Genetic Analysis of Walnut (Juglans regia L.)
Pellicle Pigment Variation Through a Novel, High-Throughput Phenotyping Platform

Gina M. Sideli,*1 Peter McAtee,† Annarita Marrano,* Brian J. Allen,* Patrick J. Brown,*
Timothy S. Butterfield,‡ Abhaya M. Dandekar,* Charles A. Leslie,* and David B. Neale*†‡
*Department of Plant Sciences, University of California, Davis, CA 95616, USA, †Plant and Food Research, Auckland, 1142, New Zealand, and ‡U.S. Department of Agriculture, ARS Crops Pathology and Genetics Unit, Davis CA 95616
ORCID IDs: 0000-0003-1056-1496 (G.M.S.); 0000-0003-1569-6680 (P.M.); 0000-0002-0534-545X (B.J.A.); 0000-0003-1332-711X (P.J.B.); 0000-0001-6829-8518 (T.S.B.); 0000-0001-7925-4086 (A.M.D.)

ABSTRACT Walnut pellicle color is a key quality attribute that drives consumer preference and walnut sales. For the first time a high-throughput, computer vision-based phenotyping platform using a custom algorithm to quantitatively score each walnut pellicle in L* a* b* color space was deployed at large-scale. This was compared to traditional qualitative scoring by eye and was used to dissect the genetics of pellicle pigmentation. Progeny from both a bi-parental population of 168 trees (‘Chandler’ × ‘Idaho’) and a genome-wide association (GWAS) with 528 trees of the UC Davis Walnut Improvement Program were analyzed. Color phenotypes were found to have overlapping regions in the ‘Chandler’ genetic map on Chr01 suggesting complex genetic control. In the GWAS population, multiple, small effect QTL across Chr01, Chr07, Chr08, Chr09, Chr10, Chr12 and Chr13 were discovered. Marker trait associations were co-localized with QTL mapping on Chr01, Chr10, Chr14, and Chr16. Putative candidate genes controlling walnut pellicle pigmentation were postulated.

English walnut (Juglans regia L.) is an economically important edible seed that is diploid 2x = 32 with 16 chromosomes and a genome size of approximately 600 Mbp. In the U.S., the English walnut industry generates 1.5 billion dollars annually with over 99% of production occurring in California (CA Dept Food Ag 2018). Interestingly, consumer preference is for a lighter color seed coat, or pellicle, thereby commanding higher prices in the U.S market. The seed coat, pellicle and/or testa covers the zygotic embryo (seed) and is formed from maternally derived tissue arising from the inner and outer ovular integuments (Mohamed-Yasseen et al. 1994). Each seed coat layer contributes to various metabolic functions which include: producing, transporting, and accumulating metabolites; photosynthetic assimilates for zygote development; defense; and physical structure (Smykal et al. 2014).

Walnut (Juglans regia L.) pellicle pigmentation can vary significantly between cultivars and developmental stages, and is influenced by environmental stressors. Flavonoids or phenolic compounds synthesized through the phenylpropanoid pathway are localized in seed coats to defend the embryo against biotic stressors (Dixon et al. 2013), and generally contain yellow pigments. Currently, pellicle color is scored qualitatively on an ordinal scale of 1 – 4 (extra light, light, light amber, amber), and subjectively by eye using a color chart published by the California Dried Fruit Association (DFA) (Figure 1A). Color perception by the human eye is complex and has been researched since at least 1878 when Hering, a physiologist proposed the opponent color theory as function of both psychological and neurological processes and Hunter in 1979 explained that receptors in human eye perceive color in pairs (Lee 2008). In particular, L* is the scale of light to dark, a* is the scale of green to red, and b* the scale of blue to yellow (Hunter and Harold 1987) (Figure 1B). The Lab color scale is arguably one of the most perpetually uniform color spaces and best approximates color difference as perceived by the human eye.
Human perception of color is largely influenced by the number and types of photoreceptors an individual possesses. Differences in the photoreceptors between individuals contribute to variability in the perception of a given color. This forms the basis of subjectivity of human quality assessments within a color perception dataset. As such, color(s) quantified by human perception can be an unreliable source for obtaining quality data (Brosnan and Sun 2004), however since the trait is influenced by consumer acceptance, human perception cannot be ignored completely. Computer vision systems (CVS) have been proposed as a large-scale replacement to quality assessments by human users. A CVS is any platform that acquires and processes images with computer software by using an algorithm for calibration and color extraction. The CVS interprets color in Hunter’s L* a* b* space (Figure 2) by using a method adopted by Leon 2006 where pixels in RGB from an image are fit in a linear model and transformed to Lab scores. This is vastly more high-throughput and consistent than a human user qualitative score. Computer vision systems have been used in a variety of crops for quality grading in carrot (Deng et al. 2017), detection of mechanical damage in blueberry (Wang et al. 2018), table grape rachis browning (Bahar et al. 2017), and sweet lemon mechanical damage destruction (Firouzjai et al. 2018).

Using traditional scoring, walnut pellicle color is labor intensive to measure, strongly influenced by environmental variation, and is a polygenic trait (Marrano et al. 2019). Combining quantitative phenotypic measurements of a CVS with genotyping data will facilitate identification of genetic loci that explain trait variation. Causal single nucleotide polymorphisms (SNPs) can be validated with marker-assisted breeding or genomic prediction models. This makes it possible to identify desirable parents and select preferred progeny at the young seedling stage rather than at sexual maturity 4-5 years post-germination. Substantially reducing the seedling-selection
time will hasten the development of new cultivars, decrease costs in allocation of resources, and accelerate the genetic gain.

The first objective of this study was to measure walnut pellicle color quantitatively using a CVS developed by (Donis-González et al. 2020) that discriminates color in the \( L^* \ a^* \ b^* \) channels. The second objective was to use quantitative-trait loci (QTL) mapping and genome-wide association study (GWAS) to detect marker trait associations combined with SNP genotyping.

**MATERIALS AND METHODS**

**Walnut source and experimental design**

The experiment was performed with *Juglans regia* seedling trees ranging from age 4 – 12 which were derived from multiple crosses within the UC Davis Walnut Improvement Program.

The genome wide association design consisted of seedling trees from 31 full-sib families with an average of 16 per family. These trees ranged in age from 4 – 9 years, and included 20 founder trees, and one cultivar ‘Robert Livermore’ released by UC Davis for a total of 528 trees. A bi-parental F1 population ‘Chandler’ × ‘Idaho’ (\( n = 168 \)) was used for QTL mapping with a range in tree age of 8 – 12. Trees in this study were planted across eight blocks; five of them were across the same field location and all trees were spaced six feet apart and irrigated with micro-sprinklers. Blocks consisted of trees planted in row plots with multiple family crosses made in one year. A full-sib family had between 10 and 50 individuals with 31 unique parents. Most male parents were only used once, with a few male parents used 2-4 times, one male was used for 13 different crosses, and two males were used reciprocally as females.

At least 30 nuts per tree were hand harvested as each tree reached fruit maturity (hull-split) during August and September of 2015, and two males were used reciprocally as females. Most male parents were only used once, with a few male parents used 2-4 times, one male was used for 13 different crosses, and two males were used reciprocally as females.

At least 30 nuts per tree were hand harvested as each tree reached fruit maturity (hull-split) during August and September of 2015, 2016, and 2017. These were hung in polypropylene mesh bags, air-dried for two weeks, and then stored at -4°C until measurements were taken. Walnut kernels were placed on commercial scoring trays obtained from the Dried Fruit Association of California which contain 100 wells each. Each tray held ten kernels from each of 10 distinct trees, and 10 kernels were evaluated. When kernels were being evaluated for color, they were kept at 0°C for short-term storage, and kernels were kept at - 4°C for long-term storage.

**Measuring pellicle color**

Qualitative color scoring by eye was done by examining each kernel for its separation into color category 1 - 4, according to the Dried Fruit Association (DFA) scoring sheet (Figure 1A), (Safe Food Alliance, Sacramento, CA, (DFA of California 2017), in accordance with the USDA Standards for Grades of Shelled Walnuts as designated by the California Walnut Board under Title 7 of the United States Code (U.S.C.) (7 C.F.R. § 51.2277-2296).

**The DFA color scores are**

ELIT = Extra light, 1;
LITE = light, 2;
LAMB = light amber, 3;
AMBR = Amber, 4

**Analysis of phenotypic variation**

Multi-year phenotypic data collected for DFA scores and the CVS scores were analyzed with ANOVA using R packages ‘car’, ‘agricolae’, ‘lmef’ with equation (1) for the model,

\[
Y_{ijkl} = \mu + f_i + g_j + y_k + (fj)_{ik} + a_{ijk} + e_{ijkl}
\]

where \( Y \) is the color measurement taken on the \( i \)-th genotype, with \( j \) number of measurements in the \( k \)-th year with \( l \) number of families; \( \mu \) is a constant value (overall mean), \( f_i \) is the effect on the color measurement of the \( i \)-th family, \( g_j \) is the effect on color measurement of the \( i \)-th genotype, \( y_k \) is the effect on color measurement of the \( k \)-th year, \((fj)_{ik}\) is the interaction between family and year, \( a_{ijk} \) is the effect...
of age on color measurement for the i-th genotype, j number of measurements in the k-th year, and \( e_{ijk} \) is the random error term with mean 0 and variance \( \sigma^2_e \).

For the ‘Chandler’ × ‘Idaho’ population, the adjusted means were calculated in R package ‘lsmeans’ to account for the year differences and one value per individual tree was used in the QTL analysis with the equation (2) for linear model,

\[
Y_{ijk} = \mu + \alpha_i + \beta_k + (\alpha\beta)_{ik} + e_{ijk}
\]

where \( y \) is the color measurement taken on the i-th genotype, j number of measurements in the k-th year, \( \alpha \) is the effect on color measurement of the i-th genotype, \( \beta_k \) is the effect on color measurement of the k-th year, \( (\alpha\beta)_{ik} \) is the interaction between genotype and year, and \( e_{ijk} \) is the random error term with mean 0 and variance \( \sigma^2_e \).

For the larger dataset of the 31 full-sib families, breeding values (BV) per individuals were estimated using ASReml-R v4 (Butler et al. 2017) in order to remove environmental effects of years and blocks on the phenotypes. In particular, the mixed model for equation (3),

\[
y = X\tau + Zu + e
\]

where \( y \) denotes the \((n \times 1)\) vector of observations, \( \tau \) is a \((p \times 1)\) vector of fixed treatment effects (block and year), \( X \) is an \((n \times p)\) design matrix, \( u \) is the \((q \times 1)\) vector of random effects (pedigree), \( Z \) is an \((n \times q)\) design matrix, \( e \) is the \((n \times 1)\) vector of residual error. The pedigree utilized in the model was reconstructed to identify corrected kinship relationships from the original UC Davis Walnut Improvement Program pedigree using the Axiom J. regia 700K SNP array (Marrano et al. 2018). Narrow sense heritability per trait was estimated with equation (4),

\[
h^2 = \frac{\sigma^2_A}{\sigma^2_Y} = \frac{\sigma^2_A}{(\sigma^2_A) + (\sigma^2_R)}
\]

where the additive genetic variation \( \sigma^2_A \) is divided by the total phenotypic variation \( \sigma^2_Y \) (additive + environmental variation or residual).

**QTL mapping**

Quantitative-trait loci mapping was performed using the genetic maps (File S5) of the bi-parental population ‘Chandler’ × ‘Idaho’ described in (Sideli et al. 2020). In particular, the ‘Chandler’ genetic map spanned 998.31 cm with 1,165 markers in total, while the ‘Idaho’ map contained 1,753 markers for a total length of 1,693.88 cm. All measurements of pellicle color were evaluated in R/QTL software v1.4.9 (Broman et al. 2003). Phenotypic measurements from each year and the adjusted mean of the two years were first assessed using simple interval mapping in a single QTL model with scanone and ‘Haley Knott’ regression (Haley and Knott 1992). Significance thresholds were defined using 1,000 permutations for estimating critical values of alpha = 0.05 (Churchill and Doerge 1994). When multiple QTL were detected per trait, their precise location was adjusted and tested using ANOVA by considering both QTL locations. Briefly, utilizing the function sim.gen, multiple imputations were performed to fill in missing genotype data with a step of 1 (cm) and 100 imputations. A multiple QTL model was then fitted with the function fitQTL where both trait QTL identified in scanone were added in the model. The function refineQTL was applied and adjusted if the new positions in the ANOVA the model had a greater fit (LOD and p-values). The Bayesian credible interval for 5% error rate was implemented as the confidence QTL interval.

**Genome-wide association analysis**

The whole population was genotyped using the Axiom J. regia 700K SNP array as described in (Marrano et al. 2018). Association mapping was performed using the high-quality Poly High Resolution SNPs (335K; for details see (Marrano et al. 2018). Quality filtering was performed in PLINK 1.9 (Purcell et al. 2007); individuals with less than 90% genotypic data were filtered, along with SNPs having a minor allele frequency less than 5%, resulting in a total of 217K SNPs to apply for analysis. The maximum genotypic call error rate was set at 5%, and SNPs were retained if their genotype frequencies corresponded with Hardy Weinberg expectations (p-value < 0.001).

Principal components of SNP markers were obtained by running the mixed linear model (MLM) algorithm in GAPIT. The first component explained 9.39% variation, the second PC explained 6.97% variation, the third PC explained 5.31%, the fourth PC explained 3.41%, and the fifth PC explained 2.63%. The number of principal components (PCs) to include as covariates in the models was defined based on the screeplot (Figure S2).

Genome-wide associations were conducted using the Fixed and Random Model Circulating Probability Unification (FarmCPU) algorithm (Liu et al. 2016) and the multi-locus mixed linear model (MLMM) algorithm (Segura et al. 2012) in the R package genomic association and prediction integrated tool (GAPIT v3) (Wang and Zhang 2018). The MLMM is a multi-locus mixed linear model (MLM) (Yu et al. 2006) where both Q (population structure) + K (kinship matrix) are fitted to the model as random effects, so that type 1 errors are reduced from spurious associations due to relatedness and population structure. A VanRaden kinship matrix (VanRaden 2008), and PCs were calculated in GAPIT and added as covariates in the model. FarmCPU is also a multi-locus model developed to control false positives by utilizing markers as covariates in a stepwise MLM to remove markers that are confounding with kinship (Lui et al. 2016). This model was set to perform 10 iterations, with four to six PC’s added as covariates, depending upon the goodness of fit in the QQ-plots for each color phenotype, a minor allele frequency threshold of 5%, and the default parameters for bin size. A 5% Bonferroni threshold (2.29 × 10^-7) was used to assess significance, and Q-Q plots and Manhattan plots were inspected for evidence of inflation. A multiple corrections test (Gao et al. 2008) was applied in order to assess SNPs with a less stringent threshold (4.70 × 10^-6).

**Mapping of candidate genes**

Since large LD blocks were observed in the GWAS population, due to high relatedness of individuals, a 100 kb window around significant loci were scanned to identify putative candidate genes. This was determined by examining LD decay plots and there was a large drop in R^2 between 100-120 kb. Within these genomic regions the predicted mapped genes and functional annotation as described in (Marrano et al. 2020) were utilized from the new chromosome-level assembly of the ‘Chandler’ reference genome v2.0, found at https://www.ncbi.nlm.nih.gov/assembly/GCA_001411555.2 and annotations found at https://hardwoodgenomics.org/Genome-assembly/2539069?tripl pane=group_downloads.

**Data availability**

All data required to replicate the analyses are available as supplements cited in-text or in Supplemental data files 1-7. Computer vision

4414 | G. M. Sideli et al.
system methods and macros are available in supplemental (File S1, File S2, File S3) respectively. Output tables from ANOVA in the GWAS population provided in File S4. ‘Chandler’ and ‘Idaho’ genetic maps with phenotypic data are available in supplemental (File S5). Estimated breeding value phenotypes and principal components used in GWAS analysis are available in supplemental (File S6, File S7). Supplemental data file S8 contains raw color score data for both the GWAS population and ‘Chandler’ x ‘Idaho’ population. Supplemental material available at figshare: https://doi.org/10.25387/g3.12276539.

**RESULTS**

**Phenotypic variation in pellicle color**

In general, ‘Chandler’ produces kernels with an extra-light pellicle color, while Idaho’s kernels are light amber (Figure S1); these differences are consistently observed in both the DFA and CVS analyses. For the ANOVA linear models significant ($P < 0.0$) factors were family, year of harvest, family x year interaction, individual, and age; however age was not a significant factor in the linear models for DFA and L* phenotypes (File S8). For the calculation of adjusted means with ‘Chandler’ x ‘Idaho’ dataset, all terms and their interactions were significant ($P < 0.0001$) (File S4).

The phenotypic distribution of color scores in this study revealed that the highest frequency of DFA color score is found in the category ‘2’, while the CVS color scores are on a continuous scale (Figure 3). Phenotypic values for L*/C3 and a*/C3 are highly correlated, 0.80, while a*/C3 and DFA/C3 score show a moderate correlation of 0.62 (Figure 3).

**Heritability and breeding value estimates**

Pellicle color measurements acquired using the new image-based phenotyping method yielded values of the narrow-sense heritability

<table>
<thead>
<tr>
<th>Trait</th>
<th>Population</th>
<th>Chr</th>
<th>Pos (cM)</th>
<th>Interval</th>
<th>Physical Position (bp)</th>
<th>Pvalue</th>
<th>$R^2$</th>
<th>Year of detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>L*</td>
<td>Chandler</td>
<td>1</td>
<td>9.58</td>
<td>5.00 - 18.00</td>
<td>4620011</td>
<td>4.87 -6</td>
<td>0.107</td>
<td>17</td>
</tr>
<tr>
<td>L*</td>
<td>Chandler</td>
<td>1</td>
<td>10.18</td>
<td>2.98 - 22.59</td>
<td>4756788</td>
<td>9.07 -6</td>
<td>0.116</td>
<td>16, 17</td>
</tr>
<tr>
<td>L*</td>
<td>Chandler</td>
<td>16</td>
<td>27.27</td>
<td>16.24 - 36.89</td>
<td>20234419</td>
<td>2.58 -5</td>
<td>0.090</td>
<td>17</td>
</tr>
<tr>
<td>L*</td>
<td>Idaho</td>
<td>10</td>
<td>59.11</td>
<td>41.25 - 65.66</td>
<td>14732679</td>
<td>3.53 -5</td>
<td>0.090</td>
<td>16, 17</td>
</tr>
<tr>
<td>L*</td>
<td>Idaho</td>
<td>14</td>
<td>34.02</td>
<td>27.00 - 42.00</td>
<td>6515088</td>
<td>1.11 -4</td>
<td>0.076</td>
<td>17</td>
</tr>
<tr>
<td>a*</td>
<td>Chandler</td>
<td>1</td>
<td>9.60</td>
<td>5.39 - 18.00</td>
<td>4532719</td>
<td>3.63 -6</td>
<td>0.103</td>
<td>16, 17</td>
</tr>
<tr>
<td>a*</td>
<td>Chandler</td>
<td>7</td>
<td>1.19</td>
<td>0 - 70.00</td>
<td>993462</td>
<td>9.29 -6</td>
<td>0.105</td>
<td>16</td>
</tr>
<tr>
<td>a*</td>
<td>Chandler</td>
<td>16</td>
<td>27.27</td>
<td>16.24 - 36.89</td>
<td>20234419</td>
<td>1.16 -6</td>
<td>0.066</td>
<td>17</td>
</tr>
<tr>
<td>a*</td>
<td>Idaho</td>
<td>10</td>
<td>59.11</td>
<td>41.25 - 65.66</td>
<td>14732679</td>
<td>9.14 -5</td>
<td>0.090</td>
<td>17</td>
</tr>
<tr>
<td>a*</td>
<td>Idaho</td>
<td>14</td>
<td>35.21</td>
<td>27.00 - 42.00</td>
<td>6764116</td>
<td>4.29 -5</td>
<td>0.087</td>
<td>16, 17</td>
</tr>
<tr>
<td>b*</td>
<td>Chandler</td>
<td>1</td>
<td>9.57</td>
<td>2.98 - 22.59</td>
<td>4620011</td>
<td>6.29 -7</td>
<td>0.130</td>
<td>17</td>
</tr>
<tr>
<td>b*</td>
<td>Idaho</td>
<td>1</td>
<td>48.12</td>
<td>35.10 - 53.00</td>
<td>13073793</td>
<td>7.16 -5</td>
<td>0.082</td>
<td>16, 17</td>
</tr>
<tr>
<td>DFA</td>
<td>Chandler</td>
<td>1</td>
<td>9.96</td>
<td>0 - 10.00</td>
<td>4223640</td>
<td>1.88 -5</td>
<td>0.096</td>
<td>16, 17</td>
</tr>
<tr>
<td>DFA</td>
<td>Chandler</td>
<td>4</td>
<td>13.69</td>
<td>8.93 - 25.00</td>
<td>17907379</td>
<td>8.09 -5</td>
<td>0.087</td>
<td>16, 17</td>
</tr>
<tr>
<td>DFA</td>
<td>Idaho</td>
<td>10</td>
<td>88.89</td>
<td>73.00 - 104.00</td>
<td>31577799</td>
<td>5.33 -5</td>
<td>0.086</td>
<td>16, 17</td>
</tr>
<tr>
<td>DFA</td>
<td>Idaho</td>
<td>14</td>
<td>38.00</td>
<td>28.00 - 41.76</td>
<td>6764116</td>
<td>3.63 -5</td>
<td>0.092</td>
<td>17</td>
</tr>
</tbody>
</table>

Table 1 Results from QTL mapping for 168 individuals in the ‘Chandler’ x ‘Idaho’ mapping population for different color phenotypes. Interval was determined by Bayes 95% Credible Interval, Pvalues converted from LOD score cutoff determined from 1000 permutations of phenotypic data, $R^2$ is the variance explained for QTL; if two years of data then $R^2$ taken on multi-year data.
Figure 4 Manhattan plots from QTL mapping of 'Chandler' x 'Idaho' population. A. SNPs segregating in the 'Chandler' population. QTL identified on Chr01 for L*, a*, DFA, scores detected two-year dataset, while b* only detected in 2017, and on Chr04 for the DFA score in both years. B. 'Idaho' dataset contains QTL identified on Chr10 for L*, a*, and DFA averaged scores over 2016 and 2017, and QTL on Chr14 for the DFA score in 2017.
Figure 5 Manhattan plots from genome-wide association analysis with CVS phenotypes. A1. L results with MLMM model A2. L results with FarmCPU model. A3. QQ-plots for both models. B1. a results with MLMM model B2. L results with FarmCPU model. B3. a QQ-plots for both models. C1. b results with MLMM model B2. b results with FarmCPU model. B3. b QQ-plots for both models. SNPs that did not anchor to any chromosome during the assembly were left as undefined in the orange after Chr16.
similar to the DFA scoring system ($h_2^{\text{DFA}} = 0.56 \pm 0.05, h_2^{L*} = 0.58 \pm 0.15, h_2^{a*} = 0.60 \pm 0.04, h_2^{b*} = 0.53 \pm 0.07$). 'Scharsch Franquette', a selection made in California which is genetically identical to 'Franquette', a French variety, is a founder of the UC Davis Walnut Improvement Program, displayed the highest breeding value (47.35) for the lightness $L^*$ score. Families that possessed multiple individuals with top $L^*$ score breeding values included 10-024 (03-001-977 x 'Ivanhoe'), 11-030 (04-003-107 x 'Ivanhoe') and 09-25 (95-026-16 x 94-019-85) along with founder 'Waterloo', and 'Chandler' (Table S1).

### QTL mapping

Color scores often had overlapping regions of significant marker trait associations (Table 1, Figure 4). For the 'Chandler' map, QTL explaining $L^*$ and $a^*$ score phenotypes were identified on Chr01 and Chr16, where only on Chr01 was a stable QTL for both 2016 and 2017 thereby explaining 11.62% of variance for $L^*$ and 10.53% for $a^*$ (Figure 4, Figure S3). Additional QTL’s for $a^*$ score phenotype in the ‘Chandler’ map were detected on Chr07 in the year 2016 only (Figure S4). For the DFC score in the ‘Chandler’ map, QTL were located on Chr01, in an overlapping interval with $L^*$ and $a^*$ (Figure 4), and Chr04 (Figure S5). In ‘Idaho’ map, a QTL for both $L^*$ and $a^*$ score phenotypes was detected on Chr10 in both 2016 and 2017 explaining 9% of the variance for those traits. A second QTL on Chr10 was identified for the DFC score (Figure 4). In addition, a QTL for DFC scores was detected on Chr14 in correspondence with the QTL identified for the $L^*$ and the $a^*$ phenotypes (Figure S6). For the $b^*$ score, two QTL were identified on Chr10 in both populations ‘Chandler’ and ‘Idaho’, and they overlapped for some of their length (‘Chandler’ QTL interval 2,966,643 – 9,763,773 bp; ‘Idaho’ QTL interval 8,826,101 – 14,691,279 bp) (Figure 4, Figure S7).

### Genome-wide association mapping

Principal component analysis revealed four to six principal components which explained SNP variation for the 31 full-sib families in the Walnut Improvement Program, excluding ‘Chandler’ x ‘Idaho’ (Figure S2). See Figure 5 for an overview of Manhattan plots for CVS phenotypes. The most significant marker trait association for the $L^*$ score was on Chr01 with a MAF of 0.20 (Table 2). There were significant SNPs detected for both $L^*$ and $b^*$ scores that were identical (Chr01 and Chr12). For the $a^*$ score, the most significant marker trait association was found on Chr08, while significant trait-to-trait correlation were also identified on Chr10, Chr11, Chr14, Chr15, Chr16.

---

**Table 2 Most significant marker-trait associations from genome-wide association scan of (n = 528)**

<table>
<thead>
<tr>
<th>Trait</th>
<th>Chr</th>
<th>Position (bp)</th>
<th>Pvalue</th>
<th>$^a$MAF</th>
<th>$^b$R²</th>
<th>Effect (est)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L*</td>
<td>1</td>
<td>24588560</td>
<td>2.05E-15</td>
<td>0.20</td>
<td>0.05</td>
<td>-1.02</td>
</tr>
<tr>
<td>L*</td>
<td>4</td>
<td>3580183</td>
<td>5.48E-8</td>
<td>0.16</td>
<td>0.02</td>
<td>0.73</td>
</tr>
<tr>
<td>L*</td>
<td>5</td>
<td>7700218</td>
<td>2.73E-11</td>
<td>0.10</td>
<td>0.07</td>
<td>-1.08</td>
</tr>
<tr>
<td>L*</td>
<td>8</td>
<td>28491012</td>
<td>2.49E-11</td>
<td>0.07</td>
<td>0.04</td>
<td>-1.28</td>
</tr>
<tr>
<td>L*</td>
<td>9</td>
<td>14110666</td>
<td>8.11E-8</td>
<td>0.47</td>
<td>0.07</td>
<td>0.63</td>
</tr>
<tr>
<td>L*</td>
<td>12</td>
<td>1886325</td>
<td>5.18E-6</td>
<td>0.14</td>
<td>0.05</td>
<td>0.66</td>
</tr>
<tr>
<td>L*</td>
<td>12</td>
<td>3406040</td>
<td>2.66E-6</td>
<td>0.47</td>
<td>0.06</td>
<td>0.88</td>
</tr>
<tr>
<td>L*</td>
<td>15</td>
<td>15274696</td>
<td>9.08E-10</td>
<td>0.09</td>
<td>0.02</td>
<td>1.05</td>
</tr>
<tr>
<td>a*</td>
<td>1</td>
<td>37670750</td>
<td>7.41E-7</td>
<td>0.11</td>
<td>0.05</td>
<td>-0.30</td>
</tr>
<tr>
<td>a*</td>
<td>8</td>
<td>24731099</td>
<td>5.22E-8</td>
<td>0.09</td>
<td>0.07</td>
<td>-0.81</td>
</tr>
<tr>
<td>a*</td>
<td>8</td>
<td>18643646</td>
<td>1.64E-14</td>
<td>0.10</td>
<td>0.04</td>
<td>-0.28</td>
</tr>
<tr>
<td>a*</td>
<td>10</td>
<td>34936156</td>
<td>7.66E-9</td>
<td>0.45</td>
<td>0.04</td>
<td>-0.25</td>
</tr>
<tr>
<td>a*</td>
<td>11</td>
<td>36386668</td>
<td>5.44E-6</td>
<td>0.05</td>
<td>0.03</td>
<td>0.23</td>
</tr>
<tr>
<td>a*</td>
<td>14</td>
<td>3094777</td>
<td>5.70E-7</td>
<td>0.28</td>
<td>0.04</td>
<td>-0.26</td>
</tr>
<tr>
<td>a*</td>
<td>15</td>
<td>16729948</td>
<td>3.28E-9</td>
<td>0.07</td>
<td>0.01</td>
<td>0.45</td>
</tr>
<tr>
<td>b*</td>
<td>1</td>
<td>24588560</td>
<td>1.15E-12</td>
<td>0.20</td>
<td>0.05</td>
<td>-1.02</td>
</tr>
<tr>
<td>b*</td>
<td>1</td>
<td>35482903</td>
<td>1.83E-7</td>
<td>0.41</td>
<td>0.04</td>
<td>0.46</td>
</tr>
<tr>
<td>b*</td>
<td>7</td>
<td>5027345</td>
<td>7.56E-6</td>
<td>0.06</td>
<td>0.05</td>
<td>-0.24</td>
</tr>
<tr>
<td>b*</td>
<td>7</td>
<td>43498845</td>
<td>2.61E-6</td>
<td>0.09</td>
<td>0.06</td>
<td>0.44</td>
</tr>
<tr>
<td>b*</td>
<td>8</td>
<td>24680376</td>
<td>5.88E-6</td>
<td>0.30</td>
<td>0.06</td>
<td>0.31</td>
</tr>
<tr>
<td>b*</td>
<td>9</td>
<td>11456318</td>
<td>7.44E-7</td>
<td>0.24</td>
<td>0.06</td>
<td>-0.29</td>
</tr>
<tr>
<td>b*</td>
<td>10</td>
<td>12202442</td>
<td>1.03E-8</td>
<td>0.43</td>
<td>0.04</td>
<td>0.38</td>
</tr>
<tr>
<td>b*</td>
<td>11</td>
<td>5382614</td>
<td>1.62E-5</td>
<td>0.16</td>
<td>0.02</td>
<td>-0.35</td>
</tr>
<tr>
<td>b*</td>
<td>12</td>
<td>3406040</td>
<td>6.27E-8</td>
<td>0.47</td>
<td>0.06</td>
<td>0.88</td>
</tr>
<tr>
<td>b*</td>
<td>13</td>
<td>26311574</td>
<td>6.23E-6</td>
<td>0.09</td>
<td>0.01</td>
<td>-0.24</td>
</tr>
<tr>
<td>b*</td>
<td>13</td>
<td>34566210</td>
<td>2.73E-7</td>
<td>0.06</td>
<td>0.05</td>
<td>0.63</td>
</tr>
<tr>
<td>b*</td>
<td>14</td>
<td>3126302</td>
<td>6.48E-9</td>
<td>0.29</td>
<td>0.04</td>
<td>0.53</td>
</tr>
<tr>
<td>DFC</td>
<td>6</td>
<td>13137859</td>
<td>8.99E-9</td>
<td>0.36</td>
<td>0.006</td>
<td>0.26</td>
</tr>
<tr>
<td>DFC</td>
<td>7</td>
<td>51050704</td>
<td>4.70E-11</td>
<td>0.14</td>
<td>0.08</td>
<td>-0.13</td>
</tr>
<tr>
<td>DFC</td>
<td>8</td>
<td>28588006</td>
<td>4.25E-6</td>
<td>0.07</td>
<td>0.02</td>
<td>0.17</td>
</tr>
<tr>
<td>DFC</td>
<td>9</td>
<td>14771377</td>
<td>1.62E-6</td>
<td>0.43</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>DFC</td>
<td>12</td>
<td>25920027</td>
<td>4.26E-7</td>
<td>0.49</td>
<td>0.05</td>
<td>-0.07</td>
</tr>
<tr>
<td>DFC</td>
<td>13</td>
<td>38230982</td>
<td>1.11E-9</td>
<td>0.49</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td>DFC</td>
<td>16</td>
<td>26455353</td>
<td>3.47E-7</td>
<td>0.14</td>
<td>0.04</td>
<td>0.09</td>
</tr>
</tbody>
</table>

$a^a$MAF minor allele frequency.

$b^b$R² variance explained by significant SNP, P value significance from Bonferroni taken from $L^*$ = model with 4 PCs, $a^*$ = model with 4 PCs, $b^*$ = model with 6 PCs, DFC = model from 5 PCs.
Chr16 with different models. For the b* score, both the MLMM and FarmCPU models detected significant associations on Chr07 and Chr08. Additional associations for the b* score were found on Chr09 and Chr10, with many other small QTL just below threshold (Chr01, Chr04, Chr11, Chr12). For the DFA scores, a significant marker trait association on Chr08 was detected with both MLMM and FCPU models (Figure S8). Significant associations were identified in similar chromosomal regions for both the CVS and DFA scores, even though many of them did not reach the significance level of $2.29 \times 10^{-7}$ (i.e., Chr 01, Chr09 and Chr13, Chr16) (Figure S8).

The SNP associated with the DFA scores were in moderate LD with SNP associated with the L* scores on Chr01 ($r^2 = 0.37$) and with SNP associated with the a* score on Chr16 ($r^2 = 0.63$). Pairwise LD measures between these SNPs associated with DFA and L* scores and SNPs associated with DFA and a* scores are visualized to show close proximity to each other 16.29 kb and 213 kb respectively (Figure 6).

Co-localization of marker trait associations in QTL mapping and GWAS

In many cases there were overlapping regions between marker trait associations detected in both QTL mapping and GWAS (Figure 7). For Chr01, both QTL intervals in association with L* score and DFA score detected in ‘Chandler’ map (220,484 - 9,763,773 bp) and (1,495,995 - 4,819,547 bp) respectively, and SNPs identified from GWAS within those intervals (5,824,300 bp) and (2,674,553 bp) respectively overlapped. Chromosome 10 displayed two co-localized regions for both associations found for L* and DFA score in the ‘Idaho’ map and the GWAS; L* score phenotype associated with SNP at (14,011,277 bp) and L* score phenotype associated with SNP at (14,011,277 bp) fell within (9,513,973 - 15,033,186 bp) QTL region, and DFA score associated with SNP at (26,668,826 bp) associated with the a* score lies 4.47 Mbp (22,190,531 bp) away from the end of the QTL interval identified in ‘Chandler’ map.

Similarly, (Marrano et al. 2019) performed a GWAS analysis on 584 walnut trees of the UC Davis Walnut Improvement Program utilizing 30 years of historical data based upon the DFA scoring method. There were two marker trait associations; Chr07 and Chr09 which co-located with three marker trait associations in this study for DFA and L* scores, and were in moderate LD ($r^2 = 0.54, 0.64, 0.50$) respectively (Table S2).

Complete L*, a*, b* information for putative gene functions

Six hundred, twenty-nine (629) putative candidate genes were identified within 100 kb windows surrounding the significant SNPs (Table S3). These putative candidate genes are predicted to have functions related to abiotic stress, pathogen resistance, hormone signaling and the biosynthesis of pigmented metabolites (Table S3, Table 3, Figure 8). Three putative gene functions were identified in association with SNPs on Chr01. The first of these gene functions is the transcription factor MYB113 which is associated with the L* score. The second gene function appears to encode cinnamoyl alcohol dehydrogenase (CAD), which is associated with the b* score. Lastly, the third gene function which lies on Chr01 encodes tryptophan synthase (TRP), which is also associated with the b* score. One gene function (transcription factor MYB1) associated with the DFA score was identified on Chr07. For the a* score, on Chr10 flavonol synthase/flavonone 3-hydroxylase (FLS) was identified to be associated with a SNP, as well as on Chr11 chorismate mutase (CM). Two putative gene functions were identified on Chr12; the first of these genes was associated with both an L* and b* score, and encoded 4-coumarate—CoA ligase while the second gene encoded flavonoid 3*-monooxygenase, was associated with a DNA score. Four gene functions were identified to be associated with SNPs on Chr13; the first gene function is anthranilate phosphoribosyltransferase (PAT) which was associated with a b* score, the second gene function was also associated with a b* score.
and encoded caffeic acid 3-O-methyltransferase (COMT), and the final two gene functions on Chr13 both encoded FLS genes, were both associated with L*/C3 scores (Table 3).

**DISCUSSION**

The use of semi-automated and high throughput phenotyping of walnut pellicles uncovered novel SNPs associated with color in walnut pellicles. The CVS was able to deconstruct the three main color components contributing to trait variation which can be used as a fast and reliable method for data collection and processing walnut pellicles.

By combining two complementary approaches, QTL mapping and genome-wide association, we performed genetic analysis for pigmentation in walnut pellicles clearly indicating a quantitative nature of the trait by observing a few large effect loci, and many small effects loci contributing to trait variation. In our QTL mapping study with a population size of about 200, only Chr01 and Chr07 harbored QTL explaining greater than or equal to 10% total variance, while most QTL found explained between 6–9% of the variance, and the average genetic variance explained by SNP was 5%. In our GWAS, with a population size of about 530, there were several small and one or two large effect QTL for each color phenotype.

For the QTL mapping analysis, the ‘Idaho’ genetic map which was constructed in (Sideli et al. 2020) was found to have nearly a twofold increase in map size of ‘Chandler’ genetic map. Shorter genetic

![Figure 7 Co-localization of QTL mapping and genome-wide association results. Associations for the L*, DFA and a* phenotypes on Chr01, 10, 14, 16 displayed with corresponding colored boxes or lines.](image-url)
maps of female parent in walnut and other species has also been observed (Luo et al. 2015; Kefayat et al. 2019; Marrano et al. 2019) thereby suggesting a lower recombination rate in the female parent of walnut. ‘Chandler’ was also found to have many homoygous regions which further supports lower recombination (Marrano et al. 2020). On Chr01 of ‘Chandler’ genetic map, multiple QTL were identified in the same genomic region for all color phenotypes. This can suggest that the color determination is detecting the same differences contributing to pigmentation due the correlation between color scores; L* (lightness) was found to be in high association with the b’ (amount of yellow), and a’ (amount of red) in moderate association with DFA. The DFA scoring phenotype has low resolution power resulting in a QTL curve that is less significant than the other three color phenotypes. Consequently, it is less sensitive because it lacks the ability to discriminate specific color components L*, a*, b* simply because it encompasses all three measurements. The fact that the L* and a’ were found to have overlapping regions of association is due to the fact that in the biosynthetic pathway, flavonol synthase (FLS/F3H) is bifunctional to perform different reactions forming flavonols (FLS) or proanthocyanidins (PAs) (Lukacin et al. 2003), all which are pigmented in a different way suggesting a complex regulation for production of flavonols. Further exploration is needed to rule out possible pleiotropy. For the GWAS analysis we tested two models, FarmCPU and MLM, in order to account for population structure and/or kinship in our population as to reduce any false positive or false negative associations. As applied in the FarmCPU model, PCs will help account for population structure, while for MLM model, both PCs and a kinship matrix is applied to account for levels of relatedness among individuals (Yu et al. 2006). The FarmCPU method has a higher detection for marker trait associations and the MLM method is known to miss associations if they are related with population structure (Cui et al. 2018). In our study we detected QTL that overlapped with methods (QTL mapping and GWAS) and models, and each method was able to detect additional QTL with large effects. Here, FarmCPU method was more sensitive in that it detected more marker-trait associations with greater effects than the MLM method. However, a plausible explanation is that some of these associations can be related to population relatedness, since they were not detected by the MLM, which explicitly accounts for levels of relatedness using the Q and K matrices as covariates.

Several candidate genes acting in pathways linked with the production of pigmented metabolites in the flavonoid biosynthetic pathway were identified by association with SNPs generated by the GWAS and QTL analysis. Two putative genes in early flavonoid biosynthesis, CM and FLS/F3H were associated with the a’ score (9.5, 9.4 respectively), indicating that these genes may contribute to a slight reddish/brown pigmentation. CM is involved in the biosynthesis of phenylalanine and tyrosine as well as their derivatives, the phenylpropanoids, lignins, flavonoids and anthocyanins, most of which are pigmented (Cotton and Gibson 1965; Cotton and Gibson 1968). FLS/F3H is a bifunctional enzyme (Lukacin et al. 2003); F3H generates the dihydroflavonols which are precursors to anthocyanins and red-brown proanthocyanidins (Britsch et al. 1980; Holton et al. 1993). Similarly, the putative genes PAT, TS, 4CL, CAD and CMOT were associated with the b’ score (26.5, 27.1, 27.6, 27.5, 27.4, respectively), suggesting that these genes promote yellow color. PAT and TS both act in the biosynthesis of tryptophan, a light yellow AAA and its yellow/orange derivative tryptamine (Lovenberg et al. 1962; Baxter and Slaytor 1972; Gibson et al. 1972). 4CL yields a precursor for brown pigmented lignins and the rest of the flavonoid biosynthesis pathway (Lindl et al. 1973; Gross and Zenk 1974). COMT and CAD both act in modifying lignin precursors (Poulton et al. 1976; Wyrambik and Grisebach 1979; Sarni et al. 1984; Gowing et al. 1991). As previously described 4CL and FLS/F3H, also correspond to association with the L’ score (41.5, 42.0 respectively), suggesting that these genes are associated with a pigmentation of the pellet; bifunctional FLS/F3H- F3H desaturates dihydroflavonols to yield yellow/brown flavonols (Forkmann et al. 1980; Charrier et al. 1995). Finally, the putative gene F3’H was associated with the DFA score (2.2), indicating a possible diversity of colored pigments. F3’H hydroxylates flavanones, dihydroflavonols, flavones and flavonoids to increase the diversity of each of the aforementioned categories of yellow flavonoids (Forkmann et al. 1980; Bruguiera et al. 1999).

The results from this study will enable molecular breeders to develop competitive allele specific pcr (KASP) markers from the most significant SNPs, assisting in efficient selection at the seedling stage rather than the selection process over a 5-10 years period requiring trees to reach sexual maturity. Early selection of progeny will dramatically increase the efficiency in developing new cultivars. Due to the fact that walnut growers are compensated based on human grader’s perception of color, an improvement to the CVS algorithm would include human input factors such as discrimination for veining, speckling or size can be considered.

For further investigation, the creation and evaluation of genomic prediction models will enable the integration of all the small QTL effects to predict progeny with lighter pellets, as the industry demands. To solidify results pertaining to the putative candidate genes for flavonoid production, an experiment setup to clone each gene and express them in bacteria would verify gene function and
Figure 8 Flavonoid biosynthesis overview. Shown is a summary of flavonoid biosynthesis, including the shikimate, aromatic amino acid, phenylpropanoid and flavonoid biosynthesis pathways. Gene abbreviations in black were found to be associated with SNPs in this study. The expression of genes in red are known to be regulated by MYB1 and/or MYB113 transcription factors, but not associated with the SNPs reported in this study.
identify the origin of the chemical reactions occurring in response to developmental stages, environmental stress or autophagy.

ACKNOWLEDGMENTS

This work was funded solely by the walnut growers of California through the California Walnut Board. We want to thank both Shira Bergman and Irwin Donis-Gonzalez for their engineering assistance with the CVS, Sara Montanari for consultation with QTL mapping, and Monica Britton for consultation with candidate gene analysis.

LITERATURE CITED


Butler, D., B. Gogel, B. Cullis, and R. Thompson, 2017 Navigating from through the California Walnut Board. We want to thank both Shira Bergman and Irwin Donis-Gonzalez for their engineering assistance with the CVS, Sara Montanari for consultation with QTL mapping, and Monica Britton for consultation with candidate gene analysis.


Poulton, J., K. Halibrock, and H. Grisebach, 1976 Enzymic synthesis of lignin precursors: Purification and properties of the S-adenosyl-l-methionine:


Communicating editor: T. Pyhäjärvi