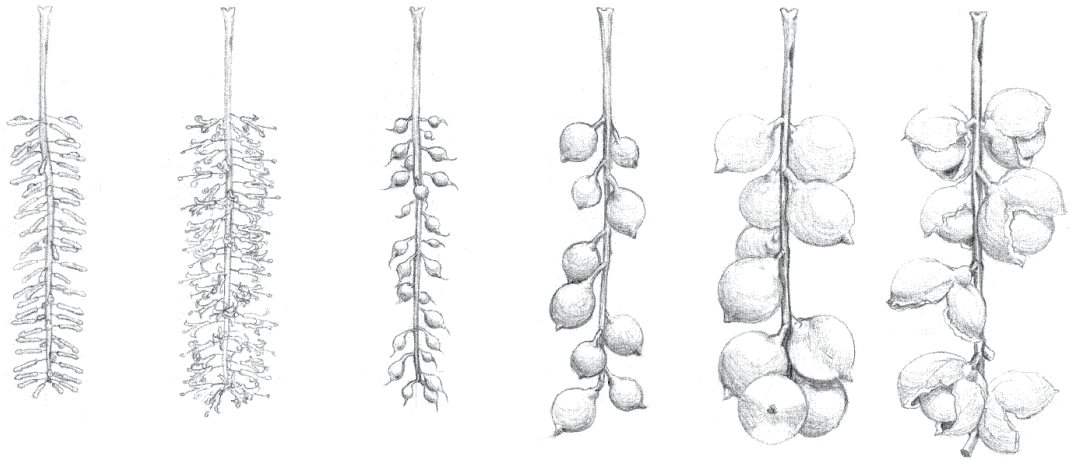
THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Selection strategies to improve yield in macadamia using component traits and genomics

Katie Merryn O'Connor

Bachelor of Environmental Science, First Class Honours



A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2019

Centre for Horticultural Science

Queensland Alliance for Agriculture and Food Innovation

Abstract

Macadamias (*Macadamia integrifolia*, *M. tetraphylla*, and their hybrids) are native to the east coast of Australia, and are grown commercially around the world for their high quality edible nut kernel. Breeding new cultivars for high nut yield is a lengthy and laborious process that can take over two decades. Evaluations are time-consuming due to the long juvenile period of four or more years, and the low correlation between young and mature tree yield means that at least eight years of evaluations are required. Furthermore, yield has low heritability, and the large tree size necessitates low field planting density, and thus increases land use and evaluation costs. It is hypothesised that genetic gain for yield may be increased, compared to traditional breeding approaches, through the use of strategies including (i) indirect selection using yield component traits, (ii) marker-assisted selection and (iii) genomic selection.

This thesis employed an experimental population of 295 seedling progeny and their 29 parents, at four sites across south-east Queensland, that were genotyped for 4,113 SNPs and 16,171 silicoDArT markers detected using Diversity Arrays Technology (DArT) methods. Population structure, genetic diversity and linkage disequilibrium (LD) between SNP markers was quantified to inform subsequent genomics analysis. LD decay was initially rapid at short distances (r^2 = 0.124 for SNPs within 1 kb of each other), but low level LD persisted for long distances. The seedling population was relatively genetically diverse (H_E = 0.255), and very similar in diversity to that of the 29 parents (0.250). Furthermore, progeny with *M. integrifolia* x *M. tetraphylla* ancestry were more genetically diverse (H_E = 0.278) than *M. integrifolia* seedlings (H_E = 0.189). Progeny were moderately differentiated and clustered into three distinct groups, which represented *M. integrifolia* germplasm and two hybrid groups.

Flowering and nut characteristics, tree growth and yield were measured on each tree. Estimations of trait heritability and genetic correlations between each component trait and yield were used to calculate selection efficiency of indirectly selecting for yield using component traits. Kernel recovery, an economically important trait, had high heritability (h^2 = 0.76) but was negatively genetically correlated with yield (r_g = -0.27). Trunk circumference correlated strongly with yield (r_g = 0.72) and was moderately heritable (h^2 = 0.44); however, a breeding aim is to reduce tree size without compromising yield. No component traits were appropriate or effective for indirectly selecting for high yield.

A genome-wide association study (GWAS) was used to identify genetic markers associated with component traits that were genetically correlated with yield or are considered economically important. SNP markers were significantly associated (after correction for false positives) with nut weight (n = 7), percentage of whole kernels (n = 4), and trunk circumference (n = 44). Multiple regression analysis found that some markers were detecting the same QTLs, and, thus, rendered some redundant. These significantly associated markers could be used for marker-assisted selection (MAS) at the seedling stage to identify trees with desirable nut characteristics prior to fruiting.

Genomic selection (GS) used markers across the whole genome to predict estimated breeding values for yield and yield stability of individuals. Predictions with models using four years of yield data were more accurate (r = 0.12 to 0.94) than those using only one or two years of data (-0.29 to 0.39). Predictions in related families were more accurate (r = 0.57) than in unrelated population predictions (r = 0.22) using four years of yield data, confirming previously reported results in other crops that using GS models will be most beneficial when the target population is closely related to the training population. Predicted genetic gain of yield for related family predictions (421–438 g/year, at 2.5% selection intensity) was more than double that for traditional breeding (202 g/year), and was low for unrelated population predictions (12 to 162 g/year). A comparison of selection strategies indicated that assessing thousands of seedlings in the nursery using genomics-assisted breeding is unachievable due to the current costs of genotyping, though this may decrease in the future. A more feasible approach could be to screen thousands of seedlings for precocity (early flowering) and kernel recovery in the field, and reduce costs by only genotyping the elite individuals to predict yield using GS.

Estimates of heritability and correlations will inform the ease and ability to breed and select for key component traits. MAS could predict phenotypes for yield component traits based on allelic states at key markers, whilst GS could be used to predict yield and yield stability, though this will depend on genotyping costs. Future research requires validation in a separate population, with more individuals and SNP markers to increase accuracy and improve certainty of results. Early identification of elite germplasm would reduce time and labour involved in evaluating progeny, and increase genetic gain by decreasing selection cycle time.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Katie Merryn O'Connor

15th August 2019

Publications during candidature

Peer-reviewed papers

O'Connor, K., Hardner, C., Alam, M., Hayes, B., and Topp, B. (2018). Variation in floral and growth traits in a macadamia breeding population. In: R. Drew (ed.), International Symposia on Tropical and Temperate Horticulture ISTTH2016 (Cairns, Queensland). Acta Horticulturae 1205(77): 623-630. https://doi.org/10.17660/ActaHortic.2018.1205.77

O'Connor, K., Hayes, B., Topp, B. (2018) Prospects for increasing yield in macadamia using component traits and genomics. Tree Genetics & Genomes 14(1): Article 7. https://doi.org/10.1007/s11295-017-1221-1

O'Connor, K., Kilian, A., Hayes, B., Hardner, C., Nock, C., Baten, A., Alam, M., and Topp, B. (2019) Population structure, genetic diversity and linkage disequilibrium in a macadamia breeding population using SNP and silicoDArT markers. Tree Genetics & Genomes 15(2): Article 24. https://doi.org/10.1007/s11295-019-1331-z

O'Connor, K., Hayes, B., Hardner, C., Alam, M., and Topp, B. (2019) Selecting for nut characteristics in macadamia using a genome-wide association study. HortScience 54(4): 629-632. https://doi.org/10.21273/HORTSCI13297-18

Submitted manuscripts included in this thesis

No manuscripts submitted for publication

Other publications during candidature

Peer-reviewed papers

Alam, M., Neal, J., **O'Connor, K.**, Kilian, A., and Topp, B. (2018). Ultra-high-throughput DArTseq-based silicoDArT and SNP markers for genomic studies in macadamia. PLOS One 13, e0203465. https://doi.org/10.1371/journal.pone.0203465

Alam, M., Hardner, C., Nock, C., **O'Connor, K.**, and Topp, B. (2018). Historical and molecular evidence of genetic identity of macadamia cultivars HAES741 and HAES660. HortScience. 54(4): 616-620. https://doi.org/10.21273/HORTSCI13318-18

Book chapter

Topp, B., Nock, C., Hardner, C., Alam, M., and **O'Connor, K.**, Advances in Plant Breeding Strategies. Volume 4: Nut and Industrial Crops. J.M. Al-Khayri, S.M. Jain and D.V. Johnson (Editors). Springer. *Accepted.*

Conference presentations

Selecting for yield component traits in macadamia using a genome-wide association study. International Macadamia Research Symposium. Hawaii, USA. September 2017.

Using DNA markers to predict yield and nut characteristics in macadamia. Australian Macadamia Industry Conference. Gold Coast, Queensland. November 2018.

Conference abstracts and posters

O'Connor, K., Hayes, B., Hardner, C., Alam, M., Kilian, A., Topp, B. Prospects for genomic selection in macadamia, an outcrossing perennial rainforest tree. V International Conference on Quantitative Genetics ICQG5. Wisconsin, USA, June 2016.

O'Connor, K., Hardner, C., Hayes, B., Russell, D., Alam, M., Topp, B. Exploring component traits to identify high yield potential in the Australian macadamia breeding program. II International Symposium on Tropical and Temperate Horticulture, Cairns, Queensland. November 2016.

O'Connor, K., Hayes, B., Alam, M., Topp, B. Identifying markers associated with disease-harbouring stick-tights in macadamia. International Tropical Agriculture Conference (TropAg). Brisbane, Queensland. November 2017.

O'Connor, K., Hayes, B., Hardner, C., Henry, R., Alam, M., Topp, B. Genomic prediction for nut yield in a macadamia breeding population using GBLUP. Queensland Alliance for Agriculture and Food Innovation Research Meeting. Brisbane, Queensland. November 2018.

Publications included in this thesis

O'Connor, K., Hayes, B., Topp, B. (2018) Prospects for increasing yield in macadamia using component traits and genomics. Tree Genetics & Genomes 14(1): Article 7. https://doi.org/10.1007/s11295-017-1221-1

A review article, incorporated as Chapter 1

Contributor	Statement of contribution
Katie O'Connor	Conception of study (20%) Wrote paper (100%) Edited paper (60%)
Ben Hayes	Conception of study (20%) Edited paper (20%)
Bruce Topp	Conception of study (60%) Edited paper (20%)

O'Connor, K., Kilian, A., Hayes, B., Hardner, C., Nock, C., Baten, A., Alam, M., and Topp, B. (2019) Population structure, genetic diversity and linkage disequilibrium in a macadamia breeding population using SNP and silicoDArT markers. Tree Genetics & Genomes 15(2): Article 24. https://doi.org/10.1007/s11295-019-1331-z

Incorporated as Chapter 2

Contributor	**Statement of contribution**
Katie O'Connor	Conception of study (50%) Statistical analysis (95%) Interpretation of results (65%) Wrote paper (95%) Edited paper (60%)
Andrzej Kilian	Provided genotyping services with Diversity Arrays Technology Pty Ltd Wrote paper (5%) (genotyping methods only)
Ben Hayes	Statistical analysis (5%) Interpretation of results (20%) Edited paper (10%)
Craig Hardner	Interpretation of results (5%) Edited paper (10%)
Mobashwer Alam	Interpretation of results (5%) Edited paper (10%)

Contributor	Statement of contribution
Bruce Topp	Interpretation of results (5%) Conception of study (60%) Edited paper (10%)

Contributions by others to the thesis

Contributor	Statement of contribution
Craig Hardner (and team at CSIRO)	Generated the experimental B1.2 progeny population of the Australian macadamia industry breeding program Supervised collection of phenotypic data from young trees (age 5 to 6) Advised and assisted in genetic modelling in this study
Andrzej Kilian (and Diversity Arrays Technology team)	Sequencing of genetic markers, and writing of detailed sequencing methods
Ben Hayes	Assisted in the selection of progeny families to use in the study Advised and assisted in genetic modelling in this study
Catherine Nock Abdul Baten	Involved in constructing the v2 macadamia genome assembly, of which unpublished scaffolds were used in the study
Cameron McConchie Bruce Topp Dougal Russell Jodi Neal Rod Daley (at Queensland Department of Agriculture and Fisheries)	Supervision and collection of phenotypic data from young trees (ages 7 and 8)

Statement of parts of the thesis submitted to qualify for the award of another degree

No works submitted towards another degree have been included in this thesis.

Research involving human or animal subjects

No animal or human participants were involved in this research.

Acknowledgements

Firstly, I would like to extend my sincere gratitude to my advisors Bruce Topp, Mobashwer Alam, Craig Hardner, Ben Hayes and Robert Henry for their generous input of ideas, guidance and support. Thanks to Bruce and Mo for being a solid point of contact throughout the project, having an "open-door" policy for both small questions and large discussions, and for their ongoing encouragement and keeping me calm when I felt overwhelmed. Thank you to Craig and Ben for countless emails back and forth discussing theory, analyses and code – I have learned so much.

This project would not have been possible without the many people who helped me with my fieldwork. I thank Rachel Abel for her ideas, wisdom, logic and problem solving, which helped make my fieldwork much less stressful and more smooth-sailing. Huge thanks to Jasmine Nunn, Codie Murphy, Leanne Bridges and Sonia Slegers for their laughter and stories during tiresome fieldwork. Thanks also to everyone else who helped me with field and lab work, including Eddie Howell, Dougal Russell, Rod Daley, Dale McKenna, Thuy Mai, and others. I am sincerely grateful for the cooperation for land-owners, farmers and managers: Gary and Sue Kelly, Adrian Walsh, Clayton Mazziati, Mark Finlay, Ian McConachie, and, in particular, Les Gain for his knowledgeable on-farm insights and discussions.

Many thanks to Mark Dieters and Jodi Neal for their ideas and encouragement, to Andrzej Killian and the team at Diversity Arrays Technology for teaching me the processes involved in genotyping, to Bob Mayer for advice with statistical analyses, and everyone at Maroochy Research Facility for making our workplace a great place to be. Thanks also to José and Lynnette Chaparro for making me feel welcome in their home during my trip to the USA.

For their patience, support and (sometimes-feigned) interest during my studies, I thank my family and friends: Dad, Lauren, Alise, Todd, Bec, Dan, and Maddy. I am indebted to my husband, Nik, for his unwavering support and belief in me, as well as providing me with a steady supply of chocolate and cups of tea. One of my biggest supporters was my aunty Kerry who passed away due to illness during my PhD. I would like to take this opportunity to acknowledge everyone who helped me get through this very painful time, and to say that there is no weakness in talking about mental health. Please reach out for help if you need it, and look after those around you.

Financial support

This research has been funded by Hort Innovation, using the Macadamia research and development levy and contributions from the Australian Government. Hort Innovation is the grower owned, not-for-profit research and development corporation for Australian horticulture.

This research was supported by an Australian Government Research Training Program Scholarship and the Charles Morphett Peglar Scholarship.

Keywords

macadamia, horticulture, tree, genomic selection, genome-wide association study, components of yield, yield, Diversity Arrays Technology

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 070602, Horticultural Crop Improvement (Selection and Breeding), 50%

ANZSRC code: 060408, Genomics, 30%

ANZSRC code: 060412, Quantitative Genetics (incl. Disease and Trait Mapping Genetics), 20%

Fields of Research (FoR) Classification

FoR code: 0706, Horticultural Production, 50%

FoR code: 0604, Genetics, 50%

Table of Contents

List of Tables

List of Figures

List of Abbreviations

°C	degrees Celsius
ΔG	genetic gain
AFLP	amplified fragment length polymorphism
AL	Alloway, Queensland, Australia
AM	Amamoor, Queensland, Australia
AMOVA	analysis of molecular variance
BLAST	basic local alignment search tool
BLUP	best linear unbiased prediction
bp	base pair
BV	breeding value
cm	centimetres
CR	call rate
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CV	cross-validation
cv.	cultivar
DArT	Diversity Arrays Technology
DNA	deoxyribonucleic acid
DNIS	dry nut-in-shell
E	selection efficiency
EG	East Gympie, Queensland, Australia
ENF	estimated number of florets per raceme
F5	number of florets per 5 cm at terminal end of raceme
FDR	false discovery rate
FSN	flowers that set nuts
g	grams
GBLUP	genomic best linear unbiased prediction
GCV	genetic coefficient of variation
GEBV	genomic estimated breeding value
GRM	genomic relationship matrix
GS	genomic selection

GWAS	genome-wide association study
G x E	genotype by environment interaction
H^2	broad-sense heritability
h^2	narrow-sense heritability
ha	hectare
HAES	Hawaii Agricultural Experiment Station
HP	Hinkler Park, Queensland, Australia
Indel	insertion and deletion
kb	kilobase
kg	kilogram
KR	kernel recovery
KW	kernel weight
L	generation length in years
LD	linkage disequilibrium
LMM	linear mixed models
MAF	minor allele frequency
MAS	marker-assisted selection
Mb	megabase
MCMC	Markov chain Monte Carlo
min	minutes
mm	millimetre
mt	megaton
NCBI	National Centre for Biotechnology Information
NIS	nut-in-shell
NPR	number of nuts per rachis
NW	nut weight
NS	non-significant
PCoA	principal coordinates analysis
PCR	polymerase chain reaction
PD	nut pedicel diameter
PIC	polymorphic information content
PPCA	probabilistic principal components analysis

QPMR	quality with scaling per million reads
QQ	quantile-quantile
QTL	quantitative trait loci
RAF	randomly amplified DNA fingerprinting
RAMiFi	randomly amplified microsatellite fingerprinting
RAPD	random amplified polymorphic DNA
RDN	rachis diameter at nut set
r_g	genetic correlation
RL	raceme length
r_p	phenotypic correlation
RSN	racemes surviving from flowering to nut set
RVT	regional variety trial
SD	standard deviation
s.e.	standard error
sec	seconds
SMC	simple matching coefficient
SNP	single nucleotide polymorphism
SPT	seedling progeny trial
sq	square root transformed
SSR	simple sequence repeat
SVD	single value decomposition
TC	trunk circumference
WK	whole kernels
WNIS	wet nut-in-shell

For Ker

Chapter 1. General Introduction and Literature Review

The following chapter has been published as a review article.

O'Connor, K., Hayes, B., Topp, B. (2018) Prospects for increasing yield in macadamia using component traits and genomics. Tree Genetics & Genomes 14(1): Article 7. https://doi.org/10.1007/s11295-017-1221-1

The chapter has been updated to reflect recently published research in the field.

My contribution to the publication

I wrote, edited and performed calculations included in the chapter.

1.1 Abstract

Selection of candidate cultivars in macadamia requires extensive phenotypic measurements over many years and trials. In particular, yield traits such as nut-in-shell yield and kernel yield are economically vital characteristics and therefore guide the selection process for new cultivars. However, these traits can only be measured in mature trees, resulting in long generation intervals and slow rates of genetic gain. In addition, these traits are expensive to measure. Strategies to reduce the generation interval and increase the intensity of selection include using yield component traits, identification of markers associated with component traits, and genomic selection for yield. Yield component traits that contribute to resource availability for fruit formation include floral and nut characteristics. In this review, these traits will be investigated to estimate their relative importance in macadamia breeding, their heritability and correlations with yield. Furthermore, the usefulness of genome-wide association studies regarding yield component traits will be reviewed. Genetic-based breeding techniques could exploit this information to increase yield gains per breeding cycle and estimate the quantitative nature of yield traits. Genomic selection uses genome-wide molecular markers to predict the phenotype of individuals at an early age before maturity, thereby reducing the cycle time and increasing gain per unit time in plant breeding programs. This review evaluates the potential for measurement of yield component traits, genome-wide association studies and genomic selection to be employed in the Australian macadamia breeding program to accelerate gains for nut yield.

1.2 Introduction

In the past few decades, substantial increases in yield have resulted from genetic improvement in many crops, including maize (Duvick 1984) and apple (Igarashi et al. 2016). However, genetic improvement is still in its infancy in many tree species due to their long generation times and the cost of screening new cultivars (Cros et al. 2015; Isik 2014; Kumar et al. 2013a; van Nocker and Gardiner 2014; Khan and Korban 2012). High yield is often the focus in crop breeding programs, yet selection gains can be hindered since yield is commonly difficult to select due to its complex nature. The process of yield genetic gain in fruit tree crops can be accelerated in a number of ways.

One method of improving yield is by mining for yield component traits. Component traits that are correlated with yield, and are more heritable and easier to measure, may be used to indirectly select for high yield (Fraser and Eaton 1983; Sparnaaij and Bos 1993; Piepho 1995). This indirect selection may increase breeding gains by reducing cycle times if the component traits are measured earlier in the process than yield.

Other methods to increase yield in crop breeding programs include employing DNA-based technologies. This includes combining genome-wide association studies (GWAS) with marker-assisted selection (MAS), and using genomic selection (GS) (Varshney et al. 2005; Isik et al. 2015; Lande and Thompson 1990; Khan and Korban 2012; Endresen 2010; van Nocker and Gardiner 2014). GWAS can help identify genetic markers associated with key yield component traits, which can then be screened for in a population, and elite candidates selected using MAS. GS can be used to select for the more complex trait yield by modelling genetic markers across the genome and their effect on the trait to predict the yield of each candidate.

Luby and Shaw (2000) proposed that fruit crops have more to gain from MAS than annual crops due to their large tree size and long generation times, and the time and cost involved in maintaining the trees. However, they recognised that this may be true only if the trait in question is simply inherited, is economically important, and is conventionally very expensive to measure (Luby and Shaw 2000). Since that time, the technology of molecular markers has dramatically expanded and advanced. Genomics-based methods for improving the efficiency of breeding programs such as GWAS and GS are now particularly pertinent for fruit trees (Iwata et al. 2016; Kumar et al. 2012b; Yamamoto and Terakami 2016; Wong and Bernardo 2008; Peace 2017). These methods have advanced from the more fundamental marker-assisted breeding and trait mapping, have higher accuracies and wider applications (Iwata et al. 2016), and have potential use in breeding for increased yield in a crop such as macadamia.

This review investigates genomic improvement in crop breeding, with specific reference to fruit and nut tree crops including macadamia. The potential use of yield component traits, GWAS and GS in improving yield in macadamia will be explored.

1.3 Macadamia: a native Australian nut

Macadamia (Proteaceae) is a subtropical rainforest tree, native to the east coast of Australia between Mount Bauple, Queensland, and Lismore, New South Wales (Gross 1995; Hardner et al. 2009). The genus contains four species: *Macadamia integrifolia, M. tetraphylla, M. ternifolia,* and *M. jansenii* (Hardner et al. 2009; Peace et al. 2008). Individual trees are produced predominantly from outcrossing, similar to many other rainforest species, are large in size, and have a long juvenile period (Figure 1-1; Sedgley et al. 1990; Trueman and Turnbull 1994). Both *M. integrifolia* and *M. tetraphylla* and their hybrids are cultivated around the world for their edible nuts (Figure 1-2; Hardner et al. 2009).

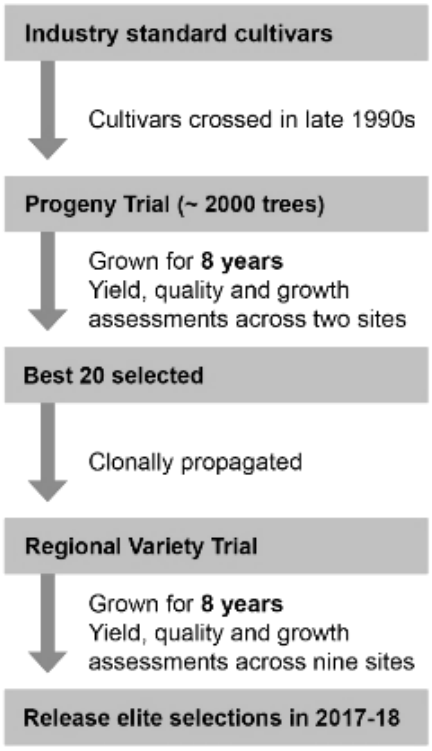


Figure 1-1 Timeline of the Australian macadamia breeding program's first generation, showing evaluation steps indicative of traditional breeding practices

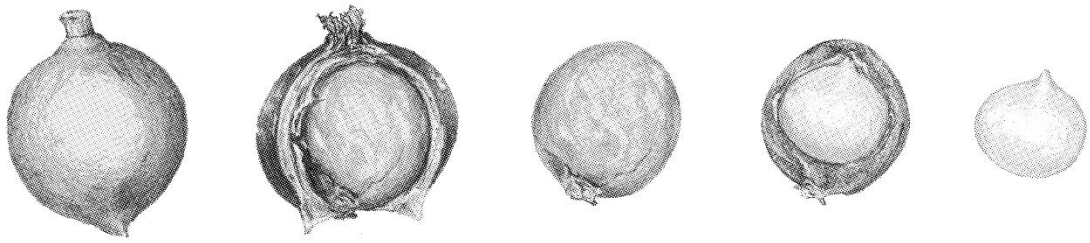


Figure 1-2 Macadamia nuts – the edible kernel is enclosed in a hard, woody shell and the outer husk. Left to right: nut in husk, split husk, nut in shell, cracked shell, and kernel. Illustration by Todd Fox

Macadamias are diploid (2n = 28), highly heterozygous, with genome size estimates ranging from 652 Mb (Nock et al. 2016) to 780 Mb (Chagné 2015). A draft genome assembly of short-read Illumina sequences from cultivar 'HAES 741' covers 79% of the total estimated genome, at 518 Mb in length (Nock et al. 2016). Nock et al. (2016) discussed ongoing work to improve genome coverage by incorporating deeper, long-read PacBio sequence data, and develop a high-density linkage map, which will be advantageous for future genomics studies. The Australian National Macadamia Germplasm Collection contains accessions across all four Macadamia species, which is also available as a genomics resource for sequencing (Hardner et al. 2004).

1.3.1 Domestication and cultivation

Domestication of macadamia is only relatively recent, with cultivation beginning in the late-1800s (Hardner et al. 2009; Peace et al. 2008; Hardner 2016). Two early importations of *M. integrifolia* nuts from Australia to Hawaii occurred in the 1880s and 1890s (Hamilton and Fukunaga 1959). It has been suggested that the first exports originated from near Mount Bauple (Hardner 2016). Planting of seedlings began around 1920 by the Hawaii Agricultural Experiment Station (HAES), whilst evaluation and selection of new cultivars commenced in the mid-1930s (Hamilton and Fukunaga 1959; Hardner 2016). This program provided the majority of commercial cultivars currently grown around the world (Hardner 2016). In Australia, the first orchards were established near Lismore, New South Wales in the 1880s and in Queensland in 1910 (Hardner et al. 2009). As such, cultivated varieties are only a few

generations removed from their wild relatives. Macadamias are mainly produced in Australia, South Africa, USA (Hawaii), and Kenya (Australian Macadamia Society 2012).

Trees begin to bear nuts after four to five years, are fully mature after ten to 15 years, and can be commercially productive for up to 60 years (Hardner et al. 2009). In 1997, the Commonwealth Scientific and Industrial Research Organisation (CSIRO) launched an Australian macadamia breeding program from which subsequent selections have been made for parental crossing and regional variety trials (RVTs) to evaluate elite selections (Figure 1-1; Hardner et al. 2002). Large areas are required over various environments to evaluate new cultivars. RVTs in macadamias are usually maintained for eight years from planting date, with many traits measured each year (Hardner et al. 2002). Several years of data are required in order to select high-yielding cultivars (Hardner et al. 2001). Yield and growth data up to year eight are available for ~2,000 trees in Australia from these trials.

High yield is the primary trait used to select new cultivars (Howlett et al. 2015; Hardner et al. 2009; Stephenson et al. 1986). Other important traits selected in the Australian breeding program are high kernel recovery, small tree size, and high proportion of intact kernels (Topp et al. 2012). Nut-in-shell (NIS) yield refers to the weight of de-husked nuts at 1% moisture content with the shell intact (Hardner et al. 2002). Kernel recovery (KR) is the ratio of kernel to nut mass (Hardner et al. 2002; Kester and Asay 1975); high KR is desired as this indicates that the kernel is relatively large compared with the weight of the shell. However, cultivars with high KR have thin shells, which are susceptible to pests and diseases (Hardner et al. 2009). Depending on the use of the product, whole unbroken kernels may be desirable, so this is also an important trait (Hardner et al. 2009; O'Hare et al. 2004). Industry standards for these traits are: 5 t/ha NIS of >18 mm diameter, >36% KR, 2-3 g kernels, and >50% whole kernels (O'Hare et al. 2004); however production can fall short of these standards.

1.4 Gain from selection

Gain from selection efforts in breeding can be predicted using the formula:

$$R = \frac{h^2 S}{Y}$$ Equation 1-1

where R is response per year, or genetic gain; h^2 is narrow-sense heritability, and also a function of trait measurement accuracy; S is the selection differential, the amount by which the average parents' performance exceeds the average breeding population performance; and Y is selection cycle length in years (Hansche 1983). In macadamia, genetic gain is impeded by a long breeding cycle with the current breeding program requiring eight years to evaluate NIS yield (Figure 1-1). Reducing the selection cycle is thus an important aim for macadamia breeding. Selection intensity and accuracy are costly to improve due to the large plant size, which adds to the cost of increased population size and replication. The following sections address these factors with reference to improvements in macadamia.

1.5 Yield and its component traits

Yield is the highest priority in macadamia breeding; yield was consistently ranked by the industry as the most important future cultivar characteristic (O'Hare and Topp 2010), and was economically weighted highest of all traits in the selection index used in Australian macadamia breeding (Hardner et al. 2006). Macadamia yield can be quantified as wet nut-in-husk, wet nut-in-shell, nut-in-shell dried to 1% kernel moisture, and expressed on a per tree or per hectare basis (Hardner et al. 2009). However, selecting for yield in the breeding program is difficult as it is a complex trait (variation is the result of small effects at many loci). Hardner et al. (2002) found that broad-sense heritability for annual NIS yield ranged from 0.06 to 0.18, whilst cumulative NIS ranged from 0.11 to 0.20 (Table 1-1).

Table 1-1 Heritability and correlations between various flower and fruit characteristics in macadamia and other nut crops. r_g, genetic correlation; r_p, phenotypic correlation; H^2, broad-sense heritability; h^2, narrow-sense heritability

Crop	**Trait/s**	**Heritability**	**Correlation between traits**	**Source**
Macadamia *Macadamia integrifolia* and *M. tetraphylla*	Annual nut-in-shell yield (10 yrs) Cumulative nut-in-shell yield (10 yrs) Kernel recovery and cumulative kernel yield (10 yrs) Stem girth and cumulative nut-in-shell yield (10 yrs) Nut weight Kernel weight Kernel recovery Nut weight and kernel weight Kernel recovery and kernel mass Kernel recovery and shell thickness	$H^2 = 0.14$ $H^2 = 0.17$ $H^2 = 0.63$ $H^2 = 0.63$ $H^2 = 0.63$ 	 $r_g = 0.08$, $r_p = 0.12$ $r_g = 0.22$, $r_p = 0.59$ $r_g = 0.79$, $r_p = 0.68$ $r_g = 0.48$, $r_p = 0.49$ $r_p = -0.70$	Hardner et al. 2002 Hardner et al. 2001 Leverington, 1962
Pecan *Carya illinoinensis*	Nut weight Kernel weight Kernel recovery and kernel weight Nut yield (kg/tree) Kernel recovery Kernel recovery and kernel weight	$h^2 = 0.35$ $h^2 = 0.38$ $H^2 = 0.855$ $H^2 = 0.897$ 	 $r_p = 0.394$ r = 0.569	Thompson and Baker, 1993 Kumar et al., 2013a
Hazelnut *Carylus avellana*	Kernel weight Kernel recovery Relative husk length	$h^2 = 0.67$ $h^2 = 0.87$ $h^2 = 0.91$		Yao and Mehlenbacher, 2000
Cashew *Anacardium occidentale*	Tree nut yield and whole nut weight Tree nut yield and number nuts per panicle Tree nut yield and number hermaphrodite flowers per panicle		$r_p = 0.108$ $r_p = 0.844$ $r_p = 0.863$	Aliyu, 2006
Walnut *Juglans regia*	Crop yield Nut weight Crop yield and nut weight	$h^2 = 0.07$ $h^2 = 0.86$ 	 $r_p = -0.20$	Hansche et al. 1972

Yield or other complex traits may be indirectly selected through correlated component traits that are more heritable (Fraser and Eaton 1983; Sparnaaij and Bos 1993; Piepho 1995). It is best to initially explore simple component traits related to yield that may be easier and/or cheaper to measure (Sparnaaij and Bos 1993). The investigation of component traits can reduce cycle times if they can be measured earlier in the tree's life, and selection intensity can be increased if the traits are efficiently measured, allowing evaluation of a larger number of plants. This is particularly true when the trees are young, as less land and fewer resources are required. However, Fraser and Eaton (1983) noted that in broadacre and horticultural crops it may be ineffective to rely on component traits correlated with the complex target trait as many components are often linked. Other sequential and path analyses have been proposed to overcome this difficulty (e.g. Thomas and Grafius 1976; Li 1975; Sparnaaij and Bos 1993; Piepho 1995; Eaton and Kyte 1978).

It is important to recognise the relationship between different traits and how they affect yield (Samonte et al. 1998). Understanding genetic parameters such as heritability and correlations between various traits can help select parents in breeding programs (Falconer 1989; Bodzon 2004). In *Prunus*, traits affecting fruit quality such as flavour, colour and shape are often related (Cantín et al. 2010; de Souza et al. 1998). Correlations of component traits and yield in macadamia and other nut crops are presented in Table 1-1.

There are many components of yield in macadamia, some of which have been evaluated and others that need further exploration (Hardner et al. 2009). This review focuses on those factors that affect flower and nut development, and hence resource utilisation for the nuts. Further research is needed to understand the different components of yield in macadamia and to identify important related traits that can easily be measured.

1.5.1 Flowering and growth traits

Flowering plays a critical role in fruit production, and so it is necessary to understand the factors affecting flower development (Westwood 1993). A review of flowering and fruiting in macadamia was conducted by Trueman (2013). The flowers are initiated on inflorescences called pendant racemes, varying from 6-30 cm in length (Huett 2004; Figure 1-3). A mature tree can produce about 2,500 racemes, with 100-300 flowers (florets) on each raceme.

Macadamia flowers are pollinated predominantly by native stingless bees and European honeybees (Trueman 2013; Howlett et al. 2015).

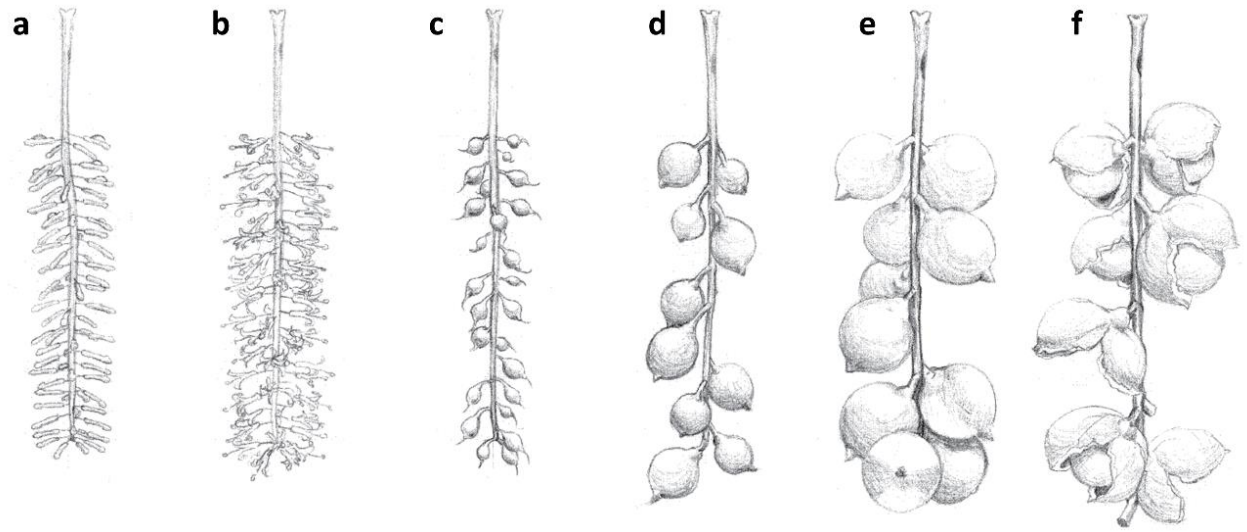


Figure 1-3 Stages of flower and nut development on a raceme in macadamia. **a** Developing florets, with looping stage shown near base of raceme. **b** Anthesis. **c** Initial nut set, with fewer nutlets than florets. **d** Developing nuts, fewer than previous stage. **e** Nuts in husk at full size. **f** Nuts dehisce from husk and fall to ground. Illustration by Todd Fox

Macadamia is generally self-incompatible through mechanisms including protandry, though there is evidence of self-compatibility in some cultivars (Urata 1954; Sedgley et al. 1990; Sedgley et al. 1985). Self-incompatibility in plants can be controlled by several multi-allelic genes acting at different stages of flower development (Seavey and Bawa 1986; de Nettancourt 1977). For example, self-fertility in almond (*Prunus dulcis*) is controlled by a major gene, operating in a quantitative manner (Kester and Asay 1975). Sedgley et al. (1990) found that several macadamia cultivars (predominantly *M. integrifolia*) presented inferior pollen tube growth from self-pollen compared with outcrossed pollen, as well as lower fruit set. *Macadamia tetraphylla* also showed some self-compatibility, though again, cross-pollen produced higher seed set per raceme (Pisanu et al. 2009). Self-compatibility should be investigated in various genotypes to identify if it is a heritable trait in macadamia, as this may be a target for breeding and selection to increase pollination success. Furthermore, the

relationship between self-fertility and nut yield could be a focus of research in macadamia genotypes to determine if inbreeding level affects seed set.

In Australia, flower development occurs from May to October. Bud initiation begins in May, followed by bud dormancy (50-96 days), raceme and floret elongation, style elongation and looping, and anthesis (Moncur et al. 1985). Fertilisation occurs one week after anthesis; however, most flowers abscise in the following two weeks. Fruits develop and some premature fruit drop occurs; nuts are mature about 28 weeks after anthesis (Nagao and Sakai 1990). Research is required to investigate the heritability of raceme and floret characteristics, and their correlation with yield.

Several studies have investigated the relationship between yield and flowering with variable results. Since the racemes have many florets, there are many opportunities for nuts to be set; floret number does not appear to be a limiting factor. Ito (1980) stated that about 0.3% of the flowers develop into mature, saleable nuts. However, this analysis was based on estimates of the numbers of racemes and flowers and of yield, and not replicated measurements. Further, Urata (1954) argued that it was unreasonable to count the number of flowers per length of raceme due to the low percentage of flowers setting nuts and the low correlation between the two characters.

Trueman and Turnbull (1994) found that number of flowers per raceme affected initial fruit set in different pairs of cross-pollinated cultivars. Both cv. ‘H2’ and cv. ‘333’ racemes bearing 200 flowers when cross-pollinated with cv. ‘246’ had higher initial fruit set (22.3% and 40.3%, respectively), than control racemes (9.1% and 24.0%). For cv. ‘660’, racemes with 50 flowers set more fruits (21.6%) than those with 200 flowers (15.9%) when cross-pollinated with cv. ‘344’. In comparison, number of fruits per raceme at final nut set increased with number of flowers per raceme for cv. ‘660’. Trueman and Turnbull (1994) also found that for cv. ‘660’ nut and kernel fresh weights as well as KR were higher in cross-pollinated (with cv. ‘333’ and ‘246’) fruits than control racemes. These results demonstrate significant variation in pollination success between macadamia cultivars. Further research is required across many genotypes to determine if raceme and flower production and fruits per raceme limit yield in macadamia.

Xylem and phloem transport water, nutrients and photosynthates throughout plants (Campbell and Reece 2002), and thus the size of these vessels may influence the growth of

limbs and fruit. Hardner et al. (2002) found that cumulative NIS yield up to year ten was positively phenotypically correlated with girth of the trunk stem (0.59) in 40 cultivars (Table 1-1). The rachis (raceme stem) in macadamia enlarges after anthesis, and is wider in inflorescences with high numbers of nuts than in inflorescences with low numbers of nuts (Urata 1954). Fruiting wood with larger diameters produced larger fruits in two out of six peach *(Prunus persica*) cultivars (Porter et al. 2002). In oil palm (*Elaeis guineensis*), leaf area was highly correlated with oil yield due to photosynthetic ability and biomass production (Bai et al. 2018). The inheritance of the diameter of tree trunk, raceme stems and fruit pedicels should be investigated in macadamia, along with the relationship between yield and these characteristics.

1.5.2 Fruit characteristics

Nut development and abscission affect yield and profitability in macadamia (Boyton and Hardner 2002). After pollination, the pollen tube grows down to the ovary and fertilises one of the two ovules. Occasionally both ovules are fertilised, resulting in twin nuts (Sedgley 1981). The immature nut expands and oil accumulates from 80 to 165 days after anthesis (Trueman et al. 2000; McConchie et al. 1996). Nuts can drop from just after fertilisation to when they are mature (Boyton and Hardner 2002). Nuts are mature when they reach their maximum oil content, about 135 to 165 days after anthesis, or from February to March in Queensland and New South Wales, Australia (McConchie et al. 1996); commercial nut drop can continue through to September, depending on the cultivar (Boyton and Hardner 2002; McConchie et al. 1997). The husk can dehisce (split open) on the tree and the nut-in-shell falls to the ground (Figure 1-3), or the husk may fall with the shell (Hardner et al. 2009). It is economically advantageous to have cultivars in which the nut-in-husk abscises from the tree. This results in higher yield recovery due to the improved mechanical harvesting efficiency and reduced carry-over due to disease-harbouring stick-tight husks.

Saleable kernel yield in macadamia is related to several component traits. Parents and their progeny are usually selected for NIS yield and KR. The nuts consist of an edible kernel enclosed by the shell, a woody testa, and husk (an outer pericarp) (Hardner et al. 2009; Figure 1-2). Nut size is also an important yield component trait in almond: larger nuts correspond with higher yields per acre (Kester and Asay 1975). Topp et al. (2012) suggested selecting for high KR after

four or five years as an indirect indication of future precocity. Precocity is a desirable trait in macadamia (Hardner et al. 2009), though as found in the cultivar 'Ikaika' (also known as '333') precociousness may mean lower yields at later ages compared with other cultivars (Hamilton and Ito 1984).

It may be useful to investigate the partitioning of the tree's resources in husk, shell and kernel. Harvest index was initially described by Donald (1962) regarding the ratio of economic yield to total biomass in grain crops. Cannell (1985) proposed that harvest index in perennial fruit trees should relate to the ratio of harvested fruit to total aboveground dry biomass. However, as only a portion of the macadamia nut is edible, then perhaps an index based on the kernel rather than the nut-in-shell and husk should be used. As much as 30% of the moisture in a macadamia nut may be in the husk (Rosengarten 2004). Husk hardness, which influences the level of pest damage (Hardner et al. 2009), differs between cultivars (Campbell et al. 2005). It is not known whether the size of the husk affects yield. It is of interest to investigate if energy used by the tree in producing husk occurs at the expense of kernel production.

Bazzaz et al. (1987) suggested that perennial plants may not invest as much energy into reproduction as annuals as they have more opportunities to reproduce and can allocate resources to other activities such as defence. Flowering intensity has been inconsistently correlated with reserves of carbohydrates in macadamia trees (McFadyen et al. 2012). Carbohydrate resources may be depleted during flowering and fruit development, meaning that fruit set is negatively affected (Stephenson et al. 1989; Wilkie 2010).

Previous studies have reported correlations between different components of yield. Nut and kernel weight were strongly correlated in macadamia ($r_g = 0.79$, $r_p = 0.68$; Table 1-1) (Hardner et al. 2001; Peace 2005), and KR decreased significantly with increased shell thickness ($r_p = -0.70$) (Leverington 1962). Kernel recovery and kernel mass were moderately correlated ($r_g = 0.48$, $p < 0.05$; $r_p = 0.49$) in different cultivars (Hardner et al. 2001). Hansche et al. (1972) found that walnut crop decreased with increased nut weight ($r_p = -0.20$), after adjusting for year effect. In other species like cashew nut (*Anacardium occidentale*) yield per tree was highly correlated with both number of nuts per panicle ($r_p = 0.844$, $p < 0.01$) and number of hermaphrodite flowers per panicle ($r_p = 0.863$, $p < 0.01$) (Aliyu 2006). In pecan (*Carya illinoinensis*), Thompson and Baker (1993) found a moderately low phenotypic correlation ($r_p = 0.394$, $p < 0.002$) between KR and kernel weight, whilst Kumar et al. (2013a) found a high

correlation (r = 0.569). The differences between these pecan studies may be due to alternate fruit bearing in the crop, differences in the study populations or year of data collection (Thompson and Baker 1993; Kumar et al. 2013a).

No information is available on the link between macadamia yield and raceme length or number of nuts per cluster. However, fruit set per raceme varied in different cultivars (McConchie et al. 1997; Boyton and Hardner 2002). As previously mentioned, the heritability of NIS yield per tree in macadamia is low. Broad-sense heritability based on individual trees ranged from 0.06 to 0.18 for annual NIS, 0.11 to 0.20 for cumulative NIS, and 0.11 to 0.21 for cumulative kernel yield, between four and ten years after planting, respectively (Hardner et al. 2002; Table 1-1). Quantification of these traits in a wider group of genotypes will be beneficial for the breeding program.

Nut size and ratio of edible nut to shell are important breeding factors in nut tree crops, and estimates of their heritability is vital. Estimates of broad-sense heritability include total genetic variance, whereas narrow-sense heritability involves only additive genetic variance and, thus, gives an indication of the ease of measuring the trait and determines the response to selection to improve the trait in breeding (Falconer 1989). Nut weight, kernel weight and KR in macadamia were all found to have the same broad-sense heritability of 0.63 by Hardner et al. (2001; Table 1-1). These characteristics were also measured in 152 pecan families (Thompson and Baker 1993). Pecans are selected for thin shells and high KR, similar to macadamia. Estimates of narrow-sense heritability for nut and kernel weight in pecan were 0.35 and 0.38, respectively. In contrast, in hazelnut (*Corylus avellana*), kernel weight, KR and relative husk length (husk:nut length) were highly heritable (Yao and Mehlenbacher 2000; Table 1-1). Walnut (*Juglans regia)* nut weight was also very highly heritable, though crop heritability was extremely low (Hansche et al. 1972; Table 1-1). Kumar et al. (2013a) selected nuts from 34 pecan selections and three standard cultivars and found that broad-sense heritability for nut yield and KR was extremely high (>0.85; Table 1-1) compared with macadamias (0.14). However, this may have been due to favourable environmental conditions during the study, rather than genetic influence in pecan (Kumar et al. 2013a).

Genetic gain for yield may be hastened by selecting for yield component traits instead of selection for yield per se. However, indirectly selecting for high yield through component traits depends on a number of factors. Firstly, success depends on the genetic variance of the

component trait and its heritability, which also encompasses the accuracy of measuring the trait. Traits with high heritability will be more easily bred and selected for than traits with low heritability. Component traits should be highly correlated with yield and more easily and cheaply measured than yield (Sparnaaij and Bos 1993). The relative efficiency of indirect selection on a trait X via direct selection for trait Y depends on the ratio of correlated response (CR_X) to direct response (R_X) (Falconer 1989):

$$\frac{CR_X}{R_X} = \frac{i_Y h_Y r_A \sigma_{AX}}{i_X h_X \sigma_{AX}} \qquad \text{Equation 1-2}$$

Or, if the selection intensities are the same, more simply $h_y r_A / h_x$.

Hardner et al. (2001, 2002) estimated heritability for kernel mass (H^2 = 0.66) and cumulative kernel yield to ten years (H^2 = 0.14) with a genetic correlation between these traits of 0.30. Thus, using Equation 1-2 above, for this population the ratio of the two responses was 0.65, indicating that indirect selection using kernel mass was only 65% as efficient as direct selection for yield. A genetic correlation of >0.46 would be needed for indirect selection of kernel mass to be more efficient than direct selection for yield. These estimates were from a population of clonally propagated elite selections and cultivars. Genetic estimates are required from segregating progeny populations to allow conclusions of the use of indirect selection in stage one of breeding. More combinations of yield and component traits should be investigated to determine if correlated response to selection is promising in macadamia populations.

Genetic gain will also be affected by the stage at which the component trait can be measured and assessed in the tree: if it can be measured when the trees are juvenile then costs may be reduced by elimination of inferior individuals prior to expensive field evaluations. Breeders selecting for early traits focus on secondary yield traits that are highly heritable (Borghi et al. 1998), where replication is not needed to accurately measure the trait (Topp et al. 2012). Additionally, trees must flower before crosses can be made to produce the next generation of seedlings, which further restricts selection cycles. Therefore, selection for component traits should be coupled with selecting for early flowering (Huett 2004).

1.6 Using genomic information to accelerate genetic gains in tree crops

1.6.1 Identifying target genes through genome-wide association studies followed by marker-assisted selection

Economically important traits in fruit trees such as yield and quality are likely to be controlled by several multi-allelic genes or a very large number of genes (Iwata et al. 2016; Khan and Korban 2012). Genome regions identified through linkage (family-based) mapping as being associated, or in linkage disequilibrium (LD), with the target or component traits are called quantitative trait loci (QTL) (Iwata et al. 2016). These QTLs can be used to predict individuals with high breeding values, which can be used in marker-assisted selection (MAS) (Iwata et al. 2016; Myles et al. 2009; Lynch and Walsh 1998; Muranty et al. 2014; Hayes and Goddard 2010). Breeding values (BVs) are the sum of the mean additive effects of all alleles in an individual (Heffner et al. 2009).

Since QTLs can be thousands of kilobases in length, multiple genes may be closely linked with the target gene (Khan and Korban 2012). Linkage drag may occur with adverse results for the breeding program as a result of undesirable traits positioned in proximity to desired ones (Khan and Korban 2012). As such, a more directed approach is desirable for capturing significant genes using genomic methods (Savolainen and Pyhäjärvi 2007).

Genome-wide association studies (GWAS) can utilise the allelic state of unrelated individuals to detect markers linked with target traits including the broader germplasm pool, rather than using family-based methods such as biparental controlled crosses, which can be impractical, laborious and expensive (Iwata et al. 2016; Myles et al. 2009). Association mapping in the form of GWAS offers a more fine-scale approach than QTLs to identify smaller, individual markers in LD with target traits. This can overcome the detrimental effects of genetic drag as the marker intervals are shorter, as well as accounting for population structure (Khan and Korban 2012; Brachi et al. 2011; Isik 2014; Myles et al. 2009; Rikkerink et al. 2007). The incorporation of population structure and kinship information can reduce a major problem of false associations between markers and phenotypes in a GWAS (Khan and Korban 2012; Brachi et al. 2011; Iwata et al. 2016).

In GWAS, each marker is tested individually for an association with the trait (Hayes and Goddard 2010; Khan and Korban 2012; Huang and Han 2014). This process relies on markers

being in LD with the genes controlling the trait (Balding 2006), and very few markers are located within that causative locus itself (Hayes and Goddard 2010). Single nucleotide polymorphisms (SNPs) are now a commonly used genetic marker for these studies, where there is a variation in the base at a location in the genome which can be compared among individuals (Hayes and Goddard 2010; Huang and Han 2014).

Recent advancements have led to high-throughput and high-density genotyping at lowered costs per marker point for use in genomic analysis (Iwata et al. 2016). Next-generation sequencing technologies, such as genotyping by sequencing, can detect molecular markers for use in genomics studies like GWAS (He et al. 2014). Marker order along the genome is not strictly required, so association studies and some other genomics studies can be performed without a reference genome. This is particularly useful in novel species (Iwata et al. 2016).

GWAS for quantitative traits such as yield have shown that often there are many markers that influence the corresponding phenotype, and these can have a minor or major effect (Lee et al. 2008; Khan and Korban 2012). The gain from MAS is proportional to the variance of the trait captured by the markers and the significance of the association (Collard et al. 2005). As such, MAS has little value for traits that are complex (that is, affected by a large number of mutations, all of small effect); MAS is more effective for monogenic or oligogenic traits (Hayes and Goddard 2010; Luby and Shaw 2000; Huang and Han 2014). Screening for target markers using GWAS and MAS can occur before the plants flower as only DNA is needed for the selection, and thus can substantially reduce the selection cycle and increase genetic gain by eliminating undesirable genotypes early (van Nocker and Gardiner 2014).

'DNA-informed breeding', a term coined by Peace (2017), is becoming the convention driving breeding direction in Rosaceae crops in the USA. Previously, however, Ru et al. (2015) reviewed the opportunities and constraints of using MAS in Rosaceae breeding. They found that MAS was not yet widely applied in fruit trees, but that affordable and program-specific testing of DNA for major trait loci at the seedling stage could be effective for the adoption of the technique. There are relatively few published studies employing GWAS in fruit trees. A study by Minamikawa et al. (2017), investigating fruit quality traits in 676 citrus individuals using 1,841 SNPs, found that correlated traits were controlled by several common SNPs. GWAS also detected markers significantly associated with six fruit quality traits in Japanese pear (*Pyrus pyrifolia*), with higher significance achieved when parent and seedling populations

were combined (Minamikawa et al. 2018). In apple (*Malus* x *domestica*), Kumar et al. (2013b) found significant associations in six fruit quality traits using 2,500 SNPs across 1,200 seedlings. SNP markers with the largest effect across linkage groups individually explained only 2% of the phenotypic variation for fruit firmness and 17% for red-flesh, which was reasonably low, yet substantially more than that explained by pedigree-based analysis in many other traits (Kumar et al. 2013b). Kumar et al. (2013b) also found two genomic regions that were linked with two pairs of fruit quality traits, suggesting a pleiotropic effect.

GWAS can also be employed to indirectly detect high yields through other secondary traits not associated with fruit quality, if the two traits are highly correlated. In oil palm, Bai et al. (2018) identified two potential QTLs for leaf area, a trait which is highly correlated with oil yield. Thus, breeders could indirectly select trees with high oil yield through MAS by identifying trees with large leaf area and reducing the laborious and destructive phenotyping previously required to measure oil yield (Bai et al. 2018).

Further studies have employed GWAS utilising genetic markers other than SNPs. For example, Iwata et al. (2013) and Cao et al. (2012) investigated fruit quality in Japanese pear and peach, respectively, using simple sequence repeat (SSR) markers. Iwata et al. (2013) detected significant associations between markers and resistance to black spot disease, spur number and harvest time, which indicated links to major QTLs, despite the small scale of the study in terms of number of markers (n = 162) and cultivars (n = 76). Using 53 SSR markers distributed across linkage groups, Cao et al. (2012) found that the significantly associated markers detected for peach fruit quality were located nearby previously known QTLs. Between 8.1–14.5% of the variation in red flesh pigment was explained by four SSRs. Similar to the findings of Kumar et al. (2013b) in apple, two of the pigment markers were associated with two other sets of traits: ripening time and fruit development period, and fruit weight and flowering time (Cao et al. 2012).

A review of breeding progress in tree nut crops by Mehlenbacher (2002), including efforts of trait mapping and MAS, concluded that genetic improvement is limited by small breeding program size. Hardner et al. (2005) evaluated the potential for MAS to improve macadamia specifically, stating that SSR markers and pedigree will be useful in detecting marker-trait associations. However, the technology in the genomics field has vastly increased and improved since then; many more markers can now be screened at a lower cost. GWAS and

MAS regarding important component traits of yield, such as nut and kernel weights and KR, using SNP markers appears to be feasible to improve macadamia if the traits are controlled by few genes of moderate to large effect. Determining the number of markers and their effects for these traits should be the focus of future genomics studies, as this is currently unreported.

1.6.2 Yield prediction using genomic selection

Genomic selection (GS) uses genome-wide markers to capture the effects of loci that affect the target trait (Meuwissen et al. 2001). GS is best when markers such as SNPs are in high LD with genes of large effect, hence capturing a large proportion of genetic variance (Druet et al. 2014; Goddard 1991; Viana et al. 2016). A two-step process is involved. In a reference (or training) population, where individuals have both genome-wide marker genotypes and target trait phenotypes available, the effect of all markers on the trait are estimated simultaneously. The effect of each marker is used to establish a prediction equation. The equation can then be used to predict genomic estimated breeding values (GEBV) for genotyped selection candidates, likely to be seedlings or young trees. The accuracy of GS is assessed with cross-validation of the predicted GEBV against the known and accurate phenotypes in a validation (or testing) population (Meuwissen et al. 2001).

Many simulations suggest that GS is superior to MAS and traditional phenotypic selection for complex traits (Heffner et al. 2009; Bernardo and Yu 2007; Grattapaglia and Resende 2011; Iwata et al. 2011). This is because MAS only uses markers significantly associated with a target trait, yet many yield and quality traits are often controlled by numerous minor-effect genes (Iwata et al. 2016; Kumar et al. 2012a; Jannink et al. 2010). In comparison, GS utilises all available genetic markers with no significance threshold, and can therefore explain more of the genetic variability than MAS (Viana et al. 2016; Meuwissen et al. 2001). Thus, GS avoids marker effect biases and produces more highly correlated measured and predicted BVs (Meuwissen et al. 2001; Heffner et al. 2009).

GS can increase genetic gain in horticulture crops by accelerating breeding cycles (Jannink et al. 2010; Desta and Ortiz 2014; Meuwissen et al. 2001; Heffner et al. 2010). By selecting potential elite juvenile individuals, filtering candidates and only proceeding to the field with

potentially high-performing trees, time, cost and labour can be reduced. However, yield and other traits still need to be assessed over several locations before cultivars are recommended (Acquaah 2012). Selection of potential superior cultivars in the juvenile stage can also drastically reduce capital and maintenance costs (Rikkerink et al. 2007; Luby and Shaw 2000). This would be useful in macadamia where the trees do not reach full nut production until they are at least eight years old (Hardner et al. 2009). Iwata et al. (2016) and Namkoong et al. (2005) recognised that a combination of traditional and marker selection strategies should be employed.

Denis and Bouvet (2013) concluded that perennial crops may have more to gain from GS than annual crops since genetic gain per unit time in perennial crops is critical for improved cultivars. There is a paucity of published studies for tree nut crops, though there has been some work conducted in citrus (Minamikawa et al. 2017), apple (Kumar et al. 2012b), oil palm (Wong and Bernardo 2008; Kwong et al. 2017a), and pear (Iwata et al. 2013; Minamikawa et al. 2018). In a recent study of fruit quality traits in Japanese pear, model accuracy was highest when data for parents and progeny were combined, at $r > 0.7$ for most traits (Minamikawa et al. 2018). High ($r > 0.7$) prediction accuracies were also obtained for six fruit quality traits in citrus (Minamikawa et al. 2017). They found that some model accuracies were trait dependent, but the GBLUP (genomic best linear unbiased prediction) model was the highest for most traits and was more accurate in predictions than MAS based on significant SNPs. Kumar et al. (2012b) investigated the use of GS in improving fruit quality traits in apple. They used 2,500 high quality SNPs for 1,120 seedlings, and model predictions for fruit quality were high, at 0.70 to 0.90 (Kumar et al. 2012b).

Wong and Bernardo (2008) demonstrated that gain per unit cost and time can be increased in oil palm through GS. Costs for GS ranged from US$75,000 to $194,000 per unit gain, depending on cost per marker data point, population size, QTL number and heritability, compared with US$116,000 to $333,000 per unit gain for 19 years of phenotypic selection per cycle. Also in oil palm, Kwong et al. (2017a) found that for 1,218 individuals genotyped using a 200K array, GS model accuracy increased with trait heritability, ranging from 0.40 to 0.70. The results of these studies may be applicable to other tree species with long generation intervals and large planting areas.

Pear has a long juvenile period and needs to be evaluated over many years and thus Iwata et al. (2013) found the crop could benefit from GS. They investigated nine disease resistance and fruit set traits in 76 Japanese pear cultivars using 162 markers, mostly SSRs. Prediction of GEBVs was moderately high for flesh firmness and fruit weight (r = 0.60 and 0.53, respectively). They found that using all makers, rather than just those with significant associations as identified using GWAS, was more accurate (Iwata et al. 2013). In comparison, predictions of BVs in citrus for fruit weight and other fruit quality traits were high (>0.7) across 106 cultivars (Minamikawa et al. 2016). A corresponding GWAS detected the influence of major QTLs in all citrus fruit traits, the use of all markers was more accurate than using only significant SNPs (Minamikawa et al. 2016).

Given the large amount of phenotypic data available for macadamia, and documented parentage since the first domesticated cultivars, this genus is a strong candidate for GS. Macadamia has a selection cycle of 22 years and typical planting densities of 312 trees/ha (Topp et al. 2012), and would benefit greatly from this technology. Potentially, the first stage of progeny testing (Figure 1-1) could be substantially reduced by genotyping seedlings, applying GS models and only continuing further evaluations with those individuals predicted to be high yielding.

The size and structure of the reference population, relationship between training and testing populations, choice of model, marker number and density, heritability of key traits and LD span need to be assessed when employing GS in breeding. These have been considered in reviews conducted by Grattapaglia (2014) and Lin et al. (2014) on forestry and annual species. The accuracy of models can decline over subsequent generations, so it is necessary to recalibrate every few generations with new phenotypes and allelic frequencies (Viana et al. 2016; Goddard 2009; Grattapaglia et al. 2018).

The training population needs to be sufficiently large to enable accurate estimation of small effects across many loci, and to capture all the genetic variation present in the breeding program (Meuwissen et al. 2001). Ideally, the training and testing populations should be related or part of the same breeding program for best results (Habier et al. 2007). Kumar et al. (2012b) divided 1,120 apple seedlings into two groups for their GS evaluations: 90% of individuals for the training population and 10% for validation. In their simulations of GS in oil palm, Wong and Bernardo (2008) used small population sizes of 30 to 70 individuals.

Macadamia has a relatively small breeding population available for genomic studies; the mean number of seedlings per family in the Australian macadamia breeding program's "B1.2" population is 14 (n = 1,961) (Topp et al. 2016). However, almost half of these trees have been removed from the field and are no longer available for genomic analysis. A larger second generation progeny population (n ≈ 4,000) is currently being phenotyped and will be available for further genomic selection evaluation (B. Topp, pers. comm.).

A number of prediction models have been developed for GS: BLUP (best linear unbiased prediction), GBLUP (genomic BLUP), ridge regression, Lasso, reproducing kernel Hilbert space, and various Bayesian regressions (Jannink et al. 2010; Heslot et al. 2012). Heslot et al. (2012) suggested using a reduced set of models for implementing GS in breeding. These include a faster version of BayesB called weighted Bayesian shrinkage regression, Bayesian Lasso and random forest. It is possible to combine different models to improve predictions; however, Heslot et al. (2012) found that combining different models did not always improve the accuracy of predictions. Cross-validation is an essential step in GS to identify the accuracy of the model or to compare different models (Heslot et al. 2012; Crossa et al. 2010). Models such as GBLUP and BayesR (Erbe et al. 2012) are sound candidates for use in macadamia; assuming respectively, a normal distribution of SNP effects using a genomic relationship matrix among candidates (GBLUP), and allowing for a small number of moderate to large effect QTL (BayesR). Testing both strategies is advisable given the paucity of data regarding the genetic nature of macadamia yield traits.

Models are affected by effective population size, heritability of the trait and the size of the reference population (Daetwyler et al. 2008; Hayes et al. 2009a; Goddard 2009). Effective population size is calculated using marker information and population heterozygosity; genetic gains are greater in species with smaller effective populations (Goddard 2009). Thus, the genetic diversity of the crop must be determined before genomics studies can begin. GEBVs are more accurately predicted when the trait is highly heritable (Hayes et al. 2009a). Therefore, it is important to understand the heritability of the characteristics and its component traits. The accuracy of predicting BVs increases with the size of the reference population (Hayes et al. 2009a), showing the importance of the training set in GS.

Accurate phenotyping is critical for GS; if the accuracy of phenotyping is poor, many more individuals will be needed in the reference population (Desta and Ortiz 2014). Accurate

phenotyping requires multiple well-characterised environments, stringent selection criteria and large training populations (Resende et al. 2012; Xu and Crouch 2008; Desta and Ortiz 2014; Rikkerink et al. 2007). For many traits such as yield, data need to be collected across multiple years and sites, and is costly and time consuming (Resende et al. 2012; Bernardo 2008; Xu et al. 2012; Isik 2014; Stephens et al. 2009). This has been the case in macadamia where 2,000 progeny from 47 families have been evaluated for eight years at nine locations (Topp et al. 2016).

To implement GS in macadamia would involve growing progeny to their first leaf for DNA extraction and subsequent genotyping, hence reducing the labour and maintenance costs of growing trees to maturity. Topp et al. (2012) compared capital, maintenance and evaluation costs for four breeding strategies, including traditional assessment (evaluation of 1,200 hybrid seedlings for nine years followed by RVT for a total cycle length of 22 years). They found that full traditional assessment was much more expensive ($1,545,922 net present value) with a low ratio of gain to breeding cost ($570,000) compared to tandem selection, where seedlings are evaluated to age seven only ($795,508 and $1,080,000). Their cloned seedling strategy, which evaluates 200 hybrid seedlings in an RVT after only two years of initial measurements, also reduced the cycle length to 15 years, with improved breeding costs and gain to cost ratio than traditional assessment ($986,075 and $680,000) (Topp et al. 2012).

Genotyping costs continue to decline, with more data points becoming available, and therefore more markers likely to be in proximity to causal genes (Heffner et al. 2009; Khan and Korban 2012; Iwata et al. 2016). The cost of genotyping analysis varies with the volume of sequencing applied per sample, with the most popular services applying between 1 million to 5 million reads, and the price per sample varying usually between US$25 to US$55 (A. Killian, Diversity Arrays Technology Pty Ltd, pers. comm.). Thus, with advancing technology, the accessibility of large numbers of molecular markers and the declining costs, the employment of and opportunity to use GS in breeding is increasing (Heffner et al. 2009; Iwata et al. 2016). Future macadamia breeding efforts should compare the costs and benefits of traditional breeding with selection strategies involving GWAS and GS.

1.7 Conclusions

The complex nature of yield in macadamia and its low heritability, as well as long cycle times, currently hinder cultivar development. Genetic improvement of yield by indirect selection for its component traits may improve breeding efficiency. Characteristics such as nut, kernel and husk weight, raceme length and number of florets, may be more simply and accurately measured than yield in breeding populations, especially in the years before yield per tree estimates are stable. Yield component traits can be investigated with GWAS to determine if any major markers are associated with each trait. If so, this information could be used in MAS. GS models are suitable for predicting complex traits like yield in macadamia seedlings, as well as to predict important related traits. It is essential to compare the genetic gains and the costs using these different breeding strategies, to determine which method or combination of methods are most efficient.

1.8 Research objectives

This chapter demonstrates the difficulty in selecting for high yield in macadamia, mainly owing to long plant juvenility, large tree size, multiple years required to assess for key traits, and the absence of rapid and efficient selection methods. In this thesis, it is hypothesised that genetic gain for yield may be increased relative to conventional breeding approaches through the use of yield component traits and/or genomics-assisted breeding. As such, the overarching goal of this study was to investigate alternative selection strategies for high yield with comparison to traditional breeding methods. Specific objectives of the project were to:

1. Determine the level of genetic diversity within and among a subset of families and parents in the Australian macadamia breeding program, and define the population structure to inform subsequent chapters and analyses (Chapter 2).
2. Determine any yield component traits that have high heritability and are genetically correlated with yield that could be used to indirectly select for high yield (Chapter 3).
3. Perform GWAS to identify markers significantly associated with yield component traits, and determine the location of significant markers on genome assembly scaffolds (Chapter 4).

4. Determine the accuracy of genomic selection methods in predicting yield and yield stability over years, and identify strategies in which genomic selection can be employed to increase genetic gain in these traits (Chapter 5).

Chapter 2. Population structure, genetic diversity and linkage disequilibrium in a macadamia breeding population using SNP and silicoDArT markers

---

The following chapter is presented as it has been published.

O'Connor, K., Kilian, A., Hayes, B., Hardner, C., Nock, C., Baten, A., Alam, M., Topp, B. (2019) Population structure, genetic diversity and linkage disequilibrium in a macadamia breeding population using SNP and silicoDArT markers. Tree Genetics & Genomes 15(2): Article 24. https://doi.org/10.1007/s11295-019-1331-z

My contribution to the publication

I collected leaf samples, performed all analyses, wrote the paper, and made final revisions. AK wrote some methods, produced genetic data and assisted in interpretation of some results. BH offered ideas for analyses and assisted in calculations. BT, BH, MO, CH and CN were involved in the interpretation of results. CN and AB provided genome assembly scaffold data. All authors were involved in editing the paper.

2.1 Abstract

Macadamia (*Macadamia integrifolia* Maiden & Betche, *M. tetraphylla* L.A.S. Johnson and their hybrids) is grown commercially around the world for its high quality edible kernel. Traditional breeding efforts involve crossing varieties to produce thousands of progeny seedlings for evaluation. Cultivar improvement for nut yield using component traits and genomics are options for macadamia breeding, but accurate knowledge of genetic diversity and structure of the breeding population is required. This study reports allelic diversity within and among families of 295 seedling offspring from 29 parents, population structure, and the extent of linkage disequilibrium (LD) in the population. Genotyping generated 19,527 silicoDArT and 5,329 SNP markers, and after filtering 16,171 silicoDArTs and 4,113 SNPs were used for diversity analyses. LD decay was initially rapid at short distances, but low-level LD persisted for long distances, with an average r^2 = 0.124 for SNPs within 1 kb of each other. The seedling population was relatively genetically diverse, and very similar to that of the 29 parents. The diversity (H_E = 0.255 for progeny and 0.250 for parents) among these individuals indicate the level of diversity at the wider population level in the breeding program, though the population appears less diverse than other fruit crops. Macadamia progeny were moderately differentiated (F_{ST} = 0.401), and formed k = 3 distinct clusters, which represents *M. integrifolia* germplasm separating from two different hybrid groups. There was low to no relationship between heterozygosity and performance for nut yield amongst progeny. These findings will inform future genomics studies of the Australian macadamia breeding program such as genome-wide association studies and genomic selection, where knowledge and control of population structure is vital.

2.2 Introduction

Macadamia is a long-lived, perennial tree of the subtropical rainforest of eastern Australia (Hardner et al. 2009). Two of the four species of the genus, *Macadamia integrifolia* Maiden & Betche and *M. tetraphylla* L.A.S. Johnson, and their hybrids, produce edible kernels and have been used in genetic improvement of the crop (Hardner et al. 2009). The two other species, *M. ternifolia* and *M. jansenii*, have been largely under-utilised in breeding programs to date. The Australian macadamia industry has a farm-gate value of over $250 million, and

produces over 48,000 mt of nuts-in-shell from about 22,000 ha (Australian Macadamia Society 2017b). Macadamia is also grown commercially in Hawaii, South Africa, Kenya and China, and is, hence, an important horticulture crop. While cultivation of macadamia probably commenced not long after European settlement of the natural distribution, development of the crop on a commercial scale began in Hawaii in the mid-1920s (Hardner et al. 2009; Hardner 2016). Development of superior cultivars was undertaken from the mid-1930s with selection from advanced generation open-pollinated families being undertaken from the 1950s. Many of the cultivars currently used for global commercial production originated from these selection efforts (Hardner et al. 2009; Hardner 2016). Due to their long juvenility and the many years needed to breed and assess potential selections, current cultivars are only two to four generations from their wild relatives (Hardner et al. 2009).

Macadamia is diploid (2n = 28), with an estimated genome size of between 652 and 780 Mb (Nock et al. 2016; Chagné 2015). Approximately 80% of the predicted gene models are similar to the sacred lotus (*Nelumbo nucifera*), which has the closest available assembled genome. Previous studies have reported on the genetic diversity of macadamia cultivars using SSR (da Rocha Sobierajski 2012; Schmidt et al. 2006; Nock et al. 2014), isozymes (Aradhya et al. 1998; Vithanage and Winks 1992), AFLP (Steiger et al. 2003), RAPD (Vithanage et al. 1997) and RAF (Peace et al. 2003; Peace et al. 2005) markers. Reviews of previous studies by Gitonga et al. (2009) and Peace et al. (2004) found RAMiFi (randomly amplified microsatellite fingerprinting) and SSR markers to be the most resourceful and cost effective. Genetics and genomics studies now commonly use single nucleotide polymorphisms (SNPs) which are ubiquitous across the genome and can be discovered with ultra-high-throughput and low cost (Gupta et al. 2008). A recent study by Alam et al. (2018) quantified the diversity of 80 macadamia cultivars using 11,526 silicoDArT (presence/absence of restriction fragments in representation) and 3,956 SNP markers, and found that diversity varied widely among the cultivars, and four distinct clusters were observed. The genetic diversity and structure across the whole genome has not yet been measured for a large population representative of the macadamia breeding population.

Plant breeding populations require diversity to develop and improve varieties that possess desired characteristics for farmers and consumers (Govindaraj et al. 2015; Mohammadi and Prasanna 2003; Glaszmann et al. 2010). Over time with subsequent generations, crops

develop complex population structures, and bottlenecks from selecting limited elite individuals can alter the diversity of the population (Flint-Garcia et al. 2003; Glaszmann et al. 2010). It is, therefore, important to quantify the level of genetic diversity amongst modern crops to inform future breeding efforts (Tanksley and McCouch 1997; Glaszmann et al. 2010). Genetic diversity in the form of heterozygosity, kinship and population structure, as well as estimating the extent of linkage disequilibrium (LD) among markers are all vital to inform genomics studies, which are now common in plant breeding research (Khan and Korban 2012; Govindaraj et al. 2015).

Genetic diversity in a crop breeding program can be partially assessed morphologically based on phenotype (Govindaraj et al. 2015) and pedigree (Barrett et al. 1998). Currently, the most widely used method of characterising the genetic diversity in plant breeding populations is through genotypes of molecular markers (Nybom et al. 2014; Govindaraj et al. 2015). Diversity Arrays Technology Pty Ltd (DArT) are an option for high-throughput genotyping of large numbers of markers (Kilian et al. 2012). DArT can provide both bi-allelic codominant DArTseq-based SNPs as well as dominant binary (present vs. absent) silicoDArT markers, and does not require a reference genome (Kilian et al. 2003). Codominant markers such as SNPs and SSRs are generally more informative as an assessment of genetic diversity than dominant markers (Govindaraj et al. 2015). Due to the ability to produce many thousands of markers in microarrays, SNPs and silicoDArTs are now more commonly used than SSRs. For example, SNPs have been utilised for genetic diversity studies in pear (Kumar et al. 2017) and cacao (Ji et al. 2013); silicoDArT markers in citrus (Sagawa et al. 2018) and peanut (Pandey et al. 2014); whilst both have been applied to chickpea breeding (Roorkiwal et al. 2014). Furthermore, the relationship between heterozygosity and fitness is often investigated to determine if more heterozygous individuals have a higher fitness than less heterozygous individuals (Zouros and Foltz 1987).

The present study is the first to analyse the genetic diversity of a large group of macadamia full-sib progenies and their parents, representing a breeding program, using SNP and silicoDArT markers. The objectives of this study were to: (i) determine marker locations on genome assembly scaffolds and calculate the extent of LD, (ii) determine the level of genetic diversity within and among a subset of families and parents in the Australian macadamia

breeding program, (iii) analyse the extent of population structure, to inform future genomics studies, and (iv) determine if heterozygosity influences performance of nut yield.

2.3 Methods

2.3.1 Study design

This study was undertaken on a subset of a population that was part of the of the Australian macadamia breeding program's first generation population (Hardner et al. 2009; Topp et al. 2016). The larger population comprised of a total of 1,961 progeny seedlings from 141 families between crosses of 47 parents designed to maximise diversity. The number of seedlings per family ranged from 1–36, with a mean of 14 (Topp et al. 2016). These seedlings were planted between 1999–2003 in south-east Queensland and north-east New South Wales at nine sites (Topp et al. 2016; Hardner et al. 2009). The study design involved a randomised incomplete block design, with trees planted 4 m apart within rows and 8 m between rows.

Thirty-two families were chosen as the subset for this study based on the family size being approximately ten progeny per family, giving a total of 295 seedling progeny. The families were from crosses between 29 commercially available parents (reciprocal crosses combined); all with at least one parent involved in another cross so there were no isolated families. Many of the parental cultivars are represented by numbers, and here the cultivars from the Hawaiian Agricultural Experiment Station (HAES) breeding program are recorded without 'HAES' before the number. Within each family, five low-yielding and five high-yielding offspring were chosen where available to ensure a range of phenotypes across families. This was based on clonal values of cumulative nut-in-shell yield to age 8 years, derived from pedigree-based BLUPs (best linear unbiased predictions; unpublished). The trees were at four orchard sites in south-eastern Queensland: two in the Bundaberg region (Alloway, AL, and Hinkler Park, HP) and two in the Gympie region (East Gympie, EG, and Amamoor, AM; Figure 2-1).

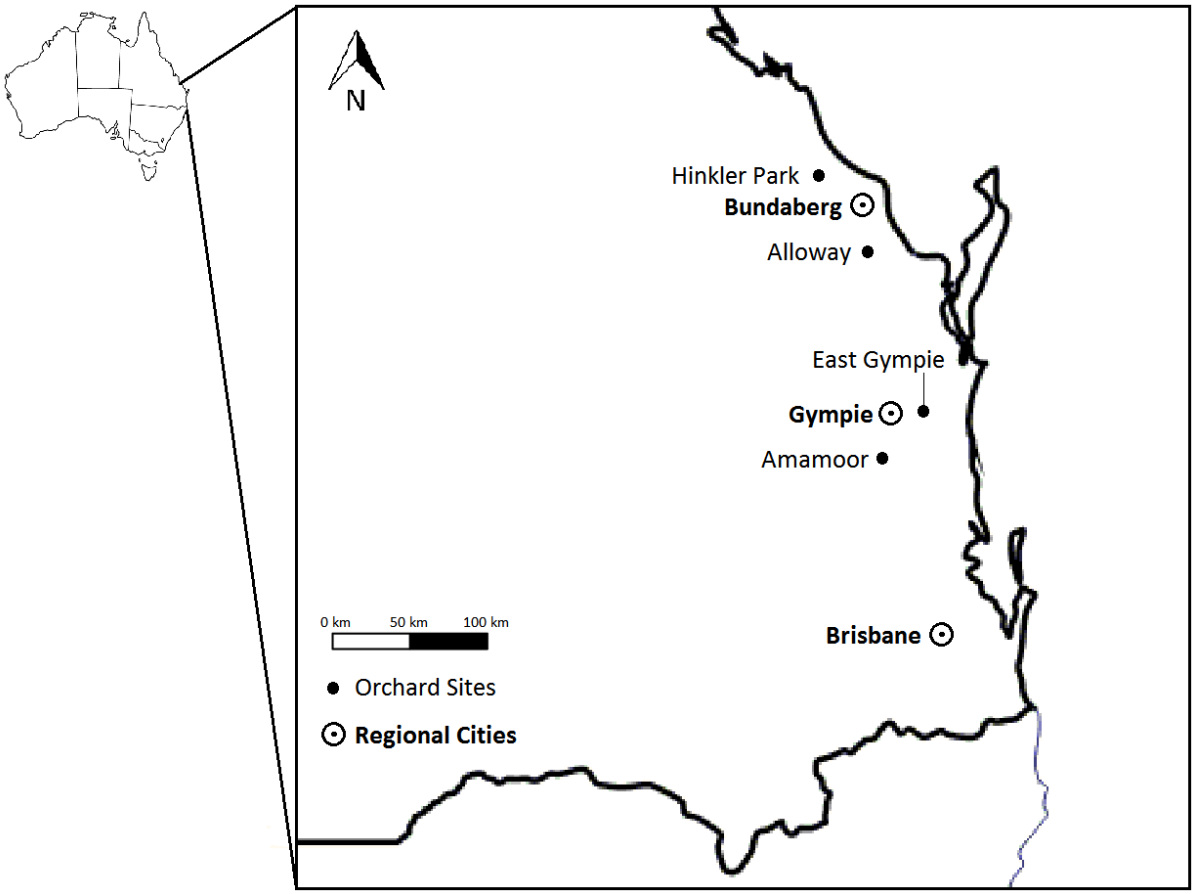


Figure 2-1 Locations of orchard sites and regional cities in south-east Queensland, Australia

2.3.2 DNA extraction and genotyping

Young leaf (where available) was sampled from each tree, stored in a sealed snap-lock bag and kept cold with ice bricks prior to DNA extraction. Samples were also collected from one tree of each parent, and three technical replicates were included. Leaf samples were processed by Diversity Arrays Technology Pty Ltd (DArT, Canberra, Australia). Genomic DNA was extracted after grinding the leaf samples using a QIAGEN Tissue Lyser, following Macherey-Nagel (Germany) protocol. Freedom EVO (Tecan) pipetting robot was used during this process. Eluted DNA samples were incubated with loading dye containing restriction enzyme buffer at 37°C for two hours, then checked on 0.8% agarose electrophoresis gel to determine DNA concentration and evaluate integrity/shearing. Some samples extracted poorly and were, therefore, purified using Zymo purification kit.

Details regarding DArT genotyping methods can be found at http://www.diversityarrays.com/dart-application and in Kilian et al. (2012), however a brief

method is described here. Digestion-ligation reaction was performed using a combination of *Pst*I and *HhaI* restriction enzymes and 4 µL of DNA with the addition of barcoded adaptors at 37°C for two hours, followed by 60°C for 20 min. Twenty-five percent technical replication of library construction was performed using another set of barcoded adapters. For the two Zymo kit purified DNA plates, a concentrated PCR mix was used, with 10 µL of DNA template. 2 µL of the digestion-ligation reaction was used as a template for PCR amplification with the primers containing sequences required for Illumina sequencing (flowcell attachment and sequencing primer sequence). Only "mixed fragments" (*Pst*I-*Hha*I) were effectively amplified in PCR using the following reaction conditions: denaturation at 94°C for 1 min, 30 cycles of 94°C for 20 sec and 58°C for 30 sec, 72°C for 45 sec, and a final extension at 72°C for 7 min. From each library, a 5 µL aliquot was run on 1.2% agarose gel to check library quality. PCR products from each plate were pooled with 10 µL from each sample using the Freedom EVO (Tecan). Replicated samples were pooled with their original sample plates. Pooled amplicons were applied to c-Bot (Illumina) bridge PCR and subsequently sequenced using Illumina Hiseq2500 for 77 cycles.

Sequences generated from each lane were processed using proprietary DArT analytical pipelines. In the primary pipeline, the fastq files were first filtered to remove poor quality sequences, applying more stringent selection criteria to the barcode region compared to the rest of the sequence. As such, the assignments of the sequences to specific samples carried in the "barcode split" step are very reliable. Approximately 2,500,000 (+/- 7%) sequences per barcode/sample were used in marker calling. Finally, identical sequences were collapsed into "fastqcall files". These files were used in the secondary pipeline for DArT Pty Ltd's proprietary SNP and silicoDArT marker calling algorithms (DArTsoft14). SNP markers were detected based on parsing sequence clusters, whilst silicoDArT markers were extracted by detecting the presence/absence of a specific sequence or sequence clusters.

2.3.3 Marker diversity and quality filtering

The data from this study will be used in future genomics studies in the breeding program; however, many genomic analyses cannot be undertaken with data that contains missing calls. Imputation was, therefore, performed on both SNP and silicoDArT markers through DArT's KDCompute Optimal Imputation plugin (http://www.kddart.org/kdcompute.html). Six

methods were compared by excluding an additional 10% of missing values, with a simple matching coefficient (SMC) being calculated from the correlation between introduced missing values and the original dataset. Missing calls were imputed using the most accurate imputation method.

Quality control was applied to both types of markers using pre-imputation read depth, call rate, one ratio, polymorphic information content (PIC), and a test of Mendelian inheritance on a subset of families. SilicoDArT markers with ≥10 average read depth and ≥1.55 QPMR (quality with scaling per million reads, average of normalized non-zero tag read counts divided by the standard deviation of read counts) were used in genetic analysis. Thresholds for SNP markers included ≥50% call rate (proportion of samples scored as "1" present or "0" absent), ≥2.5% one ratio (or minor allele frequency, proportion of samples for which the genotype was scored "1") for both SNP and reference alleles, and >0 PIC for both SNP and reference alleles. PIC was calculated by DArT as $1 - [f^2 + (1 - f)^2]$, where f is the marker frequency in the data set (A. Kilian, pers. comm.). A relatively low call rate threshold was applied because most "missing data" represent null alleles, as the average read depth of macadamia markers was very high (>50 X), and sequence volume was very consistent between the libraries/samples (CV below 10%). Given that most of the SNPs and indels responsible for such null alleles are within 30-40 bp from the called SNP (as average DArTseq fragment size is under 100 bp), the LD between called SNPs and the "secondary variants" responsible for nulls must be very high. We, therefore, anticipate very high efficiency of imputation even for markers with low call rates. A test of Mendelian inconsistencies in the form of opposing homozygotes was conducted in 16 of the families as a representation of the whole study, covering 156 (53%) of the progeny from crosses between 25 (86%) of the parents. Within full-sib families, any SNP markers in which more than two progeny had the opposing homozygote state (e.g. AA) compared with both parents (e.g. BB) were removed from the study for all families. This threshold of two or more progeny per family (18% to 29%, depending on family size) was chosen since greater than a quarter of genotyping errors are indistinguishable from inconsistencies in Mendelian inheritance (Douglas et al. 2002).

Minimum, maximum and mean values were calculated for remaining SNP and silicoDArT markers for the following quality parameters: call rate, one ratio, PIC, and reproducibility

(proportion of technical replicate assay pairs for which the marker score is consistent). Marker distribution of PIC values was graphed for both marker types.

2.3.4 Marker locations

A v2 genome assembly for *M. integrifolia* is available as a collection of 4,098 scaffolds [European Nucleotide Archive (EMBL-ENA) repository, Analysis: ERZ792049, Assembly accession: ERS2953073 (SAMEA5145324)], with a scaffold N50 of 413 kb, and the longest scaffold being 2.19 Mb. The assembly covers an estimated 93% of the genome, and the estimated repeat content is 52% (Nock, Baten *et al*. unpublished data). The tentative locations of the quality-filtered markers on these genome scaffolds were found using the NCBI stand-alone BLAST+ application (Camacho et al. 2009). A threshold of 95% identity matches between SNP and scaffold sequences was applied; top hits (based on E value) of BLAST+ analysis were used for further analyses. A correlation between number of SNPs mapped per scaffold and scaffold length was calculated. Linkage disequilibrium (LD) was estimated using only SNPs that mapped to a single scaffold, due to the repetitive nature of the macadamia genome (Nock et al. 2016). LD was calculated using the r^2 parameter between pairs of SNPs within the same scaffold using Plink v1.9 software (Purcell et al. 2007), with a window of 100 SNP and a lower limit of $r^2 = 0$. The estimated physical distance (in base pairs) was calculated between pairs of SNPs. LD was calculated using all individuals, and progeny and parents separately, and decay was plotted with a log function.

2.3.5 Analysis of genetic diversity and differentiation

Genetic diversity measures were calculated for the 295 progeny, both as one large population and as separate populations according to full-sib families, as well as for the 29 parents. Allelic frequencies were calculated in GenAlEx v6.5 (Peakall and Smouse 2012) using SNPs and the codominant option. Allelic frequencies were used to determine genetic diversity characteristics for the parental group, across all progeny, and each progeny full-sib family. The progeny families were classified into two groups: progeny from *M. integrifolia* parents and progeny from *M. integrifolia* × *M. tetraphylla* hybrid parents, based on recorded ancestry of each parent cultivar (Table 2-1) (Hardner 2016; Hardner et al. 2009; Peace 2005; Aradhya

et al. 1998). Diversity parameters included mean number of alleles per locus (A), number of effective alleles per locus (A_e), observed and expected heterozygosity (H_o and H_e), and percentage of polymorphic loci (%P) for each family.

Table 2-1 Characterisation of parent genotypes into *M. integrifolia* or *M. integrifolia* × *M. tetraphylla* hybrid groups, based on recorded ancestries and molecular evidence (Hardner 2016; Hardner et al. 2009; Peace 2005; Aradhya et al. 1998)

M. integrifolia		***M. integrifolia* × *M. tetraphylla* hybrid**	
4/7	772	1/40	L64
A38	783	A4	NG4
A9/9	797	A16	NG8
246	804	D4	NG18
333	814	695	NG35
344	816	791	
660	842		
705	849		
762	Yonik		

Nei's (1972) genetic distance was calculated between individuals, using the codominant and binary dominant options for SNPs and silicoDArTs, respectively. Analysis of molecular variance (AMOVA) was performed for both marker types to ascertain the level of partitioning of diversity within and among 32 progeny families, as well as PhiPT (diversity among families) using 999 permutations in GenAlEx v6.5 (Peakall and Smouse 2012). Wright's (1965) F-statistics were calculated for SNP markers. Principal coordinates analysis (PCoA) was computed using SNPs to identify relationships within and among progeny families. An unweighted neighbour-joining dendrogram was constructed for progeny in DARwin v6.0.13 (Perrier et al. 2003) using the SNP markers and 100 bootstraps. A genomic relationship matrix (GRM) was constructed using SNPs following VanRaden (2008) using R (R Core Team 2014) to model the kinship of parents and progeny. A heat map of the GRM was made using Excel.

STRUCTURE v2.3.4 (Pritchard et al. 2000; Falush et al. 2003) was used to determine the level of admixture between progeny families based on SNP markers. Three independent runs were performed for each value of k between 2 and 6, with a burn-in period of 10,000 and 50,000

MCMC (Markov chain Monte Carlo) iterations. No prior population information was used, and the correlated allele frequencies option was chosen. Results of each run were uploaded into Structure Harvester (Evanno et al. 2005; Earl and vonHoldt 2012; http://taylor0.biology.ucla.edu/structureHarvester/) to determine the optimal number of clusters across all families. CLUMPP v1.1.2 (Jakobsson and Rosenberg 2007) was used to determine the alignment of the independent iterations at the optimal k, and outputs were analysed and clusters graphically visualised using DISTRUCT v1.1 (Rosenberg 2004).

2.3.6 Heterozygosity and performance

Analyses were conducted to determine if there was a relationship between genetic diversity and performance for nut yield and other traits. Progeny were separated into two population groups according to their low- or high-yielding status within families, and genetic diversity measures were calculated in GenAlEx v6.5 as above. Heterozygosity was measured on a per-individual basis for progeny as the number of heterozygous markers / total number of markers. This is similar to work proposed by Diehl and Biesiot (1994). Heterozygosity was plotted against clonal values calculated for individuals for cumulative kernel yield to age 8 years, cumulative nut-in-shell yield to 8 years, total kernel recovery, height at age 6, canopy volume and yield efficiency, based on linear models for each trait as a function of an interaction between heterozygosity and family. Adjusted R-square values were obtained from linear models in R, and ANOVA (type II) was performed to determine the significance of heterozygosity, family and their interaction (R Core Team 2014).

2.4 Results

2.4.1 Marker diversity and quality

A total of 5,329 SNPs were produced through genotyping. Single value decomposition (SVD) and probabilistic principal components analysis (PPCA) imputation methods (Stacklies et al. 2007) performed highly for both datasets, and the most accurate of the six methods was used to impute missing data: SVD for silicoDArTs (0.9846 SMC) and PPCA for SNPs (0.9720 SMC). Of the 5,329 SNPs, 1,216 (22.8%) were removed through strict filtering parameters (Table 2-2). The greatest number of markers were removed due to low call rate (446). Manual

detection of progeny with a homozygous state opposite to both parents removed 349 markers from further analyses. This left 4,113 SNPs and 16,171 silicoDArTs across 295 progeny and 29 parents for genetic analyses. Had the call rate been increased to 80% or 90%, this would have resulted in far fewer markers available for genetic analysis (2,536 and 1,777, respectively; Table 2-2).

Table 2-2 Filter criteria and number of SNPs removed and remaining. Call rate, proportion of samples scored as "1" or "0"; one ratio, proportion of samples for which the genotype was scored "1" (minor allele frequency); PIC, polymorphic information content

Filter	**SNPs removed**	**SNPs remaining**
Call rate ≥50%	446	4,883
One ratio reference ≥2.5%	25	4,858
One ratio SNP ≥2.5%	175	4,683
PIC reference >0	210	4,473
PIC SNP >0	11	4,462
Opposing homozygotes	349	4,113
Total	**1216**	**4,113**
Further filtering to call rate ≥80%		2,536
Further filtering to call rate ≥90%		1,777

Call rates for the remaining SNPs ranged from 0.50 to 1.00, with a mean of 0.83, whilst mean call rate of analysed silicoDArT markers was 0.98 (Table 2-3). A much lower mean one ratio was observed for SNP than the reference allele (0.35 and 0.79, respectively). Since SNPs are generally biallelic, and silicoDArT markers are dominant (either presence or absence), PIC values range from 0 to 0.5 for both marker types (Krawczak 1999; Kilian et al. 2012). Mean PIC was higher in SNPs (0.24) than in the silicoDArT markers (0.09; Table 2-3). Very low PIC (<0.05) was observed for most silicoDArTs (66%), compared with 3% of SNPs (Figure 2-2), which may be affected by the filtering of low (≥2.5%) MAF markers and the stringent filtering by DArT during sequencing analysis. Average read depth of silicoDArT markers ranged from 10 to 990.28, with a mean of 47.44. Mean reproducibility was very high for both SNPs and silicoDArTs (0.99 and 1.00; Table 2-3).

Table 2-3 Summary of quality control and genetic diversity measures for 4,113 SNP and 16,171 silicoDArT markers used for analysis across all progeny and parents. Call rate, proportion of samples scored as "1" or "0"; one ratio, proportion of samples for which the genotype was scored "1", reported for both reference and SNP alleles (minor allele frequency); polymorphic information content, based on marker frequency; read depth, average number of samples with non-zero tag read counts; QPMR, quality with scaling per million reads, average of normalised non-zero scores divided by standard deviation; reproducibility, proportion of technical replicate assay pairs for which the marker score is consistent

SNPs	**Mean**	**Min**	**Max**
Call rate	0.83	0.50	1.00
One ratio – reference allele	0.79	0.03	1.00
One ratio – SNP allele	0.35	0.03	1.00
Polymorphic information content	0.24	0.03	0.50
Average read depth – reference allele	73.46	2.57	841.49
Average read depth – SNP allele	52.08	2.50	638.71
Reproducibility	0.99	0.93	1.00
silicoDArTs	**Mean**	**Min**	**Max**
Call rate	0.98	0.79	1.00
One ratio	0.14	0.00	1.00
Polymorphic information content	0.09	0.01	0.50
Average read depth	47.44	10.00	990.28
QPMR	11.52	1.55	4744.36
Reproducibility	1.00	0.95	1.00

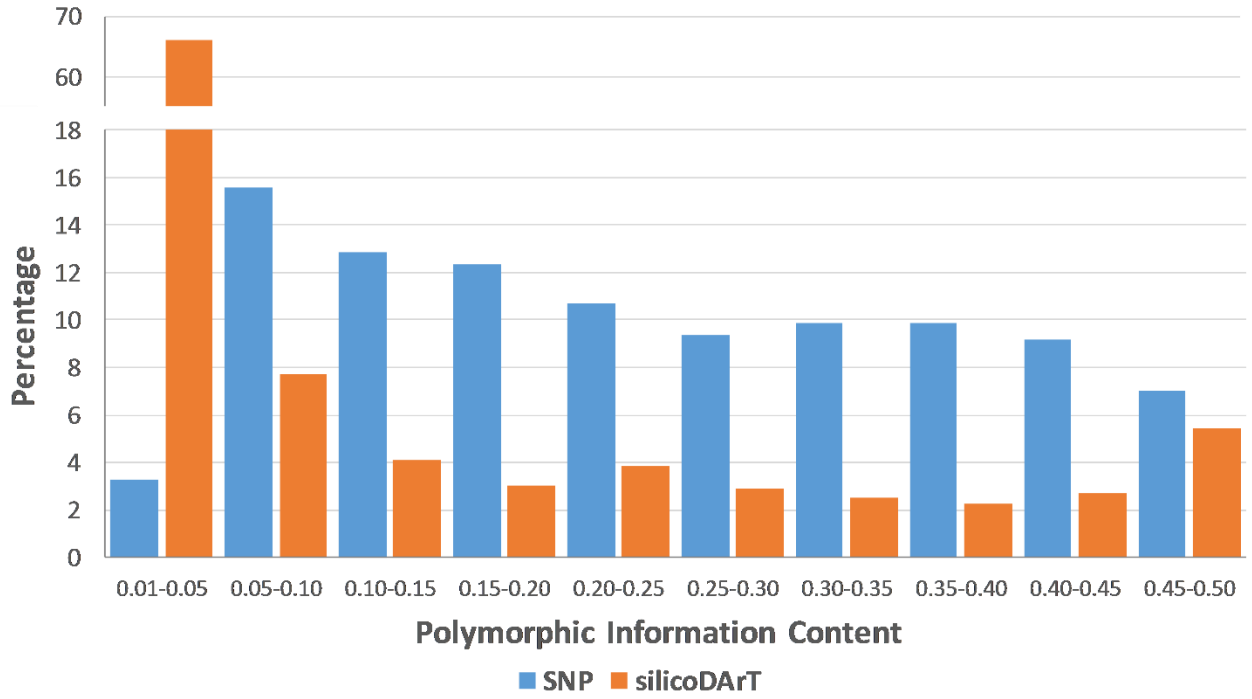


Figure 2-2 Distribution of markers across polymorphic information content ranges for 4,113 SNPs and 16,171 silicoDArT markers used for diversity analysis

2.4.2 Marker location and LD

A total of 3,700 SNPs (90%) mapped to 1,411 genome scaffolds (34%), with 5,248 alignments in total. Most SNPs were present at only one location (2,846, 80%) or mapped to two (13%) locations; the rest mapped to multiple genome scaffolds (Figure 2-3a). One SNP (s3709) mapped to 119 scaffolds, with another (s3710) mapping to 51 scaffolds. The genomic distribution of SNPs across scaffolds was not homogenous, with marker frequency ranging from one SNP per scaffold (on 440 scaffolds, 31% of SNPs) to 34 SNPs on scaffold304|size438510 (Figure 2-3b). Generally, longer scaffolds contained more SNPs than shorter scaffolds ($r^2 = 0.56$, $p < 0.001$).

LD, calculated using only the 2,846 SNPs that mapped to a unique location, decayed rapidly over short physical distances, but then declined slowly over longer pairwise distances (Figure 2-4a). Perfect LD existed between some SNPs even at large distances. Almost 60% of the r^2 values were below 0.05, 14.7% were between 0.05 and 0.10, whilst just under 1% were between 0.9 and 1.00 (Figure 2-4b). The average LD between SNPs within 1 kb of each other was 0.124 for all individuals, 0.095 for parents, and 0.054 for progeny (Figure 2-4c).

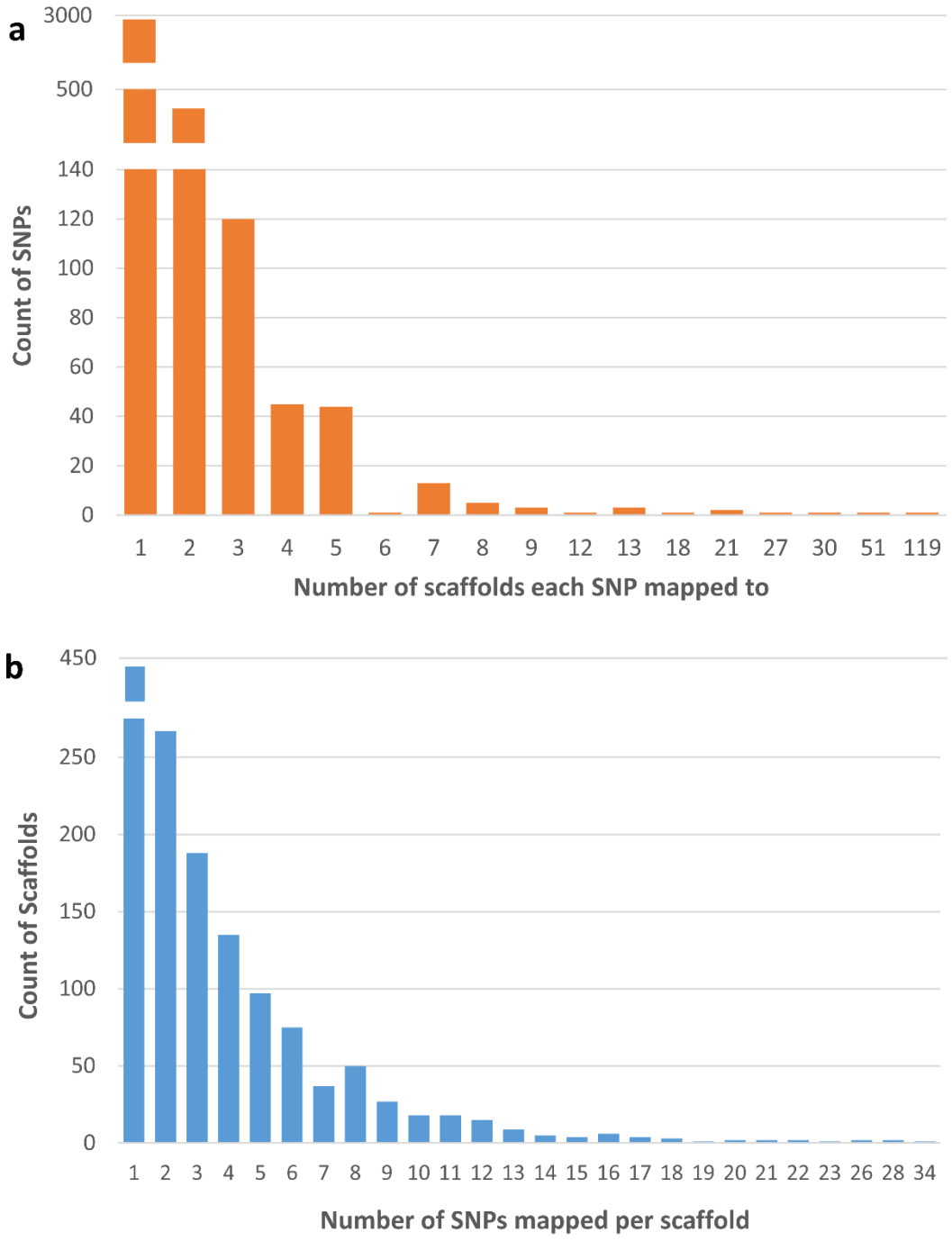
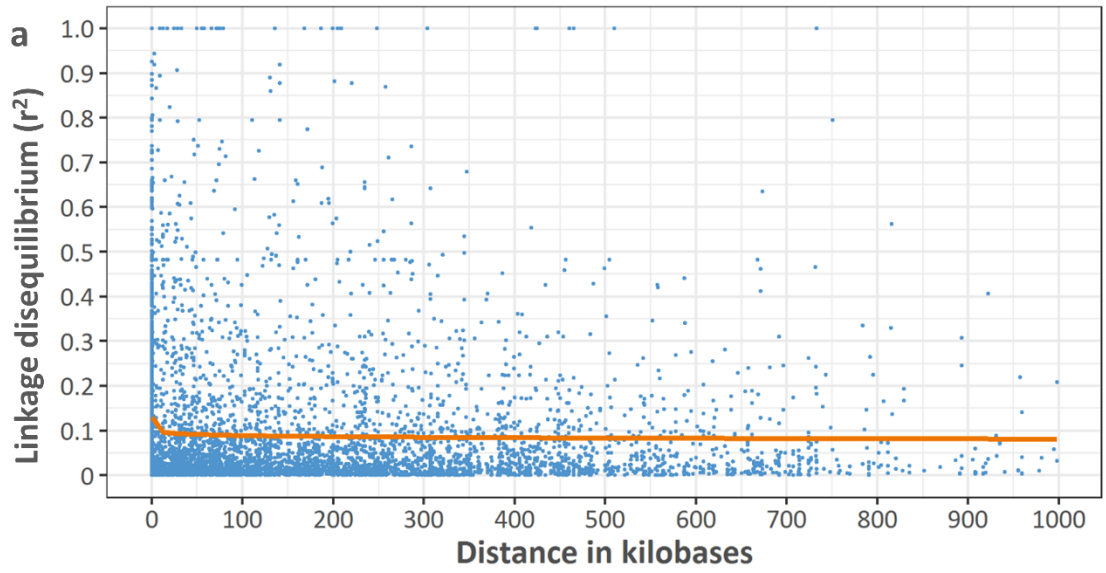
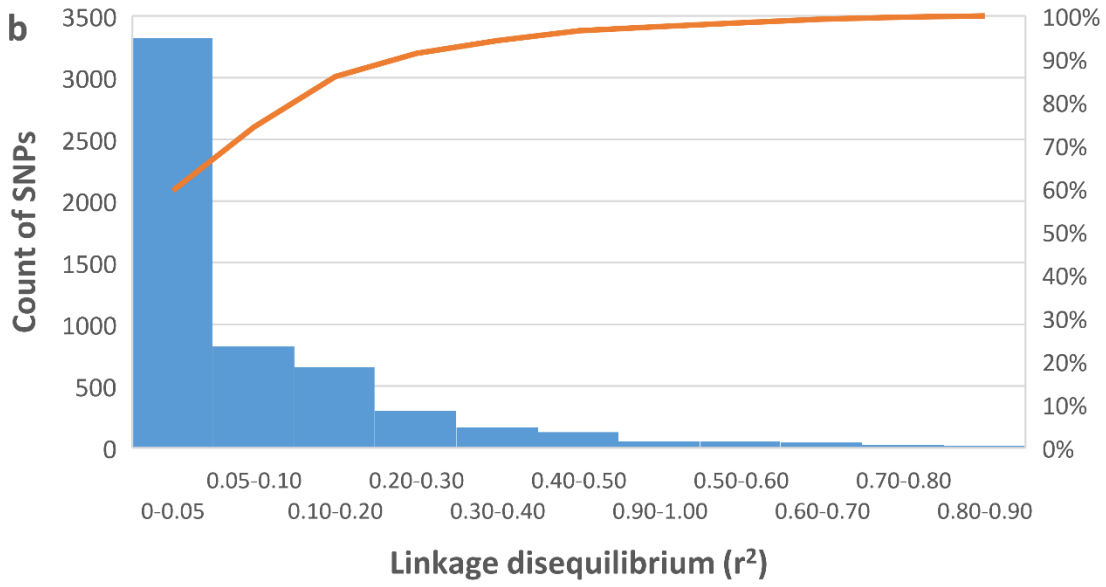


Figure 2-3 Genomic distribution of 3,700 SNPs mapped across 1,411 different genome scaffolds. **a** The number of scaffolds that each individual SNP mapped to, ranging from one unique location to repeated across 119 scaffolds. **b** The number of SNPs mapped per scaffold, ranging from one SNP to 34 SNPs per scaffold.


a
Linkage disequilibrium (r^2)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
100
200
300
400
500
600
700
800
900
1000
Distance in kilobases


b
Count of SNPs
3500
3000
2500
2000
1500
1000
500
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0-0.05
0.05-0.10
0.10-0.20
0.20-0.30
0.30-0.40
0.40-0.50
0.90-1.00
0.50-0.60
0.60-0.70
0.70-0.80
0.80-0.90
Linkage disequilibrium (r^2)

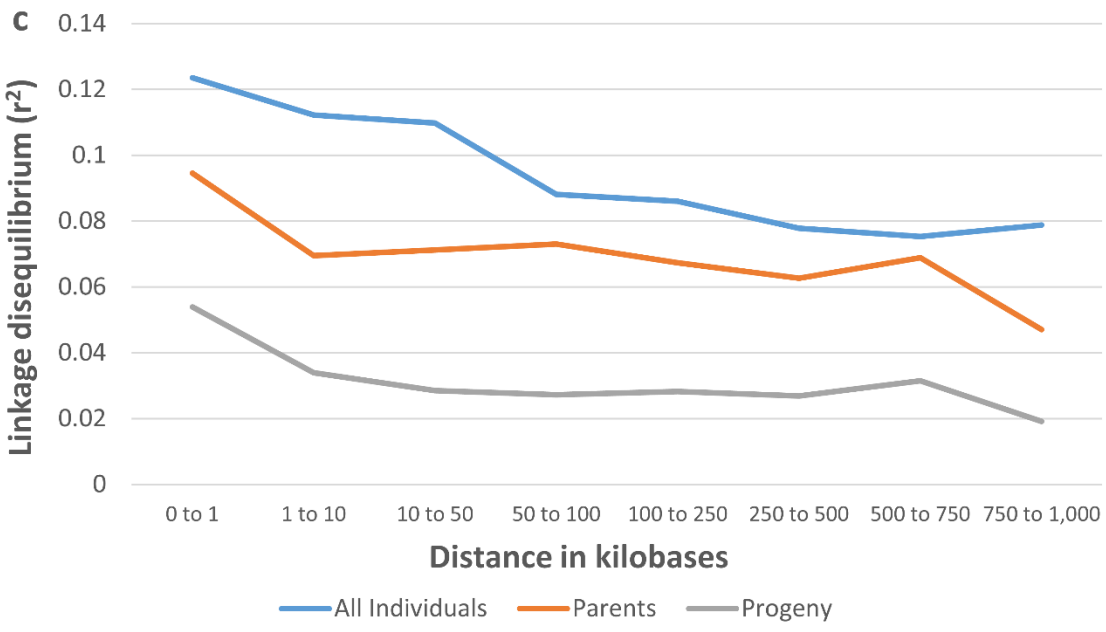

c
Linkage disequilibrium (r^2)
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0 to 1
1 to 10
10 to 50
50 to 100
100 to 250
250 to 500
500 to 750
750 to 1,000
Distance in kilobases
All Individuals
Parents
Progeny

Figure 2-4 Linkage disequilibrium (LD, r^2) between pairs of SNPs across scaffolds, measured using 2,846 SNPs that mapped to only one scaffold each (971 scaffolds total). **a** Scatter plot of pairwise LD decay between all SNPs located on the same scaffold as a function of physical distance between SNPs (kilobases) for all individuals, where the orange line represents LD decay. **b** Distribution of LD values for all individuals among r^2 bins, where the orange line represents the cumulative count of SNPs. **c** Mean pairwise LD between SNPs against pairwise physical distance (kilobases) between markers for all individuals, and parents and progeny separately

2.4.3 Genetic diversity

Genetic diversity varied slightly between progeny and parents (Table 2-4). For progeny, higher number of alleles per locus (A), number of effective alleles (A_e), expected heterozygosity (H_E) and percentage of polymorphic loci (%P) were observed than for parents, though the difference was slight. Fewer heterozygotes were observed for both parents and progeny (H_O = 0.135 and 0.124, respectively) than were expected (H_E = 0.250 and 0.255, respectively; Table 2-4). Genetic diversity was higher for all measures for progeny from hybrid parents than progeny from *M. integrifolia* parents (A = 1.997, 1.804, and %P = 99.66%, 80.43%, respectively). Observed heterozygosity was approximately half the value of expected heterozygosity for hybrid progeny (H_O = 0.131, H_E = 0.278), whilst the difference between the two measures was lower for progeny from *M. integrifolia* parents (H_O = 0.115, H_E = 0.189).

Table 2-4 Summary of genetic diversity measures averaged over 4,113 SNP markers, for the parent population, across all progeny, progeny from *M. integrifolia* parents, and progeny from hybrid parents. n, number of individuals; A, number of alleles; A_e, number of effective alleles; H_O, observed heterozygosity; H_E, expected heterozygosity; %P, percentage of polymorphic loci

		n	**A**	**A_e**	**H_O**	**H_E**	**%P**
Parents	Mean	29	1.984	1.396	0.135	0.250	98.42%
	SE		0.002	0.005	0.002	0.002	
All progeny	Mean	295	2.000	1.403	0.124	0.255	100%
	SE		0.000	0.005	0.002	0.002	
Hybrid progeny	Mean	176	1.997	1.447	0.131	0.278	99.66%
	SE		0.001	0.005	0.002	0.002	
***M. integrifolia* progeny**	Mean	119	1.804	1.303	0.115	0.189	80.43%
	SE		0.006	0.005	0.002	0.003	

Genetic diversity varied between progeny full-sib families (Table 2-5). The number of progeny per family ranged from 7–11. Progeny from the parental cross of 'D4' ('Renown') x '695' ('Beaumont') possessed the highest average values for A (1.536) and %P (53.63), whilst diversity was also high in family 'D4' x '660' (A_e = 1.324, H_E = 0.185). In comparison, families '783' x '804' and 'A9/9' x '814' exhibited low values for these measures (A_e = 1.168, H_E = 0.096, and A = 1.246, %P = 24.60, respectively). Observed heterozygosity ranged from 0.090 in '333' x '842' to 0.165 in 'A16' x '797' (Table 2-5).

Table 2-5 Summary of genetic diversity measures for progeny across 32 families, averaged over 4,113 SNP markers. n, number of progeny in the family with reciprocals combined; A, number of alleles; A_e, number of effective alleles; H_O, observed heterozygosity; H_E, expected heterozygosity; %P, percentage of polymorphic loci. Lowest and highest values for each diversity measure are bolded

Parental cross	**n**	**A**	**A_e**	**H_O**	**H_E**	**%P**
1/40 x 849	8	1.344	1.236	0.115	0.134	34.40%
333 x 842	9	1.279	1.183	**0.090**	0.104	27.86%
344 x 804	10	1.327	1.202	0.119	0.117	32.68%
4/7 x 344	10	1.407	1.272	0.126	0.156	40.68%
660 x 783	8	1.281	1.194	0.107	0.110	28.06%
705 x 816	9	1.385	1.252	0.126	0.144	38.46%
772 x 804	10	1.360	1.240	0.135	0.137	36.03%
772 x 849	8	1.403	1.252	0.141	0.145	40.31%
783 x 804	8	1.248	**1.168**	0.092	**0.096**	24.82%
A16 x 333	10	1.433	1.302	0.138	0.171	43.30%
A16 x 705	10	1.410	1.268	0.134	0.154	40.97%
A16 x 797	10	1.517	1.297	**0.165**	0.173	51.74%
A16 x NG4	9	1.389	1.272	0.113	0.153	38.88%
A38 x 246	10	1.343	1.228	0.109	0.130	34.26%
A38 x 816	9	1.358	1.240	0.117	0.137	35.81%
A4 x 791	11	1.451	1.305	0.132	0.173	45.13%
A4 x NG4	7	1.420	1.290	0.128	0.164	42.01%
A9/9 x 705	8	1.379	1.241	0.126	0.139	37.88%
A9/9 x 814	10	**1.246**	1.174	0.100	0.097	**24.60%**
D4 x 660	8	1.464	**1.324**	0.154	**0.185**	46.39%
D4 x 695	10	**1.536**	1.310	0.130	0.180	**53.63%**
L64 x 344	8	1.407	1.276	0.130	0.158	40.68%
NG18 x 660	9	1.379	1.253	0.134	0.145	37.93%
NG18 x 695	10	1.438	1.301	0.130	0.170	43.84%
NG18 x 705	9	1.340	1.226	0.112	0.129	34.01%
NG18 x 804	10	1.362	1.236	0.131	0.135	36.20%
NG35 x 791	10	1.395	1.262	0.112	0.150	39.53%
NG8 x 333	9	1.409	1.272	0.137	0.155	40.89%
NG8 x 762	10	1.407	1.267	0.126	0.153	40.65%
NG8 x 797	9	1.374	1.254	0.132	0.145	37.44%
Yonik x 814	10	1.333	1.208	0.102	0.121	33.33%
Yonik x NG8	9	1.409	1.272	0.134	0.156	40.87%

2.4.4 Population structure

Figure 2-5 demonstrates relative kinship among parents and progeny, with the family groups separated by black lines. Population structure can be visualised as progeny families related to each other more, or less, than other families, as indicated by red and blue blocks, respectively. In the principal coordinates analysis (PCoA), there are five visible clusters amongst the progeny (circled), with most progeny families clustering together in Q4, or across Q1 to Q2 (Figure 2-6a). Full-sib families that share parents clustered together, for example, 'A4' x 'NG4' and 'A16' x 'NG4' group closely together in Q2 of the PCoA. Family 'NG18' x '705' clustered separately from all other families in Q3, whilst progeny from 'NG18' x '695' and 'A16' x '705' cluster together in Q3.

Similarly to the PCoA, families that shared cultivars as parents grouped together in the dendrogram (Figure 2-6b). Families 'A4' x 'NG4' and 'A16' x 'NG4' clustered closely, as well as families with 'NG8' as a parent. Generally, progeny from hybrid parents (bolded) separated from progeny from *M. integrifolia* parents, both in the PCoA and dendrogram. Long branch lengths were observed for progeny from 'NG18' x '695' and 'D4' x '695' compared to other families. Two individual progenies grouped outside of their full-sib families. Tree HP-1-103 did not group with its family 'A16' x '797' or with any other families. In comparison, tree AL-5-32 grouped more closely with half-siblings of family 'L64' x '344' than with full-sibs from '4/7' x '344'. The clustering of families with 'NG18' and 'NG8' parents observed in the dendrogram also reflects the red blocks among these families in the heatmap (Figure 2-5).

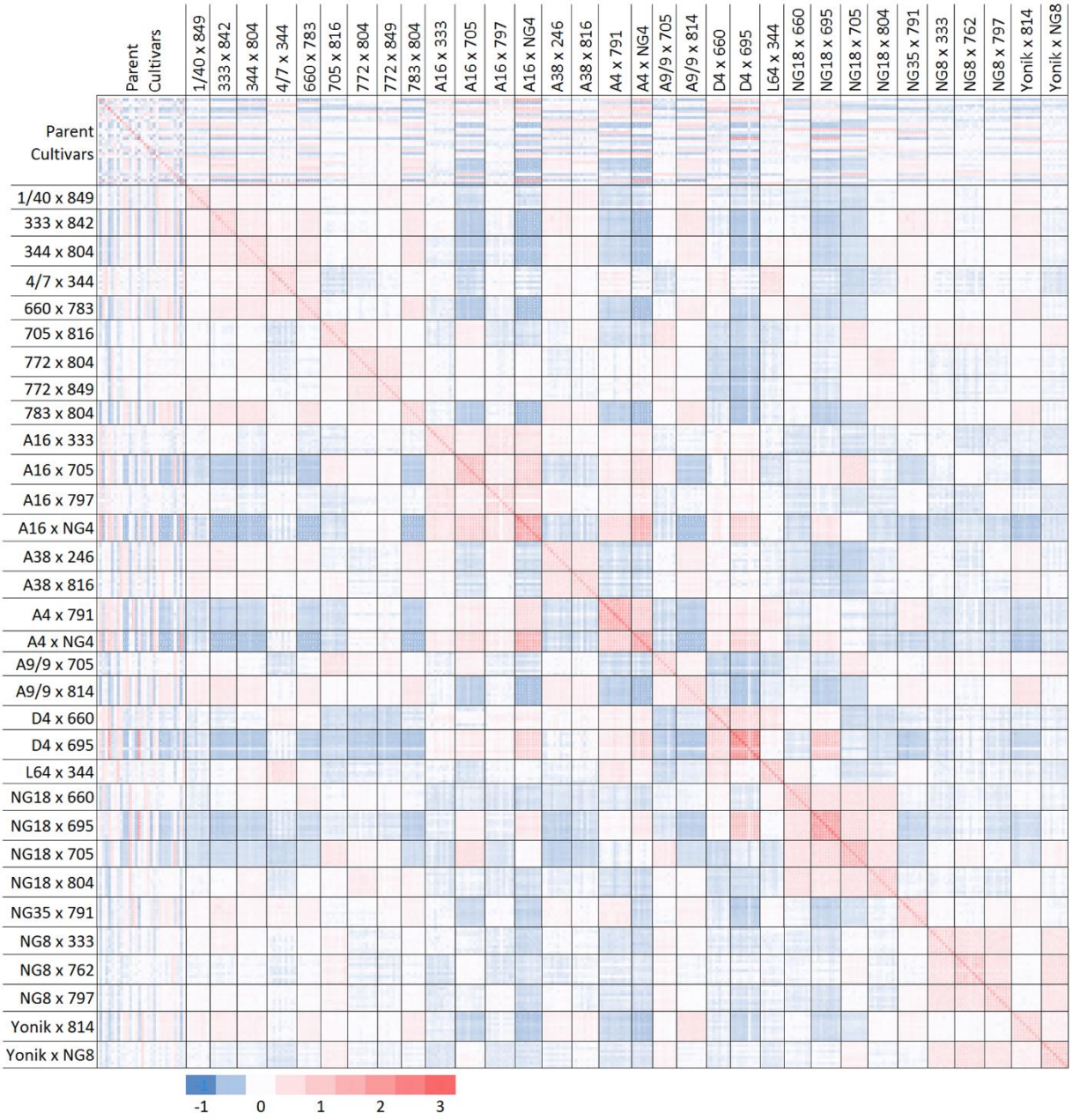


Figure 2-5 Heat map showing pairwise relationships among parents and progeny grouped by full-sib family, from a genomic relationship matrix constructed using methods of VanRaden (2008). Black lines separate parents and each full-sib family. Values range from -1 (blue) to 3 (red), representing relatedness among individuals, but are not equal to coefficients of co-ancestry

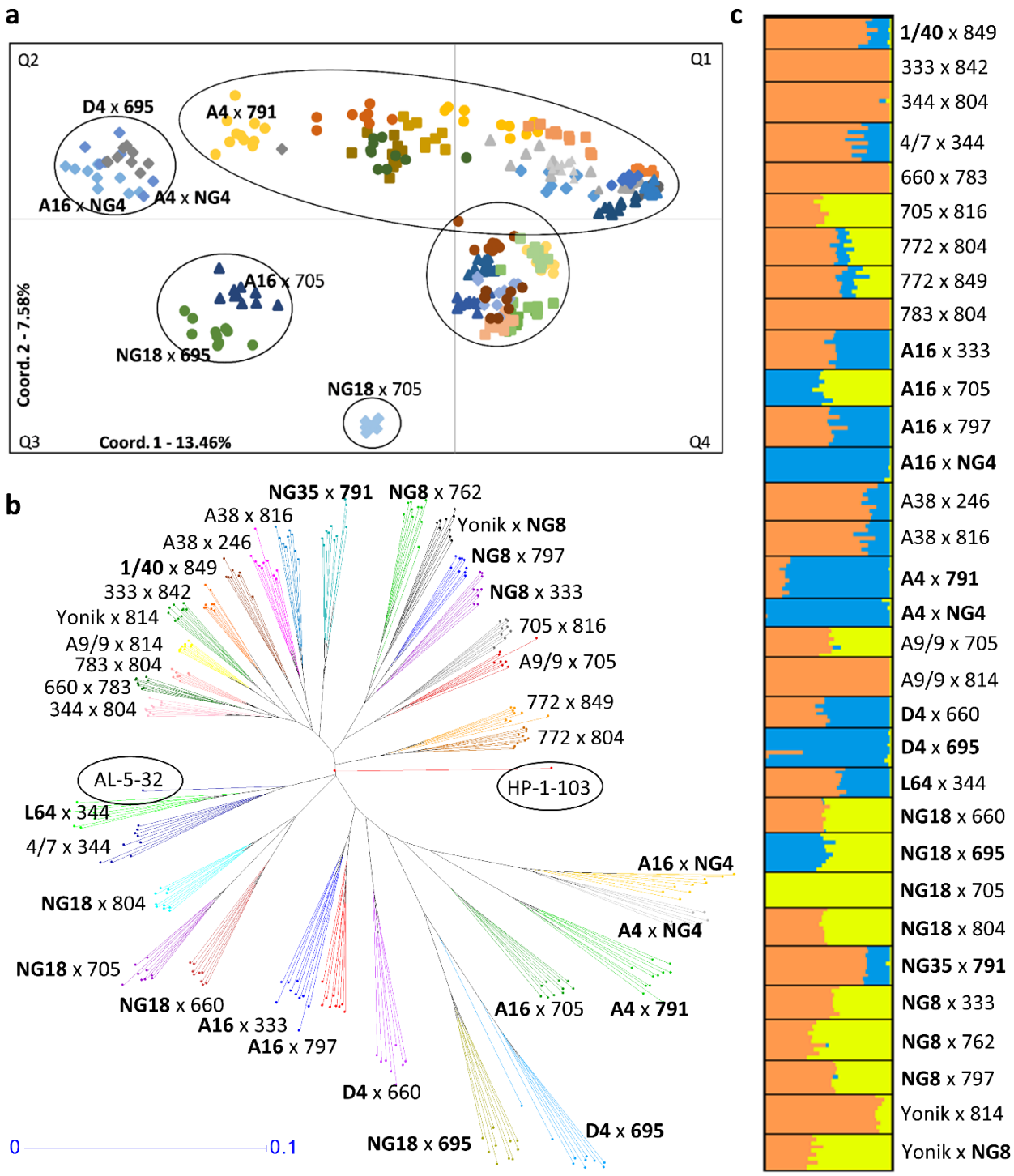


Figure 2-6 Population structure analysis of progeny families based on 4,113 SNP markers, with hybrid parent genotypes in bold. **a** Principal coordinates analysis with progeny coded according to full-sib families with five putative sub-clusters circled. The first two axes, representing the first two principal coordinates, explain 21.04% of the genetic variation. Quadrants are labelled Q1 to Q4 for reference. Full-sib progeny families of note are labelled. **b** Unweighted neighbour-joining dendrogram based on genetic distance among progeny, with

100 bootstraps. Progeny are colour-coded and labelled according to their full-sib families. Circled individuals are discussed. **c** Clustering of progeny among ancestries as calculated by STRUCTURE program and visualised using DISTRUCT software. Each horizontal line represents one genotype, and different colours represent partitioning of the genotype to each cluster, k = 3 clusters. Genotypes are grouped by full-sib family

Three ancestry clusters were identified by Structure Harvester and displayed using DISTRUCT software (Figure 2-6c). The orange group represented progeny from *M. integrifolia*, predominantly HAES parent cultivars; the blue group contains progeny from parents '1/40', '4/7', '772' ('Own Choice'), 'A16', 'A4', 'D4', 'NG4', '695', 'L64' and '791'; the yellow group includes progeny from cultivars 'NG8', 'NG18', '705', '772' and 'Yonik'. There appeared to be families with almost all ancestry attributed entirely to one of the three clusters: '333' x '842', '660' x '783' and 'A9/9' x '814' appeared all orange; 'A16' x 'NG4', 'A4' x 'NG4' and 'D4' x '695' were almost entirely blue. Progeny from family 'NG18' x '705' were the only family that were all undifferentiated in yellow in Figure 2-6c, which concurs with the separation from all other families in the PCoA (Figure 2-6a). However, family 'NG18' x '705' did not appear to separate from others in the dendrogram (Figure 2-6b). Progeny in families '772' x '804' and '772' x '849' showed ancestry from all three clusters. The two individuals that grouped outside their full-sib families in the dendrogram were not obvious in the STRUCTURE output. However, there was one individual in 'D4' x '695' family that did not share the ancestry of their full-sibs, as evidenced by the orange line amongst blue in the STRUCTURE output.

Table 2-6 Summary of partitioning of genetic variance across 32 progeny families. Results for analysis of molecular variance (AMOVA) for both SNP and silicoDArT markers, and F-statistics for SNP markers. PhiPT, diversity among groups (full-sib families); F_{ST}, diversity among subpopulations (full-sib families); F_{IS}, diversity among individuals within subpopulations; F_{IT}, diversity among individuals in the sampled population

AMOVA	**SNP markers**	**silicoDArT markers**
Among progeny families	53%	40%
Within progeny families	47%	60%
Progeny PhiPT	0.529, p = 0.001	0.403, p = 0.001
F-statistics	**Value**	**p**
F_{ST}	0.401	0.001
F_{IS}	0.196	0.001
F_{IT}	0.518	0.001

AMOVA results for progeny families were different between SNP and silicoDArT markers (Table 2-6). Among-family variance (53%) was slightly higher than within progeny families (47%) for the SNP markers, whilst the opposite was true in silicoDArT markers (40% and 60%, respectively; Table 2-6). PhiPT varied by 0.126 between the two marker sets, with the statistic being highly significant for both SNPs (0.529, p = 0.001) and silicoDArTs (0.403, p = 0.001; Table 2-6). Wright's F-statistics further subdivided the level of genetic variation amongst the progeny using SNP markers. There appeared to be a high level of variance detected among individuals (F_{IT} = 0.518) in the sampled breeding population, as well as moderately high diversity among full-sib families (F_{ST} = 0.401; Table 2-6).

2.4.5 Heterozygosity and performance

There were slightly more high-yielding progeny (n = 152) than low-yielding (n = 143), though genetic diversity measures were very similar between the two groups (Table 2-7). There was virtually no correlation between heterozygosity and yield, kernel recovery or tree size traits (r = 0.05 to 0.22; Figure 2-7). Linear models of traits as a function of heterozygosity and family (Figure 2-7) all showed that family was a significant factor, but neither heterozygosity nor the interaction were significant. While there were two individuals with high heterozygosity (>0.2, Figure 2-7), these had relatively moderate clonal values for all traits.

Table 2-7 Summary of genetic diversity measures for low- and high-yielding progeny per family, averaged over 4,113 SNP markers. n, number of individuals; A, number of alleles; A_e, number of effective alleles; H_O, observed heterozygosity; H_E, expected heterozygosity; %P, percentage of polymorphic loci

Progeny		**n**	**A**	**A_e**	**H_O**	**H_E**	**%P**
Low-yielding	Mean	143	2.000	1.399	0.121	0.252	99.98%
	SE		0.000	0.005	0.002	0.002	
High-yielding	Mean	152	1.999	1.406	0.128	0.256	99.93%
	SE		0.000	0.005	0.002	0.002	

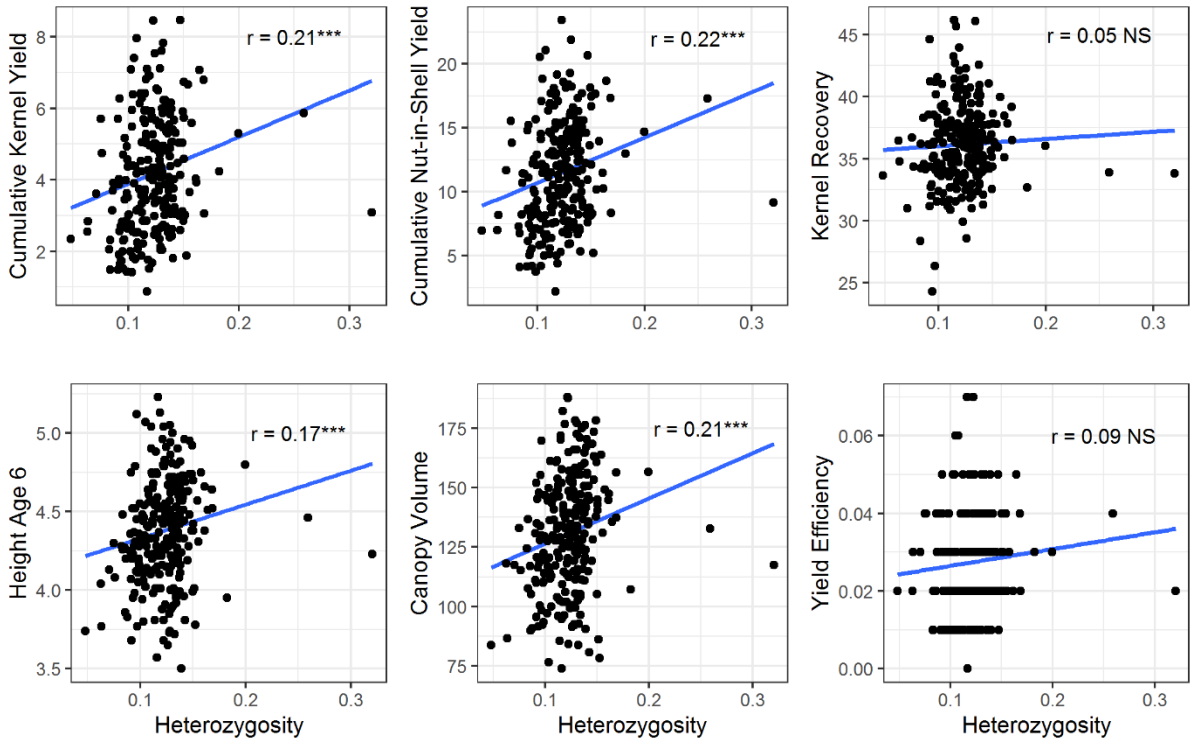


Figure 2-7 Linear models of clonal values of various yield traits as a function of individual tree heterozygosity (number of heterozygous markers / total number of markers, 4,113) and an interaction with progeny family. Correlation coefficients (r) and significance values are also shown. *** $p < 0.01$; NS, not significant.

2.5 Discussion

2.5.1 Marker location and LD

Marker locations were estimated using genome scaffolds; however, until a complete reference genome for *Macadamia* is available, the exact position and distribution of markers across chromosomes is unknown. Nock et al. (2016) estimated that 37% of the macadamia genome is repetitive. For the more recent genome assembly, to which the markers in this study were mapped, estimates of this repeat content is increased to 52% (Nock, Baten *et al*. unpublished data). This may explain why some SNPs in the current study mapped to multiple locations (up to 51 and 119 different genome scaffolds). The complexity reduction methods used by DArT in sequencing aims to reduce repetitive sequences in the representation (Kilian et al. 2012). This reduction appears to be the case here, in part, with 80% of the mapped markers being present at a single location (Figure 2-3a). Heterogeneous distribution of SNPs across scaffolds, with longer scaffolds containing more markers, was also observed in Chinese jujube (*Ziziphus jujuba*), with marker density differing across pseudo-chromosomes of similar sizes (Chen et al. 2017). Estimates of marker density across chromosomes or linkage groups will be more accurate when future versions of the macadamia genome become available.

Preliminary estimates of LD in this paper are lower than estimates for apple, a similarly cross-pollinating fruit tree. Average estimates of r^2 in apple at a distance of 100 kb ranged from 0.211 in one study (Vanderzande et al. 2017) to 0.33 in another (Kumar et al. 2012b), though population size varied among these and the current study. In comparison, within-scaffold LD in pear was $r^2 = 0.10$ for 100 kb (Kumar et al. 2017), which is similar to estimates of the current study at around 0.09 for the same distance (Figure 2-4c). Less than 1% of SNP pairs exhibited r^2 values between 0.9 and 1, compared with 4% (Vanderzande et al. 2017) and 17% (Kumar et al. 2012b) in apple. LD in the current study was lower than expected, and was probably influenced by a number of factors. Generally, out-crossing species have lower LD, and LD decays more rapidly than self-pollinating species due to recombination being more effective in heterozygous crops (Flint-Garcia et al. 2003; Semagn et al. 2010). Current macadamia cultivars are only a few generations removed from their wild relatives (Hardner et al. 2009), suggesting a limited number of recombination events during domestication. This is very different to apple and pear, which have both been cultivated for centuries and have strong genetic structures (Larsen et al. 2018). Population bottlenecks can increase LD (Semagn et al.

2010). Research suggests that the population base of the HAES germplasm is small compared to the indigenous natural populations in Australia (Hardner et al. 2009; Peace 2005). However, the parents in the current study are from many different sources, and Peace (2005) found that the wider domesticated germplasm was sourced across the wild distribution. The long extent of low level LD (r^2 ~ 0.1) is consistent with a recent reduction in effective population size, as has been observed in other domesticated species, including cattle (Hayes et al. 2003). Due to the fragmented nature of the genome assembly, caution should be employed when interpreting these preliminary marker locations and corresponding LD. LD in macadamia should be further investigated when a complete reference genome becomes available, and this analysis should take into account population structure and cryptic relatedness.

2.5.2 Marker quality

DArT marker platforms have been successfully used to quantify genetic diversity in many plant breeding studies (e.g. Sánchez-Sevilla et al. 2015; Roorkiwal et al. 2014; Grzebelus et al. 2014; Sagawa et al. 2018). Genotyping errors occur in all systems of sequencing, which can affect subsequent analyses (Saunders et al. 2007). Applying strict filter parameters can ensure that the markers used for genetic analysis are high-quality, and may decrease the number of spurious results in subsequent genomics analyses. In SNPs, heterozygosity may be under-represented due to “allelic dropouts”; when only one of two alleles is detected (Hardenbol et al. 2005), sometimes due to low coverage (Nielsen et al. 2011). Detection of null alleles or dropped alleles can be achieved through Mendelian inheritance or Hardy-Weinberg tests, rather than assay replication (Pompanon et al. 2005). This approach has rarely been applied in genetic diversity and genomics analyses (Vanderzande et al. 2017; Cros et al. 2017; Imai et al. 2018). The high read depth observed in this study suggested that most missing data were due to null alleles. However, the errors detected through Mendelian inheritance testing suggested that allelic dropouts or null alleles have occurred in the population, despite filtering protocols by Diversity Arrays Technology. Furthermore, inaccurate imputation techniques commonly replace missing data with mean or random alleles (e.g. Heslot et al. 2012; Heffner et al. 2011; Emanuelli et al. 2013). This study included a test for Mendelian inconsistencies and used an accurate method of imputation to replace missing data, and, therefore, used

more stringent genotypic quality control parameters compared with those used in many other genetic diversity studies. Given the relatively small sample size of the study population and the quality control employed to filter markers, this set of markers was considered suitable for further research including genome-wide association studies and genomic selection in macadamia. However, potential allelic dropouts could be further investigated with the aid of a complete reference genome in the future.

2.5.3 Genetic diversity

This is the first comprehensive study to investigate genetic diversity in a macadamia breeding program using SNP and silicoDArT markers over a structured population design. Despite the aim of the breeding population being to increase genetic diversity relative to previous breeding populations (C. Hardner, pers. comm.), the results of the study did indicate that the diversity of the progeny and parents was relatively low, with little difference between the two populations. However, the families in the current study are a subset of the seedling progeny trial, of which the total genetic diversity is unknown. As such, it is not known whether the two-tailed sampling (low- and high-yielding) of the progeny may have contributed to the levels of genetic diversity in the current study. Across all progeny and all parents, observed heterozygosity was almost half the expected heterozygosity. This may be explained by "allelic dropouts" giving an underrepresentation of heterozygosity (Hardenbol et al. 2005) in this study.

Ancestry of most parents in this study has been documented through pedigree records and molecular markers in previous studies (Table 2-1) (Aradhya et al. 1998; Hardner et al. 2009; Peace 2005; Hardner 2016). Progeny from hybrid parents were more genetically diverse than progeny from *M. integrifolia* parents; this was demonstrated by grouping hybrid and non-hybrid progeny, as well as by investigating full-sib families individually. The high levels of diversity in progeny from 'D4' x '695', compared with other families, as measured by number of alleles and polymorphic loci, is likely due to the fact that both varieties are *M. integrifolia* × *M. tetraphylla* hybrids (Table 2-1) (Hardner et al. 2009); therefore, introducing more genetic variation than crosses between *M. integrifolia* cultivars. Both '783' and '804' are offspring from '246' (Hardner 2016), which may have led to their subsequent progeny having lower levels of genetic diversity due to inbreeding from mating closely related individuals.

Compared to the diversity of the parental cultivars in the current study, Alam et al. (2018) found that expected heterozygosity was generally lower across four sub-populations among 80 macadamia cultivars, ranging from 0.12 to 0.22. Observed heterozygosity for progeny in the current study (0.124) was lower than that for SNPs in apple (0.36) and grape (0.34), and was comparable for Asian, European and hybrid pear groups (0.18 to 0.20). In peach (*Prunus persica*), which is predominantly self-pollinating and highly homozygous (summarised by Xie et al. 2010), observed heterozygosity (0.264) was lower than expected (0.401) (Akagi et al. 2016), though higher than the macadamia progeny (H_E = 0.255). Genetic diversity was, thus, generally lower in the current study than for other fruit crops. Furthermore, a study of genetic diversity in wild *M. tetraphylla* populations also found that expected heterozygosity was higher than observed when assessed using microsatellite markers (O'Connor et al. 2015), and both were higher than that found in this study. However, Fischer et al. (2017) found no correlation between SSR and SNP diversity estimates, and so it is difficult to compare diversity values between studies that use different marker types. That expected heterozygosity was low in the study demonstrates that the assumptions of Hardy-Weinberg law were not met, namely that of a large, randomly-mating population (Falconer and Mackay 1996). The very low correlation between heterozygosity and yield on an individual tree basis across all families reflected the minor differences in genetic diversity and heterozygosity statistics between low- and high-yielding grouped progeny (Table 2-7). This concurs with a meta-analysis of heterozygosity-fitness correlations in animal populations, where weak or no correlations are commonly found (Chapman et al. 2009). Further, there was no relationship observed at the within-family level for heterozygosity and progeny trait performance, as evidenced by the lack of significance in the interaction between heterozygosity and family for different traits in linear model analyses.

2.5.4 Population structure

The current study found disparity in partitioning of genetic variance when comparing results for SNPs and silicoDArTs. This disparity may be explained by the fact that SNPs provide more allelic information per locus, compared with silicoDArTs (presence/absence only). In wild *M. tetraphylla* populations (O'Connor et al. 2015), among-family genetic diversity, measured

using SSRs, was slightly higher (79%) than the results of the current study (40% for silicoDArTs and 53% for SNPs).

The heatmap (Figure 2-5) represents relative relationships among individuals and will inform genomics studies such as genome-wide association studies and genomic selection, where knowledge of population structure is vital. Figure 2-6 shows this population structure more explicitly through different graphical types and may reflect the pedigree of the parents. In the PCoA, progeny from hybrid parents separated from the major clusters of individuals representing the HAES germplasm. Cultivars 'A4' and 'A16' are both offspring of 'D4' (also known as 'Renown') and molecular evidence suggested they share the same pollen parent (Peace 2005). This may explain the close clustering of progeny from these parents with 'D4' x '695' ('Beaumont') in the PCoA. The long dendrogram branches of progeny from hybrid parents, for example 'D4' x '695' reflects the amount of evolutionary change (Pavlopoulos et al. 2010) compared with progeny from HAES germplasm, suggesting that the genetic diversity of hybrids is greater than those with *M. integrifolia* ancestry. This concurs with the finding that progeny from hybrid parents were more genetically diverse in terms of observed and expected heterozygosity than *M. integrifolia* progeny.

Two individuals grouped separately from their pedigree families in the dendrogram, HP-1-103 and AL-5-32. This may be due to an error somewhere between crossing and planting, however they did not separate in the PCoA or STRUCTURE output. The outliers observed in the dendrogram should not affect future genomic prediction studies using these data, as a GRM should be used rather than a pedigree matrix to account for population structure, which will provide greater accuracy (Hayes et al. 2009b).

The three ancestry clusters shown in the STRUCTURE output may represent three different germplasm groups: *M. integrifolia* HAES germplasm (orange), and two different hybrid groups. The blue group contains progeny from parent cultivars '1/40', '4/7', '772', 'A16', 'A4', 'D4', 'NG4', '695', 'L64' and '791'. Most of these cultivars had recorded hybrid ancestry (Table 2-1). There was a paucity of data regarding '4/7', and the results here suggest that the ancestry of this cultivar is partly *M. tetraphylla.* Cultivar '772', which is also known as 'Own Choice' (Hardner et al. 2009), appears to have a mixed ancestry as demonstrated by the STRUCTURE output, despite records suggesting it originated from a wild *M. integrifolia* population (Peace 2005). A previous study (Peace 2005) classified 'D4' ('Renown') into a

hybrid group separate from '695' ('Beaumont) and 'NG4' ('X4') and 'NG8' ('X8'), which conflicts with the current findings.

The other hybrid group, yellow, includes 'NG8', 'NG18', '705', 'Yonik', as well as '772' ('Own Choice'). Alam et al. (2018) also found that 'NG8' and 'NG18' were assigned mostly to the same ancestry, but that 'NG18' was more differentiated. In comparison, an earlier study inferred that 'NG18' was 5– 25% *M. tetraphylla*, despite the original genetic background being recorded as *M. integrifolia* (Peace 2005). The results from this study agree with Peace (2005) who suggested, based on SSR and RAF marker evidence, that the cultivar '705' is a hybrid. This variety was originally selected in Australia ('Stokes Siding') and introduced into Hawaii in the mid-1950s where it was given a HAES selection number (Hardner 2016; Hardner et al. 2009). It was sourced from New South Wales (C. Hardner, pers. comm.), where the local wild species is *M. tetraphylla* (Hardner et al. 2009). Progeny from family 'NG18' x '705' separated from others in the PCoA, and group entirely to one ancestry in the STRUCTURE output, suggesting that these varieties may be more distinct from other hybrid parents in the current study, though not sufficiently different to be classified into a different cluster.

Some findings of the current study also conflict with records and publications. 'Yonik' is a variety from Israel with recorded *M. integrifolia* ancestry; however, evidence from the STRUCTURE plot suggests that it may also be a hybrid. Variety 'NG35' was recorded to be of hybrid origin (Peace 2005), though, the results of this study suggest that it is more likely to be *M. integrifolia* due to its progeny clustering separately from hybrids. Finally, cultivar '791', also known as 'Fuji', was previously identified to be a tri-hybrid between *M. ternifolia*, *M. tetraphylla* and *M. integrifolia* using genetic markers (Peace 2005), which is also supported by findings by Alam et al. (2018). In the current study, families with '791' as a parent did not show clustering among three ancestries in the STRUCTURE output. Progeny of '791' did not separate from other families in the dendrogram, although 'A4' x '791' progeny were clustered between HAES and hybrid germplasm in the PCoA. It is possible that the *M. ternifolia* ancestry was, instead, attributed to the group containing *M. tetraphylla* ancestry in the current study, because there was only one parental sample containing that species. Future studies should compare genetic diversity and partitioning of multiple genotypes among the four species of the genus, and cultivar '791'.

2.6 Conclusions

This study analysed a large, representative subset of progeny of the Australian macadamia breeding population, as well as many cultivars as parents. We used 4,113 SNP and 16,171 silicoDArT markers to characterise the partitioning and structure of genetic diversity of the breeding population. High genetic variance was observed among progeny and among full-sib families, though the population appears less diverse than other tree crops. The population seems to be structured by HAES *M. integrifolia* germplasm separating from hybrid germplasm. Knowledge gained will be valuable for future studies using genetic markers in macadamia. A genome-wide association study regarding important yield traits will benefit from the number of markers available as well as awareness of population structure to avoid bias and spurious results.

Chapter 3. Genomic heritability, correlations, and selection efficiency of nut yield and component traits in a macadamia breeding population

---

Preliminary results for the following chapter have been published.

O'Connor, K., Hardner, C., Alam, M., Hayes, B., and Topp, B. (2018). Variation in floral and growth traits in a macadamia breeding population. In: R. Drew (ed.), International Symposia on Tropical and Temperate Horticulture ISTTH2016 (Cairns, Queensland). Acta Horticulturae 1205(77): 623-630. https://doi.org/10.17660/ActaHortic.2018.1205.77

My contribution to the chapter

I performed all phenotyping with assistance, wrote the paper and made final edits. CH guided me during analyses and interpretation of results. BH suggested analytical methods and interpretation of results. CH, BT and MO suggested revisions.

3.1 Abstract

Macadamia trees are grown commercially for their high quality kernel, but improving nut yield in breeding programs is inhibited due to low heritability of yield, long juvenile period and large tree size. Selection for high yield may be done indirectly through selection of yield component traits if they are correlated with yield and have higher heritability. Yield, flower, nut and tree growth characteristics were measured for 295 seedlings from 32 families, and 18 of their 29 parents, across four sites in south-east Queensland, in the Australian macadamia breeding program. Estimates of individual tree narrow-sense genomic heritability varied between traits, from 0.09 for percentage of racemes that set nuts to 0.76 for kernel recovery. Trunk circumference appeared highly genetically correlated with yield (0.72), whilst a negative correlation was observed between kernel recovery and yield (-0.27). Using estimates of heritability and genetic correlations, selection efficiency was calculated for each trait. None of the measured traits were more efficient in indirectly selecting for high yield, though trunk circumference was the highest, at 0.86 selection efficiency. Findings from this study will inform genomics research, such as the appropriateness of traits for genome-wide association studies, and be useful for future breeding strategies regarding the ease of selecting for certain traits due to heritability and correlations between traits.

3.2 Introduction

Macadamia is a long-lived, evergreen tree with a long juvenile period, native to the rainforests of south-east Queensland and north-east New South Wales, Australia. *Macadamia integrifolia* (Maiden & Betche), *M. tetraphylla* (Johnson), and their hybrids are grown commercially for their high quality nuts in countries around the world, predominantly in Australia, South Africa, Kenya and USA (Hawaii) (Hardner et al. 2009; Gross 1995; Australian Macadamia Society 2018). In 2016, Australia produced 48,000 mt of nuts-in-shell, and exported 70% of the crop around the world (Australian Macadamia Society 2017b).

Nut-in-shell (NIS) yield is the focus of breeding new varieties, and is consistently nominated as the highest priority for growers at industry meetings in Australia (O'Hare and Topp 2010). However, selecting for high yield can be difficult due to the quantitative nature of the trait, low heritability, long juvenile period, and large tree size requiring large areas of land for

evaluations. The current Australian macadamia breeding program evaluates candidate cultivars in two stages (Topp et al. 2012; Hardner et al. 2019a). Unreplicated seedlings are evaluated in field trials using pedigree and a quantitative genetic approach for growth and yield traits to predict genetic values using a weighted selection index. These traits include canopy height and width, precocity, proportion of reject and marketable kernel, with economic weights estimated for each trait based on bio-economic modelling of production and processing costs (Hardner et al. 2019a; Hardner et al. 2006). Elite individuals, identified through the selection index, are then clonally propagated and grown in replicated trials across multiple locations to determine performance in differing growing environments. These trials also allow comparison against performance of current industry standard cultivars (Hardner et al. 2009). Each of these stages extends for at least eight years due to the poor correlation between young and mature yields, and, hence, involves a large amount of labour in order to assess juvenile yield on an individual tree scale. Recently, four new varieties were released to industry, which are a result of pollination crosses made in the mid-1990s (Russell et al. 2017; Hardner et al. 2019a). Thus, the breeding of new macadamia cultivars is a costly, lengthy and laborious process (Topp et al. 2012; 2016).

Breeders could indirectly select for complex traits, like yield, through the evaluation of component traits (Simmonds 1979; Sparnaaij and Bos 1993; Piepho 1995). Here, a complex trait is determined as a function of one or more component traits, and the variation among genotypes in the complex trait is governed by component trait variation (Sparnaaij and Bos 1993). Component traits with high correlation with the target trait, high heritability, and those that are more easily measured than the complex target trait, are candidates for indirect selection (Falconer 1989; Sparnaaij and Bos 1993). Some component traits can be measured at an earlier stage in the life of the tree, which can lead to reduced cycle times, or traits can be measured more easily on a larger number of trees, meaning that selection intensity is increased. However, selection could be hindered by undesirable traits being highly correlated with the target trait, or conversely, two target traits being negatively correlated (Dicenta and Garcia 1992). In a recent review, O'Connor et al. (2018b; Chapter 1 Literature review and general introduction) suggested that nut, flowering and growth characteristics could be candidates for early-generation phenotypic selection in macadamia.

Understanding the genetic architecture of yield component traits will inform breeders of which traits could be used to indirectly select for yield. Narrow-sense heritability indicates to what level the trait is genetically controlled, and, as such, how easily it can be modified through selection (Falconer 1989). Heritability of traits and genetic correlation with yield can be estimated using linear mixed model approaches, and these estimates can then be used to calculate the efficiency of indirect selection on the target trait through the direct selection of another trait (Falconer 1989). The efficiency of selection depends on the ratio of the correlated response of the two traits to the direct response of the target trait.

Linear mixed models are a commonly used approach for the prediction of genetic effects. This method treats genetics effects as random with given covariance structure, which often must be estimated from available data prior to prediction of effects. Assuming covariance estimates are close to the real population values, genetic effects are best linear unbiased predictions (BLUPs), with environmental effects accounted for as non-genetic fixed effects. Commonly, historical pedigrees are used to describe genetic relationships among individuals to support estimation of additive genetic effects (breeding values) or other non-additive effects (Henderson 1975). Alternatively, the estimation of realised relationships among individuals using genetic markers (genomic relationship matrix, GRM) allows more accurate predictions compared with relationships estimated from pedigree, which do not take into account differences in Mendelian inheritance or undocumented cryptic relationships (Clark and Van der Werf 2013; Hayes et al. 2009b; Chapter 2 Population Structure and Genetic Diversity). The incorporation of genomic relationships has been used to predict breeding values in many crop and livestock studies, in a method termed genomic BLUP (GBLUP) (Clark and Van der Werf 2013).

Selection of elite individuals (based on certain traits) can be complicated by genotype by environment interaction (G x E); the relative performance of genotypes (for certain traits) can change depending on their environment, and so individuals that are selected for their elite performance in one environment may perform poorly in another environment (Allard and Bradshaw 1964). Cooper and DeLacy (1994) succinctly summarised different forms of G x E: where genetic variance is heterogeneous across environments, and/or the rankings of genotypes change amongst environments. G x E can be modelled using linear mixed models. In a previous study of macadamia yield (Hardner et al. 2002), G x E was evidenced by higher

genetic variance at one location, and low genetic correlation between genotype and yield at that location compared with three other environments. As such, the presence of G x E should be considered in studies of yield in macadamia. More recently, Hardner (2017) used scaled phenotypes to reduce variance heterogeneity, and explored mixed models to increase accuracy of estimating parameters and trait genetic effects in a macadamia breeding population using pedigree information.

There are many traits that may contribute to yield in macadamia. Flowers are comprised of pendant racemes, with a rachis (stem) 6–30 cm long, and an inflorescence of 100–300 florets (Huett 2004; Trueman 2013). Physical space to sustain nuts on the rachis, and opportunities for pollination may influence yield, so raceme length and floret number may be important yield components. Previous estimates suggest that only 0.3% of florets develop into nuts (Ito 1980). The percentage of flowering racemes, as well as individual florets that set nuts, could give an indication of energy investment and reproductive success among genotypes. The rachis diameter increases following successful pollination, and varies with the number of fruit per rachis (Urata 1954). Rachis diameter, as well as nut pedicel diameter, thus could also be important component traits, as they may limit nut size and yield.

Important selection traits during cultivar evaluation include NIS yield, nut weight, percentage of kernel to nut weight, and tree size (Hardner et al. 2009). Nuts are composed of an inner kernel encased by a hard shell and outer husk (Figure 3-1). Husk may dehisce with the nut, and subsequently needs to be mechanically removed, with nuts-in-shell then cracked to produce the edible kernel . Yield may be described using multiple parameters, but for this study, it is measured as the weight of NIS yield per tree in a particular year. Consistent yields of 2 t/ha of kernel or 5 t/ha of NIS (10% moisture content) are desirable in varieties from the age of 10 years (O'Hare et al. 2004). Kernels should weigh 2–3 g, preferably remain whole when the nut is cracked, and compose at least 36% of the total NIS weight (kernel recovery; KR) (O'Hare et al. 2004; Hardner et al. 2009). KR is an important trait, as higher values attract a premium price per kilogram for growers (Macadamia Processing Co. Ltd. 2018), and it directly impacts the production and processing costs; costs are higher when KR is low (Hardner et al. 2009). However, cultivars with thin shells may have lower kernel quality due to defects caused by damaging pests and diseases (Hardner et al. 2009). With the focus turning towards smaller trees with high yields, planting densities, and, consequently, yield

per hectare, are likely to increase (Toft et al. 2019). Trunk circumference (stem girth) or trunk cross-sectional area, are easily-measured traits commonly used as an estimate of tree size (Hardner et al. 2002; Toft et al. 2018); trunk size may affect nut yield due to uptake of resources available for growth and fruit production.

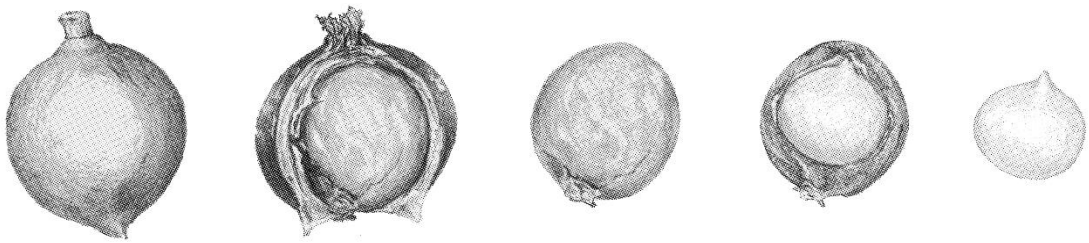


Figure 3-1 Macadamia nuts. The edible kernel is enclosed in a hard, woody shell and the outer husk. Left to right: nut in husk, split husk, nut in shell, cracked shell, and whole kernel. Illustration by Todd Fox, reproduced from O'Connor et al. (2018b)

Variance, heritability, and correlations among component traits have been investigated in various nut tree crops, including almond (Sánchez-Pérez et al. 2007; Sorkheh et al. 2010), pecan (Thompson and Baker 1993; Kumar et al. 2013a), cashew (Aliyu 2006), and walnut (Martínez-García et al. 2017), though analyses using genetic markers is lacking. To date, studies in macadamia have investigated the relationships between some component traits and yield, though this has usually included only a limited number of genotypes and without genetic marker data (e.g. Hardner et al. 2001; Hardner et al. 2002). It is of value to determine the relationships between component traits and yield in a broader genetic base, as findings will influence selection strategies in future macadamia breeding populations (O'Connor et al. 2018b). The aims of this study were to: (i) compare the phenotypic variances of nut, floral and growth component traits between progenies and families, (ii) estimate the genetic heritability of yield and component traits in the population, (iii) estimate genetic correlations among yield and component traits, and (iv) estimate efficiency of indirectly selecting for high yield through the selection of component traits. This study builds on results of a pilot study by O'Connor et al. (2018a).

3.3 Methods

3.3.1 Study population

This study involves 295 unreplicated seedling progeny from 29 parents and 32 full-sib families (reciprocal parent crosses combined), which are a subset of the Australian macadamia breeding population, as described in O'Connor et al. (2019b; Chapter 2 Population structure and genetic diversity). The studied plants were located at four sites across south-east Queensland: East Gympie (EG, n = 75 progeny) and Amamoor (AM; n = 84) near Gympie, and Alloway (AL; n = 59) and Hinkler Park (HP; n = 77) in the Bundaberg region. Methods for assessing some yield component traits were previously published in O'Connor et al. (2018a), and methods are detailed for all traits here. Each of the 295 progeny, as well as 18 of the 29 parents, were measured for yield and twelve yield component traits from August 2016 to July 2018 (Table 3-1).

Table 3-1 List of traits measured on 295 progeny and parents

Code	Trait	Unit
ENF	Estimated number of florets per raceme	count
FSN	Flowers that set nuts (NPR / ENF)	%
KR	Kernel recovery (KW / NW)	%
KW	Kernel weight	g
NPR	Number of nuts per rachis	count
NW	Nut-in-shell weight	g
PD	Nut pedicel diameter	mm
RDN	Rachis diameter at nut set	mm
RL	Raceme length	cm
RSN*	Racemes surviving from flowering to nut set	%
TC	Trunk circumference	cm
WK	Whole kernels	%
Yield*	Nut-in-shell yield	g

* denotes measured over two seasons at some/all sites

3.3.2 Phenotyping

Trunk circumference (TC) was measured 50 cm above the ground, or below the skirt of low-branching trees as an estimate of tree size. Ten racemes (approximately at looping stage, Figure 3-2) were randomly selected and removed from each tree for subsequent measurements. Raceme length (RL) was measured from first to last floret, excluding isolated florets, as a measure of the reproductive section of the raceme (Figure 3-2). Due to the impracticality of counting all florets on all racemes, O'Connor et al. (2018a) developed an equation to estimate the number of florets of a single raceme:

$$ENF = (4.92 \times RL) + (0.12 \times F5 \times RL) - 20.4 \qquad \text{Equation 3-1}$$

(adjusted $r^2 = 94.3\%$, s.e. = 18.6, two terms $p < 0.001$), where F5 denotes the number of florets per 5 cm at the terminal end.

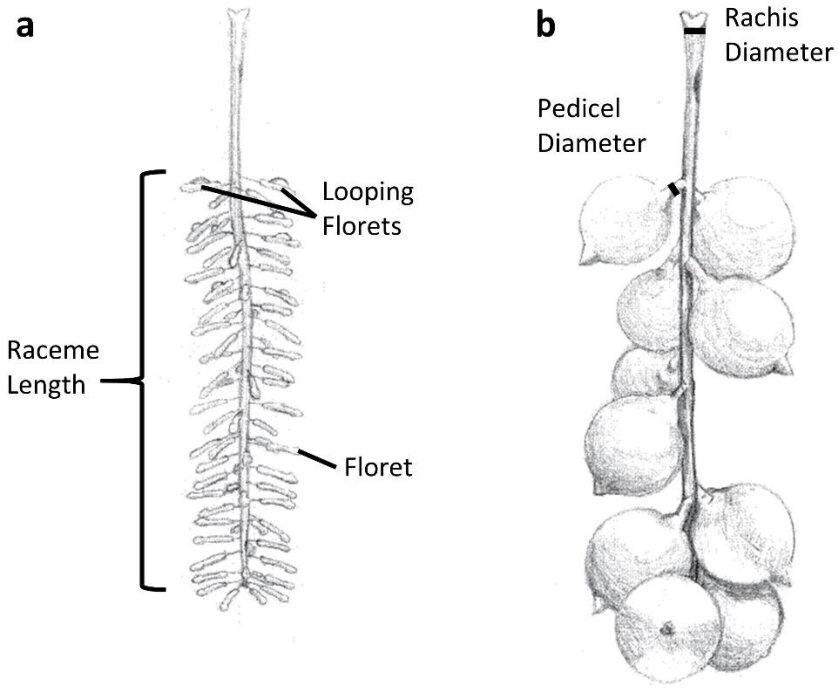


Figure 3-2 Macadamia racemes with component traits indicated. **a** Raceme with florets in flower, with some at looping stage, raceme length measured from first to last floret. **b** Rachis at nut set, rachis diameter measured at 'waist' or approximately 3 mm from base of rachis, and pedicel diameter measured at midway point. Illustrations by Todd Fox, altered from O'Connor et al. (2018b)

During flowering (September in Australia), at least ten racemes were tagged per tree. These racemes were on two branches per tree, in a length of 30 cm at a distance of 20 cm in from the end of the branch. Extra racemes were tagged outside of these branches, if necessary, to give a total of ten. At nut set, the number of racemes bearing nuts on tagged branches was counted. After adjusting for any loss of branches or tags (due to weather events and machinery), the percentage of racemes per tree surviving from flowering through to nut set (RSN) was calculated. Ten rachises and 15 nuts per tree were collected from each tree in 2017, where available. The number of nuts per rachis was recorded (NPR) in February at the beginning of harvest season, where nuts were mature but not yet dehiscing. The number of florets that set nuts (FSN) was calculated as the average NPR / ENF for each tree. Nut pedicel diameter (PD) was measured using calipers on ten nuts, along the suture line at the midway point (Figure 3-2). Rachis diameter at nut set (RDN) was measured on ten rachises using calipers, at the base (Figure 3-2).

Nuts-in-shell were harvested by hand for each tree, as a measure of yield. Site AL was not harvested in 2017 due to an extreme weather event, whilst site EG was not harvested in 2018 due to management issues. Nuts on the ground under trees were collected, and nuts remaining in the tree were removed using poles with hooks. Nuts were dehusked within three days after each harvest, and wet NIS weight recorded. A 1 kg sample was taken (where available), and dried in ovens for two days at 35°C, two days at 45°C, followed by a final two days at 55°C, as per protocol described by Prichavudhi and Yamamoto (1965). Whole harvest dry NIS weight was calculated based on the moisture content of the 1 kg sample. Total yield per tree was estimated as a sum across harvests in each year.

Nuts were stored in airtight containers, and re-dried at 50°C overnight before being cracked and measured. Nuts were weighed separately for nut weight (NW) and then cracked manually using a lever-type macadamia cracker. Nuts and kernels with visible insect damage or kernel shrivelling were discarded and not measured. For retained nuts, shell and kernel pieces were weighed separately, and kernel weight (KW) was recorded for each fruit. The percentage of whole kernels (WK) from the sample was recorded; whole kernels are those that do not split down the interface separating the two cotyledons when cracked (Walton et al. 2012) (Figure 3-1). Kernel recovery (KR) was calculated for each nut as KW / NW. A sample of 20 good-

quality nuts (without insect damage or kernel shrivelling) was measured for each tree, where available.

Means of each trait were calculated for each individual for each yield component. Data were not included for a component trait for any tree with five or fewer units (e.g. kernels, nuts, racemes) measured, meaning that for some traits very low yielding trees were not examined due to insufficient racemes or nuts. The number of trees with missing data ranged from 2–21 for primary traits, and 8–55 for traits derived from calculations of primary traits; FSN had the highest rate of missing data (n = 55) due to the trait being a derivative of RL, F5 and NPR. Yield and RSN were measured for each tree across two years. Flowering data were measured in Spring (September in Australia) of 2016, and nut data in Autumn–Winter (April–August) of 2018. Phenotypic minimum, maximum, mean and standard deviation (SD) were calculated across all trees for each trait.

3.3.3 Genotypic data

This study used genetic markers developed by Diversity Arrays Technology (DArT), as described in O'Connor et al. (2019b; Chapter 2 Population structure and genetic diversity) and briefly outlined here for completeness. Single nucleotide polymorphism (SNP) markers were detected and imputed with 97% accuracy using the PPCA method (Stacklies et al. 2007). Quality control was then performed (based on pre-imputation parameters), where markers with <50% call rate, <2.5 minor allele frequency, and those that failed a test of Mendelian inconsistencies using a comparison of parent-offspring trio homozygotes in 50% of families, were removed. A total of 4,113 high quality SNPs were used in genetic analyses. An additive GRM was constructed using methods as per VanRaden (2008), modelled in R from SNP effects, where homozygous, heterozygous and alternate homozygous genotypes were represented by 0, 1, and 2, respectively. The GRM was included in each analysis to model kinship.

3.3.4 Individual site models to test for normality

A preliminary single site analysis was undertaken with linear mixed models to confirm the normality of the distribution of residuals using a Shapiro-Wilk test (Shapiro and Wilk 1965).

All observations were first scaled per site per year, by dividing by site/year SD, to reduce the impact of heterogeneity of variances on genotype by year and genotype by site interactions (Hardner 2017; Hill 1984). Models were implemented using ASReml-R (Butler et al. 2009):

$$\mathbf{y} = \mathbf{X}\mathrm{b} + \mathbf{Z_g}\mathrm{g} + \mathrm{e} \qquad \text{Equation 3-2}$$

where $\mathbf{y}$ is a vector of (scaled) phenotypes, $\mathbf{X}$ is a design matrix that assigned fixed effects (block within site, year, tree type = seedling progeny or grafted parent clone, and interactions) to observations, $\mathbf{b}$ is a vector of fixed effects, $\mathbf{Z_g}$ is a design matrix that allocates random effects to observations, $\mathbf{g}$ is a vector of additive genetic effects of the individuals, assumed random ~N(0,$\mathbf{G}\otimes\sigma_g^2$), where G is the GRM and σ_g^2 is the genetic variance captured by the SNP (modelled as 0, 1, or 2 for AA, AB and BB genotypes, respectively, hence the model is additive), and $\mathbf{e}$ is a vector of random errors ~N(0, σ_e^2) where σ_e^2 is the error variance. For sites/traits where two years of data were available, a permanent environment effect of individual trees was included in the model by incorporating an interaction between Site and Tree as a random term.

Traits that were not normally distributed were transformed to approximate normality of within-site residual distribution, and then scaled. Transformations were either square root or $\log_{10}(x+1)$, where x represents each raw phenotype, depending on which method provided the most normal distribution of residuals. Scaled and transformed, if appropriate, observations were incorporated as phenotypes in subsequent analyses.

To test the significance of the interaction between genetic effect and assessment year for traits measured over two years (yield and RSN), the following model was used:

$$\mathbf{y} = \mathbf{X}\mathrm{b} + \mathbf{Z_g}\mathrm{g} + \mathbf{Z_{gy}}\mathrm{gy} + \mathrm{e} \qquad \text{Equation 3-3}$$

where terms are as per Equation 3-2, and $\mathbf{X}$ is a design matrix that assigned fixed effects (overall site mean, block within site, year, tree type = seedling progeny or grafted parent clone, and interactions) to observations, $\mathbf{Z_{gy}gy}$ is genotype by year interaction where $\mathbf{Z_{gy}}$ is a design matrix allocating a specific effect of an individual in one year not accounted for by the mean of the genetic effect of the individual across years, and $\mathbf{gy}$ is a vector of the deviation of breeding values at a specific year, assumed random ~N(0,$\mathbf{G}\otimes\mathbf{I}_2\otimes\sigma_{gy}^2$) where $\mathbf{I_2}$ is a 2x2 identity matrix for the two years. For both yield (at AM and HP) and RSN, the

correlation of genetic effects between years was approximately statistically not different to one (z-ratio < 1.96), so the interaction between year and genetic effects was removed from further analyses. A permanent environment effect was again included for sites/traits where two years of data were available.

3.3.5 Multi-site models to estimate variance components and heritability

To determine whether sites could be combined to represent a common environment over the four site locations, and to estimate variance components, the following model was used:

$$\mathbf{y} = \mathbf{X}\mathrm{b} + \mathbf{Z_g}\mathrm{g} + \mathbf{Z_{gs}}\mathrm{gs} + e_s \qquad \text{Equation 3-4}$$

where the terms are as per Equation 3-2, and $\mathbf{Z_{gs}}$ is a design matrix allocating a specific effect of an individual at a site not accounted for by the mean genetic effect of the individual across sites, and $\mathbf{gs}$ is a vector of the breeding values at a specific site, assumed random ~N(0,$\mathbf{G}\otimes\mathbf{I_4}\otimes\sigma^2_{gs}$) where $\mathbf{I_4}$ is a 4x4 identity matrix for the four sites, and $\boldsymbol{e_s}$ is a variance-covariance structure of residual errors for each site ~N(0,$\mathbf{I_s}\otimes\sigma^2_{es}$) where $\mathbf{I_s}$ is an identity matrix for the sth site and σ^2_{es} is the error variance at that site. Z-ratios of the variance component estimate was used as an approximate test of the significance (> 1.96) of the G x E (site) term.

Variance components were calculated for each trait using Equation 3-4 above, and partitioned among additive genetic, G x E (site), permanent environment effect (for traits/sites with two years of data), and site residual variance. Narrow-sense heritability on a single tree basis was estimated as:

$$h^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_{gs} + \sigma^2_p + \sigma^2_e} \qquad \text{Equation 3-5}$$

with standard error estimated using a linear approximation of a Taylor series expansion (pin.R function in ASReml; Kendall et al. 1987; White 2013), where σ^2_g is the additive genetic variance, σ^2_{gs} is the G x E (site) variance, σ^2_p is the site-specific permanent environment effect for traits/sites with two years of data, and σ^2_e is the residual variance at a particular site. The genetic coefficient of variation was measured for each trait as the additive genetic variance

σ_g^2 divided by the phenotypic mean across all individuals (where the means for yield and RSN were taken over both years).

3.3.6 Bivariate models to estimate genetic correlations

Multi-site models were extended to incorporate multiple traits. Pairwise bivariate models were used to estimate genetic correlations between traits, using a structured covariance matrix:

$$\begin{bmatrix} y_s \\ y_{s\prime} \end{bmatrix} = \begin{bmatrix} X_s & 0 \\ 0 & X_{s\prime} \end{bmatrix} \begin{bmatrix} b_s \\ b_{s\prime} \end{bmatrix} + \begin{bmatrix} Z_{g1} & 0 \\ 0 & Z_{g2} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} + \begin{bmatrix} e_s \\ e_{s\prime} \end{bmatrix} \quad \text{Equation 3-6}$$

where y_s is a vector of observations for both traits (i.e. yield and the other trait) at site s, $\mathbf{X}$ is a design matrix that assigned fixed effects at each site to observations, s is one site and s' is another site, $\mathbf{b}$ is a vector of fixed effects for both traits at each site, and $\mathbf{Z_g}$ is as per Equation 3-4. It was assumed that $\begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$ ~N(0,**G**⊗T), where g_1 is the vector of unknown average genetic effects across sites for yield and g_2 is the vector of unknown average genetic effects across sites for the other trait, T = $\begin{bmatrix} \sigma_{g1}^2 & \sigma_{g12}^2 \\ \sigma_{g12}^2 & \sigma_{g2}^2 \end{bmatrix}$, the structured additive genetic variance-covariance matrix of two traits, and $\begin{bmatrix} e_s \\ e_{s\prime} \end{bmatrix}$ ~N(0,**I**⊗**R**), where $\mathbf{I}$ is an identity matrix of individuals and R = $\begin{bmatrix} R_{ss} & 0 \\ 0 & R_{s\prime s\prime} \end{bmatrix}$, the structured residual variance-covariance matrix of the two traits at each site, and R_{ss} is the non-zero structured variance-covariance matrix of two traits at the s^{th} site. A permanent environment effect was again included for sites/traits where two years of data were available. The term $\mathbf{Z_{gs}gs}$ was included as per Equation 3-4 for traits where there was an interaction with site. Standard error for the genetic correlation was obtained directly from the ASReml variance component output.

3.3.7 Selection efficiency

Efficiency (E) of indirect selection of high yield using a component trait compared with direct selection of yield was calculated for each component trait using the following equation from Falconer (1989):

$$\frac{CR_X}{R_X} = \frac{h_Y \times r_g}{h_X}$$ Equation 3-7

where CR_X is the correlated indirect response of directly selecting yield, R_X is the direct response of selecting yield, h is the accuracy of individual selection the component trait (measured here as the square root of the heritability, h^2), r_g is the genetic correlation between the two traits, X is yield and Y is the yield component trait.

3.4 Results

3.4.1 Population phenotypic variation

Phenotypes for each component trait varied across families and site locations. For example, TC ranged from 14 cm in family 'NG8' x '797' to 78 cm in 'NG18' x '660' (Table 3-2). Very low NW and KW were observed in family 'NG18' x '695' (NW 4.34 g) and 'NG8' x '762' (1.46 g KW), with an average of 7.09 g and 2.73 g, respectively. Average KR was 38.7% and ranged widely from 20.2% in 'A38' x '246' to 55.6% in 'L64' x '344'. FSN ranged from 0.4 to 7.2%, whilst RSN showed the entire range of phenotypes, from zero to 100% across sites. Boxplots (Figure 3-3) strongly suggest variance of yield observations were heterogeneous among sites. Variances were also slightly heterogeneous across sites for TC. The highest average yield was observed at HP for both years, with the average and variance of yield being lower at AM and EG (Table 3-2, Figure 3-3). TC was also low on average at EG, whereas there were larger trees observed at AL and HP. Patterns for distribution of most traits, particularly PD, were similar across sites (Figure 3-3).

Table 3-2 Summary of phenotypes for yield and yield component traits across all individuals and all sites: raw untransformed maximum, minimum, average, standard deviation (SD). The family of the tree (or cultivar, cv.) with the lowest and highest values is shown. Multiple indicates that there was more than one family or cultivar with that value

Trait	Minimum		Maximum		Average	SD
	Value	Family/cv.	Value	Family/cv.		
ENF	26	'660' x '783'	376	'D4'	139	49
FSN (%)	0.4	'D4'	7.2	'783' x '804'	2.0	1.2
KR (%)	20.2	'A38' x '246'	55.6	'L64' x '344'	38.7	0.05
KW (g)	1.46	'NG8' x '762'	5.01	'1/40' x '849'	2.73	0.55
NPR	1	multiple	10.4	'333' x '842'	2.6	1.4
NW (g)	4.34	'NG18' x '695'	12.31	'NG35' x '791'	7.09	1.34
PD (mm)	1.85	'D4' x '695'	5.35	'333'	3.47	0.55
RDN (mm)	1.76	'A16' x 'NG4'	7.75	'333' x '842'	3.61	0.85
RL (cm)	3.6	'660' x '783'	32.3	'D4'	12.0	3.9
RSN (%)	0	multiple	100	multiple	25	24
TC (cm)	14	'NG8' x '797'	78	'NG18' x '660'	51	12
WK (%)	15	'NG8' x '333'	100	multiple	64	17
Yield 2017 (g)	5	'344' x '804'	26,623	'D4' x '660'	4,737	5,499
Yield 2018 (g)	36	'A9/9' x '814'	25,967	'1/40' x '849'	5,910	5,422

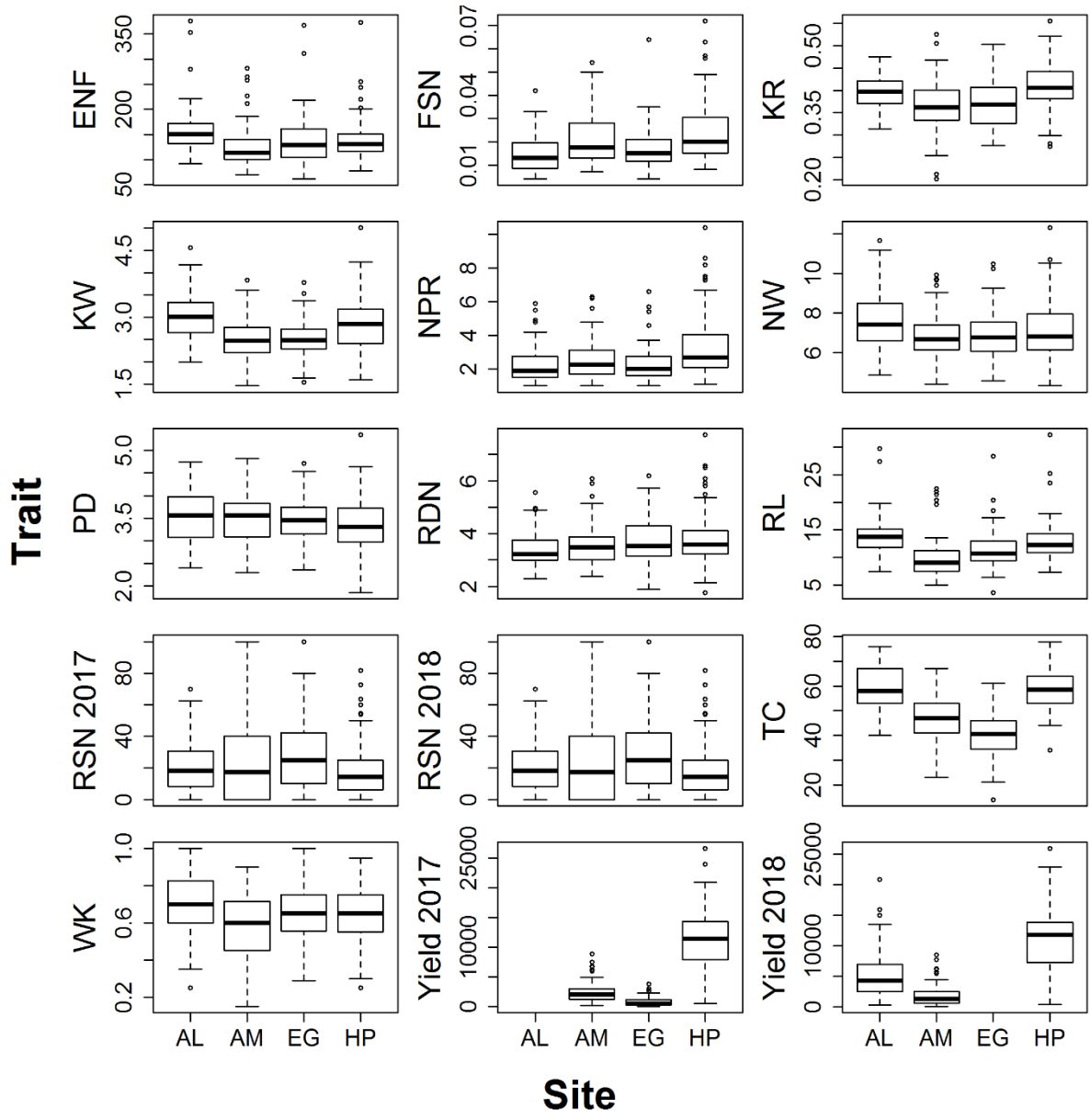


Figure 3-3 Boxplots showing distribution of phenotypes for yield and yield component traits across the four sites. Yield was not measured at AL in 2017, or at EG in 2018

3.4.2 Individual site models to investigate normality

Individual site models were successfully used to test the normality of residuals using various data transformations for each trait (data not shown). Square root transformations led to normalised residuals for FSN, KW, RL, RSN and yield. Transformation by $\log_{10}(x+1)$ was used for traits ENF, NW, NPR and RDN to achieve normality. Scaling of data was sufficient to achieve normality for KR, PD, TC and WK.

3.4.3 Multi-site models to estimate variance components and heritability

Multi-site models showed that the proportion of phenotypic variance explained by different factors varied between traits (Figure 3-4). Proportion of additive variance was similar across sites within most traits for the individual site models, particularly for PD (0.55 at AL to 0.62 at HP), RSN (0.02 at HP to 0.10 at AL), and WK (0.21 at AL to 0.23 at AM and HP). These proportions of additive variance were used to calculate individual narrow-sense genomic heritability. Heritability ranged from 0.09 for RSN to 0.76 for KR, with yield being moderately low at 0.31 (Table 3-3).

Very little G x E (site) variance was observed for most traits (Figure 3-4). In comparison, variance attributed to G x E was approximately significant (z-ratio > 1.96) for traits NPR and RDN (Figure 3-4). Permanent environment variance was measured only for RSN and for yield at AM and HP (where two years of data were available), and was negligible for RSN in all sites except AL. A substantial proportion of variance was attributed to permanent environment effect for yield at HP (0.38; Figure 3-4).

3.4.4 Bivariate models to estimate genetic correlations

Estimated genetic correlations with yield varied among traits from -0.27 for KR to 0.99 for RSN, with an average of 0.31 (Table 3-3). KR was the only trait negatively genetically correlated with yield (r_g = -0.27). TC was highly correlated with yield (0.72), whilst the estimated genetic correlation of RSN and yield was bounded at 0.99. RDN, NPR, NW and FSN were all moderately genetically correlated with yield (r_g = 0.56, 0.55, 0.45 and 0.41, respectively). WK and ENF were least correlated with yield, both at 0.03 (Table 3-3).

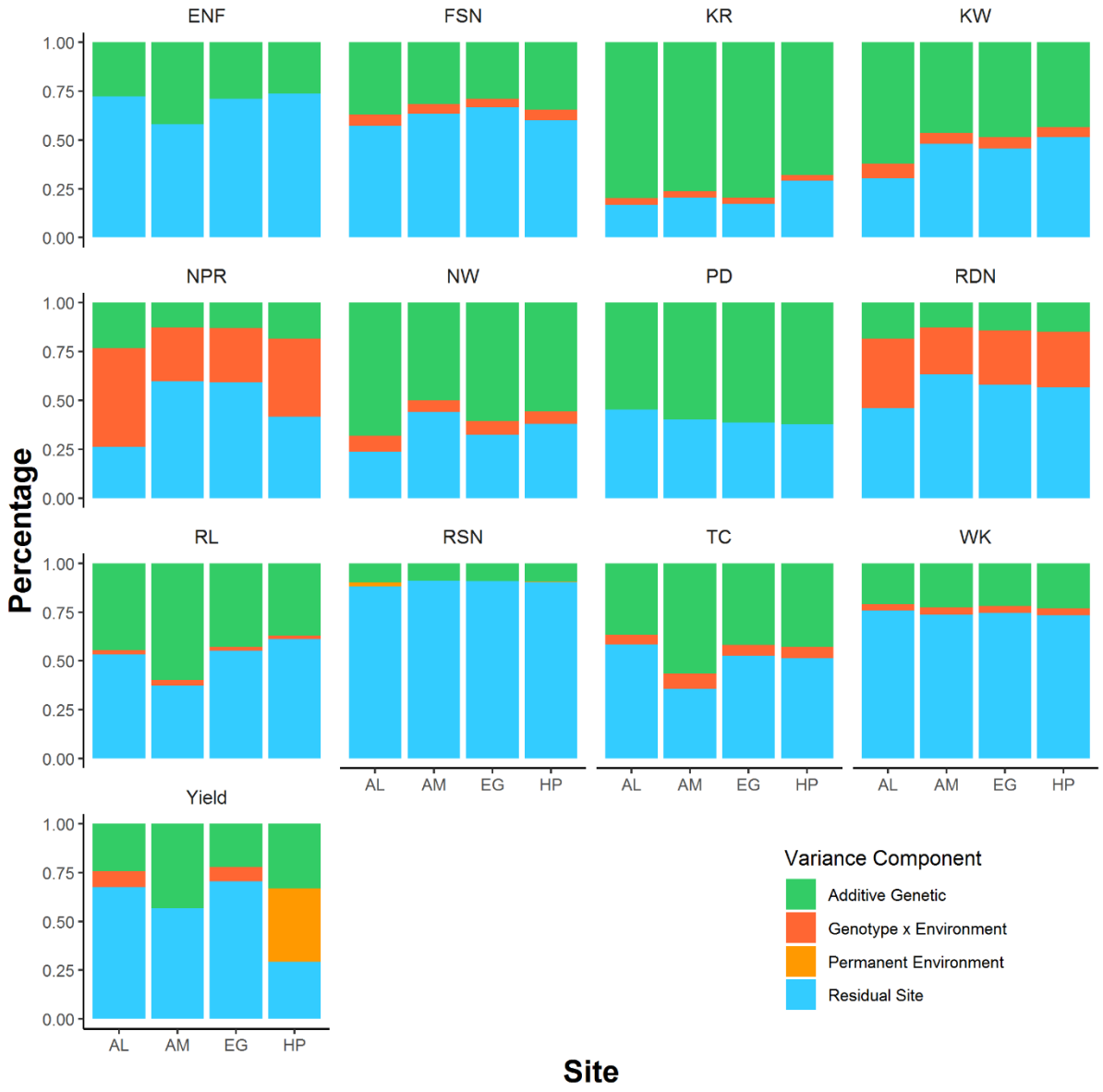


Figure 3-4 Estimated proportion of phenotypic variance due to genetic and non-genetic sources for yield and yield component traits across the four sites (from multi-site model described in Equation 3-4, where genetic and genotype x environment variances are common across all sites). Permanent environment effect was only estimated for sites/traits with two years of data available (RSN, and yield at AM and HP). Residual site variance component (blue) represents the residual (error) variance at each individual site

Table 3-3 Estimated narrow-sense heritability of each trait (h^2), genetic correlations (additive) of component traits with yield as estimated from bivariate analyses (r_g), genetic coefficient of variation measured as the additive genetic variance σ_g^2 divided by the trait phenotypic mean (GCV), and selection efficiency for indirect selection of yield through component trait (E). Standard errors of heritability and genetic correlations are shown in parentheses

Component Trait	**h^2**	**GCV**	**r_g**	**E**
ENF	0.42 (0.10)	0.00	0.03 (0.22)	0.04
FSN	0.33 (0.11)	0.14	0.41 (0.21)	0.42
KR	0.76 (0.09)	0.02	-0.27 (0.18)	-0.42
KW	0.50 (0.10)	0.15	0.23 (0.21)	0.29
NPR	0.17 (0.11)	0.06	0.55 (0.18)	0.40
NW	0.59 (0.11)	0.07	0.45 (0.18)	0.62
PD	0.60 (0.10)	0.11	0.23 (0.18)	0.32
RDN	0.15 (0.09)	0.03	0.56 (0.17)	0.39
RL	0.46 (0.10)	0.02	0.22 (0.20)	0.27
RSN	0.09 (0.05)	0.00	0.99 (NA)	0.54
TC	0.44 (0.10)	0.01	0.72 (0.12)	0.86
WK	0.22 (0.11)	0.00	0.03 (0.27)	0.03
Yield	0.31 (0.12)	0.00		1.00

3.4.5 Selection efficiency

Estimates of (absolute) selection efficiency using component traits to indirectly select for high yield varied from 0.03 (WK) to 0.86 (TC) (Table 3-3). The next highest selection efficiencies were for NW (0.62) and RSN (0.54), although the genetic correlation between RSN and yield was bounded at 0.99. Very low selection efficiency was also observed for ENF (0.04), whilst the negative genetic correlation between yield and KR resulted in -0.42 selection efficiency for that trait. None of the traits had a higher selection efficiency for indirect selection than direct selection for yield (1.00).

3.5 Discussion

3.5.1 Model fit and selection

This is the first study to use linear mixed models (LMM) combined with genetic marker data to estimate heritability and genetic correlations among a relatively large number of yield component traits for a macadamia breeding population. LMM methods have been used in a variety of crops, with relationships among individuals modelled using recorded pedigree, as in walnut (Martínez-García et al. 2017), blueberry (Cellon et al. 2018), and previously in macadamia (Hardner 2017; Hardner et al. 2019a), or using genetic markers, for example in sweet cherry (Piaskowski et al. 2018; Hardner et al. 2019b). This study used LMM to correct phenotypic observations for effects such as site and tree type. Variation of observations differed across the four study locations for some traits, and particularly for yield across two years of data. Raw observations of yield were much higher for trees at HP and AL, both in the Bundaberg region of Queensland, compared to those at AM and EG, in the Gympie region, which may be a result of different soil types and growing conditions in the regions. LMM were used to adjust for site means to combine data across sites for many traits. G x E (site) was only significant for traits NPR and RDN, which may be due to interactions between genotypes and efficient use of resources in different environments. Only one year of data were available for these traits as well as nine others, and so G x E (site) variance was confounded with residual site variance. The preliminary models evaluated in this study were used to optimise subsequent bivariate models in order to estimate genetic correlations with yield. Future analyses could incorporate data over multiple years, and results could be compared between analyses using relationship matrixes constructed using pedigree data, and additive, dominance and epistatic genomic data.

3.5.2 Variability and heritability of yield and yield component traits

Few other studies have measured these component traits on a diverse collection of germplasm. In their study of 15 varieties and selections at 4–5 years since planting, Toft et al. (2019) found that phenotypes for RL ranged from 2–46 cm, with an average of 16 cm, whilst NPR varied from 0.2–2.6. These findings do not reflect the phenotypic ranges observed in the current study (RL 3.6–32.3 cm, NPR 1–10.4), which may be explained by different germplasm,

environments and slightly different measuring techniques. RL was moderately heritable in the current study (h^2 = 0.49), which concurred with a previous study of 15 young (4–5 years since planting) cultivars (H = 0.68), despite their estimate of heritability including non-additive genetic variance (Toft et al. 2018). Heritability of NPR was lower (0.17) than expected in the current study, based on the phenotypically-observed consistency of nuts per cluster within a tree, but was similar to an estimate of broad-sense heritability by Toft et al. (2018) (H = 0.11).

Previous research investigated the proportion of six racemes that set nuts (RSN) in a population of ten varieties (Boyton and Hardner 2002); just over half (53%) of racemes successfully set nuts, compared with an average of 25% in the current study. However, very little additive genetic variance was observed for RSN, with a heritability of 0.09, and instead was largely influenced by residual site variation. Previous calculations of the percentage of florets that set nuts (FSN) were largely based on broad estimates using the number nuts per raceme and flowers per tree, at 0.3% (Ito 1980). Measuring techniques used here are presumed to be much more accurate, and phenotypes ranged from 0.4 to 7.2%. While outside the scope of this study, it would be interesting to explore the physiological reasons behind this seemingly low fruit set efficiency, to determine what the limiting factors are.

The estimate of narrow-sense heritability in this study for yield (h^2 = 0.31) is similar to that previously estimated for NIS yield at age 7 years for progeny generated from a different set of parents, ranging from 0.35–0.46 across sites using pedigree data (Hardner 2017). However, these values are much higher than estimates of broad-sense heritability for annual and cumulative NIS yield in a regional variety trial (RVT), which ranged from 0.06 to 0.20 across ages 4 to 10 (Hardner et al. 2002). This lower estimate of heritability may be due to the much lower genetic diversity represented in an RVT than a progeny trial, as far fewer genotypes are included in an RVT, and these genotypes have been selected for high yield compared to the other progeny. Narrow-sense heritability of yield in the current study was very low compared with broad-sense heritability in pecan (H^2 = 0.85) (Kumar et al. 2013a), although the broad-sense heritability for pecan incorporates additive as well as non-additive genetic variance into the estimate. Walnut yield heritability varied greatly from a moderate estimate in one study (h^2 = 0.54) (Martínez-García et al. 2017) to negligible in another (h^2 = 0.07) (Hansche et al. 1972), which may be due to differences between the populations in the studies, measurement methods, or propagation method.

KR was the most heritable of component traits in the current study (h^2 = 0.76), and was similar to Hardner et al. (2001), where broad-sense heritability was estimated at 0.66 for individual tree measurements. The high heritability for this trait means that increasing the average KR in breeding populations (38.7% in this population) towards desirable levels (>40%; Topp et al. 2019) is achievable through selection. KR was also highly heritable in hazelnut, where narrow-sense heritability, estimated through regression of offspring means on mid-parent values, was 0.87 (Yao and Mehlenbacher 2000). The estimated narrow-sense heritability for NW and KW in this study (0.66 and 0.51, respectively) was lower than broad-sense heritability estimates of these traits in another macadamia population (0.64 and 0.66, respectively) (Hardner et al. 2001), and much higher than that estimated for another population (broad-sense H = 0.27) (Toft et al. 2018). In comparison with the current study, heritability for NW and KW was similar in hazelnut (0.63 and 0.67, respectively) (Yao and Mehlenbacher 2000), and slightly lower than that in walnut (NW 0.86, KW 0.87) (Hansche et al. 1972), where both studies calculated heritability by regressing means of offspring on the average parent value.

3.5.3 Genetic correlations with yield

The slight negative correlation between KR and yield (r_g = -0.27) in the current study concurred with previous result; Hardner et al. (2002) also found that clonal genetic correlations between cumulative NIS yield and KR ranged from -0.06 at age 6 to -0.37 at age 10. In pecan, KR was virtually uncorrelated with yield (r_p = 0.09) (Kumar et al. 2013a), though this is phenotypic and not genetic correlation, and may be due to biennial bearing in the crop. The results of the current study imply that selecting for high KR may tend to lead to lower yields, and vice-versa, and may have implications for selection in the breeding program, since both high yield and high KR are selection priorities. However, this negative relationship does not imply causation, and the correlation is not strong, so it may be possible to identify candidate cultivars that possess both high yields and high KR.

Our finding of low to moderate additive genetic correlation between yield and KW (0.21) does not concur with Hardner et al. (2002), who found a negative total genetic correlation between cumulative NIS yield and kernel mass (-0.14 to -0.25). These inconsistencies in correlation estimates may be a result of population differences between the two studies (progeny trial compared with RVT, respectively), or the method of modelling kinship among individuals

(including an additive GRM compared with no inclusion of pedigree). Hansche et al. (1972) also found that yield decreased with increased walnut weight (r_p = -0.20), whilst cashew nut weight was poorly phenotypically correlated with yield (r_p = 0.11) (Aliyu 2006). Since yield in macadamia is generally based on NIS weight across the whole tree, the positive genetic correlations between yield and both NW (0.40) and KW (0.21) were expected, since heavier kernels and nuts will lead to heavier tree yields. Since NW and KW are both traits with intermediate optimums (O'Hare et al. 2004), selecting for large nuts to increase yield is not necessarily desirable, as current machinery may be unable to process very large nuts. However, marker requirements fluctuate, and in the future, large nuts may be a marketable option (B. Topp, pers. comm.).

The aim to develop smaller trees with high productivity (Toft et al. 2019) may be difficult for macadamia, due to the strong positive correlation between TC and yield (0.72), which implies that larger trees produce more nuts. The results of the current study conflict with previous estimates of correlations between stem girth and cumulative NIS, which ranged from -0.38 to 0.22, depending on tree age (Hardner et al. 2002). In comparison, recent work in young trees found that trunk cross-sectional area was significantly correlated with cumulative yield (r_p = 0.45, $p < 0.001$) (Toft et al. 2019), though this was an estimate of phenotypic rather than genetic correlation.

Our estimate of a virtually perfect correlation between RSN and yield is likely due to the very low genetic variance observed for RSN, as explained in general by Falconer (1989): genetic correlation is based on shared genetic variance between two traits, and if one trait has very low genetic variance then it is difficult to accurately estimate the amount of shared genetic covariance with other traits. This result, therefore, may not be an accurate estimate of the true genetic correlation between these traits. In comparison, the moderate genetic correlations observed between yield and FSN, RDN and NPR (0.50, 0.53 and 0.56, respectively) suggest that selecting for any of these traits may lead to an increase in yield, due to shared additive variance (Falconer 1989). The genetic correlations with yield observed for RL (0.22), ENF (0.09) and NPR differed from those of the same or similar traits in other studies. For example, RL was more highly correlated in a previous study of young macadamia trees (r_p = 0.58, $p < 0.001$) (Toft et al. 2019). Aliyu (2006) found that cashew nut yield increased significantly with both number of hermaphrodite flowers per panicle (r_p = 0.863, $p < 0.01$) and

number of nuts per panicle (r_p = 0.844, p < 0.01). However, some of these studies only estimated phenotypic correlations and did not use genetic marker data to estimate genetic correlations.

3.5.4 Selection efficiency and implications for breeding

As trait heritability and genetic correlation with yield were used to calculate selection efficiency, the moderate to high estimated values for these parameters for TC led to the highest calculated selection efficiency of all the traits, at 0.86. This was followed by NW at 0.58 with moderate values for h^2 and r_g. A previous estimate of selection efficiency for yield using NW was 0.65 (O'Connor et al. 2018b), though this was calculated using broad-sense heritability. The third highest selection efficiency was for RSN (0.53), though this is probably biased by a low estimated genetic variance resulting in very high (and imprecisely estimated) genetic correlation with yield. The selection efficiencies calculated for TC and NW suggest they could be contenders for indirectly selecting for high yield, if the cost of assessing these traits was much lower than assessing yield, as neither were equal to or more efficient than directly selecting for yield. However, as previously mentioned, it is not in the interest of the breeding program to select for larger trees or nut sizes. As such, the results here indicate that none of the studied component traits are suitable for use in indirectly selecting for high yield through correlated response to selection.

Although this study did not successfully identify any component traits that would by themselves be suitable for the indirect selection of high yield, the estimates of heritability and correlations with yield will be useful for future breeding efforts. As this study employed genetic markers to model kinship among individuals, these estimates will be more accurate than previous studies using recorded pedigree data, as suggested by Hayes et al. (2009b). The genetic correlations of yield with traits such as TC (0.72) and KR (-0.27) are informative, as breeders will need to identify seedling progeny that do not adhere to the general correlated observed here; breeding efforts should instead try to identify small trees with high yields, and trees with high KR and also high yields. Given the high heritability of KR, and only low genetic correlation with yield, it is expected that breeders will be able to actively breed for high KR and also select trees that are high yielding.

Some limitations existed in the current study that may have influenced estimates of heritability and genetic correlations with yield. The majority of yield component traits were only measured for one season, and there may be environmental changes across years that have not been accounted for here. Furthermore, measuring yield on a per-tree basis is complicated due to overlapping canopies of neighbouring trees at some sites. Since differences in site locations influenced some traits in the current study, and since yield can vary between years, future research could investigate more component traits over a number of seasons to improve accuracy of estimates. However, given that costs of genotyping are decreasing, selection efforts may become more focussed on strategies using genetic data rather than relying on time consuming and laborious phenotyping. Genomic selection for complex traits like yield, and genome-wide association studies combined with marker-assisted selection for economically important traits are suitable for use in macadamia (O'Connor et al. 2018b). A preliminary study using the same population as the current research identified genetic markers associated with NW and KW (O'Connor et al. 2019a). Using these genomics-based breeding efforts, genetic gain for yield and nut traits could be increased by selecting elite individuals earlier than previously possible, thus reducing the selection cycle.

3.6 Conclusions

The diversity of phenotypes in the population as well as the genetic variation supports the use of breeding and selection for genetic improvement of candidate cultivars. Bivariate GBLUP analyses, as used in this study, offer a method to ascertain genetic relationships among traits and predict breeding values by means of realised kinship among individuals using genetic markers. Estimates of heritability and genetic correlations between traits will inform future breeding efforts, regarding the ease of selecting for certain traits. The results here indicate that the efficiency of indirect selection for yield using the component traits measured in this study is likely to be low. The two traits with highest selection efficiencies, TC and NW, are not appropriate for indirect selection as an aim of the breeding program is to select small-sized trees with moderately sized nuts. Whilst indirect selection for high yield may not be efficient with these traits, they should still be investigated in genomic studies such

as genome-wide association studies to identify desirable and economically important traits in progeny at an early stage.

Chapter 4. Genome-wide association studies for yield component traits in a macadamia breeding population

---

The following chapter was conducted in two parts: a pilot study and the chapter presented. The two studies used the same population but different phenotypic and genotypic datasets.

The pilot study used historical nut measurements, and a set of SNP markers derived from a wider *Macadamia* population. This study has been published.

O'Connor, K., Hayes, B., Hardner, C., Alam, M., and Topp, B. 2019. Selecting for nut characteristics in macadamia using a genome-wide association study. HortScience 54(4): 629-632. https://doi.org/10.21273/HORTSCI13297-18

The chapter presented here used phenotypic data collected during the doctoral study, and a genotypic dataset using only those cultivars/progeny in the study with missing data imputed. This is the same genetic marker dataset described in Chapter 2: Population Structure and Genetic Diversity.

My contribution to the chapter and publication

I collected phenotypic data for the chapter (but not the publication), wrote the papers, performed all analyses and made final edits. BH, CH and MA assisted in interpretation of results. Cathy Nock and Abdul Baten provided genome assembly scaffold data in the chapter. BH, BT, CH, MA and CN suggested revisions. We acknowledge those involved in the collection of historical data.

4.1 Abstract

Breeding macadamias for new cultivars with high nut yield is expensive in terms of time, labour and cost. Most trees set nuts after four to five years, and candidate varieties for breeding are evaluated for at least eight years for various traits. Genome-wide association studies (GWAS) are promising methods to reduce evaluation and selection cycles by identifying genetic markers linked with key traits, potentially enabling early selection through marker-assisted selection. This study used 295 progeny from 32 full-sib families and 29 parents (18 phenotyped) which were planted across four sites, with each tree genotyped for 4,113 SNPs. ASReml-R was used to perform association analyses with linear mixed models including a genomic relationship matrix to account for population structure. Traits investigated were: nut weight (NW), kernel weight, kernel recovery (KR), percentage of whole kernels (WK), tree trunk circumference (TC), percentage of racemes that survived from flowering through to nut set, and number of nuts per raceme. Seven SNPs were significantly associated with NW (at a genome-wide false discovery rate of <0.05), and four with WK. Multiple regression, as well as mapping of markers to genome assembly scaffolds suggested that some SNPs were detecting the same QTL. There were 44 significant SNPs identified for TC although multiple regression suggested detection of 14 separate QTLs. These findings offer the opportunity to use marker-assisted selection in macadamia breeding.

4.2 Introduction

Macadamia is a large nut tree native to the coastal rainforests of southern Queensland and northern New South Wales, Australia. *Macadamia integrifolia* Maiden & Betche, *M. tetraphylla* L.A.S. Johnson and their hybrids have high-quality edible kernels, and are the first indigenous Australian food species to be commercialised internationally. The industry is largely based on cultivars developed in Hawaii in the late nineteenth century (Hardner et al. 2009). Current production is dominated by Australia, South Africa and Hawaii, and is expanding in China, Kenya and other countries around the world (Australian Macadamia Society 2018). A major focus in breeding new macadamia varieties is increasing nut-in-shell yield per unit (hectare or tree). However, yield has low heritability ($H^2 \approx 0.12$), is largely influenced by environment, and, as such, is difficult to select (Hardner et al. 2002). Limited

research has been conducted to improve yield of commercial varieties beyond conventional phenotype- and pedigree-based selection. Long juvenile periods, large tree sizes and labour involved in phenotyping over continuous years to identify elite candidate cultivars mean that fruit and nut trees may benefit from genomics to reduce selection cycles and increase genetic gain (Khan and Korban 2012).

The use of genomics in plant breeding is expanding (Iwata et al. 2016; Grattapaglia and Resende 2011; Khan and Korban 2012), including employing genome-wide association studies to identify molecular markers associated with important traits, and genomic selection for complex traits. A common approach is using genome-wide association studies (GWAS): each marker (typically single nucleotide polymorphism, SNP) is analysed individually to detect evidence of marker-trait associations (Khan and Korban 2012). This method relies on linkage disequilibrium (LD) between markers and causal polymorphisms (Khan and Korban 2012). To avoid spurious genotype-phenotype association due to population structure and family structures, linear mixed models, fitting individuals as random effects to account for relatedness, are widely used. As the realised kinship estimated from genetic markers is more accurate than recorded pedigree, fitting genomic relationships in the model can reduce false positives (Myles et al. 2009; Nejati-Javaremi et al. 1997; Hayes et al. 2009b). Findings of GWAS can be followed by marker-assisted selection (MAS) if a reasonable proportion of trait genetic variation is explained by the significant markers. In MAS, candidates are screened for target markers, their phenotypes are predicted based on allelic states, and selections can be made based on these predictions (Collard et al. 2005; Tester and Langridge 2010).

Several fruit and nut crops have employed GWAS to identify markers associated with key traits (Nishio et al. 2018a; Iwata et al. 2013; Kumar et al. 2013b; Cao et al. 2012; Minamikawa et al. 2017; Minamikawa et al. 2018; Imai et al. 2018; McClure et al. 2018). Furthermore, by mapping significant markers to reference genomes, the location of markers can be determined, although this is not necessary for MAS. GWAS coupled with MAS at these specific loci is a feasible option for improving yield component traits in macadamia (O'Connor et al. 2018b); hence, we aim to investigate this option in the Australian macadamia breeding program.

Target traits for GWAS and potential MAS in macadamia include commercially important traits, including nut and flowering characteristics, as well as tree size. Nuts consist of an inner

edible kernel, with two cotyledons, which is enclosed by a hard shell (testa) and outer husk (pericarp) (Hardner et al. 2009). Nut weight (NW), kernel weight (KW), and kernel recovery (KR) are commercially important yield component traits. For NW and KW, the industry favours intermediate optimums (6.5–7.5 g and 2–3 g, respectively) due to issues involved in handling, cracking, processing, and roasting smaller and larger nuts (Hardner et al. 2009). The selection goal for KR, which is the proportion of kernel to nut-in-shell (KW/NW), is more unclear. Whilst high (>37%) KR attracts a premium price per kilogram (Macadamia Processing Co. Ltd. 2018), very thin shells can be prone to pest and disease damage (Hardner et al. 2009). Whole kernels (WK) are those that have not split along the interface separating the two cotyledons during cracking (Walton et al. 2012); this trait can influence kernel price as some products and markets prefer whole kernels (Hardner et al. 2009; O'Hare et al. 2004).

Macadamia trees can produce about 2,500 pendant racemes 6–30 cm long, each with an inflorescence of 100–300 florets (Huett 2004; Trueman 2013). It has been estimated that less than 1% of florets produce viable nuts (Ito 1980). This estimate, therefore, indicates that many racemes and florets fail, likely due to a variety of reasons, and resource allocation may be a factor. As such, the proportion of racemes that survive from flowering through to nut set could indicate a genotype's reproduction success and energy investments, in terms of resource allocation for flowering versus nut retention (Toft 2019; Wilkie 2010). Reduced tree size is also an important selection trait to increase planting density and subsequent yield per hectare (Topp et al. 2016; Toft et al. 2018). Trunk circumference (TC) or trunk cross-sectional area can be used as an estimate of tree size in macadamia (Toft et al. 2018).

Chapter 3: Component Traits of Yield investigated heritability and correlations of yield and yield component traits measured on mature progeny. Several commercially important traits, as well as flowering and nut set characteristics that were moderately or highly correlated with yield are the focus of this study. It is hypothesised that marker-trait associations will be detected for these key traits using GWAS, and upon validation could be combined with MAS to improve breeding efforts and increase genetic gain in macadamia. The current study builds on work previously published in a preliminary study (O'Connor et al. 2019a) on the same population of trees. O'Connor et al. (2019a) found SNP markers associated with three nut characteristics (NW, KW and KR) measured on trees at the ages of 7–9 years (in 2010). In comparison, the current study uses a different set of SNP markers imputed with high

accuracy, and performs GWAS on yield component traits measured on the same trees at a mature age (aged 14–17 years, in 2016–2018). The aims of this study were to: (i) perform GWAS to identify markers significantly associated with yield component traits, and (ii) determine the location of significant markers on genome scaffolds, and calculate LD between significant marker pairs.

4.3 Methods

This study involves study material and phenotypic data collected as outlined in *Chapter 3: Component Traits of Yield*, and details are reproduced here for completeness. Methods for association analysis are similar to those by O'Connor et al. (2019a), and are also replicated here, with differences between the two studies outlined.

4.3.1 Study design

This study involved 295 seedling progeny from 32 full-sib families, as well 18 of their 29 parents (that were phenotyped), from the Australian macadamia breeding population. Trees were planted between 2001–2003 across four sites in Queensland. Clones of five of the parents were measured at all four sites as standards between the differing environments. Yield and yield component traits were measured on each tree between 2016–2018; hence, trees were mature-aged (aged 14–17 years). Details of genotyping methods for this population were reported in O'Connor et al. (2019b). Briefly, leaf samples from each genotype were sequenced by Diversity Arrays Technology (DArT) Pty Ltd. SNP markers were imputed by DArT, with 97.2% accuracy using the PPCA method (Stacklies et al. 2007). Markers were filtered for various quality control measures (based on pre-imputation genotypes), and those that passed thresholds were retained for analysis. The quality control measures included >50% call rate, >2.5% minor allele frequency, >0 polymorphic information content, and a test of Mendelian consistency between progeny-parent-parent trios in half of the studied families. This gave 4,113 SNP markers for analysis.

4.3.2 Phenotyping for yield and component traits

Phenotypic data used in this study were collected across two seasons from August 2016 to July 2018. Data are the same as in *Chapter 3: Component Traits of Yield*, and phenotyping methods are also described here. A sample of nuts was taken from each tree and dried to 1% moisture content in an oven at 35°C for 2 days, 45°C for 2 days and 55°C for a final 2 days, based on protocol by Prichavudhi and Yamamoto (1965). Twenty good quality nuts (no kernel shrivelling or pest damage) were chosen to measure four traits. Nuts were individually weighed to obtain nut weight (NW). Nuts were then manually cracked, and kernel and shell separated to record kernel weight (KW). Kernel recovery (KR) was calculated as KW / NW. The percentage of whole kernels (WK) per sample was measured as the proportion of nuts that did not split between the two cotyledons during cracking.

Tree trunk circumference (TC) was measured at a height of 50 cm above the ground, or below any low branches. Flowering racemes present in a 30 cm length of branch, 20 cm in from the edge of the branch were flagged and counted on two branches per tree. Where necessary, trees with terminal racemes were also flagged and counted, to make a total of at least ten racemes per tree. At nut maturity (around March, Australian Autumn), the number of flagged racemes that had set at least one nut was counted, and the percentage of racemes that survived from flowering through to nut set (RSN) was calculated. The number of nuts per raceme (NPR) was counted from ten racemes per tree. For each component trait, trait means were calculated for each tree for analysis where at least six observed units per tree were evaluated. For example, a tree with five or fewer nuts measured was considered to have missing data for this trait. Mean RSN was calculated for each tree over the two years.

Yield data were collected from March through to July over two successive seasons in multiple harvests. Yield was measured on each tree by manually harvesting nuts from the ground and collecting any nuts still in the tree at the end of the season. Nuts were dehusked after each harvest, weighed, and a 1 kg sample was dried to 1% moisture content. The dry nut-in-shell (DNIS) weight was estimated for each harvest using calculations of moisture content in the 1 kg sample. The DNIS weight for each harvest was summed across the whole season to give total DNIS yield. One site was not harvested in 2017 due to an extreme weather event, and in 2018 another site was not harvested due to management issues.

Histograms were used to check the distribution of phenotypes to conform with assumptions of normality for GWAS (Gondro et al. 2013). Data transformations were performed where necessary to normalise distributions. Pearson's correlations were performed between NW, KW and KR raw phenotypes in the current study and those used in O'Connor et al. (2019a) to investigate the consistency of phenotypes between the two studies.

4.3.3 Association analysis

A genomic relationship matrix (GRM) was constructed following methods of VanRaden (2008). Preliminary analysis was performed using ASReml (Gilmour et al. 2009) in R to determine the most parsimonious model for each trait:

$$\mathbf{y} = \mathbf{1\mu} + \mathbf{Xb} + \mathbf{Z_g g} + \mathbf{Z_{gs} gs} + \mathbf{e} \qquad \text{Equation 4-1}$$

where **y** is a vector of phenotypes, **1** is a vector of ones, **μ** is a fixed intercept, **X** is a design matrix allocating fixed effects (site, block within site, tree type = grafted parent or seedling progeny) to observations, **b** is a vector of these fixed effects, $\mathbf{Z_g}$ is a design matrix allocating records to the unknown average breeding value of each individual across sites; **g** is a vector of averaged breeding values of the individuals across sites, assumed random ~ N(0,$\mathbf{G}\sigma_g^2$), where **G** is the additive genomic relationship matrix (GRM) among the individuals, modelled from SNP effects (0, 1, and 2 represent homozygous, heterozygous and alternate homozygous genotypes, respectively); σ_g^2 is the genetic variance captured by the SNP; $\mathbf{Z_{gs}gs}$ describes the genotype by environment (G x E) interaction, where $\mathbf{Z_{gs}}$ is a design matrix allocating a specific effect of an individual at a site not accounted for by the mean of the individual across sites, and **gs** is a vector of the breeding values at a specific site, assumed random ~ N(0,$\mathbf{G}\otimes\mathbf{I_4}\otimes\sigma_{gs}^2$) where **I** is a 4x4 identity matrix for the four sites, and **e** is a vector of random errors ~ N(0, σ_e^2) where σ_e^2 is the error variance. This model is additive, in that two copies of the second allele will have double the effect of one copy.

Preliminary analyses determined the significance of fixed effects site, block within site, and tree type (grafted parent or seedling progeny) using the Wald statistic. After removing insignificant fixed effects (individualised for each trait), log likelihoods of models both including and excluding G x E as a random term were compared via a chi-square test to determine if the models were statistically different. The most parsimonious models were

those that fit the data best: the G x E term was excluded for a trait if the models were not statistically different, as well as any insignificant fixed effects. Narrow-sense heritability (h^2) was calculated from variance components ($h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$) for each trait using the best-fitting model. For traits where G x E was a significant factor, the G x E variance component was included in the denominator when calculating heritability.

Association analysis was performed for each trait using the most parsimonious model, as per O'Connor et al. (2019a) using ASReml (Gilmour et al. 2009) in R, using a mixed model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wm} + \mathbf{Z_g g} + \mathbf{Z_{gs} gs} + \mathbf{e} \qquad \text{Equation 4-2}$$

where **W** is a design matrix allocating records to the marker effect (modelled as 0, 1, or 2 for homozygous, heterozygous and alternate homozygous genotypes, respectively), and **m** is the effect of the marker currently being fitted in the model, as a fixed effect. All other effects are the same as per Equation 4-1.

QQ (quantile-quantile) plots were constructed for each trait to evaluate whether population structure had been accurately accounted for in the model, by comparing the observed and expected $-\log_{10}$ significance values of each SNP and ensuring that inflation had not occurred at the lower levels of significance (Gondro et al. 2013). To determine a threshold above which markers were deemed significantly associated with a trait, a false discovery rate (FDR) was calculated for each trait with the BH method (Benjamini and Hochberg 1995) using the p.adjust function in R. Markers with FDR < 0.05 were deemed significantly associated with the trait. Multiple regression was performed for traits with multiple significant associations based on the best-fit model, where significant markers were included as fixed effects, to determine if any SNPs were in LD. Markers that were no longer significant after regression were deemed to be detecting the same QTL as one of the significant markers, and as such were considered redundant.

The minor allele frequency (MAF) of each significant marker was calculated. The proportion of additive genetic variance explained by significant SNPs for each trait was derived from Falconer and Mackay (1996), and calculated as:

$$\eta = \frac{2p(1-p)a^2}{\sigma_g^2} \qquad \text{Equation 4-3}$$

where **p** is the frequency of the p allele, **a** is the effect of the marker, and σ_g^2 is the additive genetic variance for the trait.

4.3.4 Marker locations and accounting for LD between significant SNPs

Locations of significant SNPs (FDR > 0.05) on the most recent macadamia genome scaffolds (v2; 4,098 scaffolds; European Nucleotide Archive (EMBL-ENA) repository, Analysis: ERZ792049, Assembly accession: ERS2953073 (SAMEA5145324)), were estimated as per O'Connor et al. (2019b). Locations of previously identified markers associated with nut traits were also estimated on the scaffolds, using marker sequences from O'Connor et al. (2019a). LD was measured using the r^2 parameter between all pairwise significant SNPs from the current study, as well as between significant markers from O'Connor et al. (2019a) using Plink v1.9 (Purcell et al. 2007), with scaffolds as pseudo-chromosomes.

4.4 Results

4.4.1 Component traits

Raw (untransformed) phenotypes for KR, WK and TC were normally distributed (Figure 4-1). Log-transformed ($log_{10}(x)$) observations for NW, KW and NPR, as well as square root transformed observations for RSN appeared more normally distributed than raw observations (Figure 4-1). Yield (2017 and 2018) was not normally distributed, and neither log ($log_{10}(x)$, ln) nor square root transformations led to more normally distributed data, even for individual sites. This indicated that GWAS is not appropriate for yield, and association analysis was not performed for this trait.

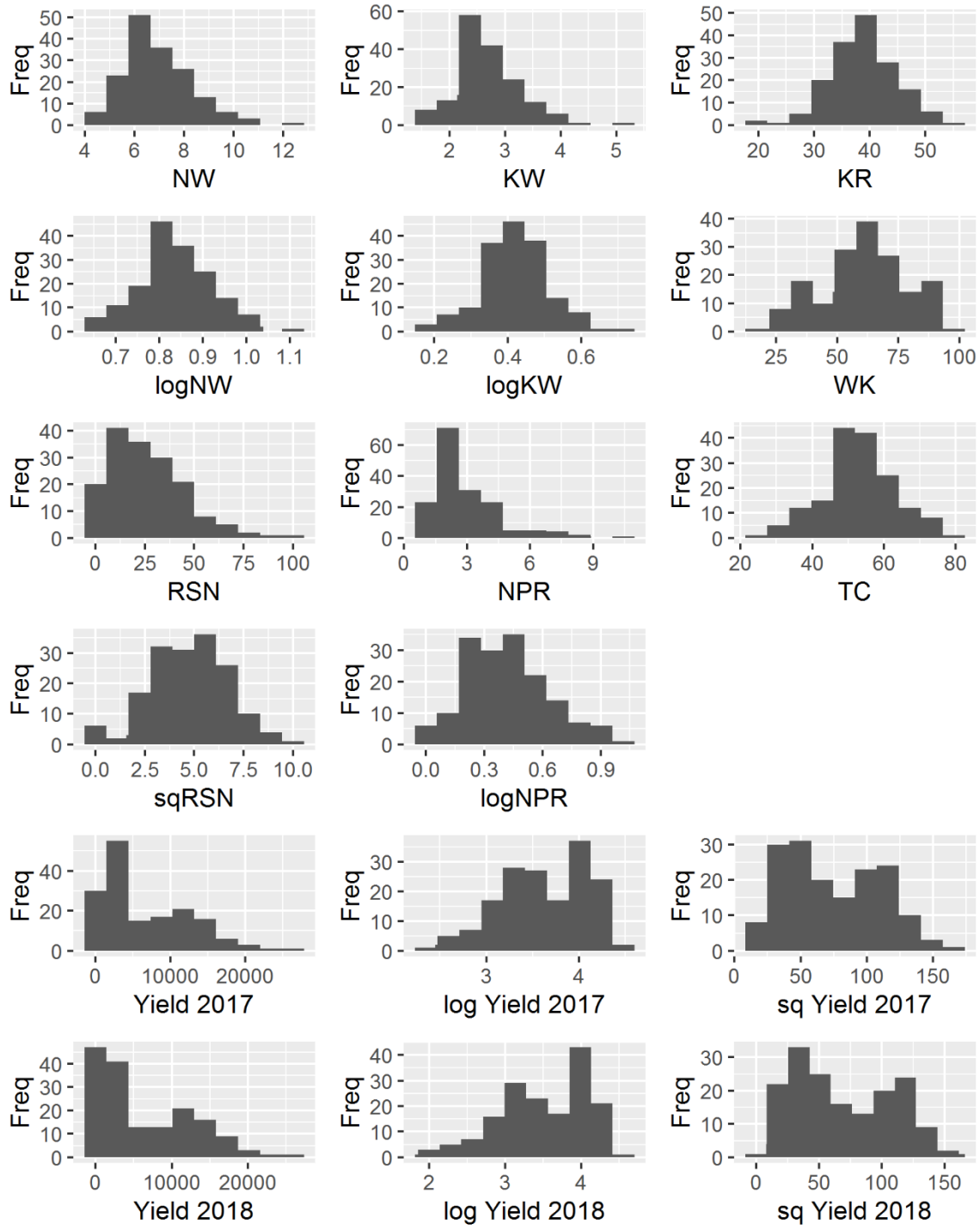


Figure 4-1 Distribution of phenotypes across all individuals for yield component traits. Freq, frequency; NW, nut weight; KW, kernel weight; KR, kernel recovery; WK, percentage of whole kernels; RSN, percentage of racemes that set nuts; NPR, number of nuts per raceme; TC, trunk circumference. Log-transformed ($log_{10}(x)$) NW, KW and NPR, and square root transformed (sq) RSN distributions are also shown, as well as both forms of transformation for yield in 2017 and 2018

Some summary statistics have been outlined in *Chapter 3: Component Traits of Yield*, and are presented here as well. Phenotypes ranged from 4.34 to 12.31 g for NW, 1.46 to 5.01 g for KW. As a derivative of these two traits, KR ranged from 20.2% to 55.6% (Table 4-1). Moderate to high correlations ($p < 0.01$) were observed between young and mature phenotypes for NW, KW and KR (0.56, 0.66 and 0.73; Table 4-1). For three genotypes, including cultivar 'Yonik', there were no broken kernels (100% WK) in the sample, whilst one tree possessed a very low WK (15%). Most small trees (small TC) were observed at site EG, with the lowest TC at 14 cm. Conversely, trees with large TC were observed at AL and HP, with a maximum TC of 78 cm at HP. An entire range of phenotypes was observed for RSN, from 0–100%, with a mean of 25%. Mean NPR was 2.6 and ranged from 1 to 10.4 (Table 4-1).

Table 4-1 Summary of raw (untransformed) phenotypes for each trait analysed in GWAS. SD, standard deviation; r_p, Pearson's correlation of current data with raw phenotypes for young trees from O'Connor et al. (2019a)

Trait	**Min**	**Max**	**Mean**	**SD**	**r_p**
NW (g)	4.34	12.31	7.09	1.34	0.56
KW (g)	1.46	5.01	2.73	0.55	0.66
KR (%)	20.2	55.6	38.7	5.4	0.73
WK (%)	15	100	64	17	-
TC (cm)	14	78	51	12	-
RSN (%)	0	100	25	18	-
NPR	1	10.4	2.6	1.4	-

4.4.2 Trait-specific models and heritability

For all traits except RSN, the most parsimonious model included site as a significant fixed effect, whilst block was also significant for NW and TC (Table 4-2). Tree type was included in the WK model, with a significance level of $p = 0.063$. The G x E term was included as a random effect for NW and NPR (Table 4-2). Narrow-sense genomic heritability varied across traits, from 0.08 for RSN to 0.74 for KR (Table 4-1). TC and NW were moderately heritable (0.45 and 0.53, respectively).

Table 4-2 Significance values of fixed and random terms included in association analysis model for each trait (log-transformed data for NW, KW and NPR, and square root data for RSN). Type, seedling progeny or grafted parents; G x E, genotype by environment (site) interaction; h^2, narrow-sense heritability. Non-significant p-values ($p > 0.05$) are not shown and were not included in models, except for Type for WK. * indicates G x E model was significantly better fitting than model without G x E term, as determined using log-likelihood ratio test. h^2 estimated from the best-fitting model with the GRM fitted

Trait	Site	Block	Tree Type	G x E	h^2
NW	0.0014	0.0025		*	0.53
KW	1.682e-13				0.37
KR	1.916e-09				0.74
WK	8.852e-05		0.063		0.24
TC	< 2.2e-16	0.0043			0.45
RSN					0.08
NPR	3.017e-08			*	0.09

4.4.3 Genome-wide associations

The GRM appeared to have effectively accounted for population structure in all traits except for TC, as no more associations than expected by chance were observed at low levels of significance in the QQ plots (Figure 4-2; Gondro et al. 2013). GWAS identified seven SNP markers significantly (FDR < 0.05) associated with NW, four with WK, and 44 with TC (Figure 4-2; Table 4-3). For both KW and KR, no markers exceeded the FDR threshold; however, there was one marker of interest in both traits that were further investigated. There were no markers significantly associated with RSN or NPR. After multiple regression, where significant SNPs were treated as fixed effects, some markers were no longer significantly associated with some traits. Only SNP s2204 remained significantly associated with NW, whilst SNP s2607 was no longer a significant association for WK (Table 4-3). The number of SNPs significantly associated with TC decreased to 14.

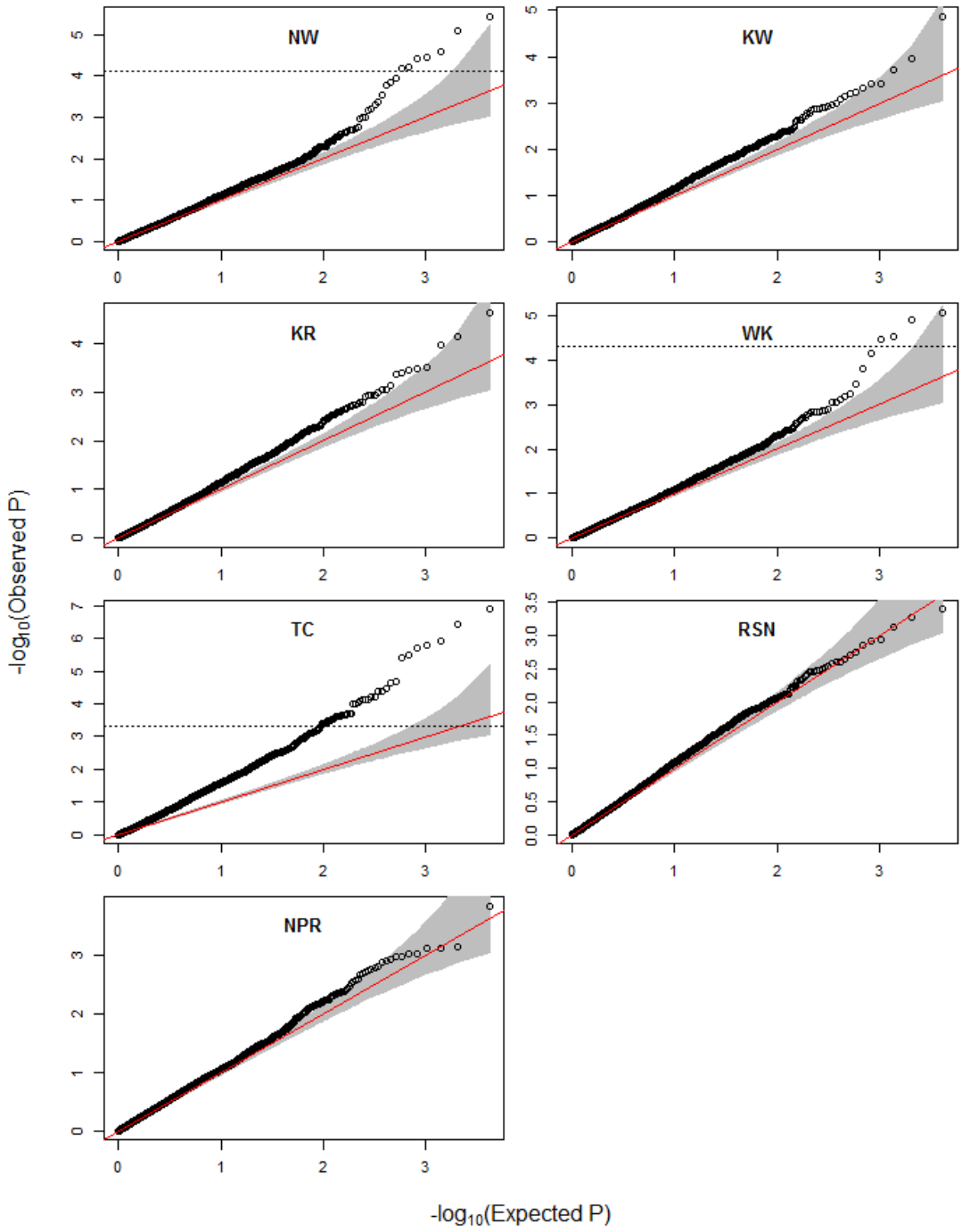


Figure 4-2 QQ plots showing expected significance levels against observed significance for 4,113 SNPs for yield component traits. Red diagonal lines indicate the null hypothesis, where observed and expected p-values would sit if there were no associations. Dashed horizontal lines indicate FDR = 0.05, SNP markers above which were deemed significantly associated with the trait; if no dashed horizontal line is present then no SNPs exceeded the FDR threshold. Shaded area indicates 95% confidence interval

Thirty-four of the 57 (60%) significant SNPs across the traits were mapped to scaffolds of the v2 macadamia genome assembly, and none mapped to the same scaffold across any of the traits (Table 4-3). Thirteen (57%) of the markers presented in Table 4-3 showed a MAF of less than 5%, whilst all markers significantly associated with NW exhibited a MAF less than 4% (Table 4-3). Almost 50% allele frequency was observed for two markers (s3540 for KW, and s3616 for TC). Linkage disequilibrium between significant associations within each trait varied from $r^2 = 0.082$ between two markers for WK, to 1.000 between six SNPs for NW (Table 4-3). Proportion of genotypic variance explained by the markers also varied between traits, from $\eta = 0.0001$ (for TC) to 0.378 (WK; Table 4-3). It should be remembered, though, that these estimates of proportion of variance explained are estimated in the discovery population, and are therefore very likely to be over-estimated due to the Beavis effect (Beavis 1994).

Table 4-3 Summary of significant SNPs associated with yield component traits identified in GWAS. Only the ten most significant markers for TC are shown. MAF, minor allele frequency of the marker; p, significance of association; pMR, significance of association as determined by multiple regression with significant SNPs as fixed effects; η, proportion of genetic variance explained by each SNP; LD, average linkage disequilibrium (r^2) between mapped markers for each trait; NS, not significant. - indicates marker was not mapped to scaffolds. [a] Scaffold in v2 genome assembly. [b] Did not pass FDR = 0.05 threshold

Trait	SNP	Scaffold[a]	Position (bp)	Alleles	MAF	p	pMR	η	LD
NW	s2204	scaffold926|size239084	212,122	A/G	0.027	3.68E-06	2.147e-05	0.118	1.000
	s4163	scaffold285|size451335	314,657	C/T	0.027	8.03E-06	NS	0.097	
	s1434	scaffold_177	804,678	T/C	0.019	2.65E-05	NS	0.134	
	s1643	scaffold44|size832018	129,241	A/C	0.021	3.46E-05	NS	0.120	
	s1121	scaffold653|size305054	6,573	A/G	0.021	3.82E-05	NS	0.103	
	s5182	-	-	A/T	0.035	6.29E-05	NS	0.088	
	s2256	scaffold710|size289053	142,496	G/T	0.026	6.45E-05	NS	0.079	
KW	s3540[b]	-	-	G/A	0.482	1.34E-05		0.116	
KR	s1707[b]	scaffold_72	587,142	C/T	0.061	2.37E-05		0.052	
WK	s0201	scaffold213|size509421	186,179	G/A	0.093	8.81E-06	4.10E-05	0.293	0.082
	s1917	-	-	A/G	0.163	1.23E-05	1.61E-02	0.221	
	s2607	-	-	T/C	0.177	2.91E-05	NS	0.241	
	s3239	scaffold361|size1112638	1,087,419	G/C	0.037	3.39E-05	5.68E-04	0.378	
TC	s3169	-	-	T/C	0.230	1.29E-07	2.92E-03	0.049	0.059
	s1053	scaffold597|size318270	258,329	G/A	0.264	3.82E-07	9.16E-03	0.077	
	s3616	scaffold364|size402398	377,926	T/C	0.473	1.15E-06	4.03E-02	0.033	
	s2631	scaffold1151|size292196	181,208	G/C	0.093	1.67E-06	NS	0.0001	
	s3332	scaffold1221|size537814	497,497	T/C	0.289	1.97E-06	1.65E-03	0.118	
	s4480	scaffold_221	469,574	A/T	0.187	3.20E-06	2.67E-02	0.018	
	s3406	-	-	G/A	0.083	3.82E-06	NS	0.001	
	s2500	-	-	G/A	0.137	2.02E-05	3.46E-02	0.080	
	s0011	-	-	G/C	0.300	2.20E-05	6.07E-02	0.024	
	s3405	-	-	G/C	0.083	3.40E-05	NS	0.002	

4.5 Discussion

4.5.1 Phenotypic data in the breeding program

Large phenotypic diversity was observed for many of the traits in this study. Mean phenotypic values observed here for NW, KW and KR (7.09 g, 2.73 g, and 38.7%, respectively) were all slightly higher compared with the same traits when the trees were young (6.21 g, 2.28 g, and 36.9%) (O'Connor et al. 2019a). The moderate heritabilities suggest that selection for a number of traits could result in good genetic progress. For example, the high narrow-sense heritability observed for KR ($h^2 = 0.74$) means that the aim to select for higher KR is possible with truncation selection. This form of selection is where trees with phenotypes or estimated breeding values below a certain threshold are excluded from parent populations, and the mean values of progeny should increase for this trait over generations (Falconer 1989). Results of this study differed to that of O'Connor et al. (2019a) which analysed the same population when the trees were younger (around 8 years of age). Heritability for KR was higher in mature trees than young trees (0.62), whilst KW was lower in mature trees (0.37) than young trees (0.53). In comparison, the difference in heritability for NW between the two studies was low (0.03), but the correlation between these phenotypes was only moderate (0.56).

This study demonstrates that linear mixed models are useful for analysing phenotypic and genetic data in macadamia to identify QTLs for target traits, which is beneficial, as developing new macadamia varieties is time-consuming, laborious and expensive. Additionally, the large tree size and numbers involved in macadamia breeding means that multiple environments are typically needed during evaluation trials. The mixed models employed in this study account for these environmental differences, as well as G x E interactions for some traits. Thus, the best model was fitted to the data on a trait-by-trait basis.

4.5.2 Genetic data

The current study used 4,113 SNP markers imputed with high accuracy, though analysis of LD found that LD declined rapidly over short distances (O'Connor et al. 2019b). The number of markers in the current study is comparable with other studies in fruit trees (Minamikawa et al. 2017; Minamikawa et al. 2018; Imai et al. 2018; Kumar et al. 2013b); however, the

fragmented nature of the macadamia genome scaffolds means the distribution of markers across the whole genome is still unknown.

Population structure affects LD, and this needs to be accounted for in GWAS to avoid spurious associations. For most traits investigated here, the QQ plots showed that only the highly significant markers deviated from the null expectation (y = x line), and did not show inflation of the observed versus expected p-values at lower significance levels. QQ plots showing this pattern demonstrate that population structure has been effectively accounted for by the GRM (Korte and Farlow 2013). One explanation for divergence from the null hypothesis (more associations detected than expected) at high p-values is polygenicity: many loci of small effect contributing to variation in the trait (Yang et al. 2011). This concept may explain the pattern observed for TC, where a large number of associated markers was detected even at low p-values. The previous study (O'Connor et al. 2019a) did not use imputed markers, and deviations from the null hypothesis line were observed. Imputation of missing data with high accuracy can, therefore, more accurately capture the realised kinship between individuals, and, as such, produce more accurate association results.

4.5.3 Association analysis

MAS, using the findings of GWAS, is effective for traits controlled by few genes, and, as such, has little value for complex traits (Hayes and Goddard 2010; Luby and Shaw 2000; Huang and Han 2014). However, Kelner et al. (2011) found two clusters of QTLs related to fruit yield and cumulative yield in apple on two different linkage groups, as well as QTLs for precocity and biennial bearing. Genomic selection may be a more appropriate and accurate method to predict yield in macadamia (O'Connor et al. 2018b).

This study identified SNP markers significantly associated with NW, WK and TC. Although no significantly associated markers were detected for KW or KR, the marker with the lowest p-value in each case should be investigated in further studies. Neither NPR nor RSN had any significant associations, which may be partly due to the very low heritability of both traits. Additionally, while there was no G x E detected in RSN, there may be a large environmental influence on the capacity of a tree to retain racemes from flowering through to nut set (Toft 2019; Wilkie 2010).

Perfect LD existed between all mapped markers for NW. Multiple regression suggested that all the markers may have detected the same or linked QTLs, with the most significant SNP (s2204) being the only non-redundant marker. However, differences in significance values among these markers suggests that the markers are not in perfect LD, and so the LD may be over-estimated with this population compared to the larger progeny population. Furthermore, closer inspection of these SNPs revealed that allelic genotypes were not perfectly correlated among markers, and so a true LD value of 1.000 seems improbable. These inconsistencies indicate that further analyses in separate populations are needed to validate the findings of this study.

Multiple regression for the four markers significantly associated with WK found that the two mapped markers (mapped to different scaffolds) and another marker remained significant, but the unmapped SNP s2607 was redundant. Furthermore, the LD between the two mapped markers for WK was low, at $r^2 = 0.082$, suggesting that they were unlinked and representing two separate causal variants. For TC, 14 of the 44 significant markers were non-redundant, suggesting that there may be 14 QTLs controlling this trait.

A direct comparison cannot be made between SNPs found to be significantly associated with nut traits in O'Connor et al. (2019a) and the current study, as two different SNP panels were used in the analyses. However, some of the significant markers could be mapped to genome assembly scaffolds. A comparison of the locations of mapped SNPs between the two studies showed that there were no markers occupying the same scaffold (data not shown). Results from GWAS are not always consistent, with variation between populations and environments altering allelic frequencies and phenotypes. Differences were found also across years in apple (McClure et al. 2018), and between QTL mapping and GWAS studies in chestnut (Nishio et al. 2018a; Nishio et al. 2018b), and this may be a consequence of limited power in these studies.

4.5.4 Markers and proportion of variance explained

Low MAF was observed for many of the significant markers. Researchers use different thresholds for determining which markers to include in their genomics studies, such as 5% MAF (Imai et al. 2018; Nishio et al. 2018a), 1% MAF within-populations (Biscarini et al. 2017), and ten copies of the minor allele across samples (McClure et al. 2018). In the present study,

markers were initially excluded with MAF <2.5%, though these statistics were calculated for each marker before imputation, and, as such, the study included markers with MAF below this threshold (MAF altered after imputation of missing calls). It was interesting, then, that all of the markers associated with NW had very low MAF. If these markers had been removed by filtering, they would not have been detected through GWAS. Associations with rare alleles should be treated with caution due to low power of detection (Gondro et al. 2013), and this is the case here. Therefore, the significant markers with low MAF in the current study should be validated in independent studies, preferably with more individuals to observe whether the MAF is similar across populations of different sizes (Hayes 2013), as this will support the findings of this study.

The proportion of variance explained (ɳ) will be overestimated due to rare minor alleles in a small sample size (Beavis 1994). Overestimations were observed in the current study; for example, the highest ɳ was calculated for marker s3239 (associated with WK) at 0.378, and MAF 0.037. With such low MAF in a relatively small population, these calculations of ɳ will be biased, and, as such, should be interpreted with caution. Again, the SNP should be validated in an independent population, and the proportion of variance the SNPs explain should be estimated in that population.

4.5.5 Demonstration of marker-assisted selection

The results of this GWAS study can be used to demonstrate the employment of MAS in the macadamia breeding program. SNPs significantly associated with commercially important traits in macadamia would be ideal candidates for use in MAS. For example, the average phenotypes of NW at SNP s2204 for AA, AG and GG genotypes were 7.03 g (n = 309, SD = 1.29), 8.20 g (n = 5, SD = 0.58), and 9.54 g (n = 6, SD = 1.73), respectively. Similarly, the average values of WK for GG, GA and AA genotypes at marker s0201 were 62.3% (n = 265, SD = 16.8), 72.9% (n = 50, SD = 15.3), and 78.0% (n = 5, SD = 11.0), respectively. The sample sizes among the three different genotypic states varies greatly in these examples, and so it is important to recognise that these findings are severely biased upwards and are only for demonstrative purposes for how MAS could be used. Simply, breeders could genotype seedling progeny from their very first leaves at these key markers. By learning the allelic states of each seedling at these markers, breeders could, for example, select for AG heterozygotes at SNP s2204 for

intermediate nuts, and AA genotype at SNP s0201 for a high percentage of whole kernels. The team could discard candidates without those favourable SNP alleles, and, instead, continue to evaluate individuals with the predicted desirable characteristics. This process would be particularly pertinent for markers with low heritability, such as WK and TC. However, the polygenic nature of TC, as well as the complexity of yield, means that these traits may be more suited for genomic selection, where many markers may have a small effect on the trait, and all markers are modelled simultaneously (Meuwissen et al. 2001), rather than one-by-one as in GWAS.

4.5.6 Further work

This study and our previous work (O'Connor et al. 2019a) provide a foundation for how the use of genomics can improve breeding in macadamia, and is among the first to analyse the potential for genomics-assisted breeding in nut crops, generally. However, the results presented here should be further explored and validated before being employed in breeding programs. Multi-trait analyses could be performed to increase the power of detection of QTLs, and also detect pleiotropy (Bolormaa et al. 2014). A separate population should be studied to determine if QTLs detected are the same as those detected here, or are new associations. Further studies should incorporate larger population sizes, to ensure that significant associations are accurate and applicable to a wider breeding population. Additionally, the low MAF observed for some markers in this study may change with sample size, which will influence estimated of the proportion of variance explained by those markers. When a more complete reference genome is assembled, the location of these markers can be estimated, and LD between markers more accurately estimated with population structure and cryptic relatedness taken into account. Due to the rapid decay of LD over short distances in macadamia (O'Connor et al. 2019b), using a larger number of markers may increase the likelihood of SNPs being in LD with causal polymorphisms. Furthermore, the potential issues posed by allelic dropouts, such as lower than expected levels of heterozygosity observed by O'Connor et al. (2019b), could be alleviated with the use of a complete reference genome in sequencing of SNPs in the future. The genome scaffolds are currently largely unannotated, and, as such, the significant SNPs cannot yet be linked to known genes or proteins, which has been achieved in other studies of GWAS in fruit trees (e.g. Kumar et al. 2013b; Minamikawa

et al. 2017; Minamikawa et al. 2018; McClure et al. 2018). Although there was a lack of significant associations in some traits in the current study, these should still be investigated in future work. Other traits that could be analysed include self-fertility, and resistance to diseases that affect nut yield, including husk spot and phytophthora. Genomic selection could be used as an alternative to GWAS for more complex traits such as yield, and perhaps TC due to the polygenic nature of this trait.

4.6 Conclusions

The findings of this study have important implications for macadamia breeding, but it also highlights the difficulties of employing GWAS in heterozygous populations with rapid LD decay. Significant associations were detected for NW, WK and TC, but no markers exceeded the significance threshold for KW, KR, RSN or NPR. Multiple regression and LD analysis determined that several significant markers were detecting the same QTL, and, as such, were redundant. By coupling validated marker-trait associations detected through GWAS with MAS, genetic gain could be increased by decreasing the selection time for economically important nut characteristics and other yield component traits. Genomic selection may be a more appropriate method to predict complex traits like yield. Overall, this study provides a foundation for genomics-assisted breeding in macadamia and nut crops more broadly, and advances our understanding of the genetic control of yield component traits.

Chapter 5. Genomic selection for nut yield and yield stability, and a comparison of selection strategies in the Australian industry macadamia breeding program

---

My contribution to the chapter

I wrote the paper, performed all analyses and made final edits. BH and CH advised in analytical methods and assisted in interpretation of results. BH, CH, BT and MA suggested revisions.

5.1 Abstract

Improving nut yield prediction and selection efficiency in macadamia trees is vital, as the time from crossing to production of new cultivars is almost a quarter of a century. Genomic selection (GS) is a useful tool for plant breeding programs, particularly perennial trees, to increase the rate of genetic gain and reduce breeding cycle times. With the aim of overcoming these bottlenecks and to accelerate selection efficiency in the Australian macadamia breeding program, for the first time we introduced GS and compared different selection strategies. The traits of focus in the study were nut yield from tree ages 5 to 8 years, and yield stability, as a basic measure of the standard deviation of yield over these four years. Narrow-sense heritability of yield and yield stability was low (h^2 = 0.23 and 0.04, respectively). Prediction accuracies in random cross-validation varied across four sites in the trial, and ranged from r = 0.57 to 0.70 using yield data for ages 5 to 8. Prediction accuracies across unrelated populations (grouped families), which are an extreme representation of real-world application, were lower than in related family predictions using randomly grouped individuals. Accuracy of prediction of yield stability was high (r = 0.79) for related family predictions. Predicted genetic gain of yield using GS methods varied across different breeding strategies, from 12 to 162 g/year for unrelated population predictions, and 421 to 590 g/year for random groups (for 2.5% selection intensity). Estimates of genetic gain for GS were comparable to traditional breeding (202 g/year) for unrelated population predictions, and more than double that for related family predictions. The incorporation of GS into the Australian macadamia breeding program may accelerate genetic gain, though the high cost of genotyping appears to be a large constraint at present.

5.2 Introduction

Nut yield is the most economically important selection trait of macadamia (Hardner et al. 2009). In 2017, the Australian industry – the world's largest – produced a crop of 46,000 tonnes of nut-in-shell (Australian Macadamia Society 2017a). With a recent average farm gate price of around AU$5 per kilogram (Australian Macadamia Society 2017c), macadamia is, thus, an economically important crop for Australia.

Although yield is the main trait of focus when selecting new macadamia varieties, it is expensive and difficult to assess in breeding programs. Nuts are comprised of an outer pericarp green husk, a hard shell testa, and an internal edible kernel. The husk either abscises from the tree along with the nut-in-shell (NIS), or dehisces (splits along a single suture) and the nut-in-shell falls to the ground (Hardner et al. 2009). After harvest, nuts are dehusked mechanically. Yield measurements are usually expressed as NIS or kernel yield per tree (Hardner et al. 2009). Yield is a complex trait affected by many processes and environmental influences, and is likely controlled by many genes (Jannink et al. 2010; Quarrie et al. 2006). As such, selection for high yield is often made difficult by environmental and genotype x environment interaction (G x E) effects (Allard and Bradshaw 1964), with G x E being previously documented in macadamia yield (Hardner et al. 2002; Hardner 2017). Kernel recovery (ratio of kernel to nut-in-shell weight; KR) (Hardner et al. 2009) is also an important economic trait, as high KR attracts a higher commission per kilogram than low KR (Macadamia Processing Co. Ltd. 2018).

In addition to increased yield and KR, precocious cultivars, those that produce nuts at an early age, would be commercially valuable by increasing early return on investment. However, it is not yet known how precocity might affect the rate at which yield begins to plateau in macadamia varieties (B. Topp, pers comm.). In coffee and apple, early-yielding varieties are desirable, particularly those with stable yields over time (Kelner et al. 2011; Cilas et al. 2011). For perennial horticulture crops, like macadamia, yield stability could be interpreted as the consistency of yield of individual trees across consecutive years (Sharma et al. 2019). Unstable yields, due to alternate bearing, is common in some perennial fruit crops, and is undesirable as regular income is vital for growers (Sharma et al. 2019; Cilas et al. 2011). Research regarding genetic architecture surrounding consistency of yield over years has been limited outside of biennial bearing in apple (e.g. Guitton et al. 2012; Durand et al. 2013). Yield stability is considered an important trait in macadamia by industry (Hardner et al. 2009). Some macadamia growers report biennial bearing in certain cultivars, such as ‘H2’ and ‘344’ (K. O’Connor, pers. comm.), which can be problematic if environmental events, such as a cyclone, occur in high-bearing years. Thus, the stability of yield is an economically important trait in macadamia that should be further investigated.

Selection of new varieties involves two stages: thousands of seedlings are produced by cross-pollination to create diversity and are assessed in a seedling progeny trial (SPT), then the best performing trees are clonally propagated and evaluated in replicated trials across multiple environments in a candidate cultivar regional variety trial (RVT) (Topp et al. 2016). Trees begin to flower and bear fruit around four to five years after planting, and are evaluated for at least eight years at each breeding stage (Hardner et al. 2002; Hardner et al. 2009). Due to the crop's long juvenile stage, as well as the need to assess yield over several years to increase the accuracy of prediction of long term yield, traditional macadamia breeding has a selection cycle of almost a quarter of a century (22 years) (Topp et al. 2016; Hardner et al. 2009). Alternative selection strategies are sought to shorten the selection cycle and increase genetic gain. Genomic selection (GS) offers an opportunity to achieve this in macadamia (O'Connor et al. 2018b).

Genomic selection utilises genome-wide markers to predict genomic estimated breeding values (GEBVs) of individuals, after which the best performers are selected (Viana et al. 2016; Meuwissen et al. 2001). GS uses a training or reference population of individuals with known genotypes and phenotypes to construct a model of each marker's effect on the trait. To determine the accuracy of prediction, the model is then applied to predict the GEBV of individuals in a validation population, for which measured phenotypes are available. The accuracy of prediction is determined by how closely the predicted values reflect the observed phenotypes; the correlation between the two. This correlation is then divided by the accuracy of these phenotypes in predicting the true breeding values, which is the square root of the heritability of the phenotype (Goddard and Hayes 2007; Dekkers 2007). The GEBV can be predicted for individuals at the seedling stage, using only genotypic data; thus, enabling early selection for elite individuals (Meuwissen et al. 2001).

Genomic selection was first used in dairy cattle, and is being increasingly used to improve genetic gain in both animal and plant breeding programs. With the potential to reduce breeding cycle times, long-lived species with slow maturation times may have the most to gain from GS (Luby and Shaw 2000; Iwata et al. 2016). Grattapaglia (2014) and Lin et al. (2014) have extensive reviews on the use of genomic selection in forestry and annual species, respectively. The main attraction of GS for perennial crops may be that it can accelerate breeding cycles, thereby increasing the gain per unit time and reducing field trial costs

(Jannink et al. 2010; Desta and Ortiz 2014; Denis and Bouvet 2013). Sweet cherry (Piaskowski et al. 2018), peach (Biscarini et al. 2017), oil palm (Kwong et al. 2017b; Kwong et al. 2017a; Wong and Bernardo 2008), citrus (Minamikawa et al. 2017), apple (Muranty et al. 2015; Kumar et al. 2012b), and pear (Minamikawa et al. 2018; Iwata et al. 2013) breeders have all investigated GS to increase genetic gain in their breeding programs. A recent study in Japanese chestnut (Nishio et al. 2018a) achieved high prediction accuracies for harvest date ($r = 0.841$) and insect infestation (0.604), though yield was not studied.

High accuracy of GS models will allow confident selection of candidates. Accuracy depends on many factors, including the model, crop, size of the reference population, extent of linkage disequilibrium (LD), marker set, and trait of interest (Crossa et al. 2010). Genetic markers should be in high LD with the genes controlling the trait, in order to capture the genetic variance (Druet et al. 2014; Goddard 1991; Viana et al. 2016). In a simulation using animal data, Calus et al. (2008) suggested that models using marker densities of LD $r^2 = 0.2$ (average distance of 0.128 cM between markers) were superior to those at lower densities. Accurate phenotyping of a large training population, preferably over multiple environments and years (allowing for the study of multiple seasons and tree ages), is required for perennial crops to derive accurate predictions, due to the interactions between these factors (Resende et al. 2012; Xu and Crouch 2008; Desta and Ortiz 2014; Rikkerink et al. 2007).

We hypothesise that using GS in the macadamia breeding program will lead to greater genetic gains than phenotypic- and pedigree-based selection, due to a substantial reduction in generation length. This study aims to: (i) determine the accuracy of GBLUP (genomic best linear unbiased prediction) methods in predicting GEBV for nut yield and yield stability across years in macadamia; and (ii) identify strategies in which GS can be employed to increase genetic gain in macadamia breeding programs. This research is the first study to utilise molecular marker technology for GS in macadamia, and to our knowledge, the first to use GS to predict yield stability over consecutive years for a fruit or nut tree crop.

5.3 Materials and methods

5.3.1 Plant material and phenotyping

This study involves a subset of progeny from the Australian industry macadamia breeding program's "B1.2" population. This population consisted of approximately 2,000 seedlings across 141 full-sib families from crosses between 47 parents, with 1–36 progeny per family (mean 14) (Topp et al. 2016). Trees were planted across nine sites in south-eastern Queensland and north-eastern New South Wales, Australia, between 2001 and 2003 in an incomplete block design (Topp et al. 2016). Trees were planted at 4 m distances within rows, and 8m between rows.

As described by O'Connor et al. (2019b), 295 macadamia seedlings from 32 full-sib families (reciprocals combined) from crosses between 29 parents (7–11 progeny per family) across four sites were chosen for genotyping. All families had at least one parent in common with another family. Progeny within families were selected for genotyping to achieve an approximately equal number of low- and high-yielding individuals per family, based on breeding values for cumulative nut-in-shell yield to age 8 years, to ensure a spread of phenotypes. The sites included Hinkler Park (HP) and Alloway (AL) near Bundaberg, Queensland, and East Gympie (EG) and Amamoor (AM) near Gympie, Queensland. Clones of five parental genotypes were planted at each of the four sites, with a further 13 parental clones planted at AL. Eleven of the 29 parents could not be phenotyped, as they were not planted in these trials.

Yield was evaluated on an individual tree basis across multiple years. Nuts-in-husk were manually harvested from the ground in multiple harvests from February to August. A final strip harvest was performed at the end of the season, in which the nuts remaining in the tree were removed with poles and hooks. Nuts were dehusked mechanically and weighed to obtain a wet nut-in-shell weight. For each tree, a 1 kg sample was taken (where available) and dried to approximately 1.5% moisture content at 35°C for 48 hours, 45°C for 48 hours and 55°C for 48 hours, based on protocol by Prichavudhi and Yamamoto (1965). Samples were then weighed to obtain a dry nut-in-shell (DNIS) weight, with the moisture content used to calculate a total DNIS weight per harvest. DNIS weights were summed across harvests to obtain the total NIS yield per tree each year. Three phenotypic datasets were used in the

analysis: yield data from young trees only (aged 5 to 8 years), yield data from mature trees only (2017 and 2018 harvests, aged 14–17 years), and young-tree and mature-tree yield data combined. Yield data were unable to be collected in 2017 at site AL due to storm damage, and in 2018 at site EG due to management issues.

5.3.2 SNP genotyping and imputation

This study used genetic markers obtained as described by O'Connor et al. (2019b). Briefly, DNA was extracted from leaves of 295 seedlings and their parents, and sequenced by Diversity Arrays Technology (DArT) Pty Ltd, to produce a total of 5,329 SNP markers and 19,527 silicoDArT markers (presence/absence, dominant). Missing calls were imputed using the PPCA method (Stacklies et al. 2007) with 97.2% accuracy, which was determined by excluding an additional 10% of missing values and calculating the correlation between the imputed calls and the original dataset. Quality control was performed using pre-imputation parameters, including 50% original call rate, 2.5% minor allele frequency, and a test of Mendelian inconsistencies (parent-offspring trio opposing homozygotes) determined using 16 (50%) of the families. This resulted in 4,113 SNPs for genomic analysis.

5.3.3 Predicting and validating GEBVs

An additive genomic relationship matrix (GRM) was constructed among all individuals using the 4,113 SNPs, as per VanRaden (2008) and detailed in O'Connor et al. (2019b). Multiple GBLUP models were used to calculate GEBVs for each tree using ASReml-R (Butler et al. 2009). In the first model, GEBVs were obtained for trees only within individual sites (e.g. each site was a separate analysis), using the following GBLUP model:

$$Yield = mean + Block + Type + Number\ Neighbours + Age + Accession + error \quad \text{Equation 5-1}$$

where **Yield** is the yield of an individual tree in one year; **mean** is the intercept of the model; **Block** is the planting block within a site, as a fixed effect; **Type** is the propagation type of the tree (seedling progeny or clonally propagated parent), as a fixed effect; **Number Neighbours** is the number of trees on either side of that tree within the row, to allow for influence on phenotype of gaps created by the death of neighbouring trees, as a fixed effect; **Age** is the

age of the tree after planting, as a fixed effect; **Accession** is the tree effect modelled as the additive genetic effect of the individual, assumed random ~ N(0,$\mathbf{G}\sigma_g^2$), where **G** is the GRM, modelled from SNP effects (where 0, 1, and 2 represent homozygous, heterozygous and alternate homozygous genotypes, respectively), and σ_g^2 is the additive genetic variance captured by the SNP; and **error** is a vector of random deviations e ~ N(0, $\boldsymbol{\sigma_e^2}$) where $\boldsymbol{\sigma_e^2}$ is the error variance. For mature data where only one year of data were available (at sites AL and EG), Age was omitted from the model. Where multiple years of yield data were used in the model, a mean yield GEBV across years was calculated.

Note that we did not fit a permanent environment in the model, owing to the complexity of the data where, for some trees and sites, only one year of data were available. To assess the impact of this, we fit a model with permanent environment effects to a subset of the data where four years of data were available. The correlation of the GEBV for the model with permanent environment effects fitted, and without, was 0.93 ($p < 0.001$). So for simplicity, permanent environment effects were omitted from the subsequent stages of analysis.

In the second model, GEBVs were calculated using data from all sites, simultaneously, as follows:

$$Yield = mean + Site + Block + Type + Number\ Neighbours + Age + Year + Year \otimes Age + Site \otimes Year + Accession + Accession \otimes Site + error \quad \text{Equation 5-2}$$

where the terms are as per Equation 5-1 above, with some additions. **Year** is the calendar year that yield was harvested, as trees were planted in different years across the sites, as a fixed effect; **Year⊗Age** is the interaction between calendar year and age of the tree, as a fixed effect; **Site⊗Year** is the interaction between site and calendar year, as a fixed effect; **Accession⊗Site** is the interaction between the additive genetic effect of the tree and the site, assumed random ~ N(0,$\mathbf{G}\otimes\mathbf{I}\otimes\sigma_{gs}^2$) where **I** is a 4x4 identity matrix for the four sites.

The inclusion of the **Accession⊗Site** term in the model meant that five GBLUP genetic effects were calculated for each tree: those for the average tree genetic effect across all sites, and the genetic effect of each tree at each of the four sites. This study investigated the accuracy of genomic prediction using GEBVs calculated in two ways: using just the average tree genetic effect (denoted here as Tree genetic effect), and using the sum of average tree genetic effect plus the genetic effect at the site being predicted (denoted here as Tree + Site genetic effect).

To determine if correcting for site effects first was more accurate than grouping all sites together, an alternative model for across all site data was performed using the following model:

$$Solutions = mean + Site + Accession + Accession \otimes Site + error \quad \text{Equation 5-3}$$

where **Solutions** are the corrected phenotypes for individual sites with no pedigree or genetic information included, and **Site** is the only fixed effect in the model.

Corrected phenotypes were calculated for individual site analyses using the following model:

$$Yield = mean + Block + Type + Number\ Neighbours + Age + Tree + error \quad \text{Equation 5-4}$$

where **Tree** is the individual tree modelled as a random effect, without any pedigree or realised genetic relationship (GRM) information. For mature data where only one year of data were available (at sites AL and EG), Age and Tree were omitted from the model. Corrected phenotypes, or solutions, were either estimated from model residual effects (when only one year of data per tree was analysed) or model random effects (when a mean of phenotypes was obtained from multiple years of data). Where multiple years of yield data were used in the model, solutions were essentially a mean yield over the multiple years.

Corrected phenotypes were also calculated using data across all sites with the following model:

$$Yield = mean + Site + Block + Type + Number\ Neighbours + Age + Year + Year \otimes Age + Site \otimes Year + Tree + Tree \otimes Site + error \quad \text{Equation 5-5}$$

Corrected phenotypes for individual years were obtained using Equation 5-5, but using only a single year of data at a time for **Yield**, and excluding **Age** and **Year** terms and interactions.

To determine if GEBVs were more accurately predicted for cumulative yield than for mean yield over multiple years, the following model was used:

$$Summed\ Solutions = mean + Accession + error \quad \text{Equation 5-6}$$

where **Summed Solutions** are solutions for individual years (corrected phenotypes calculated from Equation 5-5 with data for each tree age analysed separately) summed over multiple years to model cumulative yield, for trees aged 5 to 7 years old, aged 5 to 8, and aged 6 to 8.

To determine the accuracy of genomic prediction for yield stability over multiple years, GEBVs were obtained using the following model:

Yield SD = mean + Accession + error Equation 5-7

where **Yield SD** is the standard deviation of solutions (corrected phenotypes from Equations 5-4 or 5-5) for ages 5 to 8 across all sites.

Variance components were recorded for each model. Estimates of genomic narrow-sense genomic heritability (h^2) for yield were calculated from variance components:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{gs}^2 + \sigma_e^2}$$ Equation 5-8

where σ_g^2 is the additive genetic variance, σ_{gs}^2 is the G x E variance, and σ_e^2 is the residual variance. This was calculated using the three datasets (young-tree yield, mature-tree yield, and combined young- and mature-tree yield data), for individual sites and across all sites. Heritability was also calculated for yield stability.

5.3.4 Assessing model accuracy

The accuracy of the GEBVs from the models above were determined using five-fold cross-validation (CV). In turn, 20% of phenotypes were masked (set to missing) in a validation set, and data for the remaining 80% of individuals were used as a training set to train the model and predict the missing values. This process was repeated five times until all subsets were used in the validation set, with each individual used only once in the validation set. Note that the phenotypes of the validation set individuals were not included in the training set for that fold of CV.

Individuals were assigned to one of five groups for the five-fold CV using two grouping techniques for predictions: random and related family groups. For the random CV, individuals were selected for the training and validation group at random ("randomly grouped"). Here,

full-sibs were split across the training and validation groups, and so predictions were performed on individuals related to the training population. For the second method, individuals were grouped by family and related families (those with common parents) and grafted parents ("grouped families"), to give approximately equal-sized groups. Thus, entire full-sib families were either all in the reference set or in the validation set, and predictions were performed on families unrelated to the training population ("unrelated population"). This second method represents an extreme version of the application of GS, where the two populations are not closely related; in a real-world application, the training population is likely to be related to the target population due to overlapping parental germplasm among breeding populations. n

For each CV, prediction accuracy (r) was calculated as the correlation between GEBVs and corrected phenotypes (from Equations 5-4 or 5-5, corresponding to individual site, across sites, individual year, or across multiple years GEBVs), divided by the square root of the genomic narrow-sense heritability (as calculated using all data across all sites). Mean prediction accuracies and standard errors were calculated across the five CVs, and t-tests were performed to determine if prediction accuracies were significantly difference from zero. T-tests were also performed to determine if prediction accuracies using GEBVs for the two genetic effects (Tree and Tree + Site) were statistically different.

5.3.5 Comparison of breeding strategies and genetic gain

A simple preliminary comparison of breeding strategies was made to demonstrate how GS could be effectively incorporated into the macadamia breeding program to reduce selection time and increase genetic gain (Table 5-1). Number of trees involved in each stage and specific costs are excluded (given uncertainties of, and constantly evolving genotyping costs). Two breeding strategies were compared:

1. Traditional breeding. Progeny are evaluated in an SPT (seedling progeny trial) for at least eight years to select for yield and other economically important traits (such as KR, precocity and tree size) using a selection index. SPT is then followed by an RVT (regional variety trial) for at least eight years, where selected elites are clonally propagated and evaluated for more economically important traits across multiple environments.

2. Genomic selection. Progeny seedlings' first leaves after germination are genotyped using a large number of markers. A robust GS model is then used to predict yield (and other traits). Elite candidates are selected, using a selection index, for establishment and evaluation in the RVT.

Table 5-1 Activities involved in a traditional breeding strategy compared with a simple example of how genomic selection (GS) could be employed in a breeding program. The number of years involved in each activity for the two strategies is shown. Information for 'traditional breeding' is adapted from Topp et al. (2012). RVT, regional variety trial; SPT, seedling progeny trial

Year	Traditional breeding	Genomic selection
1	Cross parents, grow seedlings	Cross parents, grow seedlings
2	Age 1: Plant SPT	Genotype, select seedlings using GS
3	Age 2: Trial maintenance	Propagate RVT
4	Age 3: Trial maintenance	Age 1: Plant RVT
5	Age 4: Evaluations	Age 2: Trial maintenance
6	Age 5: Evaluations	Age 3: Trial maintenance
7	Age 6: Evaluations	Age 4: Evaluations
8	Age 7: Evaluations	Age 5: Evaluations
9	Age 8: Evaluations, select seedlings	Age 6: Evaluations
10	Propagate RVT	Age 7: Evaluations
11	Age 1: Plant RVT	Age 8: Evaluations
12	Age 2: Trial maintenance	Age 9: Evaluations
13	Age 3: Trial maintenance	Age 10: Evaluations
14	Age 4: Evaluations	Release
15	Age 5: Evaluations	
16	Age 6: Evaluations	
17	Age 7: Evaluations	
18	Age 8: Evaluations	
19	Age 9: Evaluations	
20	Age 10: Evaluations	
21	Release	

Genetic gain (ΔG, grams/year) was calculated for traditional breeding and GS methods using the following equation derived from Falconer (1989):

$$\Delta G = \frac{i \times r \times \sigma}{L}$$ Equation 5-9

where *i* is selection intensity as a function of the proportion of the population selected, *r* is square root of yield heritability for traditional breeding or the prediction accuracy of the GS model, σ is the genetic standard deviation (standard deviation of corrected phenotypes, from Equation 5-4 or 5-5), and *L* is generation length in years. For cumulative yield, genetic gain was further divided by the number of years summed, to give genetic gain on a per-year basis. In traditional breeding, approximately 2,000 seedlings are evaluated and 1% (20/2000) of the SPT population are further tested in an RVT (Topp et al. 2016). Here, the selected proportion of the population has been increased from 1% to 2.5% for the GS strategy, in an attempt to reduce the probability of not selecting truly elite germplasm. As such, in this equation, *i* = 2.665 and 2.338, for 1 and 2.5% selected, respectively (Falconer 1989). We assume that genetic gain for RVT selection is the same across selection strategies, and so genetic gain is only calculated here for the SPT.

5.4 Results

5.4.1 Heritability and accuracy of prediction models

Genetic variance varied between sites for the three datasets, and, thus, heritability was also variable. Site EG had low genetic variance for young-tree yield data (0.16) and high variance for mature data (0.65; Table 5-2). Variance attributed to G x E was low when data for all sites were analysed together (0.13 for young-tree data and 0.09 for combined data). Heritability was generally lowest when both young-tree and mature-tree yield data were combined, and highest when only mature-tree data were used (Table 5-2). Narrow-sense heritability for yield stability across years was 0.04 (data not shown).

Table 5-2 Variance components, as a proportion of total variance (1.00), and narrow-sense heritability (h^2) of yield using raw observations. σ_g^2 genetic variance, σ_{gs}^2 genotype x environment variance, σ_e^2 residual variance. $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_{gs}^2 + \sigma_e^2)$.

Dataset	**Site**	$\boldsymbol{\sigma_g^2}$	$\boldsymbol{\sigma_{gs}^2}$	$\boldsymbol{\sigma_e^2}$	$\mathbf{h^2}$
Young	AL	0.22		0.78	0.22
	AM	0.41		0.59	0.41
	EG	0.16		0.84	0.16
	HP	0.44		0.56	0.44
	All Sites	0.28	0.13	0.59	0.28
Mature	AL	0.09		0.91	0.09
	AM	0.39		0.61	0.39
	EG	0.65		0.35	0.65
	HP	0.50		0.50	0.50
Combined	AL	0.17		0.83	0.17
	AM	0.25		0.75	0.25
	EG	0.13		0.87	0.13
	HP	0.38		0.62	0.38
	All Sites	0.23	0.09	0.68	0.23

A heritability value of 0.23 for yield, based on the mean across sites as well as from all sites together for combined young-tree and mature-tree yield data, was used for calculations of genomic prediction accuracy. Mean genomic prediction accuracy across datasets and cross-validation methods for individual sites was r = 0.37. Model accuracies varied widely, from -0.29 for mature-tree yield data at AM to 0.97 for combined yield data at EG (Figure 5-1). For almost all analyses, higher accuracies of genomic prediction were achieved from cross-validation using randomly masked individuals compared to when families were grouped (prediction for unrelated populations), sometimes by two-fold (Figure 5-1). For example, prediction accuracy for EG using young-tree yield data was r = 0.40 for grouped families and 0.90 for randomly masked individuals. The highest prediction for grouped families was for HP using combined yield (r = 0.52, $p < 0.05$), and was similar when averaged across sites for both young-tree yield and combined yield data (r = 0.32, $p < 0.05$ and r = 0.35, $p < 0.05$, respectively).

Accuracies of genomic prediction using young-tree data were similar to combined young and mature-tree data (Figure 5-1). Site AM was consistently lower in prediction accuracies than

the other sites across all datasets. Very high prediction accuracies were observed using random masking of individuals, compared to family masking (across family prediction), for sites HP and EG using young-tree yield data ($r = 0.94$, $p < 0.001$ and $r = 0.90$, $p < 0.05$, respectively), and for EG using combined yield data ($r = 0.97$, $p < 0.05$). Site AL had moderately high prediction accuracies using random masking for young-tree yield (0.65, $p < 0.01$) and combined data (0.62, $p < 0.05$). Prediction accuracies using only mature data were very low across sites, ranging from -0.29 to 0.39 (Figure 5-1). Further analyses were, therefore, not performed using only mature-tree yield data.

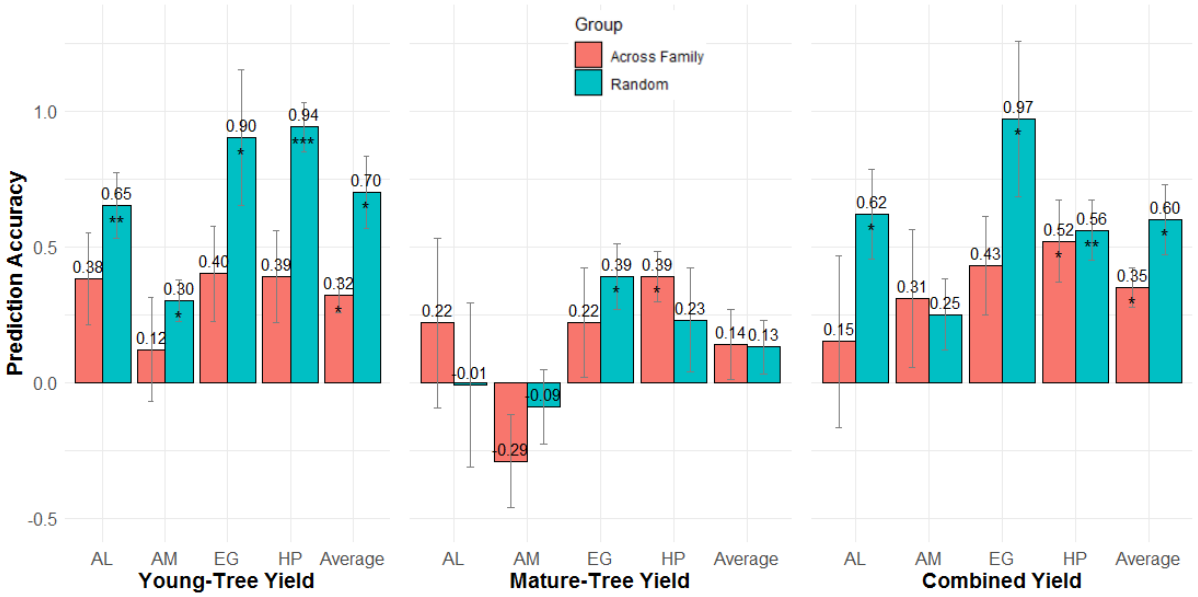


Figure 5-1 Mean prediction accuracy of yield across three datasets: young-tree yield, mature-tree yield and combined young and mature-tree yield (averaged over years). Two cross-validation (CV) grouping methods were compared: trees were randomly grouped so predictions were performed in related individuals, and trees were grouped by family so predictions were performed across unrelated populations. Accuracies are compared between individual sites, as well as an average across all sites, for each model (dataset and CV method). Prediction accuracy values are given for each model, as well as p-values indicating whether accuracies are significantly different from zero: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; blank, not significant. Prediction accuracy was measured as the correlation between predicted genotypic values (GEBVs) and corrected phenotypes divided by the square root of the heritability ($h^2 = 0.23$). Error bars indicate standard error of correlations from five cross-validations for the individual sites, or the four sites for Average.

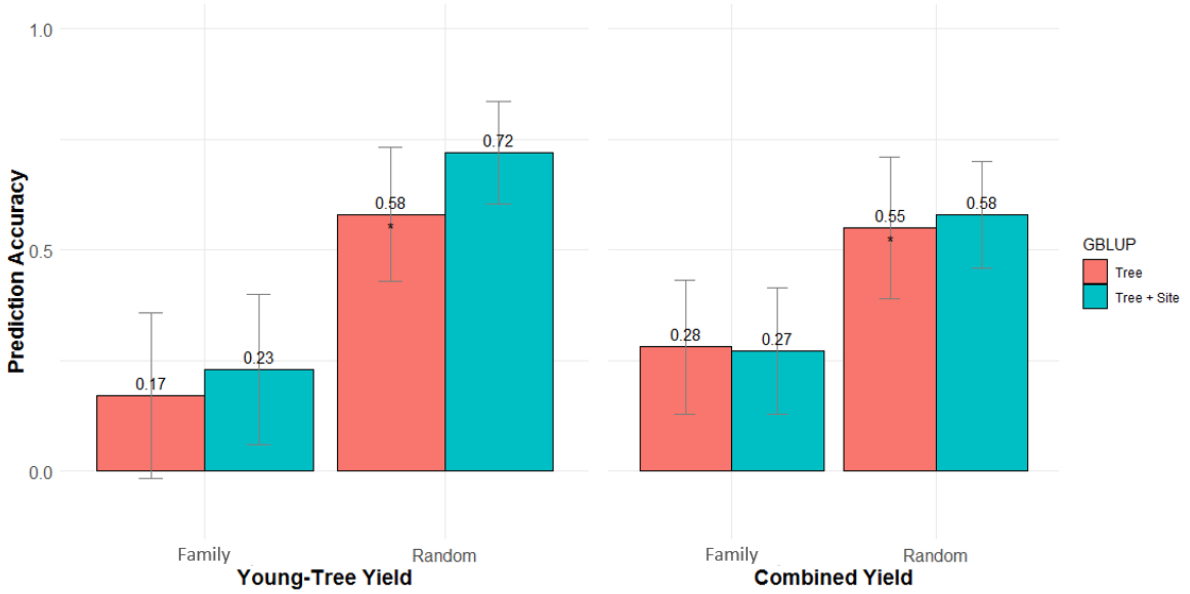


Figure 5-2 Comparison of mean prediction accuracy of yield for two different GBLUP genetic effects: average tree genetic effect (Tree), and average tree genetic effect plus the genetic effect of that tree at the corresponding site (Tree + Site). Comparisons are across young-tree yield and combined young and mature-tree yield, as well as family grouped (predictions in unrelated populations) and randomly grouped (predictions in related populations) cross-validation methods. Prediction accuracy values are given for each model, measured as the correlation between predicted genotypic values and corrected phenotypes divided by the square root of the heritability (h^2 = 0.23). Error bars indicate standard error of correlations from five cross-validations. * indicates that prediction accuracy is significantly ($p < 0.05$) different from zero.

The use of two different GBLUP genetic effects in genomic predictions were compared; using the average tree genetic effect, and the sum of the average tree genetic effect and the genetic effect of that tree at the corresponding site (i.e. a model allowing for G x E was used to obtain these predictions). Average tree genetic effect generally gave lower prediction accuracies than average tree genetic effect plus site effect, but there was no significant difference in prediction accuracy between the two GBLUP models (Figure 5-2). Further analysis was, therefore, confined to the GEBV for overall tree genetic effect.

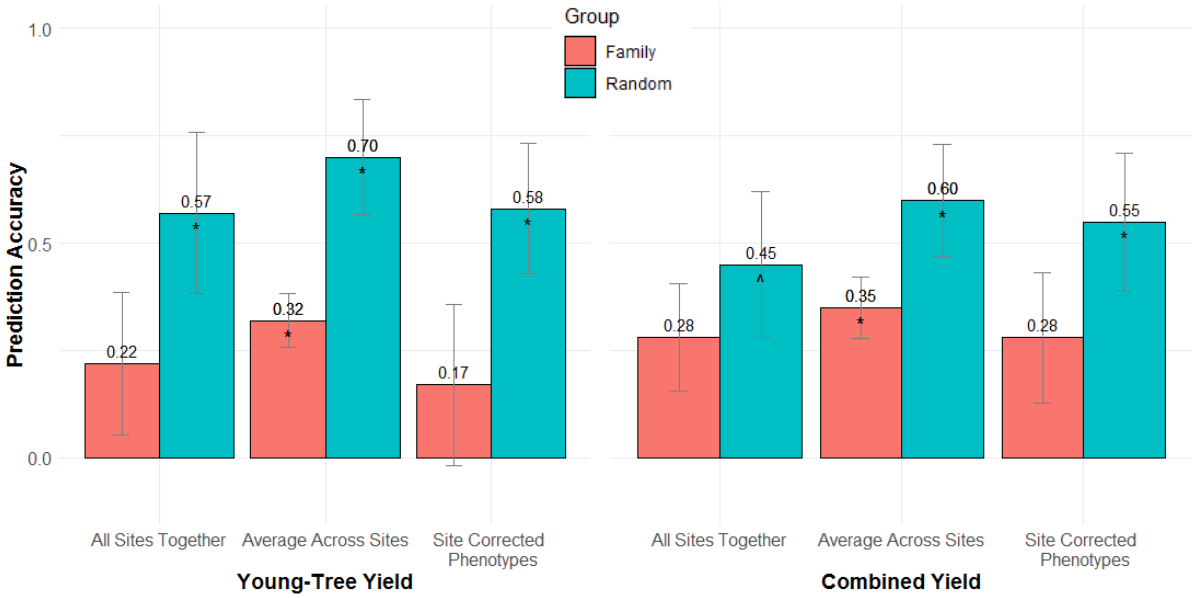


Figure 5-3 Mean prediction accuracy of yield across three methods: using phenotypes corrected with all sites analysed together (all sites together), an average accuracy across the four individual sites (average across sites), and using phenotypes corrected from individual sites and then combined (site corrected phenotypes). Prediction accuracy values for average across sites are the same as in Figure 5-1. Accuracies are compared for two datasets: young-tree yield and combined young and mature-tree yield, with two cross-validation methods: randomly grouped individuals (predictions in related populations) and individuals grouped by family (predictions in unrelated populations). Prediction accuracy values are given for each model, as well as p-values indicating whether accuracies are significantly different to zero: * $p < 0.05$, ^ $p = 0.06$; blank, not significant. Prediction accuracy was measured as the correlation between predicted genotypic values and corrected phenotypes divided by the square root of the heritability ($h^2 = 0.23$). Error bars indicate standard error of correlations from five cross-validations.

The highest yield prediction across all models was achieved when sites were analysed individually and then accuracies averaged across sites (Figure 5-3). Prediction accuracies using this method were 0.32 ($p < 0.01$) and 0.70 ($p < 0.05$) for family and random grouping, respectively, for young-tree yield, and 0.35 ($p < 0.01$) and 0.60 ($p < 0.05$) for combined yield data for family and random grouping, respectively (Figure 5-1 and Figure 5-3). Prediction accuracies were similar for young-tree data analysed with all sites together ($r = 0.57$, $p < 0.05$)

and for phenotypes corrected for individual sites and then combined (0.58, $p < 0.05$) using random groupings. Predictions in unrelated population (grouped families) was lowest when phenotypes were corrected for individual sites first and then combined, using young-tree yield data (0.17).

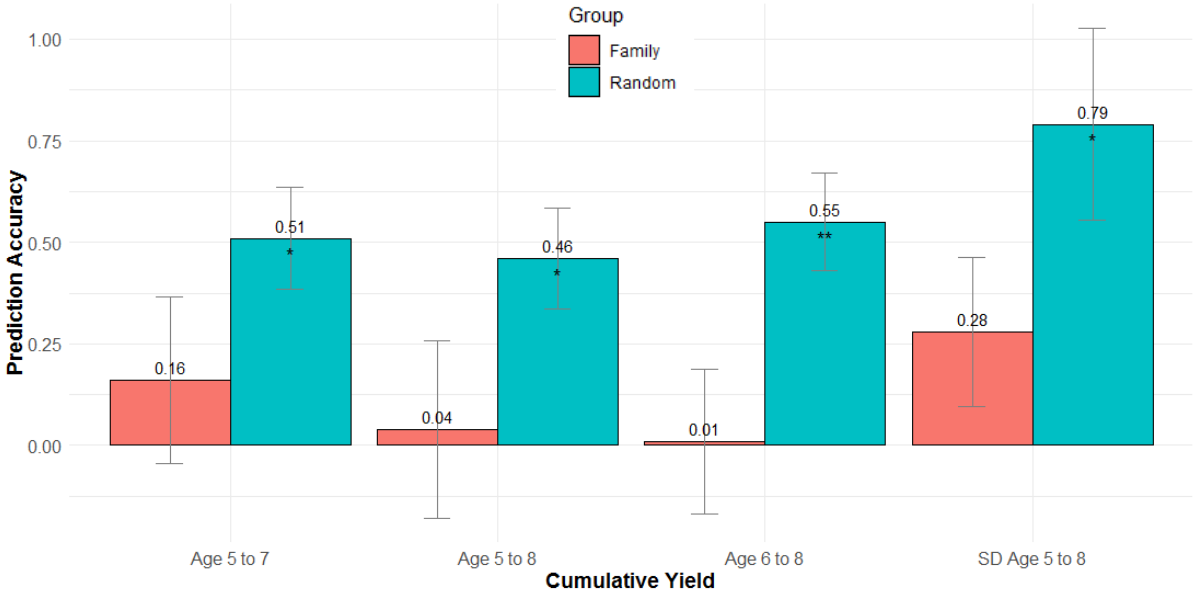


Figure 5-4 Prediction accuracy of cumulative yield and yield stability, using individual year GBLUPs across all sites together. Accuracies are compared for cumulative yield from age 5 to 7 years, age 5 to 8, age 6 to 8, and yield stability as a function of standard deviation (SD) of yield from age 5 to 8. Two cross-validation methods are shown: randomly grouped individuals (predictions in related populations) and individuals grouped by family (predictions in unrelated populations). Prediction accuracy values are given for each model, as well as p-values indicating whether accuracies are significantly different from zero: * $p < 0.05$, ** $p < 0.01$; blank, not significant. Prediction accuracy was measured as the correlation between predicted genotypic values and corrected phenotypes divided by the square root of the heritability (yield $h^2 = 0.23$, yield stability $h^2 = 0.04$). Error bars indicate standard error of correlations from five cross-validations.

Genomic prediction accuracies were moderate for cumulative yield (sum of corrected phenotypes over years) using random groups, but low for unrelated population prediction (Figure 5-4). The highest accuracies were achieved for cumulative yield from age 6 to 8 years

(r = 0.55, p < 0.01), then for age 5 to 7 (0.51, p < 0.05), with age 5 to 8 being the lowest (0.46, p < 0.05) for randomly grouped individuals. Prediction accuracies for cumulative yield (0.46–0.55) were less accurate than mean yield across years (young-tree yield data; 0.57–0.70) using random groupings (Figure 5-3). Yield stability had high prediction accuracy for random groups (0.79, p < 0.05), and moderate for unrelated population predictions (0.28), though this was not significantly different from zero (Figure 5-4).

5.4.2 Comparison of breeding strategies and genetic gain

The generation length (L) for traditional breeding was eight years, as elite individuals are identified after evaluations from age 5 to 8 and are then used as parents for the next generation (Hardner et al. 2009). By comparison, the generation length for strategies employing GS is four years. This difference is because elite individuals are identified from genetic markers using the first leaf, but cannot be used as parents until reproductive maturity (around the age of 4) (Hardner et al. 2009). The strategy using GS had a much shorter selection cycle (14 years) than traditional breeding (21 years), because it negates the SPT altogether. Both strategies employ RVTs, as it is vital to test the performance of candidate cultivars across multiple environments before commercial release.

Genetic gain using traditional breeding methods was calculated as 202 g/year for 1% selection intensity (Table 5-3). At 2.5% selection intensity, genetic gain was reduced to 177 g/year. The shorter generation cycle of GS strategies compared with traditional breeding influenced estimates of genetic gain. Generally, genetic gain for GS related family (randomly grouped) predictions was more than double that of traditional breeding, even with a lower selection intensity for GS methods (2.5%) compared to traditional breeding (1%). However, for unrelated population predictions, traditional breeding achieved higher genetic gain than GS efforts.

Genetic gain varied between GS models from 12 to 673 g/year, and was lower for 2.5% selection intensity than 1% (Table 5-3). For unrelated population predictions, the highest genetic gain was achieved using raw phenotypes across all sites (ΔG = 162 g/year for p% = 2.5), as opposed to correcting phenotypes by site first (128 g/year). The opposite was true for

randomly grouped individuals; phenotypes corrected according to site slightly outperformed analysis using all the sites together (ΔG = 438 g/year compared to 421).

Table 5-3 Genetic gain of yield and yield stability (in g/year) for each selection method and unrelated population or random cross-validation techniques. Genetic gain was calculated using accuracy of genomic selection (GS) model or square root of yield heritability for traditional breeding, standard deviation of corrected phenotypes (yield stability), generation length of four years for GS methods and eight years for traditional breeding, and 1% (i = 2.665) and 2.5% (i = 2.338) selection intensity, as per Equation 5-9.

Breeding method	**Genetic gain ΔG (g/year)***	
	p% = 1	**p% = 2.5**
Traditional Breeding	202	177
Family grouped (unrelated population) GS		
All Sites Corrected Phenotypes	146	128
All Sites Raw Phenotypes	185	162
Cumulative Age 5 to 7	163	143
Cumulative Age 5 to 8	45	40
Cumulative Age 6 to 8	14	12
Yield stability (SD Age 5 to 8)	174	153
Randomly grouped (related population) GS		
All Sites Corrected Phenotypes	499	438
All Sites Raw Phenotypes	480	421
Cumulative Age 5 to 7	534	468
Cumulative Age 5 to 8	482	423
Cumulative Age 6 to 8	673	590
Yield stability (SD Age 5 to 8)	488	428

* For cumulative yield, genetic gain is divided by the number of summed years

For cumulative yield, genetic gain was poor for age 6 to 8 with unrelated population predictions (12 g/year for p% = 2.5), whilst a much higher gain was observed for age 5 to 7 (163 g/year). Conversely, for randomly grouped individuals, cumulative yield from age 6 to 8 had very high genetic gain (590 g/year for p% = 2.5). Genetic gain was lower for cumulative yield from age 5 to 8 (423 g/year) and age 5 to 7 (468 g/year at p% = 2.5), though these were both much higher than that estimated for traditional breeding. Stability of yield had a genetic

gain comparable to mean yield across all sites for both unrelated population (153 g/year at p% = 2.5) and random grouping predictions (428 g/year; Table 5-3).

5.5 Discussion

5.5.1 Comparison of prediction models and cross-validation methods

This study is the first to investigate the use of GS to improve genetic gain for yield and yield stability in macadamia breeding, and suggests that GS offers a suitable method to select trees with predicted high GEBVS. Genomic prediction accuracy varied across models, harvest years, and for different CV methods. The most accurate prediction of GEBVs was with young-tree yield data averaged across sites using randomly grouped individuals, followed by combined young and mature-tree yield data. Calculating corrected phenotypes for individual years and then summing to obtain cumulative yield did not produce more accurate predictions than mean yield across years.

Prediction accuracy is strongly influenced by the relatedness between training and validation populations (Meuwissen et al. 2001), and unrelated population predictions are expected to perform poorly compared to related family prediction (Pszczola et al. 2012). This was observed across the models in the current study; random groupings performed consistently better than family groupings (predictions in unrelated populations). This is because with random groupings for CV, the training set includes full-sibs from the validation set (e.g. progeny from the same cross will be split across the training and validation sets), and so large blocks of chromosomes will be shared between the training and validation sets. The low to moderate prediction accuracies observed by Muranty et al. (2015) in apple were attributed to predictions across unrelated populations. In comparison, Kumar et al. (2012b) achieved high prediction accuracies (0.70 to 0.90) for apple fruit quality traits, with individuals randomly allocated to cross-validation groups. The CV method of family group prediction represents an extreme version of the potential real world application of GS in macadamia where predictions are performed across unrelated populations; it is likely that the training and target populations will actually be more closely related. This is due to the fact that there is often an overlap of industry varieties used as parents between breeding populations, and elite individuals from one population are commonly used as parental germplasm in

subsequent generations (Topp et al. 2012). It is expected that prediction of GEBVs in a breeding program will, therefore, have accuracies on a level between the random and unrelated population predictions presented in this study. Employing GS in a population closely related to that which the model is based from would provide more accurate predictions of yield. However, more research is needed using large training population sizes with validation sets of whole family groups to improve prediction accuracy before GS can be applied in macadamia breeding.

The very low prediction accuracies across sites using mature-tree yield data may be due to only one or two years' data available, as no yield data were available for AL in 2017 or for EG in 2018. As such, mature-tree yield data were only investigated for individual sites and not across all sites. Having more than two consecutive years of yield data appears to increase the accuracy of prediction in macadamia. Evaluations of seedling progeny and elite candidates currently involves at least four years of measuring yield, which makes it a sound data base for GS.

As this is the first study to investigate GS in macadamia using linear mixed models, simple models were used that did not include permanent environment effects. This may not be optimal because the model is not accounting for correlations between residuals across years. However, a test comparing models with and without permanent environment effects showed that the predicted values of the two models were highly correlated. This correlation demonstrates that there is little difference between the predicted values of the two models, and that the less complex model was still adequate for predicting GEBVs. Cedillo et al. (2018) found that permanent effects over five evaluation years among oil palm plants were highly significantly different, reflecting environmental variation among plots within site blocks. Further, permanent effect accounted for 15% of variance explained in walnut yield, compared with around 30% for both breeding value and year variance (Martínez-García et al. 2017). Future research using linear mixed models to predict yield in macadamia using genomics could include more complex models with permanent environment effects to determine if prediction accuracies are more accurate.

5.5.2 Factors affecting accuracy of genomic prediction

The prediction accuracies for yield in the current study were reasonable (r ≈ 0.55), and comparable to the accuracy of yield as measured by phenotypes (h^2 = 0.23, h = r = 0.48, where r is accuracy). This demonstrates that GS accuracy demonstrated in the current study may be as good as phenotypic analysis, regardless of the time advantage in GS strategies. In comparison with other studies, the accuracy achieved in GS here was not as high as some other predictions in other horticulture crops, which may be attributed to a number of factors. Estimates of macadamia yield involve a large degree of error, as indicated by the low heritability for this trait. Measures of yield can be inaccurate due to the overlapping canopies of neighbouring trees. Additionally, the method used to obtain DNIS weight per harvest assumes that the moisture content of the 1 kg sample is consistent through the entire harvest. For these reasons, measuring macadamia yield is very different to measuring yield in other fruit crops, which may inhibit accurate yield prediction.

This study is, to our knowledge, the first to use GS to predict stability of yield over consecutive years for a nut tree. Biennial bearing in apple has been researched by multiple authors. Guitton et al. (2012) found three QTLs associated with biennial bearing, by measuring number and mass of harvested fruit, and that these explain 50% of phenotypic variability. Additionally, Durand et al. (2013) suggested that irregular bearing may be less linked to fruit set or drop, but rather more with floral induction. Predictions using randomly grouped individuals were moderately high for yield stability, though this may be due to the low heritability of the trait upwardly biasing the calculated prediction accuracy. Nevertheless, these results will be informative for breeders to factor yield stability into a selection index when identifying elite candidates for further testing.

The population size of this study was limited compared to other studies in fruit crops, though it did consist of a large number of full-sib families. In the first study of GS in cross-pollinated fruit crop species, Kumar et al. (2012b) obtained high model accuracy for fruit quality traits in apple. They used a much larger population (1,120 seedlings) than the current study, albeit from a smaller parent population (seven full-sib families from four female and two male parents), and accuracy ranged from r = 0.70 to 0.90 using RR-BLUP and Bayesian LASSO methods. GS in citrus achieved high (r > 0.7) prediction accuracy for some fruit quality traits using around 800 individuals, with GBLUP model consistently being high performing than

other models (Minamikawa et al. 2017). Similarly, using a Japanese pear population of 86 parents and 765 progeny, prediction accuracy varied between models and cross-validation methods, and was commonly greater than 0.5 (Minamikawa et al. 2018). However, the high correlations found for citrus and Japanese pear may be slightly inflated, since negative correlation coefficients were set to zero when calculating accuracy for these studies. Increasing the size of a phenotyped and genotyped training population would increase the accuracy of yield prediction in macadamia.

LD between markers and genes controlling target traits is essential for GS (Meuwissen et al. 2001). Increasing the number of markers used in GS may not necessarily achieve better accuracies. Studies investigating the accuracy of GS in citrus, Japanese pear and apple all used fewer SNP markers than the current study (1,841, 1,502 and 2,500, respectively) (Minamikawa et al. 2017; Minamikawa et al. 2018; Kumar et al. 2012b). Recent genotyping of a macadamia breeding population with yield phenotypes produced 4,113 SNP markers (O'Connor et al. 2019b; Chapter 1 Population structure and genetic diversity). SNPs within 1 kb distance of each other on a scaffold (*M. integrifolia* v2 genome assembly, 4,098 scaffolds) had an average LD of $r^2 = 0.124$, with LD decaying rapidly over short distances and slowly over long distances (O'Connor et al. 2019b). These results are important for the current study to determine that genetic markers capture genetic variance of the target trait (Goddard 1991; Meuwissen et al. 2001). Increasing the density of markers across the genome could lead to increased prediction accuracies, as suggested by Calus et al. (2008), where models with $r^2 = 0.2$ between markers were more accurate than models with fewer markers and lower densities. Future analysis of LD in macadamia could also include corrections for population structure and cryptic relatedness.

Genetic recombination occurs with successive generations of breeding, which may affect the linkage between markers and genes controlling target traits (Khan and Korban 2012). Additionally, selection for improved individuals will also alter the frequency of alleles in the population (Falconer 1989). These changes over time will have consequences for GS accuracy. Meuwissen et al. (2001) estimated that the accuracy of GS models will decrease at around 5% per generation, due to recombination. Thus, it is necessary to recalibrate the model after every few generations, as genetic variance explained by the markers will change, along with the allelic frequencies in the population (Goddard 2009; Resende et al. 2012). To aid this,

Grattapaglia (2014) suggested that selection candidates should remain in the field and grown for five to six years to provide phenotypes for updating the model. This strategy could be employed in macadamia to ensure accuracy through subsequent generations of GS.

5.5.3 Genetic gain from genomic selection

Genetic gain was greatly influenced by the length of the breeding cycle. Genotyping seedlings using their first leaf after germination, to identify high-yielding individuals through GS, could halve the length of the SPT. Then, elite trees could be crossed to produce the next generation as soon as they begin to flower, usually around the age of four. Similarly in forestry, Grattapaglia (2014) demonstrated that GS could be used to halve the selection time from ten to five years to achieve an improved population after selecting elite trees from a candidate population. Muranty et al. (2015) also suggested that GS could increase genetic gain per year in apple compared with conventional breeding, by shortening the breeding cycle from seven to four years. In contrast, GS accuracy was not high enough for all target traits in oil palm to reduce the generation interval, meaning that breeding would still require the testing of progeny (Cros et al. 2017). They suggested that, if given the resources to increase the size of the training set, and a greater ability to model G x E interactions, GS could be a valid option to increase genetic gain in oil palm.

van Nocker and Gardiner (2014) reviewed the work of Kumar et al. (2012b; 2012a) regarding GS in apple. They proposed using MAS and GS to identify elite apple accessions, and then, to decrease time to reproductive maturity, to implement a regime to promote early flowering. Fruit would be phenotyped over two early seasons, and then BVs compared with the predicted GEBVs to analyse genetic gain. Using these methods, candidate cultivars could be clonally propagated seven years earlier than traditional breeding. However, the predicted beneficial outcomes of using GS in apple may not be as achievable if predictions were to occur across families rather than in randomly-grouped individuals, as has been shown here in macadamia.

5.5.4 Logistics of using genomic selection to increase genetic gain

The opportunity to employ GS in a wider range of crops is increasing with declining genotyping costs and advancements in technology (Heffner et al. 2009; Khan and Korban 2012; Iwata et al. 2016). However, implementing genomics-assisted breeding will be expensive due to the cost of genotyping large numbers of candidates at each cycle, though this will be a trade-off with a decrease in the costs needed for phenotyping (Heffner et al. 2010). An evaluation of costs involved in marker-assisted selection (MAS) versus GS has been made for maize and wheat, and GS outperformed MAS even when prediction accuracies were low (Heffner et al. 2010). Breeders should compare selection strategies to determine which combinations of genotyping and phenotyping is most suitable for their crop and program to maximise accuracy of trait prediction in fruit crops (Muranty et al. 2015).

To reduce genotyping costs, delaying GS to deploy on a smaller population size may be a viable option. Seedlings could be grown out as per a traditional SPT, but only evaluated to age four, and precocious (early bearing) trees evaluated for KR. Breeders could pre-select precocious seedlings with high KR, genotype this reduced number of elite individuals, and then the highest-yielding trees could be selected through GS for evaluation in RVTs. This method of delayed GS is similar to that proposed by Gardiner et al. (2014); to reduce the size of the seedling population to be genotyped, pre-screen the population for essential traits first. Longer generation intervals, due to phenotyping for a number of years initially, would lead to a lower genetic gain using this strategy than GS of more seedlings at an earlier stage; however, it may be a more cost-effective option. Additionally, whilst implementing GS in macadamia may not decrease the time from seed to reproductive maturity, selecting for precocious individuals may aid in producing more individuals with a shortened juvenile stage. Reaching reproductive maturity at an earlier stage will further increase genetic gain by reducing the generation length of four years in the GS strategy. Comparing costs of traditional breeding versus strategies using GS is not the focus of this study, though this should be evaluated to determine the prospect of implementing GS in the Australian macadamia breeding program.

5.5.5 Future research using GS in macadamia

Future work employing GS to increase genetic gain in macadamia could investigate other economically important traits, such as tree size. *Chapter 4 Genome-wide association study for yield component traits* found 14 QTLs linked with trunk circumference. The large number of markers associated with this trait, compared with other traits in the study, means that GS may be more appropriate than GWAS and MAS to increase genetic gain, given the seemingly quantitative nature of trunk circumference. GS may also be a good candidate for other traits, such as resistance to diseases, including husk spot and phytophthora (Drenth et al. 2009). Furthermore, the significant associations identified between traits and markers, as found in *Chapter 4*, could be incorporated into GS models. Genomic prediction methods including BayesR and BayesB allow the effect of some markers, such as those of significant effect, to be larger than others (Meuwissen et al. 2001; Erbe et al. 2012). Different model types could, therefore, be tested in the future to determine which are the most accurate in predictions.

Further work could also include multi-trait models, to investigate whether the inclusion of additional traits, such as trunk circumference and nut weight, increases the accuracy of yield prediction. Jia and Jannink (2012) found that prediction accuracy was increased for a trait with low heritability by including information for a correlated trait with high heritability. Estimates of heritability and genetic correlations between yield and component traits were calculated in *Chapter 3 Component traits of yield*, and, thus, this information could be used to inform multi-trait GS. Distinctions can also be made between linked QLTs (linkage between multiple QTLs affecting different traits) and pleiotropic QTL (one gene affecting multiple traits), using multi-trait methods, like those employed by Bolormaa et al. (2014).

Finally, future GS analyses should involve more genetic markers across the genome. This may ensure that small-effect loci are captured, since LD in macadamia decays rapidly over short distances (O'Connor et al. 2019b). With the aid of a complete reference genome, future sequencing of individuals for GS analysis and the calling of SNPs may be more accurate and avoid potential issues associated with allelic dropouts (O'Connor et al. 2019b).

5.6 Conclusions

In a world-first, we have found moderate to high prediction accuracies applying GS models for yield and yield stability prediction in macadamia. Highest prediction accuracies were observed for young-tree yield data and combined young and mature-tree yield data for randomly grouped individuals averaged across sites. Unrelated population predictions were generally lower than predictions for related families, but due to the relatedness in parental germplasm between subsequent breeding generations, a realistic prediction accuracy would be somewhat between those observed for the two grouping methods. Results from this study indicate that GS is a viable option to increase genetic gain in macadamia, though more research and resources are needed to increase the size of the training population, and phenotype and genotype these individuals to capture markers in LD with causal polymorphisms. With accurate genomic prediction models, future macadamia breeding can sequence seedlings and use models to predict yield. Then individuals with predicted high yield could be propagated for testing in further evaluations for yield and nut characteristics. Genomic prediction could negate the need to evaluate progeny, if accurate enough, therefore shortening the SPT and increasing genetic gain. This work could be combined with GWAS and MAS for key nut traits.

Chapter 6. General discussion and conclusions

6.1 Purpose of thesis

Development of new varieties in horticulture tree crops is often hindered by the high costs of evaluation due to long juvenile periods and hence generation time, large plants requiring large areas of land for evaluations, and high costs involved in evaluating traits over multiple seasons. Macadamia is a nut crop that faces these challenges. Alternative selection strategies could decrease the time taken to identify elite varieties for cultivar development and hence reduce costs. The thesis sought to:

1. Quantify the level of genetic diversity and population structure in a breeding population.
2. Investigate alternative selection strategies for high yield including indirect selection using yield component traits, genome-wide association studies (GWAS) for important component traits, and genomic selection (GS) for yield and yield stability.
3. Compare existing breeding strategies with alternative selection strategies in terms of time and efficiency.

The outcomes of this work will aid future breeding in macadamia, and have application to other fruit and nut crops. This chapter summarises the outcomes and conclusions of the research conducted in Chapters 2 through 5.

6.2 Achievement of thesis objectives: Outcomes and impact

6.2.1 Population structure and genetic diversity of the population

A total of 4,113 SNP and 16,171 silicoDArT markers were produced for 295 full-sib progeny and their 29 parents. Genetic diversity, analysed using GenAlEx software, was similar among seedling progeny and their parents (H_E = 0.255 for progeny and 0.250 for parents), but appeared lower than other fruit crops. Furthermore, progeny from interspecific hybrid parents were more genetically diverse (H_E = 0.278) than pure *M. integrifolia* seedlings (H_E = 0.189). Analysis conducted in STRUCTURE software found that the progeny population was

moderately differentiated (F_{ST} = 0.401), and clustered into k = 3 clusters, representing the *M. integrifolia* germplasm separating from two hybrid groups. LD decayed rapidly over short distances of genome assembly scaffolds (n = 4,098 scaffolds), whereas low level LD persisted for long distances; there was an average LD (r^2) value of 0.124 for SNPs within 1 kilobase of each other.

The low marker density and low LD have implications for GWAS and GS, where accuracy relies on linkage between genetic markers and causal polymorphisms. The genomic relationship matrix (GRM) constructed in the study using SNP data to model population structure and kinship was essential for the subsequent research chapters, as realised relationships are more accurate than recorded pedigrees (Hayes et al. 2009b). This is the first study to quantify the genetic diversity of a large group of macadamia full-sib progeny and their parents using a large array of molecular markers, and the knowledge gained will be valuable for future studies using genetic markers in macadamia.

6.2.2 Indirectly selecting for high yield using component traits

The study investigated twelve yield component traits, ranging from nut and flowering characteristics to trunk circumference (as a measure of tree size). Analyses were performed using ASReml to estimate narrow-sense heritability and additive genetic correlations between each component trait and yield. Three traits (number of nuts per raceme, rachis diameter at nut set, and percentage of flowers that set nuts) were moderately correlated with yield (r_g = 0.55, 0.56 and 0.41, respectively). However, heritability of these traits was low (h^2 = 0.17, 0.15 and 0.33, respectively), thus inhibiting improvement through selection. Nut weight (NW) and trunk circumference (TC) were the only component traits that were moderately to highly correlated with yield (r_g = 0.45 and 0.72, respectively), had moderate to high heritability (h^2 = 0.59 and 0.44, respectively), and were also easily measured. A negative genetic correlation (-0.27) was observed between yield and kernel recovery (KR), an important economical trait.

This study was the first to use genetic markers to estimate narrow-sense heritability of yield component traits and to calculate genetic correlations between these traits and yield in a macadamia breeding population. The variance of trait phenotypes and the proportion of that controlled by additive genetic variance (narrow-sense heritability) revealed the opportunity

for genetic improvement of some traits in the breeding program through selection. The correlations between yield NW (0.45), KR (-0.27), and TC (0.72), indicate that it may be difficult to breed for trees that satisfy the macadamia industry's aims for intermediate nut size, high kernel recovery, and smaller tree size (Hardner et al. 2009). Based on the genetic correlations with yield and heritabilities, none of the component traits analysed in this study were promising candidates to select for yield, as the calculated selection efficiency for each trait was lower than that for direct selection of yield. This chapter informed research efforts in the next chapter, genome-wide association study of component traits.

6.2.3 Markers associated with key component traits

GWAS was performed to identify genetic markers significantly associated with several nut characteristics, flowering traits, and TC, where individual SNP markers were modelled as fixed effects and kinship was modelled with a GRM. Statistically significant markers (at a false discovery rate of <0.05, to control for Type I errors) were mapped to genome assembly scaffolds, and LD analysis was performed to determine if any markers were linked. Significant associations were detected for NW (n = 7 SNPs), percentage of whole kernels (n = 4), and TC (n = 44). Multiple regression, as well as mapping of markers to genome assembly scaffolds, suggested that the same quantitative trait loci (QTL) region was being identified by multiple SNP loci.

This chapter provides a foundation for genomics-assisted breeding in macadamia and nut crops more broadly, and advances our understanding of the genetic control of yield component traits, as it gives a preliminary indication of the number of genes controlling each trait. As a complete reference genome is not yet available for macadamia, the locations of markers on chromosomes is not yet known, but this knowledge may help inform the location and nature of the causal genes. Nonetheless, the identification of some QTLs for key traits remains a useful resource and will provide a stepping-stone towards marker-assisted selection (MAS) in the future.

6.2.4 Genomic selection for yield and yield stability

This study was the first to assess the prospect of employing GS in macadamia. To our knowledge, it is also the first to use GS to predict yield stability for a nut tree crop. Yield stability (measured here as the standard deviation of yield over consecutive years) is an essential factor to ensure consistent income for growers over consecutive years, and so genomic prediction for this trait is important. Narrow-sense heritability, as estimated from the genomic data, was low to moderate, at $h^2 = 0.23$ for yield and 0.04 for yield stability. A number of different GS models were examined across multiple datasets, and using four consecutive years of yield data produced the most accurate results (as assessed with cross-validation). Predictions across unrelated populations were less accurate than predictions using related individuals (randomly grouped individuals) predictions. Across all sites for young-tree yield data, the accuracy of related family prediction was moderate, at 0.57, compared with 0.22 for unrelated population prediction. In comparison, prediction accuracy for yield stability was high within families, at 0.79, though this may have been upwardly biased by the low heritability for this trait.

Genetic gain per year was compared between selection strategies, and was largely influenced by the length of the breeding cycle; cycle length using GS strategies was four years, compared with eight years for traditional breeding efforts. Genetic gain using GS for related family predictions (421–438 g/year, at 2.5% selection intensity) was double that for traditional breeding (202 g/year), and was also high for cumulative yield over multiple years (423–590 g/year). Genetic gain for yield stability was much higher for related family predictions (428 g/year) than unrelated population predictions (153 g/year).

This study shows that simple GS models can achieve moderate yield prediction accuracies. GS could be incorporated into the Australian macadamia breeding program in multiple ways. However, the cost of genotyping breeding populations appears to be a large constraint for employing GS at present. One strategy to reduce genotyping costs could be to initially screen progeny for important traits that are displayed early, such as precocity and KR, and then genotyping could be performed on a subset of individuals showing desirable characteristics to identify predicted high yielding trees using GS. More research should be conducted to identify the best prediction model, and validate findings before putting into practice.

6.3 Challenges and limitations of the study

This study has examined strategies that may be used to increase yield in macadamia breeding. The significant marker-trait associations discovered in Chapter 4, and the moderate prediction accuracies achieved in Chapter 5 are encouraging for genomics to be used in future breeding efforts. However, this study suffered some limitations.

6.3.1 Population size and location

The size of the population used to discover QTLs and train genomic prediction models should be large and representative of the breeding population as a whole. The population used in this study (n ≈ 300) was representative of the larger breeding population, but was limited in size compared with other studies of tree crops, for example, 1,120 in apple (Kumar et al. 2012b), 676 in citrus (Minamikawa et al. 2017) and 765 in pear (Minamikawa et al. 2018). As such, the conclusions of this study may not be accurate enough to be readily applied to the whole Australian macadamia breeding population. Further work in genomics should use a larger population base, if funding permits, to increase accuracy by using a wider germplasm base to develop prediction models.

Genotype by environment interactions were important considerations in the research chapters of this thesis. Due to the large size and number of progeny seedlings evaluated in macadamia breeding, progeny have been planted across multiple sites. As such, analyses need to consider if there is an impact of the locations, though this was not found to be the case in Chapter 3. Furthermore, having trees located at multiple locations was essential for robust experimental design; however, multiple environments meant that phenotyping was compromised by weather, equipment availability and orchard owners' priorities. This resulted in complex analyses with yield data being unable to be collected in two instances, which meant that fewer data were available for analysis, particularly in Chapter 5. Despite the limitations in sampling, the data collected is representative of the wider population base, and shows promising results that lay the groundwork for future studies.

6.3.2 Logistics of using genomics and macadamia breeding

The results of Chapters 4 and 5 suggest that the breeding program could be improved by using genomics-assisted breeding: GWAS was successfully used to identify significant associations between markers and key traits, and GS achieved moderate prediction accuracies for yield and yield stability. However, one major limitation of this approach would be the cost involved in genotyping thousands of seedling progeny. While genotyping and sequencing technologies accelerate in their development, prices of services and/or data points may decrease in the future. The price of genotyping and the accuracy of prediction should be compared with the price of evaluations of progeny using traditional breeding methods to determine strategy efficiency, such as that which has been investigated by other research groups (e.g. Wong and Bernardo 2008; Harshman et al. 2016; Ru et al. 2016; Heffner et al. 2010). Furthermore, increasing the number of good quality markers used in genomics-assisted breeding could ensure that causal polymorphisms are captured by LD with markers across the genome. For example, Calus et al. (2008), suggested that models with $r^2 = 0.2$ (0.128 cM) between markers were more accurate than models with fewer markers and lower densities. The issues regarding marker quality, particularly allelic dropouts potentially leading to lower levels of heterozygosity than expected, will need to be investigated in the future, and sequencing with the aid of a complete reference genome may help alleviate some of these issues.

Simple GS models were tested in this preliminary study. More complex modelling with permanent environment effects were not included at this stage, though it was demonstrated that models with and without these effects produce highly correlated GEBVs ($r = 0.93$, $p < 0.001$). Further work on GS in macadamia could explore models with such permanent environment effects included to potentially more accurately fit the data. The results of the genomics studies are yet to be validated to determine how accurate predictions of nut characteristics and yield are in practice, and so the effectiveness of employing MAS and GS to increase genetic gain is yet to be adequately demonstrated. The efficiency and accuracy of GWAS/MAS and GS will be improved over multiple generations by incorporating added phenotypic and genotypic data to update prediction models.

Genomic prediction accuracies were low to moderate for unrelated population predictions, and considerably lower than predictions for related family groups, where family members were included in both the training and validation sets. This demonstrates that the relatedness

between the training and validation populations is important, with closely related populations achieving higher prediction accuracies, as observed in numerous studies of other crops. Seedling populations may share parental genotypes across different generations, but this may not always be the case. Until unrelated population prediction accuracies are increased, perhaps with a larger population base or high marker density, the application of GS to predict yield might only be viable between closely related populations.

Future genomic prediction models will need to be updated regularly to maintain accuracy, as repeated cycles of selection will change allele frequencies in the population and breakdown LD between markers and causal polymorphisms. Therefore, the breeding program should retain some seedling progenies to phenotype over time and retrain the model by updating phenotypic and genotypic data and recalibrate allelic frequencies. Periodical phenotyping will add to the labour and cost involved in breeding, but an accurate model leading to increased genetic gain will be an asset to the breeding program.

A final obstacle in employing genomics to increase genetic gain may be the response of growers to a new technology. Macadamia growers often prefer to see candidate varieties in person at field days before they will agree to plant clones on their property. Viewing of the trees will still be possible with regional variety trials (RVT) evaluating candidate varieties across multiple environments, even if the seedling progeny trial is negated using genomics. The RVT was included in all selection strategies proposed in Chapter 5. Therefore, if some growers are reluctant to trust the efficiency of new technologies, they can still view trees and data from the RVT, regardless of how GS and MAS is adopted in the breeding program.

6.4 Further research

This doctoral study has covered multiple facets of breeding for high yield using component traits and genomics to improve accuracy and efficiency in macadamia, and has laid the groundwork for future advancements. Though this work investigated heritabilities of, and correlations between many component traits and yield, the list of traits is far from exhausted. There are other important traits that, whilst may not be used to indirectly select for high yield, may be economically important in themselves. Harvest index, or the allocation of resources to edible biomass (Donald and Hamblin 1976), is an example of a trait that should be

investigated using the approaches in this study. Since nuts are covered in husk, and husk can dehisce from the tree and be harvested along with the nut-in-shell, a large proportion of the biomass harvested from each tree is actually inedible, and, therefore, not beneficial for the grower. Further research should investigate whether the investment of energy used to produce husk occurs at the expense of kernel production. Genotypes that can allocate resources to large kernels and thin husks would be beneficial; harvested biomass would be targeted towards the portion that is economically important, rather than 'by-products'.

The marker-trait associations detected in this study should be further explored and validated in a separate study. GWAS can also be used to identify marker-trait associations for other important traits in macadamia, such as self-fertility and resistance to pests and diseases. Macadamia is predominantly out-crossing (Hardner et al. 2009), but genotypes that are self-fertile may produce more nuts due to the larger amount of pollen available for successful pollination. It would also be highly beneficial to identify the causal polymorphisms associated with resistance to diseases that reduce yield, including husk spot and phytophthora (Drenth et al. 2009). With genotyping and sequencing technology advancing, future studies will likely have many thousands of markers available. Combined with the completion of a macadamia reference genome, the locations of these markers on the genome can be confirmed, which can then be used to help identify causal genes. As such, genomics studies will have more genotypic information available to pair with high-quality phenotypic data. For example, with more markers available across the entire genome, future GWAS studies will likely be more effective in detecting QTLs controlling each trait.

The GS models for yield constructed in this research also need to be validated in a separate population, preferably with more individuals to represent wider phenotypic variance, and with more genetic markers to ensure that more of the small-effect genes controlling yield are captured by LD. GS models could also be constructed for other economically important traits that may be controlled by many genes, such as trunk circumference or tree size. Multi-trait genomic prediction models may improve the accuracy of prediction by incorporating more data, including that for yield component traits. Multivariate analyses could be undertaken to increase the power of marker detection, such as methods proposed by Bolormaa et al. (2014), where a distinction can be made between pleiotropic QTL (one gene affecting multiple traits) and linked QLT (linkage between multiple QTLs affecting different traits). Furthermore, the

results of GWAS could be accommodated in genomic predictions by using a method other than GBLUP, such as BayesR or BayesB, which allow the effect of some markers (e.g. those of significant effect) to be larger (Meuwissen et al. 2001; Erbe et al. 2012).

At this stage, a major limitation of using genomics in macadamia breeding is the cost of genotyping. An economic analysis is required to compare the costs involved in genotyping a seedling progeny population with traditional breeding methods, and possibly alternative genotyping methods. An analysis of costs could also incorporate alternative strategies such as the delayed progeny genomic selection described in Chapter 5. With an assessment of cost and genetic gain of each selection strategy, the relative advantage of using genomics could be understood and quantified in the breeding program.

6.5 Conclusions

This research has made a substantial contribution to the use of component traits and genomics for increased genetic gain in fruit and nut tree crops. Chapter 2 quantified the genetic diversity and population structure of the subset of the breeding program used in this work, which was informative for the genomics research chapters. The hypothesis that component traits could be used to indirectly select for yield was investigated in Chapter 3; however, no examined traits could be identified with both high heritability and high genetic correlation with yield. Chapters 4 and 5 evaluated the use of employing genomics in macadamia breeding. The potential to use GWAS and MAS to select seedlings predicted to have desirable characteristics was analysed in Chapter 4. Genetic markers were found to have significant associations for several key traits. After validation in a separate population and further estimates of the proportion of variance explained by key markers, these associations are contenders for use in MAS when screening future seedling populations. The accuracy of predicting yield and yield stability using GS was evaluated in Chapter 5, and while accuracies varied across datasets and models, moderate accuracies suggested that this might be a potential method to increase genetic gain in macadamia. Genetic gain was comparable to traditional breeding methods for unrelated population predictions, and greater for randomly grouped individuals. Different strategies for employing GS were compared, with options to drastically reduce the length of the seedling progeny trial selection phase. With further

research and development, the application of MAS and GS to fruit and nut breeding around the world may allow this practice to become more sustainable and accelerate genetic gain into the future.

References

Acquaah G (2012) Principles of Plant Genetics and Breeding. 2nd edn. John Wiley & Sons, West Sussex, UK

Akagi T, Hanada T, Yaegaki H, Gradziel TM, Tao R (2016) Genome-wide view of genetic diversity reveals paths of selection and cultivar differentiation in peach domestication. DNA Research 23 (3):271-282

Alam M, Neal J, O'Connor K, Kilian A, Topp B (2018) Ultra-high-throughput DArTseq-based silicoDArT and SNP markers for genomic studies in macadamia. PLOS One 13 (8):e0203465. doi:10.1371/journal.pone.0203465

Aliyu O (2006) Phenotypic correlation and path coefficient analysis of nut yield and yield components in cashew (*Anacardium occidentale* L.). Silvae Genetica 55 (1):19-24

Allard RW, Bradshaw AD (1964) Implications of genotype-environmental interactions in applied plant breeding. Crop Science 4 (5):503-508

Aradhya MK, Yee LK, Zee FT, Manshardt RM (1998) Genetic variability in *Macadamia*. Genetic Resources and Crop Evolution 45 (1):19-32. doi:10.1023/A:1008634103954

Australian Macadamia Society (2012) Global Macadamia Production. Accessed 15/07/17

Australian Macadamia Society (2017a) 2017 Australian macadamia crop reaches 46,000 tonnes in-shell. http://australian-macadamias.org/industry/site/industry/industry-page/industry-news-archive/latest-news-industry/2017-australian-macadamia-crop-reaches-46000-tonnes-in-shell?Itemid=133&lang=en. Accessed 16/01/18

Australian Macadamia Society (2017b) Australia's Macadamia Industry in Numbers. https://app-ausmacademia-au-syd.s3.ap-southeast-2.amazonaws.com/factfigure/wNu2i3awkACVahT73qwcNtHTrw71qgY0bj2LMt2r.pdf . Accessed 22/05/2018

Australian Macadamia Society (2017c) Australian Macadamia Society Factsheet: Farm Gate Prices. Accessed 11/01/19

Australian Macadamia Society Estimated World Macadamia Production. In: XXXVII International Nut and Dried Fruit Congress, Spain, 2018.

Bai B, Zhang YJ, Wang L, Lee M, Ye BQ, Alfiko Y, Purwantomo S, Suwanto A, Yue GH (2018) Mapping QTL for leaf area in oil palm using genotyping by sequencing. Tree Genetics & Genomes 14 (2):31-39

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nature Reviews Genetics 7 (10):781-791

Barrett B, Kidwell K, Fox P (1998) Comparison of AFLP and pedigree-based genetic diversity assessment methods using wheat cultivars from the Pacific Northwest. Crop Science 38 (5):1271-1278

Bazzaz FA, Chiariello NR, Coley PD, Pitelka LF (1987) Allocating resources to reproduction and defense. BioScience 37 (1):58-67

Beavis W The power and deceit of QTL experiments: Lessons from comparative QTL studies. In: Proceedings of the 49th annual corn and sorghum industry research conference, 1994. Chicago, IL, pp 250-266

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57 (1):289-300

Bernardo R (2008) Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Science 48 (5):1649-1664

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Science 47 (3):1082-1090

Biscarini F, Nazzicari N, Bink M, Arús P, Aranzana MJ, Verde I, Micali S, Pascal T, Quilot-Turion B, Lambert P (2017) Genome-enabled predictions for fruit weight and quality from repeated records in European peach progenies. BMC Genomics 18 (1):432-446

Bodzon Z (2004) Correlations and heritability of the characters determining the seed yield of the long-raceme alfalfa (*Medicago sativa* L.). Journal of Applied Genetics 45 (1):49-60

Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, Tier B, Savin K, Hayes BJ, Goddard ME (2014) A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. PLOS Genetics 10 (3):e1004198

Borghi B, Accerbi M, Corbellini M (1998) Response to early generation selection for grain yield and harvest index in bread wheat (*T. aestivum* L.). Plant Breeding 117 (1):13-18

Boyton S, Hardner C Phenology of flowering and nut production in macadamia. In: Drew R (ed) International Symposium on Tropical and Subtropical Fruits, Cairns, Queensland, 2002. Acta Horticulturae, pp 381-387

Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: The missing heritability is in the field. Genome biology 12 (10):1-8

Butler D, Cullis B, Gilmour A, Gogel B (2009) Asreml: asreml () fits the linear mixed model. R package, version 3

Calus M, Meuwissen T, De Roos A, Veerkamp R (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178 (1):553-561

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10 (1):421

Campbell AJ, Maddox CD, Morris. SC (2005) Assessment protocols for nut-borer resistance - Macadamia husk hardness 1999–2000. In: McConchie CA (ed) Macadamia Improvement by Breeding Stage, vol 2. Horticulture Australia Limited, Sydney, pp 205–242

Campbell NA, Reece JB (2002) Biology. 6th edn. Pearson Education, Inc., San Francisco, CA

Cannell M (1985) Dry matter partitioning in tree crops. In: Cannell MGRJ, J.E (ed) Attributes of trees as crop plants. Institute of Terrestrial Ecology, Huntington, England, pp 160-193

Cantín CM, Gogorcena Y, Moreno MÁ (2010) Phenotypic diversity and relationships of fruit quality traits in peach and nectarine [*Prunus persica* (L.) Batsch] breeding progenies. Euphytica 171 (2):211-226

Cao K, Wang L, Zhu G, Fang W, Chen C, Luo J (2012) Genetic diversity, linkage disequilibrium, and association mapping analyses of peach (*Prunus persica*) landraces in China. Tree Genetics & Genomes 8 (5):975-990

Cedillo DO, Barrera CF, Cedillo JO, Carrera JO, de Resende MDV, Cruz CD (2018) Estimates of parameters, prediction and selection of an oil palm population in Ecuador. Revista Facultad Nacional de Agronomía Medellín 71 (2):8477-8487

Cellon C, Amadeu RR, Olmstead JW, Mattia MR, Ferrao LFV, Munoz PR (2018) Estimation of genetic parameters and prediction of breeding values in an autotetraploid blueberry breeding population with extensive pedigree data. Euphytica 214:1-13

Chagné D (2015) Whole Genome Sequencing of Fruit Tree Species. In: Christophe P, Anne-Françoise A-B (eds) Advances in Botanical Research, vol 74. Academic Press, pp 1-37

Chapman J, Nakagawa S, Coltman D, Slate J, Sheldon B (2009) A quantitative review of heterozygosity–fitness correlations in animal populations. Molecular ecology 18 (13):2746-2765

Chen W, Hou L, Zhang Z, Pang X, Li Y (2017) Genetic diversity, population structure, and linkage disequilibrium of a core collection of *Ziziphus jujuba* assessed with genome-wide SNPs developed by genotyping-by-sequencing and SSR markers. Frontiers in Plant Science 8:575-588

Cilas C, Montagnon C, Bar-Hen A (2011) Yield stability in clones of *Coffea canephora* in the short and medium term: Longitudinal data analyses and measures of stability over time. Tree Genetics & Genomes 7 (2):421-429. doi:10.1007/s11295-010-0344-4

Clark S, Van der Werf J (2013) Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. In: Gondro C, Van der Werf J, Hayes B (eds) Genome-Wide Association Studies and Genomic Prediction. Springer Science, London, UK,

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142 (1-2):169-196

Cooper M, DeLacy I (1994) Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theor Appl Genet 88 (5):561-572

Cros D, Bocs S, Riou V, Ortega-Abboud E, Tisné S, Argout X, Pomiès V, Nodichao L, Lubis Z, Cochard B (2017) Genomic preselection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses. BMC Genomics 18:839-855

Cros D, Denis M, Bouvet JM, Sánchez L (2015) Long-term genomic selection for heterosis without dominance in multiplicative traits: Case study of bunch production in oil palm. BMC Genomics 16:651-668

Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun H-J (2010) Prediction of genetic values

of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186 (2):713-724

da Rocha Sobierajski G (2012) Development and use of SSR and DArT genetic markers to study genetic diversity in macadamia (*Macadamia integrifolia*). PhD Thesis, Universidade de São Paulo, São Paulo

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLOS One 3 (10):e3395

de Nettancourt D (1977) Incompatibility in Angiosperms. Springer-Verlag, Berlin

de Souza VA, Byrne DH, Taylor JF (1998) Heritability, genetic and phenotypic correlations, and predicted selection response of quantitative traits in peach: I. An analysis of several reproductive traits. Journal of the American Society for Horticultural Science 123 (4):598-603

Dekkers J (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. Journal of Animal Breeding and Genetics 124 (6):331-341

Denis M, Bouvet J-M (2013) Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. Tree Genetics & Genomes 9 (1):37-51

Desta ZA, Ortiz R (2014) Genomic selection: Genome-wide prediction in plant improvement. Trends in Plant Science 19 (9):592-601

Dicenta F, Garcia J (1992) Phenotypical correlations among some traits in almond. Journal of Genetics and Breeding 46:241-246

Diehl WJ, Biesiot PM (1994) Relationships between multilocus heterozygosity and morphometric indices in a population of the deep-sea red crab *Chaceon quinquedens* (Smith). Journal of Experimental Marine Biology and Ecology 182 (2):237-250. doi:https://doi.org/10.1016/0022-0981(94)90054-X

Donald C (1962) In search of yield. CIMMYT 28 (No. REP-10660):171–178

Donald C, Hamblin J (1976) The biological yield and harvest index of cereals as agronomic and plant breeding criteria. Advances in Agronomy 28:361-405

Douglas JA, Skol AD, Boehnke M (2002) Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. The American Journal of Human Genetics 70 (2):487-495. doi:10.1086/338919

Drenth A, Akinsanmi OA, Miles A (2009) Macadamia diseases in Australia. Southern African Macadamia Growers' Association Yearbook 17:48-52

Druet T, Macleod IM, Hayes BJ (2014) Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112 (1):39-47

Durand J-B, Guitton B, Peyhardi J, Holtz Y, Guédon Y, Trottier C, Costes E (2013) New insights for estimating the genetic value of segregating apple progenies for irregular bearing during the first years of tree production. Journal of Experimental Botany 64 (16):5099-5113

Duvick DN (1984) Genetic contributions to yield gains of U.S. hybrid maize, 1930 to 1980. In: Fehr WR (ed) Genetic Contributions to Yield Gains of Five Major Crop Plants. CSSA

Special Publication, vol 7. Crop Science Society of America and American Society of Agronomy, Madison, WI, pp 15-47

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genet Resour 4 (2):359-361. doi:10.1007/s12686-011-9548-7

Eaton G, Kyte T (1978) Yield component analysis in the cranberry. Journal of the American Society for Horticultural Science 103 (5):578-583

Emanuelli F, Lorenzi S, Grzeskowiak L, Catalano V, Stefanini M, Troggio M, Myles S, Martinez-Zapater JM, Zyprian E, Moreira FM (2013) Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biology 13 (1):39-55

Endresen DTF (2010) Predictive association between trait data and ecogeographic data for Nordic barley landraces. Crop Science 50 (6):2418-2430

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 95 (7):4114-4129

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Molecular ecology 14 (8):2611-2620. doi:10.1111/j.1365-294X.2005.02553.x

Falconer DS (1989) Introduction to Quantitative Genetics. 3rd edn. Longman Scientific & Technical, Essex, England

Falconer DS, Mackay TF (1996) Introduction to Quantitative Genetics. 4th edn. UK Longman Group, Sussex, UK

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164 (4):1567-1587

Fischer MC, Rellstab C, Leuzinger M, Roumet M, Gugerli F, Shimizu KK, Holderegger R, Widmer A (2017) Estimating genomic diversity and population differentiation - an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. BMC Genomics 18 (1):69-83

Flint-Garcia SA, Thornsberry JM, Buckler IV ES (2003) Structure of linkage disequilibrium in plants. Annual review of plant biology 54 (1):357-374

Fraser J, Eaton GW (1983) Applications of yield component analysis to crop research. Field Crops Abstracts 36 (10):787-797

Gardiner S, Volz R, Chagné D Tools to breed better cultivars faster at Plant & Food Research. In: Proceedings of the 1st International Rapid Cycle Crop Breeding Conference, 2014. pp 7-9

Gilmour AR, Gogel B, Cullis B, Thompson R, Butler D (2009) ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Gitonga L, Muigai A, Kahangi E, Ngamau K, Gichuki S (2009) Status of macadamia production in Kenya and the potential of biotechnology in enhancing its genetic improvement. Journal of Plant Breeding and Crop Science 1 (3):49-59

Glaszmann J-C, Kilian B, Upadhyaya HD, Varshney RK (2010) Accessing genetic diversity for crop improvement. Current Opinion in Plant Biology 13 (2):167-173

Goddard M (1991) Mapping genes for quantitative traits using linkage disequilibrium. Genetics Selection Evolution 23 (Suppl 1):131s-134s

Goddard M (2009) Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136 (2):245-257

Goddard ME, Hayes BJ (2007) Genomic selection. Journal of Animal Breeding and Genetics 124 (6):323-330. doi:10.1111/j.1439-0388.2007.00702.x

Gondro C, Lee SH, Lee HK, Porto-Neto LR (2013) Quality Control for Genome-Wide Association Studies. In: Gondro C, van der Werf J, Hayes B (eds) Genome-Wide Association Studies and Genomic Prediction. Springer Science, London,

Govindaraj M, Vetriventhan M, Srinivasan M (2015) Importance of genetic diversity assessment in crop plants and its recent advances: An overview of its analytical perspectives. Genetics Research International 2015:Article 431487

Grattapaglia D (2014) Breeding forest trees by genomic selection: Current progress and the way forward. In: Tuberosa R, Graner A, Frison E (eds) Genomics of Plant Genetic Resources, vol Volume 1. Managing, Sequencing and Mining Genetic Resources. Springer, London, pp 651-682

Grattapaglia D, Resende MD (2011) Genomic selection in forest tree breeding. Tree Genetics & Genomes 7 (2):241-255

Grattapaglia D, Silva-Junior OB, Resende RT, Cappa EP, Müller BS, Tan B, Isik F, Ratcliffe B, El-Kassaby YA (2018) Quantitative genetics and genomics converge to accelerate forest tree breeding. Frontiers in Plant Science 9:Article 1693

Gross C (1995) *Macadamia*. Flora of Australia 16 (Proteaceae):419-425

Grzebelus D, Iorizzo M, Senalik D, Ellison S, Cavagnaro P, Macko-Podgorni A, Heller-Uszynska K, Kilian A, Nothnagel T, Allender C, Simon PW, Baranski R (2014) Diversity, genetic mapping, and signatures of domestication in the carrot (*Daucus carota* L.) genome, as revealed by Diversity Arrays Technology (DArT) markers. Molecular Breeding 33 (3):625-637. doi:10.1007/s11032-013-9979-9

Guitton B, Kelner J-J, Velasco R, Gardiner SE, Chagne D, Costes E (2012) Genetic control of biennial bearing in apple. Journal of Experimental Botany 63 (1):131-149

Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity 101:5-18. doi:10.1038/hdy.2008.35

Habier D, Fernando R, Dekkers J (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177 (4):2389-2397

Hamilton RA, Fukunaga ET (1959) Growing Macadamia Nuts in Hawaii, vol 121. Hawaii Agricultural Experiment Station, University of Hawaii

Hamilton RA, Ito PJ (1984) Macadamia nut cultivars recommended for Hawaii, vol 023. College of Tropical Agriculture and Human Resources, University of Hawaii, Cooperative Extension Service (USA). Hawaii Institute of Tropical Agriculture and Human Resources, University of Hawaii

Hansche P, Beres V, Forde H (1972) Estimates of quantitative genetic properties of walnut and their implications for cultivar improvement. Journal of the American Society for Horticultural Science 97 (2):279-285

Hansche PE (1983) Response to Selection. In: Moore JNJ, J. (ed) Methods in Fruit Breeding Purdue University Press, West Lafayette, Indiana, pp 154-171

Hardenbol P, Yu F, Belmont J, MacKenzie J, Bruckner C, Brundage T, Boudreau A, Chow S, Eberle J, Erbilgin A (2005) Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. Genome research 15 (2):269-275

Hardner C (2016) Macadamia domestication in Hawai'i. Genetic Resources and Crop Evolution 63 (8):1411-1430

Hardner C (2017) Exploring opportunities for reducing complexity of genotype-by-environment interaction models. Euphytica 213 (11):248-264

Hardner C, Costa e Silva J, Williams E, Meyers N, McConchie C (2019a) Breeding new cultivars for the Australian macadamia industry. HortScience 54 (4):621-628. doi:https://doi.org/10.21273/HORTSCI13286-18

Hardner C, Greaves B, Coverdale C, Wegener M Application of economic modelling to support selection decisions in macadamia. In: Mercer C (ed) 13th Australasian Plant Breeding Conference, Christchurch, New Zealand, 2006. pp 426-431

Hardner C, Peace C, Henshall J, Manners J Opportunities and constraints for marker-assisted selection in macadamia breeding. In: Drew R (ed) International Symposium on Harnessing the Potential of Horticulture in the Asian-Pacific Region, Coolum, Queensland, 2005. Acta Horticulturae, pp 85-90

Hardner C, Pisanu P, Boyton S (2004) National macadamia germplasm conservation program. Horticulture Australia Limited, Sydney

Hardner C, Winks C, Stephenson R, Gallagher E (2001) Genetic parameters for nut and kernel traits in macadamia. Euphytica 117 (2):151-161

Hardner CM, Hayes BJ, Kumar S, Vanderzande S, Cai L, Piaskowski J, Quero-Garcia J, Campoy JA, Barreneche T, Giovannini D (2019b) Prediction of genetic value for sweet cherry fruit maturity among environments using a 6K SNP array. Horticulture Research 6 (1):6-20

Hardner CM, Peace C, Lowe AJ, Neal J, Pisanu P, Powell M, Schmidt A, Spain C, Williams K (2009) Genetic resources and domestication of macadamia. Horticultural Reviews 35:1-126

Hardner CM, Winks CW, Stephenson RA, Gallagher EG, McConchie CA (2002) Genetic parameters for yield in macadamia. Euphytica 125 (2):255-264

Harshman JM, Evans KM, Hardner CM (2016) Cost and accuracy of advanced breeding trial designs in apple. Horticulture Research 3:Article 16008

Hayes B (2013) Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). In: Gondro C, Van der Werf J, Hayes B (eds) Genome-Wide Association Studies and Genomic Prediction. Springer Science, London,

Hayes B, Daetwyler H, Bowman P, Moser G, Tier B, Crump R, Khatkar M, Raadsma H, Goddard M (2009a) Accuracy of genomic selection: Comparing theory and results. Proceedings of the Association for the Advancement of Animal Breeding and Genetics 18:34-37

Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal breeding. Genome 53 (11):876-883

Hayes BJ, Visscher PM, Goddard ME (2009b) Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research 91:47-60

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome research 13 (4):635-643

He J, Zhao X, Laroche A, Lu Z-X, Liu H, Li Z (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Frontiers in Plant Science 5:Article 484

Heffner EL, Jannink J-L, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome 4 (1):65-75

Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME (2010) Plant breeding with genomic selection: Gain per unit time and cost. Crop Science 50 (5):1681-1690

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Science 49 (1):1-12

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31 (2):423-447

Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic selection in plant breeding: A comparison of models. Crop Science 52 (1):146-160

Hill W (1984) On selection among groups with heterogeneous variance. Animal Science 39 (3):473-477

Howlett BG, Nelson WR, Pattemore DE, Gee M (2015) Pollination of macadamia: Review and opportunities for improving yields. Scientia Horticulturae 197:411-419

Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. Annual review of plant biology 65:531-551

Huett DO (2004) Macadamia physiology review: a canopy light response study and literature review. Australian Journal of Agricultural Research 55 (6):609

Igarashi M, Hatsuyama Y, Harada T, Fukasawa-Akada T (2016) Biotechnology and apple breeding in Japan. Breeding Science 66 (1):18

Imai A, Nonaka K, Kuniga T, Yoshioka T, Hayashi T (2018) Genome-wide association mapping of fruit-quality traits using genotyping-by-sequencing approach in citrus landraces, modern cultivars, and breeding lines in Japan. Tree Genetics & Genomes 14 (2):24-38

Isik F (2014) Genomic selection in forest tree breeding: The concept and an outlook to the future. New Forests 45 (3):379-401

Isik F, Kumar S, Martínez-García PJ, Iwata H, Yamamoto T (2015) Acceleration of Forest and Fruit Tree Domestication by Genomic Selection. In: Plomion C, Adam-Blondon A-F (eds) Advances in Botanical Research, Land Plants - Trees, vol 74. Elsevier, Oxford, UK, pp 93-124

Ito PJ (1980) Effect of style removal on fruit set in macadamia. HortScience 15 (4):520-521

Iwata H, Hayashi T, Terakami S, Takada N, Sawamura Y, Yamamoto T (2013) Potential assessment of genome-wide association study and genomic selection in Japanese pear *Pyrus pyrifolia*. Breeding Science 63 (1):125-140

Iwata H, Hayashi T, Tsumura Y (2011) Prospects for genomic selection in conifer breeding: a simulation study of *Cryptomeria japonica*. Tree Genetics & Genomes 7 (4):747-758

Iwata H, Minamikawa MF, Kajiya-Kanegae H, Ishimori M, Hayashi T (2016) Genomics-assisted breeding in fruit trees. Breeding Science 66 (1):100-115

Jakobsson M, Rosenberg NA (2007) CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23 (14):1801-1806

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: From theory to practice. Briefings in Functional Genomics 9 (2):166-177

Ji K, Zhang D, Motilal LA, Boccara M, Lachenaud P, Meinhardt LW (2013) Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. Genetic Resources and Crop Evolution 60 (2):441-453

Jia Y, Jannink J-L (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192 (4):1513-1522

Kelner J-J, Costes E, Guitton B, Chagné D, Gardiner SE, Velasco R (2011) Genetic control of biennial bearing in apple. Journal of Experimental Botany 63 (1):131-149. doi:10.1093/jxb/err261

Kendall M, Stuart A, Ord J (1987) Kendall's advanced theory of statistics, vol 3. design and analysis, and time series, 4th edn. Oxford University Press, New York,

Kester DE, Asay R (1975) Almonds. In: Janick J, Moore JN (eds) Advances in Fruit Breeding. Purdue University Press, Indiana, USA,

Khan MA, Korban SS (2012) Association mapping in forest trees and fruit crops. Journal of Experimental Botany 63 (11):4045-4060

Kilian A, Huttner E, Wenzl P, Jaccoud D, Carling J, Caig V, Evers M, Heller-Uszynska K, Cayla C, Patarapuwadol S The fast and the cheap: SNP and DArT-based whole genome profiling for crop improvement. In: Tuberosa R, Phillips RL, Gale M (eds) Proceedings of the

International Congress In the Wake of the Double Helix: From the Green Revolution to the Gene Revolution, Bologna, Italy, 2003. Avenue Media, pp 443-461

Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H, Caig V, Heller-Uszynska K, Jaccoud D, Hopper C (2012) Diversity Arrays Technology: A Generic Genome Profiling Technology on Open Platforms. In: Pompanon F, Bonin A (eds) Data Production and Analysis in Population Genomics. Methods in Molecular Biology (Methods and Protocols), vol 888. Humana Press, Totowa, NJ, pp 67-89

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: A review. Plant Methods 9 (1):29-37

Krawczak M (1999) Informativity assessment for biallelic single nucleotide polymorphisms. Electrophoresis 20 (8):1676-1681

Kumar K, Nagi M, Singh D, Kaur R, Gupta R (2013a) Heritability estimates, correlation and path analysis studies for nut and kernel characters of Pecan (*Carya illinoensis* [Wang] K. Koch). African Journal of Agricultural Research 8 (18):1915-1919

Kumar S, Bink MC, Volz RK, Bus VG, Chagné D (2012a) Towards genomic selection in apple (*Malus* × *domestica* Borkh.) breeding programmes: Prospects, challenges and strategies. Tree Genetics & Genomes 8:1-14

Kumar S, Chagne D, Bink MC, Volz RK, Whitworth C, Carlisle C (2012b) Genomic selection for fruit quality traits in apple (*Malus* x *domestica* Borkh.). PLOS One 7 (5):e36674

Kumar S, Garrick DJ, Bink MC, Whitworth C, Chagné D, Volz RK (2013b) Novel genomic approaches unravel genetic architecture of complex traits in apple. BMC Genomics 14 (1):393-406

Kumar S, Kirk C, Deng C, Wiedow C, Knaebel M, Brewer L (2017) Genotyping-by-sequencing of pear (*Pyrus* spp.) accessions unravels novel patterns of genetic diversity and selection footprints. Horticulture Research 4:Article 17015. doi:10.1038/hortres.2017.15

Kwong QB, Ong AL, Teh CK, Chew FT, Tammi M, Mayes S, Kulaveerasingam H, Yeoh SH, Harikrishna JA, Appleton DR (2017a) Genomic selection in commercial perennial crops: applicability and improvement in oil palm (*Elaeis guineensis* Jacq.). Scientific Reports 7:2872-2881

Kwong QB, Teh CK, Ong AL, Chew FT, Mayes S, Kulaveerasingam H, Tammi M, Yeoh SH, Appleton DR, Harikrishna JA (2017b) Evaluation of methods and marker systems in genomic selection of oil palm (*Elaeis guineensis* Jacq.). BMC Genetics 18 (1):107-115

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124 (3):743-756

Larsen B, Gardner K, Pedersen C, Ørgaard M, Migicovsky Z, Myles S, Toldam-Andersen TB (2018) Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. PLOS One 13 (8):e0201889. doi:10.1371/journal.pone.0201889

Lee SH, van der Werf JH, Hayes BJ, Goddard ME, Visscher PM (2008) Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLOS Genetics 4 (10):e1000231

Leverington RE (1962) Evaluation of macadamia nut varieties for processing. Queensland Journal of Agricultural Science 19:33-46

Li C (1975) Path analysis - a primer. Boxwood Press, California

Lin Z, Hayes BJ, Daetwyler HD (2014) Genomic selection in crops, trees and forages: A review. Crop and Pasture Science 65 (11):1177-1191

Luby JJ, Shaw DV (2000) Does marker-assisted selection make dollars and sense in a fruit breeding program? HortScience 36 (5):872-879

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits, vol 1. Sinauer Sunderland, MA

Macadamia Processing Co. Ltd. (2018) 2018 Notional Price Table for NIS at 10% Moisture Content Accessed 5/12/2018

Martínez-García PJ, Famula RA, Leslie C, McGranahan GH, Famula TR, Neale DB (2017) Predicting breeding values and genetic components using generalized linear mixed models for categorical and continuous traits in walnut (*Juglans regia*). Tree Genetics & Genomes 13 (5):109-120

McClure KA, Gardner KM, Douglas GM, Song J, Forney CF, DeLong J, Fan L, Du L, Toivonen P, Somers DJ (2018) A genome-wide association study of apple quality and scab resistance. The Plant Genome 11 (1):1-14

McConchie C, Meyers N, Anderson K, Vivian-Smith A, O'Brien S, Richards S Development and maturation of macadamia nuts in Australia. In: Stephenson R, Winks C (eds) Proceedings of the 3rd Australian Society of Horticultural Science and the first Australian Macadamia Society Research Conference, Gold Coast, Australia, 1996. pp 234-238

McConchie CA, Meyers N, Vithanage V, Turnbull C (1997) Pollen parent effects on nut quality and yield in macadamia. Australian Macadamia Society News Bulletin.

McFadyen L, Robertson D, Sedgley M, Kristiansen P, Olesen T (2012) Time of pruning affects fruit abscission, stem carbohydrates and yield of macadamia. Functional Plant Biology 39 (6):481-492

Mehlenbacher SA Progress and prospects in nut breeding. In: Janick J (ed) XXVI International Horticultural Congress: Genetics and Breeding of Tree Fruits and Nuts, Toronto, Canada, 2002. Acta Horticulturae, pp 57-79

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157 (4):1819-1829

Minamikawa M, Nonaka K, Kaminuma E, Kajiya-Kanegae H, Onogi A, Goto S, Yoshioka T, Imai A, Toyoda A, Fujiyama A, Hayashi T, Nakamura Y, Shimizu T, Iwata H Genomic selection in citrus breeding: Accuracy of genomic prediction. In: Plant and Animal Genome XXIV Conference, San Diego, 2016.

Minamikawa MF, Nonaka K, Kaminuma E, Kajiya-Kanegae H, Onogi A, Goto S, Yoshioka T, Imai A, Hamada H, Hayashi T (2017) Genome-wide association study and genomic prediction in citrus: Potential of genomics-assisted breeding for fruit quality traits. Scientific Reports 7:4721-4734

Minamikawa MF, Takada N, Terakami S, Saito T, Onogi A, Kajiya-Kanegae H, Hayashi T, Yamamoto T, Iwata H (2018) Genome-wide association study and genomic prediction using parental and breeding populations of Japanese pear (*Pyrus pyrifolia* Nakai). Scientific Reports 8 (1):11994

Mohammadi S, Prasanna B (2003) Analysis of genetic diversity in crop plants—salient statistical tools and considerations. Crop Science 43 (4):1235-1248

Moncur MW, Stephenson RA, Trochoulias T (1985) Floral development of *Macadamia integrifolia* Maiden & Betche under Australian conditions. Scientia Horticulturae 27 (1):87-96

Muranty H, Jorge V, Bastien C, Lepoittevin C, Bouffier L, Sanchez L (2014) Potential for marker-assisted selection for forest tree breeding: Lessons from 20 years of MAS in crops. Tree Genetics & Genomes 10 (6):1491-1510

Muranty H, Troggio M, Sadok IB, Al Rifaï M, Auwerkerken A, Banchi E, Velasco R, Stevanato P, Van De Weg WE, Di Guardo M (2015) Accuracy and responses of genomic selection on key traits in apple breeding. Horticulture Research 2:Article 15060

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell 21 (8):2194-2202

Nagao MA, Sakai WS (1990) Effects of gibberellic acid, ethephon or girdling on the production of racemes in *Macadamia integrifolia*. Scientia Horticulturae 43 (1):47-54

Namkoong G, Lewontin RC, Yanchuk AD (2005) Plant genetic resource management: the next investments in quantitative and qualitative genetics. Genetic Resources and Crop Evolution 51 (8):853-862

Nei M (1972) Genetic distance between populations. The American Naturalist 106 (949):283-292. doi:10.2307/2459777

Nejati-Javaremi A, Smith C, Gibson J (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science 75 (7):1738-1745

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12 (6):443-451

Nishio S, Hayashi T, Yamamoto T, Terakami S, Iwata H, Imai A, Takada N, Kato H, Saito T (2018a) Bayesian genome-wide association study of nut traits in Japanese chestnut. Molecular Breeding 38 (8):99-114

Nishio S, Terakami S, Matsumoto T, Yamamoto T, Takada N, Kato H, Katayose Y, Saito T (2018b) Identification of QTLs for agronomic traits in the Japanese chestnut (*Castanea crenata* Sieb. et Zucc.) breeding. The Horticulture Journal 87 (1):43-54. doi:10.2503/hortj.OKD-093

Nock CJ, Baten A, Barkla BJ, Furtado A, Henry RJ, King GJ (2016) Genome and transcriptome sequencing characterises the gene space of *Macadamia integrifolia* (Proteaceae). BMC Genomics 17 (1):937-948

Nock CJ, Elphinstone MS, Ablett G, Kawamata A, Hancock W, Hardner CM, King GJ (2014) Whole genome shotgun sequences for microsatellite discovery and application in

cultivated and wild *Macadamia* (Proteaceae). Applications in Plant Sciences 2 (4):1300089. doi:10.3732/apps.1300089

Nybom H, Weising K, Rotter B (2014) DNA fingerprinting in botany: Past, present, future. Investigative Genetics 5 (1):1-35

O'Connor K, Hardner C, Alam M, Hayes B, Topp B Variation in floral and growth traits in a macadamia breeding population. In: Drew R (ed) International Symposia on Tropical and Temperate Horticulture ISTTH2016, Cairns, Queensland, 2018a. vol 77. Acta Horticulturae, pp 623-630

O'Connor K, Hayes B, Hardner C, Alam M, Topp B (2019a) Selecting for nut characteristics in macadamia using a genome-wide association study. HortScience 54 (4):629-632. doi:https://doi.org/10.21273/HORTSCI13297-18

O'Connor K, Hayes B, Topp B (2018b) Prospects for increasing yield in macadamia using component traits and genomics. Tree Genetics & Genomes 14 (1):Article 7. doi:10.1007/s11295-017-1221-1

O'Connor K, Kilian A, Hayes B, Hardner C, Nock C, Baten A, Alam M, Topp B (2019b) Population structure, genetic diversity and linkage disequilibrium in a macadamia breeding population using SNP and silicoDArT markers. Tree Genetics & Genomes 15 (2):Article 24. doi:https://doi.org/10.1007/s11295-019-1331-z

O'Connor K, Powell M, Nock C, Shapcott A (2015) Crop to wild gene flow and genetic diversity in a vulnerable *Macadamia* (Proteaceae) species in New South Wales, Australia. Biological Conservation 191:504-511. doi:http://dx.doi.org/10.1016/j.biocon.2015.08.001

O'Hare P, Topp B (2010) Industry consultation helps guide macadamia breeding objectives. Australian Macadamia Society News Bulletin, vol 38.

O'Hare PJ, Stephenson RA, Quinlan K, Vock NT (2004) Macadamia Growers Handbook. Queensland Government, Queensland, Australia

Pandey MK, Upadhyaya HD, Rathore A, Vadez V, Sheshshayee M, Sriswathi M, Govil M, Kumar A, Gowda M, Sharma S (2014) Genomewide association studies for 50 agronomic traits in peanut using the 'Reference set'comprising 300 genotypes from 48 countries of the semi-arid tropics of the world. PLOS One 9 (8):e105228

Pavlopoulos GA, Soldatos TG, Barbosa-Silva A, Schneider R (2010) A reference guide for tree analysis and visualization. BioData Mining 3:1-24. doi:10.1186/1756-0381-3-1

Peace C (2005) Genetic characterisation of *Macadamia* with DNA markers. PhD Thesis, University of Queensland, Brisbane, Queensland

Peace C, Ming R, Schmidt A, Manners J, Vithanage V (2008) Genomics of *Macadamia*, a Recently Domesticated Tree Nut Crop. In: Moore P, Ming R (eds) Genomics of Tropical Crop Plants, vol 1. Plant Genetics and Genomics: Crops and Models. Springer New York, pp 313-332

Peace C, Vithanage V, Neal J (2004) A comparison of molecular markers for genetic analysis of macadamia. Journal of Horticultural Science and Biotechnology 79 (6):965-970

Peace CP (2017) DNA-informed breeding of rosaceous crops: Promises, progress and prospects. Horticulture Research 4:Article 17006

Peace CP, Allan P, Vithanage V, Turnbull CN, Carroll BJ (2005) Genetic relationships amongst macadamia varieties grown in South Africa as assessed by RAF markers. South African Journal of Plant and Soil 22 (2):71-75

Peace CP, Vithanage V, Turnbull CGN, Carroll BJ (2003) A genetic map of macadamia based on randomly amplified DNA fingerprinting (RAF) markers. Euphytica 134 (1):17-26. doi:10.1023/A:1026190529568

Peakall R, Smouse P (2012) GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research – an update. Bioinformatics 28 (19):2537-2539. doi:10.1093/bioinformatics/bts460

Perrier X, Flori A, Bonnot F (2003) Data Analysis Methods. In: Hamon P, Seguin M, Perrier X, Glaszmann JC (eds) Genetic Diversity of Cultivated Tropical Plants. Enfield, Science Publishers, Montpellier, France,

Piaskowski J, Hardner C, Cai L, Zhao Y, Iezzoni A, Peace C (2018) Genomic heritability estimates in sweet cherry reveal non-additive genetic variance is relevant for industry-prioritized traits. BMC Genetics 19 (1):23-38

Piepho H-P (1995) A simple procedure for yield component analysis. Euphytica 84 (1):43-48

Pisanu PC, Gross CL, Flood L (2009) Reproduction in wild populations of the threatened tree *Macadamia tetraphylla*: Interpopulation pollen enriches fecundity in a declining species. Biotropica 41 (3):391-398

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: Causes, consequences and solutions. Nature Reviews Genetics 6 (11):847-859

Porter G, Sherman W, Beckman T, Krewer G (2002) Fruit weight and shoot diameter relationship in early ripening peaches. Journal of the American Pomological Society 56 (1):30-33

Prichavudhi K, Yamamoto HY (1965) Effect of drying temperature on chemical composition and quality of macadamia nuts. Food Technology 19 (7):1153-1156

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155 (2):945-959

Pszczola M, Strabel T, Mulder H, Calus M (2012) Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of Dairy Science 95 (1):389-400

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81 (3):559-575. doi:http://dx.doi.org/10.1086/519795

Quarrie S, Rancic D, Radosevic R, Pekic Quarrie S, Kaminska A, Barnes J, Leverington M, Ceoloni C, Dodig D (2006) Dissecting a wheat QTL for yield present in a range of environments: From the QTL to candidate genes. Journal of Experimental Botany 57 (11):2627-2637. doi:10.1093/jxb/erl026

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Resende MF, Jr., Munoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MD, Kirst M (2012) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytologist 193 (3):617-624

Rikkerink EH, Oraguzie NC, Gardiner SE (2007) Prospects of association mapping in perennial horticultural crops. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (eds) Association mapping in plants. Springer, New York, pp 249-269

Roorkiwal M, Von Wettberg EJ, Upadhyaya HD, Warschefsky E, Rathore A, Varshney RK (2014) Exploring germplasm diversity to understand the domestication process in *Cicer* spp. using SNP and DArT markers. PLOS One 9 (7):e102016

Rosenberg NA (2004) DISTRUCT: A program for the graphical display of population structure. Molecular ecology resources 4 (1):137-138

Rosengarten FJ (2004) The Book of Edible Nuts. Dover Publications, Inc, New York

Ru S, Hardner C, Carter PA, Evans K, Main D, Peace C (2016) Modeling of genetic gain for single traits from marker-assisted seedling selection in clonally propagated crops. Horticulture Research 3:Article 16015

Ru S, Main D, Evans K, Peace C (2015) Current applications, challenges, and perspectives of marker-assisted seedling selection in Rosaceae tree fruit breeding. Tree Genetics & Genomes 11 (1):8-19

Russell D, De Faveri J, Hardner C, Bell D, Mulo S, Bignell G, Topp B Four new macadamia varieties for the Australian industry. In: International Macadamia Conference, Hilo, Hawaii, USA, 2017.

Sagawa C, Cristofani-Yaly M, Novelli V, Bastianel M, Machado M (2018) Assessing genetic diversity of Citrus by DArT_seq™ genotyping. Plant Biosystems 152 (4):593-598

Samonte SOP, Wilson LT, McClung AM (1998) Path analyses of yield and yield-related traits of fifteen diverse rice genotypes. Crop Science 38 (5):1130-1136

Sánchez-Pérez R, Ortega E, Duval H, Martínez-Gómez P, Dicenta F (2007) Inheritance and relationships of important agronomic traits in almond. Euphytica 155 (3):381-391

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

Saunders IW, Brohede J, Hannan GN (2007) Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. Genomics 90 (3):291-296

Savolainen O, Pyhäjärvi T (2007) Genomic diversity in forest trees. Current Opinion in Plant Biology 10 (2):162-167

Schmidt AL, Scott L, Lowe AJ (2006) Isolation and characterization of microsatellite loci from *Macadamia*. Molecular Ecology Notes 6 (4):1060-1063. doi:10.1111/j.1471-8286.2006.01434.x

Seavey SR, Bawa KS (1986) Late-acting self-incompatibility in angiosperms. The Botanical Review 52 (2):195-219

Sedgley M (1981) Early development of the Macadamia ovary. Australian Journal of Botany 29 (2):185-193

Sedgley M, Bell F, Bell D, Winks C, Pattison S, Hancock T (1990) Self-and cross-compatibility of macadamia cultivars. Journal of Horticultural Science 65 (2):205-213

Sedgley M, Blesing MA, Vithanage HIMV (1985) A developmental study of the structure and pollen receptivity of the macadamia pistil in relation to protandry and self-incompatibility. Botanical Gazette 146 (1):6-14

Semagn K, Bjørnstad Å, Xu Y (2010) The genetic dissection of quantitative traits in crops. Electronic Journal of Biotechnology 13 (5):16-17. doi:10.2225/vol13-issue5-fulltext-21

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52 (3/4):591-611

Sharma N, Singh SK, Mahato AK, Ravishankar H, Dubey AK, Singh NK (2019) Physiological and molecular basis of alternate bearing in perennial fruit crops. Scientia Horticulturae 243:214-225. doi:https://doi.org/10.1016/j.scienta.2018.08.021

Simmonds NW (1979) Principles of crop improvement. Longman Scientific & Technical, Essex, England

Sorkheh K, Shiran B, Khodambashi M, Moradi H, Gradziel T, Martinez-Gomez P (2010) Correlations between quantitative tree and fruit almond traits and their implications for breeding. Scientia Horticulturae 125 (3):323-331

Sparnaaij LD, Bos I (1993) Component analysis of complex characters in plant breeding: I. Proposed method for quantifying the relative contribution of individual components to variation of the complex character. Euphytica 70 (3):225-235

Stacklies W, Redestig H, Scholz M, Walther D, Selbig J (2007) pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics 23 (9):1164-1167

Steiger DL, Moore PH, Zee F, Liu Z, Ming R (2003) Genetic relationships of macadamia cultivars and species revealed by AFLP markers. Euphytica 132 (3):269-277. doi:10.1023/A:1025025522276

Stephens MJ, Scalzo J, Alspach PA, Beatson RA, Connor AM (2009) Genetic variation and covariation of yield and phytochemical traits in a red raspberry factorial study. Journal of the American Society for Horticultural Science 134 (4):445-452

Stephenson R, Cull B, Mayer D (1986) Effects of site, climate, cultivar, flushing, and soil and leaf nutrient status on yields of macadamia in south east Queensland. Scientia Horticulturae 30 (3):227-235

Stephenson RA, Gallagher EC, Rasmussen TS (1989) Effects of growth manipulation on carbohydrate reserves of macadamia trees. Scientia Horticulturae 40 (3):227-235

Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: Unlocking genetic potential from the wild. Science 277 (5329):1063-1066

Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818-822. doi:10.1126/science.1183700

Thomas R, Grafius J Prediction of heterosis levels from parental information. In: Proceedings of the 7th Congress of Eucarpia, Budapest, Hungary, 1976. pp 173-180

Thompson T, Baker J (1993) Heritability and phenotypic correlations of six pecan nut characteristics. Journal of the American Society for Horticultural Science 118 (3):415-418

Toft B (2019) Phenotypic and genotypic diversity in macadamia canopy architecture, flowering and yield. PhD Thesis, University of Queensland, Brisbane, Australia

Toft BD, Alam M, Topp B (2018) Estimating genetic parameters of architectural and reproductive traits in young macadamia cultivars. Tree Genetics & Genomes 14 (4):50-59

Toft BD, Alam MM, Wilkie J, Topp B (2019) Phenotypic association of multi-scale architectural traits with canopy volume and yield: moving towards high-density systems for macadamia. HortScience 54 (4):596-602

Topp B, Hardner C, Kelly A Strategies for breeding macadamias in Australia. In: Leitao JM (ed) XXVIII International Horticultural Congress on Science and Horticulture for People (IHC2010): International Symposium on New Developments in Plant Genetics and Breeding, Lisbon, Portugal, 2012. Acta Horticulturae, pp 47-53

Topp B, Hardner CM, Neal J, Kelly A, Russell D, McConchie C, O'Hare PJ Overview of the Australian macadamia industry breeding program. In: Onus N, Currie A (eds) XXIX International Horticultural Congress on Horticulture: Sustaining Lives, Livelihoods and Landscapes (IHC2014): International Symposium on Plant Breeding in Horticulture, Brisbane, Australia, 2016. Acta Horticulturae, pp 45-50

Topp B, Nock C, Hardner C, Alam M, O'Connor K (2019) Macadamia (*Macadamia* spp.) Breeding. In: Al-Khayri JM, Jain SM, Johnson DV (eds) Advances in Plant Breeding Strategies: Nut and Beverage Crops, vol 4. Springer International Publishing, Switzerland,

Trueman S, Richards S, McConchie C, Turnbull C (2000) Relationships between kernel oil content, fruit removal force and abscission in macadamia. Australian Journal of Experimental Agriculture 40 (6):859-866

Trueman SJ (2013) The reproductive biology of macadamia. Scientia Horticulturae 150:354-359

Trueman SJ, Turnbull CGN (1994) Effects of cross-pollination and flower removal on fruit set in *Macadamia*. Annals of Botany 73 (1):23-32

Urata U (1954) Pollination requirements of macadamia. Hawaii Agricultural Experiment Station Technical Bulletin 22:1-40

van Nocker S, Gardiner SE (2014) Breeding better cultivars, faster: Applications of new technologies for the rapid deployment of superior horticultural tree crops. Horticulture Research 1:Article 14022

Vanderzande S, Micheletti D, Troggio M, Davey MW, Keulemans J (2017) Genetic diversity, population structure, and linkage disequilibrium of elite and local apple accessions from Belgium using the IRSC array. Tree Genetics & Genomes 13 (6):125-140

VanRaden PM (2008) Efficient methods to compute genomic predictions. Journal of Dairy Science 91 (11):4414-4423

Varshney RK, Graner A, Sorrells ME (2005) Genomics-assisted breeding for crop improvement. Trends in Plant Science 10 (12):621-630

Viana AP, Resende MDVd, Riaz S, Walker MA (2016) Genome selection in fruit breeding: Application to table grapes. Scientia Agricola 73 (2):142-149

Vithanage V, Hardner C, Anderson K, Meyers N, McConchie C, Peace C Progress made with molecular markers for genetic improvement of macadamia. In: Drew R (ed) International Symposium on Biotechnology of Tropical and Subtropical Species Part 2, Brisbane, Australia, 1997. Acta Horticulturae, pp 199-208

Vithanage V, Winks C (1992) Isozymes as genetic markers for Macadamia. Scientia Horticulturae 49 (1):103-115

Walton DA, Wallace HM, Webb R (2012) Ultrastructure and anatomy of *Macadamia* (Proteaceae) kernels. Australian Journal of Botany 60 (4):291-300

Westwood MN (1993) Temperate-zone Pomology: Physiology and Culture. 3rd edn. Timber Press, Portland, Oregon

White I (2013) Pin function for asreml-R. http://www.homepages.ed.ac.uk/iwhite//asreml/

Wilkie J (2010) Interactions between the vegetative growth, flowering and yield of macadamia (*Macadamia integrifolia, M. integrifolia* × *M. tetraphylla*), in a canopy management context. PhD, University of New England, Armidale, Australia

Wong C, Bernardo R (2008) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor Appl Genet 116 (6):815-824

Wright S (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19 (3):395-420. doi:10.2307/2406450

Xie R, Li X, Chai M, Song L, Jia H, Wu D, Chen M, Chen K, Aranzana MJ, Gao Z (2010) Evaluation of the genetic diversity of Asian peach accessions using a selected set of SSR markers. Scientia Horticulturae 125 (4):622-629. doi:https://doi.org/10.1016/j.scienta.2010.05.015

Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: From publications to practice. Crop Science 48 (2):391-407

Xu Y, Lu Y, Xie C, Gao S, Wan J, Prasanna BM (2012) Whole-genome strategies for marker-assisted plant breeding. Molecular Breeding 29 (4):833-854

Yamamoto T, Terakami S (2016) Genomics of pear and other Rosaceae fruit trees. Breeding Science 66 (1):148-159

Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'Connell JR, Mangino M, Magi R, Madden PA, Heath AC, Nyholt DR, Martin NG, Montgomery GW, Frayling TM, Hirschhorn JN, McCarthy MI, Goddard ME, Visscher PM (2011) Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics 19 (7):807-812

Yao Q, Mehlenbacher S (2000) Heritability, variance components and correlation of morphological and phenological traits in hazelnut. Plant Breeding 119 (5):369-381

Zouros E, Foltz D (1987) The use of allelic isozyme variation for the study of heterosis. Isozymes 13:1