Quantitative trait locus mapping of oil yield and oil quality related traits in the biofuel crop *Jatropha curcas*

Jasper Gregory Clarke, BSc with Honours of the 1st class

PhD

University of York

Biology

November 2016

Abstract

Jatropha curcas is a perennial shrub from the Euphorbiacea family. It is known for its stress resilience and high seed oil content, however little selective breeding has been carried out to fully domesticate this species. The aim of this project is to identify and map quantitative trait loci (QTL) for seed oil content, seed oil composition (oil quality), and oil yield, in order to identify loci suitable for introgression into an economically viable cultivar. In this study, an F_2 population (G51xCV) consisting of 229 plants for linkage analysis, and 145 plants for QTL analysis, was used to identify and position 312 genetic markers and 8 quantitative traits onto a genetic linkage and QTL map. Over 288 short sequence repeat (SSR) markers were mined from genome sequence to complement single nucleotide polymorphism (SNP) markers from genomic and transcribed DNA. 132 of the mined SSRs were physically linked to candidate genes, leading to the mapping of a substantial portion of genes that form the seed oil biosynthetic pathway in *Jatropha curcas*. Integration of phenotypic datasets collected over 2 independent years, enabled the identification of 15 QTL regulating seed oil content (2QTL), seed oil composition; palmitate, stearate, oleate, linoleate content (10 QTL), seed weight (1 QTL), number of branches (1 QTL) and seed yield (1QTL). Combined PVE for these QTL accounted for between 9.34 % (palmitate content year 2) to 32.26 % (seed oil content Year 2) of observed variation. Analysis of final oil yield per plant, showed that seed yield (number of seeds) was most important for regulating oil yield in this mapping population, however seed oil content and seed weight were also important traits, highlighting that selection of both seed oil and vegetative traits are of utmost importance for optimising oil yield in *Jatropha curcas*.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisor, Professor Graham, for his continued support and direction throughout my research project both during my time at the University of York and during the subsequent write up at the Royal Military Academy Sandhurst and beyond. I would like to thank him for providing a dynamic and challenging work environment that promoted attention to detail and excellence. I would like to thank Dr. Andrew King for technical expertise and knowledge specific to *J. curcas* and molecular biological research. I would like to thank Judith Mitchell, for her continued support throughout all stages of the project and particularly the administrative side of the project. Lastly I would like to thank members of the CNAP research group and fellow research students for providing a stimulating and interesting work environment.

Declaration

I declare that I am the sole author of this thesis and that it is a presentation of original work, except where due reference has been made in the text. This work has not been previously presented for an award, at this, or any other, University or institution. All sources are acknowledged as references.

Publications

King, A. J. Montes, L. R. Clarke, J. G. Affleck, J. Li, Y. Witsenboer, H. van der Vossen, E. van der Linde, P. Tripathi, Y. Tavares, E. Shukla, P. Rajasekaran, T. van Loo, E. N. Graham, I. A. (2013). Plant Biotechnology Journal, Vol 11, issue 8, p986. Linkage mapping in the oilseed crop Jatropha curcas L. reveals a locus controlling the biosynthesis of phorbol esters which cause seed toxicity.

Andrew J. King, Luis R. Montes, Jasper G. Clarke, Jose Itzep, Cesar A. A. Perez, Raymond E. E. Jongschaap, Richard G. F. Visser, Eibertus N. van Loo, Ian A. Graham. (2015). Biotechnology for Biofuels, Vol 8: 160. Identification of QTL markers contributing to plant growth, oil yield and fatty acid composition in the oilseed crop Jatropha curcas L.

Chapter 1: Introduction

1.1: Background and context – Population growth leads to challenges for food production, energy supply and climate change

The world population is currently 7 billion, and is expected to increase to 9 billion by 2050 (Godfray et al., 2010) and up to 12 billion by 2100 (Gerland et al., 2014). Such a large population places significant pressures on critical resources such as energy, land, food and water (Steinbuks and Hertel, 2016, Newbold et al., 2016, DeFries et al., 2015, Larsen et al., 2016, Jaramillo and Destouni, 2015, Fedoroff et al., 2010, Godfray et al., 2010, Tilman et al., 2009). At the same time, competition for these finite resources is exacerbated by climate change, which imposes constraints on how additional resources are made available to meet increased demand (Tester and Langridge, 2010).

Food production, the majority of which is derived from plants either directly or indirectly, will require large gains in global crop yields in order to feed the growing population (Tester and Langridge, 2010). Whilst historically crop yields have kept pace with population growth through technological advancement ('the green revolution') (Khush, 2001), optimisation of agricultural practices, refinement of genetics and intensification of farming means that in the developed world, a theoretical maximum yield per hectare is being approached using conventional technologies (Neumann et al., 2010), which places a premium on available agricultural land. In the developing world, adoption of better farming practices and improved crop cultivars may yield further gains, however there are other problems such as availability and access to chemical inputs, water, suitable land and permissive climatic conditions, that mean even with full utilisation of agricultural land for food production, meeting current and predicted food requirements will be a significant challenge (Godfray et al., 2010).

Climate change is expected to exacerbate this problem (Wheeler and von Braun, 2013, Tester and Langridge, 2010). Whilst modelling has shown some improved yields under different climate change scenarios at the regional level, overall, changes in climate are expected to decrease crop yields and available arable land; increasing arid land and desertification, and producing greater abiotic stress for plant growth in the form of unpredictable and more extreme weather conditions (Lobell et al., 2008). This will include drought conditions and semi-arid soils in many areas (Varshney et al., 2011).

One of the buffers to atmospheric CO_2 and climate change; plant biomes, means that expansion of agricultural land in pursuit of greater crop production is not recommended, as the negative effects on climate change will far outweigh any shorter-term yield increases (Steinbuks and Hertel, 2016, Newbold et al., 2016). Plant biomes also represent a rich resource of biodiversity that is vital for adapting to the unknown challenges of the future using untapped alleles, genes, germplasm, and natural products. As an example of this concept in practice, significant efforts are being made to expand the gene pool of highly domesticated crop species using wild genomes and landraces (Feuillet et al., 2008, Brozynska et al., 2016, Tester and Langridge, 2010, Zamir, 2001).

The majority of population growth will be seen in the developing world and the emerging economies, which places a significant demand on energy requirements (British Petroleum, 2016, Chu and Majumdar, 2012). Whilst fossil fuels have driven development for centuries, these are also finite resources and significant contributors to climate change through the release of greenhouse gases (Intergovernmental Panel on Climate

Change, 2014). The need to switch to renewable, low carbon energy sources in order to fuel development of emerging economies to more sustainable energy-use models and population demographics, as occurred during the industrial revolution in the developed world, is clear (Chu and Majumdar, 2012).

The cost of fossil fuels, such as crude oil, has risen steadily since their adoption and will continue to rise as sources become harder and more dangerous to extract (National Academy of Sciences, 2009, Kerr, 2011). Price fluctuations highlight the political aspect to supply and demand, and the need for energy security for many developed nations (National Academy of Sciences, 2009, US Energy Information Administration, 2016).

The majority of the infrastructure surrounding the world economy has been built up around liquid transportation fuels, such as petrol and diesel, and so there is an immediate requirement for renewable replacements (Blanch, 2010). For example, of the distillates of crude oil, which account for a third of global energy consumption (British Petroleum, 2016), 90 % is used for liquid transportation fuels, with the remaining 10 % as feedstocks for chemical manufacturing and other industrial processes (Dyer and Mullen, 2008). This includes the manufacture of important materials and chemicals, such as plastics, fertilisers and pesticides (Carlsson, 2009).

A number of renewable energy technologies are available to replace fossil fuels, and both the US and EU are pursuing a diverse renewable energy strategy including the use of nuclear, solar, wind, geothermal, hydroelectric and biomass technologies (The European Parliment, 2009, US Department of Energy, 2014). However biomass is the only renewable technology that can provide direct replacements for liquid transportation fuels (Blanch, 2010). Transition to different technologies for automotives e.g. electric, is expected to take some time, and for larger transportation systems, such as heavy-freight and aircraft, liquid transportation fuels are the only energetically-feasible energy source (Chu and Majumdar, 2012). Plant biomass is therefore an essential part of the renewable energy toolbox.

With increases in food crop production required to meet population demands, and the growing area not likely to expand due to the effects of climate change and biodiversity/ecosystem considerations, plant biomass feedstocks are required to minimise competition with food crops, particularly for agricultural land (Tilman et al., 2009).

Due to advances in genetics and plant breeding, such as the exponential rise in the power and accessibility of genome sequencing (Edwards and Batley, 2010, Davey et al., 2011, Feuillet et al., 2011, Langridge and Fleury, 2011, Morrell et al., 2012), novel plant-based solutions are now feasible. One approach that has become available is the rapid domestication of wild plant species (Langridge and Fleury, 2011). Whilst the creation of genetically improved cultivars traditionally took decades, often for small incremental improvements, with advanced breeding technologies and access to powerful sequencing and genomics technologies, this process can be significantly accelerated (Ragauskas et al., 2006). This opens up the repertoire of available plant species and germplasm from which to create new cultivars; expanding on the traditional high yielding annual crops that require high input farming, to intrinsically more efficient biomass options, such as perennials (Kantar et al., 2016, Fargione et al., 2008), that can tolerate a greater range of soil types, and use nutrients and water more efficiently (Tester and Langridge, 2010).

Quantitative Trait Locus (QTL) mapping is an approach that is greatly enhanced by genome sequencing technologies throughout its multistage process (Langridge and Fleury, 2011). The efficiency with which the genomic resources required for QTL mapping can be attained, and the repertoire of available techniques to mine and exploit novel genomes, substantially increases with access to genome sequence (Feuillet et al., 2011, Morrell et al., 2012). This includes the production of DNA markers, genetic linkage mapping, comparative mapping using synteny with sequenced relatives, the identification and mapping of candidate genes, and the delineation of QTL intervals.

Once genomic resources such as DNA markers, genetic linkage and Quantitative Trait Loci (QTL) maps are attained for novel crop species, advanced breeding technologies, such as Marker Assisted Selection (MAS), can be implemented to accelerate the breeding cycle by enabling larger populations to be screened at a much earlier stage for desirable genetics (Dekkers and Hospital, 2002). In this way the time taken to combine favourable QTL in a single cultivar is reduced, shortening the timeline between a wild plant and a genetically-improved cultivar.

Jatropha curcas, a perennial oilseed crop from the Euphorbiacea family, is a species that has generated interest for use as a biodiesel feedstock plant (Fairless, 2007). On the one hand it is a potentially high-value crop that combines a high seed oil content and valuable by-products, with significant plasticity to different soils, water and nutrient conditions. However, it is also a wild long-life perennial species that presents significant challenges to conventional selective breeding approaches; these challenges have thus far prevented domestication or genetic improvement of Jatropha. This project contributes towards the genomic resources required for the rapid domestication and genetic improvement of Jatropha, through QTL mapping of a number of oil yield and oil quality related traits. The DNA markers, genetic linkage maps, QTL maps and mapped candidate genes, will provide a basis for the breeding of improved varieties of *Jatropha curcas*.

1.2: The current state of plant-based biorenewable fuels

1.2.1: Global trends in energy consumption drive demand for oil and its derivatives

Over the past 50 years there has been a steady increase in world energy consumption from 3730.2 Mtoe (million tonnes of oil equivalent) in 1965, to 13,147.3 Mtoe in 2015 (British Petroleum, 2016). World energy consumption has increased faster than population growth, indicating that energy consumption has been driven by both population growth and a transition to more energy-intensive societies, as reflected by ever greater urbanisation and energy consumption per capita (British Petroleum, 2016, Food and Agriculture Organization of the United Nations, 2016).

Current (2015) figures show that energy consumption in OECD[1] countries accounts for 41.9 % of global energy consumption (5503 Mtoe); an increase of 1.18 % from 2000, whereas energy consumption in non-OECD countries accounts for 58.1 % of global energy consumption (7644 Mtoe); an increase of 93.57 % over 2000 usage (British Petroleum, 2016). With Non-OECD countries containing 82.6 % of the world's population, and the greatest population growth and economic development expected in these regions (World Bank, 2016), this trend is expected to continue, leading to a considerable increase in global energy demands over the coming decades (US Energy Information Administration, 2016).

World energy is supplied from 6 major sources; oil (32.9 %), natural gas (23.9 %), coal (29.2 %), nuclear energy (4.44 %), hydroelectric (6.79 %) and renewables (2.78 %) (British Petroleum, 2016). Of the three fossil fuel sources, oil, gas and coal, which together account for 86 % of global energy consumption, oil is most heavily used and least amenable to replacement by other fuel sources (Blanch, 2010).

Oil is fractionated into different hydrocarbons for different uses; approximately 90 % is used as liquid transportation fuels, and the remaining 10 % used as chemical feedstocks and for manufacturing (for example for the manufacture of plastics) (Dyer and Mullen, 2008, Carlsson, 2009). With the infrastructure of the world's economy built up around liquid transportation fuels (Chu and Majumdar, 2012), and the products of petro-chemical feedstocks vital for the functioning of society, demand for oil and its derivatives is expected to remain high.

1.2.2: Climate change is a leading driver for the switch to renewable energy sources

Anthropogenic emissions of greenhouse gases, predominantly the release of carbon dioxide from fossil fuel use, is widely accepted to contribute to climate change, and is a key driver for the switch to renewable energies (Intergovernmental Panel on Climate Change, 2014, US Energy Information Administration, 2016, The European Parliment, 2009). Carbon dioxide emissions from fossil fuel combustion and industrial processing alone accounted for 78 % of all greenhouse gas emissions between 1970-2010 (Intergovernmental Panel on Climate Change, 2014). The predicted impacts of climate change on food production, means that food security in particular provides significant incentive for the reduction of greenhouse gas emissions (Godfray et al., 2010).

[1] The Organisation for Economic Co-operation and Development (OECD) countries includes those in Europe and Australia, Canada, Chile, Israel, Japan, Mexico, New Zealand, South Korea, UK and US.

Effects of climate change on food crop production, which are pertinent to plant-based biorenewable fuels, include increasing global temperatures (Intergovernmental Panel on Climate Change, 2014), lower water availability (Larsen et al., 2016, Jaramillo and Destouni, 2015, Fedoroff et al., 2010), decreasing arable and agricultural land (Steinbuks and Hertel, 2016, DeFries et al., 2015), increasing arid and semi-arid land, more extreme and unpredictable weather conditions including droughts in many areas (Varshney et al., 2011), and these effects are thought to be proportional to the extent of climate change (Lobell et al., 2008). These effects point towards a different model of agricultural farming in the future, moving towards more efficient ways of farming with fewer inputs and under less favourable conditions (Fedoroff et al., 2010).

1.2.3: The economics of supply and demand present additional drivers for renewable replacements for oil

Despite advances in oil exploration and extraction technologies that have been able to meet increased energy demand with ever greater oil reserves and production rates (Chu and Majumdar, 2012, British Petroleum, 2016), the current reserve-to-production (R/P) ratio means on current consumption rates, world oil reserves will be exhausted within 50 years (British Petroleum, 2016). Similarly, there are concerns that the difficulty of extracting harder to reach reserves combined with ever increasing demand in the future, will lead to an oil production peak and an escalation of costs (Kerr, 2011). Oil consumption and oil production are geographically dislocated for many countries, leading to a reliance on oil imports and a susceptibility to geopolitically-caused price fluctuations (National Academy of Sciences, 2009, US Energy Information Administration, 2016). As a result, both energy supply and energy security provide significant additional drivers for renewable replacements for oil.

1.2.4: Plant-based biofuels offer the potential of a renewable, low-carbon alternative to petrochemical liquid transportation fuels

A diverse renewable energy portfolio is being established by both the US and the EU (The European Parliment, 2009, US Department of Energy, 2014) including the use of nuclear, solar, wind, geothermal, hydroelectric and biomass technologies. Despite this diversity of options however, biomass is the only renewable energy source that can be used as a direct replacement for liquid transportation fuels (Blanch, 2010). Liquid transportation fuels account for 90 % of oil use (Dyer and Mullen, 2008), over 14 % of annual greenhouse gas emissions (Intergovernmental Panel on Climate Change, 2014), and are the only energetically-feasible energy source to power critical infrastructure such as aviation and heavy-freight (National Academy of Sciences, 2009).

Plants are an attractive biomass feedstock due to their renewability and intrinsically low carbon footprint (Hill et al., 2006, Durrett et al., 2008). Plants accumulate biomass using atmospheric CO_2, water and sunlight over a short time scale, such that the carbon released through their subsequent combustion when used as a biofuel, should be less or equal to the amount of carbon fixed, with the additional lifecycle carbon-costs associated with their farming, processing and transportation (Hill et al., 2006). Plant-based biofuel feedstocks with low lifecycle greenhouse gas emissions, that minimise competition with food crops, are therefore the industry target (Tilman et al., 2009).

1.2.5: Research streams for exploitation of plant biomass for liquid transportation biofuels

Two major avenues of research currently exist for exploitation of plant biomass for liquid transportation biofuels (Guo et al., 2015). Bioethanol; produced from the fermentation of sugar to ethanol using carbohydrates derived from plant biomass, is the current biofuel of choice to replace petroleum. Bioethanol can be blended up to 10 % ethanol-to-petrol by volume, without the need to modify existing petrol combustion engines (Coyle, 2007). Biodiesel; used either as crude plant storage oil or more commonly manufactured by transesterification of plant triglycerides to simpler constituent fatty acids using methanol or ethanol, is chemically very similar to diesel and can be used as a complete replacement with little or no modification to diesel combustion engines[2] (Murugesan et al., 2009), although the most common blend in use in the US is a 20 % biodiesel-to-diesel by volume (Guo et al., 2015).

The two biofuels; bioethanol and biodiesel, can be further split into technical streams according to how their feedstock compounds are produced (Albers et al., 2016). The first stream for bioethanol production is the use of high starch/sugar-containing plants, such as sugarcane, that can be directly converted to ethanol through anaerobic fermentation. These are classed as 'first generation' biofuels, as typical high starch/sugar plant feedstocks are food crops, such as sugarcane, maize or corn.

The second bioethanol stream; cellulosic bioethanol, uses cellulose found in plant cell walls as a feedstock for producing carbohydrates from fibrous and woody plant biomass (Somerville et al., 2010). This process uses biological means to convert cellulose to ethanol, including hydrolytic enzymes to breakdown the cellulose matrix to simple sugars, and anaerobic fermentation to convert the liberated sugars to ethanol as before. Cellulosic bioethanol is a 'second generation' or 'advanced' biofuel as it uses non-edible plant biomass.

Finally thermochemical conversion of plant biomass, using pyrolysis and gasification reactions, is the third stream. This breaks down unrefined plant biomass using heat and pressure in the presence of specialised industrial catalysts, to synthesise a range of industrial chemicals including ethanol through non-biological means.

Biodiesel production occurs almost exclusively from oilseed crops, although a minor amount is converted from waste animal fat (Albers et al., 2016). The technical stream for biodiesel production, converts plant seed oil, which is a triglyceride, to biodiesel, which are single chain fatty acids, through transesterification with methanol, to yield biodiesel and the chemical by-product, glycerol. Biodiesel feedstock crops are classified as 'first generation' if the oilseed feedstock is an existing food crop, for example soybean, rapeseed or palm oil, or 'second generation' or 'advanced', if the oilseed feedstock is a non-edible or a non-food species.

[2] Biodiesels must meet the fuel standards ASTMD6751 (US) or EN14214 (EU) to be sold as pure biodiesels (B100), which is dependent on the biodiesel fatty acid composition (Murugesan et al., 2009, King et al., 2009). *Jatropha curcas* meets the US standard, and with most provenances the EU standard (King et al., 2009).

Critically, 'second generation' or 'advanced' oilseed and bioethanol crops, avoid competition with food crop production[3].

1.2.6: Current state of 'second generation' or 'advanced' plant-based biofuel feedstocks

Second generation biofuels seek to minimise competition with food crops in a number of ways (Ho et al., 2014). While non-edible crop sources were initially proposed as a way of reducing market pricing and production issues for crops that could be used for food *and* fuel (Graham-Rowe, 2011, Fairley, 2011), in many ways this was an over-simplification of the issue, since non-edible biofuel crops still fundamentally compete with food crops for land space, particularly if their market value per hectare is competitive with the food crops they displace.

The non-edible concept has been advanced to ways that more effectively avoid displacement of food crops. Current efforts are focused on using non-agricultural land for example marginal crop land, or low biodiversity semi-arid land for biofuel production (Tilman et al., 2009, Fargione et al., 2008). For this purpose perennials have a number of advantages for both bioethanol and biodiesel production; either as biomass accumulators or as oilseed crops respectively (Ragauskas et al., 2006, Somerville et al., 2010).

Another strategy is to use crop residues (the non-edible by-products of food crop harvests) as a source of lignocellulosic biomass (Sims et al., 2010). Research aims to improve the efficiency of all steps in the conversion process from lignocellulosic biomass to bioethanol (Sticklen, 2008, Blanch, 2010), from improving the intrinsic efficiency of biomass crops by making them perennial (Somerville et al., 2010), to engineering cellulose more amenable to breakdown by hydrolytic enzymes for example (National Academy of Sciences, 2009, Sticklen, 2008).

Oilseed crops for biodiesel production are typically enhanced by increasing both the oil yield and oil quality. Oil yield is a complex trait made up of a variety of component traits ('hierarchical' traits) (Alonso-Blanco and Mendez-Vigo, 2014) such as seed yield, seed oil content and seed mass. These contributing traits themselves may be determined by a number of other traits – for example in Jatropha, seed yield is affected by the amount of branching and ratio of female to male flowers, and other traits. Therefore the traits and genetic factors, or together the genetic architecture, controlling oil yield, are a diverse area of study and are often dependent, and specific to, the oilseed crop species, and the environment in which it will be grown.

Oil quality is linked to seed oil composition, a genetically controlled trait that determines the ratio of different fatty acids in the seed storage oil (Dyer and Mullen, 2008). Altering the ratio of fatty acids, affects the kinetic properties of the resultant biodiesel, including cetane number, cold flow point, and oxidative stability to name but a few (Knothe, 2009). Industry standards for each of these fuel properties determine the

[3] Whilst previously 'non-edible' was the definition of an advanced biofuel crop, it is now more accurate to define advanced biofuels as those that minimise competition with existing food crops, since this is the implied significance of the term 'non-edible', rather than the plant necessarily being inedible for human consumption. This is pertinent to *Jatropha curcas* since efforts are being made to breed non-toxic varieties, for the purpose of increasing the value of its seed meal for use as an animal feed, however even if it is technically 'edible' once made non-toxic, it is still classed as an advanced biofuel since it is not an existing food crop and it minimises competition with food crop production.

suitability of the biodiesel for different applications. With specific oil compositions achievable through genetic manipulation, the creation of 'designer' oils for industry is now an active area of research (Napier and Graham, 2010).

1.3: *Jatropha curcas*, a perennial oilseed of the Euphorbiacea family

Jatropha curcas (known as 'Jatropha') is a member of the Euphorbiacea family, which contain a number of agronomically-important species that are known to accumulate biomass efficiently; castor bean (Ricinus communis), rubber (Hevea brasiliensis), cassava (Manihot esculenta), sacha peanut (Plukenetia volubilis) and other less well-known oilseeds such as Chinese tallow (Triadica sebifera) and tung (Aleurites fordii) (Wu et al., 2015, King et al., 2013).

Jatropha is a small tree/shrub that grows to approximately 3-5 m (Heller, 1996), and up to 10 m (Divakara et al., 2010), although when pruned or managed for agricultural production, it can take a denser, more heavily branched morphology, that can be readily trained to a range of sizes (Gour, 2006). It is a 'hardy' and adaptable species, able to tolerate a wide range of soil types, day lengths and precipitation levels (Achten et al., 2010). Jatropha is a perennial plant, meaning that it grows year round, rather than being planted from seed each year as is required of annual crops. Once seedlings are established, Jatropha grows vegetatively for approximately 1-2 years, before it flowers and produces oilseed-containing fruits. In tropical regions where wet and dry seasons are observed, Jatropha typically flowers and produces seed twice a year; with vegetative growth occurring predominantly during the wet season (Heller, 1996). In permanently humid tropical regions, Jatropha flowers year round (Heller, 1996). Typically its first substantive yield is in the 2nd year of growth, although minor yields are sometimes reported in the 1st year of growth. Seed yield and oil yield continue to increase as it grows towards full maturity after ~5 years from seedling (Heller, 1996). Jatropha has a lifespan of up to 50 years (Heller, 1996). Jatropha is monoecious, asynchronous and self-compatible, leading to both crossing and self-fertilisation as reproductive strategies (Heller, 1996, Achten et al., 2010).

1.3.1: Jatropha distribution

Jatropha is a pan-tropical species originating from Meso-America (Heller, 1996, Achten et al., 2007). It has a wide geographic range and can be grown beyond the tropics of Cancer and Capricorn (latitudes of ±23.5^{o}N) (King et al., 2009), which is a greater range than the only major tropical-oilseed crop in cultivation; oil palm, which is distributed between ~±15^{o}N (Leff et al., 2004). It is reported to tolerate both dry and moisture-rich soils (Kumar and Sharma, 2008, Makkar and Becker, 2009), giving it significant climatic-plasticity, although prolonged or extremes of both, result in reduced growth and yields (Makkar and Becker, 2009, Abou Kheira and Atta, 2009, Edrisi et al., 2015).

1.3.2: *Jatropha curcas* genetics

Jatropha curcas is a diploid species with 2n=22 chromosomes, and a genome size of C=416 Mb (Carvalho et al., 2008). This is classed as a very small plant genome (Michael, 2014), and is small when compared to other species within the same phylum, order or family (Angiosperms/Malpighiales/Euphorbiacea respectively) (Bennett and Leitch, 2012). Of the crop species, only rice (389 Mb), cucumber (367 Mb), peach (220 Mb), orange (367 Mb), papayas (372 Mb) have smaller genomes that have been sequenced (Goodstein et al., 2012, Feuillet et al., 2011). Jatropha has a GC content of 38 % which is typical of dicots and similar to the model

organism Arabidopsis (*Arabidopsis thaliana*), which facilitates genome assembly and annotation (Carvalho et al., 2008).

Jatropha's nearest sequenced relative is castor bean, which has a similar genome size (C=509 Mb) (Bennett and Leitch, 2012) and a high-quality open-access genome sequence (Chan et al., 2010). A high level of synteny and gene co-linearity exists between castor bean and Jatropha, as has been demonstrated during genome sequence assembly (Wu et al., 2015, Sato et al., 2011) and genetic linkage mapping (King et al., 2013). This is pertinent to Jatropha research, as the Jatropha genome sequence has, until recently, been available only at the low-quality draft level[4], although recently this has been improved to a level comparable to that of castor bean (Wu et al., 2015).

Genetic diversity is particularly low for Jatropha germplasm distributed globally (Yue et al., 2014, He et al., 2011, Montes Osorio et al., 2014, Pecina-Quintero et al., 2014, Qi-Bao Sun et al., 2008). Genetic characterisation and diversity studies suggest that the majority of Jatropha material found outside of Meso-America, is descended from a narrow subsection of the Jatropha gene pool (He et al., 2011), as a result of limited sampling during its introduction to global trade markets in the 16th century (Heller, 1996). A tendency for inbreeding in *Jatropha curcas*, due to self-compatibility and monoecious flowering, is also thought to contribute to genetic homogenisation of populations (Achten et al., 2010). Discovery of germplasm with greater genetic and phenotypic diversity in recent times (He et al., 2011, Montes Osorio et al., 2014, Pecina-Quintero et al., 2014), has overcome a major bottleneck in the breeding of improved varieties of Jatropha.

1.3.3: Genomic resources for Jatropha research

Genomic resources for Jatropha research have increased considerably following the advancement of genome sequencing technologies. Early research assessed genetic diversity of Jatropha germplasm using molecular genetic approaches, identifying suitable breeding material and greater genetic diversity in its centre of origin (Yue et al., 2014, Montes Osorio et al., 2014, Pecina-Quintero et al., 2014, He et al., 2011, Achten et al., 2010, Graham, 2009, Basha et al., 2009, Sun et al., 2008). Following advances in genome sequencing, a number of transcriptome sequencing studies were published using different sequencing platforms; dye-terminator capillary sequencing (Costa et al., 2010); 454 pyrosequencing (King et al., 2011), and the FLX titanium platform for 454 pyrosequencing (Natarajan and Parani, 2011), enabling gene discovery to occur in key tissues such as the developing seed. This was followed shortly afterwards by the publication of a draft

[4] For the castor bean genome sequence published in Sept 2010 (Chan et al., 2010), mean and median (N50) scaffold lengths were 93 kb and 561.4 kb respectively, for a gene density (total sequence span/number of gene models) of 11,220 bp/gene, meaning that the number of genes per scaffold was relatively high. For the Jatropha genome sequence, mean and median (N50) sequence element lengths were 1,900 bp, and 3,833 bp respectively, for release JAT_r3.0 (Sato et al., 2011). Mean and median (N50) sequence element lengths were marginally increased to 7,597 bp, and 15,950 bp respectively, for release JAT_r4.5 (Hirakawa et al., 2012). With a gene density (total sequence/number of gene models) of 9,855 bp/gene for release JAT_r4.5, the number of genes per sequence element is low, highlighting the utility of the castor bean genome sequence for physically ordering the smaller sequence elements of the Jatropha genome sequence within syntenous regions. The most recent Jatropha genome sequence, published in 2015, generated mean and median (N50) scaffold lengths of 168 kb and 746 kb respectively (Wu et al., 2015), substantially increasing the quality of available genome sequence.

genome sequence (Hirakawa et al., 2012, Sato et al., 2011), significantly enhancing DNA marker and gene discovery. An interspecific cross with *j. integerrima*, led to an interspecific draft genetic linkage map for Jatropha (Wang et al., 2011), and a number of QTL and eQTL (Liu et al., 2011, Sun et al., 2012). Data from this thesis study contributed to a collaborative project that published the first high density genetic linkage map for *Jatropha curcas*, using intraspecific genetic diversity identified from its centre of origin (King et al., 2013). This was followed by publication of an integrated QTL map and updated genetic linkage map (King et al., 2015). A vastly improved Jatropha genome sequence was published in 2015 (Wu et al., 2015), increasing average contig and scaffold size, and integrating previously published linkage maps.

1.3.4: *Jatropha curcas* as a valuable biodiesel feedstock crop

Jatropha curcas has several characteristics that make it an attractive feedstock crop for biodiesel production, as an animal feed, for medicinal/pharmacological natural products, and for soil improvement/land reclamation/utilisation of marginal land (Abhilash et al., 2011, Achten et al., 2007, Becker and Makkar, 2008, Devappa et al., 2010, Devappa et al., 2013, Divakara et al., 2010, Heller, 1996, King et al., 2009, Kumar and Sharma, 2008, Makkar and Becker, 2009, Openshaw, 2000, Thomas et al., 2008, Sabandar et al., 2013, Mukherjee et al., 2011).

Jatropha curcas is an oilseed crop that produces seeds of between 30-40 % seed oil content (Achten et al., 2010). Jatropha oil has a favourable 'high-oleic' seed oil composition, making it suitable for use as biodiesel (Knothe, 2009, Durrett et al., 2008). Transesterification of crude Jatropha oil produces a biodiesel that meets both European and US fuel standards (King et al., 2009), that can be used in diesel combustion engines without further modification. Reported Jatropha oil yields vary considerably, but can be up to 2000kg/ha/year (Yue et al., 2013), which are yields from an undomesticated lineage with little genetic improvement. This compares to 3680kg/ha/year for oil palm, and 360kg/ha/year for soybean (Gupta, 2015), as examples of commercial, high yielding oilseed crop cultivars that have undergone considerable genetic improvement.

The seed meal of Jatropha; a by-product of the seed oil extraction process, is high in protein and is suitable for use as an animal feed. The presence of toxins in non-edible varieties; namely curcin and phorbal esters, requires treatment of the seed meal before it can be ingested (Aregheore et al., 2003). Curcin is readily broken down by heat treatment, however phorbal esters are more recalcitrant to detoxification methods (Kumar and Sharma, 2008, King et al., 2009), and are a known purgative, co-carcinogen and a handling risk for agricultural cultivation (Makkar et al., 2011, Aregheore et al., 2003, Makkar et al., 1998, Makkar et al., 1997, King et al., 2009). Edible varieties that do not contain phorbal esters are present in Meso-America (Pecina-Quintero et al., 2014, He et al., 2011, Makkar and Becker, 2009, Makkar et al., 1998), and so identification of the QTL for phorbal ester production is the focus of genetic research (King et al., 2013) (although beyond the scope of this thesis study). Creation of high yielding, phorbal-free cultivars will enhance economical cultivation of Jatropha, and enable it to be used as a multipurpose feedstock crop for fuel and animal feed, following similar models used for soybean cultivation (Cromwell, 2012, Food and Agriculture Organization of the United Nations, 2016).

Jatropha curcas has also long been used as a medicinal crop by the indigenous people of Meso-America (Heller, 1996). Like other species within the Euphorbiacea (such as castor bean), many parts of the plant have biological activity (Hecker, 1968, Ernst et al., 2015, Evans and Taylor, 1983). This has generated interest in using Jatropha as a feedstock for bioactive compounds or as a source of natural products for the

pharmaceutical industry (Devappa et al., 2010, Thomas et al., 2008, Sabandar et al., 2013). Studies demonstrate the anti-bacterial, anti-mollusc, anti-fungal, purgative, and latent-HIV stimulating (Wender et al., 2008) activities of compounds found in Jatropha (Kumar and Sharma, 2008).

1.3.5: Jatropha as a perennial, intrinsically stress-tolerant species

Jatropha is a perennial crop, and there are significant benefits associated with a permanent and deep root system (Becker and Makkar, 2008, King et al., 2009, Kantar et al., 2016). Perennial root systems fix more carbon in the soil, improve soil structure and retain water, nitrogen and other beneficial soil components by establishment of an extensive and robust physical root structure (Cox et al., 2006, Jerry D. Glover, 2007). Over time this improves the soil and surrounding land by increasing carbon content, locking in more water and nutrients, and increasing both above- and below-ground biodiversity, in comparison to annual crops (Cox et al., 2006, Jerry D. Glover, 2007, Kantar et al., 2016), or when planted on marginal or arid land (Cox et al., 2006, Jerry D. Glover, 2007, Becker and Makkar, 2008).

Perennial growth and a deep root system confers significant abiotic stress tolerance to Jatropha, particularly drought and low nutrient conditions, since it is able to reach deeper parts of the soil and use water and nutrients more efficiently (King et al., 2009). This characteristic makes Jatropha compatible with the 'state-of-the-art' for plant biofuels, since it conforms to two recognised strategies to avoid competition with food crops; it can be grown on degraded/non-agricultural land, and it is suitable for mixed/double cropping systems (for example as a border plant) (Tilman et al., 2009). The use of Jatropha as a border plant for example, is well documented due to its anti-herbivory properties for grazing livestock (Heller, 1996).

Its ability to grow in, and improve degraded or arid land, with little external inputs makes it a viable option for developing world agricultural systems, to utilise land that is otherwise dormant and low in biodiversity (Makkar and Becker, 2009).

As with all known perennial species, little or no genetic improvement has been applied to domesticate or create cultivars (Kantar et al., 2016). For oilseed perennials such as Jatropha, the requirements for partitioning of energy between vegetative growth and seed production is likely to be very different for a wild species compared to a crop cultivar (Van Tassel et al., 2010, Cox et al., 2006, Kantar et al., 2016, Jerry D. Glover, 2007), therefore seed and oil yield related traits are expected to have significant scope for genetic improvement. (Kantar et al., 2016)

1.3.6: Challenges for the development of Jatropha as a biodiesel feedstock

Most of the challenges for development of Jatropha as a biofuel feedstock, stem from the fact that Jatropha is a wild species that has not been through any stringent selective breeding. At present oil yields from Jatropha are sub-optimal and highly variable (Yue et al., 2013).

1.3.6.1: Challenges to conventional selective breeding

Despite the value of Jatropha being known for some time as shown in historical accounts of Jatropha use (Heller, 1996), this lack of genetic improvement is a problem commonly associated with perennial species (Kantar et al., 2016). Whilst annual species are suited to conventional selective breeding; they complete a full lifecycle within a short period of time and each year new seed must be selected and sown for the next

generation, the very characteristics that make perennial species prolific biomass accumulators and intrinsically stress tolerant, also make it difficult for conventional selective breeding to occur (Cox et al., 2006). Jatropha has a long lifespan (~50 years) and reaches maturity after ~5 years. Cross breeding is challenging as Jatropha is self-compatible and monoecious; self-fertilisation is a frequent event and hard to detect without genetic characterisation (as a result harvested seed is often a mix of cross- and self-pollinated seed -discussed in results chapter 4). Direct-domestication using advanced breeding technologies such as marker assisted selection (MAS), is a recognised strategy for the rapid domestication of wild perennials (Kantar et al., 2016), and is particularly pertinent to *Jatropha curcas*, due its generation time (9 months seedling to seed), time to maturity (~5 years), and self-compatible reproductive strategy.

Further factors confound conventional selective breeding approaches in Jatropha. The majority of material found outside its centre of origin is genetically very similar, to the point where it has been described as almost clonal (He et al., 2011, Montes Osorio et al., 2014, King et al., 2015). With some conventional selective breeding efforts in Jatropha lacking genetic characterisation of starting material, and subsequent selection of plants occurring at the phenotypic rather than genotypic level (He, 2011, Sato et al., 2011), a great deal of observed variation in Jatropha material outside its centre of origin is thought to be of epigenetic origin (Yi et al., 2010), rather than being underpinned by stable genetics suitable for selective breeding. This lack of genetic variation in starting material, and lack of adequate breeding technologies to assess and guide selection based on genetics, has until recently hampered genetic improvement of *Jatropha curcas*.

1.3.6.2: Target traits for the genetic improvement of *J. curcas*

With the implementation of an advanced breeding technology, such as MAS, to overcome the challenges of conventional selective breeding in Jatropha, a number of target traits are amenable to genetic improvement for the economic cultivation of Jatropha as a biodiesel feedstock. For oilseed crop species, traits affecting both oil yield and oil quality are of fundamental importance.

For a trait to be amenable to genetic improvement there must be phenotypic variation in the trait of interest across the population, and this variation must have a heritable component. Phenotypic variation suggests plasticity in the trait of interest that could be optimised towards certain values within its distribution (typically towards the high or low phenotypic values within its normal, bell-shaped distribution), and its heritability suggests a genetic component that could be selected for reproducible phenotypic effects that are heritable. A number of traits meet these criteria in Jatropha.

1.3.6.2.1: Oil yield related traits

Oil yield per plant[5] for an oilseed crop such as Jatropha, may be split into the component traits; *oil yield per seed*, and, *seed yield* (the number of seeds produced per plant). *Oil yield per seed* and *seed yield* themselves may be split into further component traits.

Oil yield per seed is the product of; *seed oil content* (% of oil per seed), and *seed mass* (average mass of each seed). Both *seed oil content* and *seed mass* are traits that show significant variation and heritability between

[5] Oil yield *per hectare*, as a measure of oilseed crop performance, is dependent on additional agronomy factors such as plant spacing, soil and growth management (Achten et al., 2010), which is beyond the scope of this thesis study. Hence *oil yield per plant* is used to investigate oil yield.

Jatropha lines (Achten et al., 2010), suggesting a genetic component to observed variation and scope for optimisation through selection. Increasing seed oil content not only increases the proportion of oil produced per seed, but could also increase the efficiency of oil extraction from harvested seed using mechanical or chemical extraction methods. Seed mass may increase the oil yielded per seed if oil content as a proportion does not decrease, or conversely, if seed mass does not increase and seed oil content increases, more resources could be seen to be partitioned into seed storage oil over other seed components such as proteins or carbohydrates. This interplay between seed mass and seed oil content determines the oil yielded per seed, and so these are both important seed traits for optimisation. Due to the economic and industrial value of oilseed crops, oilseed metabolism, particularly seed fatty acid biosynthesis, is an extensively studied area of research. The identity and function of key metabolic players and their corresponding genes are well known, particularly within the sequenced oilseed model species *Arabidopsis thaliana* (Li-Beisson et al., 2013). In the age of comparative genomics and crop genome sequencing, *in silico* elucidation of the fatty acid biosynthetic pathway and relevant candidate genes for oil quantity and quality traits, is both a recognised strategy for the improvement of novel biofuel cultivars (Vega-Sanchez and Ronald, 2010), and more specifically, readily-achievable for *Jatropha curcas* due to the availability of reference genome sequence (Sato et al., 2011, Wu et al., 2015), transcriptomics (King et al., 2011, Costa et al., 2010), and closely related sequenced model and crop species (Chan et al., 2010, Li-Beisson et al., 2013).

Seed yield is a highly variable trait in current Jatropha material, and is therefore a key target trait for optimisation of oil yields (Achten et al., 2010). This is a hypothesis supported by current thinking on resource partitioning of undomesticated perennial oilseed crops and the likely scope they possess for genetic improvement (Kantar et al., 2016). Seed yield is a complex trait that encompasses a variety of component vegetative- and plant architecture-related traits (King et al., 2009).

Two key traits that are thought to affect seed yield are the ratio of female to male flowers and the extent of branching (Achten et al., 2010). Jatropha is monoecious; it produces both male and female flowers, and the ratio of female to male flowers varies between lines and under different environmental conditions (Fresnedo-Ramirez, 2013, Luo et al., 2007, Wu et al., 2011). Female flowers, once fertilised, produce the oilseed-containing fruits. Therefore it has been hypothesised that the ratio of female to male flowers is one way in which seed yield can be increased in Jatropha (Fresnedo-Ramirez, 2013, Divakara et al., 2010, Mukherjee et al., 2011, Achten et al., 2010, King et al., 2009). Flowering is known to be highly dependent on environmental conditions, for example some species flower in response to stress (Wada et al., 2010), and the effects of exogenous application of plant signalling hormones on flowering in Jatropha (Makwana et al., 2010, Pan and Xu, 2011, Gargi Joshi, 2011), suggest a strong interplay with environmentally-regulated signalling mechanisms. Whilst the relative contribution of the environmental component of flower ratio variation (the E of GxE) is still to be determined (Achten et al., 2010), it is possible that genetic variation may be present that could modulate this response in naturally occurring populations.

Flower inflorescences occur at terminal and auxiliary nodes (ends of branches and branch points respectively) (Fresnedo-Ramirez, 2013, Luo et al., 2007, Wu et al., 2011), therefore the extent of branching is also thought to be a key trait that regulates seed yield in Jatropha. Due to the known effects of branching on yields of many agronomic and commercial crops (Wang and Li, 2006, Zhang et al., 2013), branching is a relatively well studied trait and the identity and function of a number of key gene classes has been elucidated (Domagalska and Leyser, 2011, Wang and Li, 2008, Ongaro and Leyser, 2008). Increasing the number of

branches increases the number of available flowering points, which, particularly if the ratio of female to male flowers remains favourable, could significantly increase the seed yield in Jatropha (Achten et al., 2010).

Other vegetative traits are likely to regulate seed yield (although beyond the scope of this thesis study). These include traits related to plant stature, biomass accumulation, and 'vigour' (often exploited in crop breeding as 'hybrid vigour'). For inbreeding and homogenised populations such as those found for Jatropha outside its centre of origin, cross breeding of genetically distinct lines to stimulate hybrid vigour could be a productive approach (Achten et al., 2010), since material from such populations are ideal candidates for releasing hybrid vigour.

The two components of oil yield per plant; seed oil yield (seed oil content, seed mass), and seed yield (including branching and flower ratio traits), can be viewed as two distinct areas of plant metabolism and development. Seed oil yield is dependent on seed metabolism ('the seed fatty acid biosynthetic pathway'), whereas seed yield is more directly associated with vegetative and architecture-related traits. A key output of this thesis study will be determining the relative importance of optimising seed oil yield compared to vegetative traits that regulate overall seed yield, for improving overall oil yield in *Jatropha curcas*.

1.3.6.2.2: Oil quality related traits

In addition to oil yield, oil quality is an important trait for biodiesel production. For oilseed crops, fatty acid composition; the ratio of different fatty acids in the seed storage oil, determines the kinetic properties of the resultant biodiesel (Knothe, 2009, Durrett et al., 2008). Due to the economic and industrial importance of oil quality (fatty acid composition) for fuels, chemical feedstocks and in health and nutrition, fatty acid composition has been extensively studied (Knothe, 2009, Cahoon et al., 2007, Durrett et al., 2008). The kinetic properties and biodiesel performance of differing fatty acids (Knothe, 2005, Atabani et al., 2013), and the genes responsible for regulating fatty acid compositions in plants (Li-Beisson et al., 2013), are two areas that have been extensively studied.

Key kinetic properties of biodiesels include the cetane number (a measure of explosiveness), the cold flow and cloud point (viscosity at low temperature and precipitation point respectively; the effective operating temperature of the biodiesel and its suitability to different climates), and its oxidative stability (the rate at which the biodiesel oxidises and degrades) (Knothe, 2009). Studies into these properties have found that the level of desaturation of fatty acids (including the number of desaturated bonds in each fatty acid, and the mix of saturated/desaturated fatty acids) is the most critical property for biofuel kinetics (Knothe, 2009, Durrett et al., 2008). Saturated fatty acids have a higher energy content and are less reactive to oxygen (and therefore have favourable cetane numbers and oxidative stability), but are more dense and viscous and therefore have poor cold flow and cloud point characteristics. Conversely, polyunsaturated fatty acids have better cold flow properties, but have less favourable cetane numbers and oxidative stability. Current research suggests that the optimal compromise between these properties are fuels high in mono-unsaturated fatty acids (oleic-acid, 18:1 and palmitoleic-acid, 16:1) (Knothe, 2009).

The genetic basis of fatty acid composition, follows the same seed fatty acid biosynthetic pathway as that for seed oil yield, as modification reactions such as desaturation, occur as part of the pathway between *de novo* fatty acid synthesis in the plastid, and the final deposition of fatty acids as triglycerides in the seed storage oil (Li-Beisson et al., 2013). Whilst the individual steps and the genes responsible for this core metabolic

pathway are highly conserved across all oilseed crop species (Li-Beisson et al., 2013), significant inter- and intra-species variation in seed fatty acid composition is present throughout nature and in response to differing environments (Canvin, 1965, Voelker and Kinney, 2001, Flagella et al., 2002); highlighting the plasticity, and potential to manipulative and adapt this pathway to achieve specific fatty acid compositions. As a concept and approach this has been proved extensively in research (Napier et al., 2014, Napier and Graham, 2010, Bates et al., 2013, Durrett et al., 2008), and through the commercialisation of crops with altered fatty acid compositions (Burton et al., 2004).

Although Jatropha seed oil meets quality standards for both EU and US fuel markets, there is also reported plasticity in seed oil composition, for example in response to differing temperatures (King et al., 2009), suggesting that optimisation of seed oil composition and therefore oil quality, would be useful in Jatropha, particularly for material that may be grown under differing environmental conditions. Similarly, due to the importance of the seed fatty acid biosynthetic pathway for oil yield and oil quality in oilseed crops such as Jatropha, genetic mapping of candidate genes within this pathway would be a valuable genomic resource for QTL mapping and genetic improvement of *Jatropha curcas*.

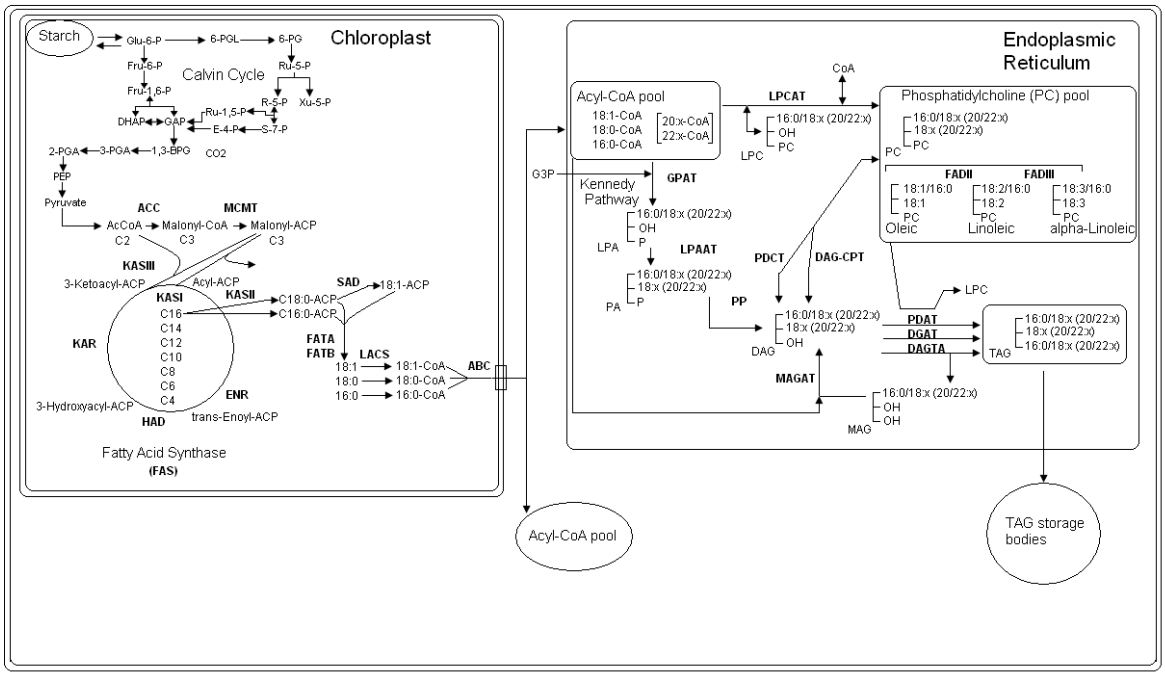


Figure 1-1 The major steps fatty acid synthesis in seed storage oil in plants.
Adapted from 'The Arabidopsis Book' (Li-Beisson et al., 2013, Meng et al., 2013). An in depth analysis of candidate genes for seed oil yield and seed oil composition, using this pathway will be discussed in Chapter 3.

1.4: Marker Assisted Selection as a technology for the rapid domestication and accelerated breeding of *Jatropha curcas* varieties

Marker assisted selection (Dekkers and Hospital, 2002), is a technology that has been proposed for the genetic improvement of *Jatropha curcas*, since it enables screening and selection of genotypes harbouring beneficial QTL at the seedling stage based on genotype. This overcomes many of the challenges with breeding improved Jatropha varieties: (1) offspring resulting from the intended cross (ie a true cross or selfing event) can be identified through genotyping, (2) selection is based on seedling genotype, eliminating

the difficulties of inferring genotype from phenotype, which may be harder to detect or quantify, and that are subject to environmental effects (Dekkers and Hospital, 2002) (3) plants containing desirable genetics can be identified at the seedling stage; eliminating the time needed for phenotypes to be expressed which can be several years with Jatropha. Selection at the seedling stage enables larger populations to be created and screened, allowing rarer, more desirable genetic events to be selected, for example the inheritance of multiple QTL alleles, which in turn reduces the number of generations required to breed the desired genotype (Dekkers and Hospital, 2002). In this way MAS can significantly accelerate the breeding process.

For such an approach, genetically- and phenotypically-diverse Jatropha lines are required from which to breed from, and genomic resources are required to inform the selection process. Genetic and phenotypic diversity issues have recently been overcome by identification of diverse germplasm in the place of origin of Jatropha in Meso-America (Montes Osorio et al., 2014, Pecina-Quintero et al., 2014, He et al., 2011). The genomic resources required for Marker Assisted Selection include DNA markers, genetic linkage maps and QTL maps, that together provide information on target genomic regions ('loci') and the beneficial or unwanted alleles within them, and the markers necessary to track their movement and transmission through breeding populations (Dekkers and Hospital, 2002). Genomic resources for Jatropha have improved recently, and several groups have begun QTL mapping studies (King et al., 2015, Sun et al., 2012, Liu et al., 2011). Public dissemination of the results of such projects will significantly facilitate the creation of genetically improved Jatropha cultivars.

1.5: The Quantitative Trait Loci (QTL) mapping process

The function of QTL mapping is to understand the genetic basis of simple and complex traits in a population or family (Mackay et al., 2009). QTL mapping associates genotype with phenotypic variation, in order to determine the regions of the genome ('loci') and the genetic variants contained within them ('alleles') that are responsible for observed variation in a quantitative trait, using natural or experimental populations. A variety of QTL mapping approaches, population structures and methods of statistical analysis, have been developed with differing advantages and limitations (Wurschum, 2012, Staub et al., 1996, Doerge, 2002, Morrell et al., 2012). The biparental, F_2 mapping population is an approach that is particularly useful for investigating traits of interest with pre-identified variation in two distinct parental lines, for example a high-oil and low-oil line.

In a biparental mapping population, the two parental lines that differ phenotypically for the trait of interest are crossed to establish a mapping population. The parental lines are ideally genetically distinct and homozygous at all loci, to enable genomic regions from each parent to be tracked and differentiated throughout the mapping population, and to ensure observed phenotypic differences have a genetic basis. Crossing of two homozygous parents, creates a heterozygous, genetically-uniform F_1 population, with 1 allele at each locus (in a diploid species such as Jatropha) originating from each parent. Selfing or crossing of an F_1 plant, creates an F_2 population consisting of plants with a genetic mosaic of alleles from each parent, due to meiotic recombination in F_1 gametes. Genetic diversity of the F_2 population is used to inform both recombination rates of loci for genetic linkage mapping, and linkage or association of particular loci and alleles (genotypes) to particular phenotypic values for QTL mapping. As a result, F_2 population size, along with marker density (Morrell et al., 2012) and appropriate and accurate phenotyping (Alonso-Blanco and Mendez-Vigo, 2014), determines the power to detect and locate QTL in this approach.

Several key stages are present in the QTL mapping approach after selection of suitable parental material and creation of a mapping population:

1) The development of DNA markers
2) Genotyping of the F_2 population
3) Genetic linkage mapping
4) Collection of phenotypic data
5) QTL mapping

1.6: Aims of the study

1.6.1: The development of SSR markers

SSR markers are a class of co-dominant marker that are hypervariable, abundant (Schlotterer, 2004, Agarwal et al., 2008) and identifiable *in-silico* using a reference genome sequence and search algorithms (Sato et al., 2011, Stieneke, 2007, Martins et al., 2009). As such SSRs are ideal for marking specific regions of the genome in a targeted manner.
The aim of this part of the study was to develop SSR markers to complement existing genome-wide non-selective markers (discussed in detail in chapters 3 and 4), for the purpose of gap filling during later-round genetic linkage mapping, and to mark identified candidate genes for relevant traits of interest.

1.6.2: Identification and mapping of candidate genes for oil yield and oil quality related traits

The principle mapping population under study in this thesis, G51xCV, was created from parental lines with reported variation in oil yield and oil quality related traits; including seed oil content, seed oil composition, seed mass, seed yield and branching. The relevance of oil yield- and oil quality-related traits, to the genetic improvement of the oilseed crop, *Jatropha curcas*, makes the identification and mapping of candidate genes for these traits a useful genomic resource for QTL mapping in this study and others.

1.6.3: Genetic linkage mapping in the G51xCV F_2 mapping population

Genotyping and genetic linkage mapping of the F_2 population, is an integral part of the QTL mapping process. Accurate ordering and positioning of genetic markers on a genetic linkage map, facilitates the detection and location of QTL during subsequent QTL mapping (Doerge, 2002). QTL can be associated with genetic intervals rather than individual markers (‘interval mapping’ versus ‘single marker analysis’) (Doerge, 2002). Positioning of QTL between flanking markers, and within confidence intervals, increases the accuracy and utility of identified QTL for crop breeding (and also other applications), in comparison to QTL associated with single markers, unless the single marker is completely linked and the QTL is monogenic (Mackay et al., 2009).

A key aim of genetic linkage mapping in the G51xCV mapping population, was to contribute data towards the first intraspecific genetic linkage map published for *Jatropha curcas* (King et al., 2013), thereby establishing a key genomic resource for Jatropha development. Genotyping and genetic linkage mapping in the G51xCV mapping population was based on analysis of 229 F_2 plants.

1.6.4: Seed-related phenotyping

G51xCV F_2 plants were phenotyped for the seed related traits; seed oil content, seed mass and seed oil composition, using Nuclear Magnetic Resonance (NMR) spectroscopy and Fatty Acid Methyl Ester (FAME) gas chromatography (see materials and methods). Seed oil composition was determined by measurement of the 4 major fatty acid moieties in *Jatropha curcas* seed oil; palmitate, stearate, oleate, linoleate. Seed related phenotyping occurred for 3 datasets collected over 2 years.

1.6.5: QTL mapping of oil yield, and oil quality related traits

QTL analysis of seed oil content, seed oil composition (palmitate, stearate, oleate, linoleate content), seed mass, seed yield, and branching traits was conducted through integration of genotypic and phenotypic datasets, using interval mapping and single marker analysis. Correlation and statistical analysis of traits was conducted to determine causative relationships and interactions. The aim was to determine: (1) the presence and location of QTL responsible for regulating oil yield and oil quality related traits in the G51xCV population; (2) to provide data on their relative contribution towards phenotype; (3) to determine the QTL parent of origin; (4) to determine their mode of action (dominance/semi-dominance/recessive/over-dominant); (5) to determine hierarchical and/or causative relationships between traits; (6) to determine the relative contribution of component traits to complex traits such as overall oil yield in the G51xCV population.

Chapter 2: Materials and Methods

2.1: The collaborative Jatropha project and contributions of this thesis study

Table 2-1 The Jatropha project structure at the University of York

Four F_2 mapping populations were used for genetic linkage and QTL mapping, with specific quantitative traits under study in each population. A combined dataset from all four mapping populations was used to create a combined genetic linkage map, and phenotypic data from two populations used to create an integrated QTL map. The principle focus of this thesis study has been highlighted in red: (1) The G51xCV mapping population from DNA marker generation onwards (SSR DNA marker generation, genotyping, phenotypic data collection, and QTL mapping) (2) SSR DNA marker generation across all 4 populations (3) Contributing data for the combined genetic linkage map and integrated QTL map (marker generation all populations; G51xCV genotyping data, phenotypic data and QTL analysis) (King et al., 2013, King et al., 2015).

Mapping Population	**G51xCV**	**G33xG43**	**QV01**	**QV02**	**Combined (for linkage mapping)**
Primary Trait Variation	Oil content, oil composition. Branching Flower ratio	Toxicity	Oil Content, oil composition	Oil Content, Oil composition	n/a
Population size (plants)	229	320	220	220	989
Mapping site	Guatemala	Guatemala	Cape Verde	Cape Verde	n/a
Mapping Population Responsibility	Biocombustibles de Guatemala, S.A	Biocombustibles de Guatemala, S.A.	Quinvita	Quinvita	Authors listed left (cols 2-5)
DNA marker generation	KeyGene JG Clarke AJ King	KeyGene AJ King JG Clarke	KeyGene AJ King JG Clarke	KeyGene AJ King JG Clarke	Authors listed left (cols 2-5)
Data Collection; Genotyping	KeyGene JG Clarke	KeyGene AJ King	KeyGene J Affleck AJ King	KeyGene J Affleck AJ King	Authors listed left (cols 2-5)
Data Collection; Phenotyping	JG Clarke *(seed oil content, oil composition, seed mass)* L Montes, Biocombustibles de Guatemala, S.A *(vegetative traits)*	AJ King L Montes	Quinvita	Quinvita	n/a
Genetic Analysis	JG Clarke	AJ King	AJ King	AJ King	AJ King
QTL Analysis	JG Clarke	AJ King			

The G51xCV mapping population, consisting of 229 F2 plants, was the principle population under study for this thesis. SSR DNA markers generated as part of this thesis study were tested across all 4 mapping populations for a number of reasons: (1) to increase the chance of the marker being mapped in the combined

map, as many of these markers were linked to candidate genes (2) to increase the available recombination data for markers that could be mapped in multiple populations (3) to enable independent mapping population maps to be aligned using shared markers for comparative mapping (4) to increase the number of markers for QTL mapping in each population.

Table 2-2 DNA markers available for genetic linkage and QTL mapping

Table 2-2 lists the type and number of markers that were mapped during genetic linkage and QTL mapping for each mapping population (for the total number of markers produced for this thesis i.e. including markers that may not have been polymorphic or mapped in these populations, see chapter 3). The outputs of this thesis study have been highlighted in red: (a) The production of SSR markers for all 4 populations (b) genotyping of markers in the G51xCV mapping population (except genome sequence SNPs which were genotyped by Keygene as part of the CRoPS marker discovery process) (van Orsouw et al., 2007) and (c) genetic linkage mapping of all markers in the G51xCV population. Not shown here is phenotypic trait data collection and QTL mapping, that was conducted as part of this thesis study in G51xCV, as specified in Table 2-1. The authors responsible for creating each marker type are listed in column 3.

Mapping Population			**G51xCV**	**G33xG43**	**QV01**	**QV02**	**Combined**
Marker type & Author	Genome Sequence SNPs	KeyGene	**181**	**161**	**287**	**283**	**318**
	Simple Sequence Repeats (SSRs)	AJ King	48	87	49	48	129
		JG Clarke	62	18	9	9	74
		R Santos	7	7	4	3	10
		Total	**117**	**112**	**62**	**60**	**213**
	EST-derived SNPs	AJ King	**14**	**30**	**32**	**35**	**58**
	Total		**312**	**303**	**381**	**378**	**594**

Table 2-2 lists the final contribution of this thesis work to genetic linkage mapping within the collaborative project. Genotyping and genetic linkage mapping for this study occurred for all marker types in the G51xCV population, except the genotyping of genome sequence SNP's which was carried out by KeyGene as part of their SNP marker discovery process. In total, genotyping occurred for 131 markers in 229 F2 plants in the G51xCV mapping population, and genetic linkage mapping with 312 markers. SSR marker mining was conducted across all 4 independent populations; contributing between 9 (QV01 and QV02) and 62 (G51xCV) SSR markers in individual maps for use in genetic linkage and QTL mapping, and 74 SSR markers for the combined genetic linkage map.

Table 2-3 Phenotypic trait data collection in the G51xCV mapping population

All phenotypic traits and corresponding datasets listed below were subject to QTL analysis as part of this thesis study. Data collection for these traits was split according to the authors specified in Table 2-1. The traits and datasets highlighted in red were collected at the University of York as part of this thesis study: (1) seed oil content (3 datasets); (2) seed oil composition (1 dataset); (3) seed mass (3 datasets). The vegetative traits, branching and seed yield, were collected by collaborators at the mapping population site (Guatemala), as listed in Table 2-1.

Year	**2011**	**2012**					**2013**				
Date	13th Dec	26th Jun	13th Sep	1st Oct	12th Oct	15th Oct	10th Jan	22nd May	28th May	16th Aug	14th Oct
Years of growth	1.76	2.30	2.51	2.56	2.59	2.6	2.84	3.2	3.22	3.44	3.60
Days after transplanting	567	763	842	860	871	874	961	1093	1099	1179	1238
Measurements taken in the field	Branching	Seed harvest, Branching	Seed harvest	Seed harvest	Seed harvest	Seed harvest	Seed harvest	Seed harvest	Seed harvest	Seed harvest	Seed harvest
Sample period for seed sent to York		Batch 1						Batch 2		Batch 3	
Trait	**Dataset name and sample period**										
Seed oil content		Year 2						Year 3a		Year 3b	
Seed oil composition		Year 2									
Seed mass		Year 2						Year 3a		Year 3b	
Branching	Year 1	Year 2									
Seed Yield		Year 2					Year 3				

Table 2-3 shows the phenotypic data collected for this thesis study. Field-based vegetative traits were collected by collaborators, along with other vegetative traits of interest that were studied outside of this work. This thesis study collected 7 phenotypic datasets from 3 traits at the University of York, from biological samples collected on the dates listed above. Data from all traits and datasets listed above was subject to QTL analysis for this thesis study.

2.2: The G51xCV F_2 mapping population

The G51xCV F_2 mapping population was grown at [13°57'33.17"N and 90°23'21.89"W], Guatemala. F_2 plants were planted at a density of one plant every 4 x 2 m, or 1,250 plants per hectare. Juvenile plants were transplanted from nursery to the field on 25 May 2010, during the rainy season. Drip irrigation was applied during the dry season (November to April), and fertilisers applied according to the nutritional requirements of the plants in conjunction with soil analysis (King et al., 2015). In total, 229 F_2 plants were available for

genetic linkage analysis, and the largest sub-population that experienced the same environmental conditions, which consisted of 145 F_2 plants, was selected for QTL analysis.

2.3: Parental lines and population structure

The G51xCV F_2 mapping population was created by crossing a fully homozygous line, 'Cape Verde' (CV), with a heterogeneously heterozygous line, 'G51'. The G51 and CV lines were selected primarily on the basis of their seed oil content; G51 at 36.90 % seed oil content, and CV at 26.00 % seed oil content. In addition to seed oil content, F_2 plants were also screened for variation in a number of other traits. This study investigated: (1) seed oil content (2) seed oil composition (3) seed mass (4) number of branches and (5) seed yield (number of seeds).

DNA marker analysis showed that G51 was in the region of 36.5 % heterozygous, based on the DNA markers used in this study. In order to maximise the number of informative markers at heterozygous loci in G51, an F_1 sibcross rather than F_1 self was used to create the F_2 population. Two F_2 populations were created in consecutive years to maximise the number of F_2 plants available for genetic linkage mapping. In total 229 F_2 plants were available for genetic linkage analysis. The larger of the two populations consisting of 145 F_2 plants was used to map quantitative trait loci. Whilst an F_1 sibcross was the primary crossing strategy, due to the self-compatibility of Jatropha and its monoecious, asynchronous flowering strategy, F_1 selfing was also present, leading to further subpopulations within the F_2 population. See figure 2-1 – this population structure and the use of informative markers is explored in greater detail in the linkage mapping results chapter.

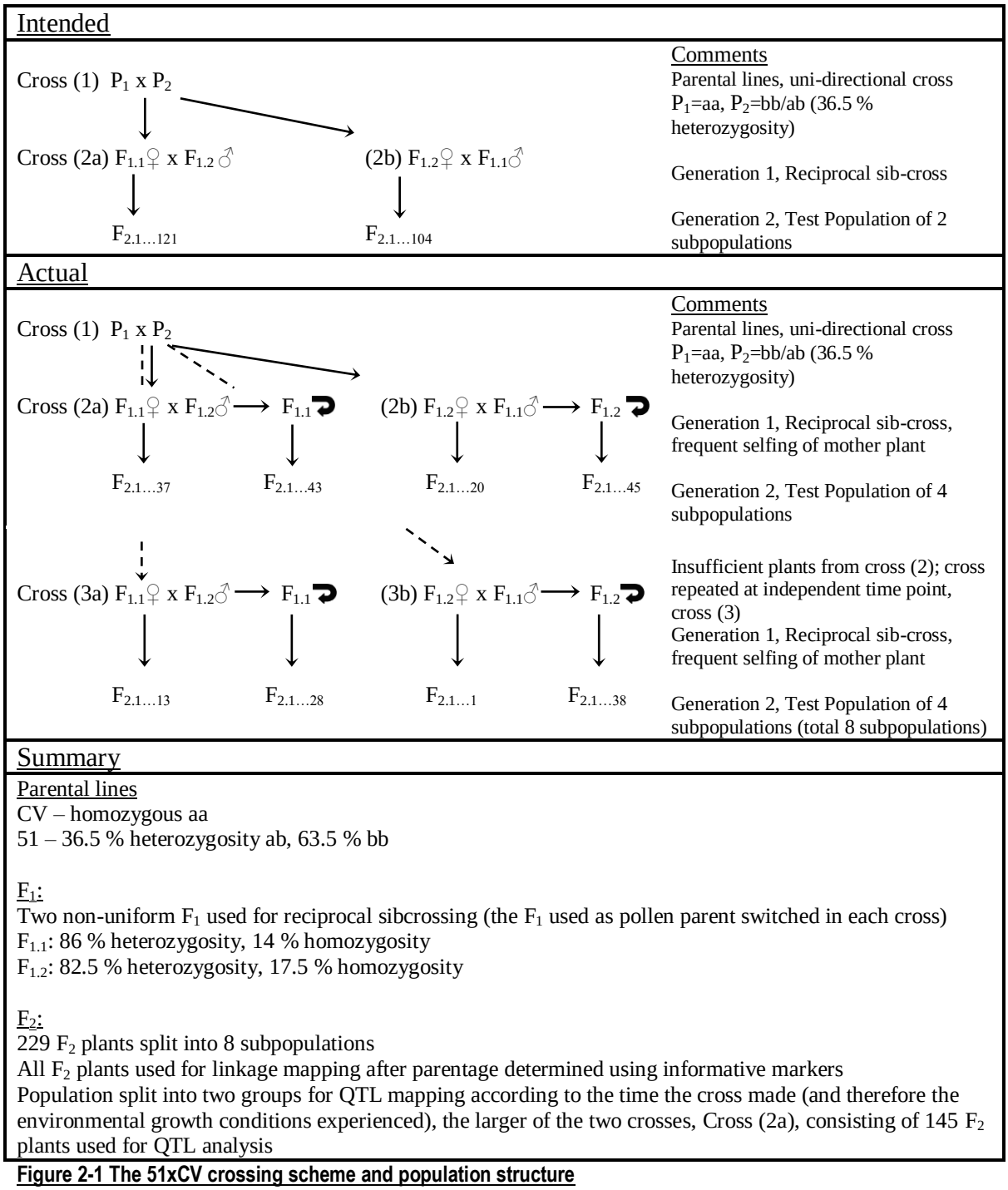
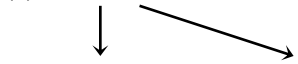


Figure 2-1 The 51xCV crossing scheme and population structure

The above diagram shows both the intended and actual population structure for the G51xCV mapping population as determined through informative marker analysis.

This complex population structure was determined through informative marker analysis; a process developed as part of this thesis study for determining the parentage (a true cross or selfing event) of F_2 plants. Figure 2-2, explains what informative markers are and how they can be used in this context.

(1) P_1aa x P_2 ab



(2) $F_{1.1}$ ab x $F_{1.2}$ aa or $F_{1.1}$ aa x $F_{1.2}$ ab

F_2 progeny

F_1 Cross:

	a	b
a	aa	ab
a	aa	ab

1:1 'aa', 'ab'

$F_{1.x}$ (ab) selfing:

	a	B
A	aa	Ab
B	ab	Bb

1:2:1 'aa', 'ab', **'bb'**

(3) Number of informative markers available to detect an F_1 selfing:

	Inf marker class A (52 markers), and F_1 genotypes			Inf marker class B (41 markers) , and F_1 genotypes		
F_1 plant	M1a	…	M52a	M1b	…	M41b
$F_{1.1}$	ab	…	ab	Aa	…	aa
$F_{1.2}$	aa	…	aa	Ab	…	ab

(3a) Detecting $F_{1.1}$ selfing (direction 1)

In the F_2 generation; the presence of informative 'bb' genotypes at markers 1a to 51a, and all 'aa' at markers 1b to 41b.

(3b) Detecting $F_{1.2}$ selfing (direction 2)

In the F_2 generation; the presence of informative 'bb' genotypes at markers 1b to 41b, and all 'aa' at markers 1a to 52a.

(3c) Detecting true cross

In the F_2 generation; a lack of 'bb' scores and a 1:1 mixture of 'aa' and 'ab' genotypes at both informative marker classes (rather than all 'aa' in one class).

Comments

(1) Informative markers begin as heterozygous loci in the G51(P_2) parent.

(2) Loci that are homozygous in one F_1 and heterozygous in the other can be informative for determining F_2 parentage

In the F_2 the presence of informative 'bb' scores indicates an F_1 selfing event

(3) There were 52 informative markers for determining selfing of $F_{1.1}$ and 41 markers for selfing of $F_{1.2:}$

(a) Detecting $F_{1.1}$ selfing (direction 1)

In the F_2 generation informative markers 1a to 52a have a 1 in 4 chance of being informative 'bb' genotypes. Informative markers 1b to 41b will all be 'aa' genotypes. Expected number of 'bb' alleles at markers 1a to 52a; 52*0.25=13'bb' scores.
A threshold of 3 or more 'bb' alleles used to assign selfing event in this dataset (with M1b to 41b all 'aa'). If cross, markers 1b to 41b; expected 0.5*41=20 'ab' scores rather than all 'aa'.

(b) Detecting $F_{1.2}$ selfing (direction 2)

In the F_2 generation informative markers 1b to 41b have a 1 in 4 chance of being informative 'bb' genotypes. Informative markers 1a to 52a will all be 'aa' genotypes. Expected number of 'bb' alleles at markers 1b to 41b; 41*0.25=10'bb' scores.
Again the threshold value of 3 or more 'bb' alleles (and all 'aa' genotypes at other marker class) is well below the expected number. If cross, expected number of 'ab' scores at M1a to 52a; 52*0.5= 26 'ab' scores rather than all 'aa'.

(c) Detecting true cross

The 'bb' score is impossible to obtain via a true cross, and between 10-13 are expected for either F_1 selfing. There is a very low probability of getting all 'aa' at the other class of informative marker since 50 % are expected to be 'ab' by chance (between 20-26 expected 'ab' scores for each cross direction).

Figure 2-2 Informative Markers available for assigning F_2 parentage in the G51xCV mapping population

The left column explains what informative loci are, and the alleles and genotypes expected at these loci in the mapping population. The right column explains how these genotype frequencies were used to determine parentage of F2 plants.

2.4: DNA extraction

Dry leaf tissue was transported from the mapping population site to the University of York on silica gel desiccant. Between 10-20 mg of dried tissue was taken for DNA extraction using the Qiagen DNEasy Plant Mini kit (Qiagen, Venlo, the Netherlands), according to the supplied protocol. DNA was eluted and stored in Qiagen AE Buffer (10 mM Tris, 0.5 mM EDTA, pH9). DNA was quantified using the DNA binding dye EvaGreen (Biotium, Hayward, CA), using salmon sperm DNA as a standard (Wang et al., 2006). DNA was transferred to 96-well plates and diluted to working DNA concentrations (2-10 ug/μl).

2.5: DNA markers

This project utilised Single Nucleotide Polymorphisms (SNPs) and Simple Sequence Repeat (SSR) markers. SNPs were mined using a reduced-representation genome-sequencing technique; Complexity Reduction of Polymorphic Sequences (CRoPS©) (van Orsouw et al., 2007) and through comparative sequencing of cDNA libraries (King et al., 2011). SSRs were mined from publically available genome sequence (Sato et al., 2011).

2.5.1: Single Nucleotide Polymorphism (SNP) markers

SNP markers obtained from the reduced representation genome sequencing strategy; Complexity Reduction of Polymorphic Sequences (CRoPs©) (van Orsouw et al., 2007) were scored using the Illumina VeraCode Assay, a high-throughput plate based assay; work carried out by Keygene. A number of SNP markers were obtained by pyrosequencing of developing seed tissue (cDNA sequencing) (King et al., 2011), and mining of publically available sequence; work carried out by Dr. Andrew King. These SNPs were scored using KASPar and allele specific PCR amplification systems (Cuppen, 2007, Bui and Liu, 2009) and an ABI3730 capillary sequencer which can also analyse fragment sizes.

SNPs identified by the CRoPs© technique were expected to be randomly distributed throughout the Jatropha genome. SNPs identified by cDNA sequencing were expected to be randomly distributed throughout transcribed DNA sequence.

2.5.2: SSR markers

SSR markers were mined from the publically available Jatropha genome sequence published in 2011 (Sato et al., 2011). Overall, the SSR mining process consisted of 4 phases.

Phase 1 was the identification of target *J. curcas* genome sequence contigs. **Phase 2** was the identification of SSRs within target contigs. **Phase 3** was amplification of SSR sites. **Phase 4** was scoring of the amplified SSRs, leading to either confirmation of polymorphism in parental lines and progression to linkage mapping, or elimination of the SSR as a marker.

SSR mining from the published reference genome sequence enabled a targeted rational approach to marker generation, which complemented the wide coverage, but less specific, distribution of SNP markers used in this project. SSR markers were used for (1) the genetic and physical mapping of candidate genes and investigation of QTL (2) the development of markers to fill gaps in the genetic linkage map following earlier rounds of genetic linkage mapping.

The 4 phases of SSR marker development is described in more detail below.

2.5.2.1: Phase 1: Identification of target genome sequence contigs

Phase 1 is split into two parallel processes, candidate gene mapping (1a) or gap filling (1b), that identify target contigs by two different approaches. After identification of the target contig, the two parallel approaches converge to follow the same phase 2 and onwards, as described below.

2.5.2.1.1: Phase 1 (a) Candidate genes

Lists of candidate genes for each trait were compiled from genomic resources such as The Arabidopsis Book (http://www.thearabidopsisbook.org)(Li-Beisson et al., 2013), the Plant Metabolic Network (http://pmn.plantcyc.org/), KEGG PATHWAY Database (http://www.genome.jp/kegg/pathway.html), Gramene (http://pathway.gramene.org/ARA/), Biocyc (http://biocyc.org/ARA/), in combination with a literature review of published genes studied in both model and crop species. Gene sequences were isolated from the databases Arabidopsis.org (http://www.arabidopsis.org) for Arabidopsis genes, and GenBank (www.ncbi.nlm.nih.gov/genbank/) for genes published in academic journals. Peptide sequences were used to search the *J. curcas* genome sequence (Sato et al., 2011) for gene homologues, using the BLAST algorithm at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and the BLAST algorithm in the *J. curcas* genome sequence viewer at the Kazusa DNA research institute webpage (http://www.kazusa.or.jp/jatropha/). All Jatropha genome sequence contigs that contained gene homologues with high sequence similarity to the reference candidate gene were scanned for SSRs.

Following successful linkage mapping of *J. curcas* candidate genes identified by this method, evidence to support gene functionality/activity was generated by BLAST searching specific nucleotide data repositories of NCBI:

1. Nucleotide Collection (nr/nt) for sequenced and characterised *J. curcas* genes
2. Expressed Sequence Tags (EST) and Transcriptome Shotgun Assembly (TSA) for *J. curcas* $_{m}$RNA submissions for evidence of expression

2.5.2.1.2: Phase 1 (b) Gap filling during linkage mapping using comparative mapping and castor bean microsynteny

Additional mapping was carried out to reduce gaps on the linkage map that were larger than 15 cM, or to reduce the QTL interval for high significance QTL.

Linkage maps from 4 Jatropha mapping populations were physically aligned *in silico* using shared markers (defined as a single marker that was independently positioned in more than one mapping population linkage maps). Markers in one map that corresponded to gaps in another linkage map could then be used to locate *J. curcas* genome sequence contigs in the target region, and the contig sequence mined for polymorphic SSRs as described in **Phase 2**.

After this approach had been completed, castor microsynteny was utilised as a way to reach remaining gaps. To identify syntenous regions, *J. curcas* transcribed amino acid gene sequences from contigs mapped by polymorphic DNA markers were blasted against the castor genome to find the most similar gene homologues, using BLASTP. This was repeated for all genes on mapped *J. curcas* contigs. Castor bean genes mapping to the same contig and in the same order suggested a syntenous region with gene co-linearity.

Where neighbouring *J. curcas* map positions aligned to the same castor contig, the region between neighbouring positions could be assumed to be syntenous. Castor bean genes in this adjoining region could then be used to as probes to search for homologous transcribed amino acid gene sequences in the *J. curcas* genome sequence, and retrieved contigs could then be mined for additional SSR markers as described in **Phase 2**. A graphical representation of this strategy is presented below.

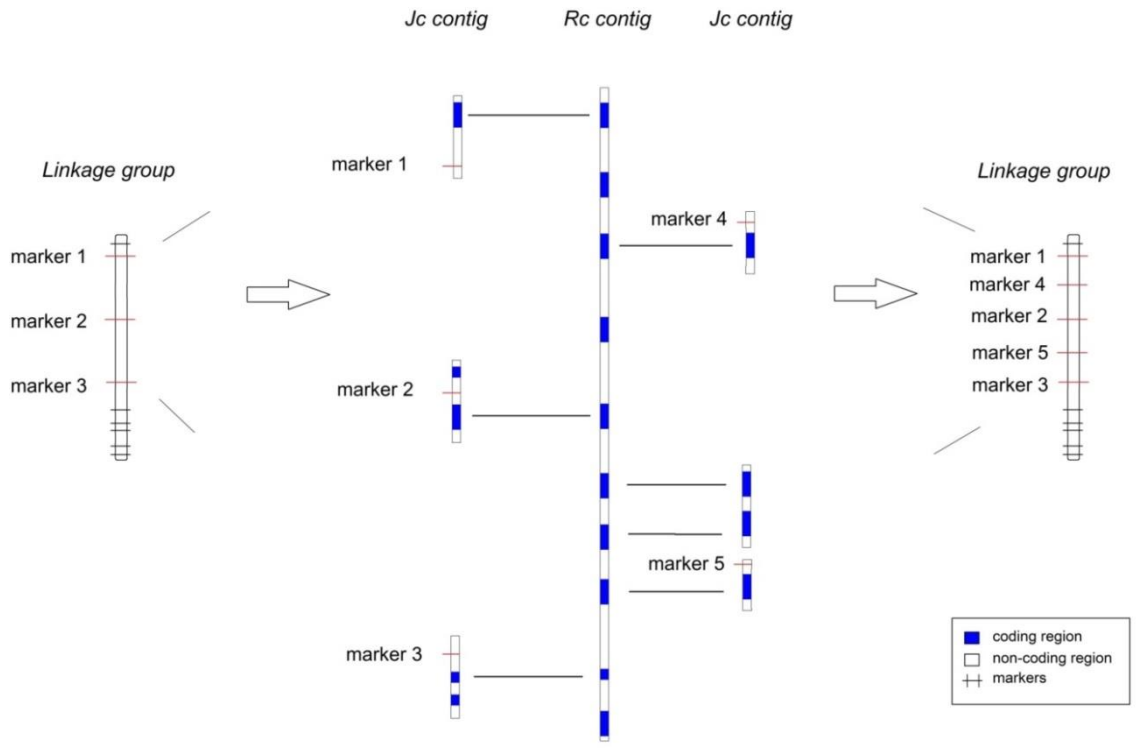


Figure 2-3 Diagrammatic representation of interspecific comparative mapping, conducted between *J. curcas* and *R. communis* genomes

The process of comparative mapping to generate markers in gaps of the *J. curcas* linkage map. From left to right: a *J. curcas* linkage group requiring further mapping; *J. curcas* genome sequence contigs retrieved using markers flanking the target regions; *R. communis* genome sequence contig retrieved using transcribed gene models found on *J. curcas* contigs (note *J. curcas* contigs map to the same *R. communis* contig -synteny); *J. curcas* contigs corresponding to the target region retrieved using *R. communis* transcribed gene models; SSRs found on retrieved *J. curcas* contigs enable markers to be mapped in the linkage group target regions.

2.5.2.2: Phase 2: the identification of SSRs within target contigs

Target *J. curcas* contigs were scanned for SSRs using web software, Websat (Martins et al., 2009) and ImperfectSSR (Stieneke, 2007). For SSR's of (x)n, where x is the length of the repeat motif, and n is the number of repeats of that sequence, stringency was set at $n \geq (12,6,6,5,4)$ number of repeats, for x= (2,3,4,5,6) repeat motif size respectively. Primer3 (Untergasser et al., 2012) software was used to design flanking PCR primers. In order to allow multiplexing, all PCR primers were designed with a T_m of 55-60 ^oC according the nearest-neighbour method (SantaLucia, 1998). The primers were designed so that amplicon sizes were between 80 to 450 bp, suitable for fragment analysis by capillary electrophoresis. A standardised

M13 18 bp sequence [5′-TGTAAAACGACGGCCAGT-3′] was appended to the 5' end of the shortest of the two primers to allow fluorescent labelling during the PCR reaction (Hayden et al., 2008).

2.5.2.3: Phase 3: the amplification of SSRs

SSR marker PCR primers were tested on 4 mapping population parental accessions to increase the chance of the marker loci being mapped in a combined genetic linkage map and to facilitate QTL mapping across all populations. Polymorphic SSR primers were multiplexed according to product size with a >10 bp difference between primer product sizes. Multiplexes contained between 5-10 primers.

Multiplexed PCR reactions were performed on the mapping population DNA using the QIAGEN Type-it Microsatellite PCR kit. Reaction components included: the PCR reagents; the primer multiplex at 0.5 µM; either M13 VIC or FAM fluorescent dye at 2 µM; and F_2 DNA at 10-20 µg/µl; in 96 well plate format.

PCR cycling conditions were:
1) Initial denaturation at 95 ^{o}C for 5 mins
2) 32 cycles of:
    (a) denaturation at 95 ^{o}C for 30 s
    (b) annealing at 55 ^{o}C for 1 min 30 s
    (c) elongation at 72 ^{o}C for 30 s;
3) Final elongation at 60 ^{o}C for 30 mins.

PCR plates were subject to a serial dilution equivalent to 100X dilution. 2 µl of diluted PCR product was combined with 9 µl Hi-Di Formamide for analysis using the Applied Biosystems (ABI) 3730 DNA Analyzer (Life Technologies). GeneScan 500 LIZ Size standard was used as an internal ladder. The standard plate injection time was 15 s.

2.5.2.4: Phase 4: scoring of the amplified SSRs, leading to confirmation of polymorphism, or elimination of the SSR marker

ABI3730 data was exported to Genemarker software (SoftGenetics, LLC, CA, USA) for scoring. Initial automatic scoring was run with a scoring window of 80-450 bp and threshold intensity of >500. The panel editor function was used to manually assign allele positions before re-running the scoring with the adjusted allele positions. Following automatic scoring each allele position was individually checked for each plant in the mapping population, due to complexities with scoring SSRs such as complex patterning and the generation of PCR artefacts (Schlotterer, 2004).

2.6: F2 Genotype/marker score processing and analysis

2.6.1: Assignment of parentage

Comparison of F_2 and F_1 informative marker scores allowed F_2 parentage to be assigned as either the product of an F_1 intercross or an F_1 selfing event. The original crossing scheme was an F_1 sibcross however *J. curcas* is self-compatible leading to frequent selfing events under field conditions; see results section chapter 4 for a full analysis of the F_2 population structure as determined by informative marker scores.

2.7: Linkage mapping

2.7.1: Assignment of markers to linkage groups by Two Point Linkage Analysis

The genotype/ F_2 matrix was converted into Crimap (Green, 1990) compatible files. Two point linkage analysis was carried out to assign markers into linkage groups. A variety of LOD thresholds values were tested to see what effect this had on the grouping of markers. The greatest LOD (logarithm of odds) threshold value that produced 11 linkage groups, as supported by cytological evidence on *J. curcas* chromosome number (Carvalho et al., 2008), that also linked the majority of markers, was used to maximise stringency of linkage between markers (to avoid false linkages) whilst producing the correct number of linkage groups. Markers were then separated into separate linkage group Crimap files for map building.

2.7.2: Linkage group mapping

Each linkage group was built independently according to the following process.

The crimap 'build' function was used to build the linkage group, starting with the two most likely markers and adding consecutive markers in order of likelihood. The build analysis output file was scanned for markers listed as completely linked to eachother and for markers that had more than one possible position. The highest likelihood score was used to assign markers where they had more than one possible position, and, where a marker had a possible position either side of another marker with equal likelihood, that marker was assigned as completely linked to that marker.

After all such markers were assigned to positions the 'build' function was re-run, specifying the known order and inputting completely linked markers. Following this, the 'flips' function was run, which switches around (or 'flips') a set number of markers (2,3,4 or 5) giving a likelihood score of the new order. Any orders that were more likely than the previous arrangement were taken as the true order, and the 'build' function rerun to apply these changes.

Following a flips analysis where the order given was most likely, the 'chrompic' function was run to look at crossovers across the linkage group for each F_2 plant. The chrompic output file for each plant is a series of 0's and 1's, or o's and i's, with each integer specifying which parent the genotype score at that locus came from. 0's and 1's represent one haplotype of a parent, o's and i's the other haplotype. In this way crossover and recombination events can be visualised for each F_2 plant.

Any positions with a double crossover across a single locus was highlighted as a potential error as the likelihood of such an event occurring by chance was low for closely linked markers (<15 cM), due to crossover interference (Ooijen et al., 2013). Marker scores obtained by capillary electrophoresis were rechecked by examining ABI3730 traces in GeneMarker. Incorrect scores were corrected , or where ambiguous, the score for that marker left as undetermined. After each round of chrompic, the crimap genotype and linkage group files were updated and the linkage mapping process repeated until a consensus map emerged.

2.7.3: χ^2 Segregation Distortion Analysis

F_2 plants were separated into sub populations according to assigned parentage in order to carry out a χ^2 test for marker segregation distortion. This involved calculating the expected number of plants for each genotype based on Mendelian genotype ratios, before using a χ^2 test to compare the observed number of plants to the expected number of plants. The expected and observed number of plants for each genotype were totalled across all subpopulations before the χ^2 comparison, so that the test was across the entire dataset rather than individual subpopulations. Regions of the genome can and do exhibit segregation distortion in nature due to mitotic or embryonic selective pressures, and inclusion of segregating distorted markers improves genetic linkage maps (Zhang et al., 2010). Therefore the small number of markers showing segregation distortion were first checked for scoring errors, and after this, only markers with highly significant segregation distortion, alongside other indicators that this segregation distortion was not due to natural phenomena, were excluded from the dataset and the map rebuilt as before. Other indicators that a marker was exhibiting unusual segregation distortion, were things such as a large genetic distance between any neighbouring markers or the positioning of the marker at the end of a linkage group.

2.7.4: Linkage mapping using Joinmap software

The 'CP' population option was selected for linkage mapping as one of the parental lines was homozygous diploid (line CV) and the other parental line heterogeneously heterozygous diploid (line G51). Conversion to Joinmap (Ooijen, 2011) compatible datafiles for this population type involved assigning markers and genotype scores the correct CP code *(<abxcd>,<efxeg>,<hkxhk>,<lmxll>* or *<nnxnp>*) and separating F_2 plants into their different subpopulations based on parentage. Each subpopulation was assigned to an independent population node within a single Joinmap project file. For each subpopulation segregation distortion was checked using the 'Locus Genotype Frequency' tab. Any markers showing highly significant segregation distortion were excluded from the dataset by using the 'exclude' option under the 'Data' tab. Pairwise linkage analysis was run and visualised using the 'Groupings (text)' and 'Groupings (tree)' tabs. Groupings were selected at the highest LOD score that gave 11 linkage groups, as supported by published literature on *J. curcas* chromosome number (Carvalho et al., 2008). Groups were created using the 'Create Groups Using the Groupings Tree' function.

A preliminary map for each group was created using the 'Calculate Map' function, with calculation options set to 'Regression mapping' using 'linkages with recombination frequency smaller than 0.4' and 'LOD greater than 1', 'Kosambi' mapping function, 'ripple' (equivalent to crimap 'flips') after each locus, and 3 mapping rounds.

Within each group node, the 'Weak Linkages' (defined as a pairwise recombination frequency greater than 0.45) and 'Suspect Linkages' (defined as a pairwise recombination frequency greater than 0.6) tabs were checked for markers with 'weak' or 'suspect' linkages as defined by recombination frequency. Markers with a high number of 'Weak linkages' to other markers not due to genetic distance as estimated from the preliminary map, were excluded from analysis using the 'Data' tab as before. In effect this comparison between calculated map distance and pairwise recombination frequencies is testing 'marker stress', as described by Ooijen in 'Genetic Mapping in Experimental Populations' (Ooijen et al., 2013).

After this process had been repeated for all linkage groups within subpopulations, corresponding linkage groups between subpopulations were combined using the 'Combine Groups for Map integration' function. The combined marker data was checked using the 'Heterogeneity Test' tab (which totals pairwise recombination frequencies across all subpopulations and carries out a χ^2 test for statistical significance), for the theoretical scenario that markers may not have reached statistically significant segregation distortion in subpopulation analysis but when totalled over all subpopulations were showing statistically significant segregation distortion. The 'Calculate Map' function was run as before to create an integrated linkage group map from each combined linkage group node.

The integrated map order was then used to check data in individual subpopulations by rebuilding the subpopulation linkage maps using the specified marker order from the integrated linkage maps. This was carried out by applying the 'map in fixed order format' at the end of the 'Session Log' tab from the integrated linkage group node, to the 'Fixed Orders' Tab of each subpopulation group node. The 'Calculate Map' function was then run on each subpopulation linkage group node with the specified fixed marker order. In this way subpopulation linkage groups were rebuilt using a marker order derived from the larger combined subpopulations dataset. Fixed order subpopulation maps were then checked for suspect double crossover events using the 'Genotype Probabilities' tab and by visualising the 'Data' tab with genotype colours on (equivalent to the 'chrompic' function of crimap). Any markers with 3 or more suspect double crossovers were excluded and the map rebuilt, using the same process of combining corresponding subpopulation group nodes to form a combined linkage group node, and using the 'Calculate Map' function to produce an integrated linkage group map. Crimap and Joinmap derived linkage maps were visualised and compared using MapChart (Voorrips, 2002) software to confirm a consensus map order derived from two independent builds using independent software.

2.7.5: Integration of multiple mapping populations into a single combined map

Genotype datasets were combined from several mapping populations to create a combined linkage map (work carried out by Dr. Andrew King, the University of York) using Crimap software as described above.

2.7.6: Additional Genetic Linkage Mapping: Gap filling using comparative mapping and castor bean microsynteny

Additional mapping was carried out to reduce gaps on the linkage map above 15 cM, see Phase 1 (b) of SSR marker development, Chapter 2.5.2.1.2:

2.8: Phenotypic data collection

2.8.1: Seed Traits

2.8.1.1: Collection of 'Seed oil content' and '100 seed mass' datasets using Nuclear Magnetic Resonance (NMR) spectroscopy

F_2 seed was received in a semi-dried state in paper sachets from the mapping population field site for Nuclear Magnetic Resonance (NMR) spectroscopy analysis using an Oxford Instruments MQC Benchtop NMR analyser (Abingdon, Oxfordshire).

Known amounts of pure *J. curcas* oil were pipetted into glass vials and used as calibration standards. The resultant oil content calibration curve was used to calibrate the NMR spectrophotometer for oil content. Calibration standards were retained and two standards re-measured at the start of every NMR session to ensure correct oil content calibration.

To calibrate the NMR measurements for water content, 2 samples were used from (a) freshly harvested seeds (b) seeds stored > 1year at ambient humidity (c) seeds stored at 20 % relative humidity and (d) seeds stored at 60 % relative humidity. Seeds were first measured for water content using NMR, before being placed in an oven at 103 ^{o}C overnight to remove water. Seeds were cooled to room temperature in silica gel desiccant before measuring dry weight to calibrate the NMR for water content.

Differing relative humidity was achieved by placing seeds in hermetically sealed chambers containing salt solutions (KCl) of differing concentrations. The salt concentration determines the surrounding air moisture content by affecting the equilibrium between water in solution with the salt ions and water as gas particles in the surrounding chamber. Seeds were left in these chambers for at least four weeks to ensure moisture diffusion and equilibrium throughout all seed tissue.

F_3 seed (seed produced by the F_2 plants) were analysed in batches of 5-6 seeds to ensure correct positioning in the NMR magnetic field. Seed mass was measured using scales accurate to 2 decimal places, prior to oil and water content measurement using NMR. This generated datasets containing seed mass, oil content & water content. Typically 50, but at least 20 seeds, from each F_2 plant were measured. Individual seed measurements were integrated into single values for each F_2 plant and normalised to 7 % water content.

2.8.1.2: Analysis of the seed fatty acid composition dataset using Gas Chromatography

For each F_2 plant, 20-25 whole seeds were mechanically ground using domestic coffee grinders until a fine homogenous powder was formed. Triplicate samples were taken from this material for Gas Chromatography (GC) according to published methodology (He et al., 2011). Briefly 10-30 mg of material was transferred to 2 ml glass vials containing 1 m HCL (in methanol), hexane, and a 15 carbon internal standard. Vials were sealed with Teflon lined screw caps, vortexed to ensure thorough mixing and subject to an 2 hr incubation at 85 ^{o}C. After cooling, cell components were partitioned by the addition 0.9 % KCL, before removal and transfer of the hexane layer containing fatty acids, into tapered vials for GC analysis. Negative controls containing hexane and the external standard SUPELCO 37 FAME mix were included for every GC run. Each sample was injected in triplicate as a control against machine variation. Raw GC data was scored by first creating a template trace containing each fatty acid peak from the external standard. Automatic scoring was then carried out using software, before manual checks of each sample to ensure correct assignment of each peak. Raw data was processed to produce % of each fatty acid compared to total seed oil in each sample.

2.8.2: Non-seed traits: branching and seed yield

Non-seed traits were collected in the field by Luis Montes (Biocombustibles de Guatemala, Guatemala Ciudad, Guatemala, and Plant Breeding Wageningen UR, Wageningen, The Netherlands) at the mapping population field site in Guatemala.

2.9: QTL mapping

2.9.1: Trait Analysis

Two approaches were used to integrate genetic and phenotypic datasets for QTL analysis. In the first approach the complex population structure consisting of multiple subpopulations and semi-informative markers were used, with data analysis occurring in GridQTL (Seaton G., 2006) software which is compatible with such population structures/markers. This approach assumed that shared alleles in the parental lines e.g. 'a' alleles at semi-informative loci, are due to these loci being conserved or the same in both parents. As a result this QTL mapping approach places greater emphasis on loci alleles rather than parent of origin effects.

In the second approach the complex population structure was converted to a standard F_2 population by converting semi-informative markers to informative markers, either by inferring genotype by closely flanking (<15 cM) informative marker genotypes or by conversion to dominant markers in non-informative subpopulations where flanking markers were not available. This results in all alleles from a particular parent being labelled one genotype class e.g. 'a' and the alleles from the other parent the other genotype class e.g. 'b' as in a standard F_2 population. This approach therefore assumes all loci from the parental lines are unique and ignores any difference between the non-uniform F_1 used in the cross. Therefore in contrast to the first approach this places greater emphasis on parentage of origin effects rather than loci alleles. By integrating subpopulations it also increases the QTL mapping population size and gives all markers the same relative information content. The mapping population size in this second approach was compatible with MapQTL (Ooijen, 2004) software.

2.9.2: GridQTL

Genotype, map and trait data were converted into GridQTL compatible files as described by GridQTL protocols. QTL analysis was run using both additive and dominative models, with experimental and chromosomal wide permutation analysis set at the maximum 10,000 iterations. F-values were calculated every 1 cM along each linkage group. F-values were converted to LOD scores to allow -1 LOD and -2 LOD QTL boundaries to be calculated, using the following formula:

$$LOD = \frac{LRT}{2\ln(10)}$$

where,

$$LRT = \{df(RSSf) - 0.5(2 - df(QTL))\}\ln\left(\frac{RSSr}{RSSf}\right)$$

And,

$$\frac{RSSr}{RSSf} = \{1 + Fvalue \times \frac{df(QTL)}{df(RSSf)}\}$$

Where, *RSSf* is the residual SS for the full model (including QTL)

RSSr is the residual SS for the reduced model (without QTL)

df(QTL) are the degrees of freedom for the QTL

df(RSSf) are the degrees of freedom associated with the residual for the full model

Source: The above formulae were derived from 'Multivariate statistical analysis for biologists' (1964), personal communications with Dr. Sarah Knott, contributing author and software support for GridQTL.

The author of this thesis study integrated the above formulae for excel data manipulation:

F-value to LOD conversion:

$$LOD = \frac{df(RSSf) \times \ln\left\{1 + \left(\frac{Fvalue \times 2}{df(RSSf)}\right)\right\}}{2\ln(10)}$$

Excel format: *LOD* =(*df(RSSf)**LN(1+((*Fvalue**2)/*df(RDDf)*)))/(2*(LN(10)))

-1 LOD QTL boundary, F-value:

$$Fvalue(at - 1\ LOD) = \frac{df(RSSf)}{2} \times \left\{\left(e^{\left(\frac{(LODmax)-1}{df(RSSf)} \times 2\ln(10)\right)}\right) - 1\right\}$$

Excel format: *Fvalue(at-1 LOD)* =(*df(RSSf)*/2)*(EXP(((*LODmax*-1)/*df(RSSf)*)*(2*LN(10)))-1)

Where, *LODmax* is the LOD score at the QTL position

-2 LOD QTL boundary, F-value:

$$Fvalue(at - 2\ LOD) = \frac{df(RSSf)}{2} \times \left\{\left(e^{\left(\frac{(LODmax)-2}{df(RSSf)} \times 2\ln(10)\right)}\right) - 1\right\}$$

Excel format: *Fvalue(at-2 LOD)* =(*df(RSSf)*/2)*(EXP(((*LODmax*-2)/*df(RSSf)*)*(2*LN(10)))-1)

Where, *LODmax* is the LOD score at the QTL position

2.9.3: MapQTL

2.9.3.1: Data manipulation to convert subpopulations and semi-informative markers to a standard F_2 population and informative markers

Semi-informative markers were informative in 1 out of 4 subpopulations (where the heterozygous F_1 was selfed). All other subpopulations were either semi-informative (for F_1 sib crosses where 'H' genotypes were produced), or non-informative for the homozygous selfed F_1 containing indistinguishable 'a' alleles from each parent. The genotype scores in these semi- and non-informative populations were deleted and the parent of origin genotype inferred through flanking informative markers, under the assumption that double crossovers do not occur within a distance of <15 cM due to crossover interference (Ooijen et al., 2013). In some cases the distance between flanking informative markers was greater than 15 cM, or flanking markers had different genotypes and so in these cases genotype scores were not able to be inferred. Where these markers were previously 'h', they were converted to dominant scores e.g. –b, since a minimum of one 'b' allele had to be present for the 'h' score to occur. Semi-informative markers with at least one 'b' allele, that were flanked by an 'a' informative and either a 'h' or 'b' informative marker (within 15 cM) were converted to the 'h' or 'b' genotype of the informative marker, since the 'a' genotype was excluded by the presence of the 'b' informative allele in the original marker score, and a differing 'h' or 'b' score excluded due to crossover interference. Where flanking informative markers within 15 cM were 'h' and 'b' the semi-informative marker score was changed to a dominant score since the marker had to be 'h' or 'b' and so contain a minimum of one 'b' allele represented by the dominant '-b' genotype score. After conversion of the

dataset into fully informative markers and a single F_2 population, the genotype and phenotype datasets were converted to MapQTL format and imported into MapQTL for analysis.

2.9.3.2: QTL mapping

Following import of MapQTL compatible datafiles, a Kruskall-Wallis (One way ANOVA) test was carried out to analyse single marker/trait associations. A permutation test was carried out to determine QTL significance thresholds for interval and composite interval mapping. Both linkage group wide and genome wide significance tables were created for a range of p-values using the 'Permutation' function with 1,000 iterations and 1,000 repeats.

Composite interval mapping was carried out for all Quantitative traits. After an initial mapping round, markers nearest the maximum of significant QTL were selected as co-factors, before repeating the composite interval mapping function. Markers flanking the co-factors that were still above the significance threshold after this second mapping round were also selected as co-factors and the mapping analysis repeated until the minimum QTL region was produced.

2.9.4: Cosegregation analysis

Cosegregation analysis was carried out using visual box and whisker plots and by statistical means.

For Box and whisker plots, the nearest marker to the QTL position, or markers flanking the QTL position (the QTL interval), were used to group phenotypic scores according to the marker genotype. Box and whisker plots were generated using SPSS software (IBM, 2013). This enabled a visualisation of the difference between genotype means, interquartile ranges and ranges. Percentage increase between the different genotype means, and Percentage of Variation Explained by genotype (PVE) was used to quantify the strength of the QTL on phenotype.

An analysis of variance (ANOVA) test was carried out to test cosegregation statistically. ANOVA looks at the level of variation within and between the grouped data. The ratio of variation between genotype groups compared to the total variation within the dataset gives an indicator as to the strength of the genetic component of the variation present. To test which genotype classes were statistically different from eachother, and therefore indicate if the QTL was dominant/recessive/semi-dominant/overdominant, a post hoc Tukey's test was carried out using SPSS (IBM, 2013).

2.9.5: Correlation and Linear Regression Analysis

SPSS software (IBM, 2013) was used to carry out correlation analysis to look at association between different traits. A Pearson two way correlation analysis was used to quantify the strength of association between pairwise traits. After integration of multiple traits to calculate oil yield per plant (seed oil content x seed mass x seed yield), linear regression analysis was used to calculate the relative contribution of each trait to oil yield, to determine the most important trait.

Chapter 3: Identification and validation of SSR markers from *J. curcas* genotypes selected primarily on the basis of seed oil quantity and quality

3.1: Introduction

3.1.1: Identification and validation of SSR markers

One of the requisites of QTL mapping and marker assisted selection (MAS), is the production of DNA markers (Dekkers and Hospital, 2002). An advantage of co-dominant markers is that they display both dominant and recessive alleles at a given locus (Staub et al., 1996). Short Sequence Repeat markers (SSRs) are one class of co-dominant marker that can be genotyped using polymerase chain reaction (PCR) (Schlotterer, 2004).

SSR's are repeat sequences that expand and retract over time due to a number of processes related to DNA replication (Li et al., 2002). They are both abundant and hypervariable (Oliveira et al., 2006); consequently they serve as excellent genetic markers.

SSR positions, or microsatellites, can be identified by searching for repeat sequences in genome sequence (Stieneke, 2007, Martins et al., 2009). This search can be carried out in a reference genome sequence, preventing the need to sequence the genome of parent plants of a given mapping population. Once identified in the reference genome sequence, SSR positions can be checked for polymorphism in the mapping population by PCR amplification.

Since the physical positions of SSRs are known in a reference genome sequence, specific SSRs can be targeted to map regions of interest, which is particularly useful for mapping gaps in a linkage map or for the mapping of candidate genes. This approach is dependent on the coverage of the reference genome sequence, whether SSRs exist close enough to the region of interest, and if the SSRs are polymorphic in the mapping population. Candidate genes; genes known or suspected to regulate a trait of interest (Pflieger et al., 2001), such as seed oil content or quality, can be mapped in this way.

The candidate gene approach is reliant on information being present on the genes associated with the trait of interest, including reliable sequence data (Pflieger et al., 2001). Well studied model species with characterised genes, such as Arabidopsis, serve as excellent resources from which to compile candidate gene lists. Since Arabidopsis is an oilseed species it is particularly useful for identifying candidate genes associated with oil quantity and quality in Jatropha, using a reference genome sequence and search algorithms.

Once a candidate gene is mapped on a genetic linkage map its involvement with identified QTL can be hypothesised based on whether it falls within the QTL confidence interval. This process can occur in both directions. Once a QTL is identified, markers within the confidence interval can be used to pull out genome sequence in that region and scanned for likely candidate genes. For long generation plants such as Jatropha (9 months seedling to seed, first substantive harvest in Year 2), where phenotypic data for QTL analysis may take considerable time to obtain, candidate genes involved in the regulation of important traits can be mapped first, so that once QTL are identified the position of candidate genes has already been determined. Phenotypic variation in parental lines can inform which traits are likely to have QTL associated with them. If a mapped

candidate gene does fall within a QTL confidence interval, the gene sequence and any surrounding regulatory sequences, can then be amplified and sequenced to detect potentially causative polymorphisms.

The seed oil biosynthetic pathway is particularly important for a biofuel crop such as Jatropha for obvious reasons. It is fortunate that, due to the economic value of plant oils for both food and industry, the seed oil biosynthetic pathway is one of the most well studied and understood pathways in plants (Li-Beisson et al., 2013). Genes within this pathway are known to regulate both seed oil content (a component trait of oil yield) and oil quality (fatty acid composition); two traits that are vital for developing economically viable biofuel cultivars. Vegetative and plant architecture traits such as the amount of branching, seed yield, or seed mass, can also contribute towards final oil yield and so are important traits to map if variation is present in parental lines. The mapping of candidate genes associated with oilseed metabolism, provides a basis to investigate QTL that impact on oil yield and oil quality in this study and others, since it is relatively easy to align different genetic maps using bridging and anchor markers.

Genetic linkage mapping benefits from large population sizes and high numbers of markers (Mackay et al., 2009). Since genetic mapping counts the crossover/recombination frequency between different markers; an event that is proportional to the genetic distance between those markers and independent from external, environmental conditions, genetic data from independent mapping populations can be combined to increase the amount of genetic data available to calculate recombination frequency. For this reason the same marker can be used in multiple mapping populations to increase the data available for its recombination frequency, whilst increasing marker density in the individual maps. There is also a greater chance that a marker will be polymorphic if tested across multiple populations rather than single populations, which is particularly important for the mapping of candidate genes, since once positioned on one map, the candidate gene position can be inferred on all other maps.

3.2: Results

3.2.1: SSR mining leads to the identification of over 300 SSR positions, of which 288 had flanking sequence suitable for validation by PCR amplification

Figure 3-1, shows the results of SSR mining for the Jatropha project. Over 300 SSR positions were identified from reference genome sequence, of which 288 had flanking sequence suitable for validation via PCR, according to criteria required for PCR multiplexing and scoring using an ABI3730 capillary sequencer (80-450 bp amplicon size, and a melting temperature (T_m) of 55 ^oC; see materials and methods). As outlined in the materials and methods, the majority of this search occurred *in silico*, using the Jatropha reference genome sequence and web-based programmes to identify repeat sequences, candidate gene homologues and suitable primer binding sites (Sato et al., 2011, Stieneke, 2007, Martins et al., 2009, Untergasser et al., 2012).

3.2.2: 39.59 % of validated SSRs were polymorphic in 1 or more mapping populations, providing data for these loci to be mapped in a combined genetic linkage map and subsequent QTL analysis

Of the 288 SSRs tested, 39.59 % (114) were polymorphic in 1 or more mapping populations: 46 SSRs in G51xCV only (the principle population of this thesis study), 25 SSRs in G33xG43 only, and 43 SSRs in 2 or more populations (34 SSRs in 2 populations, 4 SSRs in 3 populations, and 5 SSRs in all 4 populations). Mapping of markers in multiple populations increases the amount of genetic data available to calculate

marker recombination frequency; increasing the accuracy of the resulting genetic linkage map. It also enables independent maps from different mapping populations to be aligned for the purpose of gap filling/comparative mapping. Post-genetic linkage mapping, it also provides usable genetic markers for QTL mapping in each of the populations in which the markers are mapped.

3.2.3: SSRs were developed primarily for the mapping of candidate genes (58.33 %, 168 SSRs) or for gap filling during linkage mapping (41.67 %, 120 SSRs)

The primary function of the SSRs identified and validated were for the mapping of candidate genes (133 for seed oil-related genes, 25 for branching genes, 10 for flowering genes) or for gap filling during linkage mapping (120 SSRs). SSRs are ideally suited for mapping specific regions of the genome in a targeted way. A query sequence can be used to find SSRs in a specific region using a reference genome sequence. For example, candidate genes can be used as query sequences, or markers corresponding to gaps in a linkage map when aligning different genetic linkage maps (comparative mapping).

Genes that carry out a core metabolic function, such as seed oil candidate genes, tend to be highly conserved at the protein level, enabling homologues to be found relatively easily using BLAST algorithms. For gap filling, due to the presence of markers that were polymorphic in multiple populations, their resulting genetic linkage maps could be aligned, and markers corresponding to regions requiring additional mapping in one map, used as the query sequence to mine for additional SSRs in that region in other maps. This approach also works across species; castor bean, a close Euphorbiacea relative of Jatropha, can be aligned using transcribed amino acid sequences from annotated gene models (which are more highly conserved than nucleotide sequence), and transcribed amino acid sequences corresponding to the target region used to search the Jatropha genome for nearby SSRs; provided synteny and gene-colinearity exists between the species in the target region.

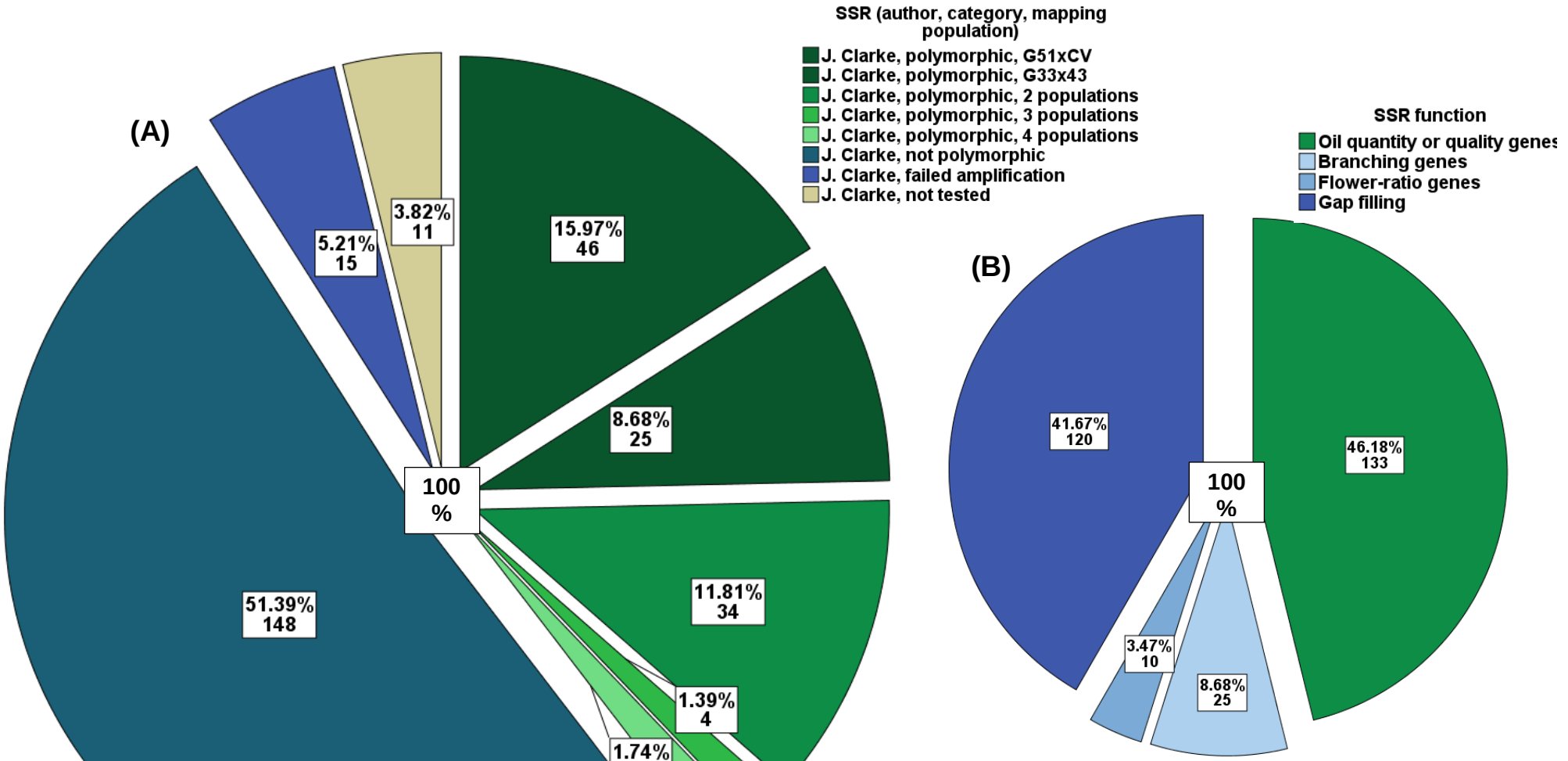


Figure 3-1 Short Sequence Repeat markers developed as part of this thesis study for Jatropha mapping populations at the University of York.

In total, 288 SSRs were identified and validated from reference genome sequence. **(A)** 114 (39.59 %) SSRs were polymorphic in the principle mapping population (G51xCV) and up to 3 other populations, providing data for a combined linkage map and subsequent QTL analysis. **(B)** SSRs were developed for the mapping of candidate genes or for gap filling during linkage mapping.

3.2.4: Candidate genes were identified for seed oil related traits (seed oil content and seed oil composition), and branching

As can be seen by figure 3-1, the majority of SSRs (168, 58.33 %) were developed for the mapping of candidate genes for the traits seed oil content, fatty acid composition, branching and flower ratio. Genes relevant to these traits were mined from Jatropha genome sequence using comparative genomics. A table listing the identified gene-linked SSRs is provided in the appendix to this chapter. The majority of candidate genes were first identified in the model species Arabidopsis, before sequence homologues were identified in *Jatropha curcas* genome sequence, using transcribed amino acid sequences which are more highly conserved across species than nucleotide sequence (due to redundancy of triplicate codons).

Once genes were identified in *Jatropha curcas* genome sequence, nucleotide sequence was used to search for nucleotide and $_{m}$RNA Jatropha accessions in GenBank, in order to provide one indicator of likely functionality (gene expression) of the identified gene sequences.

3.2.4.1: Seed oil content candidate genes

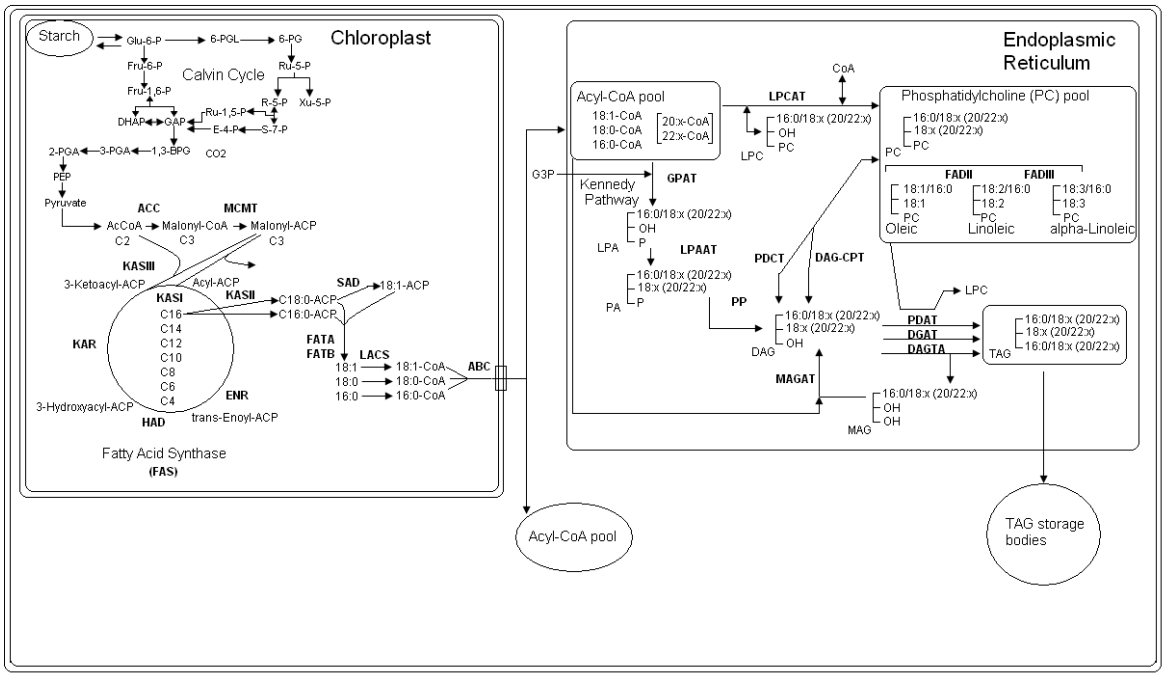


Figure 3-2. The major steps fatty acid synthesis in seed storage oil in plants.

Compiled from 'The Arabidopsis Book' (Li-Beisson et al., 2013). Major steps of the seed fatty acid pathway (represented by arrows), and the genes responsible (in bold) are presented. In general, fatty acid biosynthesis can be divided into *de novo* synthesis in the chloroplast, and four separate fatty acid pools; the acyl-CoA pool (both cytoplasmic and endoplasmic reticulum pools), the phosphatidylcholine pool, and the tri-acyl-glycerol (TAG) pool. Fatty acids from each fatty acid pool or compartment have unique carrier proteins; acyl carrier protein (ACP), co-enzyme A (CoA), phosphatidylcholine (PC) or glycerol. Detail on relevant candidate genes in this pathway are explained in the text below.

3.2.4.1.1: Acetyl-CoA Carboxylase (*ACC*)

ACC converts the main output of the Calvin cycle, pyruvate (after its been hydrogenated to Acetyl-CoA), to Malonyl-CoA using bicarbonate ions and ATP, and represents the first committed step in fatty acid synthesis in the plastid (Li-Beisson et al., 2013). It is made up of 4 subunits; BC, BCCP, CAC2 –α, CAC2-β (Sasaki

and Nagano, 2004). Since there is competition for pyruvate for other metabolic processes such as amino acid production, and since pyruvate, and hence Acetyl-CoA production, can be high at points for instance during periods of high rates of photosynthesis and respiration, the activity of ACC is thought to affect the proportion of pyruvate that is committed to fatty acid synthesis (Paul K. Stumpf, 2012, Li-Beisson et al., 2013), and is also thought to be a rate limiting step in fatty acid synthesis (Sasaki and Nagano, 2004). Its activity is known to be regulated by certain mechanisms, such as the amount of free fatty acid and light/dark (Shintani and Ohlrogge, 1995, Li-Beisson et al., 2013), suggesting that a mutation that overrides its negative regulation by other signals (such as the amount of free fatty acid), could provide a mechanism by which to continue shuttling pyruvate into fatty acid synthesis to increase seed oil content. Alternatively a mutation that increases its substrate turnover rate (enzyme activity) could also be envisaged to increase seed oil content. There are examples where overexpression of ACCase has increased seed oil content in rapeseed (*Brassica napus*) (Roesler et al., 1997). There are also crop breeding examples where seed oil content QTL associate with *ACC* polymorphisms, in plants such as oat (Kianian et al., 1999).

3.2.4.1.2: Acyl carrier protein (*ACP*)

Acyl carrier protein is responsible for binding to and shuttling malonyl, the two carbon building block, to the fatty acid synthase (FAS) complex, before shuttling the growing fatty acid chains through progressive elongation cycles, and then shuttling them out of the FAS complex, onto modification steps (such as desaturation) before ACP is exchanged for CoA for export to the cytoplasm. ACP can therefore be thought of as one of the main players that are associated with fatty acids throughout the plastid stage of synthesis. One can envisage that a mutation in ACP could potentially affect its interaction with the host of metabolic enzymes and transport proteins that make up fatty acid synthesis in the plastid, potentially affecting both seed oil content and seed oil composition.

3.2.4.1.3: The Keto-Acyl Synthases (*KASI, KASII, KASIII*)

The fatty acid synthase (FAS) complex, made up of 4 independent catalytic subunits (KASI, KAR, HAD, ENR), and two additional enzymes that control fatty acid entry and exit into FAS (KASIII, KASII), controls the elongation reactions of fatty acid synthesis in the plastid. The initial condensation reaction is carried out by KASIII, linking Malonyl-ACP to Acetyl-CoA to form the 4 carbon fatty acid, 3-Ketoacyl-ACP. Subsequent steps are carried out by KASI, through the addition of Malonyl-ACP, from a 4 carbon fatty acid up to a 16 carbon fatty acid, in two carbon increments. A final elongation reaction from a 16 carbon to an 18 carbon fatty acid is carried out by the KASII enzyme.

There is an extensive array of studies proving the central roles of the KAS genes in regulating both fatty acid synthesis and fatty acid composition. From a mechanistic point of view the relative activities of these three enzymes to eachother and other genes, such as the thioesterases (that release fatty acids from ACP to halt elongation), are proposed as a mechanism by which they can affect the proportion of different fatty acids in seed storage oil, such as the relative amounts of Palmitate (16C) and Stearate (18C) for example. Upregulated activity of KASIII tends to push more fatty acid into the FAS complex, and has been proposed as a rate limiting step in fatty acid synthesis and a target for increasing seed oil content with success (Yu et al., 2015, JUN LI, 2008, Stoll et al., 2006). The activity of KASI compared to KASII, seems to compensate better by this increased flux, elongating fatty acids up to the 16 carbon length and out of the FAS complex, quicker

than KASII can carry out the final elongation to 18 carbon fatty acid, resulting in higher levels of palmitate in the seed storage oil.

There have also been studies looking at these genes in plants with naturally differing seed oil compositions (Voelker and Kinney, 2001). Plants high in palmitate compared to stearate often have highly active *KASI* genes to produce high levels of 16 carbon fatty acid, and *FATB* thioesterase genes (that release palmitate from the FAS complex for export) with a much higher activity than the *KASII* gene. *KASII* and *FATB* can be thought to be in competition for the 16 carbon palmitate substrate so their relative activities determine what proportion of fatty acid is released as palmitate or converted to stearate. They work in conjunction with different thioesterases which have differing substrate preferences, determining at what stage elongation is halted and preventing metabolic bottlenecks that could be self-limiting.

3.2.4.1.4: Stearoyl-ACP desaturases (*SAD*)

After KASII has elongated palmitate-ACP to Stearoyl-ACP, the first fatty acid modification step can occur by desaturation of the ninth carbon by SAD (a delta 9 desaturase) in the plastid. Whilst the majority of fatty acid modification occurs on the endoplasmic reticulum, this step differs in that it occurs in the plastid. A mutation resulting in a change of activity of the *SAD* gene could be hypothesised to cause a change in fatty acid composition. Reduced activity could reduce the proportion of oleic acid, and downstream linoleic, and alpha linoleic acids, and increase the proportion of saturated fatty acids in seed storage oil. Increased activity could be hypothesised to increase the proportion of desaturated fatty acids, by converting more stearate into oleate, and downstream polyunsaturated fatty acids. Demonstration of this principle can be seen by the anti-sense suppression of a *SAD* gene in *Brassica napus* and *Brassica rapa*, resulting in dramatically increased stearate levels (Knutzon et al., 1992).

3.2.4.1.5: The Thioesterases (*FATA* and *FATB*)

The thioesterases are responsible for removing the ACP carrier proteins that are associated with fatty acids as they are elongated and modified in the plastid. Removal of ACP prevents further modifications in the plastid and, with the addition of a CoA carrier protein, enables their transport to the cytosol for the next stage of the seed storage oil pathway. The two isoforms, FATA and FATB have differing substrate preferences that affect seed oil composition. FATB hydrolyses shorter chain, saturated fatty acids preferentially (palmitic-ACP), although they can also hydrolyse stearoyl-ACPs and oleic-ACP to a lesser extent. FATA hydrolyses oleic-ACP preferentially, with lower activity towards palmitic-ACP and stearoyl-ACP. Their relative activities in conjunction with the KAS genes has a large effect on the fatty acid output of the plastid during synthesis (Voelker and Kinney, 2001). Plants with high KASII, SAD and FATA activities channel more fatty acid through to oleic acid. Plants with lower FATB and SAD activity tend to channel more fatty acid towards stearic acid. Those with lower KASII activity and higher FATB activity channel more fatty acid towards palmitic acid. Mutations resulting in changes of activities of these enzymes have been proven to substantially alter fatty acid composition (Moreno-Perez et al., 2012).

3.2.4.1.6: Long Chain Acyl-CoA Synthases (*LACS*)

After the fatty acid thioesterases have liberated fatty acids from their acyl carrier proteins, the LACS are responsible for conjugating them to CoA proteins for export to the cytosol. Since it has been proven that free fatty acids can have negative feedback regulation on fatty acid synthesis as a whole, including the activity of

ACC, the efficiency of conjugation and subsequent export into the cytosol is a critical factor in seed oil content (Zhao L, 2010). Similarly the *LACS* gene family is known to have different substrate turnover rates in Arabidopsis (Shockey et al., 2002), and along with other acyl modifying genes, regulates fatty acid compartmentation and subsequent modification (Chapman and Ohlrogge, 2012), and so it could be hypothesised that changes in the activity of different LACS could affect oil composition in plants such as Jatropha.

3.2.4.1.7: Compartmentation and shuttling genes on the cytoplasmic, endoplasmic reticulum (*LPCAT, GPAT, LPAAT, PP, MAGAT, PDCT, DAG-CPT, PDAT, DGAT, DAGTA, Pla2g4b*)

Once exported into the cytosol, Acyl-CoAs are shuttled between 3 fatty acid pools, that act as compartmentation mechanisms to carry out further modification, or storage, mechanisms. This is controlled by the associated carrier protein. The Acyl-CoA pool represents newly exported fatty acids from the plastid. The phosphotidyl-Choline (PC) pool represents another compartment where modification steps such as desaturation can occur. Finally the Triacylglycerol (TAG) pool is the storage pool that enables fatty acids to accumulate to high levels without interfering with other cellular reactions. The exchange of the CoA and PC carrier proteins, and the incorporation of both pools onto the glycerol backbone for TAG storage, is carried out in a number of steps by different genes (see figure 3-2). The relative activities of these genes can easily be hypothesised to control both seed oil content and seed oil composition (Chapman and Ohlrogge, 2012, Bates et al., 2013). For example the efficiency with which newly synthesised Acyl-CoAs from the plastid are incorporated into TAG storage oil could regulate seed oil content, by controlling the amount of excess fatty acid in solution that is known to negatively feedback on overall fatty acid synthesis. Oil composition could be affected by the rates at which differing fatty acids are incorporated into TAG in a substrate specific manner, and also the rate at which fatty acids are shuttled to the Acyl-PC pool for further modification. There are a number of examples where this has already been hypothesised and experimentally tested (Chapman and Ohlrogge, 2012, Li-Beisson et al., 2013, Sharma and Chauhan, 2012, Xu et al., 2012, Andrianov et al., 2010, Zheng et al., 2008, Lardizabal et al., 2008).

3.2.4.1.8: Endoplasmic fatty acid desaturases (*FAD2, FAD3*)

Further modification can occur once fatty acids have been transported out of the plastid onto the endoplasmic reticulum. The desaturation reactions that convert oleate (18:1) to linoleate (18:2) and downstream linolenate (18:3), occur in the Acyl-PC pool, by the action of desaturases. FAD2, a delta 12 desaturase, is responsible for desaturating the twelfth carbon position of oleate to convert it to the polyunsaturated linoleate, followed by FAD3 which desaturates linoleate to linolenate. Since the majority of these modified fatty acids end up in TAG storage oil in seed tissue, altering the activities of these genes has been shown to substantially alter fatty acid composition of TAG (Qu et al., 2012, Belo et al., 2008, Sandhu et al., 2007, Schuppert et al., 2006, Hu et al., 2006, Hernandez et al., 2005, Patel et al., 2004). Similarly because these desaturated fatty acids are predominantly stored, particularly when looking at seed specific FAD isoforms, modification has not been found to affect overall cell metabolism or plant fitness. In plants, the *FAD2* gene seems to be the sole pathway for oleate desaturation, with a seed specific isoform that regulates seed oil desaturation. Therefore mutations that affect *FAD2* genes seem to be very effective at altering seed oil composition. Active site mutations, and other naturally occurring knockout polymorphisms have been shown to produce high oleate oil, which is the preferred seed oil fatty acid profile for biofuel production. This is an attractive target for

manipulation to produce designer oil, and has been exploited for a wide range of plant species both experimentally and commercially.

3.2.4.1.9: Seed oil body associated storage proteins (Oleosins, Caleosins)

Once free fatty acids and their associated carrier proteins, PC or CoA, have been exchanged for a glycerol backbone, to form TAG, the compound becomes insoluble and starts to form inclusion bodies. This groups TAG molecules together, effectively storing them as large bodies separate from cellular reactions. Caleosins and Oleosins bind to the outside of these bodies, providing an interface between the hydrophilic cytoplasm and the hydrophobic oil body and ultimately stabilising them, regulating flux in and out of these bodies, and enabling them to become bigger without breaking up (Hyun et al., 2013, Jolivet et al., 2013, Parthibane et al., 2012, Gitte I. Frandsena, 2001). These molecules have been shown to be critical for allowing higher concentrations of seed oil to accumulate, with studies showing QTL association of oleosins with high seed oil content in crops including Jatropha (Liu et al., 2011).

3.2.4.1.10: Fatty acid synthesis master regulator, *Wrinkled 1* (*WRI1*)

WRI1, is a transcription factor that has been proved to be important for regulating fatty acid synthesis as well as other metabolic processes (Baud and Lepiniec, 2009)(Tajima et al., 2013). It is an APETELA-ethylene responsive binding element, that has been proved to regulate genes of late glycoloysis, and the plastidial fatty acid synthesis gene network (Baud and Lepiniec, 2009), including the fatty acid synthase (FAS) machinery directly. The *WRI* gene is necessary for normal seed storage oil accumulation, with seed storage oil severely impaired in the *wri1* mutant (although basal fatty acid synthesis is maintained to enable vegetative growth) (To et al., 2012). Conversely, *WRI1* overexpressors accumulate higher levels of seed oil, both at the per seed and per hectare level in maize (Shen et al., 2010), and in other species (Vanhercke et al., 2013). Also, unlike upstream regulators such as *LEC1*, the storage fatty acid specificity of *WRI1* means that modulation has little effect on overall plant fitness and does not have any known pleiotropic effects on other processes.

3.2.4.2: Branching candidate genes and flower ratio genes

A number of candidate gene classes were identified for the branching and flower ratio traits.

Key genes, and gene families, identified for branching included: (1) the *MAX* gene family (Bennett et al., 2006); (2) genes encoding F-box proteins; *TIR1, AFB* (Kepinski and Leyser, 2005, Dharmasiri et al., 2005); (3) *AXR1* gene (Stirnberg et al., 1999, Ongaro and Leyser, 2008); (4) the *PIN1* gene (Bennett et al., 2006); (5) the *MOC1* gene of rice and the Arabidopsis equivalent *LAS* (Sun et al., 2010, Wang and Li, 2008); and (6) the transcription factor *ABI3* (McSteen and Leyser, 2005, Ehrenreich et al., 2007, Ongaro and Leyser, 2008).

Genes and gene families identified for flower ratio include: (1) the lipoxygenase gene family (*LOX*) (Caldelari et al., 2011, Feussner and Wasternack, 2002); (2) the maize sex-determination *TASSELSEED* genes, and Arabidopsis homologues, *ATA1, ADH1* (DeLong et al., 1993, Thompson and Hake, 2009, Barazesh and McSteen, 2008); (3) the MADS-box flower developmental genes; *PI, SHP2, AG* (Dornelas et al., 2011, Adam et al., 2007, Becker and Theissen, 2003, Liljegren et al., 2000, Favaro et al., 2003).

3.3: Discussion

A key requirement of QTL mapping, after parental lines have been selected, is the development of DNA markers. Ideally DNA markers should be spread throughout the genome so that all regions of the genome can be tracked throughout the mapping population, and spaced so that minor effect QTL (<10 cM) (Darvasi et al., 1993) and all crossover events (<15 cM) (Ooijen et al., 2013), including double crossovers, can be detected(Darvasi et al., 1993). To get this coverage, genome-wide non-selective marker strategies can be used (Davey et al., 2011). In this project the Complexity Reduction of Polymorphic Sequences (CRoPS) technique (a modified AFLP sequencing, genome reduction strategy) (Davey et al., 2011, van Orsouw et al., 2007) was used to develop SNPs dispersed across the genome in a non-selective manner. SNPs mined from comparative sequencing of transcribed DNA (cDNA) (King et al., 2011), developed a small number of SNPs spread throughout transcribed genes, increasing the chance of picking up a mutation in functional DNA, although still in a non-selective manner within the Jatropha transcriptome. To complement this approach, SSR markers were used to place additional markers on the genetic linkage map in a more specific and targeted manner.

As can be seen, the majority of SSRs were developed to map specific genes, or to fill in remaining gaps in the linkage map. As a result of this work, over 300 SSR positions were identified, of which 288 had flanking sequence suitable for amplification via PCR. Validation testing across parental lines, showed that 114 (39.59 %) of amplified SSRs were polymorphic in at least 1 mapping population and could be used for genetic linkage mapping. In addition, a significant proportion were polymorphic in more than one population, increasing the utility of the marker for QTL mapping in multiple populations, and providing additional data on its recombination frequency for a combined genetic linkage map. Markers that were mapped in more than one mapping population, also enabled accurate alignment of individual maps and facilitated subsequent gap filling using comparative mapping strategies.

Such SSR markers were designed to complement the less specific genome wide SNP markers. Each SSR marker either marked a potential candidate gene or metabolic gene identified through research, or corresponded to gaps in the linkage maps after genetic mapping had been carried out using the genome wide SNPs. This targeted rational approach enhanced the robustness and information content of the overall DNA marker set for this project. In addition since it is relatively easy to anchor new linkage maps onto existing ones, by mapping shared markers, the position of all mapped candidate genes can be used to inform future QTL mapping projects.

Candidate genes identified for this project were primarily obtained for marking seed oil biosynthetic genes, for obvious reasons in the Jatropha biofuel crop. Identification of candidate genes, either *a priori* or post QTL analysis and gene sequencing, requires existing knowledge to be available, if predictions on gene function are to be achieved before investing in functional characterisation; a core function of the candidate gene approach (Pflieger et al., 2001). It is fortunate that, due to the high economic and industrial value of plant seed oils for food and industry, the seed oil biosynthetic metabolic pathway is one of the most well-known and characterised pathways in plants (Li-Beisson et al., 2013). The main challenge is translating existing knowledge into a previously understudied species such as Jatropha (at least at the gene functional level) (Morrell et al., 2012). Essentially it is dependent on the degree of sequence conservation between functional homologues (Peregrin-Alvarez et al., 2009). The sheer diversity of mutations that have been

shown to modulate seed oil content and composition across different species (Gupta, 2015, Napier et al., 2014, Bates et al., 2013, Sanyal and Randal Linder, 2012, Sharma and Chauhan, 2012, Weselake et al., 2009), with new mutations being discovered all the time, means mapping as much of the metabolic pathway as possible, is more likely to capture all potential mutations. In the age of comparative genomics (Morrell et al., 2012) this type of approach is readily achievable.

It is advantageous that firstly the core metabolic pathway for seed oil biosynthesis has been extensively studied, and that the majority of translative processing to find functional homologues can occur *in silico*. With SSR marker development also occurring predominantly *in silico*, a comprehensive approach can be used to maximise the chance of capturing as many relevant genes as possible. Since Jatropha is a relatively novel crop under study, particularly at the genetic level, marking of these vitally important genes has on-going utility both inside and outside of this project, and complements the rapidly improving genomic resources available for Jatropha development (King et al., 2015, Wu et al., 2015, Yue et al., 2013, King et al., 2013, Sun et al., 2012).

Whilst the majority of metabolic pathway and candidate gene research can occur *in silico*, using model species (Li-Beisson et al., 2013), online gene databases, published literature and the Jatropha reference genome sequence (Li-Beisson et al., 2013, Hirakawa et al., 2012, Sato et al., 2011), there are further advantages of using SSRs as markers.

SSRs have the advantage of being easily identified in a reference genome sequence due to their repetitive sequence (Stieneke, 2007, Martins et al., 2009). SNPs on the other hand require comparative sequencing to be detected. Once identified, SSRs can be validated and checked for polymorphism in parental lines via PCR amplification. Since SSR polymorphisms result in different SSR repeat sizes, and hence different PCR fragment sizes, sequencing is not needed and the actual nucleotide sequence itself is irrelevant. This reduces the technological requirements of developing reliable DNA markers, which for lower numbers of bespoke markers targeted to specific regions of the genome e.g. to mark candidate genes or to fill gaps, may be a more suitable approach than the implementation of genome-wide sequencing approaches.

Genes that carry out a core metabolic function, such as seed oil candidate genes, tend to be highly conserved at the protein level (Peregrin-Alvarez et al., 2009), enabling functional homologues to be found relatively easily using BLAST algorithms (Gish and States, 1993). For gap filling, due to the presence of shared markers that were mapped in multiple populations, independent linkage maps could be aligned, and markers corresponding to regions requiring additional mapping could be used as query sequences to search the Jatropha genome for additional SSRs in the target region, highlighting the utility of polymorphism testing of SSR markers across multiple mapping populations.

SSR markers provide substantial value to the marker set available for genetic linkage mapping in the G51xCV mapping population. Marker coverage and information content is increased through the targeting of SSRs to gaps in the linkage map after initial rounds of genetic linkage mapping, and by the marking of candidate genes related to a number of agronomically-relevant traits.

3.4: Appendix

List of candidate gene linked SSR markers.

Table 3-1 Candidate gene linked SSR markers

	Source		*Arabidopsis thaliana* gene model			Nearest sequenced relative	Evidence for Expression (*Jatropha curcas* mRNA)					
Marker	**Contig (r3.0)**	**Contig (r4.5)**	**Gene Symbol (At)**	**Gene description**	**At Homologue**	**Rc Homologue**	**NCBI accession**	**Nucleo. Coverage (%)**	**Nucleo. Similarity (%)**	**Accession Title**	**Authors**	**Journal**
3.2.4.1: Acetyl-CoA Carboxylase (*ACC*)												
JcSSR_G352		Jcr4S02200	*ACC1*	acetyl-CoA carboxylase 1	AT1G36160.2	29908.m005991	DQ632746.1	98	99	Amplification and sequencing of cytosolic ACCase gene from Jatropha curcas	Krishna Kumar,R., Jain,D., Parameswaran,S. and Johnson,T.S.	NCBI submission (2009)
JcSSR_G353		Jcr4S01232	*BCCP1*	Acetyl-CoA carboxylase BCCP subunit	AT5G16390.1	29929.m004560	HQ153098.1	96	100	Molecular cloning and expression of heteromeric ACCase subunit genes from Jatropha curcas	Gu,K., Chiam,H., Tian,D. and Yin,Z.	Plant Sci. 180 (4), 642-649 (2011)
JcSSR_G354	`	Jcr4S01222	*BCCP2*	biotin carboxyl carrier protein 2	AT5G15530.1	29630.m000809	GQ241721.1	80	100	Identification and characterization of a novel biotin carboxyl carrier protein subunit from Jatropha curcas L.	Wei,Q., Wu,P.Z., Zeng,L., Li,M.R., Chen,Y.P., Jiang,H.W. and Wu,G.J.	NCBI submission(2009)
JcSSR_G355		Jcr4S03449	*CAC2 -α*	acetyl Co-enzyme a carboxylase biotin carboxylase subunit	AT5G35360.1	30185.m000954	FJ952146.1	80	100	Identification and characterization of a novel ACCase from Jatropha curcas L.	Wei,Q., Wu,P.Z., Zeng,L., Chen,Y., Li,M.R., Jiang,H.W. and Wu,G.J.	NCBI submission(2009)
`JcSSR_G359		Jcr4S02200	*ACC1*	acetyl-CoA carboxylase 1	AT1G36160.2	29908.m005991	DQ632746.1	98	99	Amplification and sequencing of cytosolic ACCase gene from Jatropha curcas	Krishna Kumar,R., Jain,D., Parameswaran,S. and Johnson,T.S.	NCBI submission (2009)
JcSSR_G360		Jcr4S00416	*CAC2-β*	acetyl Co-enzyme a carboxylase carboxyltransferase alpha subunit	AT2G38040.2	27798.m000585	GQ845013.1	98	99	Identification and characterization of a novel alpha-carboxyltransferase subunit from Jatropha curcas L.	Wei,Q., Zeng,L., Wu,P.Z., Chen,Y.P., Li,M.R., Jiang,H.W. and Wu,G.J.	NCBI submission (2009)
JcSSR_G361		Jcr4S00075	*CAC2-β*	acetyl Co-enzyme a carboxylase carboxyltransferase alpha subunit	AT2G38040.2	30174.m008999	GAHK01016038.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)

JcSSR_G366		Jcr4U29862	*ACCD*	acetyl-CoA carboxylase carboxyl transferase subunit beta	ATCG00500.1	28890.m000006	HQ153096.1	96	99	Molecular cloning and expression of heteromeric ACCase subunit genes from Jatropha curcas	Gu,K., Chiam,H., Tian,D. and Yin,Z.	Plant Sci. 180 (4), 642-649 (2011)
3.2.4.1.2: Acyl Carrier Protein (*ACP*)												
JcSSR_G38	JCCA0308711	Jcr4S00106	*ACP*	Acyl carrier protein	AT4G25050.1	29726.m003980	EZ418424.1	100	100	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G42	JcCB0042491	Jcr4S00742	*mtACP*	Acyl carrier protein	AT5G47630.2	29826.m000732	GAHK01027354.1	98	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G353		Jcr4S01232	*mtACP2*	Acyl carrier protein	AT1G65290.1	29929.m004551	EF179617	100	100	Cloning and characterization of Jatropha curcas ACP gene	Jiang,L.D., Zhang,Y., Wang,Y.X., Wang,Y.C., Xu,Y. and Chen,F.	NCBI submission (2009)
JcSSR_G401		Jcr4S00649	*ACP3*	acyl carrier protein 3	AT1G54630.1	29739.m003654	GAHK01026238.1	30	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G408		Jcr4S00546	*ACP3*	acyl carrier protein 3	AT1G54630.1	30128.m008670	EZ412266.1	100	100	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G411		Jcr4S00546	*ACP3*	acyl carrier protein 3	AT1G54630.1	30128.m008670	EZ412266.1	100	100	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G415		Jcr4S00190	*ACP4*	acyl carrier protein 4	AT4G25050.1	30147.m014425	GAHK01013596.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
3.2.4.1.3: The Keto-Acyl Synthases (*KASI, KASII, KASIII, FAE*)												
JcSSR_G33	JcCA0143871	Jcr4S04655	*KAS2*	Beta-ketoacyl-ACP synthase II	AT1G74960.3	29739.m003711	DQ987700.2	100	99	Cloning and characterization of a beta-ketoacyl-acyl carrier protein synthase II from Jatropha curcas	Wei,Q., Li,J., Zhang,L., Wu,P., Chen,Y., Li,M., Jiang,H. and Wu,G.	J. Plant Physiol. 169 (8), 816-824 (2012)
JcSSR_G34	JCCB0043371	Jcr4S00903	*KAS3*	Beta-ketoacyl-ACP synthase III	AT1G62640.2	28455.m000368	DQ987701.1	100	99	Molecular cloning and expression analysis of a gene encoding a putative beta-ketoacyl-acyl carrier protein (ACP) synthase III (KAS III) from Jatropha curcas	Li,J., Li,M.R., Wu,P.Z., Tian,C.E., Jiang,H.W. and Wu,G.J.	Tree Physiol. 28 (6), 921-927 (2008)
JcSSR_G386		Jcr4S02541	*KAS1*	3-ketoacyl-acyl carrier protein synthase I	AT5G46290.1	29693.m002034	GAHK01017251.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G389a		Jcr4S04655	*KAS2*	Beta-ketoacyl-ACP synthase II	AT1G74960.3	29739.m003711	DQ987700.2	100	99	Cloning and characterization of a beta-ketoacyl-acyl carrier protein synthase II from Jatropha curcas	Wei,Q., Li,J., Zhang,L., Wu,P., Chen,Y., Li,M., Jiang,H. and Wu,G.	J. Plant Physiol. 169 (8), 816-824 (2012)
JcSSR_G396		Jcr4S08397	*KAS1*	3-ketoacyl-acyl carrier protein synthase I	AT5G46290.1	30068.m002515	GAHK01007871.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)

JcSSR_G409		Jcr4S27123	*KAS2/FAB 1*	fatty acid biosynthesis 1	AT1G749 60.3	29739.m00371 1	GT976545 .1	64	91	Transcriptome analysis of the oil-rich seed of the bioenergy crop Jatropha curcas L	Costa,G.G.L., Cardoso,K.C., Del Bem,L.E.V., Lima,A.C., Cunha,M.A.S., de Campos-Leite,L., Vicentini,R., Papes,F., Moreira,R.C., Yunes,J.A., Campos,F.A.P. and Da Silva,M.J.	BMC Genomics 11 (1), 462 (2010)
JcSSR_G413		Jcr4S00903	*KAS3*	Beta-ketoacyl-ACP synthase III	AT1G626 40.2	28455.m00036 8	DQ98770 1.1	100	99	Molecular cloning and expression analysis of a gene encoding a putative beta-ketoacyl-acyl carrier protein (ACP) synthase III (KAS III) from Jatropha curcas	Li,J., Li,M.R., Wu,P.Z., Tian,C.E., Jiang,H.W. and Wu,G.J.	Tree Physiol. 28 (6), 921-927 (2008)
JcSSR_G418		Jcr4S00288	*FAE/KCS 2*	3-ketoacyl-CoA synthase 2	AT1G042 20.1	29844.m00318 6	GAHK010 19920.1	95	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G419		Jcr4S00731	*FAE/KCS 4*	3-ketoacyl-CoA synthase 4	AT1G194 40.1	30190.m01090 3	GW87937 4.1	27	99	Profiling gene expression of the reproductive organs of Jatropha curcas	Wang,W., Wei,B., Sing,P., Jin,Q.D., Wong,W.S., Zhang,S.H. and Li,N.	NCBI submission (2010)
JcSSR_G421		Jcr4S00865	*FAE/KCS 19*	3-ketoacyl-CoA synthase 19	AT5G045 30.1	29690.m00041 2	No evidence for expression					
3.2.4.1.4: Stearoyl-ACP desaturases (*SAD, FAB2*)												
JcSSR_G36	JcCB0395 461	Jcr4S01370	*FAB2*	Plant stearoyl-acyl-carrier-protein desaturase	AT2G437 10.2	30020.m00020 3	DQ08449 1.1	100	99	Jatropha curcas stearoyl-ACP desaturase cDNA	Luo,T., Xu,Y., Deng,W., Wang,S., Tang,L., Xiao,M., Zeng,N., Guo,L., Zhang,Y. and Chen,F.	NCBI submission (2005)
JcSSR_G393		Jcr4S13936	stearoyl-ACP desaturase	Plant stearoyl-acyl-carrier-protein desaturase family protein	AT3G026 30.1	29929.m00451 5	EZ418900 .1	67	99	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G398		Jcr4S01370	*FAB2*	Plant stearoyl-acyl-carrier-protein desaturase	AT2G437 10.2	30020.m00020 3	DQ08449 1.1	100	99	Jatropha curcas stearoyl-ACP desaturase cDNA	Luo,T., Xu,Y., Deng,W., Wang,S., Tang,L., Xiao,M., Zeng,N., Guo,L., Zhang,Y. and Chen,F.	NCBI submission (2005)

JcSSR_G399		Jcr4S03070.3	stearoyl-ACP desaturase	Plant stearoyl-acyl-carrier-protein desaturase family protein	AT3G02630.1	29929.m004514	GAHK01004511.1	100	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G400		Jcr4S03070.4	stearoyl-ACP desaturase	Plant stearoyl-acyl-carrier-protein desaturase family protein	AT3G02630.1	29929.m004514	GAHK01026581.1	82	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G407		Jcr4S03522	stearoyl-ACP desaturase	Plant stearoyl-acyl-carrier-protein desaturase family protein	AT1G43800.1	27985.m000877	EZ412266.1	100	100	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
3.2.4.1.5: The Thioesterases (*FATA* and *FATB*)												
JcSSR_G39	JcCB0017291	Jcr4S00539	*FATA*	fatA acyl-ACP thioesterase	AT3G25110.1	30217.m000262	EU267122.2	100	96	Identification and characterization of a novel acyl-ACP thioesterase (FATA) from Jatropha curcas L.	Wu,P.Z., Li,J., Li,M.R., Jiang,H.W. and Wu,G.J.	NCBI submission (2010)
JcSSR_G416		Jcr4S00062	*FATB*	fatty acyl-ACP thioesterases B	AT1G08510.1	29841.m002744	JX966083.1	100	96	Cloning and characterization of an acyl-acyl carrier protein thioesterase like from Jatropha curcas	Zhang,L., Wu,P.Z., Jiang,H.W. and Wu,G.J.	NCBI submission (2012)
JcSSR_G417		Jcr4S02908	*FATB*	fatty acyl-ACP thioesterases B	AT1G08510.1	29660.m000782	JX966081.1	100	99	Cloning and characterization of an acyl-acyl carrier protein thioesterase like from Jatropha curcas	Zhang,L., Wu,P.Z., Jiang,H.W. and Wu,G.J.	NCBI submission (2012)
3.2.4.1.6: Long Chain Acyl-CoA Synthases (*LACS*)												
JcSSR_G27	JcCB0175451	Jcr4S01110	*LACS9*	Long-chain-fatty-acid CoA ligase	AT1G77590.1	29908.m006186	GAHK01014712.1	100	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G40_L1	JcCB0030361	Jcr4S00733	*LACS2*	long-chain acyl-CoA synthetase	AT1G49430.1	29851.m002473	GAHK01016749.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G356		Jcr4S05261	*LACS1*	Long-chain-fatty-acid CoA ligase	AT2G47240.2	30076.m004616	GW611327.1	27	100	Profiling gene expression of the reproductive organs of Jatropha curcas	Wang,W., Wei,B., Sing,P., Jin,Q.D., Wong,W.S., Zhang,S.H. and Li,N.	NCBI submission (2010)
JcSSR_G362		Jcr4S00096	*LACS4*	Long-chain acyl-CoA synthetase 4	AT4G23850.1	30190.m010831	GAHK01004006.1	97	92	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G363		Jcr4S00818	*LACS7*	long-chain acyl-CoA synthetase 7	AT5G27600.1	30128.m008777	GAHK01002069	94	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G367		Jcr4S00733	*LACS2*	long-chain acyl-CoA synthetase	AT1G49430.1	29851.m002473	GAHK01016749.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G368		Jcr4S05261	*LACS1*	Long-chain-fatty-acid	AT2G472	30076.m00461	GW61132	27	100	Profiling gene expression of the reproductive organs of Jatropha curcas	Wang,W., Wei,B., Sing,P.,	NCBI submission

				CoA ligase	40.2	6	7.1				Jin,Q.D., Wong,W.S., Zhang,S.H. and Li,N.	(2010)
3.2.4.1.7: Compartmentation and shuttling genes (*LPCAT, GPAT, LPAAT, PP, MAGAT, PDCT, DAG-CPT, PDAT, DGAT, DAGTA*)												
JcSSR_G41	JcCA0009631	Jcr4S00582	*GPAT3*	glycerol-3-phosphate acyltransferase	AT4G01950.1	30076.m004618	No evidence for expression					
JcSSR_G45	JcCA0084251	Jcr4S08388	*ATS1*	Plastid glycerol-3-phosphate acyltransferase	AT1G32200.2	30068.m002660	GAHK01004097.1	72	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G356		Jcr4S05261	*GPAT3*	glycerol-3-phosphate acyltransferase 3	AT4G01950.1	30076.m004618	GAHK01029922.1	70	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G357		Jcr4S00121	*GPAT3*	glycerol-3-phosphate acyltransferase 3	AT4G01950.1	29908.m005967	GT969993.1	30	99	Transcriptome analysis of the oil-rich seed of the bioenergy crop Jatropha curcas L	Costa,G.G.L., Cardoso,K.C., Del Bem,L.E.V., Lima,A.C., Cunha,M.A.S., de Campos-Leite,L., Vicentini,R., Papes,F., Moreira,R.C., Yunes,J.A., Campos,F.A.P. and Da Silva,M.J.	BMC Genomics 11 (1), 462 (2010)
JcSSR_G358		Jcr4S08802	*GPAT3*	glycerol-3-phosphate acyltransferase 3	AT4G01950.1	29908.m005967	GAHK01031716.1	19	90	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G364		Jcr4S00582	*GPAT3*	glycerol-3-phosphate acyltransferase	AT4G01950.1	30076.m004618	No evidence for expression					
JcSSR_G368		Jcr4S05261	*GPAT3*	glycerol-3-phosphate acyltransferase 3	AT4G01950.1	30076.m004618	GAHK01029922.1	70	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G369		Jcr4S01535	*GPAT8*	glycerol-3-phosphate acyltransferase 8	AT4G00400.1	30174.m008615	GAHK01015603.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G370		Jcr4S00361	*GPAT5*	glycerol-3-phosphate acyltransferase 5	AT3G11430.1	29736.m002070	No evidence for expression					
JcSSR_G371		Jcr4S00686	*GPAT6*	glycerol-3-phosphate	AT2G381	29736.m00207	No					

				acyltransferase 6	10.1	0	evidence for expression					
JcSSR_G373		Jcr4S11460	*GPAT7*	glycerol-3-phosphate acyltransferase 7	AT5G06090.1	27568.m000266	No evidence for expression					
JcSSR_G380		Jcr4S03010	*GPAT6*	glycerol-3-phosphate acyltransferase 6	AT2G38110.1	29969.m000267	GAHK01013941.1	70	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G427		Jcr4S01398	*ATS1*	seed gene 1	AT4G26740.1	30008.m000820	EZ409221.1	100	100	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G28	JcCA0153351	Jcr4S00343	*LPAT1*	Lysophosphatidic acid acyltransferase	AT4G30580.1	29687.m000572	GAHK01003393.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G29	JcCA0151201	Jcr4S01477	*LPAT4*	Lysophosphatidic acid acyltransferase	AT1G75020.2	30170.m013990	GAHK01019129.1	61	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G30	JcCB0037171	Jcr4S00971	*LPAT5*	Lysophosphatidic acid acyltransferase	AT3G18850.4	29851.m002448	GAHK01023279.1	51	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G365		Jcr4S00971	*LPAT5*	Lysophosphatidic acid acyltransferase	AT3G18850.4	29851.m002448	GAHK01023279.1	51	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G372		Jcr4S00017	*LPAT2*	1-acylglycerol-3-phosphate acyltransferase	AT3G57650.1	30169.m006432	No evidence for expression					
JcSSR_G374		Jcr4S01622	*LPAT1 (ATS2)*	Phospholipid/glycerol acyltransferase family protein	AT4G30580.1	29666.m001430	GAHK01020046.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G375		Jcr4S22362	*LPAT2*	lysophosphatidyl acyltransferase 2	AT3G57650.1	27810.m000646	GAHK01004915.1	95	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G376		Jcr4S01477	*LPAT4*	Lysophosphatidic acid acyltransferase	AT1G75020.2	30170.m013990	GAHK01019129.1	61	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G31	JcCA0311711	Jcr4S09416	*LPP3*	lipid phosphate phosphatase 3	AT3G02600.1	29586.m000620	GAHK01020421.1	78	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G377		Jcr4S02521	*LPP2*	lipid phosphate phosphatase 2	AT1G15080.1	29747.m001075	GAHK01017362.1	42	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G378		Jcr4S09416	*LPP3*	lipid phosphate	AT3G026	29586.m00062	GAHK010	78	100	Global Analysis of Transcriptome Responses and Gene Expression	Wang,H., Zou,Z., Wang,S.	PLoS ONE 8 (12), E82817

				phosphatase 3	00.1	0	20421.1			Profiles to Cold Stress of Jatropha curcas L	and Gong,M.	(2013)
JcSSR_G37	JcCB0022101	Jcr4S04735	*MCAAT*	Malonyl-CoA : ACP Acyltransferase (MCAAT)	AT2G30200.1	30113.m001448	GAHK01018914.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G32	JcCA0249341	Jcr4S00514	*ROD1 (PDCT)*	phosphatidic acid phosphatase-related / PAP2-related	AT3G15820.1	29841.m002865	GAHK01011904.1	97	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G382		Jcr4S00514	*PDCT (ROD1), FAD5*	phosphatidic acid phosphatase-related / PAP2-related	AT3G15820.1	29841.m002865	GAHK01011904.1	97	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G381		Jcr4S01903	*AAPT1 (DAG-CPT)*	Diacylglycerol Cholinephosphotransferase	AT1G13560.2	30138.m003845	GAHK01012179.1	95	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G385		Jcr4S00804	*PDAT1*	Phosphatidylcholine: Diacylglycerol Acyltransferase	AT3G44830.1	29706.m001305	GAHK01010861.1	98	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G388		Jcr4S01037	*PDAT1*	phospholipid:diacylglycerol acyltransferase	AT5G13640.1	29912.m005286	HQ827796.1	91	100	Expression profiles of genes involved in fatty acid and triacylglycerol synthesis in developing seeds of Jatropha (Jatropha curcas L.)	Xu,R., Wang,R. and Liu,A.	Biomass Bioenergy 35 (5), 1683-1692 (2011)
JcSSR_G403		Jcr4S08851	*PDAT2*	Phospholipid:diacylglycerol acyltransferase 2	AT3G44830.1	29991.m000626	No evidence for expression					
JcSSR_G404		Jcr4S19008	*PDAT2*	Phospholipid:diacylglycerol acyltransferase 2	AT3G44830.1	29991.m000626	No evidence for expression					
JcSSR_G432		Jcr4S01037	*PDAT1*	phospholipid:diacylglycerol acyltransferase	AT5G13640.1	29912.m005286	HQ827796.1	91	100	Expression profiles of genes involved in fatty acid and triacylglycerol synthesis in developing seeds of Jatropha (Jatropha curcas L.)	Xu,R., Wang,R. and Liu,A.	Biomass Bioenergy 35 (5), 1683-1692 (2011)
JcSSR_G379		Jcr4S00709	*DGAT3*	Acyltransferase-like protein	AT1G54570.1	30128.m008656	GAHK01002587.1	90	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G383		Jcr4S01935	*DGAT3*	transferases, transferring acyl groups other than amino-acyl groups;acyltransferases	AT3G02030.1	30131.m007010	GAHK01000294.1	99	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G384		Jcr4S04966	*DGAT3*	transferases, transferring acyl groups other than amino-acyl groups;acyltransferases	AT3G02030.1	30131.m007010	GAHK01000294.1	97	91	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G387		Jcr4U29423	*DGAT2*	diacylglycerol acyltransferase	AT3G51520.1	29682.m000581	JQ319813.1	100	99	Characterization of DGAT1 and DGAT2 from Jatropha curcas and their nonredundant functions in storage lipid biosynthesis	Xu,R. and Liu,A.	NCBI submission (2011)

JcSSR_G394		Jcr4S00406	*DGAT*	Diacylglycerol acyltransferase	AT1G54570.1	30128.m008656	GAHK01002587.1	95	85	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G395		Jcr4S03244	*DGAT*	Diacylglycerol acyltransferase	AT3G02030.1	30131.m007010	GAHK01000294.1	96	91	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G402		Jcr4S04033	*DGAT*	Phospholipid/glycerol acyltransferase family protein	AT1G80950.1	30170.m014002	GAHK01010064.1	81	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G43	JcCB0021271	Jcr4S15605, Jcr4S05766	*Pla2g4b*	Phospholipase A2, group IVB (cytosolic)	AT3G45880.1	29489.m000170	GAHK01037072.1	24	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G365		Jcr4S00971	*n/a*	phospholipase A-2-activating protein	AT3G18860.2	29851.m002449	GAHK01018279.1	71	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G422		Jcr4S02881	*DGD1*	UDP-Glycosyltransferase superfamily protein	AT3G11670.1	28726.m000069	GAHK01014347.1	99	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
3.2.4.1.8: Endoplasmic reticulum desaturases (*FAD2, FAD3*)												
JcSSR_G32		Jcr4S00514	*FAD5*	fatty acid desaturase 5	AT3G15850.1	29841.m002863	GAHK01015790.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G35		Jcr4S04563	*FAD8*	Omega-3 fatty acid desaturase, endoplasmic reticulum	AT5G05580.1	29681.m001360	EU267121.1	95	84	Functional characterization of two microsomal fatty acid desaturases from Jatropha curcas L	Wu,P., Zhang,S., Zhang,L., Chen,Y., Li,M., Jiang,H. and Wu,G.	J. Plant Physiol. 170 (15), 1360-1366 (2013)
JcSSR_G35	JcCA0269921	Jcr4S27172	*FAD3/7*	Microsomal omega-3 fatty acid desaturase	AT3G11170.1	29681.m001360	EU267121.1	67	99	Identification and characterization of a novel microsomal omega-3 fatty acid desaturase from Jatropha curcas L.	Wu,P.Z., Li,J., Li,M.R., Jiang,H.W. and Wu,G.J.	NCBI submission (2007)
JcSSR_G390		Jcr4S09407	*FAD2*	fatty acid desaturase 2	AT3G12120.2	29613.m000358	EZ409947.1	100	98	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G391		Jcr4S01187	*FAD3*	Microsomal omega-3 fatty acid desaturase	AT5G05580.1	29681.m001360	EU267121.1	100	97	Functional characterization of two microsomal fatty acid desaturases from Jatropha curcas L	Wu,P., Zhang,S., Zhang,L., Chen,Y., Li,M., Jiang,H. and Wu,G.	J. Plant Physiol. 170 (15), 1360-1366 (2013)
JcSSR_G392		Jcr4S01217	*FAD4*	fatty acid desaturase A	AT4G27030.1	29666.m001456	GAHK01013623.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G397		Jcr4S03307	*FAD8*	fatty acid desaturase 8	AT5G05580.1	29814.m000719	DQ452089.1	100	100	A plastidial omega-3 fatty acid desaturase from Jatropha curcas	Guo,L., Qing,R., Huang,M., He,W., Xu,Y., Tang,L. and Chen,F.	NCBI submission (2006)

JcSSR_G405		Jcr4S02981	*FAD3*	fatty acid desaturase 3	AT2G29980.1	29681.m001360	EU267121.1	100	94	Functional characterization of two microsomal fatty acid desaturases from Jatropha curcas L	Wu,P., Zhang,S., Zhang,L., Chen,Y., Li,M., Jiang,H. and Wu,G.	J. Plant Physiol. 170 (15), 1360-1366 (2013)
JcSSR_G406		Jcr4S04563	*FAD8*	Omega-3 fatty acid desaturase, endoplasmic reticulum	AT5G05580.1	29681.m001360	EU267121.1	95	84	Functional characterization of two microsomal fatty acid desaturases from Jatropha curcas L	Wu,P., Zhang,S., Zhang,L., Chen,Y., Li,M., Jiang,H. and Wu,G.	J. Plant Physiol. 170 (15), 1360-1366 (2013)
JcSSR_G410		Jcr4S00514	*FAD5*	fatty acid desaturase 5	AT3G15850.1	29841.m002863	GAHK01015790.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G414		Jcr4S03452	*FAD6*	fatty acid desaturase 6	AT4G30950.1	29696.m000105	EU106889.1	95	100	Identification and characterization of a novel chloroplast omega-6 fatty acid desaturase from Jatropha curcas L.	Wu,P.Z., Li,J., Li,M.R., Jiang,H.W. and Wu,G.J.	NCBI submission(2007)
3.2.4.1.9: Seed oil body-associated storage protein genes; Oleosins and Caleosins												
JcSSR_G412		Jcr4S06252	Oleosin	Oleosin family protein	AT3G18570.1	30174.m008728	GW619162.1	100	100	Profiling gene expression of the reproductive organs of Jatropha curcas	Wang,W., Wei,B., Sing,P., Jin,Q.D., Wong,W.S., Zhang,S.H. and Li,N.	NCBI submission (2010)
JcSSR_G420		Jcr4S01276	*OLEO1*	oleosin 1	AT4G25140.1	30147.m014333	EZ417041.1	100	100	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G423		Jcr4S05922	Caleosin	Caleosin-related family protein	AT1G70670.1	29673.m000932	GAHK01023742.1	97	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G424		Jcr4S28232	Oleosin	Oleosin family protein	AT2G25890.1	29794.m003372	EZ412177.1	100	100	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G428		Jcr4S00534	Oleosin	Oleosin family protein	AT2G25890.1	29794.m003372	EZ412177.1	96	100	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G429		Jcr4S05992	Oleosin	Oleosin family protein	AT3G01570.1	29917.m001992	EZ418548.1	100	100	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
3.2.4.1.10: Fatty acid synthesis master regulators												
JcSSR_G425		Jcr4S00084	*WRI1*	Integrase-type DNA-binding superfamily protein	AT3G54320.3	29736.m002029	No evidence for expression					
JcSSR_G426		Jcr4S07197	*WRI1*	Integrase-type DNA-binding superfamily protein	AT3G54320.3	29736.m002029	No evidence for expression					

JcSSR_G430		Jcr4S03855	*WRI1*	Integrase-type DNA-binding superfamily protein	AT3G54320.3	30069.m000440	JF703666.1	100	99	Isolation and characterization of JcWRI1 gene from Jatropha curcas	Zhang,L., Yang,Z. and Shen,S.	NCBI submission (2011)
JcSSR_G431		Jcr4S05417	*WRI1*	Integrase-type DNA-binding superfamily protein	AT3G54320.3	29822.m003477	No evidence for expression					
3.2.4.2: Branching candidate genes												
JcSSR_G56	JcCA0154071.10	Jcr4S00260	*TIR1*	TRANSPORT INHIBITOR RESPONSE 1 (F-box/RNI-like superfamily protein)	AT3G62980.1	29647.m002022	GAHK01014281.1	100	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G57	JcCA0074811.10	Jcr4S05147	*AFB5*	auxin F-box protein 5	AT5G49980.1	29908.m006223	GAHK01004749.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G58	JcCA0133141.20	Jcr4S00529	*ABI3*	AP2/B3-like transcriptional factor family protein	AT3G24650.1	30204.m001803	EZ411848.1	57	98	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G59	JcCB0165711.10	Jcr4S02351	*MAX2*	RNI-like superfamily protein	AT2G42620.1	29451.m000049	GAHK01032867.1	16	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G60	JcCA0141181.10	Jcr4S16834	*MAX4*	carotenoid cleavage dioxygenase 8	AT4G32810.1	29794.m003382	No evidence for expression					
JcSSR_G61_A1	JcCA0306791.10	Jcr4S22672	*PIN1*	Auxin efflux carrier family protein	AT1G73590.1	29651.m000296	GAHK01024416.1	73	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G428		Jcr4S00534	*MAX4*	carotenoid cleavage dioxygenase 8	AT4G32810.1	29794.m003382	No evidence for expression					
JcSSR_G435		Jcr4S05086	*AXR1*	NAD(P)-binding Rossmann-fold superfamily protein	AT1G05180.1	29600.m000552	GAHK01025428.1	68	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G436		Jcr4S00105	*TIR1*	TRANSPORT INHIBITOR RESPONSE 1 protein	AT3G62980.1	29933.m001427	GAHK01007122.1	79	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G437		Jcr4S03655	*AFB2*	auxin signaling F-box 2	AT3G26810.1	30131.m006863	GAHK01014896.1	99	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G438		Jcr4S00335	*MAX1*	Fatty acid/sphingolipid desaturase	AT2G46210.1	29794.m003308	EF208109.1	100	100	Characterization of D8-sphingolipid desaturase from Jatropha curcas	Qing,R., Guo,L. and Chen,F.	NCBI submission

												(2007)
JcSSR_G439		Jcr4S01087	*MAX2*	RNI-like superfamily protein	AT2G42620.1	29682.m000601	GAHK01022914.1	45	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G440		Jcr4S01272	*MAX3*	carotenoid cleavage dioxygenase 7	AT2G44990.1	30174.m008796	No evidence for expression					
JcSSR_G441		Jcr4S10204	*AtPIN1*	Auxin efflux carrier family protein	AT1G73590.1	30180.m001064	No evidence for expression					
3.2.4.2: Flower ratio genes												
JcSSR_G46	JcCB0031471	Jcr4S00778	*PI* like	K-box region and MADS-box transcription factor family protein	AT5G20240.1	29648.m001978	GW614527.1	100	100	Profiling gene expression of the reproductive organs of Jatropha curcas	Wang,W., Wei,B., Sing,P., Jin,Q.D., Wong,W.S., Zhang,S.H. and Li,N.	NCBI submission (2010)
JcSSR_G47	JcCB0402301	Jcr4S01776	*SHP2* like	K-box region and MADS-box transcription factor family protein	AT2G42830.2	30026.m001501	EZ417572.1	88	100	Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing	King,A.J., Li,Y. and Graham,I.A.	Bioenergy Res (2011)
JcSSR_G48	JcCB0462581	Jcr4S00750	*LOX5*	PLAT/LH2 domain-containing lipoxygenase family protein	AT3G22400.1	30178.m000859	GAHK01038599.1	10	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G49	JcCB0292551.10	Jcr4S06599	*LOX2*	lipoxygenase 2	AT3G45140.1	30152.m002449	GW618166.1	27	100	Profiling gene expression of the reproductive organs of Jatropha curcas	Wang,W., Wei,B., Sing,P., Jin,Q.D., Wong,W.S., Zhang,S.H. and Li,N.	NCBI submission (2010)
JcSSR_G50_A1	JcCA0044701.10	Jcr4S09208	*LOX5*	PLAT/LH2 domain-containing lipoxygenase family protein	AT3G22400.1	29726.m003891	GAHK01003986.1	58	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G52	JcCA0308701.10	Jcr4S00435	Lipooxygenase	Lipase/lipooxygenase, PLAT/LH2 family protein	AT4G39730.1	30147.m014251	GW617962.1	97	100	Profiling gene expression of the reproductive organs of Jatropha curcas	Wang,W., Wei,B., Sing,P., Jin,Q.D., Wong,W.S., Zhang,S.H. and Li,N.	NCBI submission (2010)
JcSSR_G53_A1	JcCA0317961.10	Jcr4S03289	Lipooxygenase	Lipase/lipooxygenase, PLAT/LH2 family protein	AT2G22170.1	28206.m000101	GO247614.1	96	97	Expressed sequence tags from Jatropha curcas root cDNA library	Nalini,E., Parmeshwaran,S., Balaji,S., Bhagyam,A. and Johnson,T.S.	NCBI submission (2009)
JcSSR_G54	JcCA0154881	Jcr4S00285	*LAS, AGAMOUS*-like	GRAS family transcription factor	AT1G55580.1	28966.m000535	GAHK01043718.1	10	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G55	JcCB0046	Jcr4S01162	*MOC1*-	Mitochondrial	AT4G146	30190.m01096	GAHK010	80	100	Global Analysis of Transcriptome	Wang,H.,	PLoS ONE 8

	511.10		like	transcription termination factor family protein	05.1	5	20523.1			Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Zou,Z., Wang,S. and Gong,M	(12), E82817 (2013)
JcSSR_G58		Jcr4S00529	*ADH1*	alcohol dehydrogenase	AT1G77120.1	27985.m000885	GT969949.1	30	99	Transcriptome analysis of the oil-rich seed of the bioenergy crop Jatropha curcas L	Costa,G.G.L., Cardoso,K.C., Del Bem,L.E.V., Lima,A.C., Cunha,M.A.S., de Campos-Leite,L., Vicentini,R., Papes,F., Moreira,R.C., Yunes,J.A., Campos,F.A.P. and Da Silva,M.J.	BMC Genomics 11 (1), 462 (2010)
JcSSR_G363		Jcr4S00818	*LOX1*	lipoxygenase 1	AT1G55020.1	30128.m008781	GAHK01014084.1	100	100	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M.	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G433		Jcr4S06869	*LOX3*	lipoxygenase 3	AT1G17420.1	29468.m000030	GAHK01018559.1	50	99	Global Analysis of Transcriptome Responses and Gene Expression Profiles to Cold Stress of Jatropha curcas L	Wang,H., Zou,Z., Wang,S. and Gong,M	PLoS ONE 8 (12), E82817 (2013)
JcSSR_G434		Jcr4S00095	*ATA1*	Short chain alcohol dehydrogenase	AT3G42960.1	29780.m001371	No evidence for expression					

Chapter 4: Linkage mapping in an F_2 population derived from parents with high and low seed oil phenotypes

Genetic linkage mapping uses genetic markers to measure recombination frequency, or crossover events, that occur during DNA replication. The rate of recombination is proportional to the genetic distance between the markers; completely unlinked markers recombine in a diploid species 50 % of the time, whereas physically linked markers have a lower recombination frequency proportional to physical distance. Each F_2 plant in the mapping population is effectively a single measurement of recombination between the available DNA markers. Therefore the larger the F_2 population, the greater the sample size for estimating recombination frequency and the greater the accuracy of the inferred genetic linkage map.

Since DNA replication occurs independently from external, environmental conditions, it is not necessary for the F_2 plants to experience the same environmental conditions when measuring recombination frequency for genetic linkage mapping. This is in contrast to QTL mapping that is dependent on phenotypic measurements; a function of Genetic and Environmental factors (GxE). This property has been exploited to increase sample size in several ways in this project. Firstly individual mapping populations (such as the principle population under study for this thesis, G51xCV) has F_2 plants created over 2 rounds of crossing. Secondly the G51xCV data has been combined with 3 other mapping populations in order to create a combined dataset and genetic linkage map. This was possible due to all populations using the same DNA marker set.

This results chapter will have two narrative strands. Genetic linkage mapping in the G51xCV mapping population will be the central focus, as this was the principle mapping population under study for this thesis. This will also set the conditions for the G51xCV QTL analysis in the proceeding chapter.

The other strand will look at the genetic linkage mapping process overall. It will analyse the DNA marker set to determine individual marker characteristics and performance during the genetic linkage mapping process. It will also present the combined genetic linkage map, which aside from being the ultimate product of genetic data collected from the G51xCV mapping population, also acts as a reference map with which to compare this thesis linkage map. This comparison was used in this thesis work for comparative mapping strategies and to assess the quality and robustness of G51xCV results against the larger combined dataset.

4.1: Genetic linkage mapping in G51xCV

4.1.1: The G51xCV $\underline{F_2}$ mapping population has a complex population structure, due to heterozygosity in G51, and the asynchronous, self-compatible flowering strategy of *J. curcas*

The G51xCV mapping population was created from a cross between a homozygous parent ('CV') and a heterogeneously heterozygous parent ('G51'), containing 36.5 % heterozygosity. In order to map as many of these heterozygous loci as possible, a reciprocal sib-cross of two non-uniform F_1's was carried out, with alternate F_1 plants acting as the mother plant for each direction of the cross.

Further complexity was added to this population structure due reproductive characteristics of Jatropha. *J. curcas* is self-compatible, with male and female flowers found on the same inflorescence. Whilst Jatropha is protogynous; female flowers reach maturity before males flowers of the same inflorescence, the plant as a whole is asynchronous; inflorescences on the same plant flower at different times. With mature male and female flowers found on different parts of the plant during the flowering cycle, there is opportunity for self-pollination to occur (Achten et al., 2010).

Therefore whilst an F_1 sib-cross was the intended crossing strategy for the F_2 population, self-pollination was also present. In each direction of the reciprocal sibcross, the mother plant self-fertilised at a high rate, and as a result, instead of two F_2 subpopulations derived from the F_1 sibcross in each direction, a further 2 subpopulations were created from selfing of each F_1 mother plant.

This cross was repeated on two separate occasions in order to generate more F_2 plants for linkage mapping, therefore in total this population contained 8 subpopulations for genetic linkage mapping. Please refer to Figure 2-1 (materials and methods), for an illustration of the intended and actual population structures.

4.1.2: Heterozygosity in G51 enabled population structure to be determined through the use of informative marker loci

The complex population structure of G51xCV as described in fig 2-1, was determined through informative marker analysis. Whilst selfing of genetically uniform F_1 plants would be identical to sib-crossing, selfing and sib-crossing of two heterogeneously heterozygous (non-uniform) F_1 plants would have different outcomes on the genetics of the F_2 offspring. To illustrate the non-uniformity of the F_1 plants, $F_{1.1}$ was 86 % heterozygous, whereas $F_{1.2}$ was 82.5 % heterozygous.

Ironically, the heterozygosity in G51 that gave rise to this complex population structure, also created informative marker loci that enabled this population structure to be elucidated. Please refer to Figure 2-2 (materials and methods) which explains what informative marker loci are, and how they were used to inform the population structure and parentage of each F2 plant in this project.

4.1.3: Heterozygosity in G51 is likely to represent underlying genetic similarity to CV, rather than non-informative marker loci, therefore heterozygous loci have been included to maximise accuracy of downstream linkage and QTL mapping

A model F_2 population is created from two homozygous, genetically-distinct lines e.g. Parent 1 would be all 'aa' genotype, whereas parentage 2 would be all 'bb' genotype. In this population the parent G51 was partially 'ab'.

There are two implications for the heterozygosity in G51. Either these 'a' alleles indicate that the underlying genetic region is identical to the CV parent, in which case the genotype of these markers is fully informative, or that the underlying genetic region is different in the two parents and the markers developed for these regions are only 50 % informative. In either scenario the parent of origin cannot be distinguished in the F_2 generation, however if the marker allele is informative of the underlying genetic region/alleles, as in the first scenario, then the markers can be used to measure the association of genetic alleles to phenotypic measurements during QTL analysis, and therefore provide value for QTL mapping.

All markers used in this study are co-dominant, however SSRs provide an additional level of informativeness due to their continuous rather than digital nature in comparison to SNPs. SNPs can have a maximum of 4 alleles (A,C,T,G), whereas SSRs can potentially have any number of repeats, giving greater ability to differentiate the genetic origin of particular sequences. It is because of this that the first scenario would seem more likely; that the heterozygous regions in the G51 parents are 50 % identical to the CV parent (rather than the loci being completely genetically distinct but just the marker scores being shared). If the heterozygous regions in the G51 parent were heterozygous regions originating from a completely genetically distinct line from the CV parent, the chance that all heterozygous loci would have identical SSR lengths (or alleles) to the CV parent by chance is low (given that there are 93 informative loci that share the identical number of SSR repeats as the CV parent).

It is therefore more likely that the G51 parent shares some genetic relatedness to the CV parent, a scenario supported by the small centre of origin for this species and the lack of genetic diversity observed so far in comparison to most other (despite being more highly cultivated) crop species. Therefore the marker scores at these loci would seem fully informative in terms of the underlying genetic alleles present, and are included in downstream analysis for linkage and QTL mapping.

4.1.4: The G51xCV genetic linkage map, derived from 229 F2 plants, comprises 312 co-dominant DNA markers spread over 11 linkage groups

Genetic linkage mapping in the G51xCV mapping population was carried out using 229 F_2 plants, and 312 SNPs and SSR markers. Correct parentage of F_2 plants was ascertained using informative marker loci as described above, and linkage mapping carried out using Crimap software, which is able to incorporate the complex population structure of the G51xCV mapping population. A robust quality control and error checking process was also incorporated to maximise robustness of the linkage map. See 'Materials and Methods' for a detailed description of the Crimapping procedure.

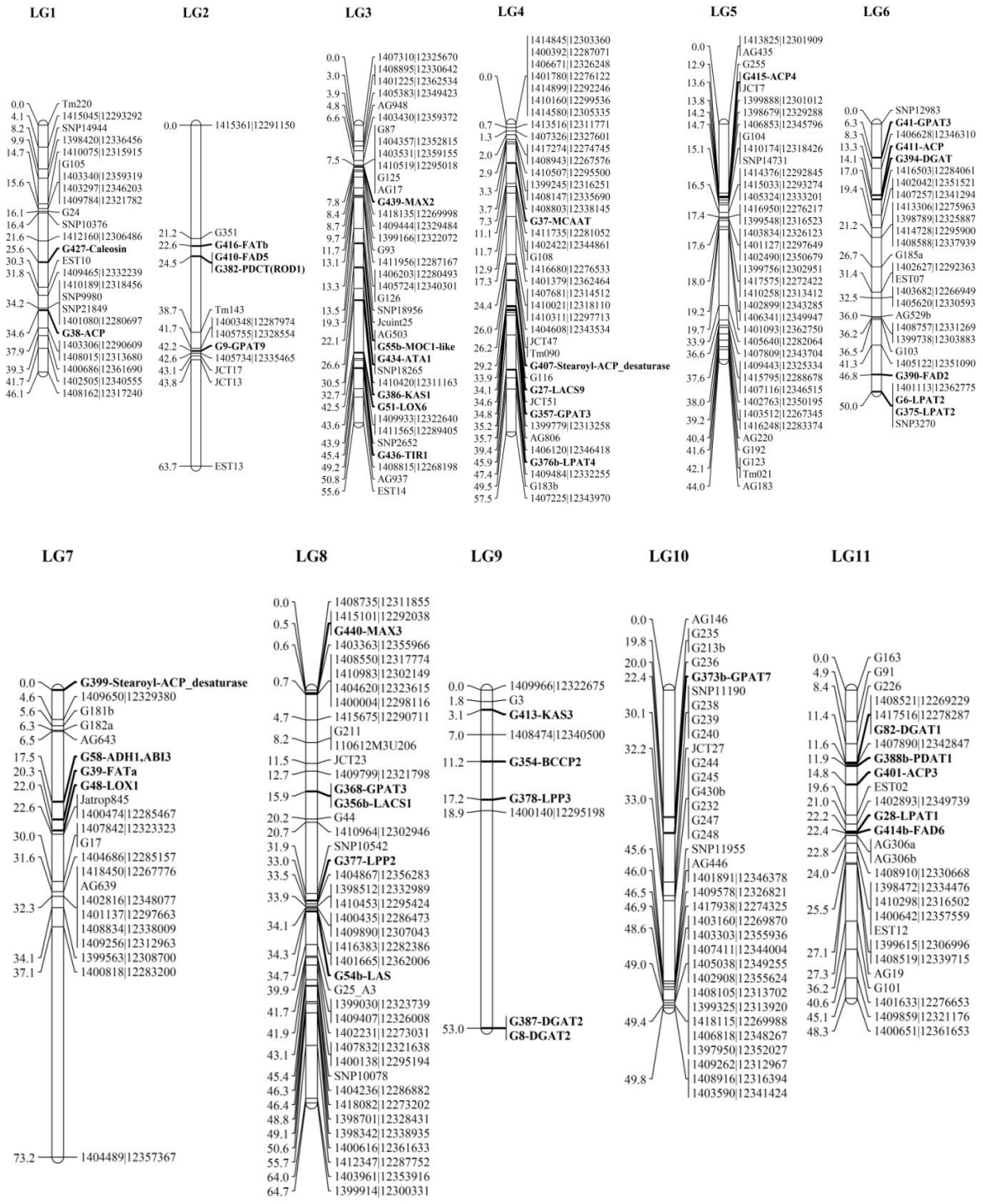


Figure 4-1 The G51xCV Genetic Linkage Map, Linkage groups 1-11.

The 51xCV linkage map is composed of 312 co-dominant markers, distributed over 11 linkage groups with a genome wide marker density for unique loci of 2.969 cM. Markers physically linked to candidate genes have been highlighted in bold. In this population, 44 candidate genes have been mapped for oil content, oil composition, branching and flower ratio traits.

Table 4-1 The G51xCV genetic linkage map statistics.

Summary statistics for the G51xCV linkage map are presented below. The G51xCV linkage map consisting of 312 co-dominant markers, 11 linkage groups and a total genetic distance of 621 cM, has a mean marker density of between 2.21 cM (LG04) and 7.58 cM (LG09) for unique loci.

Linkage Group		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**All**
Markers	SSRs	6	9	15	12	10	10	10	9	6	16	14	117
	SNPs	15	5	18	27	27	15	12	30	3	16	13	181
	EST SNPs	4	0	3	0	1	2	0	2	0	2	0	14
	Total	**25**	**14**	**36**	**39**	**38**	**27**	**22**	**41**	**9**	**34**	**27**	**312**
Total Loci		18	12	26	27	21	17	15	31	8	15	20	210
Total Distance (cM)		**46.1**	**78.3**	**55.6**	**57.5**	**44.0**	**50.0**	**73.2**	**64.7**	**53.0**	**49.8**	**48.3**	**621**
Marker Density (cM)	All markers	**1.92**	**6.02**	**1.59**	**1.51**	**1.19**	**1.92**	**3.48**	**1.62**	**6.63**	**1.51**	**1.86**	**2.00**
	Unique loci	**2.71**	**7.12**	**2.23**	**2.21**	**2.20**	**3.13**	**5.23**	**2.16**	**7.58**	**3.56**	**2.54**	**2.97**

Table 4-1 shows the G51xCV genetic linkage map summary statistics. The G51xCV linkage map consisted of 312 co-dominant markers, spread over 11 linkage groups and covering a total genetic distance of 621 cM. Marker density is 2.97 cM per unique locus across all linkage groups, and ranges from 2.21 cM (LG04) to 7.58 cM (LG09) per unique loci in individual linkage groups. Markers physically linked to candidate genes identified from Jatropha genome sequence (Sato et al., 2011) have been highlighted in bold**,** Figure 4-1. In the 51xCV population, 44 candidate genes have been mapped for oil content, oil composition, branching and flower ratio traits.

The 11 linkage groups presented here are in agreement with cytological evidence on *J. curcas* chromosome number (Carvalho et al., 2008) and a previously published interspecific linkage map (Wang et al., 2011). Based on the total genetic distance mapped of 621 cM, and cytological evidence suggesting the Jatropha genome size to be 416 Mbp (Carvalho et al., 2008), a genetic distance of 1 cM is corresponds to approximately 0.7 Mbp or 700Kbp on this map.

4.1.5: Physical alignment of the G51xCV linkage map, to independent mapping populations and the combined population linkage map, confirms mapping accuracy and genome coverage for G51xCV

Figure 4-7 to 4-19, Appendix 4.6.2, shows the physical alignment of the G51xCV, and other population linkage maps that together were used to build the combined linkage map - an example of which is presented below, Figure 4-2. This alignment was made possible by the fact that all 4 populations used the same DNA marker set. Markers that were mapped in multiple populations have been connected by black lines to facilitate comparison.

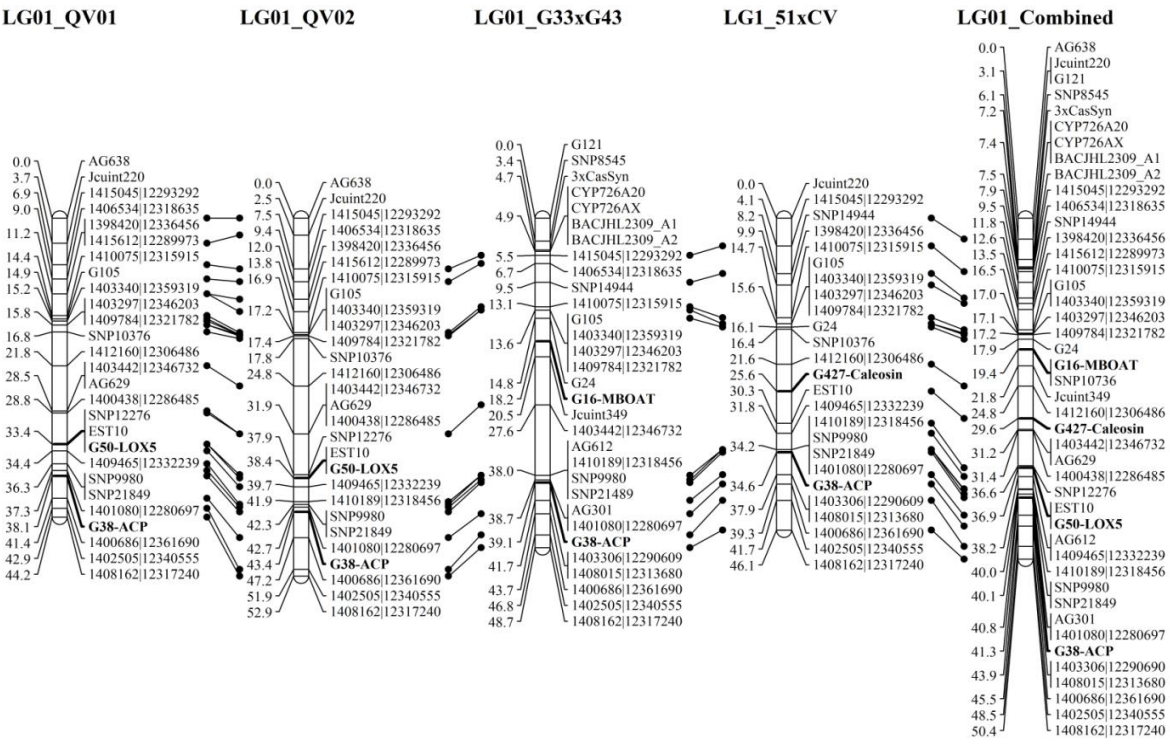


Figure 4-2 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.

Alignment of independent mapping population linkage maps was used as a way to confirm accuracy of individual maps. Linkage groups are annotated with the format (*linkage group number*)_(*Mapping population*); the above is an example of linkage group 1 alignment (LG1) for populations QV01, QV02, G33xG43, G51xCV (the principle population for this thesis study), and the combined genetic linkage map. Markers that were mapped in more than one population have been connected with dark black lines to aid comparison. As can be seen by the connecting lines, marker order and spacing is highly conserved across individual linkage groups, indicating that the markers and datasets generated from independent mapping populations are robust and accurate. Marker spacing (genetic distance) is likely to be most accurate in the combined genetic map since this represents the largest sample size from which recombination frequency has been calculated (a total of 989 F2 plants). As with marker order, genetic distance is also highly conserved across independent mapping populations. Comparisons of linkage groups 1-11 are including in the appendix.

Physical alignment enables the level of consensus between individual linkage maps to be established, and by doing so gives an indicator of the relative accuracy and robustness of individual population datasets. This comparison works on the principle that marker order and recombination rate should be conserved across each population, since the physical position of the DNA markers are constant in the Jatropha genome. Overall one can see that marker order and spacing is highly conserved, and for G51xCV, indicates that this is an accurate linkage map that contributes robust data to the combined genetic linkage map, as well as providing the foundation for accurate QTL mapping in Chapter 5. This alignment technique was used to carry out comparative mapping during later rounds of linkage mapping as outlined in the materials and methods of this thesis.

4.1.6: Quantification of gaps on linkage maps highlights regions requiring further mapping and also suggests areas of low polymorphism and regions identical by descent

Physical alignment of individual maps, as presented above, enables visual comparison of marker order, spacing and coverage between individual linkage maps and the combined genetic linkage map and also

highlights areas of low marker density for future improvement of the maps. Whilst this visual comparison enables an overall impression to be gained, a quantitative approach enables gaps to be identified more systematically, according to predetermined thresholds.

For example it is reckoned that a spacing of 10-20 cM between markers represents a robust marker coverage that is capable of capturing all QTL effects. Similarly for accurate genetic linkage mapping, double crossovers are thought unlikely to occur within distances smaller than 15 cM due to crossover interference.

Quantification of internal gaps can be done by map position of flanking markers. For missing regions at the end of linkage groups, comparison to the combined linkage map has been carried out, since this represents the most accurate and robust linkage map based on sample size (989 F2 plants). To carry out this comparison, the position of the nearest shared end marker between the individual and combined map is used as the reference point. The distance of the marker to the end of the linkage group on the two linkage maps gives an indication of the amount of the missing region on the individual map, compared to the combined map.

Table 4-2 Quantitative analysis of the number of regions in individual population maps that could be targeted with additional markers

The table below shows regions that could be targeted with additional markers and highlights that certain regions show low polymorphism across multiple independent mapping populations, suggesting regions identical by descent. Gaps are measured by flanking markers in the population map containing the gap, whereas end regions are calculated by comparison to the combined genetic map, since this represents the best estimation of genetic distance. Numbers are in cM, and symbols represent location; G = internal gap, T = top of linkage group, B = bottom of linkage group. Regions of low marker density across all maps, suggesting regions identical by descent, are highlighted in red.

	Mapping Population			
Linkage group	**G51xCV**	**G33xG43**	**QV01**	**QV02**
Map author	JG Clarke	AJ King	AJ King	AJ King
1				
2	12.9 B 21.2 G 19.9 G	20.3 G	19.8 B 23.1 G	19.8B 23.1 G
3		36.3T		
4				
5				
6				
7	36.1G	12.7B		
8				
9	14.4T 34.1G	16.3G		
10	19.8G	16.9G	17.9G	
11		15.5G		
Count	7	6	3	2

As can be seen by the table, alignment of linkage maps using shared markers has enabled regions that require further mapping to be identified. Quantitative analysis of the G51xCV map shows that there are just 3 regions spanning greater than 20 cM, and 5 regions greater than 15 cM (LG 2,7,9 & 10), and so the probability of missing double crossover events, and therefore underestimating genetic distance, is very low for most areas of this linkage map. Similarly, high significance QTL are expected to be detected by this marker density, although further mapping is required to ensure low significance QTL are not present in the larger gaps on linkage groups 7 & 9.

It is interesting to note that many of the identified regions tend to cluster around the same linkage groups in different populations (highlighted red), suggesting that these regions are areas of low intraspecific polymorphism or regions identical by descent from a shared ancestral line. Comparative mapping using aligned linkage maps from individual populations, and comparative mapping using castor bean microsynteny, did not yield any polymorphic SSR markers in these regions. Similarly, randomly or unbiasedly selected genotyping by sequencing SNPs did not fall in these regions by chance. This would suggest that these could be low polymorphism regions (similar in both parental lines), or so called 'identical by descent' regions (i.e. they have remained unchanged from the common ancestral line). Identical by descent regions are thought to be maintained by stabilising or positive selection, often at a very early stage in plant development such that they can persist through the selective pressures during a plant's life cycle (Jordan et al., 2005).

4.1.7: In addition to identifying regions of low marker density in G51xCV, comparative mapping also highlights isolated markers that are accurate, that otherwise would have been excluded during the genetic linkage mapping process.

The 5 regions greater than 15 cM in G51xCV, are all present towards the end of linkage groups, and are marked by single markers in some cases. Such markers would normally be suggestive of an erroneous marker showing a spurious linkage, however alignment of maps from other mapping populations suggest that these markers are correct, as either these particular markers have been mapped to the same region in an independent mapping population, or they map a region covered by different markers in another population. For this reason they have been included on the G51xCV map.

4.2: Incorporation of G51xCV data and SSR markers, contributes towards the combined genetic linkage map; a robust and comprehensive linkage map for *J. curcas*

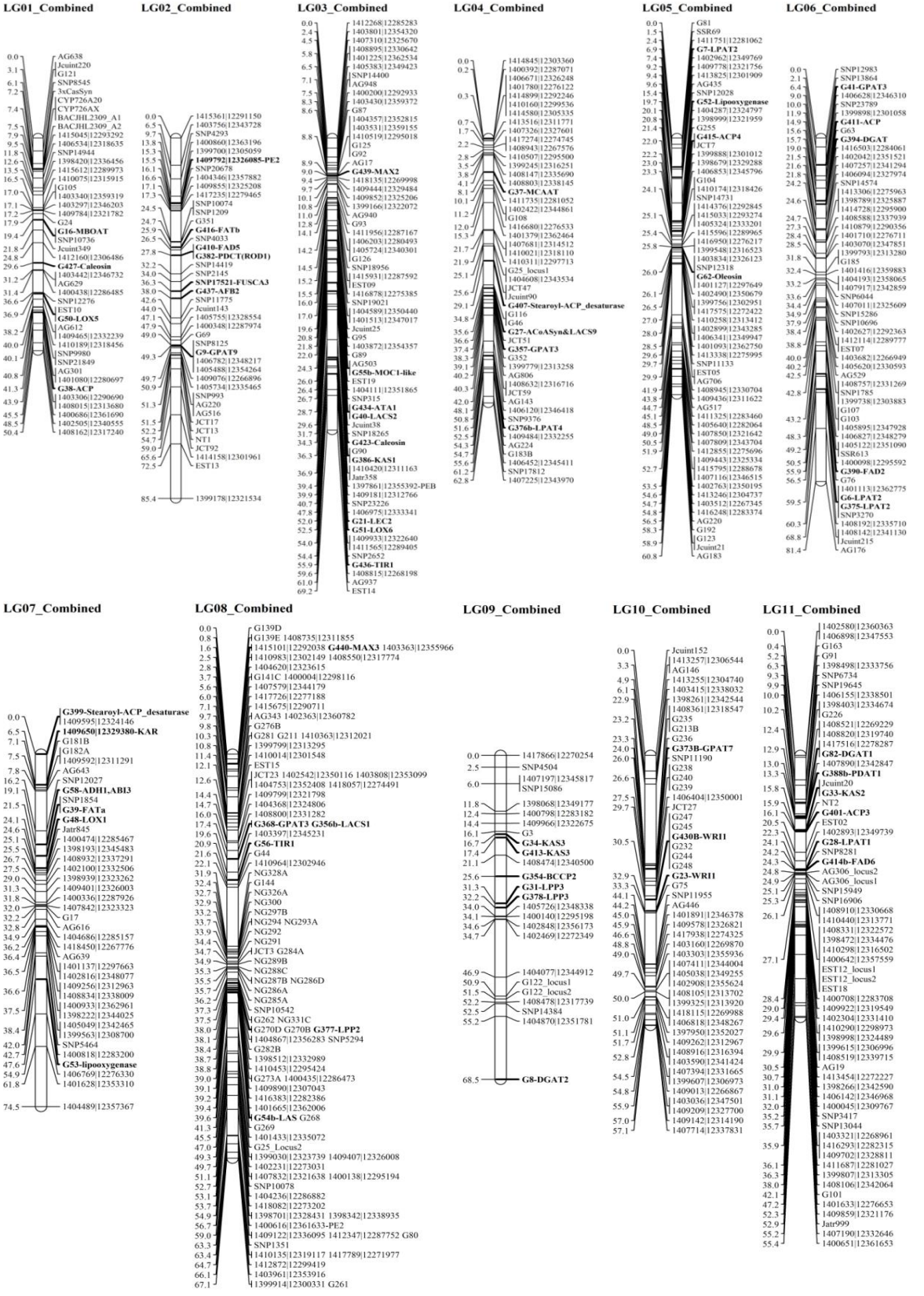


Figure 4-3 The Combined Genetic Linkage map derived from four F_2 mapping populations (989 F2 plants).

In total, 589 co-dominant markers map 11 linkage groups and a genetic distance of 733 cM. Average marker density is 1.62 cM per unique loci. A total of 67 candidate genes and trait-related metabolic genes are mapped (highlighted in bold).

Table 4-3 The Combined Linkage Map; Marker and Map Statistics.

Summary statistics for the combined genetic linkage map. The 594 co-dominant markers, spread over 11 linkage groups, map a total distance of 733 cM. Genome wide marker density for unique loci is 1.62 cM, individual linkage group marker densities range from 0.96 cM (LG11) to 2.98 cM (LG09).

Linkage group		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**All**
Markers	SSRs	19	15	28	18	18	16	12	39	9	19	20	213
	SNPs	19	16	31	29	40	31	25	48	13	29	37	318
	EST SNPs	6	11	7	2	4	9	3	4	3	2	7	58
	Total	**44**	**42**	**66**	**49**	**62**	**56**	**40**	**91**	**25**	**50**	**64**	**589**
Total loci		34	35	52	39	49	43	34	71	24	34	49	464
Total distance (cM)		**50.4**	**85.5**	**69.2**	**62.8**	**60.8**	**81.4**	**74.5**	**67.2**	**68.5**	**57.1**	**55.4**	**733**
Marker density (cM)	All markers	**1.17**	**2.08**	**1.07**	**1.31**	**1.00**	**1.48**	**1.91**	**0.74**	**2.86**	**1.16**	**0.88**	**1.24**
	Unique loci	**1.53**	**2.51**	**1.36**	**1.65**	**1.27**	**1.94**	**2.26**	**0.96**	**2.98**	**1.73**	**1.15**	**1.62**

Table 4-3 and Figure 4-3, show the combined genetic linkage map and statistics. As can be seen 589 co-dominant markers were mapped across 11 linkage groups, with a mean marker density of 1.62 cM per unique locus. The lowest marker density for individual linkage groups was 2.98 cM (LG09). This map was made from 4 independent F_2 mapping populations, each containing between 220 to 320 F_2 plants. Since all four mapping populations used the same set of DNA markers, recombination data for the combined map was calculated from a combined population size of 989 F_2 plants. Markers physically linked to candidate and trait-related metabolic genes have been highlighted in bold. In total 67 candidate gene markers were mapped.

4.3: DNA Marker analysis

4.3.1: DNA markers used throughout the genetic linkage mapping process show differing performance

During the course of genetic linkage mapping for this project, a number of properties were of interest relating to how efficiently a DNA marker could be mined, genotyped and mapped. In this section, the DNA marker set will be analysed to obtain insight into marker performance within this genetic mapping process.

Table 4-4 Comparison of EST SNP and SSR marker performance

(a) Marker Type					
Marker type	**EST SNPs**	**Short Sequence Repeats (SSRs)**			
Author	AJ King	All	AJ King	JG Clarke	R Santos
Number of marker primers designed	**104**	**683**	**371**	**288**	**24**
(b) Marker outcome					
Polymorphic	73	317	190	115	12
Not polymorphic	5	263	114	147	2
Failed	22	82	57	15	10
Not tested	4	21	10	11	0
Total	104	683	371	288	24
Mapped	**58**	**213**	**129**	**74**	**10**
(c) Success rate (%)					
Fail rate	**22**	**12.4**	**15.8**	**5.42**	**41.7**
Polymorphism rate	**93.6**	**54.7**	**62.5**	**43.9**	**85.7**
Map rate	**79.5**	**67.2**	**67.9**	**64.3**	**67.2**

Table 4-6, gives attribution for the DNA markers developed, and enables a comparison of EST SNP and SSR marker performance during the linkage mapping process. Part (a) lists the number of marker sequences mined and primers designed over the course of the project. In total 104 EST SNP markers were mined and designed (Andy King), of which 73 were polymorphic and 58 were successfully mapped (Andy King, J Clarke). In total 683 SSR marker sequences were mined and primers designed (Andy King, Jasper Clarke, Roberto Santos), of which 317 were polymorphic and 129 were successfully mapped (Andrew King, Jasper Clarke). Individual attribution for each marker type is given in the table. Part (c) highlights the success rate of markers throughout the linkage mapping process. Fail rate measured as 'Total failed/Total tested (Total tested = Total designed-Not tested)', looks at the proportion of markers that failed PCR amplification. This can therefore be thought of as a reflection of both the accuracy of the target sequence on which the primers were designed, and the success of the primer design since both PCR cycling conditions and PCR reagents were kept constant throughout this work. Polymorphism rate, measured as 'Total Polymorphic/Total successfully tested (Total successfully tested = Polymorphic + Not polymorphic)', looks at the proportion of markers that, following successful PCR amplification, were found to be polymorphic in the populations tested. This can therefore be thought of as a measure of the polymorphism of the target sequence. Polymorphism rate of the target sequence is itself influenced by the type of SSR targeted and level of polymorphism between the lines tested. Map rate, measured as 'Total mapped/total Polymorphic', asks the

question, ‘of the markers found to be polymorphic, what proportion successfully made it through to the final linkage map?’ Factors that can affect this rate are the ease of scoring during genotyping, and the robustness of amplification throughout the F_2 population.

Map rate for the EST SNPs is considerably higher than for the SSR markers, probably reflecting the ease and unambiguity of scoring, particularly for KASPAR markers. In contrast, SSRs can have complex patterning, particularly for complex or compound repeat sequences and also contain PCR artefacts that can be difficult to accurately distinguish (Schlotterer, 2004). Multiplexed primers can sometimes interact in unforeseen ways meaning a portion of markers fail the F2 genotyping or mapping process despite being polymorphic when tested individually.

The fail rate of SSR markers is low, averaging 12.4 % across the combined dataset. Polymorphism rate is substantially lower than the EST SNPs, as expected, since SSRs are not identified as polymorphic prior to PCR amplification. Polymorphism rate, instead, reflects a combination of the level of polymorphism or genetic relatedness of the parental lines tested, and the type of SSR targeted. Different SSR sequence lengths can differ in polymorphism rate due to the ease in which DNA polymerase slippage and DNA mismatching can occur during DNA replication, a point that will be explored later. Since this data has not been grouped by the lines tested or type of SSR targeted, these values should represent an averaged value across all markers and populations according to those tested by the listed authors.

The observed variation in polymorphism rate between the individual authors for SSR markers reflects the difference in function that these markers were developed for, in terms of the proportion of the total for each author. Mapping of gaps in the linkage map represents a drop in polymorphism rate for several reasons. By definition these regions tend to have lower polymorphism rates as otherwise the randomly distributed markers would have been expected to fall within these regions by chance. Secondly, since mapping of gaps is specific to individual maps from individual populations, these markers are only tested in single populations, effectively decreasing the chance of polymorphism by 4. Mapping of candidate genes can also have a similar effect, since particular candidate genes are developed for specific populations. It is an important aim of this type of analysis to deconstruct and differentiate between these various influences on marker performance, so that the individual effects of each factor can be clearly seen, as well as the underlying/intrinsic performance rate of this particular marker type when such influences are removed. Such data is then informative for future use if such markers are used again.

Influences of the sequence source, function/application for which the markers were developed and the type of SSR targeted on SSR marker performance will be explored in greater detail in this chapter.

This analysis highlights the difference in performance statistics between EST SNPs and SSRs. General rules can be identified, such as the difference in fail rate between markers developed from transcribed DNA and genomic DNA (due to the presence of introns and regulatory elements), the difference in polymorphism rate between markers developed from comparative sequencing data compared to genome sequence (and whether polymorphism is identified *in silico* prior to PCR amplification), and the difference in mapping rate between EST SNPs and SSRs (due to the differing systems used to amplify and score these markers). SSR performance rates indicate the average rates over the course of the project, without differentiating marker source, function or type of SSR targeted.

Table 4-5 SSR Marker Source analysis

(a) SSR Marker Source									
SSR Source	**SSR Enriched Library**	**Publically available mRNA/cDNA**	**Publically available nucleotides**	**Previous study (1)**	**Previous study (2)**	**Previous study (3)**	**Genome Sequence (4)**	**BAC Sequencing**	**Total**
Number of Marker Primers Designed	**74**	**19**	**3**	**9**	**17**	**23**	**530**	**8**	**683**
Author(s)	(A)	(A)	(A)	(A)	(A)	(A)	(A),(B)	(A)	
(b) Primer outcome									
Polymorphic	31	12	2	2	11	14	240	5	317
Not Polymorphic	28	4	1	1	5	8	215	1	263
Failed	15	3	0	6	1	1	54	2	82
Not tested	0	0	0	0	0	0	21	0	21
Total	74	19	3	9	17	23	530	8	683
Mapped	**28**	**11**	**2**	**2**	**10**	**13**	**143**	**4**	**213**
(c) Success rate (%)									
Fail rate	**20.3**	**15.8**	**0**	**66.7**	**5.88**	**4.35**	**10.6**	**25.0**	**12.4**
Polymorphism rate	**52.5**	**75.0**	**66.7**	**66.7**	**68.8**	**63.6**	**52.7**	**83.3**	**54.7**
Map rate	**90.3**	**91.7**	**100**	**100**	**90.9**	**92.9**	**59.6***	**80.0**	**67.2**

*Map rate excluding markers developed for gap filling = 77.6 %

Authors

SSR mining, primer design and polymorphism testing carried out by:

(A) Dr. Andrew J. King

(B) Jasper G. Clarke

Previous studies

(1) (Sun et al., 2008)

(2) (Phumichai et al., 2011)

(3) (Wang et al., 2011)

(4) (Sato et al., 2011)

Table 4-5 analyses the SSR markers according to the sequence source from which they were mined. Part (a) lists the different marker sources in the order that they were carried out during the project (from left to right), the number of primers designed and the author(s) attributed with this work. Part (b) lists the outcome of primer testing, and part (c) the performance statistics including fail rate, polymorphism rate and map rate, calculated in the same way as described in the table 4-6 analysis.

SSR markers were mined from a variety of sources. This included transcribed DNA in the form of EST enriched libraries (column 2) or publically deposited mRNA/cDNA submissions (column 3), GenBank nucleotide submissions (column 4), markers used to characterise *J. curcas* genetic variation in previously published studies (column 5 & 6), markers mined from the Jatropha genome sequence (column 8), markers to anchor previously published linkage maps (column 7), and markers mined from Bacterial Artificial

Chromosome (BAC) sequencing for fine mapping applications (column 9). The majority of markers were mined from EST enriched libraries and the Jatropha genome sequence (together accounting for 88.4 % of total markers designed).

Comparison of markers developed from transcribed DNA (columns 2 & 3) to markers developed from genomic DNA sequence (columns 6-8) shows the characteristically higher fail rate (as observed with EST SNPs table 2) due to the absence of introns and other regulatory sequences in transcribed DNA, leading to a greater PCR primer fail rate when amplifying from genomic DNA. Polymorphism rate should not have a bias between transcribed and genomic DNA marker sources, since neither of these marker groups have been identified as polymorphic prior to polymorphism testing.

Here we see polymorphism rates that vary between a lower end of just over 50 % (for EST enriched library and genome sequence), to the majority of sources being between mid to high 60's (66.7 %, 66.7 %, 68.8 %, 63.6 %) to a high score of 75 % (for publically deposited $_{m}$RNA/cDNA). Polymorphism rate does seem to be lower for EST enriched library SSR markers and genome sequence SSRs. In part this can be explained by the function for which the markers were developed. For genome sequence derived SSRs, a significant proportion were developed to fill in gaps during later rounds of linkage mapping, map candidate genes or fine map QTL. All of these applications are specific to individual populations, reducing the number of populations each marker was tested in and therefore the chance of polymorphism. In addition, markers designed for gap filling are specifically targeting low polymorphism regions, which otherwise would be expected to be covered by the randomly distributed markers by chance. The fact that subsequent SSR markers targeted to these regions through comparative mapping techniques also show a low polymorphism rate confirm that these gaps tend to be low polymorphism regions. Polymorphism rate is expected to be lower in the EST enriched library source since greater evolutionary pressures operate on coding DNA, shortening average SSR repeat sequence length in comparison to genomic SSRs and limiting the frequency of repeat sequence expansions/contractions that would be expected to have an impact on protein function. Given that we know these influences are reducing polymorphism rates in genome sequence sourced SSRs, it seems likely that the rate of between 65-70 % polymorphism as observed for the majority of the other sources reflects the most accurate indicator of underlying polymorphism across these populations for SSR markers.

Map rate (defined as 'Total markers mapped/total markers polymorphic'; see table 1.3) should also be constant since all markers were scored in the same manner, and each category is composed of a mixture of different SSR types. Here we see the majority of markers are within the 90 % or greater range, with the exception of Genome Sequence derived SSRs. The drop in map rate observed for genome sequence SSRs can be explained by the fact that this group contains markers developed for the mapping of gaps in later rounds of linkage mapping, and mapping of candidate gene markers specific to particular populations. The approach used to map gaps was to target multiple SSRs per locus with only 1 polymorphic marker needing to be mapped during linkage mapping. To illustrate the effect that the application for which the SSR primer was developed has on map rate, when markers developing for gap filling are removed from this group, map rate increases to 77.6 %. Candidate gene mapping can again be specific to individual populations such that if the marker is not polymorphic in that specific population there is no point in mapping in other populations that the marker may be polymorphic, particularly if the region is already mapped. It seems likely that a map rate of ~90 % represents the most accurate reflection of the rate of mapping SSRs mined from genome sequence in this dataset when all other specific influences are removed.

Taken together, these points suggest that the most representative performance statistics for SSR markers mined from genome sequence, are a fail rate of between 5-10 % (excluding the influence of markers designed from transcribed DNA), a polymorphism rate of ~60-65 % (excluding the influence of markers targeted to low polymorphism regions and tested in single populations), and a map rate of around 90 % (excluding markers where several SSRs were targeted per locus or markers were developed for specific populations).

Table 4-6 SSR Marker Function/Application and performance

(a) SSR Marker Function (ordered by number mapped)						
Function	**Candidate Genes**	**Gap filling (comparative mapping)**	**Non-specific gene mapping (transcribed sequences)**	**Non-specific mapping**	**Map anchoring**	**Total**
Designed	**197**	**284**	**96**	**83**	**23**	**683**
Author	(A)(B)	(A)(B)	(A)	(A)	(A)	
(b) Primer outcome						
Polymorphic	100	116	45	42	14	317
Not Polymorphic	71	128	33	23	8	263
Failed	15	30	18	18	1	82
Not tested	11	10	0	0	0	21
Total	197	284	96	83	23	683
Mapped	**67**	**57**	**41**	**35**	**13**	**213**
(c) Success rate (%)						
Fail rate	**8.1**	**10.9**	**18.8**	**21.7**	**4.3**	**12.4**
Polymorphism rate	**58.5**	**47.5**	**57.7**	**64.6**	**63.6**	**54.7**
Map rate	**67.0**	**49.1**	**91.1**	**83.3**	**92.9**	**67.2**

Authors

SSR mining, primer design and polymorphism testing carried out by:

(A) Dr. Andrew J. King

(B) Jasper G. Clarke

Table 4-8 shows the SSR markers grouped according to function/application in the project. Part (a) lists application in order of number of markers mapped (from left to right), the number of primers designed and the authors responsible for the work. As can be seen, a total of 683 markers were developed, of which 197 were for candidate gene mapping, 284 for second generation linkage mapping, 96 were for non-specific gene mapping based on transcribed sequence data, 83 were for non-specific mapping and 23 for anchoring linkage groups to a previously published linkage map. Part (b) lists the outcome of primer testing. In summary, the following number of SSRs were mapped for each application: 67 for candidate gene mapping, 57 for second generation linkage mapping, 41 for non-specific gene mapping using transcribed sequence, 35 for unspecified mapping, and 13 for anchoring linkage groups to previously published maps. In total this gives 213 SSR markers mapped. Part (c) lists the performance statistics for each category of SSR, in fail rate, polymorphism rate and map rate.

As can be seen by the statistics listed in parts (a) and (b), Candidate gene mapping was the predominant applications of SSR markers used in this project. Total primers for these categories combined account for 70.4 % of all markers designed. Looking at the fail rates across the groups, we see that the typical ~5-10 %

fail rate (4.3 % map anchoring to 10.9 % for gap filling) is observed with the exception of the markers mined from transcribed sequences (18.8 %), which, as already discussed, have a characteristically higher fail rate due to the presence of introns and other regulatory elements not present in the genomic DNA that PCR primers are tested on. This fail rate of 18.8 % is very similar to the EST-SNP marker fail rate (22 %, table 2), which were also designed from transcribed DNA.

Map rate is at the typical ~90 % rate for all groups except candidate genes and gap filling, ranging from 83.3 % for non-specific mapping to 92.9 % for Map anchoring. As previously discussed, a drop in map rate in gap filling is expected due to multiple SSRs being targeted per locus, and for candidate genes, such gene markers only being required in the populations harbouring the specific traits of interest.

To summarise this analysis, SSR markers have been predominantly used for the mapping of candidate genes, later stage linkage mapping to improve areas of low marker density. A significant proportion of markers were also developed from an EST enriched library to map expressed genes in a less specific manner. As observed with previous analyses, markers designed from transcribed DNA show a characteristically higher fail rate compared to markers mined directly from genomic DNA. Markers designed for gap filling has a lower polymorphism and map rate, due to markers targeting lower polymorphism regions and being specific to individual mapping populations.

Table 4-7 SSR SSR repeat sequence size and performance

(a) SSR Repeat Sequence										
Repeats seq size $(x)_n$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Complex**	**Unknown**	**Total**
Number of marker primers designed	**2**	**332**	**119**	**14**	**3**	**10**	**2**	**66**	**135**	**683**
(b) Author										
AJ King	0	183	43	4	0	5	0	25	135	395
JG Clarke	2	149	76	10	3	5	2	41	0	288
Total	**2**	**332**	**119**	**14**	**3**	**10**	**2**	**66**	**135**	**683**
(c) Primer outcome										
Polymorphic	1	169	41	0	0	5	0	27	74	317
Not Polymorphic	1	99	64	12	3	4	2	33	45	263
Failed	0	51	10	1	0	1	0	4	15	82
Not Tested	0	13	4	1	0	0	0	2	1	21
Total	2	332	119	14	3	10	2	66	135	683
Total Mapped	**1**	**123**	**32**	**0**	**0**	**5**	**0**	**18**	**34**	**213**
(d) Success rate (%)										
Fail rate	0	16.0	8.70	7.69	0	10	0	6.25	11.2	12.4
Polymorphism rate	**50**	**63.1**	**39.0**	**0**	**0**	**55.6**	**0**	**45**	**62.2**	**54.7**
Map rate	**100**	**72.8**	**78.0**	**0**	**0**	**100**	**0**	**66.7**	**45.9**	**67.2**

Table 4-7 shows the performance of individual types of SSR SSR. Part (a) lists the number of SSRs mined and primers designed for each SSR sequence type. Part (b) lists the authors responsible for the work, and the number of primers designed for each SSR type. Part (c) lists the outcome of primer testing and part (d) lists the overall success rate of each SSR type in fail rate, polymorphism rate and map rate.

In this project, repeat sequences of 1 to 7 nucleotides were targeted for simple repeat SSRs, as well as complex SSRs consisting of either or both elements of (a) multiple repeat sequence sizes (compound SSRs) or (b)interrupted SSRs (repeat sequences interspersed with non-repeat sequence). Complex SSRs were mined using 'Imperfect SSR' software as described in Materials and Methods. The most commonly used SSR was the dinucleotide $(xx)_n$ and trinucleotide $(xxx)_n$ repeat sequences with 332 and 119 sequences mined respectively. There was also a significant quantity of complex SSRs (66 sequences). Unknown SSRs reflect a number of unannotated SSRs included in order to maintain continuity in overall/total marker statistics. As can be seen, all authors predominantly targeted the shorter repeat sizes of two and three nucleotides, reflecting the fact that these were often the most common SSR type found and often the largest in total number of repeats units.

It can be expected that the shorter the repeat motif and the greater the number of repeat units, the more likely the chance of polymorphism (Lai and Sun, 2003, Schlotterer and Tautz, 1992). SSR expansion is thought to be caused by a combination of DNA polymerase slippage and DNA mismatching during DNA replication, leading to the formation of loop structures that enable these repeat stretches to expand or contract over time (Li et al., 2002). Both DNA polymerase slippage and DNA mismatching are more likely to occur with shorter repeat motif sequences due to the smaller physical distances between each repeat unit making replication errors, such as slippage and mismatching, more likely to occur. Longer repeat stretches facilitate the formation of loop structures during replication and can also be expected to facilitate changes in repeat length.

Part (c) lists the outcome of primer testing and part (d) an analysis of these values to indicate overall marker performance. Fail rate was expected to have been constant across all groups since the different marker sequence sources (such as transcribed or genomic DNA) were distributed across all groups. Fail rate ranges from 6.25 % for complex SSRs to 16 % for dinucleotide repeats, roughly keeping to the 5-10 % fail rate observed previously, with a variation from the mean rate of 12.4 % of -6.15 % and +3.6 % for these two most extreme groups respectively.

Polymorphism rate would have been expected to vary depending on the type of SSR targeted. The three largest groups, 2, 3 and complex SSRs, are most informative in this regard due to the higher number of markers tested. Here we see that dinucleotide repeats have a significantly higher polymorphism rate compared to trinucleotide and complex SSRs, as expected. Interestingly, complex SSRs have a higher polymorphism rate than trinucleotide repeats, suggesting that elements such as short non-SSR sequences in the middle of SSRs, or mixtures of different repeat sequence sizes, are not in themselves a major factor limiting repeat expansion over time.

Map rate, as previously discussed can be thought of as an indicator of the ease of scoring and robustness of amplification across the F2 population, after other factors such as the application for which they were developed are taken into account. Since this data is not grouped according to application, the effects of markers developed for gap filling which have a lower map rate, for example, should be distributed across all of the groups being compared here, enabling the relative map rates of each group to be fairly compared.

Amplified SSRs create a characteristic amplification pattern depending on size and type of repeat sequence, and the alleles present. Allele sizes that are very similar can interact to form additional patterns due to overlapping peaks. It could therefore be expected that different repeat sizes and types could have different emergent map rates, as the number of primers tested increases to greater numbers. We know that the

function/application for which the SSR were developed affects map rate, however this effect should be reduced since the groupings in this analysis do not have any bias towards function/application, and so the effect of these markers should be split between the different SSR repeat categories, such that the relative rates of each category can be compared.

We do see a difference in map rate between the three largest categories, 2, 3 and complex SSRs which have a 72.8 %, 78 % and 66.7 % map rate observed respectively. This data would suggest that trinucleotide repeats are slightly easier to score than dinucleotide repeats, which can be true when the polymorphism level is a single repeat unit difference. Here a 3 bp difference in PCR product compared to a 2 bp difference could be expected to be easier to differentiate when scoring in terms of less overlapping sequences trace. However, it is the authors' experience that these two categories do not present a substantial difference in ease of scoring, and so the differences could be due to the number of markers in each category that were developed for different applications, although a slight drop can be observed for complex SSRs which may be due to the more complex trace patterning that these class of marker produce. A substantial difference in map rate is not observed between different SSR repeat sequence types, suggesting that there is not a great difference in the ease of scoring and the robustness of amplification between the different categories.

A remaining factor that is not covered in this analysis is the number of repeat units and its effect on polymorphism rate. As previously discussed, the longer the repeat sequence stretch the higher the polymorphism rate is expected to be, due to the greater number of opportunities available for DNA slippage and mismatching to occur and the ease at which loop structures would be expected to be formed. However, data for this property was not available for analysis.

Nevertheless, this analysis supports the hypothesis that shorter repeat motifs have a higher polymorphism rate, and therefore should be preferentially targeted for the development of DNA markers. This analysis also highlights the fact that complex SSRs, particularly those consisting of short repeat sequence sizes, should also be considered for SSR marker development, highlighting the importance of using software that is capable of recognizing these complex sequence types in genomic or transcribed DNA sequence.

4.4: Discussion

Breeding programmes that intend to utilise Marker Assisted Selection (MAS), require a number of genetic resources. Genetically and phenotypically distinct lines are required to form mapping populations. Phenotypic variation should be present in the trait of interest, and genetic variation should be present to ensure that such traits are heritable, and that they can be tracked throughout a segregating population. A comprehensive set of DNA markers are required to track individual loci, assess their performance on phenotype and ultimately inform selection of lines containing beneficial QTL. DNA markers should cover the complete genome, and ideally be placed as close to, and flanking, any QTL present. The chance of marking functionally relevant loci and causative mutations is increased by mapping expressed genes and candidate genes. DNA markers need to be grouped, ordered and their relative distances calculated so that the resulting linkage map can be used to most accurately detect and locate QTL.

In this study, genotyping and genetic linkage mapping of the G51xCV mapping population was used to generate a genetic linkage map and to contribute data towards the combined genetic linkage map along with data from three other populations.

DNA marker production across these populations was facilitated by the genetic similarity of one parental line in each population. This reference line, to which the other lines containing beneficial traits was crossed, was genetically very similar to the majority of material used across the world as determined by genetic characterisation work (Yue et al., 2014, Montes Osorio et al., 2014, Pecina-Quintero et al., 2014, He et al., 2011). By using a single reference line in each population, the chances of DNA markers being polymorphic across multiple populations was increased. This increased the number of markers available per population, and increased the value per marker versus the resources required to produce it. The genetic similarity of the reference lines to widely used material will facilitate introgression of beneficial QTL identified by this study, as the DNA markers disseminated should be directly transferable, and QTL performance will have been established in a genetically-similar background.

Markers with polymorphism in multiple populations not only improve marker density and facilitate QTL mapping in individual populations, but such shared markers have numerous benefits for comparative mapping and linkage mapping from combined datasets.

Alignment of the individual mapping population linkage maps was possible due to these shared markers. Since the physical location of these markers in the *J. curcas* genome remains constant in each population, marker ordering and recombination rate should be similar, which should lead to identical marker ordering and similar marker spacing in all individual population linkage maps. Alignment and comparison of shared marker positions between maps, enables relatively accuracy and consensus to be determined across each population, and serves as an excellent quality control for the overall mapping process. Marker ordering, spacing and total distance mapped remains highly conserved between all independent population maps, indicating the robustness and accuracy of each dataset.

Following alignment, such shared markers also become useful for comparative mapping. After linkage maps are aligned, markers in one map that correspond to gaps in another, can be used to target additional markers to the regions required. The markers corresponding to the gap are used as probes to retrieve *J. curcas* genome sequence contigs; these contigs can then be scanned for additional SSR markers.

Lastly, shared markers substantially improve the accuracy of the combined linkage map by increasing the recombination data available for these markers, which account for a significant portion of the total markers mapped. This substantially improves the accuracy of the calculated genetic distances and marker ordering, since sample size is effectively increased from individual population numbers to multiple populations, for calculating recombination rates of these markers. Marker ordering and spacing for shared markers in the combined map, represents data collected from up to 989 F2 plants. The combined map produced from this data represented a considerable improvement on the only other available Jatropha linkage map at the time of publication; an inter-specific linkage map produced from 93 plants (Wang et al, 2011).

The 51xCV mapping population utilised a variety of DNA marker types and sequence sources. A genome wide, randomly distributed selection of genome sequencing derived SNPs were developed to cover as much of the genome as possible. The CRoPs® technique utilised here, harnesses next generation sequencing to comparatively sequence amplified fragments produced from a modified DNA fingerprinting technique, AFLP (van Orsouw et al., 2007). As with genotyping by sequencing approaches, SNPs can be selected from a panel of thousands that offer most use to the user and are genotyped using high throughput methods (work conducted by Keygene) (Davey et al., 2011).

Whilst this approach amplified random segments of the genome without differentiating between coding and non-coding regions leading to a genome-wide non-selective marker set, markers in expressed gene were also available for genotyping and mapping in the G51xCV mapping population (King et al., 2011). Such expressed sequence tagged (EST) markers were randomly distributed through expressed genes, adding selectivity towards functional DNA, although still randomly distributed throughout the Jatropha transcriptome (King et al., 2011, Varshney et al., 2014).

Simple Sequence Repeat (SSR) markers were utilised for a range of specific tasks. Mapping of markers from previous studies enabled comparison of parental material to previously studied germplasm, and linkage maps to be anchored to a previously published interspecific linkage map (Wang et al 2011). The main applications for SSRs, however, were for gap filling of individual linkage maps during later round linkage mapping, and for the mapping of candidate genes.

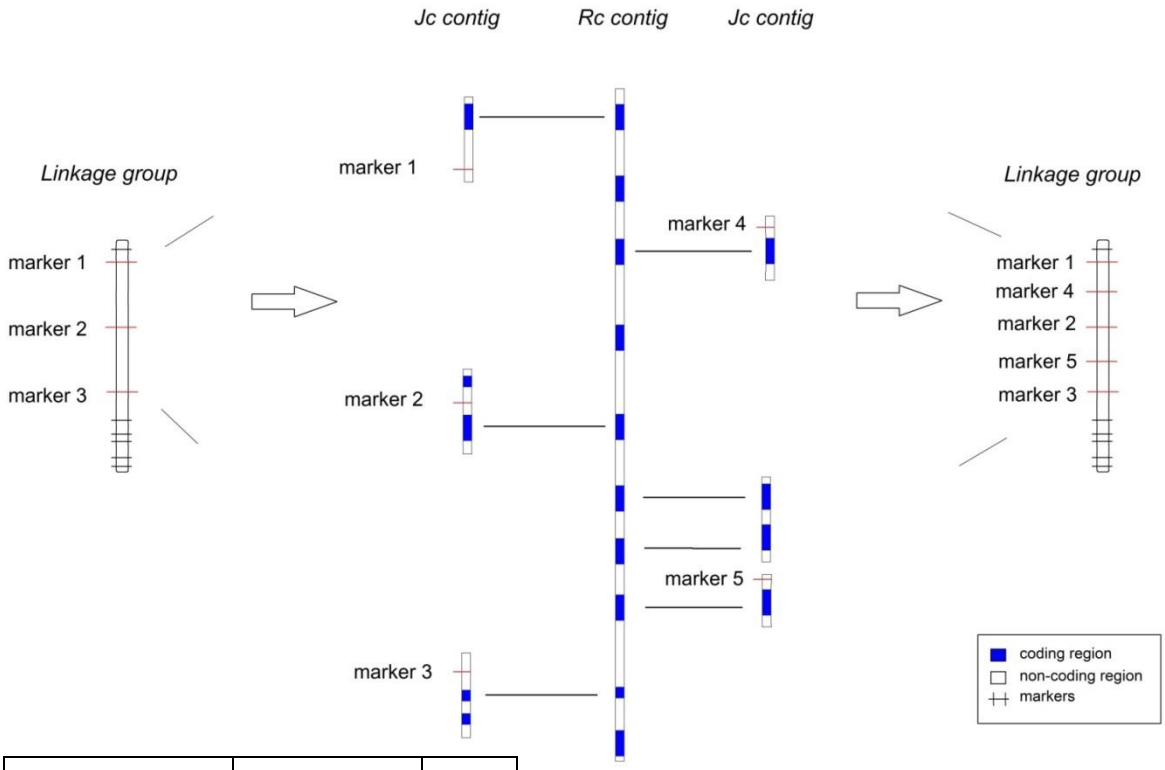
SSR mining benefitted from the publication of the Jatropha genome sequence in 2011 (Sato et al, 2011). SSR SSRs can be efficiently mined from genome sequence using software (Martins et al., 2009, Stieneke, 2007). In contrast to SNPs that require comparative sequencing to identify markers, SSRs are identifiable from a single sequence such as a reference genome sequence. Validation of identified SSRs is conducted by PCR amplification in parental lines, which is perhaps the biggest drawback of SSRs since the design of PCR primers, amplification and scoring in parental lines (polymorphism testing) is more user-intensive than higher throughput approaches that utilise genome sequencing.

For this reason rational selection of SSRs to increase the chance of polymorphism is important to maximise efficiency of marker production. As has been explored in this chapter, shorter repeat motifs have a higher polymorphism rate, due to the ease at which DNA polymerase slippage and DNA mismatching can occur during DNA replication (Li et al., 2002). The length or number of repeat units also increases the probability of polymorphism since there are more opportunities for polymerase slippage and mismatching to occur, and more sequence in which loop structures can form (Oliveira et al., 2006, Lai and Sun, 2003, Li et al., 2002). As highlighted by the analyses in this chapter, the presence of compound or interrupted SSRs (together 'complex SSRs'), does not in itself represent a major influence on polymorphism rate, such that complex SSRs, particularly those consisting of shorter repeat motif sizes, should be readily included in ones SSR mining strategy. Software such as Imperfect SSR is recommended over standard SSR mining software in this regard.

This final point is particularly important for draft genome sequences, which are differentiated from complete genome sequences by a lack of physical mapping. Draft genome sequences consist of a series of unordered contigs of varying sizes. For the majority of applications the draft genome state is more than adequate, such that there is very little incentive to produce physical maps which add another (significant) level of technical difficulty, time and cost. In an age where next generation sequencing is becoming cheaper and more accessible, and where a greater range of crops are becoming sequenced, draft genome sequences will continue to represent the majority of genomes for all but the most widely studied species (Feuillet et al., 2011). Mining of all effective marker types from potentially small contigs is important to increase the utility of such resources. Comparative mapping techniques, utilising comparative genomics, were also extensively used in this project for the same reason.

SSR markers were utilised for gap filling and the mapping of candidate genes in the G51xCV mapping population. Comparative mapping techniques, such as alignment of individual population maps, was used as a method of targeting SSRs to gaps in the G51xCV linkage map. Comparative mapping and synteny with castor bean (Chan et al., 2010) was also utilised for SSR marker development as an extension of this process.

For example after G51xCV linkage groups were aligned to other population maps, markers corresponding to gaps were used to pull out *J. curcas* contigs as before. From there, amino acid sequences from all predicted gene models in the contig, were used as probes to search the castor bean genome sequence. Castor bean transcribed amino acid sequences that mapped to the same contig and in the same order, suggested a region of synteny and gene-colinearity. Transcribed amino acid sequences further upstream or downstream in the castorbean genome sequence could then be used as a probe to retrieve *J. curcas* genome sequence in the reverse direction. For syntenous regions, the longer contig sizes of the castor bean genome sequence could be used to identify and order the more fragmented *J. curcas* genome sequence, in order to reach contigs corresponding to target regions of the linkage map. Amino acid sequence conservation for transcribed gene models was high between the two Euphorbiacea species, *Jatropha curcas* and castor bean, facilitating this approach. Figure 4-6 below, gives a visualisation of this process.



Genome Stats	Jatropha curcas		R. comm.
Release	r3.0	r4.5	1
Total length of contigs	285 Mbp	297 Mbp	325 Mbp
Total number of contigs	150,417	39,277	3,500
Max length of contigs	29,744	277,264	4.7 Mb
Av. length of contigs	1,900	7,579	13,000
N50	3,833	15,950	561,400
GC content	34.3	34	32.5

Figure 4-4 Comparative mapping between *J .curcas* and *R. communis* during later round linkage mapping.
Interspecific synteny and gene-colinearity were utilised between *J. curcas* and its nearest sequenced relative *R. communis* (Hirakawa et al., 2012, Sato et al., 2011, Chan et al., 2010). The more highly conserved transcribed amino acid sequences of gene models, from *J. curcas* and *R. communis* genome sequence, were used to establish syntenous regions and navigate back and forth between genomes in order to identify SSR markers in target regions. Note the difference in contig length between *J. curcas* and *R. communis*, which made this technique possible. Average contig length displayed in the statistics table is misleading, due to the high number of small fragments in each genome sequence. More informative is the median contig length (N50) which more accurately reflects the difference in contig sizes between the two genome sequences. *J. curcas* had two genome sequence releases r3.0 and r4.5 (Hirakawa et al., 2012, Sato et al., 2011).

As explored in this chapter, mining of SSR markers for different applications produced differing polymorphism, and map rates. Analysis of the various components influencing SSR marker success rates, such as target sequence source, application, and repeat sequence motif size, showed that with all distorting influences removed, SSR performance characteristics were approximately a 5-10 % fail rate, 60-70 % polymorphism rate, and a 90 % map rate, highlighting the efficiency and utility of using this readily available marker type for the applications suggested. The mapping of candidate genes, as discussed in Chapter 3 was a particularly useful addition of developed SSR markers to the genome wide non-selective markers available in the G51xCV mapping population.

Overall the marker mining strategies employed in this project were designed to produce a comprehensive set of DNA markers. A panel of randomly (unbiasedly) distributed SNPs were generated using a reduced representation genotyping by sequencing approach (Davey et al., 2011, van Orsouw et al., 2007). Genotyping by sequencing based approaches remain the most thorough and powerful marker mining technology available, since both SNP discovery and genotyping can be carried out in a single run and the density of markers produced is theoretically maximal depending on whether a genome reduction strategy is employed and the depth of sequencing (Davey et al., 2011). The power of this approach should only increase as sequencing technology improves and costs reduce. Comparative EST library sequencing was used to increase the number of mapped expressed genes and increase the chance of mapping functionally relevant loci. Since expressed genes make up the majority of functional DNA, mapping of expressed genes would seem like a critical component of a comprehensive marker set (depending on the level and coverage of genotyping by sequencing if it is employed). SSR markers were mined from the Jatropha genome sequence to improve the linkage maps during later round linkage mapping, as well as to map candidate genes and other important trait-related metabolic genes. These SSR markers represented an efficient, cost effective and readily available source of marker to complement and further improve on the randomly distributed, non-selective (either across the entire genome or the transcriptome) markers.

Genetic linkage mapping in the G51xCV mapping population, produced a robust and reliable genotypic dataset and genomic resource that was validated through alignment to 3 other independent dataset linkage maps using shared markers. The G51xCV genetic linkage map underpins both accurate QTL detection and QTL location in this population, and has also contributed towards the first published intraspecific genetic linkage map for *Jatropha curcas*. Mapping of SSRs linked to seed oil biosynthetic genes in particular, is a useful genomic resource for further study into the genetic basis of oil yield and oil quality variation in *Jatropha curcas* populations.

4.5: Appendix

Contents:

(1) The 51xCV linkage map

(2) Physical Alignment and Comparison of individual mapping population linkage maps

(3) The Combined linkage map

(4) Candidate and metabolic gene linked SSR markers

4.5.1: The G51xCV Genetic Linkage Map

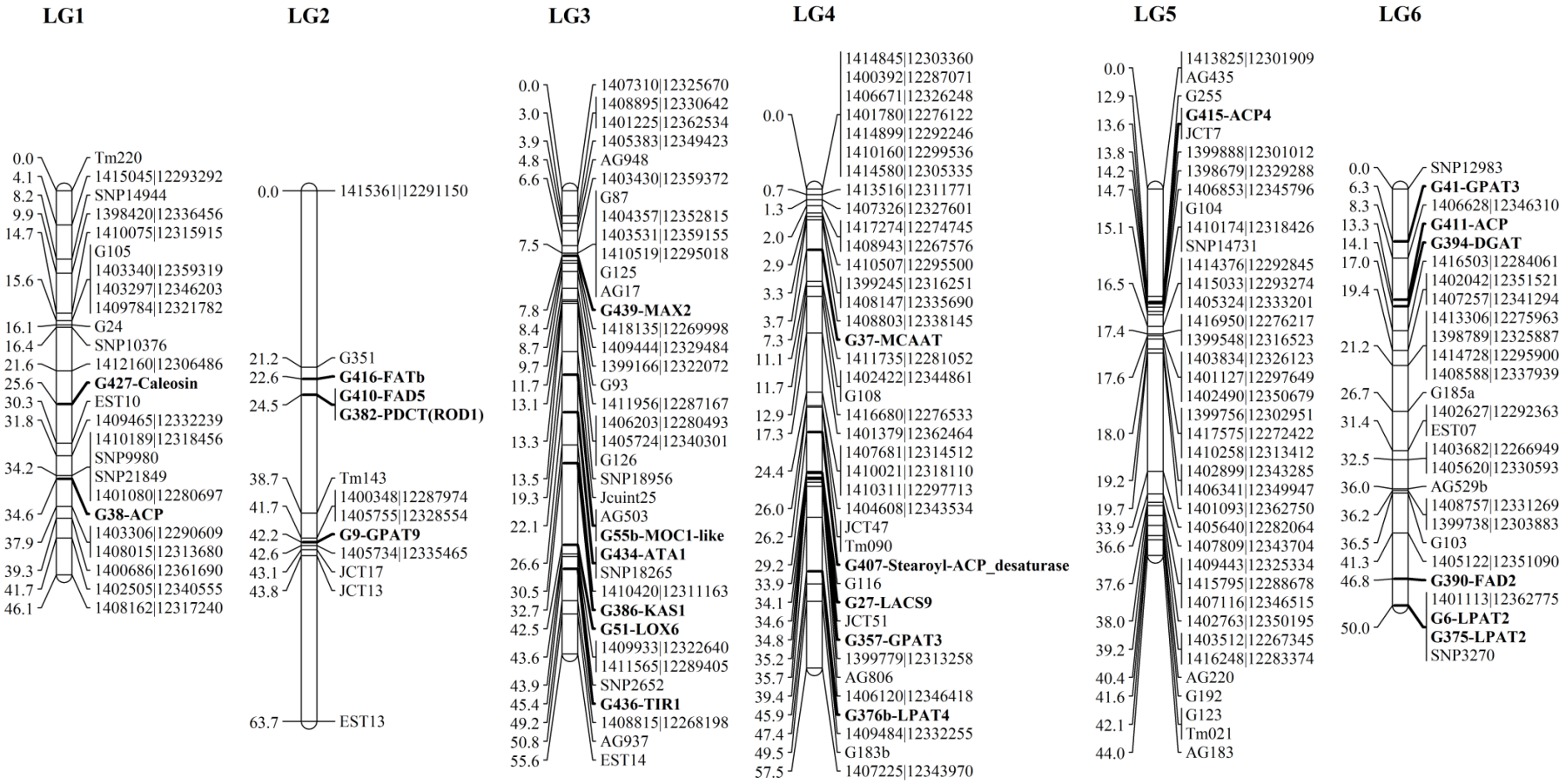


Figure 4-5 The G51xCV Genetic Linkage Map, Linkage groups 1-6.

The 51xCV linkage map is composed of 312 co-dominant markers, distributed over 11 linkage groups at a density of 2.969 cM per unique loci. Markers linked to candidate genes have been highlighted in bold.

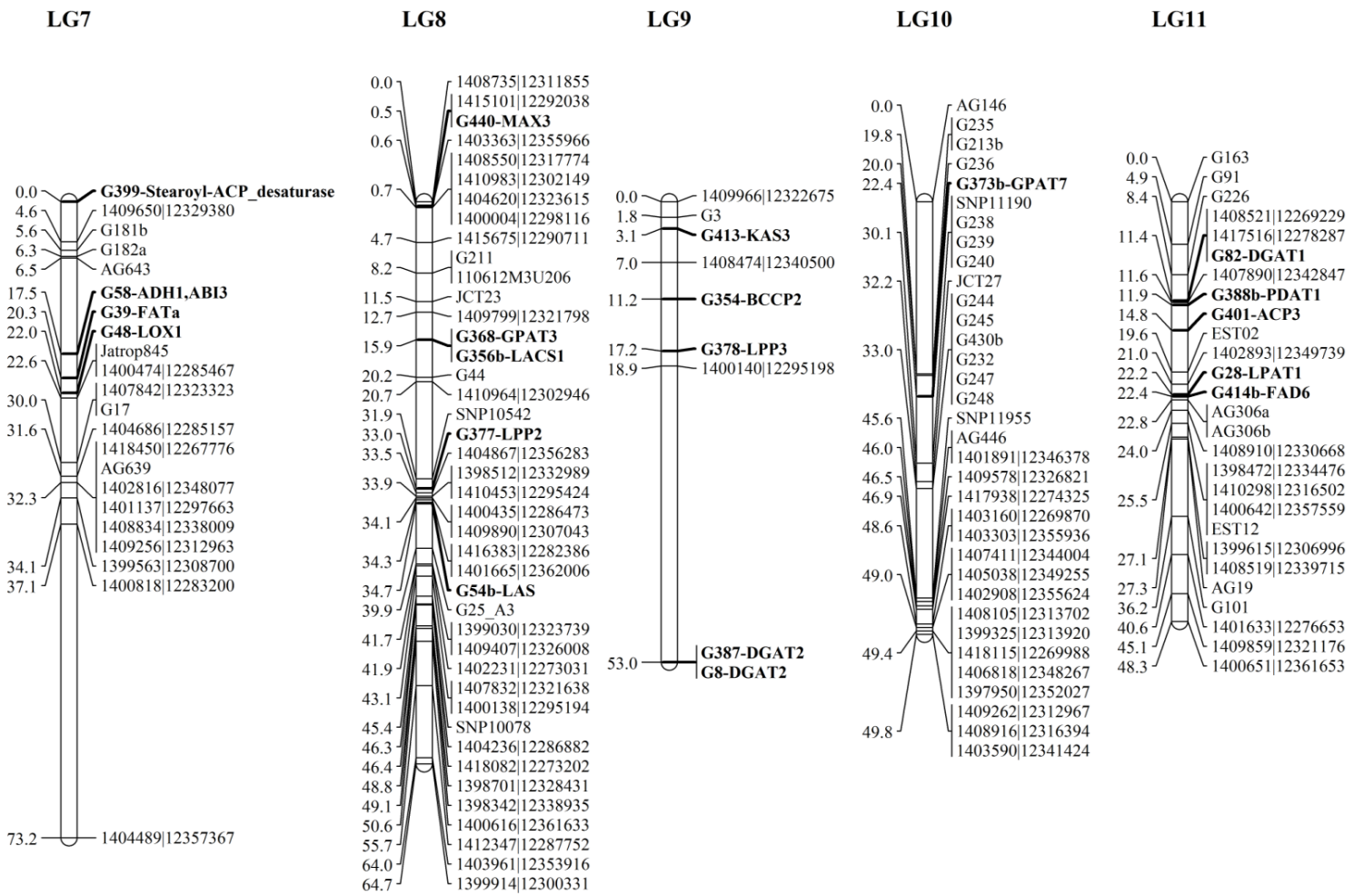


Figure 4-6 The G51xCV Genetic Linkage Map, Linkage groups 7-11.

The 51xCV linkage map is composed of 312 co-dominant markers, distributed over 11 linkage groups at a density of 2.969 cM per unique loci. Markers linked to candidate genes have been highlighted in bold.

4.5.2: Physical Alignment and Comparison of individual mapping population linkage maps

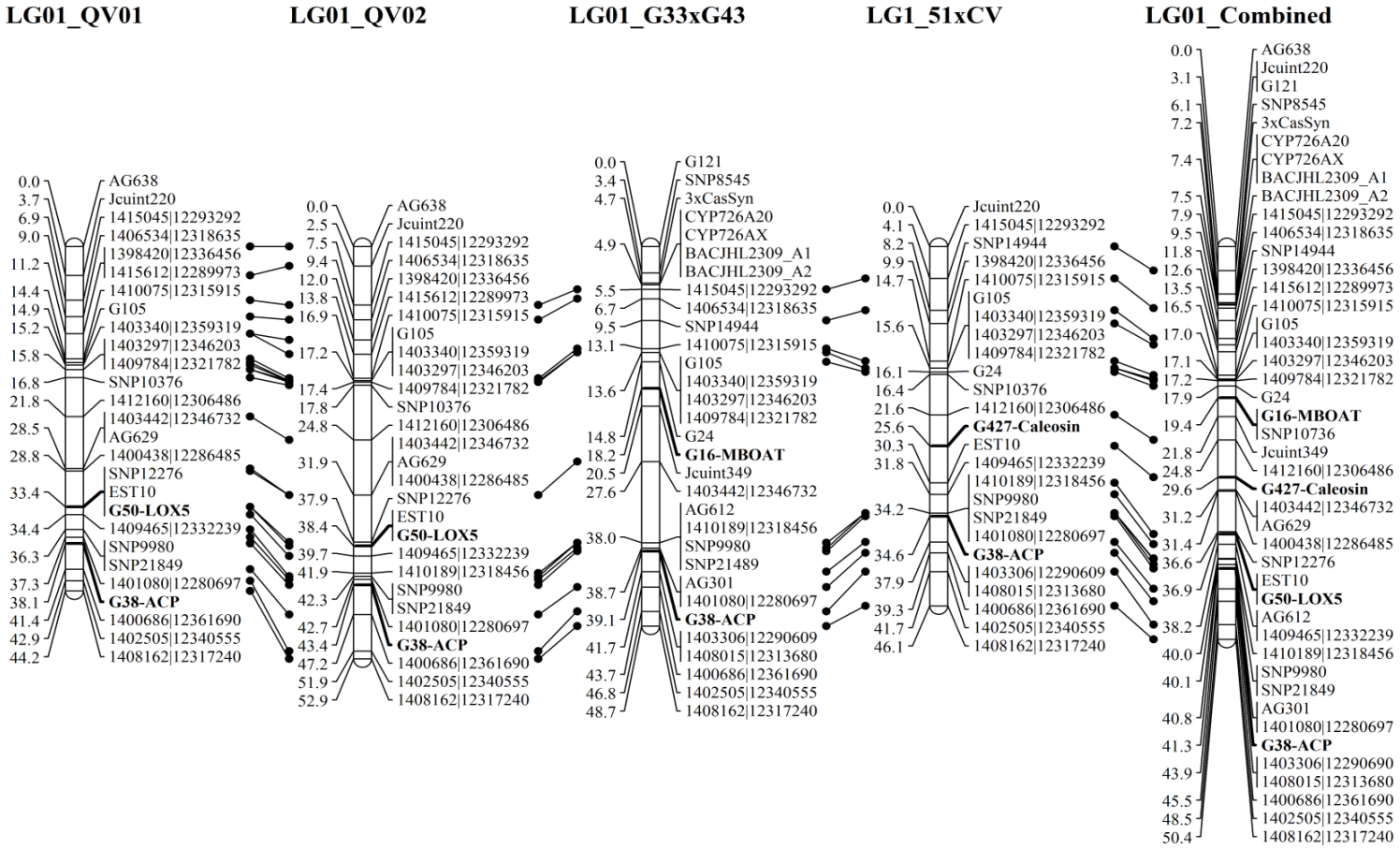


Figure 4-7 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.
Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

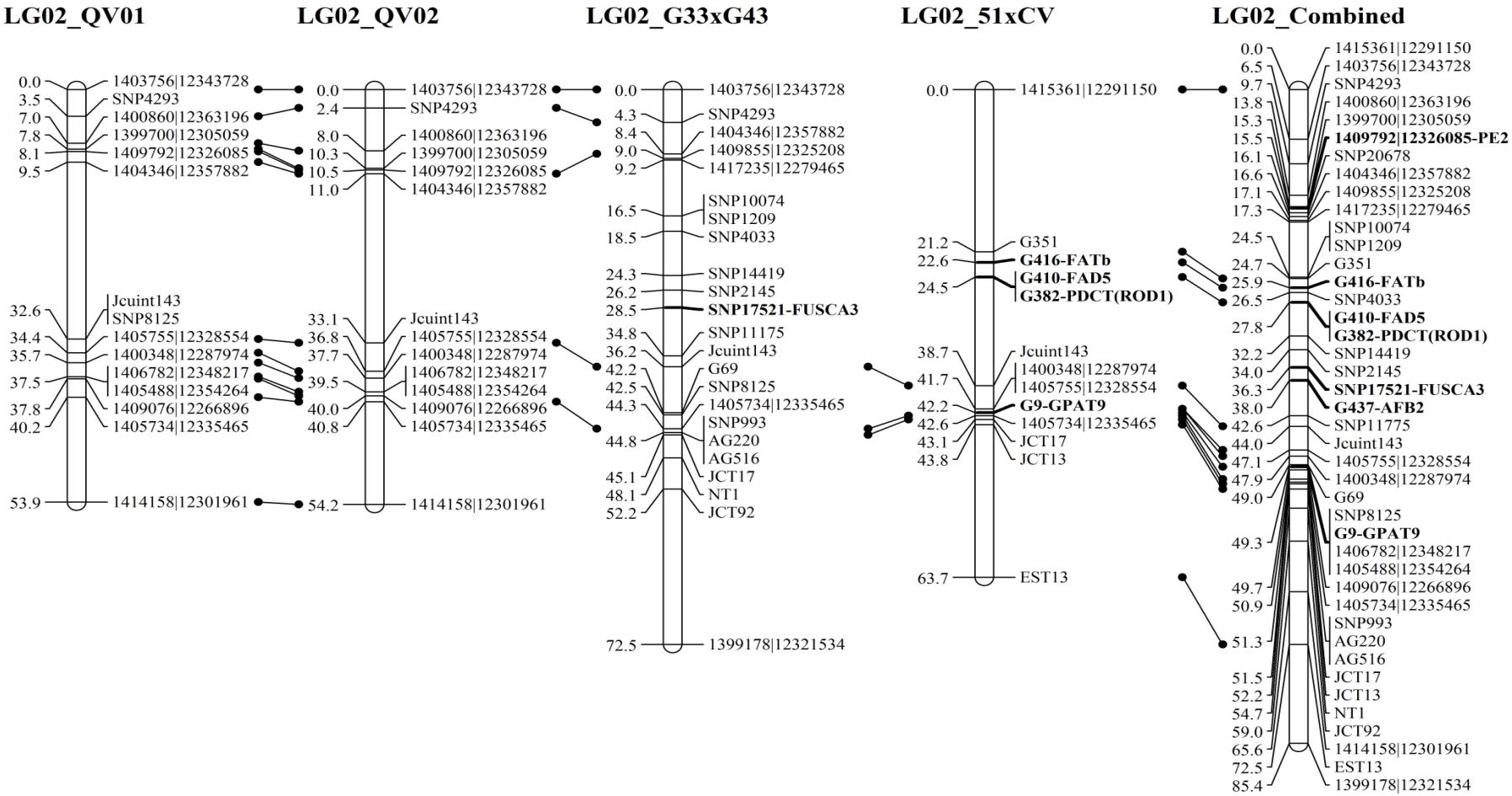


Figure 4-8 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.

Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

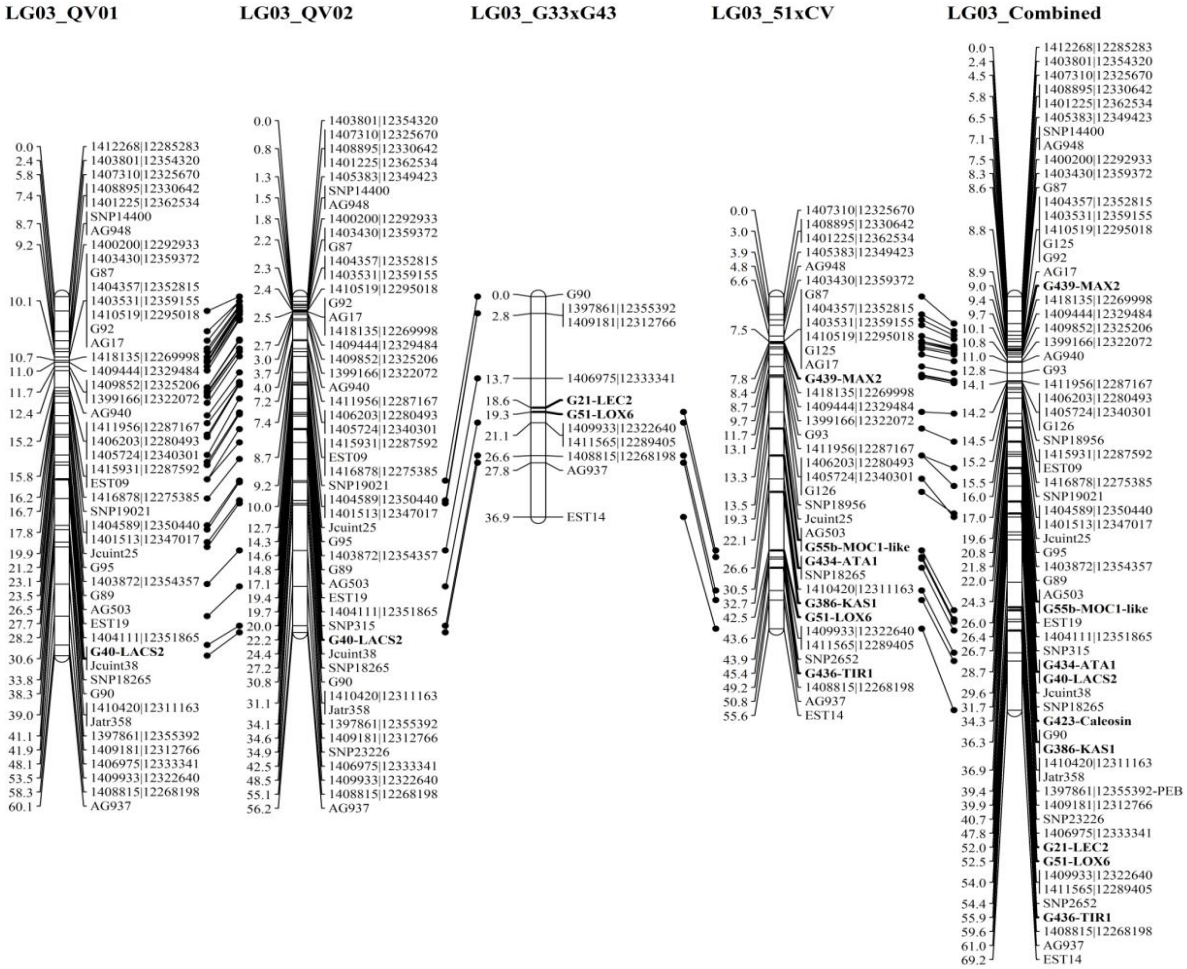


Figure 4-9 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.

Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

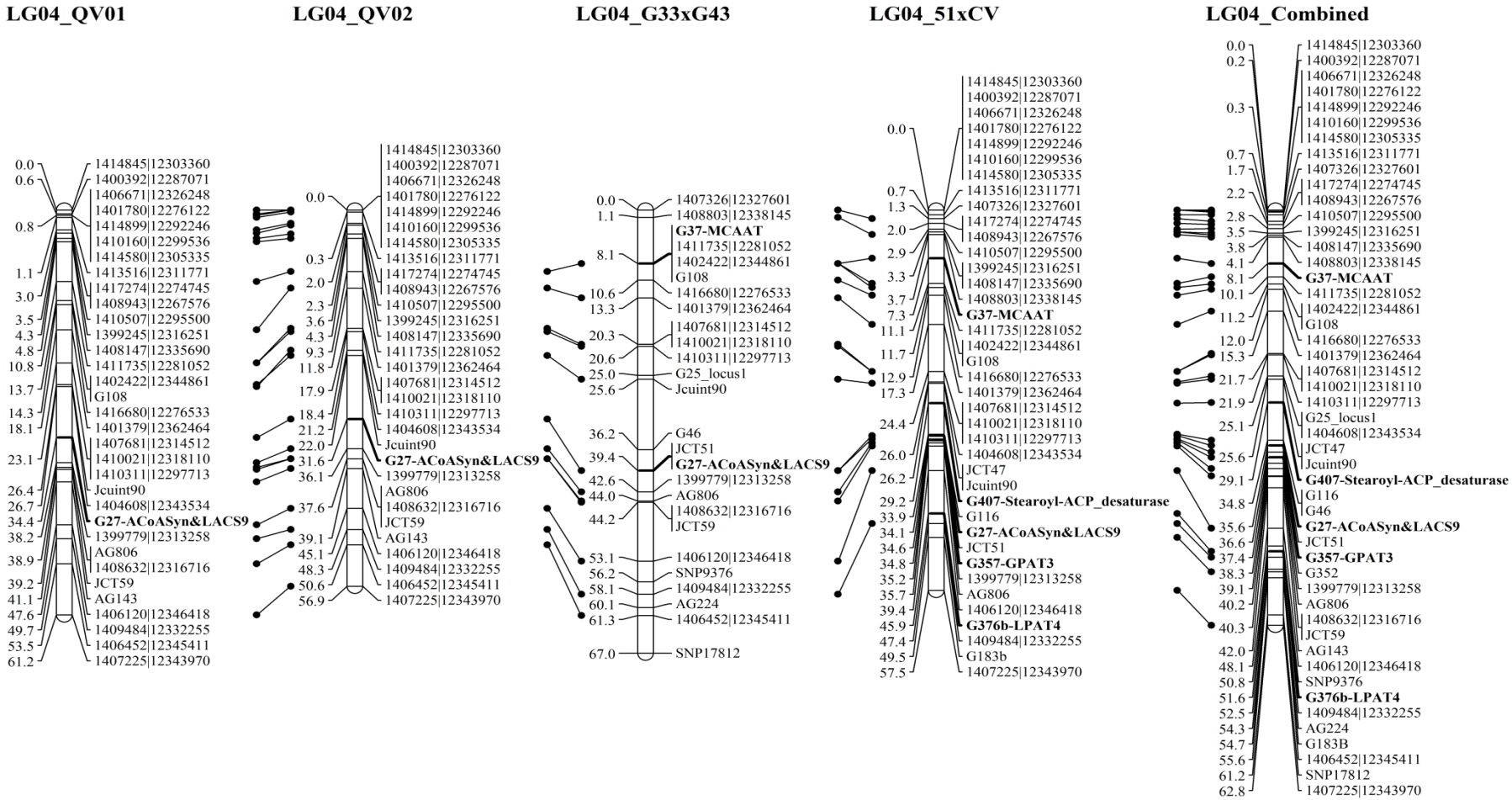


Figure 4-10 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.

Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

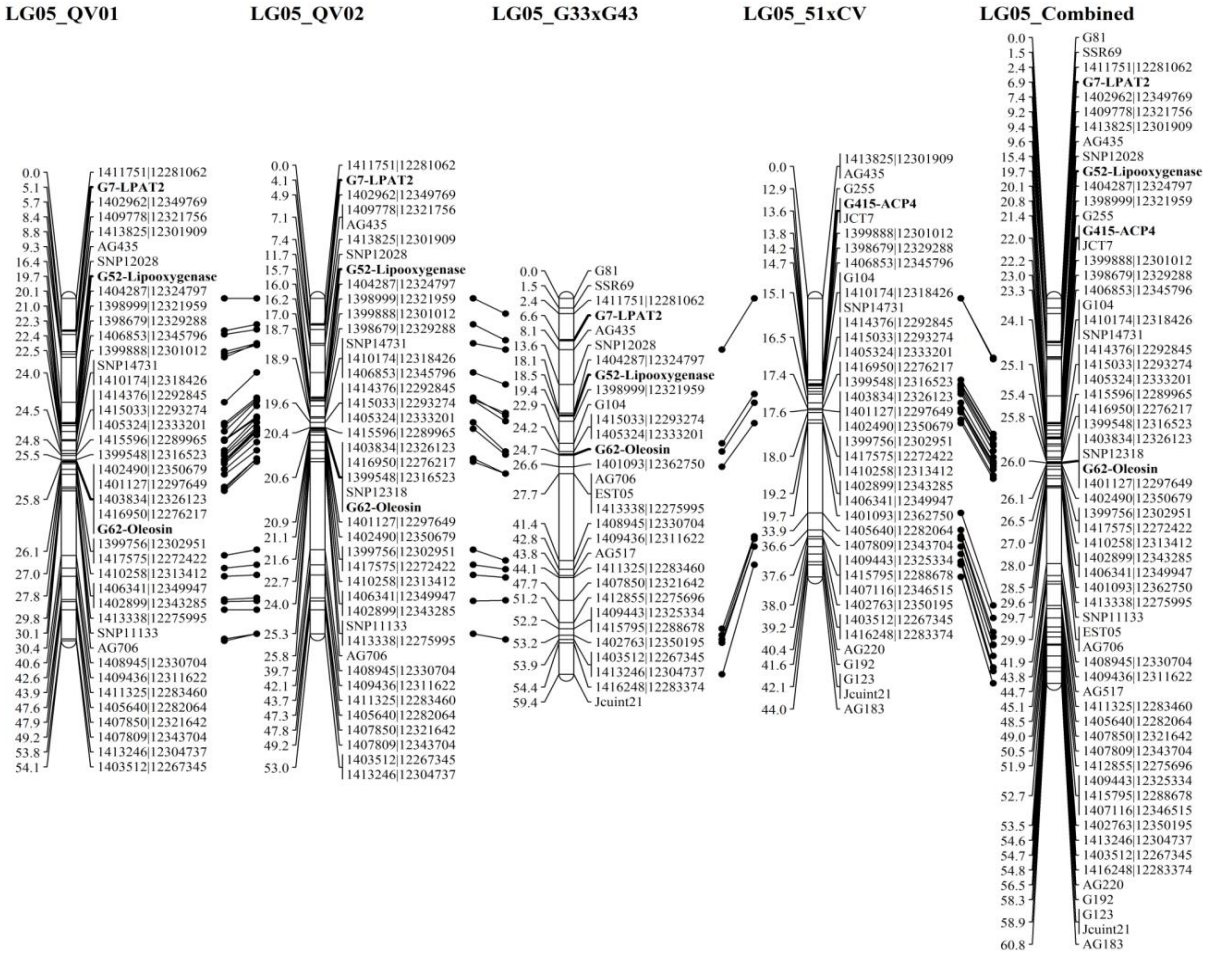


Figure 4-11 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.

Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

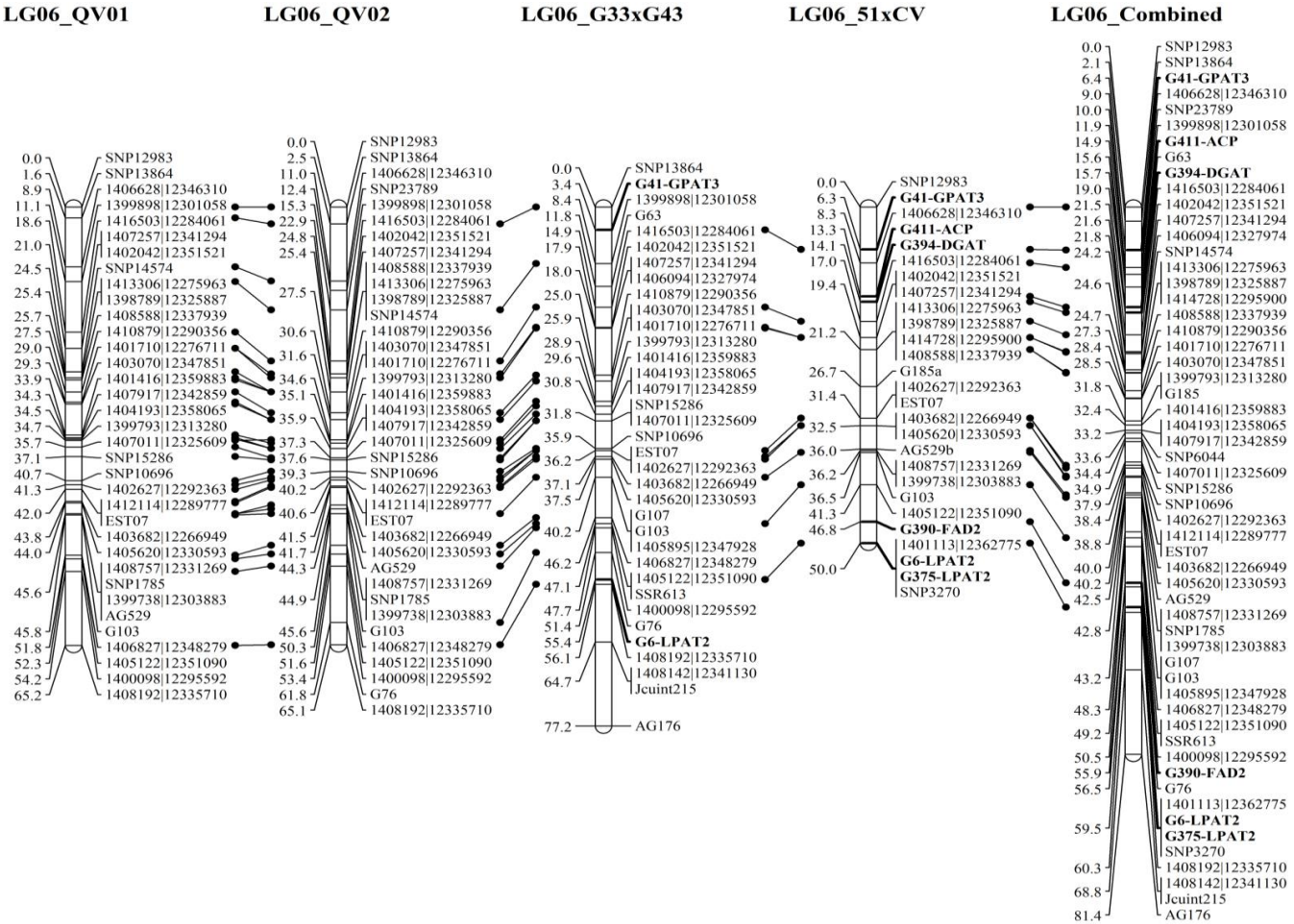


Figure 4-12 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.

Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

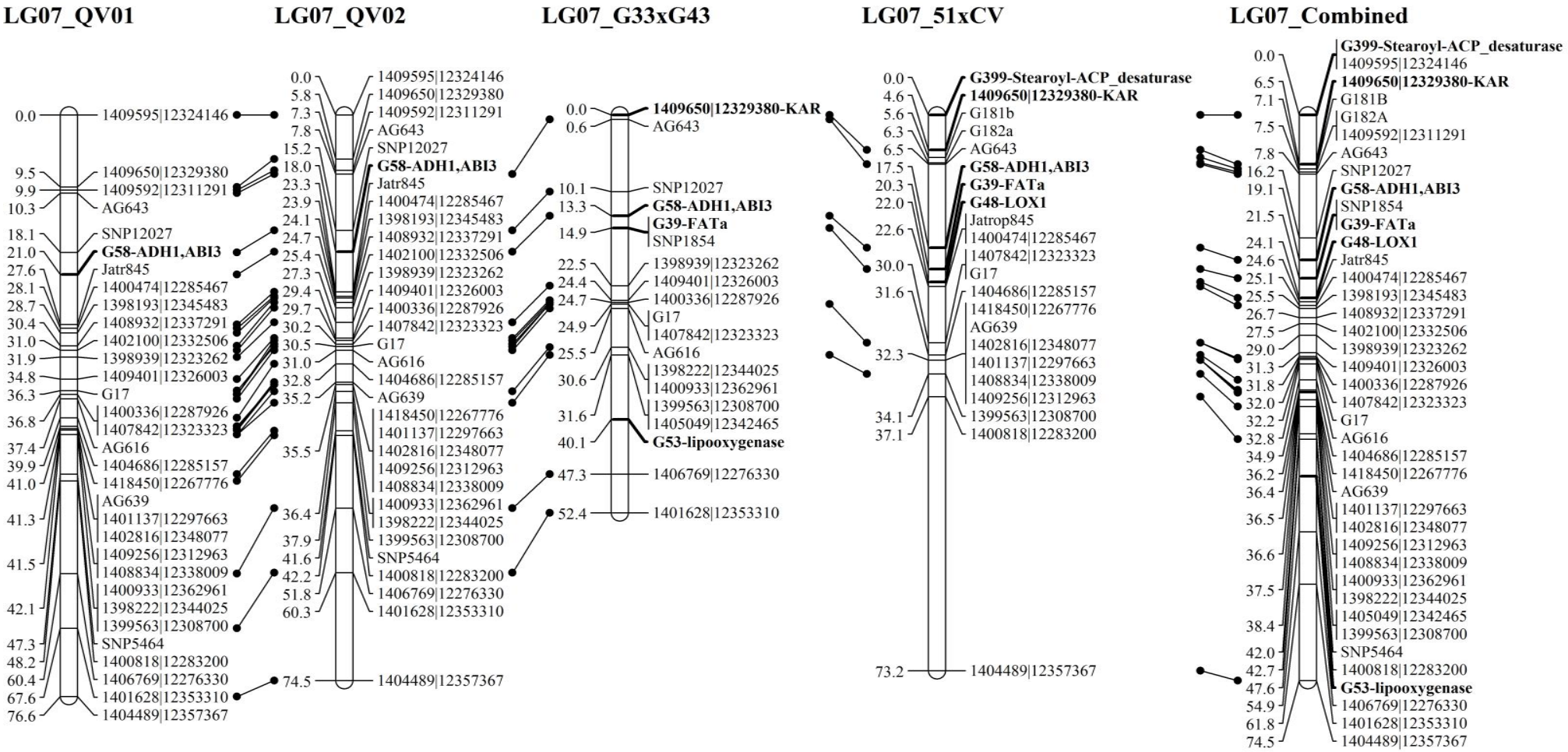


Figure 4-13 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.

Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

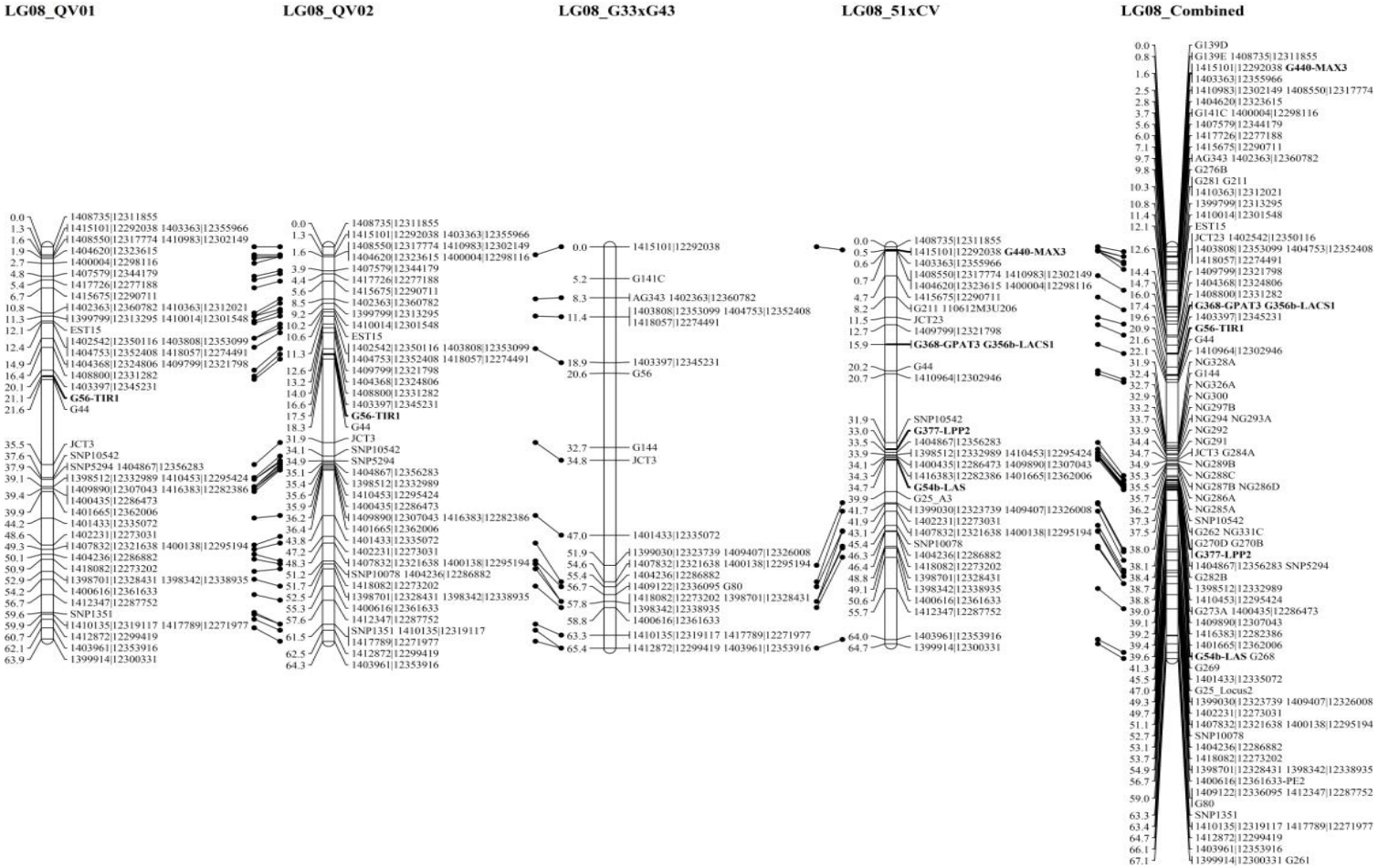


Figure 4-14 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.

Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

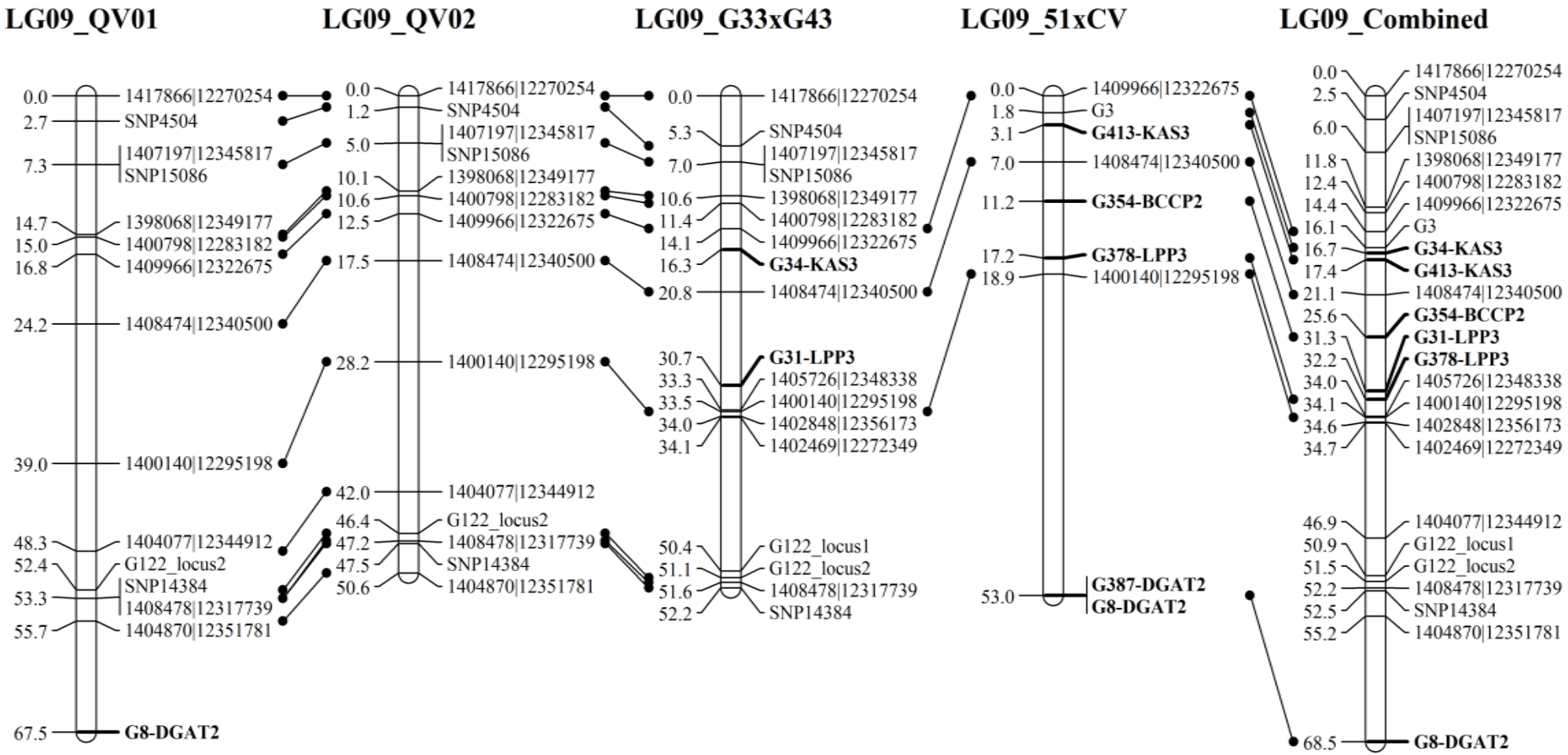


Figure 4-15 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.

Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

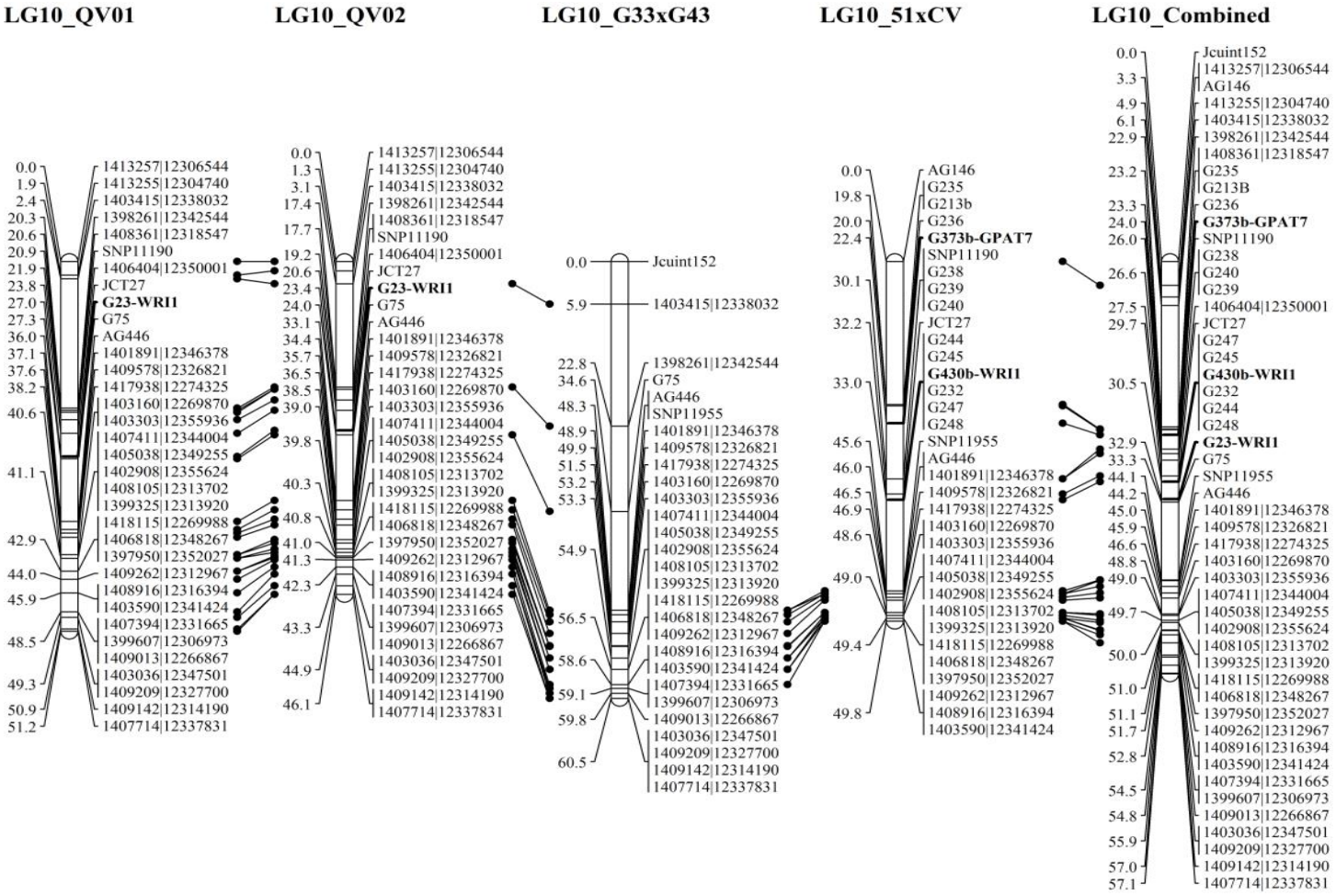


Figure 4-16 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.
Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

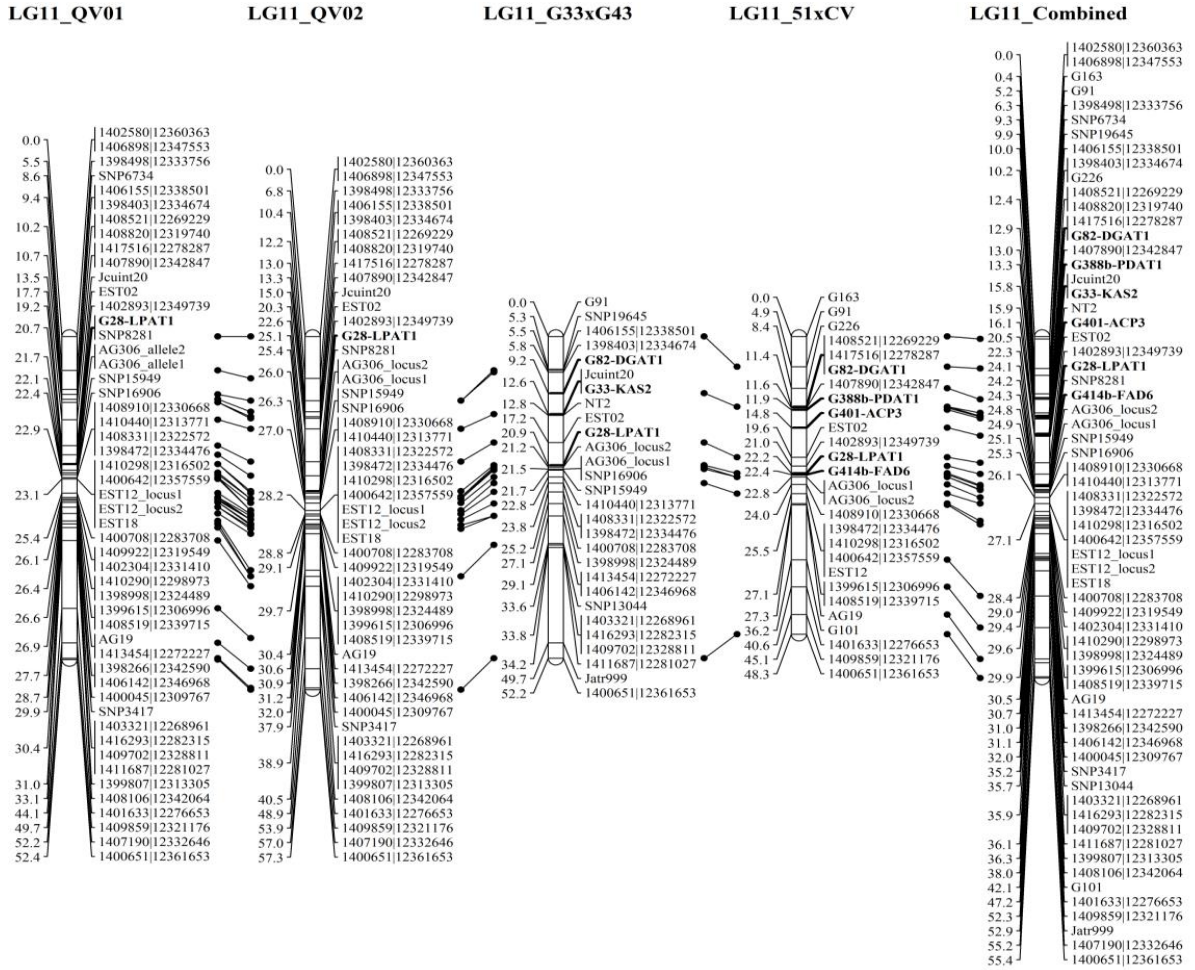


Figure 4-17 Physical alignment and comparison of linkage maps from independent mapping populations and the combined population dataset.

Shared markers (markers that were mapped in multiple populations) are linked with black lines to indicate the level of consensus between individual population maps. Note that the author of this study was responsible for developing markers that were mapped in all 5 linkage maps, whereas genotyping and linkage mapping was carried out solely in the 51xCV population. This comparison enables the accuracy of individual linkage maps to be compared, and the contribution of individual maps to the combined map to be visualised.

4.5.3: The Combined Genetic Linkage Map

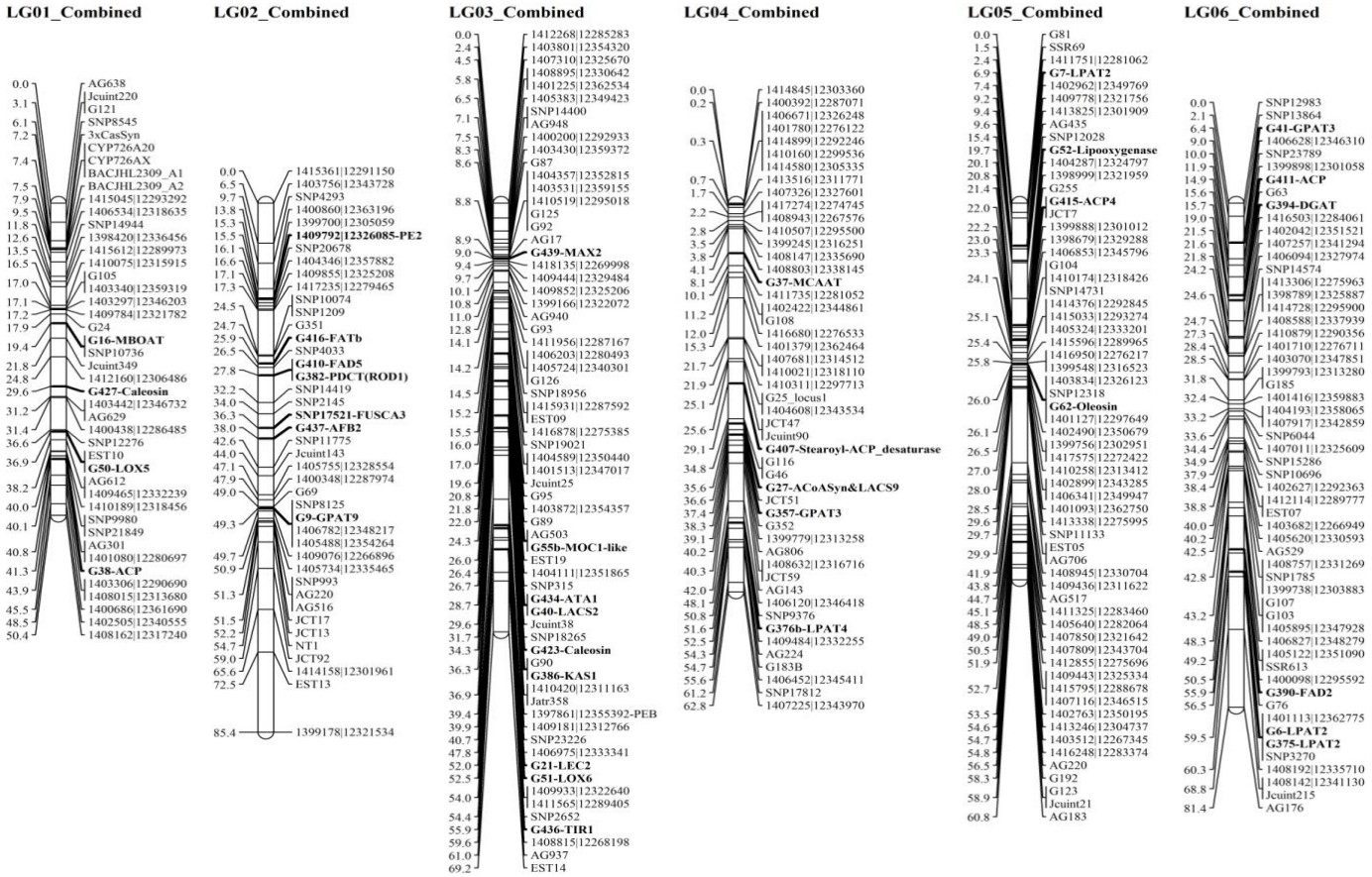


Figure 4-18 The Combined Genetic Linkage map, groups 1-6.

This map was generated from four F_2 mapping populations. In total, 589 co-dominant markers map 11 linkage groups and a distance of 733 cM. Average marker density is 1.62 cM per unique loci. A total of 67 candidate genes and trait-related metabolic genes were mapped (highlighted in bold).

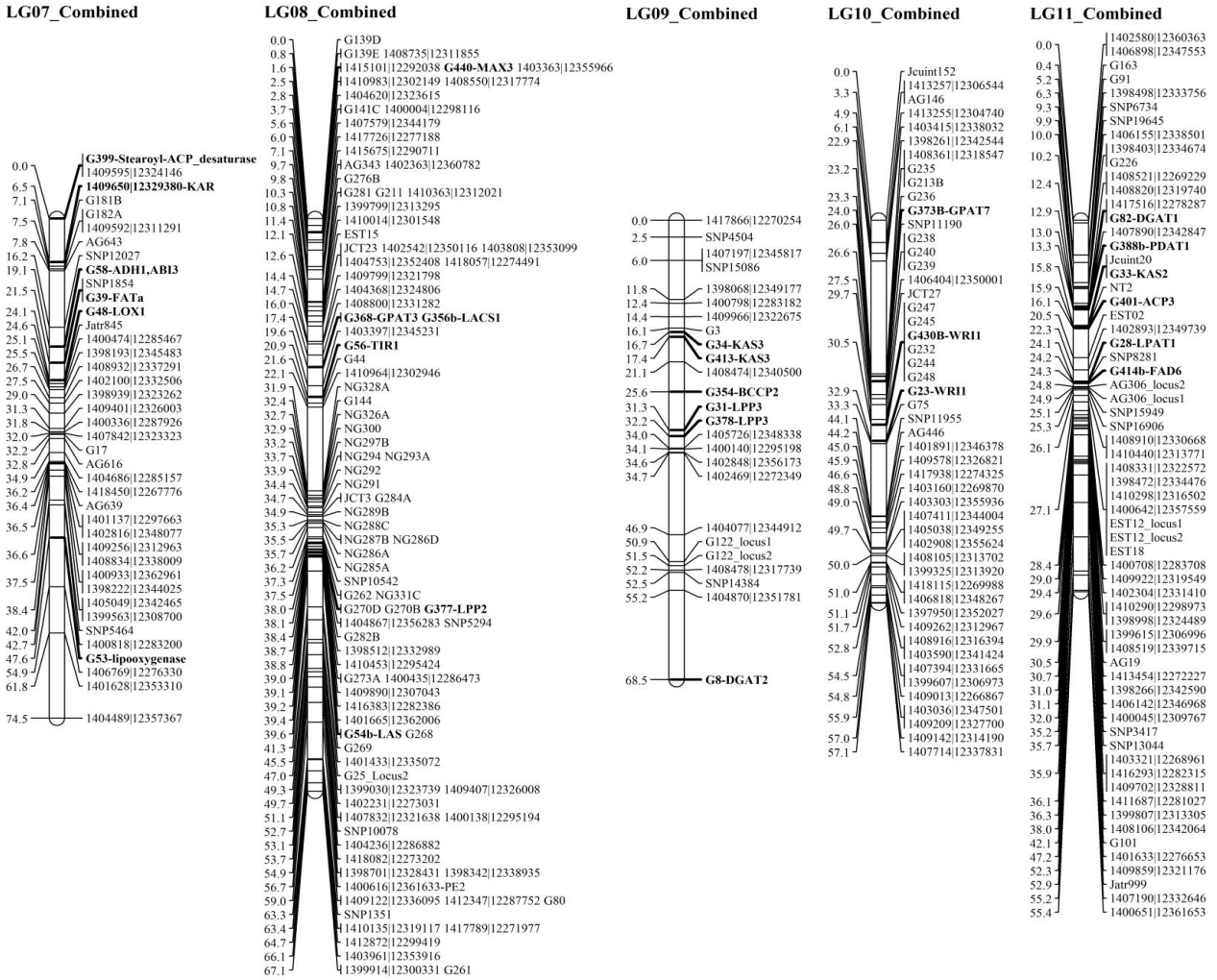


Figure 4-19 The Combined Genetic Linkage map, groups 7-11

This map was generated from four F_2 mapping populations. In total, 589 co-dominant markers map 11 linkage groups and a distance of 733 cM. Average marker density is 1.62 cM per unique loci. A total of 67 candidate genes and trait-related metabolic genes were mapped (highlighted in bold)

Chapter 5: Integration of phenotypic datasets identifies several QTL that contribute to oil yield and oil quality in the G51xCV mapping population.

5.1: Introduction

5.1.1: Target traits for the genetic improvement of *J. curcas*

After a mapping population and linkage map have been produced Quantitative Trait Locus (QTL) mapping can be carried out through the integration of phenotypic datasets. QTL mapping uses association to correlate genotype with phenotype (Mackay et al., 2009). Two commonly used methods to calculate QTL are (1) single marker analysis; the association between phenotypic values and genotype class at each marker position independently, and (2) interval mapping, which calculates the association between phenotypic values against set intervals across each linkage group (Doerge, 2002).

There are many traits that are desirable in a crop. Traits that affect overall yield, such as vigour, plant architecture, size and biomass, would be desirable within most agronomic plant species. For oilseed crop species intended for biofuel production, seed oil content and seed oil composition are also important traits. Increasing seed oil content increases the oil yield per seed, and seed oil extraction efficiency. Seed oil composition, the relative amounts of different fatty acids in the seed storage oil, in terms of carbon chain length and the presence/absence of different functional groups or bonds, has a large effect on the end biofuel performance characteristics and therefore determines oil quality (Durrett et al., 2008, Vega-Sanchez and Ronald, 2010, Balat, 2011, Atabani et al., 2013, Knothe, 2009). The properties of the different fatty acids affect the biofuel performance in combustion engines in several ways. Cetane number (the speed at which a fuel combusts, also a measure of explosiveness), the cloud point (the temperature at which the biofuel precipitates), coldflow point (a measure of viscosity at low temperatures), and oxidative stability (the rate at which fatty acids oxidise and degrade) are all defined by the biofuel fatty acid composition.

The amount of saturated and unsaturated bonds has been shown to substantially alter these properties. Biofuels high in saturated fatty acids such as 16:0 and 18:0 show favourable cetane numbers and oxidative stability, but they have reduced cold flow properties due to the increased density of saturated fatty acids. Conversely, oils high in unsaturated or polyunsaturated fatty acids have improved cold flow properties due the presence of carbon double bonds, that reduce packing density by introducing a kink into the fatty acid backbone. However this improved cold flow property is at the expense of both oxidative stability and CN number. Current opinion is that oil high in oleic acid (with a single double bond), is the optimal compromise between these properties (Vega-Sanchez and Ronald, 2010, Durrett et al., 2008, Graef et al., 2009, Knothe, 2009, King et al., 2009).

For Jatropha, a monoecious, self-compatible species, seed yield is thought to be highly correlated with the amount and type of branching, and the ratio of female to male flowers (Divakara et al., 2010, King et al., 2009). Inflorescences (flower nodes) develop at branch points, and within inflorescences, only the female flower produces seed (Wu et al., 2011, Fresnedo-Ramirez, 2013).

Both seed yield and oil yield could be affected by seed mass. Integration of seed mass, seed oil content and seed yield datasets, enabled oil yield to be calculated, and the relative contribution of each of these traits

could be determined within the G51xCV mapping population. Optimisation of traits that affect final oil yield will be vital for the economical production of *Jatropha curcas* as a biodiesel feedstock.

5.1.2: The G51xCV mapping population, and phenotypic dataset generation

Jatropha is perennial and asynchronous in flowering, and does not have a distinct growth and harvest season in contrast to the majority of widely cultivated crops. To enable seed to mature and accumulate *in situ* to a level suitable for harvest, seed was collected at 5 time points throughout the year. Similarly, although Jatropha is reported to produce seed from the 1st year of growth, seed yield increases with size and maturity of plant (Fresnedo-Ramirez, 2013), up to the reported full maturity stage that is reached after 5 years (Fresnedo-Ramirez, 2013, Atabani et al., 2013). Seed batches were sent for analysis from years 2 and 3 (datasets Year 2 and Year 3a). These two collection points roughly equate to the same growth and harvest time for each seed batch.

Due to the high number of seeds produced in Year 3, a second batch of seed was sent which was sampled from later in the year (dataset Year 3b). Since Year 3 seed batches experienced slightly different environmental conditions (separated by several months growth time between harvests), each seed batch was treated as a separate dataset.

Seed oil related phenotypic datasets; Seed oil content, Seed oil quality (fatty acid composition) and Seed weight (100 seed mass), were collected at the University of York using Nuclear Magnetic Resonance (NMR) and Gas Chromatography (GC), described in further detail in the Materials and Methods.

Number of branches datasets were collected in years 1 and 2, prior to pruning of non-seed producing branches in year 3 in line with recommended agronomic practice for Jatropha cultivation (personal communications Dr. Luis Montez). Seed yield (number of seeds) data was collected at 5 time points per year, and a sample period of 1st February to 31st January was used to create a Year 2 and Year 3 seed yield dataset (expressed as total number of seeds produced per plant).

Seed oil content (% of total seed mass), seed mass (average mass per seed) and seed yield (number of seeds produced per plant), enabled oil yield per plant to be calculated for Year 2 and 3 datasets, as a product of these three traits.

Table 5-1 Seed and branching sample dates, and dataset naming.

Year	2011	2012					2013				
Date	13th Dec	26th Jun	13th Sep	1st Oct	12th Oct	15th Oct	10th Jan	22nd May	28th May	16th Aug	14th Oct
Years of growth	1.76	2.30	2.51	2.56	2.59	2.6	2.84	3.2	3.22	3.44	3.60
Days after transplanting	567	763	842	860	871	874	961	1093	1099	1179	1238
Measurements taken in the field	Branching	Seed harvest, Branching	Seed harvest	Seed harvest	Seed harvest	Seed harvest	Seed harvest	Seed harvest	Seed harvest	Seed harvest	Seed harvest
Sample period for seed sent to York		Batch 1						Batch 2		Batch 3	
Trait	**Dataset name and sample period**										
Seed oil content		Year 2						Year 3a		Year 3b	
Seed oil composition		Year 2									
Seed mass		Year 2						Year 3a		Year 3b	
Branching	Year 1	Year 2									
Seed Yield		Year 2					Year 3				
Oil yield		Year 2					Year 3				

5.2: Results

5.2.1: Phenotypic trait population distributions

5.2.1.1: Phenotypic traits are normally distributed in the G51xCV mapping population.

Phenotypic traits in the G51xCV mapping population have Normal population distributions (Figure 5-1, p120), indicative of quantitative, continuous traits, suitable for QTL mapping. The equal and balanced distribution of values around the mean, as indicated by the symmetry and low skewness values for the lines of best fit, Figure 5-1, p120 and Table 5-2, p122, shows that this is an unbiased mapping population that is not under any distorting selective pressure. In some cases, the dataset means are lower or higher than the mean for the Normal line of best fit, due to the presence of outliers that sit outside the normal distribution and decrease or increase the dataset mean respectively. In most cases the median value is a more accurate reflection of the centre of the distribution.

5.2.1.2: Phenotypic traits in the G51xCV population, show a high level of variation in seed oil content, seed oil composition and other oil yield related traits that can be investigated by QTL mapping

The G51xCV mapping population was created from parents selected primarily for oil content; CV at 26.00 % seed oil content, and G51 at 36.90 % seed oil content. Therefore it is of interest to see if this trait shows variation, and segregates, in the F2 population. Seed oil content ranges from 26.00-39.75 % in Year 2, 19-40.0 % in Year 3a, and 23.60-40.30 % in Year 3b (Table 5-2, p122). The middle 90 % of plants fall between 30.80-39.40 % in Year 2, 27.50-39.10 % in Year 3a, 29.68-39.21 % in Year 3b. For all three datasets, the middle 90 % of plants are within typical seed oil content ranges reported in Jatropha (between ~30-40 % seed oil content) (Balat, 2011, Achten et al., 2007, Wang et al., 2008, Yue et al., 2013, Heller, 1996). The variation between the minimum and maximum, and the middle 90 % of plants, show that there is significant variation in this trait, that is consistent across independent harvests. There are also a number of consistently low outliers that are between ~20-25 % seed oil content in all three datasets (Figure 5-1, Parts A-C, p120), which is similar to the phenotype of the CV parent at 26 % seed oil content, and a number of high oil F2 plants at 37 % seed oil content or above, which is similar to the phenotype of the G51 parent at 36.90 % seed oil content.

There is also significant variation in seed oil composition for the four most common fatty acids; stearate, palmitate, oleate and linoleate. The middle 90 % of values; Stearate 11.23-13.41 %, Palmitate 6.76-8.71 %, Oleate 43.93-49.11 %, Linoleate 28.09-33.87 %, are ranges typically found within Jatropha material (Atabani et al., 2013, Balat, 2011, King et al., 2009). For high oleate biodiesel fuels, changes of a few percent oleate content have significant effects on biofuel performance (Durrett et al., 2008, Knothe, 2009). In the G51xCV mapping population, oleate varies by 5.18 % for the middle 90 % of plants, and 7.90 % between the minimum and maximum values.

Seed yield (number of seeds), seed mass and branching also show significant variation. In contrast to seed oil content which has a number of low outliers, both branching and seed yield show a consistent number of high outliers that are outside the normally distributed F2 plants, Figure 5-1, parts G-J, p120.

In summary there is a high level of phenotypic variation in the traits measured, including a portion of plants that segregate close to the parental phenotypes for seed oil content, and outliers that are consistent across independent harvests, together suggesting that there is a strong genetic component to this phenotypic variation.

5.2.1.3: Correlation analysis shows a number of significant correlations between seed and vegetative traits in the G51xCV mapping population

There are a number of significant correlations between seed and vegetative traits in the G51xCV mapping population as identified by Pearson's correlation analysis, Table 5-3, p123. For all traits measured, there is highly significant (p<0.01) correlation between independent harvests/datasets for each trait, showing that phenotypes are consistent across multiple harvests in the F_2 population.

The seed related traits, 100 seed mass and seed oil content, show either strong highly significant correlation or weak highly significant correlation to each other depending on the harvest, Table 5-3, p123. The positive correlation between seed mass and seed oil content, suggests that variation in seed mass can be partly

attributable to variation in seed oil content; either partly attributable for weak correlations, or predominantly attributable for strong correlations. The difference in magnitude of correlation between different harvests, could be due to environmental effects. For example correlation between Seed oil content Y2 and 100 seed mass Y2, and seed oil content Y3b and 100 seed mass Y3b, is R=0.440 and R=0.454 respectively, whereas correlation between seed oil content Y3a and 100 seed mass Y3a is R=0.700. As can be seen by Table 5-1, p110, these seeds were harvested at different times of the year. This correlation data suggests that the year 3a dataset experienced environmental conditions where seed mass was most affected by the rate of seed storage oil deposition rather than other seed constituents that could affect seed mass, for example proteins, fibre or polysaccharides. As mentioned in Materials and Methods, seed mass data was normalised to 7 % water content to eliminate this as a variable.

Weak but significant correlations were detected between the seed related traits; seed oil content and 100 seed mass, and the vegetative traits; branching and number of seeds, Table 5-3, p123. Both seed oil content (Y3a) and 100 seed mass (Y3a and Y3b) showed weak but significant correlation with number of seeds Y3. The positive correlation between seed yield and seed oil content, shows that seed oil content is not mutually exclusive with seed yield. Plants that have high seed oil content also tend to have high seed yields in this population. This correlation is present in Year 3 datasets but not Year 2 datasets, suggesting that there could be an environmental effect or an effect attributable to the maturity of the plants.

Branching is weakly but significantly correlated with seed yield (number of seeds) in both Y2 and Y3 (p<0.000). Flower nodes are known to occur at branch points, therefore the more branched a plant the more sites available for flower and seed development. The weak correlation despite the high significance, could be due to the fact that branching does not take into account the ratio of female to male flowers at each branch node, meaning an additional factor is missing when directly correlating branching and seed yield (number of seed) traits.

5.2.1.4: Fatty acid moieties are most highly correlated with other fatty acid moieties suggesting that metabolic pathway regulation and the shuttling of fatty acids from different pools to seed storage oil is the major regulator of seed oil composition (oil quality) in the G51x CV mapping population.

Correlation between the four most common Jatropha seed oil fatty acids, is negative for nearly all pairwise comparisons in this mapping population, Table 5-4, p124. This could be expected due to the fact that these 4 moieties are part of the same biosynthetic pathway (see chapter 3, figure 3-1), therefore an increase or decrease in one fatty acid would be expected to be at the expense of the other fatty acids in the pathway, assuming a limited fatty acid pool or limiting rate of *de novo* synthesis. An exception to this, is the correlation between stearate and oleate, which is positive. In this mapping population plants with high oleate content also tend to have higher stearate content and visa versa. A possible explanation for this, is that if the rate of oleate to linoleate conversion is reduced, it could cause upstream moieties oleate and stearate to accumulate and visa versa, since more of these fatty acids would be available for incorporation into the seed storage oil.

Whilst most correlations are weak, the correlation between oleate and linoleate is much stronger (R=-0.835 at p<0.000 significance). Since conversion of oleate to linoleate occurs in a single step by the action of the fatty acid desaturase 2 gene (of which there is one endoplasmic homologue in *J. curcas* (Ye et al., 2009, Liu et al.,

2011, King et al., 2011, Sato et al., 2011, Gu et al., 2012, King et al., 2013, Jiang et al., 2012, Wu et al., 2013, Costa et al., 2010, Qu et al., 2012, Utomo et al., 2015), this suggests that manipulation of this gene would be highly effective at altering the ratio of oleate to linoleate in Jatropha seed storage oil.

5.2.1.5: Linoleate seed oil content is positively correlated with overall seed oil content, and seed mass, suggesting that there is a weak bias towards storing excess fatty acids as the linoleate moiety in the seed storage oil of G51xCV F2 plants.

Linoleate content is positively correlated with both seed oil content and seed mass in the G51xCV mapping population, Table 5-4, p124. As mentioned in section 5.2.1.3: p111, seed mass is correlated with, and partly determined by, seed oil content. The positive correlation between linoleate content and seed oil content, shows that as the amount of oil per seed increases so does the relative amount of linoleate, suggesting that there is a bias towards storing excess fatty acids as linoleate in the seed storage oil pool. The bias towards storing excess fatty acids as linoleate could be due to the promiscuity of the fatty acid desaturase 2 gene located on the endoplasmic reticulum fatty acid pool, which is supported by the strong negative correlation between linoleate content and oleate content. This is further supported by the correlations between oleate, linoleate and 100 seed mass year 2 datasets. The oleate content is negatively associated with seed mass, whilst the linoleate content is positively associated, suggesting that as the seed mass increases, in part due to seed oil content, the relative amount of linoleate compared to oleate increases.

5.2.1.6: Palmitate content is weakly correlated with branching and seed yield, suggesting that vegetative traits that contribute towards seed yield can have minor effects on seed oil composition (oil quality) in the G51xCV mapping population

Palmitate content has weak but significant, positive correlation with branching and seed yield in the G51xCV mapping population, Table 5-4, p124. This suggests that there is a tendency for the relative proportion of palmitate to increase as the amount of branching and seed yield increases, through an unknown mechanism. This demonstrates that, although a weak association, it is possible for vegetative traits such as branching and seed yield to be correlated with the fatty acid composition of the seed storage oil.

5.2.1.7: Oil yield is correlated with nearly all seed and vegetative traits measured, however it is most strongly correlated with number of seeds per plant, showing that seed yield is the strongest determinant of oil yield in the G51xCV mapping population.

Oil yield in Years 2 and 3 is significantly correlated with all seed and vegetative traits measured in the G51xCV mapping population with the exception of Seed oil content Year 2 and Seed oil content Year 3b, and the seed oil composition fatty acids Stearate, Oleate and Linoleate, Table 5-5, p124. Correlations are weak but significant with the exception of seed yield (number of seeds) which is strongly and significantly correlated (Oil yield Year 2 and Number of seeds Year 2, R=0.972, and Oil yield Year 3 and Number of seeds Year 3, R=0.948). This strong correlation shows that of all the vegetative and seed traits measured, seed yield is the strongest determinant of oil yield in the G51xCV mapping population. This stronger correlation may be due to the greater level of variation in number of seeds compared to seed oil content and seed mass in this population.

Given that seed mass is an indirect measurement of seed oil content, the traits seed yield and seed oil content are the two most important determinants of oil yield in this population. Since variation in seed yield is much greater than variation in seed oil content, seed yield has greater influence on final oil yield in this population. Seed oil content and seed mass do have a significant effect on oil yield, therefore they remain important traits for QTL analysis, particularly for introgression into cultivars that may be high yielding in terms of seeds, but require improvement of seed oil content or seed mass traits.

5.2.1.8: Palmitate content is weakly associated with oil yield in the G51xCV mapping population

Palmitate content is weakly correlated with oil yield in both year 2 and year 3 datasets, Table 5-5, p124. Palmitate content is also correlated with seed yield and branching in this mapping population. In turn seed yield and branching are highly correlated with oil yield. There is insufficient evidence to determine whether Palmitate and Oil yield are directly regulating each other through an unknown mechanism, or whether the co-correlation of Palmitate and Oil yield to traits such as seed yield or branching, mean they are indirectly correlated. Either way in this population, palmitate is a (weak) marker/predictor of oil yield.

5.2.2: Quantitative trait locus mapping

5.2.2.1: Interval mapping reveals a number of QTL for oil quality and oil yield-related traits in the G51xCV mapping population.

5.2.2.1.1: Seed oil content analysis identifies two QTL on linkage groups 4 and 10

Figure 5-2 part A, p126, shows two QTL for seed oil content across the Year 2, 3a and 3b datasets. QTL are located on linkages groups 4 and 10 (Year 2), linkage group 4 only (Year 3a) and linkage group 10 only (Year 3b). Association reaches LOD4.624, 4.414, 3.527, 3.503 respectively, which is significant at experiment wide thresholds ($p<0.05$ for Year 2, and $p<0.01$ for Year 3a, and 3b), Table 5-6, p125. PVE values for these loci, are 16.56, 15.75, 13.42, and 15.95 percent respectively. The fact that the QTL on linkage groups 4 and 10 are both significant in the Year 2 dataset, whereas only one of each is significant in the Year 3a and 3b datasets could be due to environmental effects that limit the effects of one of the QTL in each dataset. As previously outlined, all three datasets were harvested at different times of the year, Table 5-1, p110. Alleles at these QTL loci are dominant, as shown by Tukey's comparison of means tests, with the high oil allele originating from the G51 parent, Figure 5-3 parts A, B, C, p129.

Figure 5-3, part A, p129, examines the effect of both QTL (linkage group 4 & 10) on seed oil content in the Year 2 dataset. From left to right, the number of G51 'b' alleles increases at both loci, from AA (homozygous A at both QTL positions) to BB (homozygous B at both QTL positions). As can be seen, as the number of 'b' alleles increase at both loci, so does the seed oil content. However, in part due to the low number of plants in each genotype class (N=145, Table 5-2, p122, for 9 genotype classes, Figure 5-3, part A, p129), Tukey's post hoc comparison of means test shows that we can only be statistically confident that AA, and HH/HB/BH are statistically different. Nevertheless, we can see from these groups alone that as the number of 'b' alleles at each loci increases, so does the average and minimum and maximum oil content values. This shows that these QTL loci act in a synergistic manner, and that they could be stacked or pyramided in a single cultivar for greater enhancement of seed oil content.

5.2.2.1.2: 100 seed mass analysis identifies one QTL on linkage group 4

One QTL was detected for 100 seed mass on linkage group 4 in Years 2, 3a and 3b datasets, Figure 5-2 part A, p126. This QTL reaches experiment wide significance thresholds ($p<0.01$ for Year 2 and 3a, and $p<0.05$ for Year 3b). The calculated QTL position differs by ≤8 cM across the 3 datasets. For Year 2, 3a and 3b datasets association reaches LOD7.776, 4.964, 3.285 with PVE values of 29.39, 19.39, and 14.89 percent respectively, Table 5-6, p125. Alleles at these loci are dominant, with the beneficial allele originating from the G51 parent Figure 5-3 parts A, B, C, p129.

5.2.2.1.3: Branching (number of branches) analysis identifies one QTL on linkage group 1

One QTL was detected for branching @763days on linkage group 1 at the $p<0.05$ experiment wide significance threshold, Figure 5-2 part D, p126. Association at the QTL is LOD3.477, which accounts for a PVE of 12.115 percent, Table 5-6, p125. The allele at this QTL is dominant as determined by Tukey's comparison of means test, with the beneficial allele originating from the CV parent, Figure 5-3 part G, p129, which is in contrast to the seed oil content and seed mass QTL where the high allele is G51.

5.2.2.1.4: Seed yield (number of seeds) identifies one QTL on linkage group 10

One QTL was detected for number of seeds Year 3, linkage group 10 at the $p<0.05$ experiment wide significance threshold, Figure 5-2 part E, p126. Association was LOD 3.966 with a PVE of 14.15 percent, Table 5-6, p125. As with branching, the beneficial allele is dominant and originates from the CV parent, Figure 5-3 part H, p129.

5.2.2.2: Oil quality (composition) traits

5.2.2.2.1: Palmitate content analysis identifies two QTL on linkage groups 5 and 7

Two QTL were detected for palmitate content on linkage groups 5 and 7 at the $p<0.01$ and $p<0.05$ experiment wide significance thresholds respectively, Figure 5-2, Part C, p126. Association reached LOD7.929 and 3.24 which accounted for 30.54 and 11.51 percent PVE respectively, Table 5-6, p125. The linkage group 5 QTL has a semidominant character according to Tukey's post hoc comparison of means test, with the high palmitate allele originating from the CV parent, , p130. For the linkage group 7 QTL, the high palmitate allele is recessive and also originates from the CV parent according to the Tukey's comparison of means test.

5.2.2.2.2: Stearate content analysis identifies three QTL on linkage groups 1, 4 and 7

Three stearate QTL were identified on linkage groups 1, 4 and 7, Figure 5-2, Part C, p126. All three QTL surpassed the $p<0.01$ experiment wide significance threshold, with peak association at LOD4.144, 4.226 and 6.606 respectively, Table 5-6, p125. PVE for the 3 QTL were 14.95, 15.26 and 24.86 percent respectively. The linkage group 1 QTL was dominant according to Tukey's comparison of means test, with the high stearate allele originating from the G51 parent, Figure 5-4, p130. The linkage group 4 QTL is semi-dominant with the high stearate allele originating from the CV parent. The linkage group 7 QTL is also semi-dominant but the high stearate allele originates in the G51 parent.

5.2.2.2.3: Oleate content analysis identifies one QTL on linkage group 6

A single oleate QTL was identified on linkage group 6, at LOD3.398 which surpassed the $p<0.05$ experiment wide significance threshold, Figure 5-2, Part C, p126. This QTL was attributable for a PVE value of 12.10 percent, Table 5-6, p125 . The high oleate allele originated from the CV parent, and has a recessive nature, Figure 5-4, p130, however interestingly the homozygous G51 genotype ('B') was not significantly different from either the homozygous CV genotype ('A', high oleate) or the heterozygous genotype ('H', low oleate). This may be due to a heterosis effect whereby when one of each allele is present in the same genotype a low oleate phenotype is generated, presumably due to interaction between the two alleles or factors within physically linked DNA.

5.2.2.2.4: Linoleate content identifies three QTL on linkage groups 4, 6 and 8

Three linoleate content QTL were identified on linkage groups 4, 6 and 8 at the $p<0.05$ (linkage groups 4 and 8 QTL) and $p<0.01$ (linkage group 6 QTL) experiment wide significance thresholds, Figure 5-2, Part C, p126. Association at the QTL loci reached LOD3.995, 4.307 and 3.287, which accounted for 14.37, 15.58 and 11.68 percent PVE, for linkage groups 4, 6 and 8 QTL respectively, Table 5-6, p125. For linkage group 4 and 6 QTL, the high linoleate content allele originated from the G51 parent and was dominant, whereas for the linkage group 8 QTL the high linoleate content allele originated from the CV parent but was also dominant, Figure 5-5, p130.

5.2.2.2.5: Multiple QTL for palmitate, stearate, oleate and linoleate co-locate to the same linkage groups, providing evidence that these loci contain genes responsible for the shuttling of fatty acids through parts of the seed fatty acid synthesis and modification metabolic pathway.

5.2.2.2.6: A Palmitate and Stearate QTL co-locate to linkage group 7

Conversion of Palmitate to Stearate occurs in a single step by the action of the Keto-Acyl Synthase 2 (KASII) gene, through the addition of a 2 carbon acyl ACP group (see chapter 3, fig 3-1). The KASII gene would be the most likely candidate gene to be contained within QTL regions that co-locate for both traits, since conversion of one moiety into the other, mediated through a change in KASII activity, would affect both traits. This is supported by the strong negative correlation between Palmitate and Stearate content according to Pearson's correlation analysis, 5.2.1.4: p112 and Table 5-4, p124. In the G51xCV mapping population a Palmitate and Stearate QTL co-locate to linkage group 7, Figure 5-2, p126. The Jatropha KASII gene was mapped using a nearby SSR marker (Marker G33, J. Clarke) which maps to linkage group 11. One explanation is that a causative gene on linkage group 7 could regulate the expression of the KASII gene on linkage group 11. A way to investigate this would be to carry out expression QTL mapping (eQTL mapping) since if this hypothesis was correct both the linkage group 7 and linkage group 11 loci would associate with palmitate and stearate phenotypes.

5.2.2.2.7: An oleate and linoleate QTL co-locate to linkage group 6

Oleate is converted to linoleate through desaturation of the delta-12 carbon of the fatty acid backbone, through the action of the Fatty Acid Desaturase II gene (FAD2), (see chapter 3, fig 3-1). A single locus that associates with changes in both oleate and linoleate would suggest that this locus is controlling the

conversion of oleate to linoleate, or the activity of the FAD2 gene. The linkage group 6 oleate and linoleate QTL map to within 2 cM of each other (2 cM and 4 cM respectively), supporting the hypothesis that a single locus is regulating both traits. The FAD2 gene was mapped using SSR's (marker G390, J. Clarke, see chapter 3 appendix), and maps to linkage group 6 at position 55.9 cM, which unfortunately is outside of the Bayes' 95 % confidence interval for these QTL. Given the known activity and predominance of the FAD2 gene for this metabolic conversion, this QTL could be a trans regulator of the FAD2 gene located on the same linkage group. Other mapped candidate genes, that lie within the QTL region are a GPAT and DGAT (markers G41, G394, J. Clarke, see chapter 3 appendix). These genes are responsible for incorporating fatty acids onto a glycerol backbone to form triglycerides, that are subsequently stored in the seed storage oil bodies. Some GPATs and DGATs have been known to show selectivity towards certain types of fatty acids (in terms of the rate at which they incorporate different molecules into triacylglycerol) (Snyder et al., 2009, Cahoon et al., 2007, Graham et al., 2007, Yen et al., 2008, Zheng et al., 2008, Baud and Lepiniec, 2010, Li-Beisson et al., 2013, Bates and Browse, 2012, Bates et al., 2012) which could lead to accumulation of one fatty acid over the other.

5.3: Discussion

Overall, a number of agronomically-relevant QTL were identified in the G51xCV mapping population. The primary purpose of this mapping population was the identification of seed oil content QTL, due to the variation present in the parents; G51 (36.9 % seed oil) and CV (25 % seed oil), of which 2 QTL were identified that were responsible for a combined PVE of 32.31 % seed oil content. Due to its intended use as a biofuel crop, any trait that contributes towards final oil yield was also of interest in this mapping population. Of particular note was the identification of branching and seed yield (number of seeds) QTL, that were responsible for PVE of 12.115 % and 14.15 % respectively, and which were strongly correlated with final oil yield in this population. Lastly a number of oil quality (oil composition) QTL were identified, of which, a single QTL regulating oleate to linoleate conversion was identified opening up the possibility of producing high oleate designer oil in future research through investigation of this QTL.

Analysis of the phenotypic datasets showed that all had normal distributions. This indicated that the phenotypes measured were complex traits with a continuous distribution. The low skewness values for the normal distributions also demonstrated that the mapping population was unbiased and not under any distorting selective pressure. There were a number of outliers for most datasets, and some indications of segregation according to parental phenotypes in traits such as seed oil content. Plant phenotypes tended to be very consistent across the independent harvests as indicated by the high level of correlation between individual datasets, suggesting that a genetic component contributed to the variation displayed. Often there were a number of outliers both above and below the rest of the distribution. For seed oil content these were similar phenotypic values to the parents, suggesting a level of segregation in the F_2 population.

Correlation between seed oil content and seed mass suggested that seed mass was to a large extent determined by seed oil content, which is hardly surprising given that Jatropha seed is typically 30-40 % oil by weight. There was also significant correlation between the number of branches and the number of seeds produced per plant, which can also be expected given that flower inflorescences develop at branch nodes. Although highly significant (p<0.001), correlation was weak which could be due to the fact that an additional factor; female to male flower ratio at each inflorescence, also determines the number of seeds that can be

produced per plant. The four most common fatty acid moieties in Jatropha seed oil, palmitate, stearate, oleate and linoleate, which are all produced by the same metabolic pathway were nearly all negatively correlated with eachother, which supports the hypothesis that plants high in a particular fatty acid tend to have reduced amounts of the other fatty acids due to greater partitioning of the available fatty acid sink into the 'high' fatty acid at the expense of the other fatty acids. To highlight the metabolic relationship between these 4 moieties, each fatty acid is the building block of the next fatty acid in the pathway, so this is an expected result. Oil yield, calculated as the product of seed oil content (average percentage of oil per seed) x seed mass (the average mass of a seed) x seed yield (the number of seeds produced per plant), gave insight into which of these 3 traits had most impact on final oil yield. All 3 traits significantly affected final oil yield, with seed yield having the greatest influence on oil yield in the G51xCV mapping population. This could be partly explained by the greater level of variation in seed yield in this mapping population, in comparison to seed oil content or seed mass variation. This suggests that vegetative traits that affect seed yield are of utmost importance for creating high oil-yielding Jatropha. In addition both seed oil content and seed mass have significant impact on oil yield, and for material that has optimised vegetative traits but may require optimisation of seed oil content and seed mass, introgression of QTL for these traits will also be of great importance for maximising final oil yield.

The specialisation of different Jatropha material can be seen even within the G51xCV mapping population. High seed oil content and high seed mass alleles originated from the G51 parent, whereas high branching and high seed yield alleles originated from the CV parent, Chapter 5.2.2.1: p114. The G51 line contains QTL alleles that enhance seed related traits, whereas the CV parent contains QTL alleles that enhance vegetative traits. Correlation analysis of the F2 progeny show that plants high in seed yield also tend to be high in seed oil content (Chapter 5.2.1.3: p111), suggesting that these traits are synergistic and not mutually exclusive. Combining QTL alleles to optimise both vegetative growth and oil productivity into a single cultivar, would be desirable to get the best of both types of traits to optimise overall oil yield. This could also be symptomatic of a heterosis effect; hybrid material created from genetically distinct lines is well known to produce vigorous offspring in some cases. If this is the case, the generation of genetically fixed parental material from which to reliably breed heterozygous cultivars will be important, or the development of efficient propagation strategies and protocols such as the use of cuttings.

The mapping of candidate genes, particularly for well-studied metabolic pathways such as seed oil biosynthesis, has great utility in non-model species such as Jatropha. Likely candidate genes that could be responsible for observed phenotypic variation can be identified through knowledge of the biosynthetic pathway and previous examples of mutants with similar phenotypes. Immediately the candidate can be either confirmed or eliminated if they have already been mapped, by comparing its physical position to the confidence intervals for the identified QTL. Should a candidate gene lie within the confidence interval, it is simple, in theory, to clone and sequence the gene since its position and the surrounding sequence have already been identified. Unfortunately, although most of the genes in the fatty acid biosynthetic pathway were identified and mapped as part of this project, no causative mutations in these genes were present in this mapping population. However the positions of these genes have now been placed for future studies into Jatropha. Future Jatropha QTL mapping projects need only anchor their linkage map to the one developed in this study for the positions of all these genes to be of use. Or novel flanking markers could be developed using the candidate gene sequences as probes to pull out the physically linked genome sequence contig.

For the purpose of crop breeding the identification of QTL and their flanking markers are all that is required for introgression of the QTL into another cultivar. However from a conceptual perspective it would be interesting to determine the causative gene behind the QTL, and by doing so, begin to form a hypothesis on the mechanism by which they could regulate the phenotype. Of great utility for this purpose is the availability of genome sequence. Contigs harbouring a DNA marker physically linked to the QTL enables that contig to be searched for potential candidate genes. The nearest flanking marker may be some physical distance from the QTL position, therefore the greater the quality of the genome sequence and size of contig, the greater the chance of identifying the QTL sequence.

QTL confidence intervals were relatively large in this mapping population. Confidence intervals can be reduced by increasing the mapping population size which increases the number of crossovers at the genetic level, and the number corresponding phenotypic measurements to correlate them to. Confidence intervals may also be reduced by increasing the density of genetic markers to more accurately determine the position of crossovers. Marker density is partly reliant on the type of marker used and the method by which they are produced, for example genotyping by sequencing in theory should produce the highest density of markers since all sequence variation is detected down to individual SNPs (within the limitations of the sequencing technology). Genetic marker density is also restricted by the genetic diversity of the starting material used for the cross, since the more genetically similar the parental lines, the fewer sequence polymorphisms that are available for use as markers. Whilst genotyping by sequencing on its own would undoubtedly increase the density of markers on this map, using it as part of an association mapping project would have several advantages. Association mapping would use a panel of genetically diverse Jatropha from which to produce markers and populations. By starting with genetically diverse material, a greater number of genotypes and polymorphisms would be available, effectively looking at genetic crossover events and mutations accumulated over evolutionary timescales. Such an approach would also have the advantage of having a larger library of germplasm with which to start a breeding programme from after identification of desirable QTL. This combined with genotyping by sequencing markers to create a dense genetic linkage map would maximise the resolution of individual QTL and reduce the confidence interval. The smaller the confidence interval and the more precise the introgression of the QTL sequence into the recipient cultivar whilst minimising hitchhiking or unwanted physically linked sequence.

5.4: Chapter 5 Appendix

5.4.1: Phenotypic trait distributions

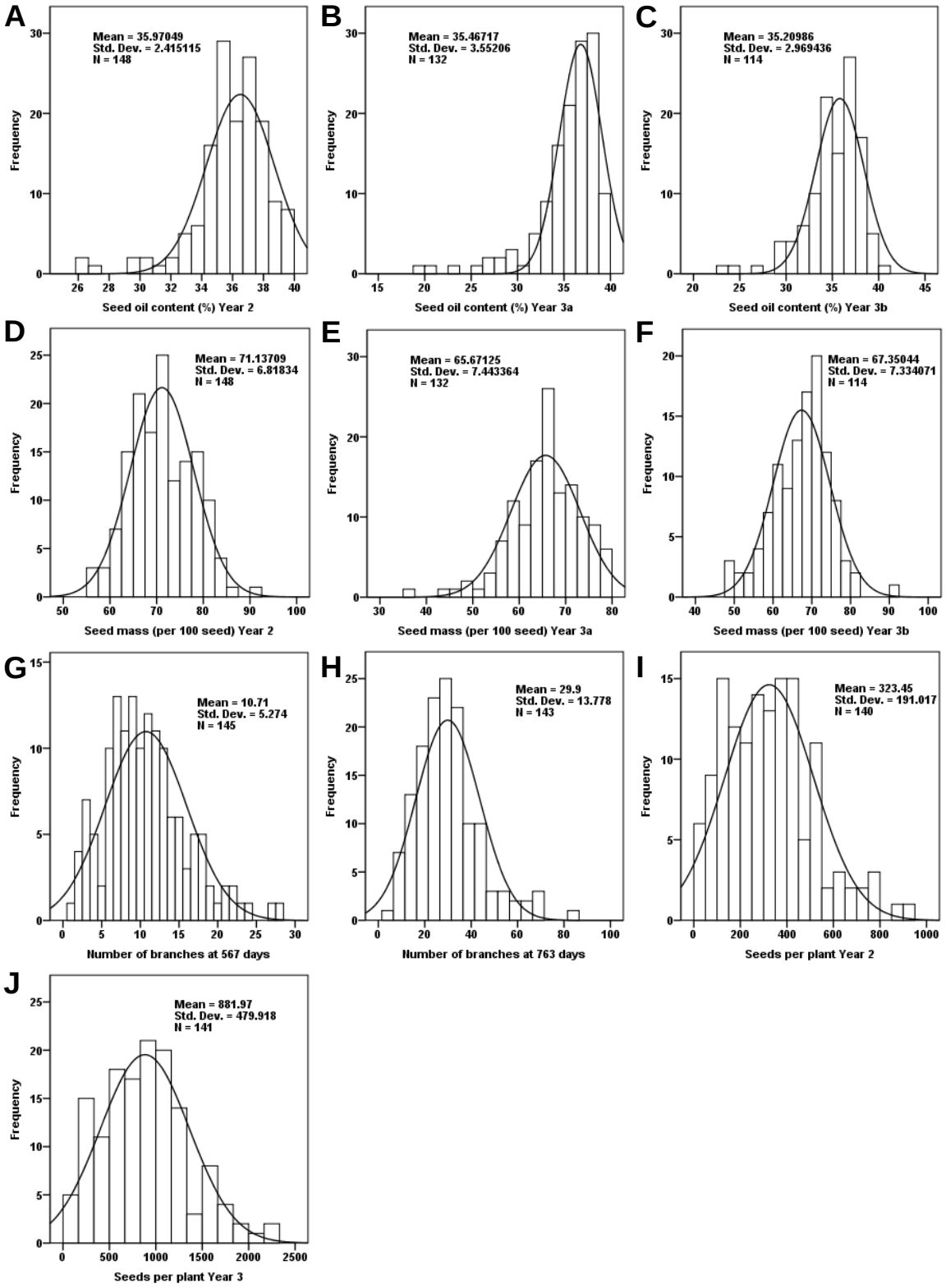


Figure 5-1. The population distribution of phenotypic traits in the G51xCV mapping population.
The distributions are plotted as frequency (number of F2 plants) on the y-axis, over the phenotypic unit of measure on the x-axis. A Normal line of best fit has been drawn to demonstrate that these traits have a normal distribution as expected from a quantitative continuous trait. Traits measured include, from A-J, seed oil content Year 2, seed oil content Year 3a, seed oil content Year 3b, seed mass year 2, seed mass year 3a, seed mass year 3b, number of branches at

567days, number of branches at 763days, seeds per plant year 2, seeds per plant year 3. Dataset mean, standard deviations and number of measurements (N) are also included.

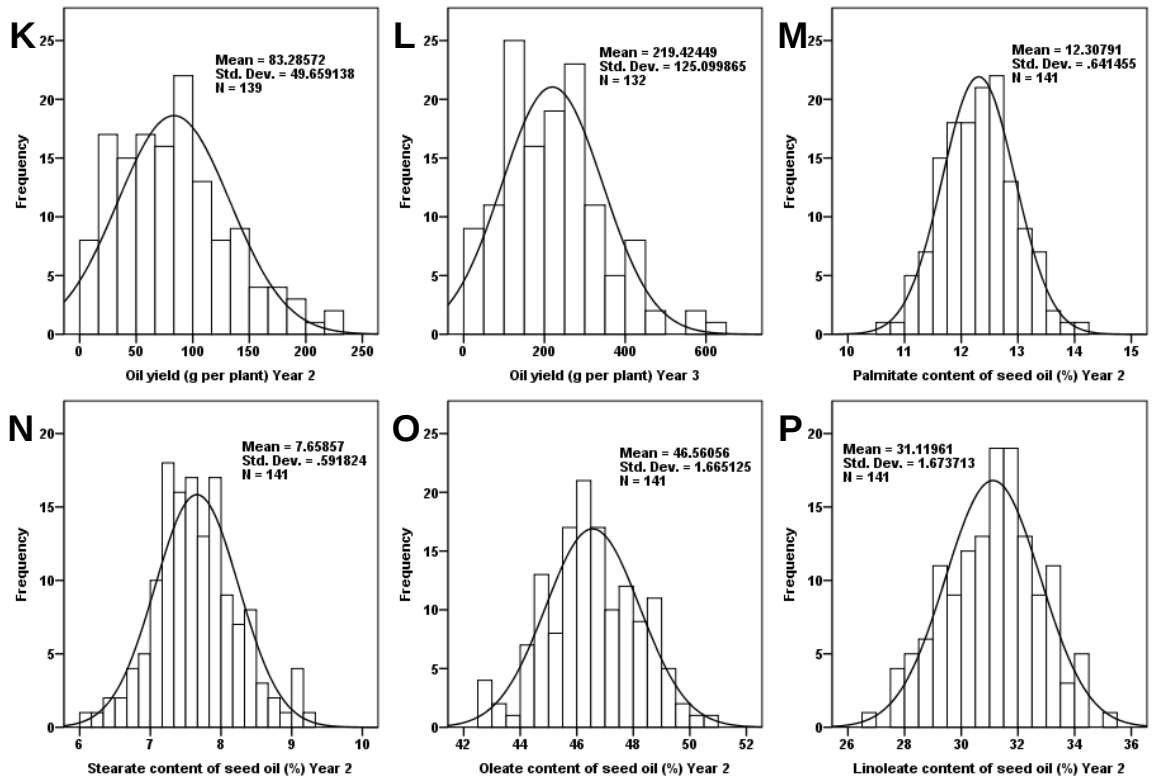


Figure 5-1. The population distribution of phenotypic traits in the G51xCV mapping population.
The distributions are plotted as frequency (number of F2 plants) on the y-axis, over the phenotypic unit of measure on the x-axis. A Normal line of best fit has been drawn to demonstrate that these traits have a normal distribution as expected from a quantitative continuous trait. Traits measured include, from K-P, oil yield year 2, oil yield year 3, palmitate content of seed oil year 2, stearate content of seed oil year 2, oleate content of seed oil year 2, linoleate content of seed oil year 2. Dataset mean, standard deviations and number of measurements (N) are also included.

Table 5-2. Phenotype dataset statistics for the G51xCV mapping population.

Statistics include N (number of F2 plants measured), Mean, Median, Std. Deviation (standard deviation), Skewness, Standard Error of Skewness, Range, Minimum value, Maximum value, 5th and 95th Percentile values. All values are correct to 2 decimal places. Trait and dataset are listed in columns from left to right (Y2= Year 2, Y3a = Year 3a, Y3b= Year 3b).

		Seed oil content (%)			100 seed mass (g)			Number of branches		Seed yield (number of seeds)		Oil yield (g/plant)		Stearate content (%)	Palmitate content (%)	Oleate content (%)	Linoleate content (%)
		Y2	Y3a	Y3b	Y2	Y3a	Y3b	@567 days	@763 days	Y2	Y3	Y2	Y3	Y2	Y2	Y2	Y2
N (F2 plants)		145	132	114	145	132	114	145	143	140	141	139	132	141	141	141	141
Mean		35.97	35.47	35.21	71.14	65.67	67.35	10.71	29.9	323.45	881.97	83.29	219.42	12.31	7.66	46.56	31.12
Median		36.16	36.3	35.61	71.07	65.98	68.68	10	28	309.5	883	81.39	215.09	12.32	7.61	46.45	31.26
Std. Deviation		2.415	3.55	2.97	6.82	7.44	7.33	5.274	13.778	191.02	479.92	49.66	125.1	0.64	0.59	1.67	1.67
Skewness		-1.55	-2.04	-1.28	0.14	-0.68	-0.27	0.696	1.075	0.64	0.54	0.67	0.73	0.12	0.22	-0.04	-0.19
Std. Error of Skewness		0.2	0.21	0.23	0.2	0.21	0.23	0.201	0.203	0.21	0.2	0.21	0.21	0.2	0.2	0.2	0.2
Range		13.76	20.94	16.7	36	42.48	41.63	27	79	894	2310	228.54	612.18	3.42	3.17	7.91	8.68
Minimum		26	19.01	23.6	55.14	37.34	48.73	1	4	10	12	2.74	12.13	10.66	6.08	42.63	26.65
Maximum		39.75	39.95	40.3	91.14	79.81	90.36	28	83	904	2322	231.28	624.31	14.08	9.24	50.53	35.33
Percentiles	**5**	30.77	27.55	29.68	60.19	53.26	52.34	3	11	53.15	195.2	14.92	40.79	11.23	6.76	43.93	28.09
	95	39.37	39.14	39.21	82.06	77.00	78.38	21	59.4	721.75	1780.5	182.77	444.11	13.41	8.71	49.11	33.87

Table 5-3. Pearson correlations of phenotypic traits in the 51xCV mapping population.

In each cell, the upper value represents the Pearson correlation coefficient (R), and the lower value the significance as a p-value, for each pairwise comparison. Significant correlations have been highlighted in dark green ($p<0.01$) and light green ($0.01<p<0.05$).

	Seed oil content Y2	Seed oil content Y3a	Seed oil content Y3b	100 seed mass Y2	100 seed mass Y3a	100 seed mass Y3b	Branches @ 567 days	Branches @ 763 days	Number of seeds Y2	Number of seeds Y3
Seed oil content Y2	1	.482 .000	.440 .000	.440 .000	.373 .000	.370 .000	.056 .510	.054 .526	-.138 .104	-.094 .271
Seed oil content Y3a	.482 .000	1	.782 .000	.431 .000	.700 .000	.575 .000	-.033 .704	.109 .212	.066 .454	.191 .028
Seed oil content Y3b	.440 .000	.782 .000	1	.322 .001	.448 .000	.454 .000	-.011 .909	.078 .414	-.075 .438	-.042 .665
100 seed mass Y2	.440 .000	.431 .000	.322 .001	1	.615 .000	.643 .000	-.006 .948	.101 .234	.014 .873	-.026 .762
100 seed mass Y3a	.373 .000	.700 .000	.448 .000	.615 .000	1	.815 .000	-.033 .711	.121 .167	.040 .655	.232 .007
100 seed mass Y3b	.370 .000	.575 .000	.454 .000	.643 .000	.815 .000	1	.141 .138	.202 .033	.124 .197	.268 .004
Branches @ 567 days	.056 .510	-.033 .704	-.011 .909	-.006 .948	-.033 .711	.141 .138	1	.731 .000	.333 .000	.357 .000
Branches @ 763 days	.054 .526	.109 .212	.078 .414	.101 .234	.121 .167	.202 .033	.731 .000	1	.312 .000	.448 .000
Number of seeds Y2	-.138 .104	.066 .454	-.075 .438	.014 .873	.040 .655	.124 .197	.333 .000	.312 .000	1	.565 .000
Number of seeds Y3	-.094 .271	.191 .028	-.042 .665	-.026 .762	.232 .007	.268 .004	.357 .000	.448 .000	.565 .000	1

Table 5-4 Pearson correlations between oil quality (oil composition) and other phenotypic traits measured in the G51xCV mapping population.

In each cell, the upper value represents the Pearson correlation coefficient (R), and the lower value the significance as a p-value, for each pairwise comparison. Significant correlations have been highlighted in dark green ($p<0.01$) and light green ($0.01<p<0.05$).

Correlations														
	Seed oil content Year 2	Seed oil content Year 3a	Seed oil content Year 3b	Seed mass Year 2	Seed mass Year 3a	Seed mass Year 3b	Branches Year 1	Branches Year 2	Seeds per plant Year 2	Seeds per plant Year 3	Palmitate content Year 2	Stearate content Year 2	Oleate content Year 2	Lineoleate content Year 2
Palmitate content Year 2	-.152	-.028	-.124	.157	.114	.070	.142	.229	.145	.238	1	-.293	-.395	.078
	.072	.754	.194	.063	.197	.464	.094	.007	.090	.005		.000	.000	.355
Stearate content Year 2	-.021	-.078	-.090	-.134	-.043	-.108	.156	.058	.115	.042	-.293	1	.267	-.494
	.808	.381	.350	.113	.632	.260	.065	.496	.180	.628	.000		.001	.000
Oleate content Year 2	-.009	-.072	-.066	-.191	-.163	-.101	-.067	-.184	-.072	-.079	-.395	.267	1	-.835
	.918	.419	.490	.023	.065	.290	.433	.031	.403	.359	.000	.001		.000
Lineoleate content Year 2	.180	.190	.208	.237	.207	.170	-.035	.076	-.015	-.010	.078	-.494	-.835	1
	.033	.031	.028	.005	.018	.075	.681	.375	.859	.911	.355	.000	.000	

Table 5-5 Pearson correlations between oil yield and other phenotypic traits measured in the the G51xCV mapping population.

In each cell, the upper value represents the Pearson correlation coefficient (R), and the lower value the significance as a p-value, for each pairwise comparison. Significant correlations have been highlighted in dark green ($p<0.01$) and light green ($0.01<p<0.05$).

Correlations																
	Seed oil content Year 2	Seed oil content Year 3a	Seed oil content Year 3b	Seed mass Year 2	Seed mass Year 3a	Seed mass Year 3b	Branches Year 1	Branches Year 2	Seeds per plant Year 2	Seeds per plant Year 3	Palmitate content Year 2	Stearate content Year 2	Oleate content Year 2	Lineoleate content Year 2	Oil yield Year 2	Oil yield Year 3
Oil yield Year 2	.050	.185	.079	.205	.185	.291	.328	.319	.972	.541	.151	.086	-.103	.040	1	.522
	.559	.035	.410	.015	.035	.002	.000	.000	.000	.000	.078	.319	.232	.639		.000
Oil yield Year 3	.100	.402	.115	.233	.470	.489	.343	.474	.477	.948	.274	-.005	-.091	.050	.522	1
	.255	.000	.236	.007	.000	.000	.000	.000	.000	.000	.002	.953	.307	.576	.000	

Table 5-6. Summary statistics for QTL identified by interval mapping in the G51xCV mapping population.

Trait	Dataset	Linkage group	Position (cM)	LOD	p-value (p<x)	PVE (%)	Bayes 95 % CI (cM)	-1 LOD interval (cM)	-2 LOD interval (cM)	Beneficial (high) allele	Effect
Seed oil content	Year 2	4	33	4.624	0.0027	16.56	2.0-56.0	7.0-36.0	6.0-39.0	G51	Dominant
		10	31	4.414	0.0044	15.75	8.0-40.0	23.0-34.0	3.0-47.0	G51	Dominant
	Year 3a	4	46	3.527	0.0291	13.42	0.0-57.0	43.0-55.0	39.0-57.0	G51	Dominant
	Year 3b	10	27	3.503	0.0322	15.95	22.0-46.0	23.0-49.0	20.0-49.0	G51	Dominant
Seed mass	Year 2	4	11	7.776	0.0000	29.39	0.0-53.0	2.0-12.0	0.0-22.0	G51	Dominant
	Year 3a	4	10	4.964	0.0011	19.39	0.0-51.0	0.0-19.0	0.0-23.0	G51	Dominant
	Year 3b	4	3	3.285	0.0507	14.89	2.0-53.0	0.0-12.0	0.0-38.0	G51	Dominant
Seed oil composition	Palmitate content Year 2	5	32	7.929	0.0000	30.54	28.0-41.0	27.0-36.0	25.0-37.0	CV	Semi-dominant
		7	58	3.24	0.0548	11.51	26.0-70.0	44.0-75.0	25.0-75.0	CV	Recessive
		10	33	2.917	0.1056	9.34	0.0-38.0	30.0-38.0	23.0-46.0	G51/CV	Heterosis
	Stearate content Year 2	1	35	4.144	0.0010	14.95	4.0 - 41.0	28.0-38.0	22.0-43.0	G51	Dominant
		4	27	4.226	0.0079	15.26	7.0-53.0	23.0-34.0	21.0-47.0	CV	Semi-dominant
		7	22	6.606	0.0000	24.86	1.0-36.0	19.0-31.0	12.0-41.0	G51	Semi-dominant
	Oleate content Year 2	6	2	3.398	0.0387	12.10	0.0-19.0	0.0-13.0	0.0-17.0	CV	Recessive
	Linoleate content Year 2	4	24	3.995	0.0117	14.37	1.0-49.0	17.0-39.0	0.0-41.0	G51	Dominant
		6	4	4.307	0.0055	15.58	0.0-15.0	0.0-14.0	0.0-17.0	G51	Dominant
		8	11	3.287	0.0505	11.68	1.0-48.0	1.0-14.0	0.0-30.0	CV	Dominant
Number of Branches	@ 763 days	1	25	3.477	0.0024	12.115	0.0-26.0	17.0-30.0	0.0-32.0	CV	Dominant
Seed yield (number of seeds)	Year 3	10	33	3.966	0.0104	14.15	0.0-33.0	0.0-39.0	0.0-49.0	CV	Dominant

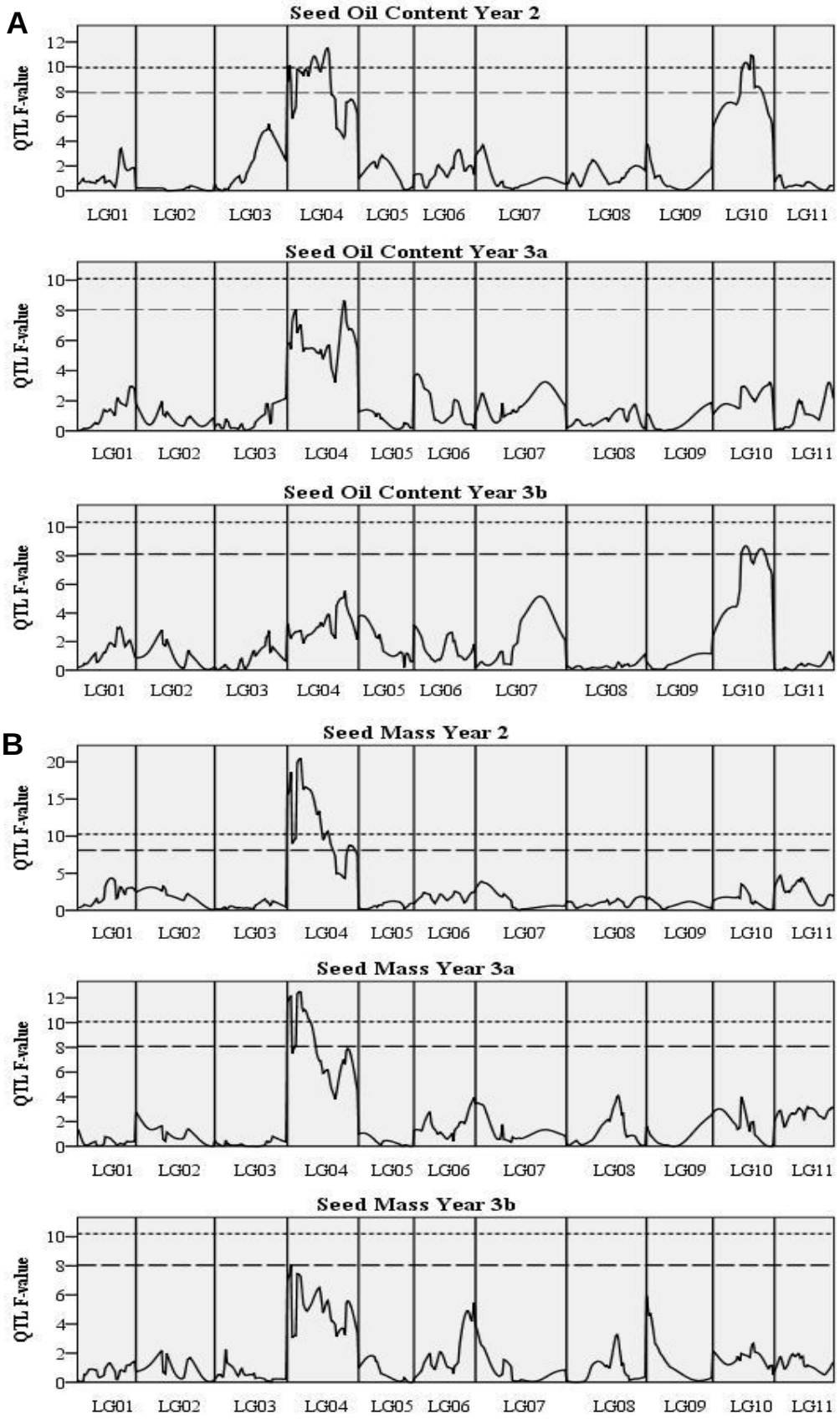


Figure 5-2. The output of a QTL analysis using GridQTL software.
The level of QTL association, as determined by Haley and Knott interval mapping, is indicated as the F statistic (y-axis). Linkage groups 1-11 (x-axis) are separated by vertical lines. Horizontal lines represent experiment wide significance thresholds (long dash, p= 0.05, short dash, p= 0.01) calculated from bootstrap analysis using 10,000 iterations. Phenotypic traits showing significant QTL association include: Seed oil content, A; Seed mass, B; Seed oil composition, C; Number of branches, D; Seed yield (number of seeds), E.

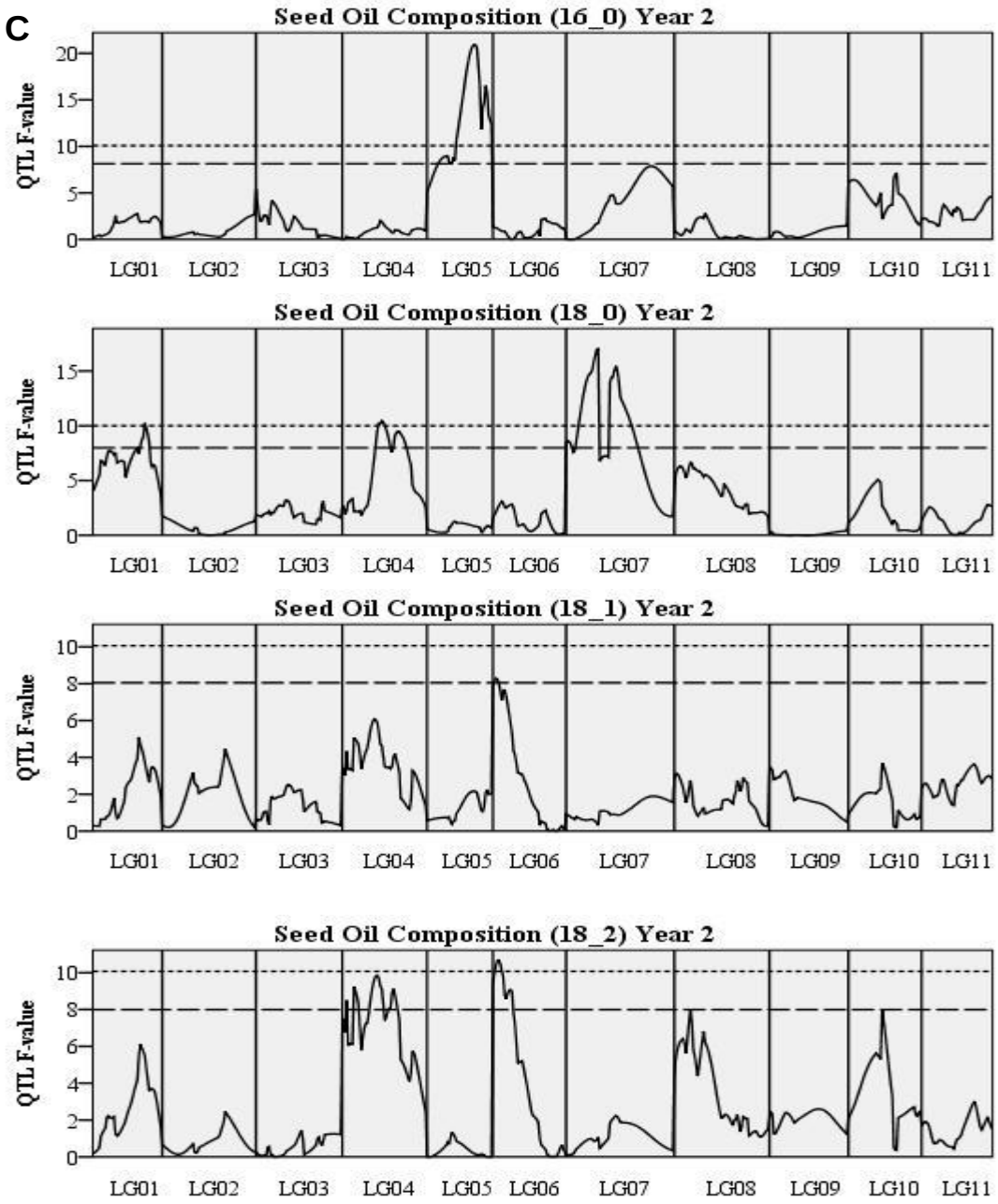


Figure 5-2. The output of a QTL analysis using GridQTL software.

The level of QTL association, as determined by Haley and Knott interval mapping, is indicated as the F statistic (y-axis). Linkage groups 1-11 (x-axis) are separated by vertical lines. Horizontal lines represent experiment wide significance thresholds (long dash, p= 0.05, short dash, p= 0.01) calculated from bootstrap analysis using 10,000 iterations. Phenotypic traits showing significant QTL association include: Seed oil content, A; Seed mass, B; Seed oil composition, C; Number of branches, D; Seed yield (number of seeds), E.

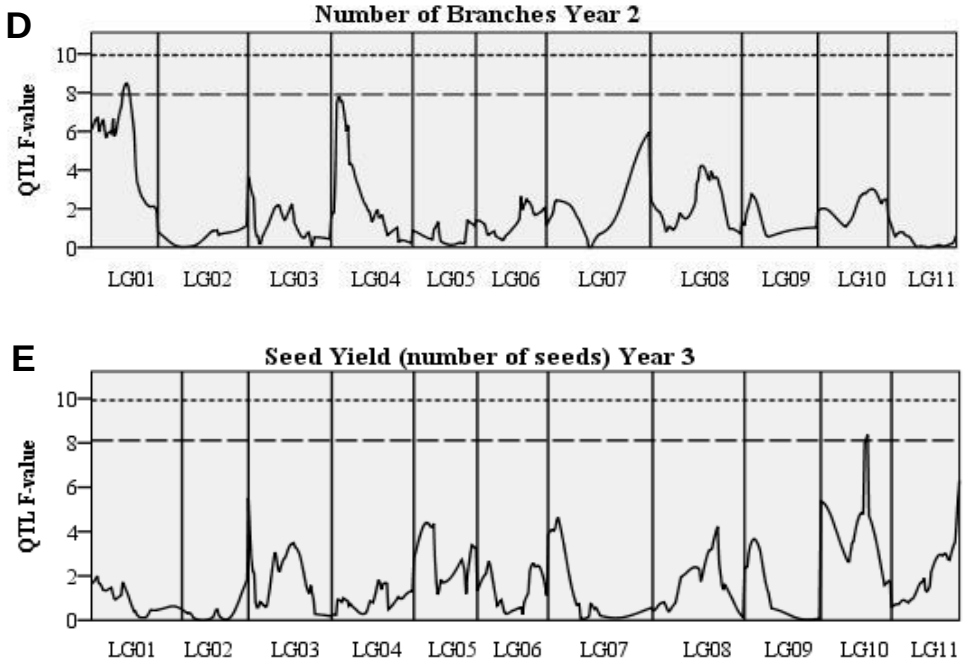


Figure 5-2. The output of a QTL analysis using GridQTL software

The level of QTL association, as determined by Haley and Knott interval mapping, is indicated as the F statistic (y-axis). Linkage groups 1-11 (x-axis) are separated by vertical lines. Horizontal lines represent experiment wide significance thresholds (long dash, p= 0.05, short dash, p= 0.01) calculated from bootstrap analysis using 10,000 iterations. Phenotypic traits showing significant QTL association include: Seed oil content, A; Seed mass, B; Seed oil composition, C; Number of branches, D; Seed yield (number of seeds), E.

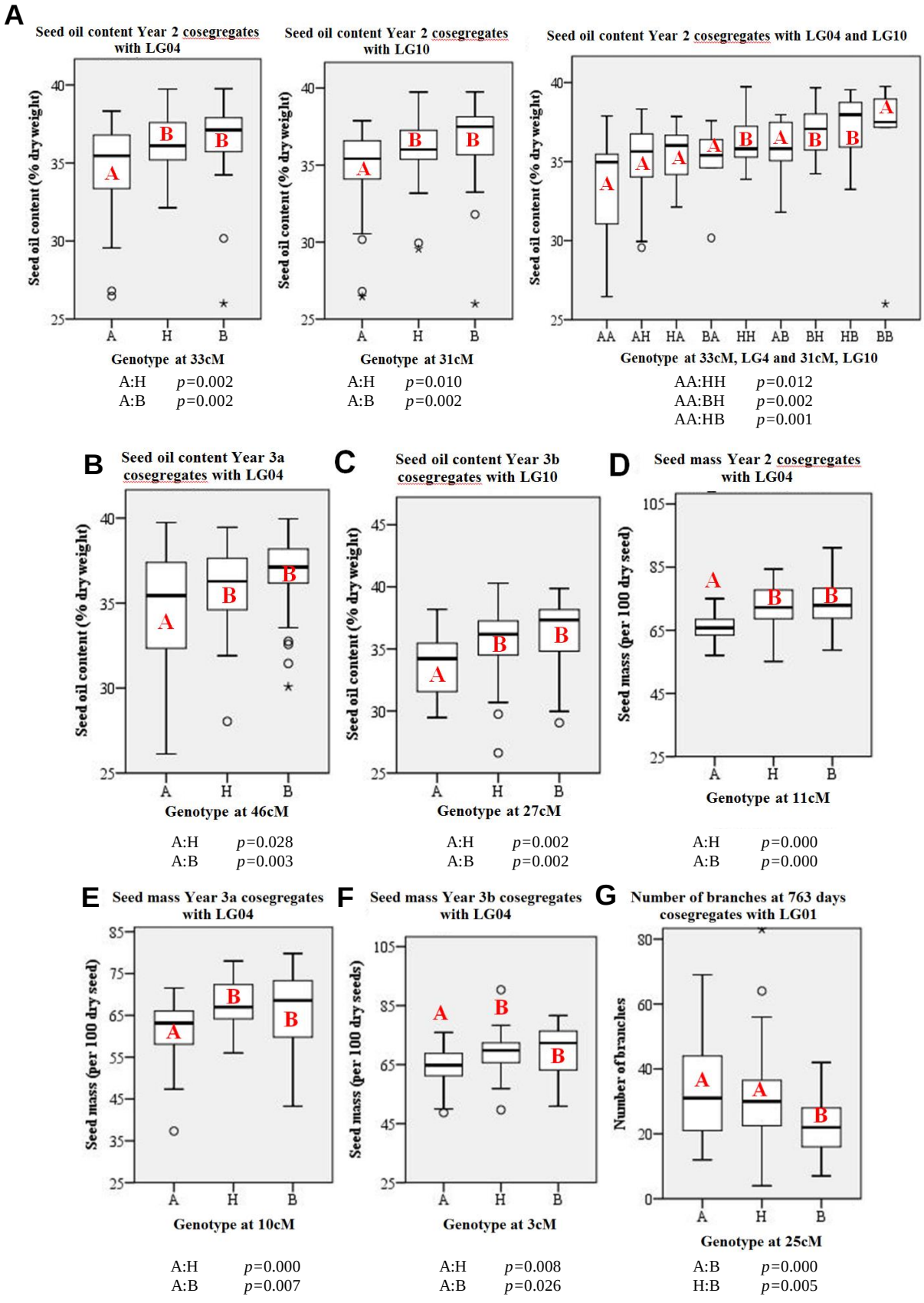


Figure 5-3. Boxplot showing correlation between phenotype and genotype at identified Quantitative Trait Loci in the G51xCV mapping population.

The whiskers represent the dataset range, with outliers shown as circles ($p<0.05$) or stars ($p<0.01$). The box edges (upper and lower) represent the interquartile range. The median value is indicated by the thick line within the box. Statistically different groups, as determined by a Tukey's post hoc comparison of means test, have been labelled in red, with the *p* value for each comparison indicated below the box and whisker plot. Phenotypic traits and datasets include: Seed oil content; Year 2 (A), Year 3a (B), Year 3b (C), Seed mass; Year 2 (D), Year 3a (E), Year 3b (F), Number of branches at 763days (G).

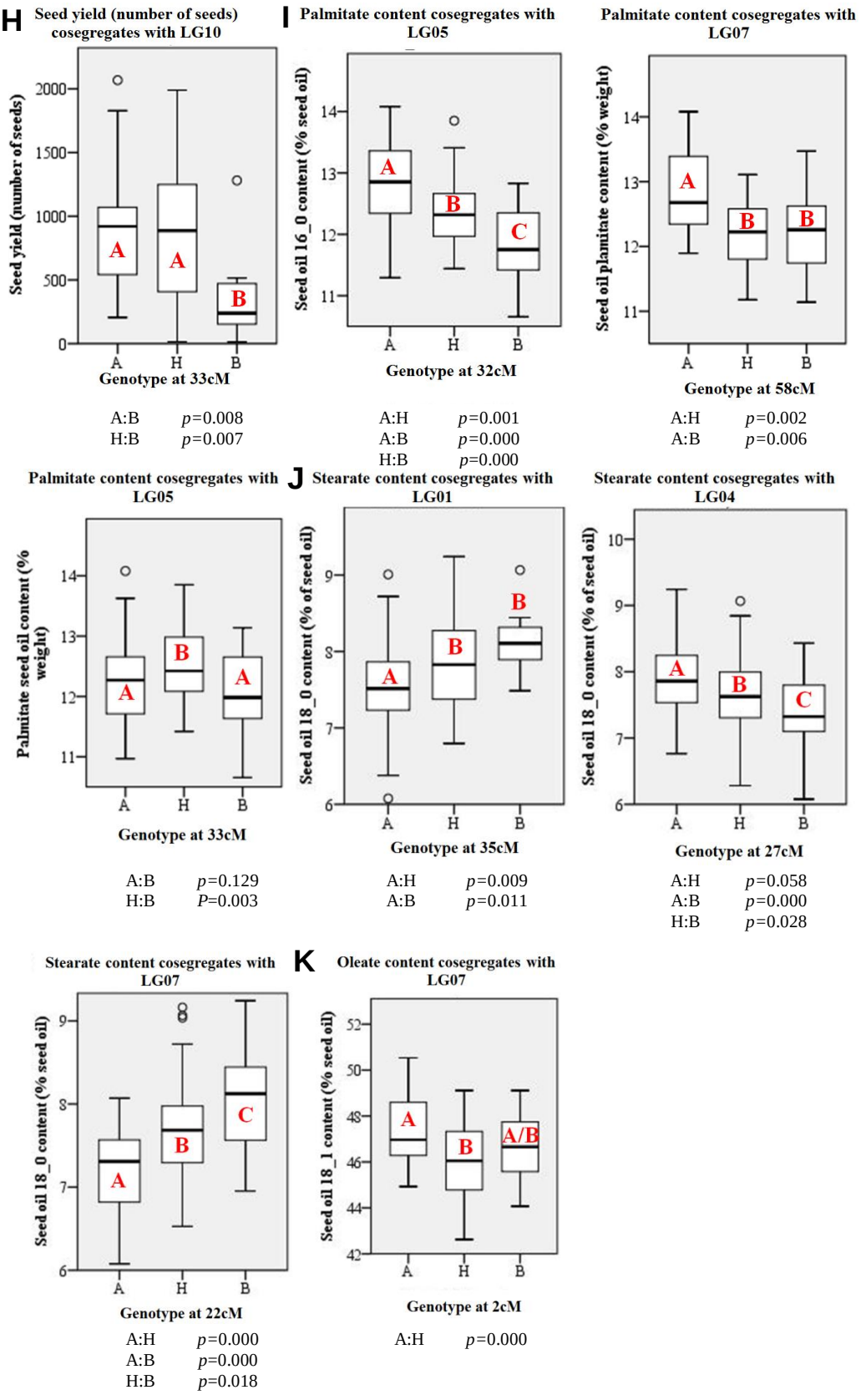
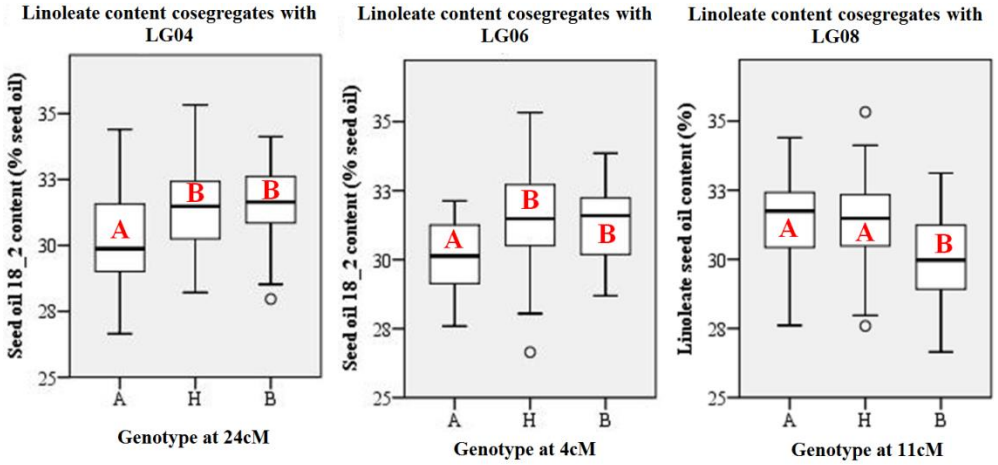


Figure 5-4 Boxplot showing correlation between phenotype and genotype at identified Quantitative Trait Loci in the G51xCV mapping population

The whiskers represent the dataset range, with outliers shown as circles (p<0.05) or stars (p<0.01). The box edges (upper and lower) represent the interquartile range. The median value is indicated by the thick line within the box. Statistically different groups, as determined by a Tukey's post hoc comparison of means test, have been labelled in red, with the p value for each comparison indicated below the box and whisker plot. Phenotypic traits and datasets include: Seed Yield (number of seeds), Year 3 (H); Palmitate seed oil content, Year 2 (I); Stearate seed oil content, Year 2 (J); Oleate seed oil content, Year 2 (K).

L



A:H	p=0.000	A:H	p=0.000	A:B	p=0.004
A:B	p=0.000	A:B	p=0.007	H:B	p=0.001

Figure 5-5 Boxplot showing correlation between phenotype and genotype at identified Quantitative Trait Loci in the G51xCV mapping population

The whiskers represent the dataset range, with outliers shown as circles (p<0.05) or stars (p<0.01). The box edges (upper and lower) represent the interquartile range. The median value is indicated by the thick line within the box. Statistically different groups, as determined by a Tukey's post hoc comparison of means test, have been labelled in red, with the p value for each comparison indicated below the box and whisker plot. Phenotypic traits and datasets include: Linoleate seed oil content, Year 2 (L).

Chapter 6: Summary and conclusions

Population growth, economic development and climate change necessitate an increase in world energy production with concurrent reductions in greenhouse gas emissions (Intergovernmental Panel on Climate Change, 2014, US Energy Information Administration, 2016). Plant based biofuels, offer the only renewable, low-carbon alternative to liquid transportation fuels (Blanch, 2010); the single largest sector of the most widely-used fossil fuel; oil (British Petroleum, 2016). In the midst of food security concerns due to population growth and predicted effects of climate change on food crop production (Godfray et al., 2010), de-confliction of biofuel and food crops particularly for agricultural land, suggests a greater utilisation of marginal land (Tilman et al., 2009), and the harnessing of novel crop species more adaptable to alternative models of farming (Tester and Langridge, 2010). For bioethanol and biodiesel production, perennial species are generating much interest as biomass and oilseed crops respectively, as they are more amenable to growth on marginal land, and could be intrinsically more-efficient at using nutrients, water and sequestering carbon, than currently-cultivated annual crops (Somerville et al., 2010, Kantar et al., 2016).

Jatropha curcas, a perennial oilseed crop from the Euphorbiacea, is a biodiesel candidate that has generated interest due to a high seed oil content, a protein rich seed meal suitable for use as an animal feed, and an adaptability to a wide range of soil types, nutrient and precipitation levels (King et al., 2009, Achten et al., 2010). Before economic cultivation of Jatropha can be assessed, optimisation of oil yield and oil quality - related traits is required in order to domesticate current semi-wild/wild material and create a genetically-improved cultivar.

Jatropha presents several challenges to selective breeding. Jatropha is monoecious and self-compatible, which leads to a propensity for self-fertilisation. Jatropha is a long-life perennial and has a life span of 50 years; Jatropha plants reach maturity and express full phenotypes after approximately 5 years from seedling. Seedling-to-seed generation time is approximately 9 months. Material distributed outside its centre of origin in Meso-America contains very little genetic variation and is almost clonal (King et al., 2015, Montes Osorio et al., 2014, Pecina-Quintero et al., 2014, He et al., 2011). These point towards the requirement for Quantitative Trait Locus (QTL) mapping, so that breeding technologies such as Marker Assisted Selection (MAS) can guide selection of breeding material based on the genetics underpinning key traits. Ultimately the stacking of multiple, desirable QTL in a single Jatropha line will be required to create a cultivar suitable for economic cultivation. The aim of this study was to identify and locate QTL underlying oil yield- and oil quality-related traits (seed oil content, seed oil composition, seed mass, seed yield, branching), in the G51xCV mapping population; a biparental F_2 population created from parental lines selected primarily on the basis of seed oil content; G51 at 36.90 % and CV at 26.00 %, seed oil content respectively.

Reduced-representation genome sequencing, using the Complexity Reduction of Polymorphic Sequences (CRoPS) approach (van Orsouw et al., 2007), was used to generate a high-coverage, genome-wide marker set and this formed the majority of markers used for genetic linkage mapping in the G51xCV population (181 markers, 58.01 %). In addition a small number of expressed sequence tagged (EST) markers were available to map expressed genes (14 markers, 4.49 %) (King et al., 2011). The development of SSR markers formed the principle output of marker development in this thesis study, and accounted for the second largest group of markers available for linkage mapping in the G51xCV population (117 markers, 37.50 %).

In total, over 300 SSR positions were identified using reference genome sequence (Hirakawa et al., 2012, Sato et al., 2011), of which 288 SSRs had flanking sequence suitable for PCR amplification. Polymorphism testing across parental lines from 4 independent mapping populations, showed that of the 288 SSRs tested, 114 SSR markers (39.59 %) were polymorphic in at least 1 mapping population, and 43 SSR markers (14.93 %) were polymorphic in 2 or more populations. Markers that were mapped in multiple populations (shared markers), increased recombination data available for these loci in a combined dataset, and also enabled alignment and comparison of individual population linkage maps, providing both a validation method for individual population datasets, and a means to conduct comparative mapping to mine additional SSRs in regions that required them. Marker ordering, marker spacing and total genetic distance mapped was highly conserved across all 4 independent linkage maps, indicating the robustness of the markers and datasets, from all 4 populations. This robust approach led to a substantial increase in the accuracy of estimated genetic map size, over a previously published interspecific linkage map (Wang et al., 2011, King et al., 2013). SSR markers developed in this study, and the dataset produced from the genotyping and linkage mapping of 229 plants from the G51xCV mapping population, contributed data towards the first intraspecific linkage map published for *Jatropha curcas* in 2013 (King et al., 2013).

SSR marker development was carried out for several functions. As can be seen from this project, SSR markers are ideally suited for the mapping of smaller numbers of loci in a more selective and targeted manner, in comparison to more highly-parallel, high-throughput sequencing-based strategies used for the other marker types in this study. As such the SSR markers developed in this study added value to the existing genome-wide marker set, by enhancing both the coverage and information content of mapped loci. For example, 120 SSR markers were developed to map gaps in the linkage map using comparative mapping, after genetic linkage mapping with the genome-wide marker set, thereby reducing the number of regions where marker spacing was greater than the recommended 10 cM required for complete QTL detection (Darvasi et al., 1993). Information content of the genetic linkage map was significantly enhanced through the mapping of candidate genes.

In total, 133 SSR markers (42.63 %) developed from this thesis study, were physically linked to candidate genes of the fatty acid biosynthetic pathway; a pathway responsible for regulating oil yield and oil quality related traits (seed oil content and fatty acid composition), as well as a smaller number of SSR markers (35 SSRs or 11.22 %) that were linked to branching and flower ratio candidate genes. The utility of such genes for plant biotechnology, and their central role in regulating both oil yield and oil quality (seed oil composition) (Napier et al., 2014, Bates et al., 2013, Vega-Sanchez and Ronald, 2010, Durrett et al., 2008), as has been extensively demonstrated in research and agriculture(Napier et al., 2014, Bates et al., 2013, Vega-Sanchez and Ronald, 2010, Durrett et al., 2008), highlights the importance of these genes in an oilseed crop such as Jatropha. *In silico* identification of candidate genes through the compilation of genes from model species such as Arabidopsis (Li-Beisson et al., 2013), and comparative genomic approaches utilising search algorithms, enables a comprehensive approach to mapping of this pathway; ensuring as much of the genetic architecture of complex traits such as oil yield, and its component trait, seed oil content, are mapped. Since alignment of genetic linkage maps is relatively easy using anchor and bridging markers, as has been demonstrated in this study, the positioning of such key genes, is of on-going utility to genetic research in *Jatropha curcas*, for example for the investigation of QTL associated with oil yield or oil quality related traits.

Integration of additional SSRs mined for gap filling and the marking of candidate genes, in the G51xCV mapping population during this thesis study, enabled the updating of the combined genetic linkage map published in 2013 (King et al., 2015, King et al., 2013); modestly but significantly increasing marker density and coverage, and substantially increasing information content of the linkage map through the addition of candidate gene markers (King et al., 2015). This led to dissemination of a refined intraspecific linkage map for *Jatropha curcas* in 2015 (King et al., 2015), that was validated through subsequent genome sequencing efforts from the Chinese Academy of Sciences (Wu et al., 2015), that together enabled the physical mapping of 51 % of the Jatropha genome sequence, and 64 % of protein encoding sequences (King et al., 2015); a key genomic resource for the investigation of QTL in *Jatropha curcas*.

Collection of seed-related phenotypic data using Nuclear Magnetic Resonance spectroscopy and FAMES gas chromatography enabled QTL mapping of traits from up to 3 separate sample points (years 2, and years 3a and 3b) in the G51xCV mapping population.

A combination of single marker analysis and interval mapping, combined with correlation analysis, enabled the investigation of phenotypic traits in the G51xCV mapping population. A hierarchical trait relationship (Alonso-Blanco and Mendez-Vigo, 2014) could be established for oil yield per plant, based on the product of seed oil content, seed mass and seed yield traits. It was therefore of interest to determine the relative importance of these component traits to overall oil yield. Similarly, positive and negative correlations between traits gave an indicator of the level of independence or causality between traits. Key findings of correlation analysis in G51xCV population was that seed yield, seed oil content and seed mass were all positively correlated with eachother, suggesting that increases in any one of these traits is not at the expense of the others, and that optimisation of all three of these key traits is compatible within a single cultivar. Of these three traits, seed yield was most strongly correlated with oil yield per plant, showing that seed yield was most important for regulating final oil yield in the G51xCV mapping population.

Negative correlations, for example between oleic and linoleic acid, suggested causality between these traits; a suggestion supported by the fact that both fatty acids exist in the same metabolic pathway. This finding is consistent with the concept of limited fatty acid pools, or limited rates of synthesis, such that conversion of one fatty acid into the other, leads to an increase in one fatty acid at the expense of the other, ultimately producing differing fatty acid ratios. The fact that QTLs for both these fatty acid moieties co-located to a single region, suggests that this conversion is controlled by a single locus and perhaps a single gene. This is supported by the known pathway for this conversion, which is the result of the single step conversion by the fatty acid desaturase gene, FAD2, in the oilseed model species, *Arabidopsis thaliana* (Li-Beisson et al., 2013). The fact that a high oleate fatty acid composition is *the* key attribute for developing biodiesels that meet international fuel standards (Durrett et al., 2008, Knothe, 2009), suggests that elucidation of this locus is of high value for optimising oil quality of *Jatropha curcas*. This has been proven recently in *J. curcas* using gene silencing of the FAD2 gene, to produce high oleate Jatropha oil (Utomo et al., 2015, Ye et al., 2009).

QTL were identified in all component traits of oil yield per plant and oil quality; seed oil content, seed mass and seed yield (including the component trait of seed yield; branching), and through QTL mapping of the major fatty acids that make up *Jatropha curcas* seed fatty acid composition. In total, 15 QTL were detected for seed oil content (2QTL), seed oil composition; palmitate, stearate, oleate, linoleate content (10 QTL), seed mass (1 QTL), number of branches (1 QTL) and seed yield (1QTL). Strongest QTL effects were

detected for seed oil content, the principle trait of interest in the G51xCV population. Combined PVE for seed oil content in year 2, accounted for 32.26 % of observed variation. Similarly large effect QTL were detected for the other component traits of oil yield per plant; seed yield, 14.15 % PVE and seed mass, between 14.89 % and 29.39 % PVE. Multiple QTL located to linkage groups 4 and 10; suggesting regions on these linkage groups are important for regulating a variety of traits (or that some QTL could be pleiotropic), and are therefore key target regions for introgression into a cultivar. Analysis of the parent-of-origin for QTL alleles, showed that beneficial vegetative/architecture QTL alleles (branching, seed yield traits) originated from the CV parent, whereas beneficial alleles for seed related QTL (seed oil content, seed mass traits) originated from the G51 parent. This suggests that breeding of hybrid plants may be an advantageous strategy to combine favourable components of both seed and vegetative/architecture related traits in a single cultivar. Since the CV parent is genetically similar to widely distributed Jatropha material, the seed related trait QTL as found in G51, may be more beneficial for introgression to improve widely distributed material. That being said, in G51xCV, seed yield was ultimately the most important trait for regulating oil yield, and so, alongside seed oil content and seed mass traits, generating Jatropha cultivars with high seed yields will be of utmost importance for creating high yielding varieties.

6.1: Future recommendations

The major limitations of this study were the confidence intervals for identified QTL which were relatively large; ranging from 13 cM (palmitate year 2) to 57 cM (seed oil content year 3a) for the Bayes 95 % confidence interval. For crop breeding delimiting QTLs to the smallest interval possible is advantageous so that QTL can be introgressed and/or stacked (combined) with greatest precision (Dekkers and Hospital, 2002). It also enables desirable alleles to be found across diverse germplasm collections with greatest accuracy, as the genotype for flanking markers of QTL, are more likely to be informative of the underlying genetic alleles, the closer the markers are together.

The resolution of QTL location is dependent on the number of recombination events and independent phenotypic measurements with which to correlate them to, and the density of DNA markers to most accurately locate recombination points (Mackay et al., 2009). During a single meiosis event, between zero and two crossover events per linkage group, were most common according to linkage mapping data used in this study (data not shown), which is consistent with findings in Arabidopsis despite differences in chromosome length between the two species (Giraut et al., 2011). Therefore in a Jatropha F_2 population, which is the product of recombination in two F_1 gametes as it is a diploid species, 0 to 2 times the F_2 population size per gamete, can be used as an estimation of the number of recombination events sampled per linkage group (i.e. between 0 and 4 recombinations, per linkage group, per F_2 plant). One can increase recombination data, by using shared markers that are mapped in independent populations, as utilised by linkage mapping in the combined genetic linkage map, however unless independent populations phenotype the same traits, and experience identical environmental conditions, there may not be corresponding phenotypic measurements to correlate these recombination events to, for QTL mapping of particular traits.

Another way to increase the number of recombinations sampled *per plant*, is to conduct association mapping using a diverse germplasm panel (Hamblin et al., 2011, Davey et al., 2011, Rafalski, 2010). By doing so recombination events are sampled over the many generations since divergence of the germplasm panel from a common ancestor, substantially increasing the resolution of QTL mapping (Hamblin et al., 2011). This,

combined with genome-wide sequencing of the germplasm collection (Davey et al., 2011), produces a maximal marker density - depending on the depth of sequencing and the sequencing approach used (i.e. complete vs reduced representation) (Davey et al., 2011), which determines the completeness of the genome sequences that can be compared for marker generation. Surveying a diverse germplasm collection also enables investigation of a wider selection of genotypes and phenotypes, potentially capturing more advantageous phenotypic and genetic variation that can be investigated (Hamblin et al., 2011). This germplasm collection can then form the basis of a breeding programme, once relevant QTL have been discovered. Whilst increasing population size, recombination events per plant, and marker density are well-known strategies to increase the power to detect and locate QTL (Mackay et al., 2009, Davey et al., 2011), the importance of accurate and suitable phenotyping to increase QTL detection, particularly for complex traits such as oil yield, is also important (Alonso-Blanco and Mendez-Vigo, 2014).

Complex traits such as oil yield are known to be the result of the interaction of a myriad of loci, affecting different areas of plant development and metabolism, each potentially contributing minor effects that are subject to modulation by the environment (Holland, 2007, Alonso-Blanco and Mendez-Vigo, 2014). As such QTL mapping of a complex trait such as oil yield, limits the power to detect minor effect QTL that nevertheless, in aggregate, may be important for regulating oil yield (Alonso-Blanco and Mendez-Vigo, 2014). One method to increase the power to detect QTL is to determine hierarchical trait relationships (Alonso-Blanco and Mendez-Vigo, 2014); splitting oil yield into component traits, such as oil content, seed mass and seed yield – as conducted in this study. Whilst QTL may have only minor effects at the oil yield level, the effects on component traits that QTL may more-directly regulate would be bigger, effectively increasing the power to detect these QTL. This can be seen by seed oil composition QTL detected in this study. Whilst variation of only a few percent seed oil composition were observed for some fatty acid moieties, strong QTL associations were able to be detected, due to precision phenotyping of individual fatty acids using gas chromatography.

Oil yield per plant in this study, was broken down into component traits, seed oil content, seed mass and seed yield. Whilst seed mass could have been broken down further, for example into seed components, such as proteins, carbohydrates and fats, or different seed tissues, to give further information on seed mass variation, this, whilst interesting from a biological perspective, would have less significance for overall oil yield and the breeding of improved Jatropha varieties. Seed yield however, could have significantly more scope for phenotypic dissection.

In this study branching was one trait that was measured as a component of seed yield; which itself was the most strongly associated with oil yield per plant in G51xCV (in years 2 and 3, seed yield and oil yield showed a correlation of R=0.972, and R=0.948 respectively). Although branching was significantly associated with seed yield variation ($p<0.000$), showing that this is an important contributor to seed yield, the fact that correlation was incomplete (between R=0.312 and R=0.448) suggested that additional factors were present. A key trait that would be expected to be correlated with branching and seed yield, would be flower ratio. Flower inflorescences, which produce seed through the female flowers, are highly dependent on branching, since flower inflorescences occur at positions along branches. Whilst branching may determine the number of possible flower points on a plant, the actual extent of flowering and the ratio of female to male flowers, may well also contribute towards seed yield and so this would be a key trait to measure in future studies along with branching. Other vegetative traits were measured in the G51xCV mapping population as

part of a wider study outside of this thesis work (King et al., 2015). Plant height and stem diameter were also shown to be significant to seed yield (King et al., 2015), demonstrating the complex genetic architecture likely to underlie seed yield in Jatropha.

As such, seed yield, could perhaps be the trait that could most benefit from future genome wide association studies, including precision and hierarchical trait phenotyping, to dissect this complex trait and determine the genetic basis of observed variation through the study of diverse germplasm collections. This will provide the basis to add to QTL discovered in this study, whilst expanding the collection of suitable starting material for the breeding of improved cultivars of *Jatropha curcas*.

Chapter 7: List of References

ABHILASH, P. C., SRIVASTAVA, P., JAMIL, S. & SINGH, N. 2011. Revisited Jatropha curcas as an oil plant of multiple benefits: critical research needs and prospects for the future. *Environ Sci Pollut Res Int,* 18**,** 127-31.

ABOU KHEIRA, A. A. & ATTA, N. M. M. 2009. Response of Jatropha curcas L. to water deficits: Yield, water use efficiency and oilseed characteristics. *Biomass and Bioenergy,* 33**,** 1343-1350.

ACHTEN, W. M. J., MATHIJS, E., VERCHOT, L., SINGH, V. P., AERTS, R. & MUYS, B. 2007. Jatropha biodiesel fueling sustainability? *Biofuels, Bioproducts and Biorefining,* 1**,** 283-291.

ACHTEN, W. M. J., NIELSEN, L. R., AERTS, R., LENGKEEK, A. G., KJÆR, E. D., TRABUCCO, A., HANSEN, J. K., MAES, W. H., GRAUDAL, L., AKINNIFESI, F. K. & MUYS, B. 2010. Towards domestication ofJatropha curcas. *Biofuels,* 1**,** 91-107.

ADAM, H., JOUANNIC, S., ORIEUX, Y., MORCILLO, F., RICHAUD, F., DUVAL, Y. & TREGEAR, J. W. 2007. Functional characterization of MADS box genes involved in the determination of oil palm flower structure. *J Exp Bot,* 58**,** 1245-59.

AGARWAL, M., SHRIVASTAVA, N. & PADH, H. 2008. Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep,* 27**,** 617-31.

ALBERS, S. C., BERKLUND, A. M. & GRAFF, G. D. 2016. The rise and fall of innovation in biofuels. *Nat Biotechnol,* 34**,** 814-21.

ALONSO-BLANCO, C. & MENDEZ-VIGO, B. 2014. Genetic architecture of naturally occurring quantitative traits in plants: an updated synthesis. *Curr Opin Plant Biol,* 18**,** 37-43.

ANDRIANOV, V., BORISJUK, N., POGREBNYAK, N., BRINKER, A., DIXON, J., SPITSIN, S., FLYNN, J., MATYSZCZUK, P., ANDRYSZAK, K., LAURELLI, M., GOLOVKIN, M. & KOPROWSKI, H. 2010. Tobacco as a production platform for biofuel: overexpression of Arabidopsis DGAT and LEC2 genes increases accumulation and shifts the composition of lipids in green biomass. *Plant Biotechnol J,* 8**,** 277-87.

AREGHEORE, E. M., BECKER, K. & MAKKAR, H. P. S. 2003. Detoxification of a toxic variety of Jatropha curcas using heat and chemical treatments, and preliminary nutritional evaluation with rats. *The South Pacific Journal of Natural and Applied Sciences,* 21**,** 51-56.

ATABANI, A. E., SILITONGA, A. S., ONG, H. C., MAHLIA, T. M. I., MASJUKI, H. H., BADRUDDIN, I. A. & FAYAZ, H. 2013. Non-edible vegetable oils: A critical evaluation of oil extraction, fatty acid compositions, biodiesel production, characteristics, engine performance and emissions production. *Renewable & Sustainable Energy Reviews,* 18**,** 211-245.

BALAT, M. 2011. Potential alternatives to edible oils for biodiesel production - A review of current work. *Energy Conversion and Management,* 52**,** 1479-1492.

BARAZESH, S. & MCSTEEN, P. 2008. Hormonal control of grass inflorescence development. *Trends Plant Sci,* 13**,** 656-62.

BASHA, S. D., FRANCIS, G., MAKKAR, H. P. S., BECKER, K. & SUJATHA, M. 2009. A comparative study of biochemical traits and molecular markers for assessment of genetic relationships between Jatropha curcas L. germplasm from different countries. *Plant Science,* 176**,** 812-823.

BATES, P. D. & BROWSE, J. 2012. The significance of different diacylgycerol synthesis pathways on plant oil composition and bioengineering. *Front Plant Sci,* 3**,** 147.

BATES, P. D., FATIHI, A., SNAPP, A. R., CARLSSON, A. S., BROWSE, J. & LU, C. 2012. Acyl editing and headgroup exchange are the major mechanisms that direct polyunsaturated fatty acid flux into triacylglycerols. *Plant Physiol,* 160**,** 1530-9.

BATES, P. D., STYMNE, S. & OHLROGGE, J. 2013. Biochemical pathways in seed oil synthesis. *Curr Opin Plant Biol,* 16**,** 358-64.

BAUD, S. & LEPINIEC, L. 2009. Regulation of de novo fatty acid synthesis in maturing oilseeds of Arabidopsis. *Plant Physiol Biochem,* 47**,** 448-55.

BAUD, S. & LEPINIEC, L. 2010. Physiological and developmental regulation of seed oil production. *Prog Lipid Res,* 49**,** 235-49.

BECKER, A. & THEISSEN, G. 2003. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Molecular Phylogenetics and Evolution,* 29**,** 464-489.

BECKER, K. & MAKKAR, H. P. S. 2008. Jatropha curcas: A potential source for tomorrow' s oil and biodiesel. *Lipid Technology,* 20**,** 104-107.

BELO, A., ZHENG, P., LUCK, S., SHEN, B., MEYER, D. J., LI, B., TINGEY, S. & RAFALSKI, A. 2008. Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. *Mol Genet Genomics,* 279**,** 1-10.

BENNETT & LEITCH 2012. Angiosperm DNA C-values database. release 8.0, Dec. 2012 ed. http://www.kew.org/cvalues/.

BENNETT, T., SIEBERER, T., WILLETT, B., BOOKER, J., LUSCHNIG, C. & LEYSER, O. 2006. The Arabidopsis MAX pathway controls shoot branching by regulating auxin transport. *Curr Biol,* 16**,** 553-63.

BLANCH, H. 2010. Addressing the Need for Alternative Transportation Fuels: The Joint BioEnergy Institute. *American Chemical Society Journal of Chemical Biology,* 3**,** 17-20.

BRITISH PETROLEUM 2016. *BP Statistical Review of World Energy June 2016*.

BROZYNSKA, M., FURTADO, A. & HENRY, R. J. 2016. Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnol J,* 14**,** 1070-85.

BUI, M. & LIU, Z. 2009. Simple allele-discriminating PCR for cost-effective and rapid genotyping and mapping. *Plant Methods,* 5**,** 1.

BURTON, J. W., MILLER, J. F., VICK, B. A., SCARTH, R. & HOLBROOK, C. C. 2004. Altering Fatty Acid Composition in Oil Seed Crops. *Advances in Agronomy.* Academic Press.

CAHOON, E. B., SHOCKEY, J. M., DIETRICH, C. R., GIDDA, S. K., MULLEN, R. T. & DYER, J. M. 2007. Engineering oilseeds for sustainable production of industrial and nutritional feedstocks: solving bottlenecks in fatty acid flux. *Curr Opin Plant Biol,* 10**,** 236-44.

CALDELARI, D., WANG, G., FARMER, E. E. & DONG, X. 2011. Arabidopsis lox3 lox4 double mutants are male sterile and defective in global proliferative arrest. *Plant Mol Biol,* 75**,** 25-33.

CANVIN, D. T. 1965. The Effect of Temperature on the Oil Content and Fatty Acid Composition of the Oils from Several Oil Seed Crops. *Canadian Journal of Botany,* 43**,** 63-69.

CARLSSON, A. S. 2009. Plant oils as feedstock alternatives to petroleum - A short survey of potential oil crop platforms. *Biochimie,* 91**,** 665-70.

CARVALHO, C. R., CLARINDO, W. R., PRACA, M. M., ARAUJO, F. S. & CARELS, N. 2008. Genome size, base composition and karyotype of Jatropha curcas L., an important biofuel plant. *Plant Science,* 174**,** 613-617.

CHAN, A. P., CRABTREE, J., ZHAO, Q., LORENZI, H., ORVIS, J., PUIU, D., MELAKE-BERHAN, A., JONES, K. M., REDMAN, J., CHEN, G., CAHOON, E. B., GEDIL, M., STANKE, M., HAAS, B. J., WORTMAN, J. R., FRASER-LIGGETT, C. M., RAVEL, J. & RABINOWICZ, P. D. 2010. Draft genome sequence of the oilseed species Ricinus communis. *Nat Biotechnol,* 28**,** 951-6.

CHAPMAN, K. D. & OHLROGGE, J. B. 2012. Compartmentation of triacylglycerol accumulation in plants. *J Biol Chem,* 287**,** 2288-94.

CHU, S. & MAJUMDAR, A. 2012. Opportunities and challenges for a sustainable energy future. *Nature,* 488**,** 294-303.

COSTA, G. G., CARDOSO, K. C., DEL BEM, L. E., LIMA, A. C., CUNHA, M. A., DE CAMPOS-LEITE, L., VICENTINI, R., PAPES, F., MOREIRA, R. C., YUNES, J. A., CAMPOS, F. A. & DA SILVA, M. J. 2010. Transcriptome analysis of the oil-rich seed of the bioenergy crop Jatropha curcas L. *BMC Genomics,* 11**,** 462.

COX, T. S., GLOVER, J. D., VAN TASSEL, D. L., COX, C. M. & DEHAAN, L. R. 2006. Prospects for developing perennial-grain crops. *Bioscience,* 56**,** 649-659.

COYLE, W. 2007. The future of biofuels: a global perspective. *Amber Waves,* 5**,** 24.

CROMWELL, D. 2012. Soybean Meal–An Exceptional Protein Source. *Soybean Meal InfoCenter*.

CUPPEN, E. 2007. Genotyping by Allele-Specific Amplification (KASPar). *CSH Protoc,* 2007**,** pdb prot4841.

DARVASI, A., WEINREB, A., MINKE, V., WELLER, J. I. & SOLLER, M. 1993. Detecting Marker-Qtl Linkage and Estimating Qtl Gene Effect and Map Location Using a Saturated Genetic-Map. *Genetics,* 134**,** 943-951.

DAVEY, J. W., HOHENLOHE, P. A., ETTER, P. D., BOONE, J. Q., CATCHEN, J. M. & BLAXTER, M. L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet,* 12**,** 499-510.

DEFRIES, R., FANZO, J., REMANS, R., PALM, C., WOOD, S. & ANDERMAN, T. L. 2015. Global nutrition. Metrics for land-scarce agriculture. *Science,* 349**,** 238-40.

DEKKERS, J. C. & HOSPITAL, F. 2002. The use of molecular genetics in the improvement of agricultural populations. *Nat Rev Genet,* 3**,** 22-32.

DELONG, A., CALDERON-URREA, A. & DELLAPORTA, S. L. 1993. Sex determination gene TASSELSEED2 of maize encodes a short-chain alcohol dehydrogenase required for stage-specific floral organ abortion. *Cell,* 74**,** 757-68.

DEVAPPA, R. K., MAKKAR, H. P. & BECKER, K. 2010. Nutritional, biochemical, and pharmaceutical potential of proteins and peptides from jatropha: review. *J Agric Food Chem,* 58**,** 6543-55.

DEVAPPA, R. K., MALAKAR, C. C., MAKKAR, H. P. & BECKER, K. 2013. Pharmaceutical potential of phorbol esters from Jatropha curcas oil. *Nat Prod Res,* 27**,** 1459-62.

DHARMASIRI, N., DHARMASIRI, S., WEIJERS, D., LECHNER, E., YAMADA, M., HOBBIE, L., EHRISMANN, J. S., JURGENS, G. & ESTELLE, M. 2005. Plant development is regulated by a family of auxin receptor F box proteins. *Dev Cell,* 9**,** 109-19.

DIVAKARA, B. N., UPADHYAYA, H. D., WANI, S. P. & GOWDA, C. L. L. 2010. Biology and genetic improvement of Jatropha curcas L.: A review. *Applied Energy,* 87**,** 732-742.

DOERGE, R. W. 2002. Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet,* 3**,** 43-52.
DOMAGALSKA, M. A. & LEYSER, O. 2011. Signal integration in the control of shoot branching. *Nat Rev Mol Cell Biol,* 12**,** 211-21.
DORNELAS, M. C., PATREZE, C. M., ANGENENT, G. C. & IMMINK, R. G. 2011. MADS: the missing link between identity and growth? *Trends Plant Sci,* 16**,** 89-97.
DURRETT, T. P., BENNING, C. & OHLROGGE, J. 2008. Plant triacylglycerols as feedstocks for the production of biofuels. *Plant J,* 54**,** 593-607.
DYER, J. M. & MULLEN, R. T. 2008. Engineering plant oils as high-value industrial feedstocks for biorefining: the need for underpinning cell biology research. *Physiol Plant,* 132**,** 11-22.
EDRISI, S. A., DUBEY, R. K., TRIPATHI, V., BAKSHI, M., SRIVASTAVA, P., JAMIL, S., SINGH, H. B., SINGH, N. & ABHILASH, P. C. 2015. Jatropha curcas L.: A crucified plant waiting for resurgence. *Renewable & Sustainable Energy Reviews,* 41**,** 855-862.
EDWARDS, D. & BATLEY, J. 2010. Plant genome sequencing: applications for crop improvement. *Plant Biotechnol J,* 8**,** 2-9.
EHRENREICH, I. M., STAFFORD, P. A. & PURUGGANAN, M. D. 2007. The genetic architecture of shoot branching in Arabidopsis thaliana: a comparative assessment of candidate gene associations vs. quantitative trait locus mapping. *Genetics,* 176**,** 1223-36.
ERNST, M., GRACE, O. M., SASLIS-LAGOUDAKIS, C. H., NILSSON, N., SIMONSEN, H. T. & RONSTED, N. 2015. Global medicinal uses of Euphorbia L. (Euphorbiaceae). *J Ethnopharmacol,* 176**,** 90-101.
EVANS, F. & TAYLOR, S. 1983. Pro-inflammatory, tumour-promoting and anti-tumour diterpenes of the plant families Euphorbiaceae and Thymelaeaceae. *Fortschritte der Chemie organischer Naturstoffe/Progress in the Chemistry of Organic Natural Products.* Springer.
FAIRLESS, D. 2007. Biofuel: The little shrub that could - maybe. *Nature,* 449**,** 652-655.
FAIRLEY, P. 2011. Introduction: Next generation biofuels. *Nature,* 474**,** S2-5.
FARGIONE, J., HILL, J., TILMAN, D., POLASKY, S. & HAWTHORNE, P. 2008. Land clearing and the biofuel carbon debt. *Science,* 319**,** 1235-8.
FAVARO, R., PINYOPICH, A., BATTAGLIA, R., KOOIKER, M., BORGHI, L., DITTA, G., YANOFSKY, M. F., KATER, M. M. & COLOMBO, L. 2003. MADS-box protein complexes control carpel and ovule development in Arabidopsis. *Plant Cell,* 15**,** 2603-11.
FEDOROFF, N. V., BATTISTI, D. S., BEACHY, R. N., COOPER, P. J., FISCHHOFF, D. A., HODGES, C. N., KNAUF, V. C., LOBELL, D., MAZUR, B. J., MOLDEN, D., REYNOLDS, M. P., RONALD, P. C., ROSEGRANT, M. W., SANCHEZ, P. A., VONSHAK, A. & ZHU, J. K. 2010. Radically rethinking agriculture for the 21st century. *Science,* 327**,** 833-4.
FEUILLET, C., LANGRIDGE, P. & WAUGH, R. 2008. Cereal breeding takes a walk on the wild side. *Trends Genet,* 24**,** 24-32.
FEUILLET, C., LEACH, J. E., ROGERS, J., SCHNABLE, P. S. & EVERSOLE, K. 2011. Crop genome sequencing: lessons and rationales. *Trends Plant Sci,* 16**,** 77-88.
FEUSSNER, I. & WASTERNACK, C. 2002. The lipoxygenase pathway. *Annu Rev Plant Biol,* 53**,** 275-97.
FLAGELLA, Z., ROTUNNO, T., TARANTINO, E., DI CATERINA, R. & DE CARO, A. 2002. Changes in seed yield and oil fatty acid composition of high oleic sunflower (Helianthus annuus L.) hybrids in relation to the sowing date and the water regime. *European Journal of Agronomy,* 17**,** 221-230.
FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. 2016. *FAOSTAT - Data Statistics for the Food and Agriculture Organisation of the United Nations.*
FRESNEDO-RAMIREZ, J. 2013. The Floral Biology of Jatropha curcas L.-A Review. *Tropical Plant Biology,* 6**,** 1-15.
GARGI JOSHI, A. S., ALOK SHUKLA 2011. Synergistic response of auxin and ethylene on physiology of Jatropha curcas L. *BRAZILIAN SOCIETY OF PLANT PHYSIOLOGY,* 23**,** 67-77.
GERLAND, P., RAFTERY, A. E., SEVCIKOVA, H., LI, N., GU, D., SPOORENBERG, T., ALKEMA, L., FOSDICK, B. K., CHUNN, J., LALIC, N., BAY, G., BUETTNER, T., HEILIG, G. K. & WILMOTH, J. 2014. World population stabilization unlikely this century. *Science,* 346**,** 234-7.
GIRAUT, L., FALQUE, M., DROUAUD, J., PEREIRA, L., MARTIN, O. C. & MEZARD, C. 2011. Genome-wide crossover distribution in Arabidopsis thaliana meiosis reveals sex-specific patterns along chromosomes. *PLoS Genet,* 7**,** e1002354.
GISH, W. & STATES, D. J. 1993. Identification of protein coding regions by database similarity search. *Nat Genet,* 3**,** 266-72.
GITTE I. FRANDSENA, J. M., JASON T. C. TZEN 2001. Oil bodies and their associated proteins, oleosin and caleosin. *PHYSIOLOGIA PLANTARUM,* 112**,** 301-307.
GODFRAY, H. C., BEDDINGTON, J. R., CRUTE, I. R., HADDAD, L., LAWRENCE, D., MUIR, J. F., PRETTY, J., ROBINSON, S., THOMAS, S. M. & TOULMIN, C. 2010. Food security: the challenge of feeding 9 billion people. *Science,* 327**,** 812-8.

GOODSTEIN, D. M., SHU, S., HOWSON, R., NEUPANE, R., HAYES, R. D., FAZO, J., MITROS, T., DIRKS, W., HELLSTEN, U., PUTNAM, N. & ROKHSAR, D. S. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res,* 40**,** D1178-86.

GOUR, V. Production practices including post-harvest management of Jatropha curcas. Proceedings of the biodiesel conference toward energy independance-Focus of Jatropha, Hyderabad, India, 2006. 223-251.

GRAEF, G., LAVALLEE, B. J., TENOPIR, P., TAT, M., SCHWEIGER, B., KINNEY, A. J., VAN GERPEN, J. H. & CLEMENTE, T. E. 2009. A high-oleic-acid and low-palmitic-acid soybean: agronomic performance and evaluation as a feedstock for biodiesel. *Plant Biotechnol J,* 7**,** 411-21.

GRAHAM-ROWE, D. 2011. Agriculture: Beyond food versus fuel. *Nature,* 474**,** S6-8.

GRAHAM, I. A., LARSON, T. & NAPIER, J. A. 2007. Rational metabolic engineering of transgenic plants for biosynthesis of omega-3 polyunsaturates. *Curr Opin Biotechnol,* 18**,** 142-7.

GRAHAM, W. H. A. J. K. A. M. K. J. A. C. D. R. I. A. 2009. Genetic and biochemical analysis of diversity in edible and non-edible accessions of Jatropha

curcas (L.) from Madagascar and Mexico. *Unpublished*.

GREEN, P. F., K.; CROOKS, S. 1990. Documentation for CRI-MAP. Version 2.4 ed. St Louis: MO: Washington University School of Medicine.

GU, K., YI, C., TIAN, D., SANGHA, J. S., HONG, Y. & YIN, Z. 2012. Expression of fatty acid and lipid biosynthetic genes in developing endosperm of Jatropha curcas. *Biotechnol Biofuels,* 5**,** 47.

GUO, M. X., SONG, W. P. & BUHAIN, J. 2015. Bioenergy and biofuels: History, status, and perspective. *Renewable & Sustainable Energy Reviews,* 42**,** 712-725.

GUPTA, S. K. 2015. *Breeding Oilseed Crops for Sustainable Production: Opportunities and Constraints,* Academic Press 2015.

HAMBLIN, M. T., BUCKLER, E. S. & JANNINK, J. L. 2011. Population genetics of genomics-based crop improvement methods. *Trends Genet,* 27**,** 98-106.

HAYDEN, M. J., NGUYEN, T. M., WATERMAN, A. & CHALMERS, K. J. 2008. Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. *BMC Genomics,* 9**,** 80.

HE, W. 2011. *Biochemical and genetic analyses of Jatropha curcas L. Seed composition.* Doctor of Philosophy, University of York.

HE, W., KING, A. J., KHAN, M. A., CUEVAS, J. A., RAMIARAMANANA, D. & GRAHAM, I. A. 2011. Analysis of seed phorbol-ester and curcin content together with genetic diversity in multiple provenances of Jatropha curcas L. from Madagascar and Mexico. *Plant Physiol Biochem,* 49**,** 1183-90.

HECKER, E. 1968. Cocarcinogenic principles from the seed oil of Croton tiglium and from other Euphorbiaceae. *Cancer Res,* 28**,** 2338-49.

HELLER, J. 1996. Physic nut. Jatropha curcas L. *International Plant Genetic Resources Institute,* Promoting the conservation and use of underutilized and neglected crops. 1.

HERNANDEZ, M. L., MANCHA, M. & MARTINEZ-RIVAS, J. M. 2005. Molecular cloning and characterization of genes encoding two microsomal oleate desaturases (FAD2) from olive. *Phytochemistry,* 66**,** 1417-26.

HILL, J., NELSON, E., TILMAN, D., POLASKY, S. & TIFFANY, D. 2006. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proc Natl Acad Sci U S A,* 103**,** 11206-10.

HIRAKAWA, H., TSUCHIMOTO, S., SAKAI, H., NAKAYAMA, S., FUJISHIRO, T., KISHIDA, Y., KOHARA, M., WATANABE, A., YAMADA, M., AIZU, T., TOYODA, A., FUJIYAMA, A., TABATA, S., FUKUI, K. & SATO, S. 2012. Upgraded genomic information of Jatropha curcas L. *Plant Biotechnology,* 29**,** 123-130.

HO, D. P., NGO, H. H. & GUO, W. 2014. A mini review on renewable sources for biofuel. *Bioresour Technol,* 169**,** 742-9.

HOLLAND, J. B. 2007. Genetic architecture of complex traits in plants. *Curr Opin Plant Biol,* 10**,** 156-61.

HU, X., SULLIVAN-GILBERT, M., GUPTA, M. & THOMPSON, S. A. 2006. Mapping of the loci controlling oleic and linolenic acid contents and development of fad2 and fad3 allele-specific markers in canola (Brassica napus L.). *Theor Appl Genet,* 113**,** 497-507.

HYUN, T. K., KUMAR, D., CHO, Y. Y., HYUN, H. N. & KIM, J. S. 2013. Computational identification and phylogenetic analysis of the oil-body structural proteins, oleosin and caleosin, in castor bean and flax. *Gene,* 515**,** 454-60.

IBM 2013. IBM SPSS Statistics for Windows. 22.0 ed.: Armonk, NY: IBM Corp.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II, and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp. in IPCC AR5 Synthesis Report website.

JARAMILLO, F. & DESTOUNI, G. 2015. Local flow regulation and irrigation raise global humanwater consumption and footprint. *Science,* 350.

JERRY D. GLOVER, C. M. C., JOHN P. REGANOLD 2007. Future Farming: A Return to Roots? *Scientific American*.

JIANG, H., WU, P., ZHANG, S., SONG, C., CHEN, Y., LI, M., JIA, Y., FANG, X., CHEN, F. & WU, G. 2012. Global analysis of gene expression profiles in developing physic nut (Jatropha curcas L.) seeds. *PLoS One,* 7**,** e36522.

JOLIVET, P., ACEVEDO, F., BOULARD, C., D'ANDREA, S., FAURE, J. D., KOHLI, A., NESI, N., VALOT, B. & CHARDOT, T. 2013. Crop seed oil bodies: from challenges in protein identification to an emerging picture of the oil body proteome. *Proteomics,* 13**,** 1836-49.

JORDAN, D. R., TAO, Y. Z., GODWIN, I. D., HENZELL, R. G., COOPER, M. & MCINTYRE, C. L. 2005. Comparison of identity by descent and identity by state for detecting genetic regions under selection in a sorghum pedigree breeding program. *Molecular Breeding,* 14**,** 441-454.

JUN LI, M.-R. L., PING-ZHI WU, CHANG-EN TIAN, HUA-WU JIANG, GUO-JIANG WU 2008. Molecular cloning and expression analysis of a gene encoding a putative B-ketoacyl-acyl carrier protein (ACP) synthase III (KAS III) from Jatropha curcas. *Tree Physiology***,** 921-927.

KANTAR, M. B., TYL, C. E., DORN, K. M., ZHANG, X., JUNGERS, J. M., KASER, J. M., SCHENDEL, R. R., ECKBERG, J. O., RUNCK, B. C., BUNZEL, M., JORDAN, N. R., STUPAR, R. M., MARKS, M. D., ANDERSON, J. A., JOHNSON, G. A., SHEAFFER, C. C., SCHOENFUSS, T. C., ISMAIL, B., HEIMPEL, G. E. & WYSE, D. L. 2016. Perennial Grain and Oilseed Crops. *Annu Rev Plant Biol,* 67**,** 703-29.

KEPINSKI, S. & LEYSER, O. 2005. The Arabidopsis F-box protein TIR1 is an auxin receptor. *Nature,* 435**,** 446-51.

KERR, R. A. 2011. Energy supplies. Peak oil production may already be here. *Science,* 331**,** 1510-1.

KHUSH, G. S. 2001. Green revolution: the way forward. *Nat Rev Genet,* 2**,** 815-22.

KIANIAN, S. F., EGLI, M. A., PHILLIPS, R. L., RINES, H. W., SOMERS, D. A., GENGENBACH, B. G., WEBSTER, F. H., LIVINGSTON, S. M., GROH, S., O'DONOUGHUE, L. S., SORRELLS, M. E., WESENBERG, D. M., STUTHMAN, D. D. & FULCHER, R. G. 1999. Association of a major groat oil content QTL and an acetyl-CoA carboxylase gene in oat. *Theoretical and Applied Genetics,* 98**,** 884-894.

KING, A. J., HE, W., CUEVAS, J. A., FREUDENBERGER, M., RAMIARAMANANA, D. & GRAHAM, I. A. 2009. Potential of Jatropha curcas as a source of renewable oil and animal feed. *J Exp Bot,* 60**,** 2897-905.

KING, A. J., LI, Y. & GRAHAM, I. A. 2011. Profiling the Developing Jatropha curcas L. Seed Transcriptome by Pyrosequencing. *Bioenergy Research,* 4**,** 211-221.

KING, A. J., MONTES, L. R., CLARKE, J. G., AFFLECK, J., LI, Y., WITSENBOER, H., VAN DER VOSSEN, E., VAN DER LINDE, P., TRIPATHI, Y., TAVARES, E., SHUKLA, P., RAJASEKARAN, T., VAN LOO, E. N. & GRAHAM, I. A. 2013. Linkage mapping in the oilseed crop Jatropha curcas L. reveals a locus controlling the biosynthesis of phorbol esters which cause seed toxicity. *Plant Biotechnol J,* 11**,** 986-96.

KING, A. J., MONTES, L. R., CLARKE, J. G., ITZEP, J., PEREZ, C. A., JONGSCHAAP, R. E., VISSER, R. G., VAN LOO, E. N. & GRAHAM, I. A. 2015. Identification of QTL markers contributing to plant growth, oil yield and fatty acid composition in the oilseed crop Jatropha curcas L. *Biotechnol Biofuels,* 8**,** 160.

KNOTHE, G. 2005. Dependence of biodiesel fuel properties on the structure of fatty acid alkyl esters. *Fuel Processing Technology,* 86**,** 1059-1070.

KNOTHE, G. 2009. Improving biodiesel fuel properties by modifying fatty ester composition. *Energy & Environmental Science,* 2**,** 759-766.

KNUTZON, D. S., THOMPSON, G. A., RADKE, S. E., JOHNSON, W. B., KNAUF, V. C. & KRIDL, J. C. 1992. Modification of Brassica seed oil by antisense expression of a stearoyl-acyl carrier protein desaturase gene. *Proc Natl Acad Sci U S A,* 89**,** 2624-8.

KUMAR, A. & SHARMA, S. 2008. An evaluation of multipurpose oil seed crop for industrial uses (Jatropha curcas L.): A review. *Industrial Crops and Products,* 28**,** 1-10.

LAI, Y. & SUN, F. 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol Biol Evol,* 20**,** 2123-31.

LANGRIDGE, P. & FLEURY, D. 2011. Making the most of 'omics' for crop breeding. *Trends Biotechnol,* 29**,** 33-40.

LARDIZABAL, K., EFFERTZ, R., LEVERING, C., MAI, J., PEDROSO, M. C., JURY, T., AASEN, E., GRUYS, K. & BENNETT, K. 2008. Expression of Umbelopsis ramanniana DGAT2A in seed increases oil in soybean. *Plant Physiol,* 148**,** 89-96.

LARSEN, T. A., HOFFMANN, S., LUTHI, C., TRUFFER, B. & MAURER, M. 2016. Emerging solutions to the water challenges of an urbanizing world. *Science,* 352**,** 928-33.

LEFF, B., RAMANKUTTY, N. & FOLEY, J. A. 2004. Geographic distribution of major crops across the world. *Global Biogeochemical Cycles,* 18**,** n/a-n/a.

LI-BEISSON, Y., SHORROSH, B., BEISSON, F., ANDERSSON, M. X., ARONDEL, V., BATES, P. D., BAUD, S., BIRD, D., DEBONO, A., DURRETT, T. P., FRANKE, R. B., GRAHAM, I. A.,

KATAYAMA, K., KELLY, A. A., LARSON, T., MARKHAM, J. E., MIQUEL, M., MOLINA, I., NISHIDA, I., ROWLAND, O., SAMUELS, L., SCHMID, K. M., WADA, H., WELTI, R., XU, C., ZALLOT, R. & OHLROGGE, J. 2013. Acyl-lipid metabolism. *Arabidopsis Book,* 11**,** e0161.

LI, Y. C., KOROL, A. B., FAHIMA, T., BEILES, A. & NEVO, E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol,* 11**,** 2453-65.

LILJEGREN, S. J., DITTA, G. S., ESHED, Y., SAVIDGE, B., BOWMAN, J. L. & YANOFSKY, M. F. 2000. SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. *Nature,* 404**,** 766-70.

LIU, P., WANG, C. M., LI, L., SUN, F., LIU, P. & YUE, G. H. 2011. Mapping QTLs for oil traits and eQTLs for oleosin genes in jatropha. *BMC Plant Biol,* 11**,** 132.

LOBELL, D. B., BURKE, M. B., TEBALDI, C., MASTRANDREA, M. D., FALCON, W. P. & NAYLOR, R. L. 2008. Prioritizing climate change adaptation needs for food security in 2030. *Science,* 319**,** 607-610.

LUO, C.-W., LI, K., CHEN, Y. & SUN, Y.-Y. 2007. Floral display and breeding system of Jatropha curcas L. *Forestry Studies in China,* 9**,** 114-119.

MACKAY, T. F., STONE, E. A. & AYROLES, J. F. 2009. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet,* 10**,** 565-77.

MAKKAR, H. P. S., ADERIBIGBE, A. O. & BECKER, K. 1998. Comparative evaluation of non-toxic and toxic varieties of Jatropha curcas for chemical composition, digestibility, protein degradability and toxic factors. *Food Chemistry,* 62**,** 207-215.

MAKKAR, H. P. S., BECKER, K., SPORER, F. & WINK, M. 1997. Studies on nutritive potential and toxic constituents of different provenances of Jatropha curcas. *Journal of Agricultural and Food Chemistry,* 45**,** 3152-3157.

MAKKAR, H. P. S., KUMAR, V., OYELEYE, O. O., AKINLEYE, A. O., ANGULO-ESCALANTE, M. A. & BECKER, K. 2011. Jatropha platyphylla, a new non-toxic Jatropha species: Physical properties and chemical constituents including toxic and antinutritional factors of seeds. *Food Chemistry,* 125**,** 63-71.

MAKKAR, H. R. S. & BECKER, K. 2009. Jatropha curcas, a promising crop for the generation of biodiesel and value-added coproducts. *European Journal of Lipid Science and Technology,* 111**,** 773-787.

MAKWANA, V., SHUKLA, P. & ROBIN, P. 2010. GA application induces alteration in sex ratio and cell death in Jatropha curcas. *Plant Growth Regulation,* 61**,** 121-125.

MARTINS, W. S., LUCAS, D. C., NEVES, K. F. & BERTIOLI, D. J. 2009. WebSat--a web software for microsatellite marker development. *Bioinformation,* 3**,** 282-3.

MCSTEEN, P. & LEYSER, O. 2005. Shoot branching. *Annu Rev Plant Biol,* 56**,** 353-74.

MENG, G. T., LI, G. X., HE, L. P., CHAI, Y., KONG, J. J. & LEI, Y. B. 2013. Combined Effects of CO2 Enrichment and Drought Stress on Growth and Energetic Properties in the Seedlings of a Potential Bioenergy Crop Jatropha curcas. *Journal of Plant Growth Regulation,* 32**,** 542-550.

MICHAEL, T. P. 2014. Plant genome size variation: bloating and purging DNA. *Brief Funct Genomics,* 13**,** 308-17.

MONTES OSORIO, L. R., TORRES SALVADOR, A. F., JONGSCHAAP, R. E., AZURDIA PEREZ, C. A., BERDUO SANDOVAL, J. E., TRINDADE, L. M., VISSER, R. G. & VAN LOO, E. N. 2014. High level of molecular and phenotypic biodiversity in Jatropha curcas from Central America compared to Africa, Asia and South America. *BMC Plant Biol,* 14**,** 77.

MORENO-PEREZ, A. J., VENEGAS-CALERON, M., VAISTIJ, F. E., SALAS, J. J., LARSON, T. R., GARCES, R., GRAHAM, I. A. & MARTINEZ-FORCE, E. 2012. Reduced expression of FatA thioesterases in Arabidopsis affects the oil content and fatty acid composition of the seeds. *Planta,* 235**,** 629-39.

MORRELL, P. L., BUCKLER, E. S. & ROSS-IBARRA, J. 2012. Crop genomics: advances and applications. *Nat Rev Genet,* 13**,** 85-96.

MUKHERJEE, P., VARSHNEY, A., JOHNSON, T. S. & JHA, T. B. 2011. Jatropha curcas: a review on biotechnological status and challenges. *Plant Biotechnology Reports,* 5**,** 197-215.

MURUGESAN, A., UMARANI, C., SUBRAMANIAN, R. & NEDUNCHEZHIAN, N. 2009. Bio-diesel as an alternative fuel for diesel engines-A review. *Renewable & Sustainable Energy Reviews,* 13**,** 653-662.

NAPIER, J. A. & GRAHAM, I. A. 2010. Tailoring plant lipid composition: designer oilseeds come of age. *Curr Opin Plant Biol,* 13**,** 330-7.

NAPIER, J. A., HASLAM, R. P., BEAUDOIN, F. & CAHOON, E. B. 2014. Understanding and manipulating plant lipid composition: Metabolic engineering leads the way. *Curr Opin Plant Biol,* 19**,** 68-75.

NATARAJAN, P. & PARANI, M. 2011. De novo assembly and transcriptome analysis of five major tissues of Jatropha curcas L. using GS FLX titanium platform of 454 pyrosequencing. *BMC Genomics,* 12**,** 191.

NATIONAL ACADEMY OF SCIENCES 2009. *Liquid Transportation Fuels from Coal and Biomass: Technological Status, Costs, and Environmental Impacts*, The National Academies Press, Washington, D. C.

NEUMANN, K., VERBURG, P. H., STEHFEST, E. & MULLER, C. 2010. The yield gap of global grain production: A spatial analysis. *Agricultural Systems,* 103**,** 316-326.

NEWBOLD, T., HUDSON, L. N., ARNELL, A. P., CONTU, S., DE PALMA, A., FERRIER, S., HILL, S. L., HOSKINS, A. J., LYSENKO, I., PHILLIPS, H. R., BURTON, V. J., CHNG, C. W., EMERSON, S., GAO, D., PASK-HALE, G., HUTTON, J., JUNG, M., SANCHEZ-ORTIZ, K., SIMMONS, B. I., WHITMEE, S., ZHANG, H., SCHARLEMANN, J. P. & PURVIS, A. 2016. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science,* 353**,** 288-91.

OLIVEIRA, E. J., PADUA, J. G., ZUCCHI, M. I., VENCOVSKY, R. & VIEIRA, M. L. C. 2006. Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology,* 29**,** 294-307.

ONGARO, V. & LEYSER, O. 2008. Hormonal control of shoot branching. *J Exp Bot,* 59**,** 67-74.

OOIJEN, J. W. V. 2004. MapQTL® 5, Software for the mapping of quantitative trait loci in experimental populations. Wageningen, Netherlands: Kyazma B.V.

OOIJEN, J. W. V. 2011. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genet Res (Camb),* 93**,** 343-9.

OOIJEN, J. W. V., KYAZMA, B. V. & JANSEN, J. 2013. *Genetic Mapping in Experimental Populations*, Cambridge Publishing.

OPENSHAW, K. 2000. A review of Jatropha curcas: an oil plant of unfulfilled promise. *Biomass & Bioenergy,* 19**,** 1-15.

PAN, B. Z. & XU, Z. F. 2011. Benzyladenine Treatment Significantly Increases the Seed Yield of the Biofuel Plant Jatropha curcas. *Journal of Plant Growth Regulation,* 30**,** 166-174.

PARTHIBANE, V., RAJAKUMARI, S., VENKATESHWARI, V., IYAPPAN, R. & RAJASEKHARAN, R. 2012. Oleosin is bifunctional enzyme that has both monoacylglycerol acyltransferase and phospholipase activities. *J Biol Chem,* 287**,** 1946-54.

PATEL, M., JUNG, S., MOORE, K., POWELL, G., AINSWORTH, C. & ABBOTT, A. 2004. High-oleate peanut mutants result from a MITE insertion into the FAD2 gene. *Theor Appl Genet,* 108**,** 1492-502.

PAUL K. STUMPF, J. B. M., W. DAVID NES 2012. *The Metabolism, Structure, and Function of Plant Lipids*, Springer Science & Business Media.

PECINA-QUINTERO, V., ANAYA-LOPEZ, J. L., ZAMARRIPA-COLMENERO, A., NUNEZ-COLIN, C. A., MONTES-GARCIA, N., SOLIS-BONILLA, J. L. & JIMENEZ-BECERRIL, M. F. 2014. Genetic structure of Jatropha curcas L. in Mexico and probable centre of origin. *Biomass & Bioenergy,* 60**,** 147-155.

PEREGRIN-ALVAREZ, J. M., SANFORD, C. & PARKINSON, J. 2009. The conservation and evolutionary modularity of metabolism. *Genome Biol,* 10**,** R63.

PFLIEGER, S., LEFEBVRE, V. & CAUSSE, M. 2001. The candidate gene approach in plant genetics: a review. *Molecular Breeding,* 7**,** 275-291.

PHUMICHAI, C., PHUMICHAI, T., KONGSIRI, N., WONGKAEW, A., SRIPICHIT, P. & KAVEETA, R. 2011. Isolation of 55 microsatellite markers for Jatropha curcas and its closely related species. *Biologia Plantarum,* 55**,** 387.

QI-BAO SUN, LIN-FENG LI, YONG LI, GUO-JIANG WU & XUE-JUN GE 2008. SSR and AFLP Markers Reveal Low Genetic Diversity in the Biofuel Plant Jatropha curcas in China. *Crop Science,* 48.

QU, J., MAO, H. Z., CHEN, W., GAO, S. Q., BAI, Y. N., SUN, Y. W., GENG, Y. F. & YE, J. 2012. Development of marker-free transgenic Jatropha plants with increased levels of seed oleic acid. *Biotechnol Biofuels,* 5**,** 10.

RAFALSKI, J. A. 2010. Association genetics in crop improvement. *Curr Opin Plant Biol,* 13**,** 174-80.

RAGAUSKAS, A. J., WILLIAMS, C. K., DAVISON, B. H., BRITOVSEK, G., CAIRNEY, J., ECKERT, C. A., FREDERICK, W. J., JR., HALLETT, J. P., LEAK, D. J., LIOTTA, C. L., MIELENZ, J. R., MURPHY, R., TEMPLER, R. & TSCHAPLINSKI, T. 2006. The path forward for biofuels and biomaterials. *Science,* 311**,** 484-9.

ROESLER, K., SHINTANI, D., SAVAGE, L., BODDUPALLI, S. & OHLROGGE, J. 1997. Targeting of the Arabidopsis homomeric acetyl-coenzyme A carboxylase to plastids of rapeseeds. *Plant Physiol,* 113**,** 75-81.

SABANDAR, C. W., AHMAT, N., JAAFAR, F. M. & SAHIDIN, I. 2013. Medicinal property, phytochemistry and pharmacology of several Jatropha species (Euphorbiaceae): a review. *Phytochemistry,* 85**,** 7-29.

SANDHU, D., ALT, J. L., SCHERDER, C. W., FEHR, W. R. & BHATTACHARYYA, M. K. 2007. Enhanced oleic acid content in the soybean mutant M23 is associated with the deletion in the Fad2-1a gene encoding a fatty acid desaturase. *Journal of the American Oil Chemists Society,* 84**,** 229-235.

SANTALUCIA, J., JR. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A,* 95**,** 1460-5.
SANYAL, A. & RANDAL LINDER, C. 2012. Quantitative trait loci involved in regulating seed oil composition in Arabidopsis thaliana and their evolutionary implications. *Theor Appl Genet,* 124**,** 723-38.
SASAKI, Y. & NAGANO, Y. 2004. Plant acetyl-CoA carboxylase: structure, biosynthesis, regulation, and gene manipulation for plant breeding. *Biosci Biotechnol Biochem,* 68**,** 1175-84.
SATO, S., HIRAKAWA, H., ISOBE, S., FUKAI, E., WATANABE, A., KATO, M., KAWASHIMA, K., MINAMI, C., MURAKI, A., NAKAZAKI, N., TAKAHASHI, C., NAKAYAMA, S., KISHIDA, Y., KOHARA, M., YAMADA, M., TSURUOKA, H., SASAMOTO, S., TABATA, S., AIZU, T., TOYODA, A., SHIN-I, T., MINAKUCHI, Y., KOHARA, Y., FUJIYAMA, A., TSUCHIMOTO, S., KAJIYAMA, S., MAKIGANO, E., OHMIDO, N., SHIBAGAKI, N., CARTAGENA, J. A., WADA, N., KOHINATA, T., ATEFEH, A., YUASA, S., MATSUNAGA, S. & FUKUI, K. 2011. Sequence analysis of the genome of an oil-bearing tree, Jatropha curcas L. *DNA Res,* 18**,** 65-76.
SCHLOTTERER, C. 2004. The evolution of molecular markers - just a matter of fashion? *Nat Rev Genet,* 5**,** 63-69.
SCHLOTTERER, C. & TAUTZ, D. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res,* 20**,** 211-5.
SCHUPPERT, G. F., TANG, S. X., SLABAUGH, M. B. & KNAPP, S. J. 2006. The sunflower high-oleic mutant Ol carries variable tandem repeats of FAD2-1, a seed-specific oleoyl-phosphatidyl choline desaturase. *Molecular Breeding,* 17**,** 241-256.
SEATON G., H. J., GRUNCHEC J.A., WHITE I., ALLEN J., DE KONING D.J., WEI W., BERRY D., HALEY C., KNOTT S. . GridQTL: A Grid Portal for QTL Mapping of Compute Intensive Datasets. Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, August 13-18 2006 Belo Horizonte, Brazil.
SHARMA, A. & CHAUHAN, R. S. 2012. In silico identification and comparative genomics of candidate genes involved in biosynthesis and accumulation of seed oil in plants. *Comp Funct Genomics,* 2012**,** 914843.
SHEN, B., ALLEN, W. B., ZHENG, P., LI, C., GLASSMAN, K., RANCH, J., NUBEL, D. & TARCZYNSKI, M. C. 2010. Expression of ZmLEC1 and ZmWRI1 increases seed oil production in maize. *Plant Physiol,* 153**,** 980-7.
SHINTANI, D. K. & OHLROGGE, J. B. 1995. Feedback Inhibition of Fatty-Acid Synthesis in Tobacco Suspension Cells. *Plant Journal,* 7**,** 577-587.
SHOCKEY, J. M., FULDA, M. S. & BROWSE, J. A. 2002. Arabidopsis contains nine long-chain acyl-coenzyme a synthetase genes that participate in fatty acid and glycerolipid metabolism. *Plant Physiol,* 129**,** 1710-22.
SIMS, R. E., MABEE, W., SADDLER, J. N. & TAYLOR, M. 2010. An overview of second generation biofuel technologies. *Bioresour Technol,* 101**,** 1570-80.
SNYDER, C. L., YURCHENKO, O. P., SILOTO, R. M., CHEN, X., LIU, Q., MIETKIEWSKA, E. & WESELAKE, R. J. 2009. Acyltransferase action in the modification of seed oil biosynthesis. *N Biotechnol,* 26**,** 11-6.
SOMERVILLE, C., YOUNGS, H., TAYLOR, C., DAVIS, S. C. & LONG, S. P. 2010. Feedstocks for lignocellulosic biofuels. *Science,* 329**,** 790-2.
STAUB, J. E., SERQUEN, F. C. & GUPTA, M. 1996. Genetic markers, map construction, and their application in plant breeding. *Hortscience,* 31**,** 729-741.
STEINBUKS, J. & HERTEL, T. W. 2016. Confronting the Food-Energy-Environment Trilemma: Global Land Use in the Long Run. *Environmental & Resource Economics,* 63**,** 545-570.
STICKLEN, M. B. 2008. Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nat Rev Genet,* 9**,** 433-43.
STIENEKE, D. L., EUJAYL, I. A. 2007. Imperfect SSR Finder. Version 1.0 ed. Kimberley.
STIRNBERG, P., CHATFIELD, S. P. & LEYSER, H. M. 1999. AXR1 acts after lateral bud formation to inhibit lateral bud growth in Arabidopsis. *Plant Physiol,* 121**,** 839-47.
STOLL, C., LUHS, W., ZARHLOUL, M. K., BRUMMEL, M., SPENER, F. & FRIEDT, W. 2006. Knockout of KASIII regulation changes fatty acid composition in canola (Brassica napus). *European Journal of Lipid Science and Technology,* 108**,** 277-286.
SUN, F., LIU, P., YE, J., LO, L. C., CAO, S., LI, L., YUE, G. H. & WANG, C. M. 2012. An approach for jatropha improvement using pleiotropic QTLs regulating plant growth and seed yield. *Biotechnol Biofuels,* 5**,** 42.
SUN, F., ZHANG, W., XIONG, G., YAN, M., QIAN, Q., LI, J. & WANG, Y. 2010. Identification and functional analysis of the MOC1 interacting protein 1. *J Genet Genomics,* 37**,** 69-77.
SUN, Q. B., LI, L. F., LI, Y., WU, G. J. & GE, X. J. 2008. SSR and AFLP markers reveal low genetic diversity in the biofuel plant Jatropha curcas in China. *Crop Science,* 48**,** 1865-1871.
TAJIMA, D., KANEKO, A., SAKAMOTO, M., ITO, Y., HUE, N. T., MIYAZAKI, M., ISHIBASHI, Y., YUASA, T. & IWAYA-INOUE, M. 2013. <i>Wrinkled</i> 1 (WRI1) Homologs, AP2-

Type Transcription Factors Involving Master Regulation of Seed Storage Oil Synthesis in Castor Bean (<i>Ricinus communis</i>L.). *American Journal of Plant Sciences,* 04**,** 333-339.
TESTER, M. & LANGRIDGE, P. 2010. Breeding technologies to increase crop production in a changing world. *Science,* 327**,** 818-22.
THE EUROPEAN PARLIMENT 2009. Directive 2009/28/EC of the European Parliament and of the Council of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC (Text with EEA relevance). *In:* PARLIMENT, T. E. (ed.).
THOMAS, R., SAH, N. K. & SHARMA, P. B. 2008. Therapeutic biology of Jatropha curcas: a mini review. *Curr Pharm Biotechnol,* 9**,** 315-24.
THOMPSON, B. E. & HAKE, S. 2009. Translational biology: from Arabidopsis flowers to grass inflorescence architecture. *Plant Physiol,* 149**,** 38-45.
TILMAN, D., SOCOLOW, R., FOLEY, J. A., HILL, J., LARSON, E., LYND, L., PACALA, S., REILLY, J., SEARCHINGER, T., SOMERVILLE, C. & WILLIAMS, R. 2009. Energy. Beneficial biofuels--the food, energy, and environment trilemma. *Science,* 325**,** 270-1.
TO, A., JOUBES, J., BARTHOLE, G., LECUREUIL, A., SCAGNELLI, A., JASINSKI, S., LEPINIEC, L. & BAUD, S. 2012. WRINKLED transcription factors orchestrate tissue-specific regulation of fatty acid biosynthesis in Arabidopsis. *Plant Cell,* 24**,** 5007-23.
UNTERGASSER, A., CUTCUTACHE, I., KORESSAAR, T., YE, J., FAIRCLOTH, B. C., REMM, M. & ROZEN, S. G. 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Res,* 40**,** e115.
US DEPARTMENT OF ENERGY 2014. Strategic plan 2014-2018
US ENERGY INFORMATION ADMINISTRATION 2016. *International Energy Outlook 2016.*
UTOMO, C., SUBROTO, A. P., DARMAWAN, C., SETYOBUDI, R. H., NUGROHO, Y. A. & LIWANG, T. 2015. Construction for Delta-12 Fatty Acid Desaturase (FAD2) Silencing to Improve Oil Quality of Jatropha curcas Linn. *New and Renewable Energy and Energy Conservation, the 3rd Indo Ebtke-Conex 2014, Conference and Exhibition Indonesia,* 65**,** 36-41.
VAN ORSOUW, N. J., HOGERS, R. C., JANSSEN, A., YALCIN, F., SNOEIJERS, S., VERSTEGE, E., SCHNEIDERS, H., VAN DER POEL, H., VAN OEVEREN, J., VERSTEGEN, H. & VAN EIJK, M. J. 2007. Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One,* 2**,** e1172.
VAN TASSEL, D. L., DEHAAN, L. R. & COX, T. S. 2010. Missing domesticated plant forms: can artificial selection fill the gap? *Evol Appl,* 3**,** 434-52.
VANHERCKE, T., EL TAHCHY, A., SHRESTHA, P., ZHOU, X. R., SINGH, S. P. & PETRIE, J. R. 2013. Synergistic effect of WRI1 and DGAT1 coexpression on triacylglycerol biosynthesis in plants. *FEBS Lett,* 587**,** 364-9.
VARSHNEY, R. K., BANSAL, K. C., AGGARWAL, P. K., DATTA, S. K. & CRAUFURD, P. Q. 2011. Agricultural biotechnology for crop improvement in a variable climate: hope or hype? *Trends Plant Sci,* 16**,** 363-71.
VARSHNEY, R. K., TERAUCHI, R. & MCCOUCH, S. R. 2014. Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS Biol,* 12**,** e1001883.
VEGA-SANCHEZ, M. E. & RONALD, P. C. 2010. Genetic and biotechnological approaches for biofuel crop improvement. *Curr Opin Biotechnol,* 21**,** 218-24.
VOELKER, T. & KINNEY, A. J. 2001. Variations in the Biosynthesis of Seed-Storage Lipids. *Annu Rev Plant Physiol Plant Mol Biol,* 52**,** 335-361.
VOORRIPS, R. E. 2002. MapChart: software for the graphical presentation of linkage maps and QTLs. *J Hered,* 93**,** 77-8.
WADA, K. C., YAMADA, M., SHIRAYA, T. & TAKENO, K. 2010. Salicylic acid and the flowering gene FLOWERING LOCUS T homolog are involved in poor-nutrition stress-induced flowering of Pharbitis nil. *Journal of Plant Physiology,* 167**,** 447-452.
WANG, C. M., LIU, P., YI, C., GU, K., SUN, F., LI, L., LO, L. C., LIU, X., FENG, F., LIN, G., CAO, S., HONG, Y., YIN, Z. & YUE, G. H. 2011. A first generation microsatellite- and SNP-based linkage map of Jatropha. *PLoS One,* 6**,** e23632.
WANG, W., CHEN, K. & XU, C. 2006. DNA quantification using EvaGreen and a real-time PCR instrument. *Anal Biochem,* 356**,** 303-5.
WANG, Y. & LI, J. 2006. Genes controlling plant architecture. *Curr Opin Biotechnol,* 17**,** 123-9.
WANG, Y. & LI, J. 2008. Molecular basis of plant architecture. *Annu Rev Plant Biol,* 59**,** 253-79.
WANG, Z. Y., LIN, J. M. & XU, Z. F. 2008. [Oil contents and fatty acid composition in Jatropha curcas seeds collected from different regions]. *Nan Fang Yi Ke Da Xue Xue Bao,* 28**,** 1045-6.
WENDER, P. A., KEE, J. M. & WARRINGTON, J. M. 2008. Practical synthesis of prostratin, DPP, and their analogs, adjuvant leads against latent HIV. *Science,* 320**,** 649-52.
WESELAKE, R. J., TAYLOR, D. C., RAHMAN, M. H., SHAH, S., LAROCHE, A., MCVETTY, P. B. & HARWOOD, J. L. 2009. Increasing the flow of carbon into seed oil. *Biotechnol Adv,* 27**,** 866-78.
WHEELER, T. & VON BRAUN, J. 2013. Climate Change Impacts on Global Food Security. *Science,* 341**,** 508-513.

WORLD BANK 2016. *World Bank Open Data*.
WU, J., LIU, Y. A., TANG, L., ZHANG, F. L. & CHEN, F. 2011. A study on structural features in early flower development of Jatropha curcas L. and the classification of its inflorescences. *African Journal of Agricultural Research,* 6**,** 275-284.
WU, P., ZHOU, C., CHENG, S., WU, Z., LU, W., HAN, J., CHEN, Y., CHEN, Y., NI, P., WANG, Y., XU, X., HUANG, Y., SONG, C., WANG, Z., SHI, N., ZHANG, X., FANG, X., YANG, Q., JIANG, H., CHEN, Y., LI, M., WANG, Y., CHEN, F., WANG, J. & WU, G. 2015. Integrated genome sequence and linkage map of physic nut (Jatropha curcas L.), a biodiesel plant. *Plant J,* 81**,** 810-21.
WU, P. Z., ZHANG, S., ZHANG, L., CHEN, Y. P., LI, M. R., JIANG, H. W. & WU, G. J. 2013. Functional characterization of two microsomal fatty acid desaturases from Jatropha curcas L. *Journal of Plant Physiology,* 170**,** 1360-1366.
WURSCHUM, T. 2012. Mapping QTL for agronomic traits in breeding populations. *Theor Appl Genet,* 125**,** 201-10.
XU, J., CARLSSON, A. S., FRANCIS, T., ZHANG, M., HOFFMAN, T., GIBLIN, M. E. & TAYLOR, D. C. 2012. Triacylglycerol synthesis by PDAT1 in the absence of DGAT1 activity is dependent on re-acylation of LPC by LPCAT2. *BMC Plant Biol,* 12**,** 4.
YE, J., QU, J., BUI, H. T. & CHUA, N. H. 2009. Rapid analysis of Jatropha curcas gene functions by virus-induced gene silencing. *Plant Biotechnol J,* 7**,** 964-76.
YEN, C. L., STONE, S. J., KOLIWAD, S., HARRIS, C. & FARESE, R. V., JR. 2008. Thematic review series: glycerolipids. DGAT enzymes and triacylglycerol biosynthesis. *J Lipid Res,* 49**,** 2283-301.
YI, C., ZHANG, S., LIU, X., BUI, H. T. & HONG, Y. 2010. Does epigenetic polymorphism contribute to phenotypic variances in Jatropha curcas L.? *BMC Plant Biol,* 10**,** 259.
YU, N., XIAO, W. F., ZHU, J., CHEN, X. Y. & PENG, C. C. 2015. The Jatropha curcas KASIII gene alters fatty acid composition of seeds in Arabidopsis thaliana. *Biologia Plantarum,* 59**,** 773-782.
YUE, G. H., LO, L. C., SUN, F., CAO, S. Y., YI, C. X., HONG, Y. & SUN, W. B. 2014. No Variation at 29 Microsatellites in the Genome of Jatropha curcas. *J Genomics,* 2**,** 59-63.
YUE, G. H., SUN, F. & LIU, P. 2013. Status of molecular breeding for improving Jatropha curcas and biodiesel. *Renewable & Sustainable Energy Reviews,* 26**,** 332-343.
ZAMIR, D. 2001. Improving plant breeding with exotic genetic libraries. *Nat Rev Genet,* 2**,** 983-9.
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. *Theor Appl Genet,* 121**,** 1071-82.
ZHANG, Y. C., YU, Y., WANG, C. Y., LI, Z. Y., LIU, Q., XU, J., LIAO, J. Y., WANG, X. J., QU, L. H., CHEN, F., XIN, P., YAN, C., CHU, J., LI, H. Q. & CHEN, Y. Q. 2013. Overexpression of microRNA OsmiR397 improves rice yield by increasing grain size and promoting panicle branching. *Nat Biotechnol,* 31**,** 848-52.
ZHAO L, K. V., LI F, HAUGHN GW, KUNST L. 2010. Insertional mutant analysis reveals that long-chain acyl-CoA synthetase 1 (LACS1), but not LACS8, functionally overlaps with LACS9 in Arabidopsis seed oil biosynthesis. *Plant Journal,* 64**,** 1048-58.
ZHENG, P., ALLEN, W. B., ROESLER, K., WILLIAMS, M. E., ZHANG, S., LI, J., GLASSMAN, K., RANCH, J., NUBEL, D., SOLAWETZ, W., BHATTRAMAKKI, D., LLACA, V., DESCHAMPS, S., ZHONG, G. Y., TARCZYNSKI, M. C. & SHEN, B. 2008. A phenylalanine in DGAT is a key determinant of oil content and composition in maize. *Nat Genet,* 40**,** 367-72.