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

Katie O'Connor (k.oconnor@uq.edu.au)
University of Queensland https://orcid.org/0000-0003-3052-8300
Ben Hayes
University of Queensland
Craig Hardner
University of Queensland
Catherine Nock
Southern Cross University
Abdul Baten
AgResearch Ltd Grasslands Research Centre
Mobashwer Alam
University of Queensland
Robert Henry
University of Queensland
Bruce Topp
University of Queensland

Research article

Keywords: horticulture, plant breeding, progeny, genomics, marker-assisted selection, nut

Posted Date: December 3rd, 2019

DOI: https://doi.org/10.21203/rs.2.17991/v1

Version of Record: A version of this preprint was published on March 4th, 2020. See the published version at https://doi.org/10.1186/s12864-020-6575-3.

Abstract

Background: Breeding for new macadamia 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 (KW), 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. Results: 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 16 separate QTLs. Conclusions: These findings have important implications for macadamia breeding, and highlight the difficulties of heterozygous populations with rapid LD decay. By coupling validated marker-trait associations detected through GWAS with MAS, genetic gain could be increased by reducing the selection time for economically important nut characteristics. Genomic selection may be a more appropriate method to predict complex traits like tree size and yield.

Background

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 (1). Current production is dominated by Australia, South Africa and Hawaii, and is expanding in China, Kenya and other countries around the world (2). A major focus in breeding new macadamia varieties is increasing nut-in-shell yield per tree. However, the heritability of yield is low ($H^2 \approx 0.12$), largely influenced by environment, and, as such, difficult to select (3). To date, conventional phenotype- and pedigree-based selection has been employed to improve yield of commercial varieties. 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 genomic approaches to reduce selection cycles and increase genetic gain (4).

The use of genomics in plant breeding is expanding (4-6), 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 tested individually to detect evidence of marker-trait

associations (4). This method relies on linkage disequilibrium (LD) between markers and causal polymorphisms (4). 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 of putative large-effect QTLs (7, 8). QTLs identified through 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 (9, 10).

Several fruit and nut crops have employed GWAS to identify markers associated with key traits (11-18). Furthermore, by mapping significant markers to reference genomes, the location of markers can be determined in order to investigate candidate genes, 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 (19); 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, such as 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) (1, 20). 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 (1). The selection goal for KR, which is the proportion of kernel to nut-in-shell (KW/NW), may not be completely clear. Whilst high (>37%) KR attracts a premium price per kilogram (21), very thin shells can be prone to pest and disease damage (1). Whole kernels (WK) are those that have not split along the interface separating the two cotyledons during cracking (22); this trait can influence kernel price as some products and markets prefer whole kernels (1, 23).

Macadamia trees can produce about 2,500 pendant racemes 6–30 cm long, each with an inflorescence of 100–300 florets (24, 25). It has been estimated that less than 1% of florets produce viable nuts (26). 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 (27, 28). Reduced tree size is also an important selection trait to increase planting density and subsequent yield per hectare (29, 30). Trunk circumference (TC) or trunk cross-sectional area can be used as an estimate of tree size in macadamia (30).

O'Connor (31) 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 (32) on the same population of trees. O'Connor, Hayes (32) 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.

Results

Component traits

Raw (untransformed) phenotypes for KR, WK and TC were normally distributed (Figure 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 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 indicates that GWAS is not appropriate for yield, and association analysis was not performed for this trait.

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 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 the AL and HP sites, 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 1).

Figure 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

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 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 2). Narrow-sense genomic heritability varied across traits, from 0.08 for RSN to 0.74 for KR (Table 2). TC and NW were moderately heritable (0.45 and 0.53, respectively).

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 2; 33). GWAS identified seven SNP markers significantly (FDR < 0.05) associated with NW, four with WK, and 44 with TC (Figure 2; Table 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 for WK, the two mapped markers (mapped to different scaffolds) and another marker remained significant, but the unmapped SNP s2607 was redundant. The number of SNPs significantly associated with TC decreased to 16 after multiple regression analysis.

Figure 2 QQ plots showing expected significance levels against observed significance for yield component traits. Each circle represents one of 4,113 SNP markers. 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

Fifty-two of the 57 (91%) significant SNPs across the traits were mapped to scaffolds of the v2 macadamia genome assembly (Table 3). Some markers mapped to multiple scaffolds, for example, SNP3710 was located on 51 different scaffolds. Most scaffolds only had one SNP mapped, though six scaffolds had two SNPs mapped each. Almost 50% allele frequency was observed for two markers (s3540 for KW, and s3616 for TC; Table 3). The BLUEs estimated for the significant markers from the multiple regression model ranged from -10.359 to 4.608 for WK, and -11.946 to 4.088 for TC (Table 3).

Figure 3 Distribution of raw phenotypes across genotypic states for nut weight and percentage of whole kernels. Numbers above each box represent the number of trees with that genotype for that marker.

The phenotypic (raw, untransformed) distributions across the three genotypic states were examined with boxplots for the most significant marker for NW and WK (Figure 3). 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 (Figure 3). Similarly, the average values of WK for AA, GA and GG genotypes at marker s0201 were 78.0% (n = 5, SD = 11.0), 72.9% (n = 50, SD = 15.3), and 62.3% (n = 265, SD = 16.8) respectively (Figure 3). A two-way unbalanced analysis of variance (ANOVA) found that for NW at s2204 there was a signficiant difference between genotypes AA/AG ($p < 0.05$) and AA/GG ($p < 0.001$) but not for AG/GG, and for WK at s0201 a significant difference existed between genotypes AA/GG and AG/GG ($p < 0.001$), but not AA/AG.

Discussion

Phenotypic data in the breeding program

Large phenotypic diversity was observed for many of the traits in this study. Average 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%) (32). The moderate heritabilities suggest that selection for a number of traits will 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 achievable 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 (34). Results of this study differed to that of O'Connor, Hayes (32) 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 the average effect of the environment, 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.

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 (35). The number of markers in the current study is comparable with other studies in fruit trees (13, 15-17); however, the fragmented nature of the macadamia genome scaffolds means the distribution of markers across the whole genome is still unknown. Genetic linkage maps have been used to anchor scaffolds to chromosomes (Langdon et al. *in preparation*), and the location of scaffolds in the genome will be informative for determining locations of genes detected by SNPs in this study.

Population structure affects LD, and this needs to be accounted for in GWAS to avoid spurious associations and over-prediction of allelic effects. 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 (36). 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 (37). This genetic model may explain the pattern observed for TC, where a large number of associated markers was detected even at low p-values. The previous study (32) 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.

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 (38-40). However, Kelner, Costes (41) performed QTL mapping and 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 (19).

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 (27, 28).

For TC, 16 of the 44 significant markers were non-redundant, suggesting that there may be 14 QTLs controlling this trait. Multiple regression suggested that all of the the markers significantly associated

with NW may have detected the same or linked QTLs, with the most significant SNP (s2204) being the only non-redundant marker. The location of scaffolds in linkage groups (Nock et al. in preparation) may further aid the understanding of whether markers are in linkage disequilibrium or are separate QTLs.

A direct comparison cannot be made between SNPs found to be significantly associated with nut traits in O'Connor, Hayes (32) 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. For example, differences were found also across years in apple (18), and between QTL mapping and GWAS studies in chestnut (11, 42), and this may be a consequence of limited power in these studies.

Researchers use different thresholds for determining which markers to include in their genomics studies, such as 5% MAF (11, 17), 1% MAF within-populations (43), and ten copies of the minor allele across samples (18). 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 (33), 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 (44), as this will support the findings of this study.

Demonstration of marker-assisted selection

The results of this GWAS study can be used to demonstrate the implementation of MAS in the macadamia breeding program. SNPs significantly associated with commercially important traits would be ideal candidates for use in MAS. The estimates of BLUEs in the multiple regression analysis indicate the additive effect of the SNP allele at that marker on that trait. For example, the estimated effect for SNP allele at s2204 was 0.084, meaning that genotypes with one SNP allele will have an added 0.084 g of nut weight controlled by genetic variance than those without. The influence of additive genetic variance of these alleles was quite different to that which was observed in the raw phenotype, as the phenotype will have been influenced by non-additive genetic effects and environment. The three genotypic states for NW at SNP s2204 and for WK at marker s0201 showed clear association with phenotypic averages, though the difference in genotypic states was much lower than the additive allele effect calculated from BLUEs. The sample sizes among the three different genotypic states varied 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 first

leaf at these key markers. Determining the allelic states at these markers would allow selection of AG heterozygotes at SNP s2204 for seedlings with predicted intermediate nuts, and AA genotype at SNP s0201 for a high percentage of whole kernels. However, with such low MAF and number of individuals in these genotypic states, these results should be interpreted with caution. Again, the SNP should be validated in an independent population, and the effect of the SNP alleles should be estimated in that population.

Further work

This study and our previous work (32) 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. However, the results presented require validation before being employed in breeding programs. Multi-trait analyses could be performed to increase the power of detection of QTLs, and also detect pleiotropy (45). 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 the proportion of variance explained by those markers.

When a more complete reference genome is assembled, the location of these markers can be determined, 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 (35), 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, Kilian (35), could be alleviated with the use of a complete reference genome in sequencing of SNPs in the future. Without genome scaffold annotation, the significant SNPs cannot be linked to known genes or proteins, which has been achieved in other studies of GWAS in fruit trees (e.g. 13, 15, 16, 18). The v2 scaffolds and chromosomes are being (Nock et al. in preparation), and so candidate genes could be identified in future studies.

Although there was a lack of significant associations in some traits in the current study, these should still be investigated in future work. 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 (46), rather than one-by-one as in GWAS.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 these traits.

Conclusions

The findings of this study have important implications for macadamia breeding, but also highlight 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. The traits with significant SNPs identified are likely to be controlled by fewer genes than the other traits. Multiple regression 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 reducing 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. 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.

Methods

Methods for association analysis are similar to those by O'Connor, Hayes (32), and are replicated here for completeness, with differences between the two studies outlined.

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, with East Gympie (EG) and Amamoor (AM) in the Gympie region, and Alloway (AL) and Hinkler Park (HP) in the Bundaberg region. Clones of five of the parents were measured at all four sites. 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, Kilian (35). 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 (47). 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.

Phenotyping for yield and component traits

Phenotypic data used in this study were collected across two seasons from August 2016 to July 2018, with all traits except RSN and yield measured only in one season. 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 (48). 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 from the branch apex 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. Component trait means were calculated for each tree for analysis where at least six observed units per tree were evaluated. For example, trees with five or fewer nuts measured were 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 (33). 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, Hayes (32) to investigate the consistency of phenotypes between the two studies.

Association analysis

A genomic relationship matrix (GRM) was constructed following methods of VanRaden (49). Preliminary analysis was performed using ASReml (50) in R to determine the most parsimonious model for each trait:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{b} + \mathbf{Z}_g\mathbf{g} + \mathbf{Z}_{gs}\mathbf{gs} + \mathbf{e} \quad \text{Eq. 1}$$

where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{1}$ is a vector of ones, μ is a fixed intercept, $\mathbf{X}$ is a design matrix allocating fixed effects (site, block within site, tree type = grafted parent or seedling progeny) to observations, $\mathbf{b}$ is a vector of these unknown 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,**G**), 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); 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,$\mathbf{G}\otimes\mathbf{I_4}\otimes$) where **I** is a 4x4 identity matrix for the four sites, and **e** is a vector of random errors ~ N(0,) where is the error variance. This model is additive, in that two copies of one 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 with the least number of parameters that fit the data as well as more complex models: 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 = / (+)) 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, Hayes (32) using ASReml (50) in R, using a mixed model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wm} + \mathbf{Z_g g} + \mathbf{Z_{gs} gs} + \mathbf{e} \qquad \text{Eq. 2}$$

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 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 (33). 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 (51) 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. FDR was again calculated for the markers included in the multiple regression. 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. An estimation of the additive allele effect of each significant SNP was estimated from fixed effects (best linear unbiased estimators; BLUEs) from the multiple regression model.

Marker locations

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, Kilian (35). Locations of previously identified markers associated with nut traits were also estimated on the scaffolds, using marker sequences from O'Connor, Hayes (32).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The data that support the findings of this study are available from The University of Queensland's Institutional Data Access/Ethics Committee, but restrictions apply to the availability of these data. Data are however available from the authors upon reasonable request and with permission of The University of Queensland for researchers who meet the criteria for access to confidential data. Contact data@library.uq.edu.au.

Competing interests

The authors declare that they have no competing interests

Funding

This research has been funded by Hort Innovation Australia, 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. KO acknowledges the Australian Postgraduate Award and Charles Morphett Peglar scholarship for financial support.

Authors' contributions

KO collected phenotypic data with help from field assistants, wrote the paper, performed all analyses and made final edits. BH, CH and MA assisted in interpretation of results. CN and AB provided genome assembly scaffold data in the chapter. BH, BT, CH, MA and CN suggested revisions. All authors read and approved the final manuscript.

Acknowledgements

KO acknowledges the many people who assisted with field work, including Rachel Abel, Jasmine Nunn and Codie Murphy.

References

1. Hardner CM, Peace C, Lowe AJ, Neal J, Pisanu P, Powell M, et al. Genetic resources and domestication of macadamia. Horticultural Reviews. 2009;35:1-126.
2. Australian Macadamia Society, editor Estimated World Macadamia Production. XXXVII International Nut and Dried Fruit Congress; 2018; Spain.
3. Hardner CM, Winks CW, Stephenson RA, Gallagher EG, McConchie CA. Genetic parameters for yield in macadamia. Euphytica. 2002;125(2):255-64.
4. Khan MA, Korban SS. Association mapping in forest trees and fruit crops. Journal of Experimental Botany. 2012;63(11):4045-60.
5. Iwata H, Minamikawa MF, Kajiya-Kanegae H, Ishimori M, Hayashi T. Genomics-assisted breeding in fruit trees. Breeding Science. 2016;66(1):100-15.
6. Grattapaglia D, Resende MD. Genomic selection in forest tree breeding. Tree Genetics & Genomes. 2011;7(2):241-55.
7. Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, et al. Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell. 2009;21(8):2194-202.
8. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research. 2009;91:47-60.
9. Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica. 2005;142(1-2):169-96.
10. Tester M, Langridge P. Breeding technologies to increase crop production in a changing world. Science. 2010;327:818-22.
11. Nishio S, Hayashi T, Yamamoto T, Terakami S, Iwata H, Imai A, et al. Bayesian genome-wide association study of nut traits in Japanese chestnut. Molecular Breeding. 2018;38(8):99-114.
12. Iwata H, Hayashi T, Terakami S, Takada N, Sawamura Y, Yamamoto T. Potential assessment of genome-wide association study and genomic selection in Japanese pear *Pyrus pyrifolia*. Breeding Science. 2013;63(1):125-40.

13. Kumar S, Garrick DJ, Bink MC, Whitworth C, Chagné D, Volz RK. Novel genomic approaches unravel genetic architecture of complex traits in apple. BMC Genomics. 2013;14(1):393-406.
14. Cao K, Wang L, Zhu G, Fang W, Chen C, Luo J. Genetic diversity, linkage disequilibrium, and association mapping analyses of peach (*Prunus persica*) landraces in China. Tree Genetics & Genomes. 2012;8(5):975-90.
15. Minamikawa MF, Nonaka K, Kaminuma E, Kajiya-Kanegae H, Onogi A, Goto S, et al. Genome-wide association study and genomic prediction in citrus: Potential of genomics-assisted breeding for fruit quality traits. Scientific Reports. 2017;7:4721-34.
16. Minamikawa MF, Takada N, Terakami S, Saito T, Onogi A, Kajiya-Kanegae H, et al. Genome-wide association study and genomic prediction using parental and breeding populations of Japanese pear (*Pyrus pyrifolia* Nakai). Scientific Reports. 2018;8(1):11994.
17. Imai A, Nonaka K, Kuniga T, Yoshioka T, Hayashi T. 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. 2018;14(2):24-38.
18. McClure KA, Gardner KM, Douglas GM, Song J, Forney CF, DeLong J, et al. A genome-wide association study of apple quality and scab resistance. The Plant Genome. 2018;11(1):1-14.
19. O'Connor K, Hayes B, Topp B. Prospects for increasing yield in macadamia using component traits and genomics. Tree Genetics & Genomes. 2018;14(1):Article 7.
20. Strohschen B. Contributions to the biology of useful plants. 4. Anatomical studies of fruit development and fruit classification of the Macadamia nut (*Macadamia integrifolia* Maiden and Betche). Journal of Applied Botany and Food Quality. 1986;60:239-47.
21. Macadamia Processing Co. Ltd. 2018 Notional Price Table for NIS at 10% Moisture Content 2018 [
22. Walton DA, Wallace HM, Webb R. Ultrastructure and anatomy of *Macadamia* (Proteaceae) kernels. Australian Journal of Botany. 2012;60(4):291-300.
23. O'Hare PJ, Stephenson RA, Quinlan K, Vock NT. Macadamia Growers Handbook. In: Fisheries DoPIa, editor. Queensland, Australia: Queensland Government; 2004.
24. Huett DO. Macadamia physiology review: a canopy light response study and literature review. Australian Journal of Agricultural Research. 2004;55(6):609.
25. Trueman SJ. The reproductive biology of macadamia. Scientia Horticulturae. 2013;150:354-9.
26. Ito PJ. Effect of style removal on fruit set in macadamia. HortScience. 1980;15(4):520-1.
27. Toft B. Phenotypic and genotypic diversity in macadamia canopy architecture, flowering and yield [PhD Thesis]. Brisbane, Australia: University of Queensland; 2019.
28. Wilkie J. Interactions between the vegetative growth, flowering and yield of macadamia (*Macadamia integrifolia, M. integrifolia* × *M. tetraphylla*), in a canopy management context [PhD]. Armidale, Australia: University of New England; 2010.
29. Topp B, Hardner CM, Neal J, Kelly A, Russell D, McConchie C, et al., editors. Overview of the Australian macadamia industry breeding program. XXIX International Horticultural Congress on Horticulture:

Sustaining Lives, Livelihoods and Landscapes (IHC2014): International Symposium on Plant Breeding in Horticulture; 2016; Brisbane, Australia: Acta Horticulturae.

30. Toft BD, Alam M, Topp B. Estimating genetic parameters of architectural and reproductive traits in young macadamia cultivars. Tree Genetics & Genomes. 2018;14(4):50-9.
31. O'Connor K. Selection strategies to improve yield in macadamia using component traits and genomics. Brisbane: University of Queensland; 2019.
32. O'Connor K, Hayes B, Hardner C, Alam M, Topp B. Selecting for nut characteristics in macadamia using a genome-wide association study. HortScience. 2019;54(4):629-32.
33. Gondro C, Lee SH, Lee HK, Porto-Neto LR. Quality Control for Genome-Wide Association Studies. In: Gondro C, van der Werf J, Hayes B, editors. Genome-Wide Association Studies and Genomic Prediction. London: Springer Science; 2013.
34. Falconer DS. Introduction to Quantitative Genetics. Essex, UK: Longman Scientific & Technical; 1989.
35. O'Connor K, Kilian A, Hayes B, Hardner C, Nock C, Baten A, et al. Population structure, genetic diversity and linkage disequilibrium in a macadamia breeding population using SNP and silicoDArT markers. Tree Genetics & Genomes. 2019;15(2):Article 24.
36. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: A review. Plant Methods. 2013;9(1):29-37.
37. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics. 2011;19(7):807-12.
38. Hayes B, Goddard M. Genome-wide association and genomic selection in animal breeding. Genome. 2010;53(11):876-83.
39. Luby JJ, Shaw DV. Does marker-assisted selection make dollars and sense in a fruit breeding program? HortScience. 2000;36(5):872-9.
40. Huang X, Han B. Natural variations and genome-wide association studies in crop plants. Annual review of plant biology. 2014;65:531-51.
41. Kelner J-J, Costes E, Guitton B, Chagné D, Gardiner SE, Velasco R. Genetic control of biennial bearing in apple. Journal of Experimental Botany. 2011;63(1):131-49.
42. Nishio S, Terakami S, Matsumoto T, Yamamoto T, Takada N, Kato H, et al. Identification of QTLs for agronomic traits in the Japanese chestnut (*Castanea crenata* Sieb. et Zucc.) breeding. The Horticulture Journal. 2018;87(1):43-54.
43. Biscarini F, Nazzicari N, Bink M, Arús P, Aranzana MJ, Verde I, et al. Genome-enabled predictions for fruit weight and quality from repeated records in European peach progenies. BMC Genomics. 2017;18(1):432-46.
44. Hayes B. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). In: Gondro C, Van der Werf J, Hayes B, editors. Genome-Wide Association Studies and Genomic Prediction. London: Springer Science; 2013.

45. Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, et al. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. PLOS Genetics. 2014;10(3):e1004198.
46. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819-29.
47. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinformatics. 2007;23(9):1164-7.
48. Prichavudhi K, Yamamoto HY. Effect of drying temperature on chemical composition and quality of macadamia nuts. Food Technology. 1965;19(7):1153-6.
49. VanRaden PM. Efficient methods to compute genomic predictions. Journal of Dairy Science. 2008;91(11):4414-23.
50. Gilmour AR, Gogel B, Cullis B, Thompson R, Butler D. ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK2009.
51. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300.

Tables

Table 1 Summary of raw (untransformed) phenotypes for each trait analysed in GWAS.

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	-

SD, standard deviation; r_p, Pearson's correlation of current data with raw phenotypes for young trees from O'Connor, Hayes (32)

Table 2 Significance values of fixed and random terms included in association analysis model for each trait

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

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. † indicates data were transformed

Table 3 Summary of significant SNPs associated with yield component traits identified in GWAS.

Trait	SNP	Scaffold[a]	Position (bp)	Alleles	MAF	p	pMR	BLUE
NW†	s2204	scaffold926|size239084	212,122	A/G	0.027	3.68E-06	4.46e-06	0.084
	s4163	scaffold285|size451335	314,657	C/T	0.027	8.03E-06	NS	
	s1434	scaffold_177|size983250	804,678	T/C	0.019	2.65E-05	NS	
	s1643	scaffold44|size832018	129,241	A/C	0.021	3.46E-05	NS	
	s1121	scaffold653|size305054	6,573	A/G	0.021	3.82E-05	NS	
	s5182	-	-	A/T	0.035	6.29E-05	NS	
	s2256	scaffold710|size289053	142,496	G/T	0.026	6.45E-05	NS	
KW†	s3540[b]	∫	∫	G/A	0.482	1.34E-05		
KR	s1707[b]	scaffold_72|size1196525	587,142	C/T	0.061	2.37E-05		
WK	s0201	scaffold213|size509421	186,179	G/A	0.093	8.81E-06	1.11E-06	4.608
	s3239	scaffold361|size1112638	1,087,419	G/C	0.037	3.39E-05	2.45E-04	-10.359
	s1917	-	-	A/G	0.163	1.23E-05	NS	
	s2607	-	-	T/C	0.177	2.91E-05	NS	
TC	s3169	scaffold146|size572432	176,797	T/C	0.230	1.29E-07	1.13E-07	-1.343
	s1885	∫	∫	C/T	0.319	8.57E-05	4.85E-05	-1.706
	s2320	scaffold81|size707423	173,614	C/A	0.083	1.02E-04	3.90E-05	4.088
	s3332	scaffold1221|size537814	497,497	T/C	0.285	1.97E-06	3.98E-04	2.167
	s1208	∫	∫	C/T	0.179	3.14E-04	6.96E-04	-2.383
	s3291	∫	∫	G/T	0.267	4.09E-05	7.52E-04	0.540
	s4709	∫	∫	G/A	0.106	4.74E-04	2.62E-03	-11.946
	s3311	-	-	A/C	0.043	3.90E-04	3.81E-03	-4.442
	s3828	∫	∫	G/A	0.093	4.03E-04	4.47E-03	-2.009
	s2230	scaffold_88	424,720	G/T	0.884	2.03E-04	6.15E-03	-2.360

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; BLUE, best linear unbiased predictor (fixed effect) of SNP, additive effect of

allele on the trait; NS, not significant. - indicates marker was not mapped to scaffolds. ∫ indicates marker was mapped to multiple scaffolds. [a] Scaffold in v2 genome assembly. [b] Did not pass FDR = 0.05 threshold. † indicates data were transformed

Figures

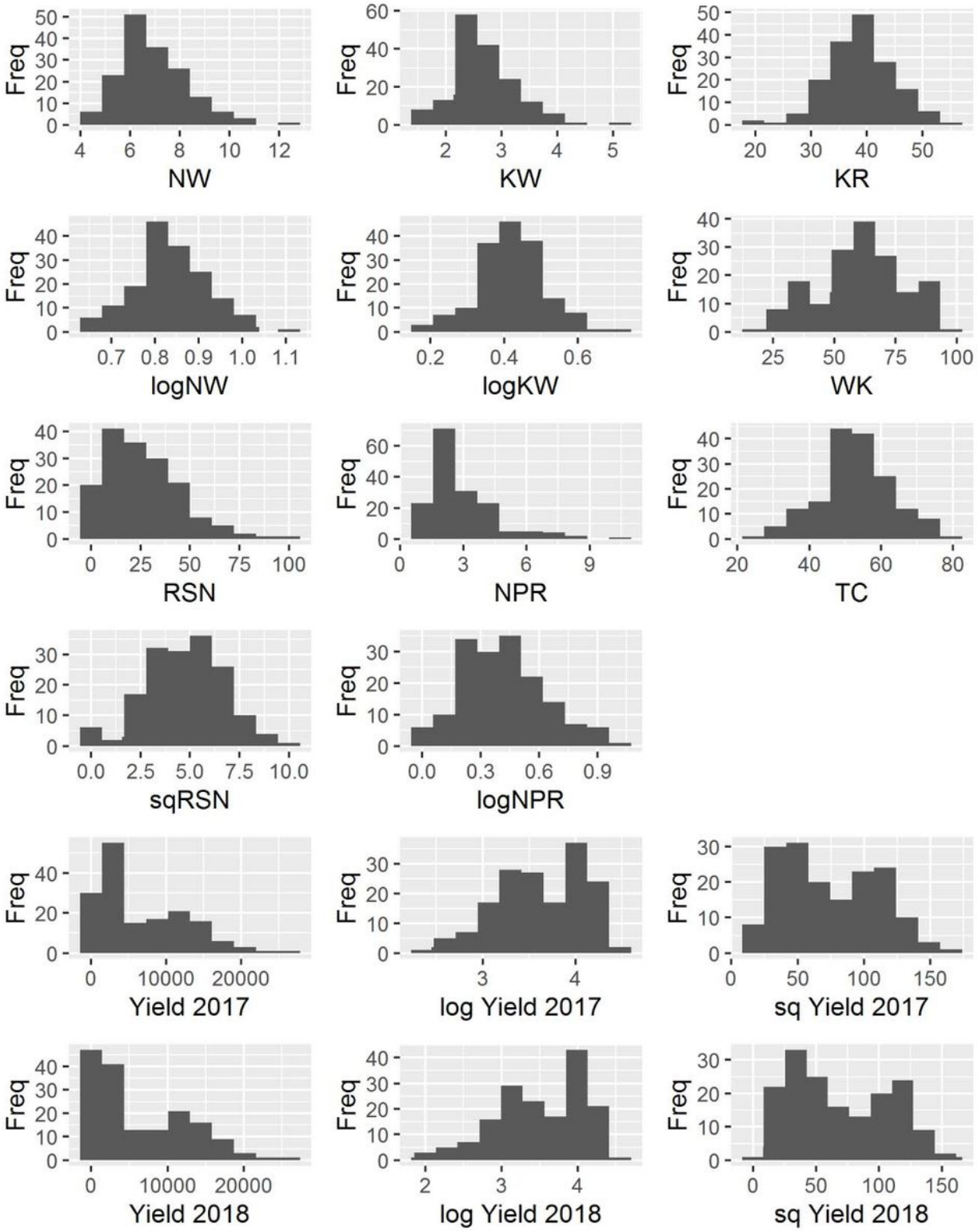

Figure 1

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 (log10(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

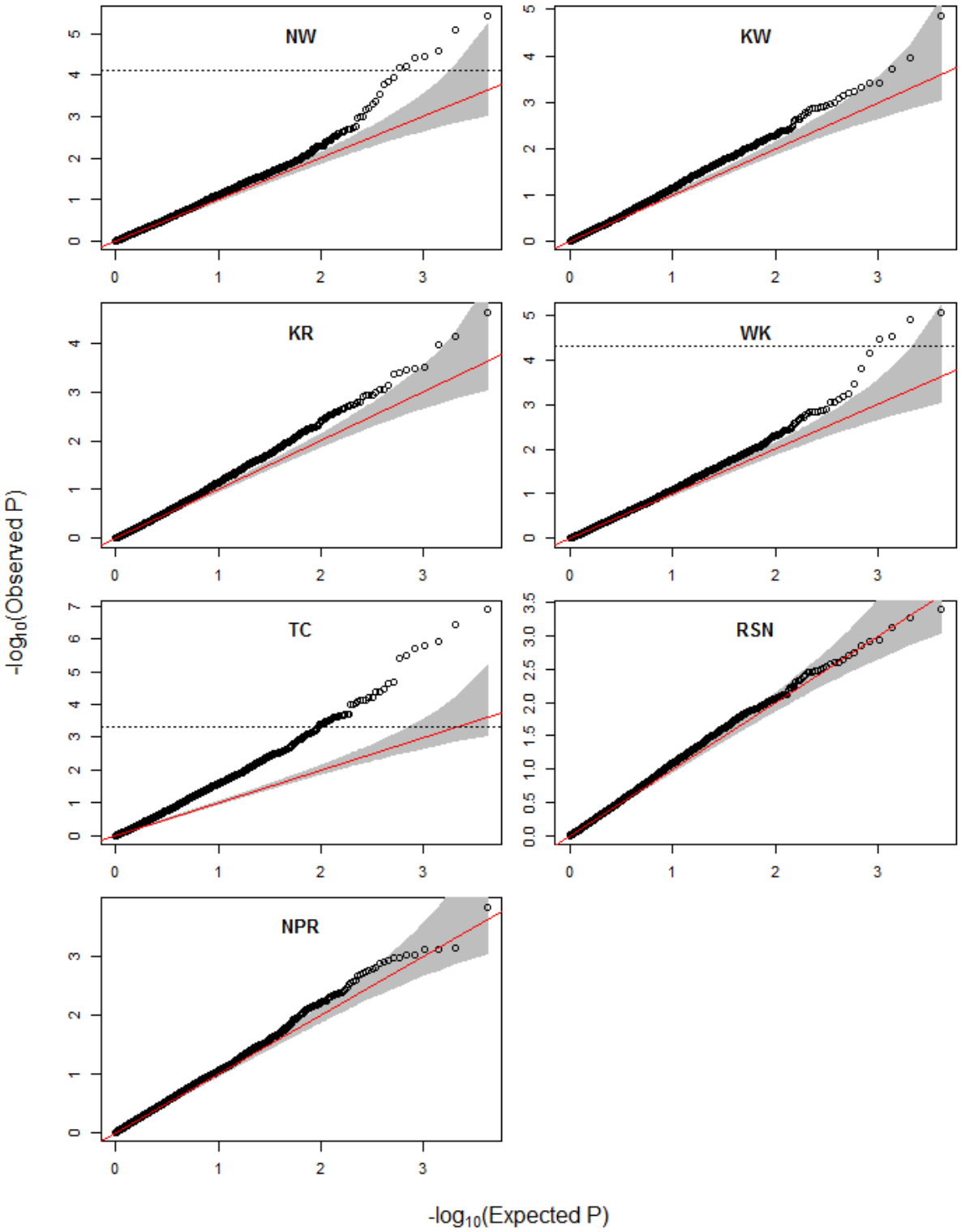

Figure 2

QQ plots showing expected significance levels against observed significance for yield component traits. Each circle represents one of 4,113 SNP markers. 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

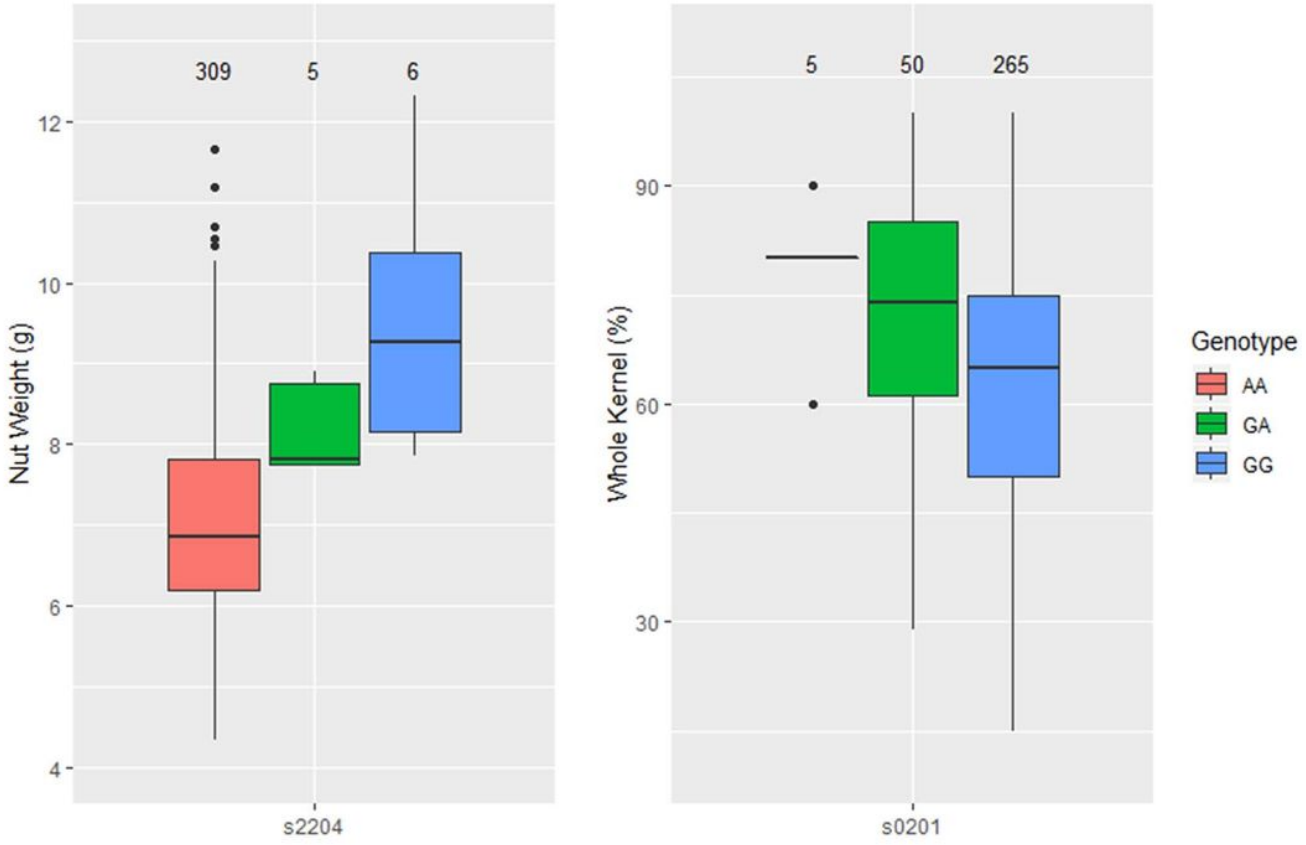

Figure 3

Distribution of raw phenotypes across genotypic states for nut weight and percentage of whole kernels. Numbers above each box represent the number of trees with that genotype for that marker.