 3	1.7	2.5
21	OsAffx.30822.1.S1_at	probe 5	1.7	3.2
22	OsAffx.2631.1.S1_at	probe 6	1.6	2.9
23	OsAffx.28750.1.S1_at	probe 9	1.4	2.8
24	OsAffx.2626.1.S1_at	probe 6	2	3.2
25	Os.9523.1.S1_at	probe 4	3.3	1.3
26	Os.54297.1.S1_at	probe 5	2.3	1.2
27	Os.51839.1.S1_x_at	probe 2	4.7	2.4
28	Os.50167.1.S1_at	probe 2	3.2	1.2
29	Os.5846.1.S1_at	probe 6	1.6	1.2
30	Os.9168.1.S1_at	probe 3	4.9	1.2
31	Os.2486.1.S1_at	probe 5	1.9	1.1
		probe 6	1.6	1.6
32	OsAffx.25789.1.S1_at	probe 3	4.7	1.1
		probe 8	1.4	1.7
33	OsAffx.16056.2.S1_x_at	probe 9	4.1	1.2
34	OsAffx.25602.1.S1_at	probe 5	3.7	2.1
35	Os.21876.1.S1_at	probe 2	4.8	1.1
36	Os.14280.1.S1_x_at	probe 4	4.8	1.4
37	OsAffx.6491.1.S1_at	probe 8	4.1	1.3
38	Os.9123.1.S1_a_at	probe 7	1.5	2.6

39	Os.49953.1.S1_at	probe 4	2	2.7
40	Os.50186.1.S1_at	probe 5	2	1.2
		probe 7	2.8	1.4
41	Os.54503.1.A1_at	probe 5	2.1	3.7
42	Os.1044.1.S1_at	probe 3	1.7	2.9
43	OsAffx.29383.1.S1_at	probe 5	1.2	3.2
		probe 6	1.5	1.6
44	OsAffx.12970.1.S1_s_at	probe 7	1.4	2.6
45	OsAffx.16707.1.S1_at	probe 4	1.6	2.6
		probe 5	1.2	1.7
46	OsAffx.2052.1.S1_at	probe 9	1.3	2.5
47	OsAffx.12538.1.S1_at	probe 7	1.3	2
		probe 8	1.3	2.6
48	Os.56450.1.S1_at	probe 4	1.2	1.4
49	Os.22683.1.S1_at	probe 7	1.4	1.7
50	OsAffx.10614.1.S1_x_at	probe 8	2.1	1.5
51	OsAffx.27688.1.S1_at	probe 4	1.4	1.3
52	OsAffx.32330.1.S1_x_at	probe 4	1.6	1.26
53	Os.38100.1.S1_at	probe 5	1.33	1.35
54	Os.23127.1.S1_s_at	probe 6	1.6	1.24
55	OsAffx.32237.1.A1_at	probe 3	1.9	1.8
		probe 10	2	1.38
56	Os.28037.1.A1_at	probe 2	6.1	1.4
		probe 9	3.3	1.5
		probe 11	3.1	1.2
57	OsAffx.32279.1.S1_at	probe 9	1.6	1.2
58	Os.57569.1.S1_at	probe 7	4.5	1.2
59	Os.12924.1.S1_s_at	probe 8	1.8	1.3
		probe 10	4.9	1.4
60	Os.33607.2.S1_x_at	probe 9	1.6	1.2

(C) Potential probe-set and probe-pairs that are generated from cross-hybridisation on rice GeneChip at signal intensity of 500 and below.

	Potential probe-set	Potential probe-pairs	Fold-change Value	
			768	Superbulk
1	OsAffx.13276.1.S1_at	probe 2	3.9	3
2	OsAffx.9753.1.S1_at	probe 3	2.9	1.3
3	Os.9523.1.S1_at	probe 4	3.3	1.26
4	Os.49922.1.S1_at	probe 4	3.1	2.13
5	Os.51235.1.S1_at	probe 4	3.8	2.3
6	OsAffx.18742.1.S1_at	probe 4	3.2	2.54
7	OsAffx.23724.1.S1_x_at	probe 10	2.9	2.1
8	OsAffx.18311.1.S1_at	probe 7	1.8	2.1
9	Os.54523.1.S1_at	probe 3	1.6	2.2
10	Os.53103.1.S1_x_at	probe 4	2.2	2.3
11	OsAffx.2690.1.S1_at	probe 5	1.3	2.5
12	Os.53248.1.A1_at	probe 4	8.1	3.5
13	Os.12010.1.S1_x_at	probe 2	1.72	2.4
14	Os.54503.1.A1_at	probe 5	2.1	3.7

Appendix 6 - The PCA diagrams for ten characters measured in the F_5 segregating populations of Bambara groundnut under (a) drought conditions and (b) irrigated conditions.

(a)

(b)

| PC1 VS PC2 | PC2 VS PC3 | PC1 VS PC3 |

Appendix 7 (a) - The distinctness graphs of the top 100 PM probes ranking from the highest to lowest distinctness score.

7.93 Gma.3025.1.S1_at	933459
7.30 GmaAffx.63504.1.S1_s_at	460879
5.31 Gma.15877.1.S1_at	41730
5.01 GmaAffx.47965.1.S1_at	414206
4.98 Gma.15384.1.S1_at	1240235
4.62 Gma.5649.1.S1_at	570186
4.55 Gma.12105.1.S1_at	171610
4.41 Gma.12776.2.S1_at	1210247
4.30 Gma.11014.1.S1_at	1270451
4.29 GmaAffx.77310.1.S1_at	34582
4.16 GmaAffx.88141.1.S1_at	1052116
4.10 Gma.3757.2.S1_s_at	1270872
4.00 Gma.10361.1.S1_at	862663
3.94 PsAffx.C86000013_at	586615
3.94 GmaAffx.16203.1.S1_at	478734
3.92 Gma.14062.1.A1_at	909357
3.91 Gma.5649.1.S1_at	869553
3.90 GmaAffx.85321.1.S1_at	642098
3.87 Gma.12895.2.S1_at	246402
3.86 Gma.13289.1.S1_at	581030
3.86 Gma.5291.1.S1_at	344023
3.84 Gma.5712.1.S1_s_at	560967
3.82 GmaAffx.71175.1.S1_s_at	1195578
3.80 Gma.5717.4.S1_a_at	268934
3.80 Gma.3967.1.S1_at	1203917

321

Appendix 7 (a) (cont.) - The distinctness graphs of the top 100 PM probes ranking from the highest to lowest distinctness score.

3.78 GmaAffx.51734.1.A1_at 392624	3.77 GmaAffx.36098.1.S1_at 1146768	3.74 Gma.7504.2.S1_a_at 356261	3.72 Gma.12360.1.S1_at 334638	3.71 GmaAffx.36098.1.S1_at 590374
3.71 GmaAffx.32555.1.S1_s_at 420458	3.70 Gma.289.1.S1_s_at 937841	3.70 GmaAffx.32555.1.S1_s_at 223406	3.66 HgAffx.10300.1.S1_at 772419	3.64 GmaAffx.6438.3.S1_s_at 660770
3.64 Gma.3572.1.A1_at 691231	3.64 GmaAffx.79730.1.S1_at 155763	3.63 GmaAffx.77400.1.S1_at 50592	3.63 Gma.5717.4.S1_a_at 1326172	3.63 Gma.1655.1.S1_at 1326937
3.61 Gma.7504.2.S1_a_at 1196417	3.60 GmaAffx.36098.1.S1_at 821143	3.60 GmaAffx.70287.1.S1_at 1201169	3.53 Gma.13332.1.S1_at 1276910	3.56 Gma.7317.2.A1_a_at 865108
3.54 Gma.3967.1.S1_at 41312	3.54 GmaAffx.80642.1.S1_at 1286643	3.53 Gma.6289.3.S1_a_at 1046679	3.53 GmaAffx.32555.1.S1_s_at 695279	3.53 Gma.10990.3.S1_at 386431

Appendix 7 (a) (cont.) - The distinctness graphs of the top 100 PM probes ranking from the highest to lowest distinctness score.

Appendix 7 (a) (cont.) - The distinctness graphs of the top 100 PM probes ranking from the highest to lowest distinctness score.

3.32 GmaAffx.45888.1.S1_at 242025	3.31 GmaAffx.82829.1.S1_at 1135008	3.31 Gma.7329.1.S1_at 1228282	3.31 Gma.3659.1.S1_at 954425	3.29 Gma.14062.1.A1_at 136562
3.29 GmaAffx.14750.1.S1_s_at 1100327	3.29 Gma.2243.1.S1_at 1330734	3.27 GmaAffx.14750.1.S1_s_at 1326613	3.27 GmaAffx.14750.1.S1_s_at 311945	3.26 Gma.3204.1.S1_at 520854
3.26 Gma.8681.2.S1_at 826017	3.26 GmaAffx.31352.1.S1_s_at 337416	3.24 GmaAffx.14750.1.S1_s_at 245850	3.24 GmaAffx.20667.1.S1_at 324060	3.24 GmaAffx.14750.1.S1_s_at 139038
3.22 GmaAffx.82968.1.S1_at 211431	3.22 Gma.7851.1.S1_at 597551	3.19 GmaAffx.67103.1.S1_at 8931	3.19 PsAffx.C393000001_at 106533	3.17 GmaAffx.83095.1.S1_s_at 187647
3.16 GmaAffx.36562.2.S1_at 33072	3.15 GmaAffx.69325.1.S1_at 1253166	3.15 GmaAffx.82934.1.S1_at 786140	3.15 GmaAffx.14750.1.S1_s_at 886163	3.15 GmaAffx.18174.1.S1_at 1021848

Appendix 7 (b) - The distinctness graphs of the top 100 MM probes ranking from the highest to lowest distinctness score.

7.10 Gma.289.1.S1_s_at 1048658	4.73 Gma.17784.1.S1_at 177722	4.47 Gma.15877.1.S1_at 42834	4.40 Gma.1310.1.S1_s_at 65264	4.39 PsAffx.C118000052_at 72378
4.10 Gma.3025.1.S1_at 334623	4.04 GmaAffx.61513.1.A1_at 854661	3.95 Gma.12105.1.S1_at 172774	3.84 HgAffx.6074.1.A1_at 1129711	3.84 GmaAffx.4391.1.S1_at 761964
3.75 Gma.13332.1.S1_at 242284	3.68 Gma.17534.1.S1_at 434038	3.60 PsAffx.C73000015_at 219374	3.54 PsAffx.psMY009iC07r_at 1176526	3.43 GmaAffx.72634.1.S1_at 272560
3.39 PsAffx.C170000002_at 987546	3.31 Gma.5278.1.S1_at 887684	3.26 Gma.1249.1.S1_at 268665	3.21 HgAffx.7722.1.A1_at 515019	3.17 HgAffx.13872.1.S1_at 836732
3.16 Gma.8034.1.A1_s_at 352456	3.12 GmaAffx.30338.1.S1_at 498581	3.12 PsAffx.C61000056_at 447196	3.11 GmaAffx.48936.1.S1_at 1344128	3.11 GmaAffx.82368.1.S1_at 212535

Appendix 7 (b) (cont.) - The distinctness graphs of the top 100 MM probes ranking from the highest to lowest distinctness score.

3.11 GmaAffx.58232.1.A1_at   226183
3.10 Gma.5442.2.A1_at   924342
3.10 GmaAffx.93629.1.S1_s_at   188812
3.09 GmaAffx.58761.1.S1_at   95633
3.09 GmaAffx.22630.1.S1_at   1292507

3.06 Gma.7504.2.S1_a_at   957425
3.04 PsAffx.C38000120_at   226920
3.04 PsAffx.C131000041_at   194116
2.98 GmaAffx.79122.1.S1_at   919855
2.97 GmaAffx.21062.1.A1_at   72558

2.97 GmaAffx.77310.1.S1_at   35746
2.93 Gma.3583.1.S1_at   1025482
2.92 GmaAffx.89095.1.S1_s_at 735695
2.91 GmaAffx.89229.1.S1_at   959880
2.90 GmaAffx.63832.1.S1_at   438145

2.90 GmaAffx.79122.1.S1_at   292051
2.89 GmaAffx.59996.1.S1_at   356696
2.89 GmaAffx.42864.1.S1_at 1344127
2.88 Gma.7851.1.S1_at   598715
2.87 Gma.8020.1.S1_at   961488

2.87 GmaAffx.77836.1.S1_at 1236443
2.87 GmaAffx.36098.1.S1_at   673496
2.86 Gma.5712.1.S1_s_at   562131
2.86 Gma.8681.2.S1_at   827181
2.86 GmaAffx.81926.1.S1_at   161048

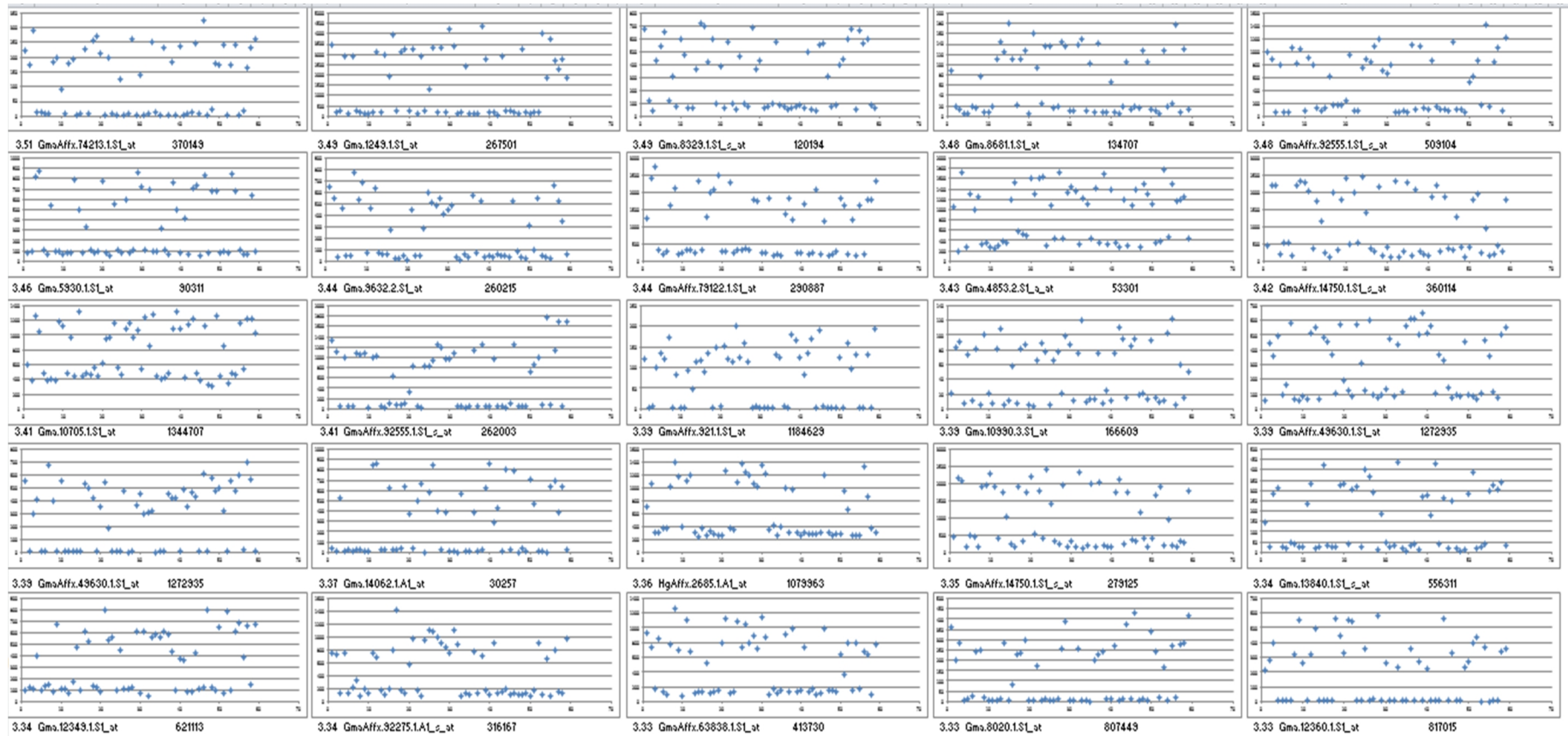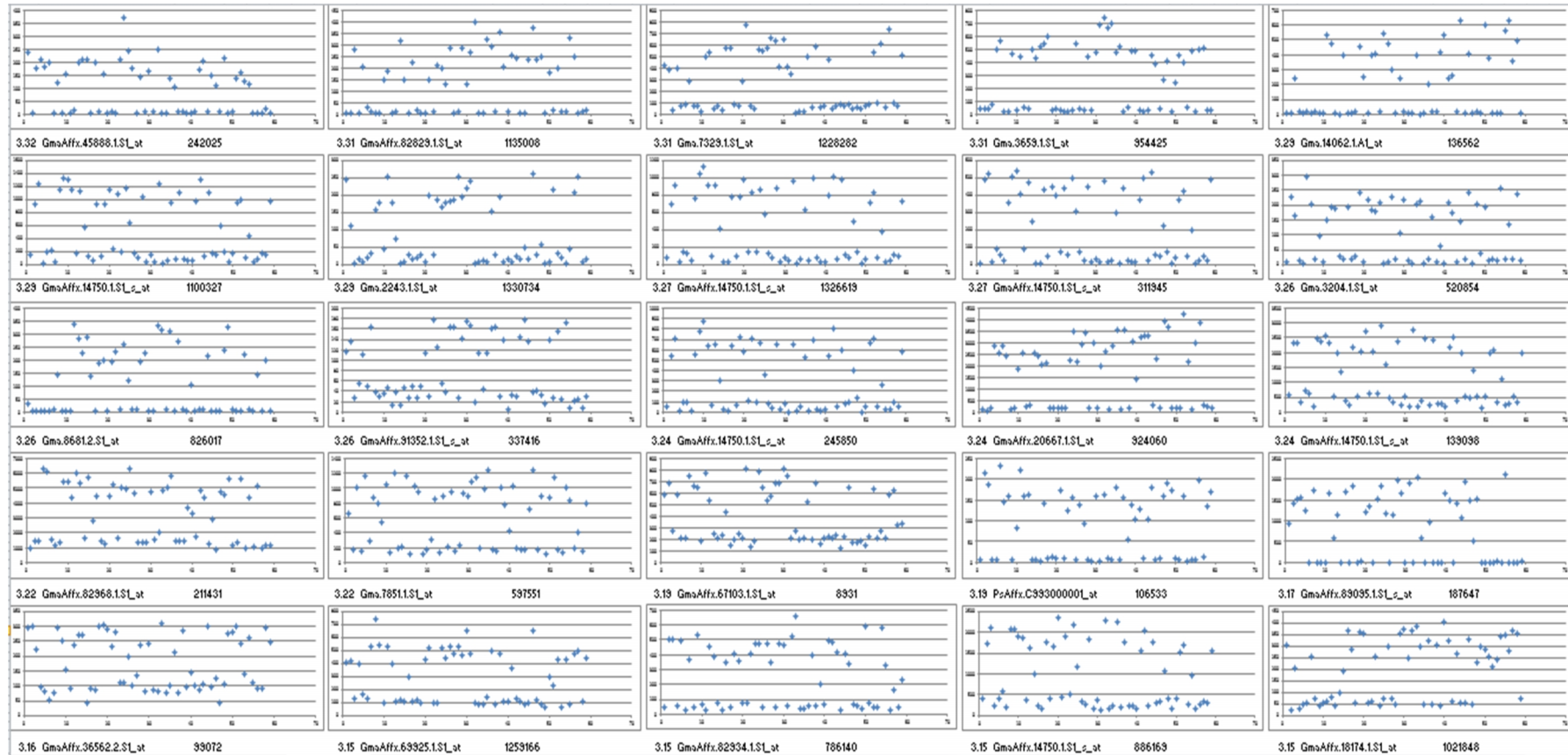**Appendix 7 (b) (cont.) -** The distinctness graphs of the top 100 MM probes ranking from the highest to lowest distinctness score.
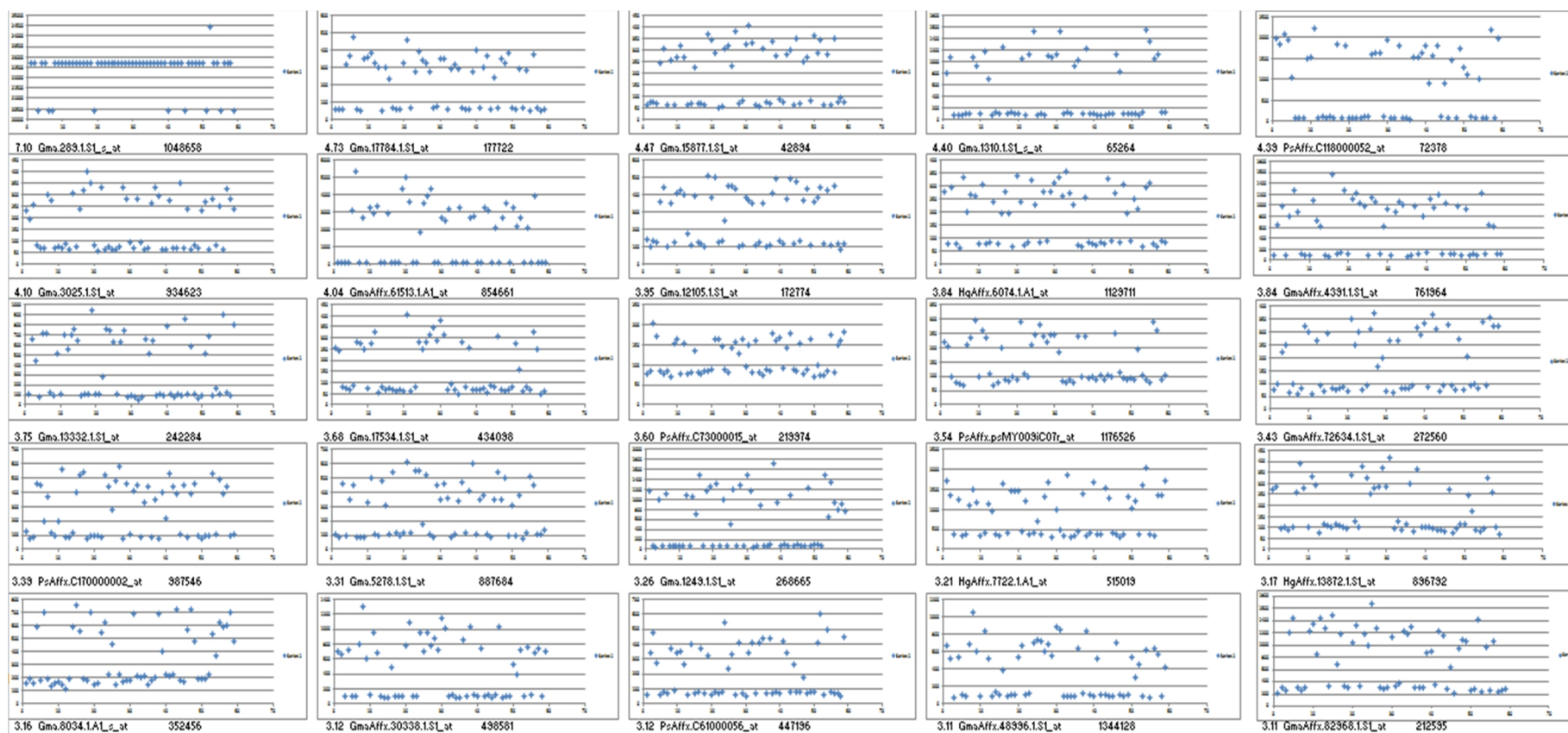

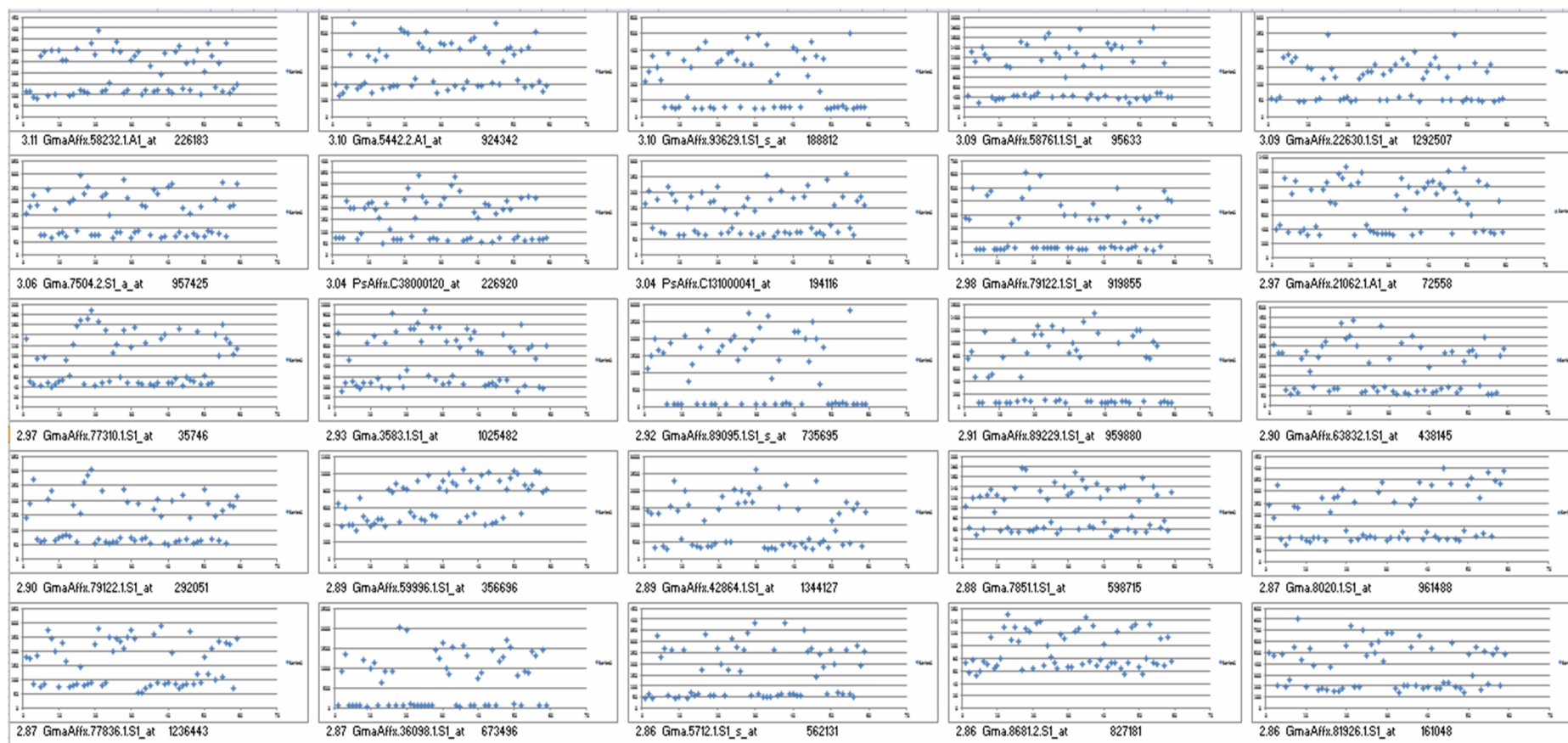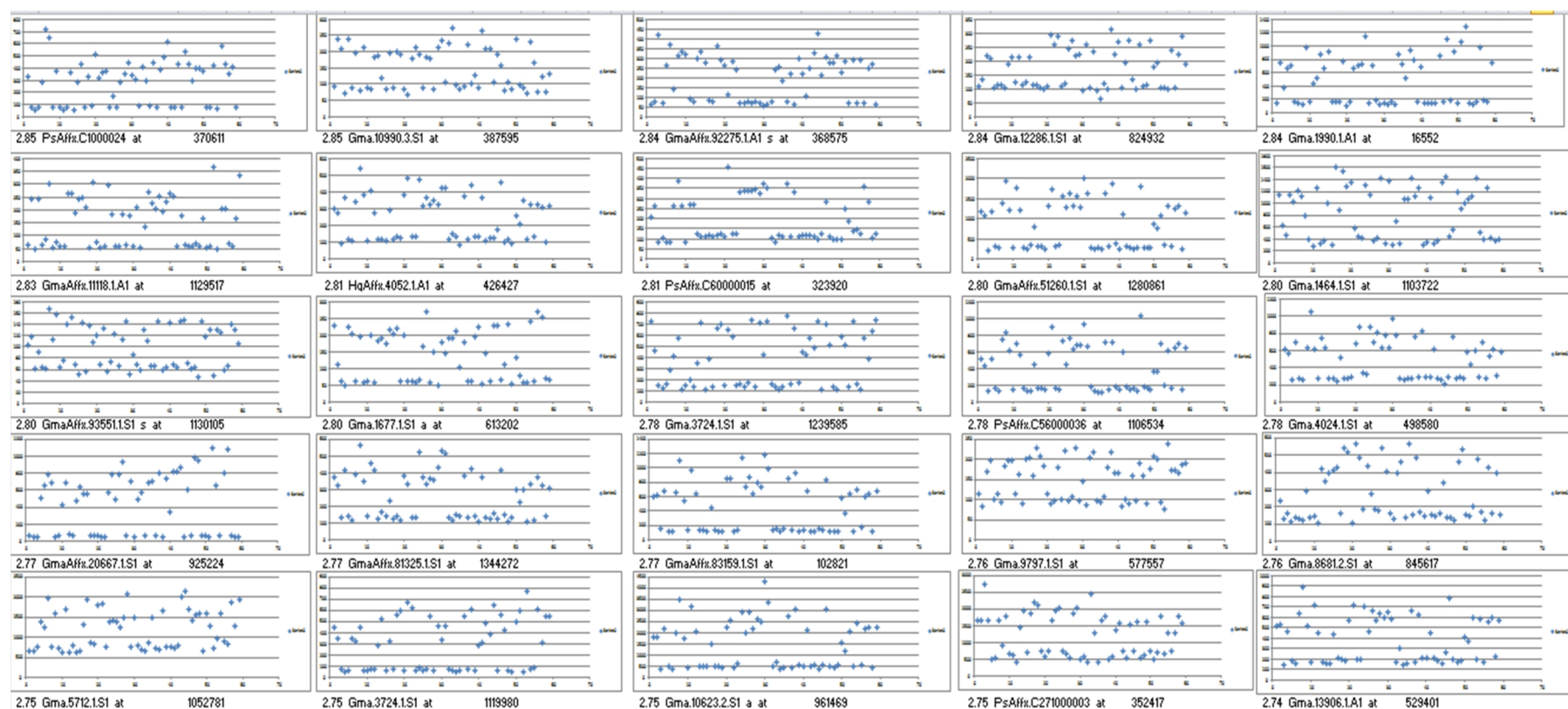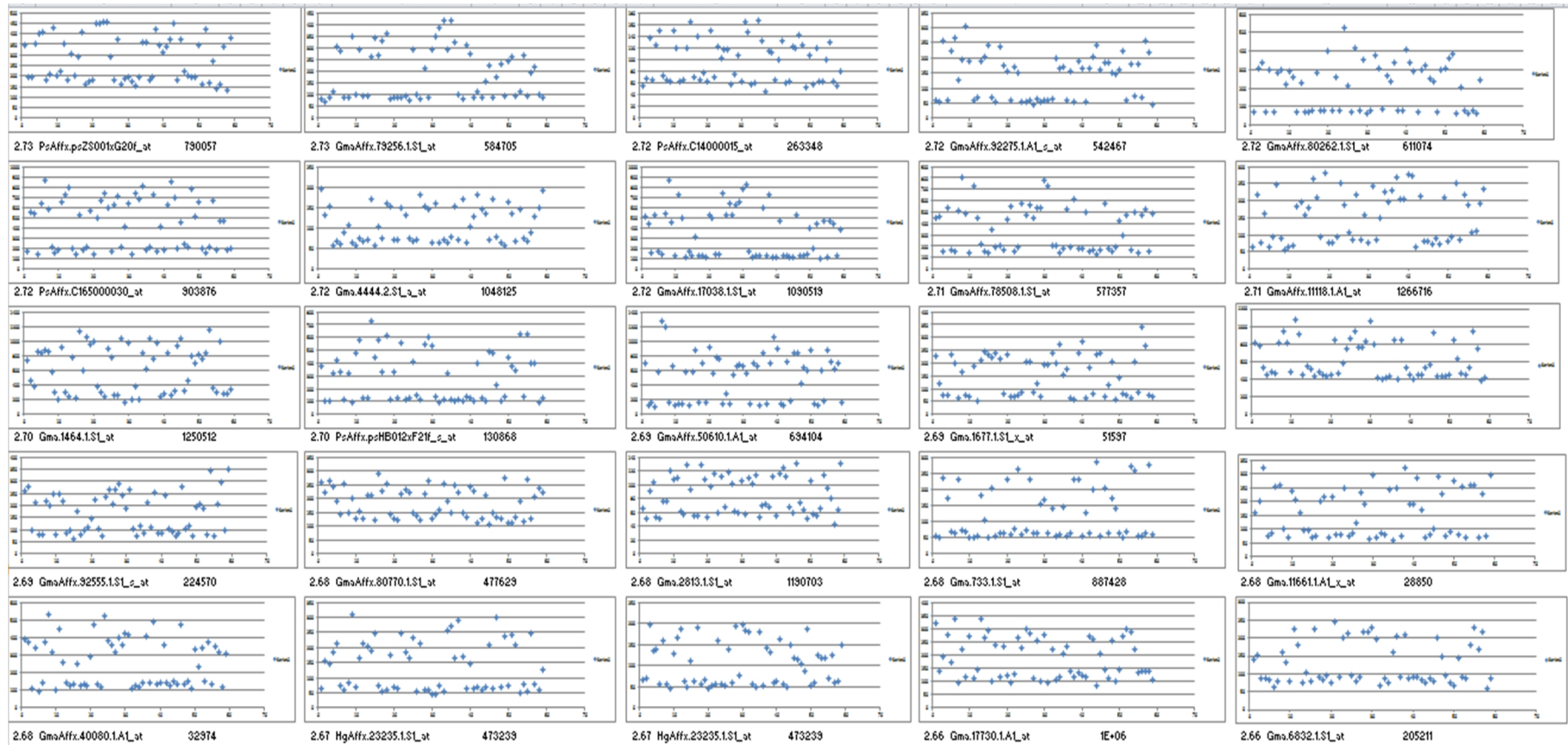|				
2.85 PsAffx.C1000024_at 370611	2.85 Gma.10990.3.S1_at 387595	2.84 GmaAffx.92275.1.A1_s_at 368575	2.84 Gma.12286.1.S1_at 824932	2.84 Gma.1990.1.A1_at 16552
2.83 GmaAffx.11118.1.A1_at 1129517	2.81 HgAffx.4052.1.A1_at 426427	2.81 PsAffx.C60000015_at 323920	2.80 GmaAffx.51260.1.S1_at 1280861	2.80 Gma.1464.1.S1_at 1103722
2.80 GmaAffx.93551.1.S1_s_at 1130105	2.80 Gma.1677.1.S1_a_at 613202	2.78 Gma.3724.1.S1_at 1239585	2.78 PsAffx.C56000036_at 1106534	2.78 Gma.4024.1.S1_at 498580
2.77 GmaAffx.20667.1.S1_at 925224	2.77 GmaAffx.81325.1.S1_at 1344272	2.77 GmaAffx.83159.1.S1_at 102821	2.76 Gma.9797.1.S1_at 577557	2.76 Gma.8681.2.S1_at 845617
2.75 Gma.5712.1.S1_at 1052781	2.75 Gma.3724.1.S1_at 1119980	2.75 Gma.10623.2.S1_a_at 961469	2.75 PsAffx.C271000003_at 352417	2.74 Gma.13906.1.A1_at 529401

327

Appendix 7 (b) (cont.) - The distinctness graphs of the top 100 MM probes ranking from the highest to lowest distinctness score.

Appendix 8 - The marker locations for LG5A, LG 8B and LG11A in genetic linkage maps using two mapping approaches, regression mapping (left) and maximum likelihood (right).
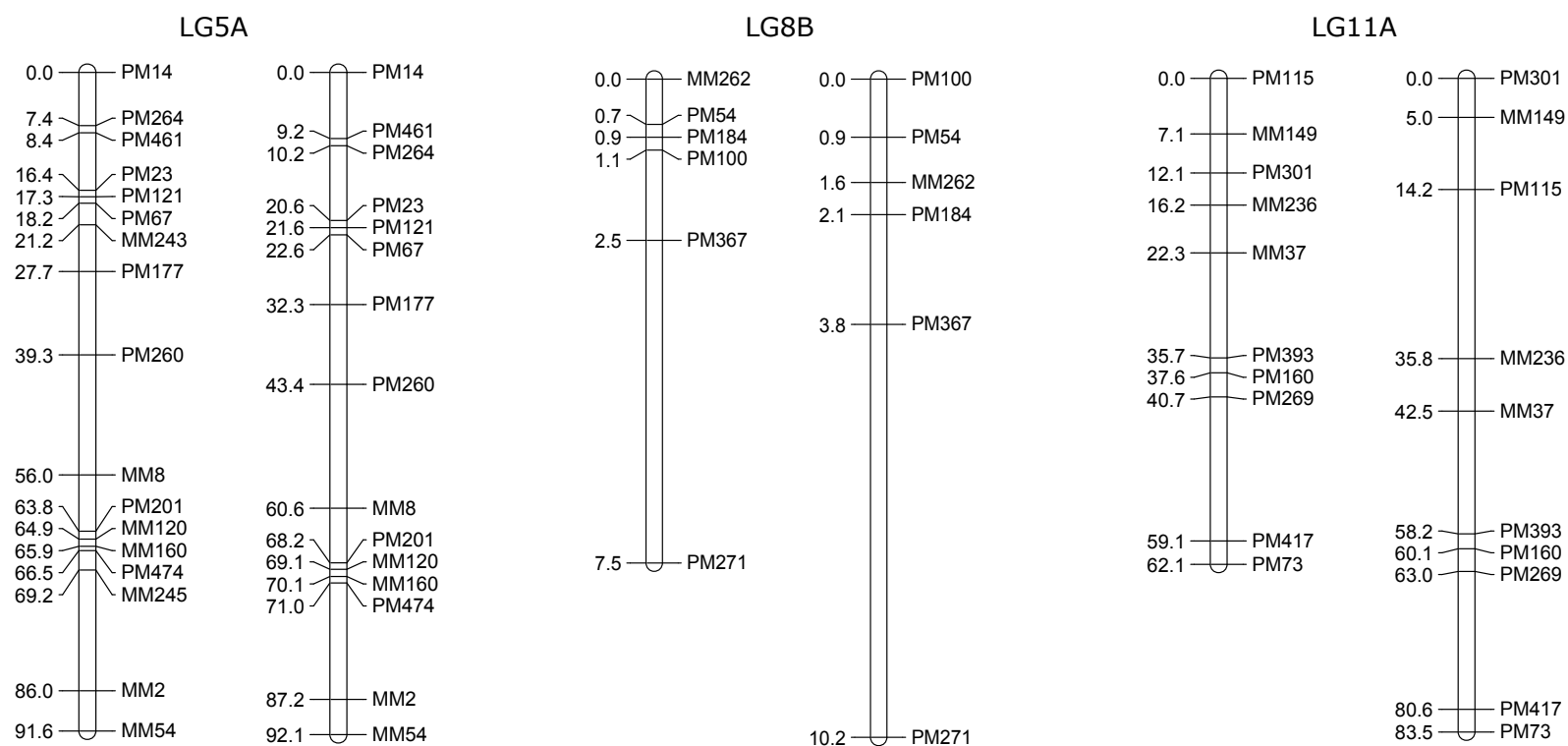
Appendix 9 The additive and dominance effects in the F_5 segregating population derived from the same cross between DipC and Tiga Nicuru.

Traits	QTL-LG	Position (cM)	LOD	PT	Additive effect	Dominance effect
DE	7	8.48	2.20	5.00	-0.05	-0.06
DF	8B	0.00	3.83	4.80	1.31	-1.15
EDP	8B	6.38	1.95	3.70	1.16	2.06
	11B	4.20	1.94	3.70	1.22	5.08
IN	1	54.73	7.28	3.70	-0.66	-0.07
PEL	1	54.73	9.52	3.70	-1.13	0.71
PN	1	53.73	3.94	3.80	-12.27	-20.32
PW	1	62.85	4.04	3.80	-10.00	-40.88
	2B	2.00	3.89	3.80	9.09	-41.03
	11A	46.89	4.16	3.80	10.46	-25.15
SN	1	53.73	4.82	3.70	-14.33	-19.93
SW	1	62.85	2.61	3.60	-5.81	-29.97
	2B	2.00	4.59	3.60	7.36	-42.87
	11A	46.89	3.60	3.60	7.07	-19.11
HSW	2B	4.00	3.14	3.90	5.79	-58.51
SDW	1	66.85	4.20	3.70	-8.90	-43.67
	11A	46.89	4.14	3.70	9.50	-22.68
HI	1	55.73	4.41	3.70	-0.12	-0.09
	2b	2.00	2.86	3.70	0.09	-1.13
RWC	4A	18.16	2.10	3.80	0.10	3.56
SC	2A	39.93	2.10	3.80	14.14	23.07
CID	2B	0.00	2.49	3.80	-0.39	2.29
SD	9	47.71	2.16	3.70	0.71	-6.55

ns: non-significance at p≤0.05 by permutation test using 10,000 reiterations.

p: putative QTLs whereby LOD score was lower than GW threshold by 0.1 to 1 interval.

PT: permutation test using 10,000 reiterations at p≤0.05.

*DF, days to flowering; IN, internode length; PEL, peduncle length; PN, pod number per plant; PW, pod weight per plant; SN, seed number per plant; SW, seed weight per plant; HSW, 100-seed weight; SDW: shoot dry weight; HI, harvest index.