Bivariate genome-wide association study reveals polygenic contributions to covariance between total carotenoid and dry matter contents in yellow-fleshed cassava *(Manihot esculenta)*

Seren S. Villwock[1*], Elizabeth Y. Parkes[2], Edwige Gaby Nkouaya Mbanjo[3], Ismail Y. Rabbi[3], Jean-Luc Jannink[1,4]

Author affiliations:

[1] Cornell University, College of Agriculture and Life Sciences, School of Integrative Plant Science, Section of Plant Breeding and Genetics, Emerson Hall, 222 Tower Rd, Ithaca, New York 14853, USA

[2] International Institute of Tropical Agriculture, Southern Africa Research & Administration Hub, Plot 1458B, Ngwerere Road, Chongwe District, Lusaka Province, Zambia

[3] International Institute of Tropical Agriculture, Headquarters & West Africa Hub, PMB 5320, Oyo Road, Ibadan 200001, Oyo State, Nigeria

[4] US Department of Agriculture-Agricultural Research Service, Plant, Soil and Nutrition Laboratory, Robert W. Holley Center for Agriculture and Health, 538 Tower Rd, Ithaca, New York 14850, USA

***Author for correspondence:**

Seren S Villwock

Tel: +1 512 820 6362

Email: ssv42@cornell.edu

Running title (<50 characters):

Polygenic covariance between carotenoids and dry matter in cassava

Keywords:

Genetic linkage, pleiotropy, GWAS, biofortification

Abstract (< 250 words)

Cassava breeders aim to increase the provitamin A carotenoid content of storage roots to help combat vitamin A deficiency in sub-Saharan Africa, but a negative genetic correlation between total carotenoid (TC) and dry matter (DM) contents hinders breeding efforts. Genetic linkage between a major-effect variant in the *phytoene synthase 2 (PSY2)* gene and nearby candidate gene(s) has been thought to drive this correlation. Evidence from molecular experiments, however, suggest there may be a metabolic relationship between TC and DM, which we predicted would create genome-wide mediated pleiotropy. Bivariate genome-wide associations were used to test the hypothesis of pleiotropy and examine the genetic architecture of the negative covariance between TC and DM. A population of 378 accessions in the yellow-fleshed cassava breeding program at the International Institute of Tropical Agriculture (IITA) in Ibadan, Nigeria was genotyped with DArTseqLD. TC measured by iCheck™ spectrometer and DM data were available from field trials over ten years across three locations in Nigeria. Mixed linear models controlling for the previously-identified *PSY2* causal variant were used to identify multiple new quantitative trait loci (QTL) jointly associated with both traits. A binomial test with 13 jointly-associated loci identified at a relaxed significance threshold indicated that loci are significantly more likely to affect TC and DM in opposite directions. Even after accounting for the effects of these 13 loci, there was significantly negative polygenic covariance between TC and DM remaining. These findings support the hypothesis that mediated pleiotropy rather than genetic linkage drives the negative genetic correlation between TC and DM in cassava and demonstrate a new application of multivariate GWAS for interrogating the genetic architecture of correlated traits.

Plain language summary (<100 words)

Increasing provitamin A in cassava roots has reduced their dry matter content, making vitamin-enriched cassava varieties less desirable. This study used multi-trait models to identify shared genetic factors, most of which had opposing effects on the two traits. The negative relationship was distributed across the genome, suggesting an inherent physiological trade-off. These findings will guide breeders in developing selection strategies for vitamin-enriched cassava and other starchy crops. More broadly, this study demonstrates the use of multi-trait associations to help distinguish whether traits are associated due to separate, nearby genes (genetic linkage) or if the same genes affect multiple traits (pleiotropy).

Introduction

Vitamin A deficiency is a global public health burden impacting around half of all children in sub-Saharan Africa despite supplementation initiatives (Stevens et al. 2015). Humans require dietary consumption of provitamin A carotenoids to support vision and immune health (Wiseman et al. 2017). Carotenoids are isoprenoid metabolites that are responsible for the yellow, orange, and red pigments in many fruits and vegetables. Certain carotenoids, including β-carotene, α-carotene, and β-cryptoxanthin, are cleaved to form vitamin A in the human body (Perveen et al. 2021). Significant efforts in the last twenty years have focused on biofortifying staple crops with increased carotenoid content to help provide nutritional security for populations at risk for vitamin A deficiencies (Asare-Marfo et al. 2013; Giuliano 2017).

Cassava (*Manihot esculenta* Crantz) has been a particular focus of biofortification efforts as a major food security crop in sub-Saharan Africa, where around 100kg is consumed annually per capita (FAO 2021). Cassava is a woody, tropical plant with starchy storage roots that are a staple food consumed by millions of people globally. Though white-fleshed cassava is generally low in micronutrients, yellow-fleshed varieties accumulate low amounts of provitamin A carotenoids, predominantly β-carotene, which increases both its nutritional value and post-harvest storability (Sánchez et al. 2006; Afolami et al. 2021; Esuma et al. 2021). There are now biofortified varieties of cassava with substantial carotenoid content that have been generated through both conventional selection and genetic engineering (Atser 2014; Beyene et al. 2018; Parkes et al. 2020; Abincha et al. 2024).

In several starchy crops, including cassava and sweetpotato, negative correlations between total carotenoid (TC) content and dry matter (DM) content have been widely reported in breeding populations (Akinwale et al. 2010; Cervantes-Flores et al. 2011; Njoku et al. 2015; Esuma, Kawuki, et al. 2016; Rabbi et al. 2017; Gemenet et al. 2020; Parkes et al. 2020; Esuma et al. 2021). Reductions in DM, which is mainly comprised of starch, negatively impact dry yield, texture, and consumer acceptance of new cassava varieties (Tahirou et al. 2015). It is unclear whether the trade-off between TC and DM is caused by genetic linkage or pleiotropy, with evidence supporting both hypotheses (Ceballos et al. 2013; Rabbi et al. 2017; Beyene et al. 2018; Gutschker et al. 2024; Villwock et al. 2024).

Previous genome-wide association studies (GWAS) in cassava have identified a locus at 24.15 Mbp on chromosome 1 that is associated with both TC and DM, contributing 70% and 37% of the variance for those traits, respectively (Rabbi et al. 2017; Rabbi et al. 2022). This locus contains *phytoene synthase 2 (PSY2)*, the rate-limiting enzyme in carotenoid synthesis, harboring a

65 functionally-validated variant ($C_{572}A$) that increases the catalytic efficiency of PSY2 and promotes
66 carotenoid accumulation (Welsch et al. 2010). There are also two candidate genes for DM at this
67 locus, *sucrose synthase* (SUS) and *ADP-glucose pyrophosphorylase (ADP-glc PPase),* although causal
68 variants in these genes have not been investigated, and the genetic resolution is not high enough to
69 distinguish whether these quantitative trait loci (QTL) are separate from *PSY2*. Similarly, a locus
70 near *PSY2* and *SUS* associated with both TC and DM has also been reported in sweetpotato breeding
71 populations (Gemenet et al. 2020). This evidence strongly supports the hypothesis that an
72 unfavorable genetic linkage at the chromosome 1 locus drives the trade-off between TC and DM. A
73 Latin American cassava population that has overcome the negative correlation through recurrent
74 selection further supports the hypothesis of a breakable genetic linkage (Ceballos et al. 2013).
75 African cassava breeding programs, however, have not been able to replicate these results despite
76 continued selection for both traits.
77 Other evidence points to a pleiotropic effect of *PSY2* on both TC and DM due to a
78 physiological relationship between the traits. Cassava lines accumulating carotenoids due to
79 expression of *PSY* and *DXS* transgenes exhibited a 50-60% reduction in starch content, along with
80 broad metabolic shifts including increases in sucrose, fatty acid content, and abscisic acid (ABA)
81 content (Beyene et al. 2018). In an African cassava breeding population, yellow-fleshed cassava
82 genotypes exhibited increased biosynthesis of cell wall components, increased sucrose, and
83 decreased starch compared to white-fleshed genotypes (Gutschker et al. 2024). Similar metabolic
84 perturbations have been observed in other crop species engineered to accumulate higher
85 carotenoid content, including potato, maize, and citrus (Diretto et al. 2010; Li et al. 2012; Cao et al.
86 2015; Decourcelle et al. 2015).
87 These data suggest the presence of an inherent metabolic interaction between TC and DM,
88 for which several possible biological mechanisms have been proposed, including apocarotenoid
89 signaling and competition for carbon precursors (Olayide et al. 2023; Gutschker et al. 2024;
90 Villwock et al. 2024). If such a metabolic interaction exists, we would expect multiple genomic
91 regions associated with TC to exhibit pleiotropic effects on DM, or *vice versa,* depending on the
92 underlying mechanism. Additionally, we would expect a significant majority of associated genetic
93 variants to affect the traits in opposite directions. Solovieff et al. (2013) define this type of
94 pleiotropy, based on a causal relationship between traits, as "mediated pleiotropy."
95 To statistically test the hypothesis of mediated pleiotropy, the identification of more small-
96 effect QTL is needed. Previous GWAS for TC have primarily detected only the *PSY2* locus on
97 chromosome 1 that is responsible for the qualitative difference in carotenoid accumulation

between white- and yellow-fleshed cassava, with a few other minor loci were reported on chromosomes 2, 5, 13, and 15 (Esuma, Herselman, et al. 2016; Rabbi et al. 2017; Ikeogu et al. 2019; Rabbi et al. 2022). The yellow-fleshed cassava breeding population at the International Institute of Tropical Agriculture (IITA) in Ibadan, Nigeria presents the opportunity to discover new alleles contributing to quantitative variation in TC and determine if they are jointly associated with DM. Since yellow-fleshed cassava is part of a separate breeding pipeline, there may be different genetic variation within the yellow-fleshed population as opposed to between the white and yellow populations that is overshadowed by the *PSY2* major effect. In addition, the yellow population has been routinely phenotyped using a spectrophotometric device that measures TC more precisely than the indirect estimates of TC in previous GWAS studies, which have used either visual color chart, chromameter measurements, or near-infrared spectroscopy predictions of TC (Jaramillo et al. 2018). Therefore, we genotyped the yellow-fleshed cassava breeding population at IITA and employed a multivariate GWAS approach to test associations between SNPs and the bivariate phenotype of TC and DM.

Multivariate associations have increased statistical power compared to univariate associations when traits are correlated, even if only one phenotype is associated with a given genetic variant (Stephens 2013; Galesloot et al. 2014). This is a counterintuitive but potentially helpful property of multivariate analysis when dealing with an unfavorable correlation between traits because it could facilitate the identification of useful loci that only affect one trait without negatively impacting the other. In some cases, QTL that are not detectable in univariate associations can be identified in multivariate associations due to this increased power (Stephens 2013).

A major limitation of most multivariate GWAS methods is the difficulty in interpreting a significant association with multiple traits, since it is not clear which traits are associated and whether the associations are direct or indirect (Solovieff et al. 2013; Stephens 2013). A common approach is to follow multivariate tests with multiple univariate tests to help interpret which of the traits are associated (Stephens 2013; Taraszka et al. 2022). This approach has limitations because it loses the statistical power benefit of the multivariate test and fails to account for the trait correlations. As a result, the univariate tests may result in false-positive associations with traits that are only indirectly associated with the variant via their correlation with a directly associated trait (Stephens 2013). Our approach overcomes these challenges of traditional multivariate GWAS by explicitly estimating the effect of the SNP on each trait and conducting a bivariate test for joint associations, which facilitates interpretation of how effects on each trait contribute to a multivariate association.

131 We further expand on a traditional multivariate GWAS. Rather than focusing on single
132 associations, we examine the pattern of associations across the genome to help infer the biological
133 relationship between correlated traits. Specifically, we assess the distribution of jointly-associated
134 SNPs across the genome to test the hypothesis that TC and DM are genetically correlated only
135 because of the *PSY2* major effect locus. If this major locus is the sole or primary cause, we expect
136 that after controlling for the *PSY2* locus and the traits' overall genetic correlation, other loci would
137 only be associated with a single trait and that their effect directionality would be random.
138 Alternatively, if TC and DM are genetically correlated due to an inherent metabolic or physiological
139 relationship, we expect that most QTL associated with one trait would also be associated with the
140 other trait and would affect the traits in opposite directions. Multivariate models further present a
141 framework for characterizing the genome-wide contributions to the genetic covariance between
142 traits.
143 To distinguish between mediated pleiotropy versus "biological" or direct pleiotropy,
144 statistical mediation analyses have been used to infer causal relationships between traits (Mackay
145 and Anholt 2024). With mediation analysis, a regression for one trait is fit conditional on an
146 associated trait to test for conditional independence and to partition causal effect into direct and
147 mediated effects (Rockman 2008). Stephens (2013) presented a Bayesian model comparison
148 framework for determining whether multiple correlated traits that are jointly associated with a
149 genetic variant are individually associated directly, indirectly, or unassociated. By using Bayesian
150 model averaging rather than directly fitting each linear model for comparison, the computational
151 time is greatly reduced, but the proposed algorithm did not allow for the correction of population
152 structure and genomic relatedness through covariates. To overcome this challenge, we
153 implemented a frequentist version of the model comparison method described in Stephens (2013)
154 to distinguish between indirect and direct trait associations while properly accounting for
155 population structure.
156 While these statistical association models may not necessarily reflect biological
157 mechanisms, examining the pattern of associations across the genome provides a framework to
158 evaluate evidence for biological or mediated pleiotropy as contrasted with genetic linkage.
159 Determining whether genetic linkage or pleiotropy is involved will inform breeding strategies for
160 increasing or maintaining both TC and DM, which is critical for the adoption of PVA biofortified
161 staple crops and shed light on potential metabolic interactions between carotenoids and
162 carbohydrates.
163

164 **Methods**

165 *Germplasm and genotypic data*

166 A sample of 380 cassava accessions were selected to represent the yellow cassava breeding
167 population at the International Institute of Tropical Agriculture (IITA) in Ibadan, Nigeria, including
168 the HarvestPlus (https://www.harvestplus.org/) and the NextGen Cassava project
169 (https://www.nextgencassava.org/) breeding pipelines. The population included recent ancestry
170 from Latin America, with 26 Latin American clones and 34 F1 or F2 hybrids between Latin
171 American and African clones. 78% of the clones were early-stage selections, while the remainder
172 were more advanced selections. 92% of the clones were classified as "cream" or "yellow" flesh types
173 while the remainder were "white" based on visual color rating.

174 Leaf tissues from the clones were sampled and sent to Intertek (Parkside, SA, Australia) for
175 DNA extraction. DNA was sequenced, aligned to cassava reference genome v6.1, and 12,981 SNPs
176 were called using the DArTSeqLD platform (Diversity Arrays Technology, Bruce, ACT, Australia).
177 These data were combined with existing genotype data for 50 accessions downloaded from
178 Cassavabase (https://cassavabase.org/). The identities of 9 pairs of duplicate genotype records for
179 the same accession were checked with 'bcftools gtcheck' (*bcftools* version 1.18) and highly similar
180 duplicates were removed. Dissimilar duplicates with a high number of mismatching genotypes
181 were both removed, excluding two accessions from further analysis.

182 Additional and missing SNPs were imputed with Beagle5.0 (Browning et al. 2021) using the
183 R package genomicMateSelectR version 0.2.0 (Wolfe 2022) with a reference panel of 21,856
184 accessions available on Cassavabase
185 (https://cassavabase.org/ftp/marnin_datasets/nextgenImputation2019/). Then, 7,096 accessions
186 available on Cassavabase with both DArTSeqLD genotypes and KASP marker genotypes for the
187 *PSY2* causal allele at S1_24155522 were joined and phased with Beagle5.0 to create a reference
188 panel for imputing the *PSY2* marker. Following imputation, SNPs were filtered with the following
189 parameters: dosage R^2 (DR2) > 0.75 (with the exception of the *PSY2* marker with DR2 = 0.74),
190 Hardy-Weinberg equilibrium p-value (P_HWE) > 1e-20, and minor allele frequency (MAF) > 0.05
191 for a total of 55,436 biallelic SNPs for 378 unique accessions. Imputed genotypes were rounded to
192 {0, 1, 2}.

193 Population structure was visualized with principal component (PC) analysis in R using
194 *prcomp()* on the genetic data matrix and *ggplot()* for plotting. The first five PCs were used to
195 account for population structure in the genome-wide associations. An additive genomic
196 relationship matrix calculated using the rrBLUP package (Endelman 2011) was used to account for

197 relatedness among the genotypes. Linkage disequilibrium (LD) patterns in the data were evaluated
198 by calculating R^2, the squared Pearson correlation coefficient, between each pair of SNPs with *cor()*
199 and *ggplot().*
200
201 *Phenotypic data*
202 Phenotypic data from field trials containing three or more of the genotyped accessions were
203 downloaded from Cassavabase (https://cassavabase.org/). 51 trials that contained carotenoid data
204 for at least three genotyped accessions and/or dry matter data for at least 10 genotyped accessions
205 were selected for analysis (Table S1). Of these, 6 were at the Clonal Evaluation Trial (CET) stage, 17
206 were at the Preliminary Yield Trial (PYT) stage, 12 were at the Advanced Yield Trial (AYT) stage, 12
207 were at the Uniform Yield Trial (UYT) stage, and 4 were genetic diversity collection trials.
208 Dry matter (DM) content percentage was measured as the ratio of the weight before and
209 after oven drying about 100g of shredded cassava mixed from 8 peeled roots as previously
210 described (Rabbi et al. 2017). Total carotenoids were measured by spectrophotometry using the
211 iCheck™ Carotene device (TCICHK) from BioAnalyt (Teltow, Germany) as previously described
212 (Rodriguez-Amaya and Kimura 2004; Udoh et al. 2017; Jaramillo et al. 2018). Briefly, a blended
213 cassava root sample was mixed with an alcohol-based extraction reagent and placed into the
214 iCheck™ spectrophotometer to estimate the concentration of total carotenoids (in μg/g fresh
215 weight (FW)) based on light absorbance. This method showed a high correlation ($r^2 = 0.98$, $p <$
216 0.001) with HPLC measurement in a previous study (Jaramillo et al. 2018). Carotenoid content was
217 also rated by visual color evaluation using a color chart (TCHART) on an ordinal scale from 1
218 (white) to 7 (deep yellow) (Rabbi et al. 2017). Additional agronomic traits measured for at least
219 300 accessions were selected for association analyses (Table 1).
220 The phenotypic data were analyzed in a two-step approach to account for environmental
221 effects before running the GWAS. Phenotypic values were transformed using the Box-Cox
222 transformation for normality using the R function *boxcox,* using the optimum λ value for each trait
223 (Box and Cox 1964). The data were then centered and scaled by dividing by the traits' standard
224 deviations. Best linear unbiased predictions (BLUPs) of the effect of genotype on each trait and
225 variance components were then calculated using the following mixed model fit with ASReml-R
226 version 4.1.0.110 (The VSNi Team 2023).
227
228 Eqn. 1). $$\boldsymbol{y} \sim \mu + X_{l:y}\boldsymbol{\beta_1} + Z_g\boldsymbol{U_1} + Z_{g:l:y}\boldsymbol{U_2} + Z_{t:b}\boldsymbol{U}_3 + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is a vector (n x 1) of Box-Cox normalized trait values for n observations; μ is the overall population mean; $X_{l:y}$ is a design matrix (n x n_{ly}) indicating the year within location for each observation; $\boldsymbol{\beta_1}$ is a vector (n_{ly} x 1) of fixed effect estimates for year nested within location; Z_g is a design matrix (n x n_g) indicating the genotype for each observation; $\boldsymbol{U_1}$ is a vector (n_g x 1) of predicted random effects of genotype for n_g genotypes, assumed to be normally distributed $\boldsymbol{U_1} \sim N(0, {\sigma_g}^2)$ with genetic variance ${\sigma_g}^2$; $Z_{g:l:y}$ is a design matrix (n x n_{gly}) indicating the genotype, location, and year combination for each observation; $\boldsymbol{U_2}$ is a vector (n_{gly} x 1) of predicted random effects of the interaction of genotype with year/location combination, assumed to be normally distributed $\boldsymbol{U_2} \sim N(0, {\sigma_{gly}}^2)$ with genotype by environment variance ${\sigma_{gly}}^2$; $Z_{t:b}$ is a design matrix (n x n_{tb}) indicating the block number within trial for each observation; $\boldsymbol{U_3}$ is a vector (n_{tb} x 1) of predicted random effects for block within trial, assumed to be normally distributed $\boldsymbol{U_3} \sim N(0, {\sigma_{tb}}^2)$ with block variance ${\sigma_{tb}}^2$; and $\boldsymbol{\epsilon}$ is the vector (n x 1) of residual effects, assumed to be normally distributed $\boldsymbol{\epsilon} \sim N(0, {\sigma_\epsilon}^2)$ with residual variance ${\sigma_\epsilon}^2$.

Outliers were identified and removed based on studentized deleted residuals with a Bonferroni corrected α = 0.05 (Kutner et al. 2005). Then, Eqn. 1 was re-fitted without the outliers (4 observations excluded for TCICHK; 3 for TCHART; and 6 for DM). Finally, BLUPs were de-regressed by dividing by their prediction reliability (1 - $\frac{PEV}{\sigma_g^2}$) as described in Garrick et al., 2009. The broad-sense heritability (H^2) of each trait was calculated as follows: $H^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \frac{\hat{\sigma}_{gly}^2}{\bar{n}_{ly}} + \frac{\hat{\sigma}_\epsilon^2}{\bar{n}_r}}$, where $\hat{\sigma}_g^2$ is the estimated genetic variance, $\hat{\sigma}_{gly}^2$ is the estimated genetic by environment variance, $\hat{\sigma}_\epsilon^2$ is the estimated residual variance, $\bar{n}_{ly}$ is the harmonic mean number of location/year combinations per accession, and $\bar{n}_r$ is the harmonic mean number of plots per accession across locations, years, and replications (Holland et al. 2002).

Statistical methods for genome-wide associations

Genome-wide associations for the bivariate phenotype were tested using the following mixed linear model fit for each SNP*:*

Eqn. 2). $$(\boldsymbol{y_1}, \boldsymbol{y_2}) \sim \boldsymbol{\mu} + \boldsymbol{X_i\beta_1} + \boldsymbol{P\beta_2} + Q\boldsymbol{\beta_3} + ZU + \epsilon,$$

where $(\boldsymbol{y_1}, \boldsymbol{y_2})$ is a matrix (n x 2) of normalized, scaled, and de-regressed TCICHK and DM BLUP

values for n genotypes; $\boldsymbol{\mu}$ is a vector (2 x 1) of overall means for each trait; $\boldsymbol{X_i}$ is a vector (n x 1) of marker dosages (0, 1, 2) for SNP i for n genotypes; $\boldsymbol{\beta_1}$ is a vector (1 x 2) of estimates for the additive fixed effect of SNP i on each trait; $\boldsymbol{P}$ is a vector (n x 1) indicating the imputed genotype at the *PSY2* causal SNP, S1_24155522; $\boldsymbol{\beta_2}$ is a vector (1 x 2) of estimates for the additive fixed effect of the *PSY2* SNP on each trait; Q is a matrix (n x 5) of the first 5 PCs for the n genotypes; $\boldsymbol{\beta_3}$ is a matrix (5 x 2) of fixed effect estimates for the first 5 PCs; Z is an identity matrix (n x n) identifying the genotype, U is a matrix (n x 2) of the genotype random effect on each trait with matrix normal distribution $U \sim N\left(\begin{smallmatrix}0\\0\end{smallmatrix}, G, F\right)$, with covariance across rows specified by G, the relationship matrix (n x n) for the genotypes calculated with genome-wide SNPs excluding the *PSY2* SNP, and covariance across columns specified by F, the additive genetic variance-covariance matrix (2 x 2) for the two traits after accounting for the SNP fixed effects; ϵ is a matrix (n x 2) of residual effects for each trait, with matrix normal distribution $\epsilon \sim N\left(\begin{smallmatrix}0\\0\end{smallmatrix}, I, E\right)$ with covariance across rows I, the identity matrix (n x n), and covariance across columns E, the variance-covariance matrix (2 x 2) of the residual effects for the two traits. This equation was fit for each SNP in *ASRemlR* with the following statements:

fixed = cbind(TCICHK, DM) ~ trait + trait:SNP + trait:PSY2 + PC1 + PC2 + PC3 + PC4 + PC5,

random = ~ corgh(trait):vm(genotype, Ginverse),

residual = ~ corgh(units).

Genome-wide associations for univariate phenotypes were tested using the following mixed linear model fit for each SNP:

Eqn. 3). $$\boldsymbol{y_j} \sim \mu + \boldsymbol{X_i}\beta_1 + \boldsymbol{P}\beta_2 + Q\boldsymbol{\beta_3} + Z\boldsymbol{U} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y_j}$ is a vector (n x 1) of normalized, scaled, and de-regressed BLUP values for trait j for n genotypes, μ is the overall trait mean; β_1 is the estimated additive fixed effect of SNP i; β_2 is the estimated additive fixed effect of the *PSY2* SNP on trait j (only included for traits TCICHK, TCHART, and DM); $\boldsymbol{\beta_3}$ is a vector (5 x 1) of fixed effect estimates for the first 5 PCs; $\boldsymbol{U}$ is a vector (n x 1) of predictions for the genotype random effect, assumed normally distributed $\boldsymbol{U} \sim N(0, G)$; $\boldsymbol{\epsilon}$ is a vector (n x 1) of residual effects, assumed normally distributed $\boldsymbol{\epsilon} \sim N(0, \sigma_e^2)$; σ_e^2 is the residual variance; and all other variables are as specified above. These models were fit for each SNP and each trait in *ASRemlR* with the following statements:

fixed = trait ~ 1 + SNP + PSY2 + PC1 + PC2 + PC3 + PC4 + PC5,

random = ~ vm(genotype, Ginverse),

residual = ~ units.

The statistical significance of SNP effects was tested with Wald tests using the R function *aod::wald.test.* SNP effect estimates, and their variance-covariance matrix in the case of the bivariate associations, were extracted from *asreml.predict.* A modified Bonferroni cut-off for α = 0.05 was calculated based on the estimated number of independent tests as described in Gao et al., (2008). The number of eigenvalues explaining 99.5% of the variation in the SNP correlation matrix per chromosome totaled to 4661, which was used as the denominator in the Bonferroni calculation to establish a significance threshold at $p < 1.07 \times 10^{-5}$.

A relaxed threshold at $p < 0.001$ was used to minimize type II error when selecting SNPs for downstream analyses on effect direction. This relaxed threshold enables the examination of patterns in small-effect loci that could not be identified among the small set of significantly-associated SNPs. While there are likely some false-positive SNPs in this set, we assume they add neutral noise around biological patterns. SNPs with the lowest p-value within a region of high LD ($R^2 > 0.75$) were selected to represent independent peaks passing the relaxed threshold. A binomial test was used to determine whether these suggestively-associated SNPs were more likely to affect TC and DM in opposite directions than would be expected by chance. The proportion variance explained by each associated SNP *i* was calculated as follows: $PEV_i = \frac{\beta_i^2 * 2p * (1-p)}{\hat{\sigma}_y^2}$, where p is the allele frequency of the alternate allele and $\hat{\sigma}_y^2$ is the estimated total phenotypic variance.

Model comparison

Models specified in Eqn. 2 and 3 were compared with the following mixed linear models using one trait as a covariate for the other:

Eqn. 4). $$\boldsymbol{y_1} \sim \mu + \boldsymbol{X_i}\beta_1 + P\boldsymbol{\beta_2} + Q\boldsymbol{\beta_3} + \boldsymbol{y_2}\beta_4 + Z\boldsymbol{U} + \boldsymbol{\epsilon}$$

Eqn. 5). $$\boldsymbol{y_2} \sim \mu + \boldsymbol{X_i}\beta_1 + P\boldsymbol{\beta_2} + Q\boldsymbol{\beta_3} + \boldsymbol{y_1}\beta_4 + Z\boldsymbol{U} + \boldsymbol{\epsilon}$$

These models were fit for each of the putatively associated SNPs in *ASRemlR* with the following statements:

fixed = trait_1 ~ 1 + SNP + PSY2 + trait_2 + SNP:trait_2 + PC1 + PC2 + PC3 + PC4 + PC5,

random = ~ vm(genotype, Ginverse),

residual = ~ units.

Applying a frequentist version of the framework outlined by Stephens (2013), we used model comparison to determine whether SNPs are directly or indirectly associated with the two traits. Models using trait 2 as a covariate for trait 1 (Eqn. 4) test whether trait 1 is independent of a SNP ($\beta_1 = 0$) conditional on trait 2. If a SNP is significantly associated with the bivariate phenotype (Eqn. 2) but is conditionally independent of trait 1 given trait 2 as a covariate, it suggests that the bivariate association is mainly driven by the overall trait correlation, representing a direct relationship with trait 2 and an indirect relationship trait 1 (Fig. 1). Conversely, a SNP that is still associated with trait 1 after accounting for trait 2 as a covariate suggests that the SNP has a direct relationship with both traits. The reasoning is the same using trait 1 as a covariate for trait 2 (Eqn. 5). These possible pleiotropic relationships are summarized in Fig. 1.

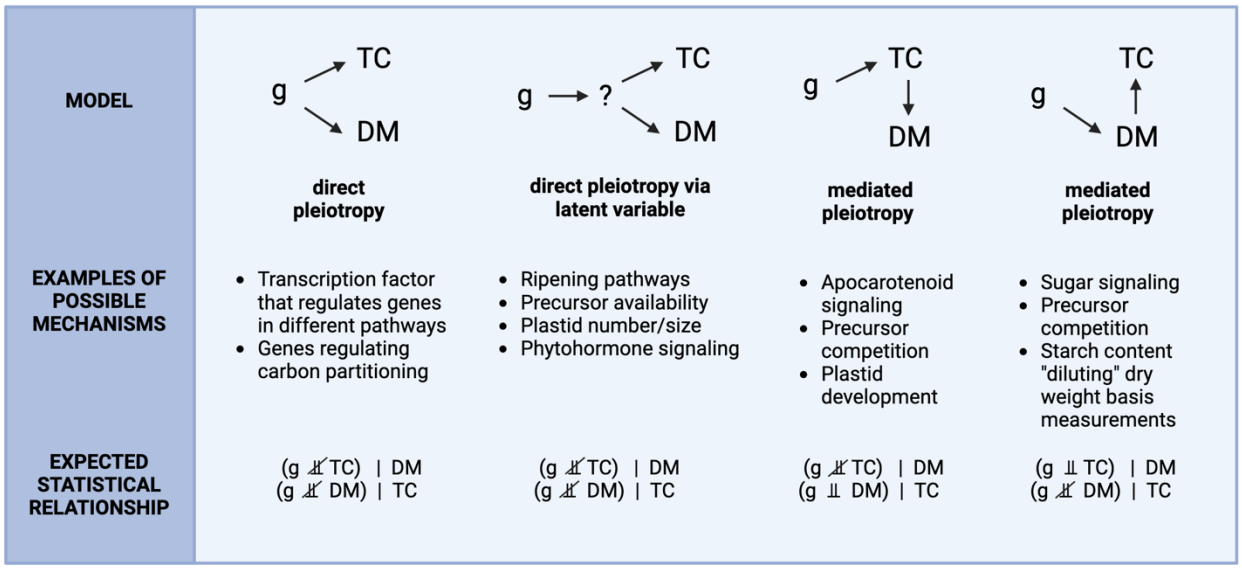


Figure 1. Illustration of direct and mediated pleiotropic relationships and their expected statistical associations, where g ⫫ A | B denotes conditional independence of a genetic variant *g* with trait A given trait B. Hypothesized biological mechanisms are described in detail in Villwock et al. (2024). Created with Biorender.com.

Estimates of polygenic contribution to negative genetic covariance between TCICHK and DM

The residual genetic covariance of TCICHK and DM after accounting for major SNP effects and the *PSY2* causal variant was examined to assess the contribution of causal loci too small to detect by GWAS. The variance-covariance matrix *F* was extracted from a bivariate model including all *z* peak SNPs identified at the relaxed significance threshold, as described in Eqn. 2 except with $\sum_{i=1}^{z} \boldsymbol{X_i}\beta_i$ in place of $\boldsymbol{X_i}\beta_1$. The variance components in *F* therefore represent the polygenic

347 variance-covariance remaining after accounting for the identified QTL. If these major SNPs
348 completely accounted for the negative genetic covariance between traits, then we would expect the
349 residual genetic covariance to be near zero. A likelihood ratio test comparing models with the
350 residual genetic covariance between traits set to zero or unconstrained was used to test whether
351 there was significant residual genetic covariance after accounting for associated SNPs.
352
353 *Candidate gene analysis*
354 *M. esculenta* reference genome v6.1 gene annotation data were downloaded from the JGI
355 data portal (https://data.jgi.doe.gov/). A list of *a priori* candidate genes involved in the carotenoid
356 synthesis (MetaCyc pathway ontology ID PWY-6475), apocarotenoid synthesis (PWY-6806), starch
357 synthesis (PWY-622), and starch degradation (PWY-6724) pathways annotated in *Manihot*
358 *esculenta v6.1* on Phytozome was assembled (Table S2) (Goodstein et al. 2012; Bredeson et al.
359 2016; Caspi et al. 2020). *A priori* candidate genes were screened for LD and physical distance to
360 significantly associated SNPs. Then, all annotated genes within 100kb of the peak SNPs were
361 extracted from the genome annotation file. Genes with missing ontology annotations were
362 supplemented with annotations from the UniProt Knowledgebase where possible (Bateman et al.
363 2023).
364
365 **Results**
366 *Trait relationships*
367 In this dataset, TCICHK and DM had moderately high broad-sense heritability of 0.49 and
368 0.50, respectively. The two traits were negatively genetically correlated with $r = -0.34$ and $p = 1.84$
369 $\times 10^{-11}$ (Fig. 2) and phenotypically correlated with $r = -0.15$ and $p = 4.48 \times 10^{-22}$. TCICHK was also
370 negatively genetically correlated with several other agronomic traits: dry yield (DYLD; $r = -0.17$, $p =$
371 1.66×10^{-3}), root weight ($r = -0.18$, $p = 5.98 \times 10^{-4}$), fresh root yield (FRYLD; $r = -0.15$, $p = 8.43 \times 10^{-}$
372 ^{3}), and sprouting percentage (SPRTPERC; $r = -0.25$, $p = 1.26 \times 10^{-6}$) (Fig. 3). The broad-sense
373 heritabilities of all traits are listed in Table 1.

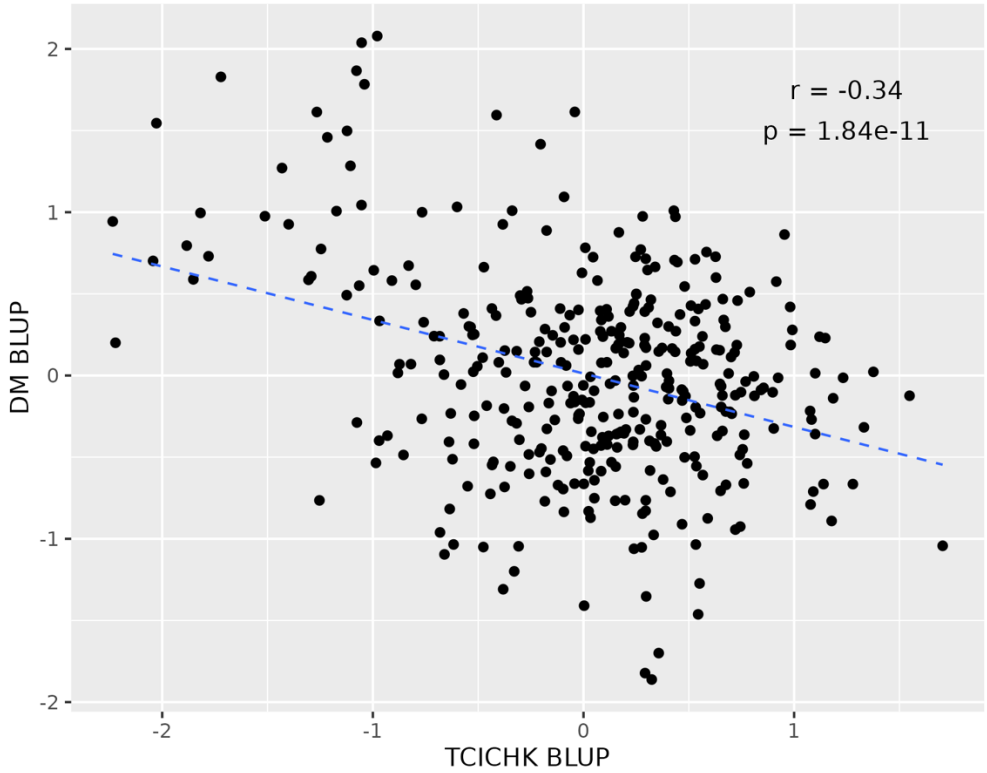


Figure 2. Negative genetic relationship between total carotenoid content as measured by iCheck™ Carotene (TCICHK) and dry matter content (DM) BLUPs among the 378 genotyped accessions after normalization of phenotypic values with Box-Cox transformation. Pearson correlation coefficient (r) and p-value are displayed at the top right.

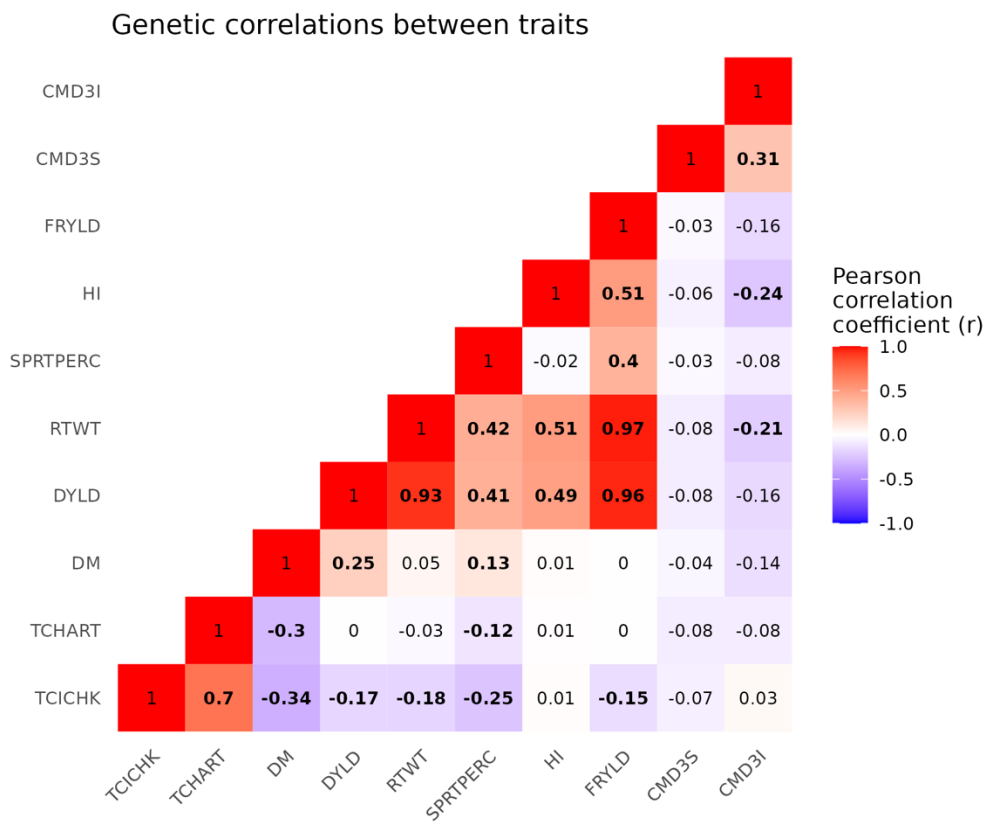


Figure 3. Pairwise genetic correlations between all analyzed traits. Each cell is labeled with the Pearson correlation coefficient (r) between the traits listed on the x and y axis. Significant correlations ($p < 0.05$) are bolded.

Table 1. Description, sample size, and broad-sense Heritability (H^2) of traits analyzed.

Trait name	**Abbreviation**	**Trait ontology ID**	**No. of genotypes**	**No. of observations**	**H^2**
Total carotenoids by iCheck™ method	TCICHK	CO_334.0000162	365	4663	0.49
Total carotenoids by color chart (1-9) method	TCHART	CO_334.0000161	374	8353	0.73
Dry matter content percentage	DM	CO_334.0000092	370	7575	0.50
Dry yield	DYLD	CO_334.0000014	332	4373	0.21
Fresh storage root weight per plot	RTWT	CO_334.0000012	373	8372	0.26
Harvest index	HI	CO_334.0000015	368	6481	0.43
Sprouting proportion	SPRTPERC	CO_334.0000008	375	7877	0.24
Fresh root yield	FRYLD	CO_334.0000013	334	5188	0.23

Cassava mosaic disease severity 3-month evaluation	CMD3S	CO_334.0000192	376	8703	0.60
Cassava mosaic disease incidence 3-month evaluation	CMD3I	CO_334.0000196	376	2904	0.49

386

387 *Bivariate genome-wide associations for TCICHK and DM*

388 The Manhattan plot from the bivariate associations reveals 13 peaks that pass the relaxed
389 threshold ($p < 0.001$), including two QTL peaks that pass the modified Bonferroni significance
390 threshold for joint association with both TCICHK and DM after controlling for the effect of *PSY2*
391 (Fig. 4). The significant joint associations are located on chromosomes 1 and 10. The top SNP on
392 chromosome 1 is located at 26.700297 Mbp, which is 2.5 Mbp away from the *PSY2* causal variant at
393 24.155522 Mbp. The two SNPs are in low LD with each other ($R^2 = 0.015$). Fig. 5 illustrates the
394 results from bivariate and univariate associations in this QTL region in higher resolution. The top
395 jointly-associated SNP on chromosome 1 in the bivariate associations is significantly associated
396 with DM but not with TCICHK in univariate analyses. The percent variance explained (PVE) by the
397 peak SNPs were relatively small ($PEV_{TCICHK} = 0.1\%$ and $PEV_{DM} = 8\%$ for the chromosome 1 peak;
398 $PEV_{TCICHK} = 6\%$ and $PEV_{DM} = 0.01\%$ for the chromosome 10 peak). Summary statistics for all
399 associated SNPs are listed in Table 2.

400 Of the 13 QTL identified at the relaxed significance threshold, 11 peak SNPs had opposite
401 effects on TC and DM, while two SNPs had effects on TC and DM in the same direction. A binomial
402 test with these counts resulted in a p-value of 0.011, rejecting the null hypothesis that SNPs are
403 equally likely to affect TC and DM in same or opposite directions. A likelihood ratio test indicated
404 there was significant negative genetic covariance between TCICHK and DM remaining after
405 accounting for all 13 of these SNPs and *PSY2* as covariates in a linear model (LR-statistic = 3.80, p =
406 0.026). Across all SNP association models fit in the bivariate GWAS, the residual genetic covariance
407 was consistently negative, ranging from -0.51 to -0.19.

408 Since the phenotypes were normalized with a Box-Cox transformation before analysis,
409 bivariate association models (Eqn. 2) were re-fit with untransformed, de-regressed BLUPs for the
410 13 peak SNPs to translate their effect size estimates to the traits' original units. For both TCICHK
411 and DM, effect estimates from untransformed data were linearly related to effect estimates from
412 Box-Cox-transformed data with slopes of 5.20 and 6.66, respectively (Fig. S3). SNP effect estimates
413 from the untransformed BLUPs ranged from -1.82 to 2.63 μg/g FW for TCICHK and -2.58 to 2.08%
414 for DM.

415 **Table 2.** Summary statistics of SNPs that are jointly-associated with TCICHK and DM. The SNP with
416 the lowest p-value was selected from each group of associated SNPs. P-values are annotated with
417 two asterisks if they pass the modified Bonferroni significance threshold and a single asterisk if
418 they pass the relaxed threshold. SNP effect estimates (additive effect of the ALT allele) and standard
419 errors are with respect to the transformed, de-regressed BLUP values.

Peak SNP (chr#_ position)	**Bivariate association p-value**	**TCICHK effect estimate**	**TCICHK effect SE**	**DM effect estimate**	**DM effect SE**	**Effect direction (TCICHK/ DM)**	**REF allele**	**ALT allele**	**Allele frequency (ALT)**
S10_2262535	7.74E-07**	0.506	0.097	-0.022	0.100	+/-	A	C	0.94
S1_26700297	1.46E-06**	0.053	0.078	-0.389	0.075	+/-	C	T	0.88
S1_23484417	2.99E-05*	0.324	0.091	-0.307	0.091	+/-	C	G	0.92
S11_25505603	6.02E-05*	-0.284	0.070	0.178	0.073	-/+	A	G	0.11
S15_15230957	8.00E-05*	-0.356	0.096	0.272	0.096	-/+	C	T	0.05
S15_1902620	8.44E-05*	0.045	0.071	0.281	0.069	+/+	G	A	0.16
S4_22936714	1.95E-04*	-0.135	0.047	0.153	0.045	-/+	G	T	0.26
S13_23695071	1.99E-04*	-0.216	0.074	0.241	0.072	-/+	A	T	0.10
S13_5292787	2.35E-04*	0.089	0.082	0.294	0.081	+/+	A	G	0.09
S3_3963505	2.66E-04*	0.283	0.073	-0.135	0.074	+/-	G	A	0.90
S14_3992190	3.23E-04*	-0.274	0.081	0.224	0.086	-/+	A	T	0.11
S5_17845586	4.79E-04*	0.130	0.054	-0.188	0.055	+/-	G	A	0.65
S9_4599621	8.61E-04*	-0.208	0.075	0.225	0.076	-/+	G	C	0.89

420

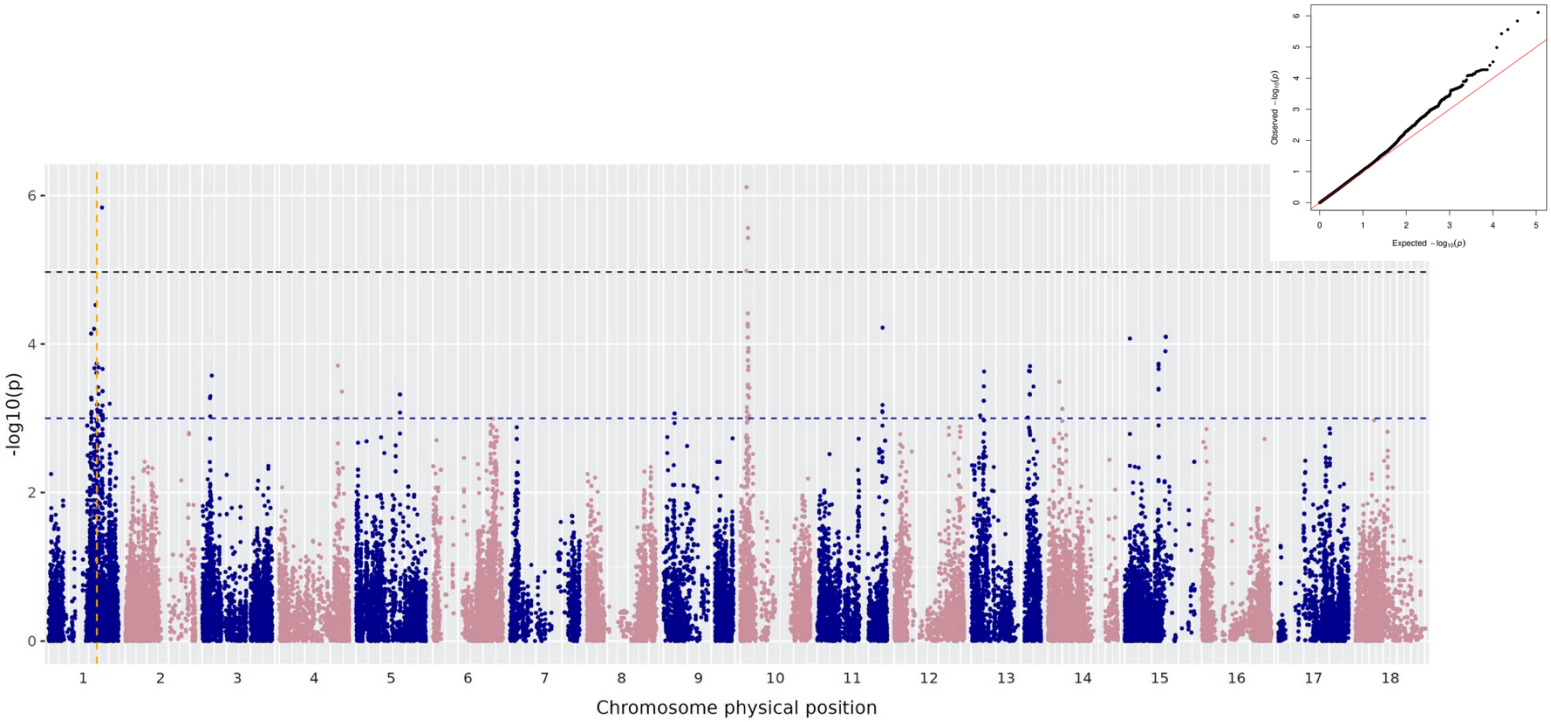
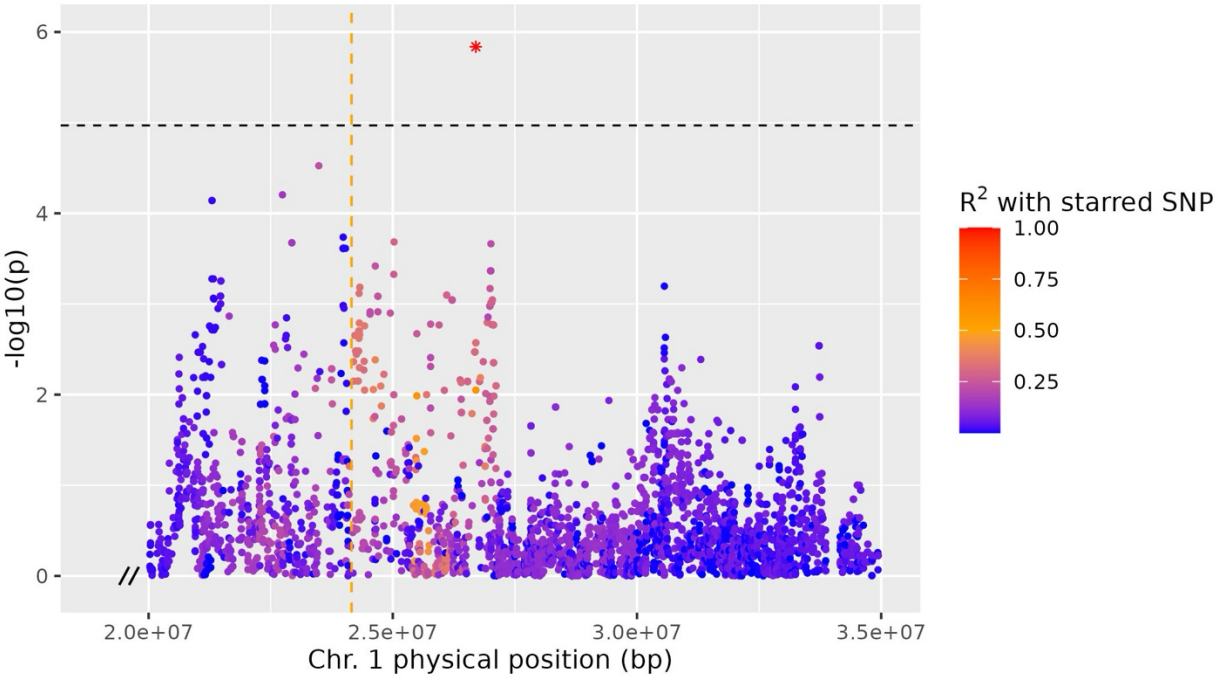
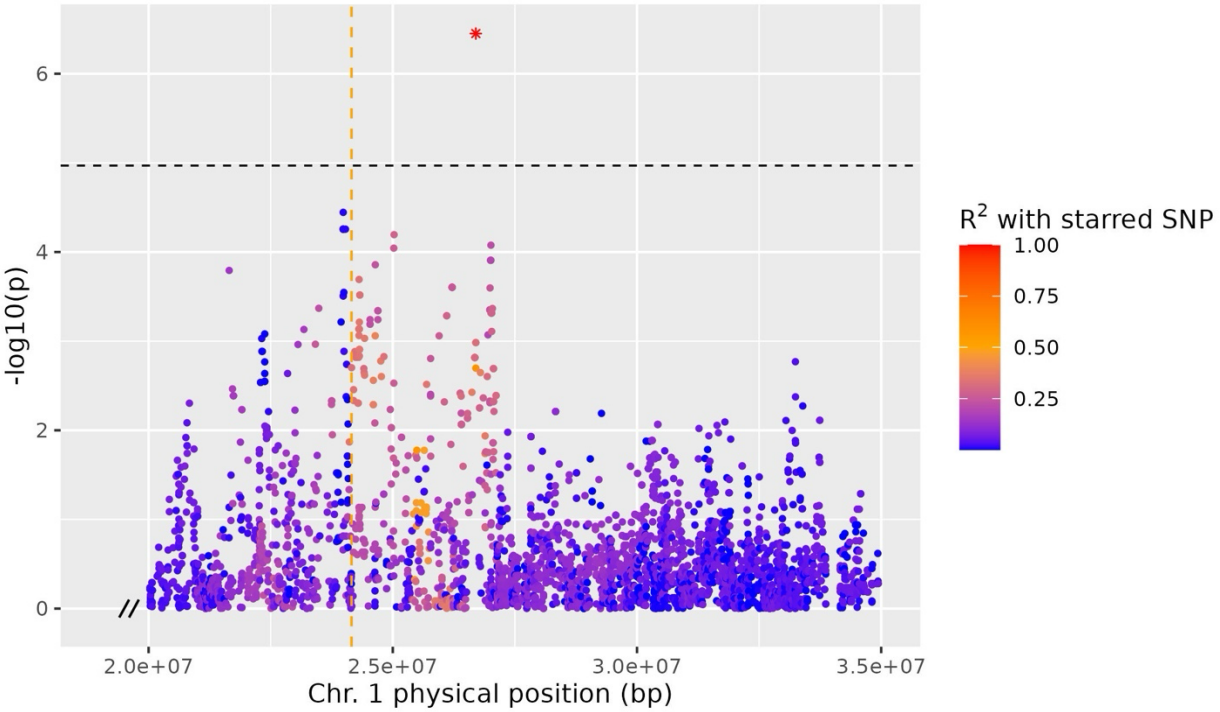


Figure 4. Manhattan plot showing p-values from bivariate Wald tests of SNP effects on TCICHK and DM, accounting for genomic relatedness and the *PSY2* causal SNP. The vertical orange dashed line indicates the position of the *PSY2* causal SNP. The black dashed line indicates the modified Bonferroni significance cut-off at $\alpha = 0.05$. The blue dashed line indicates a relaxed threshold of $p < 0.001$ used for downstream analyses. At top right, the Q-Q plot of expected versus observed p-values shows acceptable control of $\log_{10}$ p-value inflation.

A. Bivariate GWAS for TCICHK and DM



B. Univariate GWAS for DM



C. Univariate GWAS for TCICHK

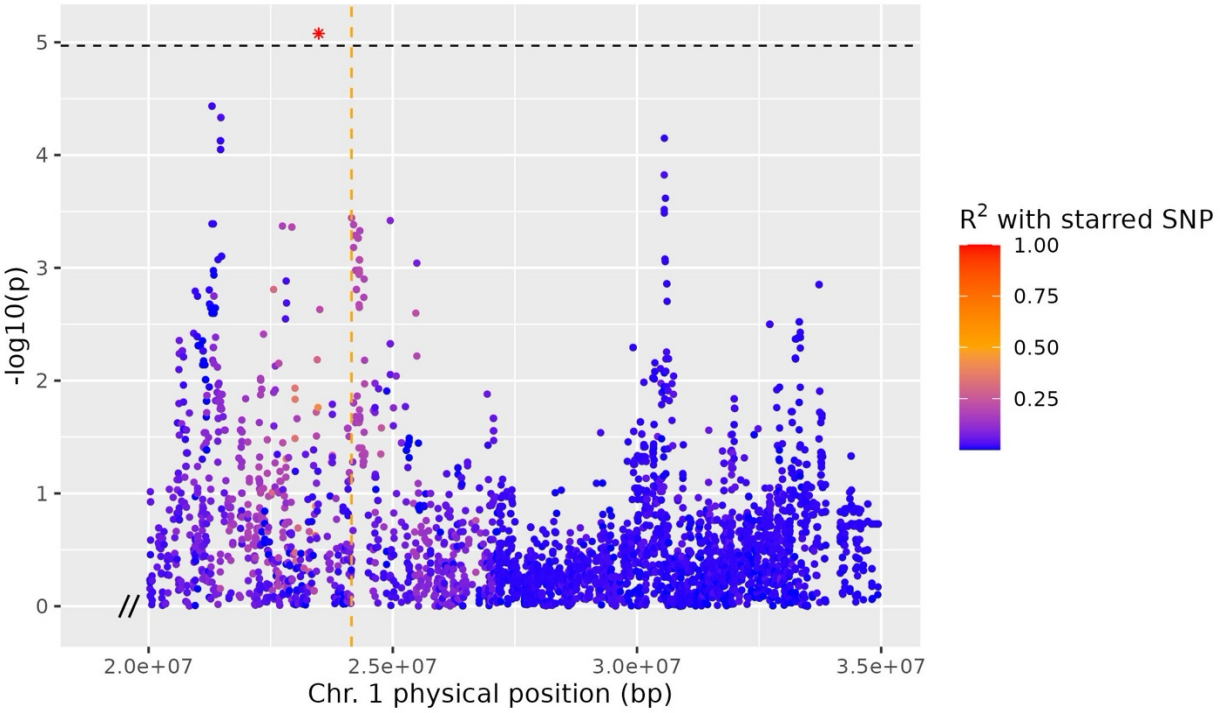


Figure 5. Manhattan plots for A) bivariate TCICHK and DM, B) univariate DM, and C) univariate TCICHK associations on chromosome 1 near the *PSY2* QTL region from 20–35 Mbp. The vertical orange line indicates the position of the *PSY2* causal SNP. Points are colored by their LD (pairwise R^2) with the top SNP, marked with an asterisk, which differs across association models. The black dashed line indicates the modified Bonferroni significance cut-off at $\alpha = 0.05$.

Model comparison

The bivariate association model was compared with univariate models with one trait as the response variable and the other as a covariate to determine whether the associations were direct or indirect. For each of the top SNPs from 13 peaks that passed the relaxed significance threshold, the bivariate model had the lowest AIC compared to the univariate models (Table 3). For 8 of the 13 SNPs, the SNP effect retained its significance at the $p < 0.001$ threshold in a univariate model for one trait with the other as a covariate. For five of these SNPs, the TCICHK univariate model retained its significance, indicating support for a direct association with TC and indirect association with DM. The remaining three SNPs retained their significance in the DM univariate model, suggesting a direct association with DM and an indirect association with TC. In no cases did the SNP retain significance in both univariate models with the other trait as a covariate.

Table 3. Comparison of bivariate and reciprocal univariate association models for the 13 peak SNPs identified as jointly associated at the relaxed threshold ($p < 0.001$). "TCICHK model" refers to an association between a SNP and TCICHK with DM as a covariate (Eqn. 4), while the "DM model" is the reverse (Eqn. 5). P-values are from univariate or bivariate Wald tests of SNP effects. P-values are annotated with two asterisks if they pass the modified Bonferroni significance threshold and a single asterisk if they pass the relaxed threshold.

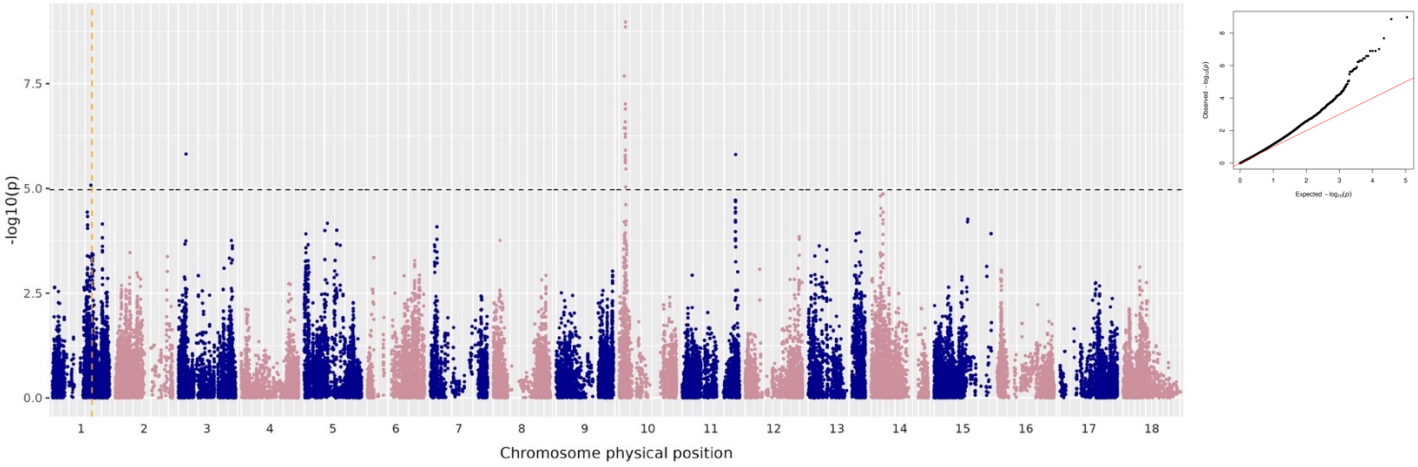
Peak SNP (chr#_ position)	**Bivariate model AIC**	**TCICHK model AIC**	**DM model AIC**	**Bivariate model p-value**	**TCICHK model p-value**	**DM model p-value**	**Supported model**
10_2262535	-39.95	-7.70	4.20	7.74E-07**	1.95E-08**	4.63E-01	Direct TC, indirect DM
1_26700297	-38.24	23.05	-14.16	1.46E-06**	9.34E-01	3.42E-07**	Direct DM, indirect TC
1_23484417	-31.46	14.28	-0.79	2.99E-05*	3.02E-04*	5.92E-03	Direct TC, indirect DM
11_25505603	-31.28	5.92	-0.25	6.02E-05*	6.89E-06**	5.83E-02	Direct TC, indirect DM
15_15230957	-30.26	7.14	1.96	8.00E-05*	3.70E-04*	4.14E-02	Direct TC, indirect DM
15_1902620	-29.86	14.73	-9.00	8.44E-05*	2.40E-01	3.08E-05*	Direct DM, indirect TC

4_22936714	-25.10	20.14	3.04	1.95E-04*	3.02E-02	5.63E-03	NA
13_23695071	-27.19	17.92	1.95	1.99E-04*	1.37E-02	6.96E-03	NA
13_5292787	-28.55	19.23	-6.99	2.35E-04*	1.04E-01	1.06E-04*	Direct DM, indirect TC
3_3963505	-26.00	6.17	3.35	2.66E-04*	2.52E-05	2.62E-01	NA
14_3992190	-25.37	16.24	0.43	3.23E-04*	5.53E-04*	5.75E-02	Direct TC, indirect DM
5_17845586	-24.36	16.93	-5.56	4.79E-04*	1.52E-02	3.61E-03	NA
9_4599621	-23.88	17.01	2.90	8.61E-04*	3.68E-02	1.47E-02	NA

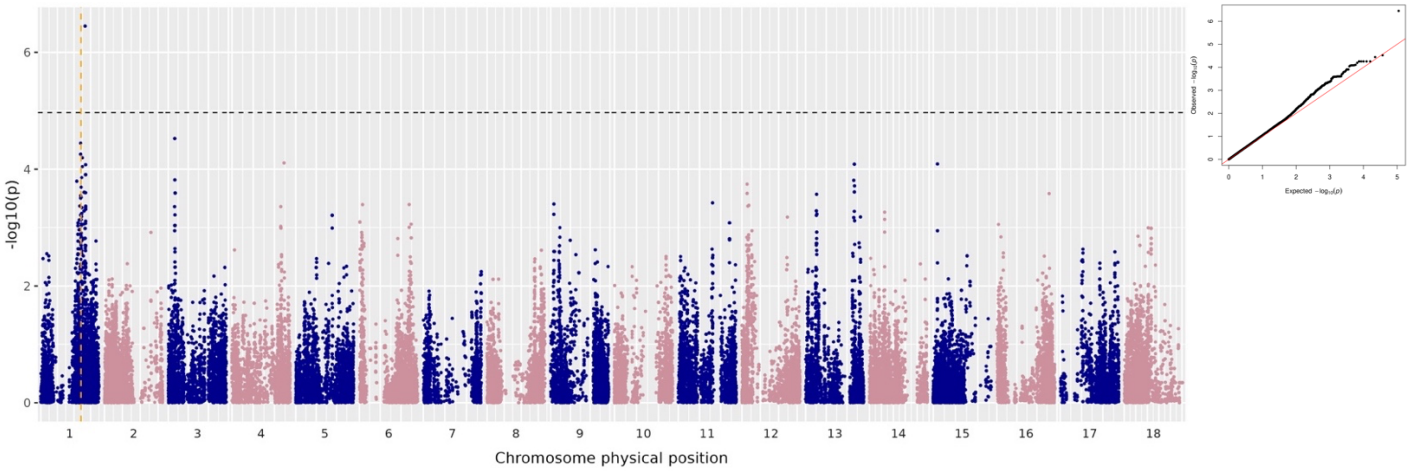
Univariate associations for all traits

Several peaks passing the modified Bonferroni significance threshold were identified in univariate analyses: TCICHK (4), DM (1), TCHART (6) (Fig. 6). TCICHK and TCHART QTL overlapped on chromosomes 3, 10, and 11. There were no overlapping univariate QTL between TCICHK or TCHART and DM. The peak SNP positions and summary statistics for univariate QTL are reported in Table S3. Among other agronomic traits analyzed (DYLD, FRYLD, RTWT, HI, SPRTPERC, CMD3I, and CMD3S), 5 other QTL were identified at the Bonferroni significance level (Fig. S4).

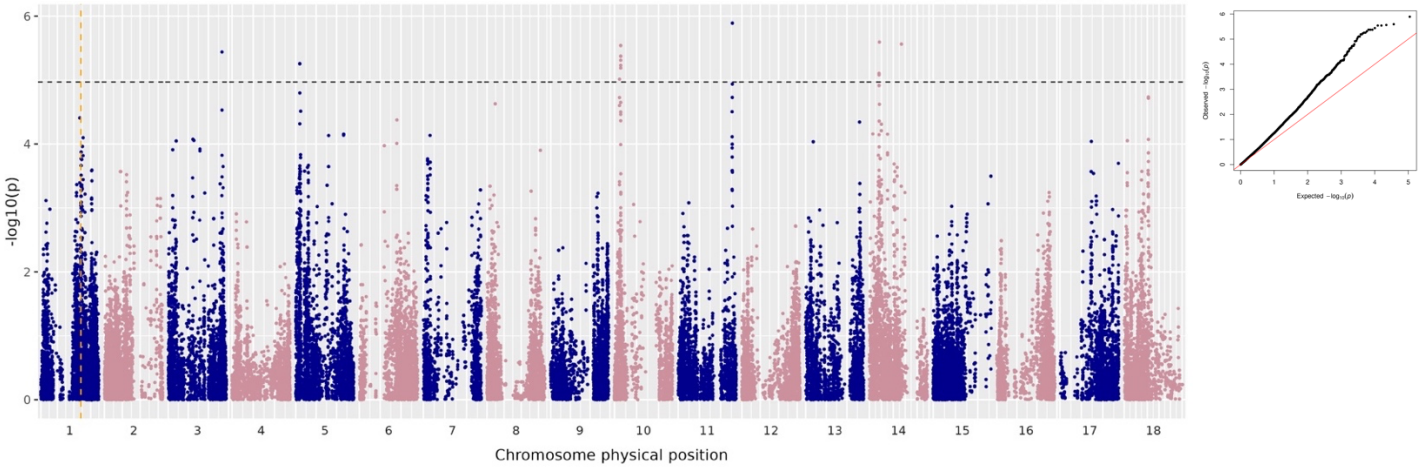
A. TCICHK



B. DM



C. TCHART



473 **Figure 6.** Manhattan plots for univariate GWAS for TCICHK (A), DM (B), and TCHART (C). The
474 dashed black line indicates the modified Bonferroni significance threshold at α = 0.05. The vertical
475 dashed orange line indicates the position of the *PSY2* causal locus used as a covariate. Q-Q plots for
476 each analysis are at right.
477
478 *Candidate gene analysis*
479 Annotated genes within a window of 100kb from the bivariate peak were identified as
480 candidates based on the rate of LD decay. On most chromosomes, LD decayed to an average R^2 of
481 0.25 at approximately 100kb physical distance, although chromosomes 1 and 4 had more extensive
482 LD, likely due to the presence of large *M. glaziovii* introgressions on these chromosomes (Bredeson
483 et al. 2016; Wolfe et al. 2019; Chan et al. 2022) (Figure S2).
484 Candidate genes within 100kb of bivariate peak SNPs are listed in Table S4. None of the *a*
485 *priori* candidate genes were identified in this list, although two genes were related to carbohydrate
486 metabolism more broadly. *Rubisco activase*, central to carbon metabolism via its key role in
487 photosynthesis, was identified 17kb away from the peak SNP for joint TC and DM on chromosome
488 10. *Alpha-galactosidase,* involved in stachyose degradation, was identified 52kb away from the
489 bivariate peak SNP on chromosome 1. Candidate genes for other agronomic traits (DYLD, FRYLD,
490 RTWT, HI, SPRTPERC, CMD3I, and CMD3S) from univariate analyses are listed in Table S5.
491
492 **Discussion**
493 The main objective of this study was to test the hypothesis of mediated pleiotropy between
494 TC and DM in cassava by examining genome-wide contributions to the negative covariance between
495 traits. Bivariate mixed linear models were used to estimate the additive effects of each SNP on TC
496 and DM while controlling for the effects of *PSY2* and genome-wide population structure. Four main
497 points of evidence for mediated pleiotropy were considered: whether there are multiple QTL jointly
498 associated with both traits, whether QTL tend to affect the traits in opposite directions, whether
499 QTL affect the traits in a direct or indirect manner, and the extent of residual genetic covariance
500 between traits after accounting for identified QTL.
501 The bivariate genome-wide associations identified two QTL passing the modified
502 Bonferroni significance threshold at α = 0.05 on chromosomes 1 (S1_26700297) and 10
503 (S10_2262535) for joint association with TCICHK and DM. To our knowledge, the QTL on
504 chromosome 10 jointly associated with TC and DM has not previously been reported. Rabbi et al.
505 (2017) show a single significant SNP associated only with TCHART on chromosome 10, but it does

506 not appear to be in the same location as the QTL peak identified here. These associations therefore
507 represent new observations of apparently pleiotropic QTL.
508 The incorporation of more diverse, early-stage selections within the yellow cassava
509 breeding population likely facilitated the identification of these new loci. None of the TCICHK or
510 TCHART univariate association peak SNPs overlapped with the DM univariate association peaks,
511 despite the presence of two significant bivariate TCICHK and DM peaks. This observation suggests
512 that the increased power of the bivariate association model facilitated the identification of these
513 joint associations which would have been missed if the two univariate associations were simply
514 compared.
515 The estimated effect sizes of the chromosome 1 and 10 loci correspond to +0.13 μg/g FW
516 TC / -2.58 % DM and +0.51 μg/g FW TC / -0.016 % DM, respectively (Fig. S3). While these effect
517 sizes are large enough to potentially warrant marker-assisted selection strategies for individual
518 traits, since these loci affect the traits in opposite directions, a genomics-assisted breeding method
519 incorporating information from many markers would likely be more useful. In addition, the allele
520 frequencies for both loci are already quite high (Table 2).
521 Eleven additional loci passed the relaxed significance threshold of $p < 0.001$ for joint
522 association with TCICHK and DM. After accounting for all 13 of these SNPs and *PSY2* as covariates in
523 a linear model, there was still statistically significant negative genetic covariance remaining
524 between TCICHK and DM. Therefore, we conclude that there are significant polygenic contributions
525 to the genetic correlation between TCICHK and DM, disrupting the hypothesis that *PSY2* is the main
526 driver of the relationship.
527 To infer whether these joint associations were direct or mediated, we used model
528 comparisons to test whether jointly-associated SNPs retained their associations with each
529 individual trait after accounting for the other trait as a covariate. For the chromosome 10 QTL, the
530 top SNP (S10_2262535) retained significance in the univariate model for TCICHK after accounting
531 for DM, but not in the univariate model for DM accounting for TCICHK. This result suggests that the
532 chromosome 10 QTL is more likely to be directly associated with TCICHK and indirectly associated
533 with DM. On chromosome 1, S1_26700297 retained significance with DM after accounting for
534 TCICHK but not vice versa, suggesting a direct association with DM and indirect association with
535 TCICHK. Among all the suggestively-associated SNPs, the model comparison analysis supported a
536 direct relationship with TCICHK for 5 SNPs and a direct relationship with DM for 3 SNPs, while no
537 SNPs seemed to be directly associated with both traits simultaneously. These results support the

538 hypothesis of mediated rather than direct pleiotropy (Fig. 1), though there is not clear evidence for
539 the directionality of the mediating mechanism.
540 Since there is similarly strong evidence for TCICHK and DM as the mediating trait, these
541 results suggest that the relationship between traits can go both ways. Such bidirectional
542 mechanisms have been proposed, including the hypothesis of a trade-off for carbon precursors
543 between starch and carotenoid synthesis, in which any variants upregulating metabolic flux
544 towards one pathway would take away from the other pathway (Villwock et al. 2024). In addition,
545 these data would support the hypothesis of a direct pleiotropy mechanism mediated by a latent
546 variable such as a phytohormone signal, in which case it would be difficult to distinguish between
547 direct/indirect and indirect/direct statistical associations (Fig. 1) (Stephens 2013).
548 Similarly, the lack of *a priori* candidate genes identified among the significant bivariate GWAS hits
549 supports that the QTL identified here play a regulatory role rather than direct involvement in
550 carotenoid or starch synthesis/catabolism pathways. This finding supports the hypothesis that the
551 interaction between carotenoids and starch is mediated indirectly through regulatory pathways or
552 competition for precursors rather than direct pleiotropic action of genes in the biosynthetic
553 pathways (Villwock et al. 2024). The two previously identified candidate genes for DM, *sucrose*
554 *synthase* and *ADP-glc PPase***,** did not show up as GWAS hits in this study, either because they were
555 not causally polymorphic or because their effects were statistically eliminated by the *PSY2*
556 covariate used in the association models.
557 Of the identified candidate genes (Table S3), several have potential biological relevance to
558 the TCICHK and DM relationship. The three nearest genes to the chromosome 1 peak SNP
559 (*WDSAM1/*Manes01.G160500, *SELENOPROTEIN T*/Manes01G160400, and *PHF5-like*/
560 Manes01G160200) have not been well-characterized but all have putative regulatory roles via
561 protein ubiquitination, redox function, and mRNA splicing, respectively. Among other candidate
562 genes, Manes.01G159700, 56.7kb away from the peak 1 SNP, is a tetratricopeptide repeat protein
563 (TPR), which mediate protein-protein interactions. A TPR protein was identified in regulating
564 chromoplast development in *Mimulus* flowers (Stanley et al. 2020). In addition, AQUAPORIN *TIP3-*
565 *1-RELATED* (Manes01.G160000), 36kb from the chromosome 1 peak SNP, is involved in water
566 uptake. Water content is an indirect component of both TC and DM measured per unit fresh weight.
567 In *Arabidopsis, TIP3* expression is induced by an ABA-responsive transcription factor (Mao and Sun
568 2015). If the same is true in cassava, *AQUAPORIN TIP3* could hypothetically be involved in
569 increasing water uptake in response to higher ABA synthesis downstream of the carotenoid
570 pathway, resulting in lower dry matter percentage.

571 The peak SNP on chromosome 10 identified in both the bivariate associations and the
572 TCICHK/TCHART univariate associations falls within the *pumilio homolog 1-related* (Manes
573 10.G027800) coding region. Pumilio is a family of RNA-binding proteins that can mediate sequence-
574 specific post-transcriptional regulation (Huh 2021). The *Arabidopsis pumilio 1* (*APUM-1)* and its
575 homologs were shown to bind with mRNA transcripts containing certain binding sequences,
576 including in the 3'UTR region of a chlorophyll α/β-binding protein (Francischini and Quaggio
577 2009). In the cassava genome, the *PUM* binding sequence is found in the *ADP-glc PPase* gene
578 (BLAST), suggesting that *PUM-1* may regulate starch synthesis. Upstream 23kb of the chromosome
579 10 peak SNP is *LOB DOMAIN-CONTAINING PROTEIN 33* (Manes.10G027400). Lateral Organ
580 Boundaries domain-containing (LBD) proteins are involved in regulating plant organ development,
581 and LBD33 specifically is involved in lateral root initiation in *Arabidopsis* (Xu et al. 2016). Variants
582 in this gene could therefore impact cassava root development and metabolite accumulation.
583 Candidate gene validation is needed to test these hypotheses.
584 Finally, we tested whether QTL are more likely to contribute negatively to the covariance
585 between traits than would be expected by chance. A binomial test of effect directions on the two
586 traits for SNPs passing the relaxed significance threshold indicated that QTL are significantly more
587 likely to affect the two traits in opposite directions. Though the 13 peak SNPs used in the binomial
588 test are not exactly independent (all SNPs were identified in only one dataset), we treated them as
589 such given that they are mostly on different chromosomes and had low R^2 with each other ($R^2 \leq$
590 0.022). In setting a more relaxed threshold for analyses on effect direction, we assume that any
591 false-positive peaks would add neutral noise around a biological pattern. In this case, we assume
592 that false positives would be equally likely to show traits being affected in the same or in opposite
593 directions, whereas the biological pattern is for opposite direction effects.
594 The *PSY2* causal SNP (1_24155522) had smaller effect sizes than expected (X for TCICHK
595 and X for DM) . Since the *PSY2* causal SNP (1_24155522) was imputed with a DR2 value of 0.74, it is
596 possible there were imputation errors such that this covariate does not fully account for the *PSY2*
597 effect. Other methods of accounting for the *PSY2* effect in the association models were tested,
598 including by fitting the closest non-imputed SNP to 1_24155522 as a fixed effect covariate, using the
599 *PSY2* region (within 20kb of the coding region) haplotype as a random effect factor covariate, and
600 using the first three PCs from a PCA of the *PSY2* region as fixed effects. None of these methods
601 significantly affected the results (data not shown). While the chromosome 1 QTL at 1_26700297 is
602 relatively near to the *PSY2* causal variant at 1_24155522, it appears to be a distinct QTL. If the
603 1_26700297 QTL was driven by residual effects of the *PSY2* variant, we would expect it to present

604 strongly in the univariate analysis for TCICHK and TCHART as well, but it does not. Therefore, we
605 assume that the *PSY2* effect was effectively captured in the model.
606 Q-Q plots from bivariate genome-wide associations with TCICHK and DM showed
607 acceptable control of inflation from the genomic relationship matrix and the *PSY2* covariate. In the
608 univariate analyses, there was considerable Q-Q plot inflation for TCICHK and TCHART. Although
609 population structure and the *PSY2* were accounted for in these models, strong selection pressure
610 for root yellowness as the primary trait in the yellow cassava breeding population likely created
611 additional structure around causal loci. In practice, white-fleshed and yellow-fleshed cassava roots
612 are easily distinguished, but quantitative variation among cream-colored roots blurs the line
613 between them. After subsetting the data to retain only the yellow subpopulation, the positions of
614 Manhattan plot peaks were similar, although with generally higher p-values reflecting lower power
615 (data not shown).
616 The precision of TCICHK phenotypes did not appear to outweigh the advantage of having
617 many more TCHART observations, since TCHART had a higher H^2 (0.73) compared to that of
618 TCICHK (0.49). (Table 1). However, negative genetic relationships with yield traits (DRYLD, FRYLD,
619 RTWT) were evident with TCICHK but not with TCHART. While TCICHK and colorimetric
620 measurements are both highly correlated with HPLC measurements of TC (Jaramillo et al. 2018),
621 TCHART may be influenced by other factors that affect visual perception of root color, while
622 TCICHK may be more specific to carotenoids. The negative correlation between TCICHK and
623 SPRTPERC (sprouting percentage; $r = -0.25$, $p < 0.05$) in these data is notable. This relationship has
624 been previously observed in the field, though not formally described to our knowledge (P. Kulakow,
625 personal communication, July 2021). We hypothesize that this negative correlation could be driven
626 by the influence of carotenoid synthesis on downstream ABA, which regulates bud dormancy (Chen
627 et al. 2020).
628 This study represents a new application for multivariate GWAS that goes beyond identifying
629 individual variants to test hypotheses about trait genetic relationships and identify mediated
630 pleiotropy. It builds on previous multivariate GWAS methods that explicitly estimate SNP effects on
631 each trait, which has a built-in interpretation framework, as opposed to the difficult-to-interpret
632 single p-values given from an overall multivariate association. A downside to our method is that the
633 computational time is high, and it may not be feasible to extend to more than two traits. More
634 efficient but less flexible implementations similar to our multivariate GWAS method are available in
635 the MTMM and GEMMA softwares (Korte et al. 2012; Zhou and Stephens 2014). Multivariate
636 GEMMA output very similar GWAS results with the data used here and runs much faster, but

637 GEMMA does not offer the same flexibility in specifying and analyzing covariance structures as
638 ASReml-based models do. Future studies could improve on these methods by improving the
639 computational time while retaining the modeling flexibility.
640 Overall, the identification of polygenic mediated pleiotropy in this study demonstrates a
641 multivariate GWAS framework for investigating the genetic architecture of correlated traits,
642 providing insights to guide both applied breeding strategies and future research into the underlying
643 biological mechanisms.
644
645 **Data availability**
646 All datasets and R code used for these analyses are available on GitHub
647 (http://www.github.com/serenvillwock/mvGWAS). Raw field trial data and imputed genotype files
648 are additionally available on http://www.cassavabase.org/ under the public list
649 'yellowGWAS23_trials_plusCASS'. The reference panel data used for genotype imputation is
650 available at
651 https://www.cassavabase.org/ftp/marnin_datasets/nextgenImputation2019/ImputationStageIII_7
652 2619/. *Manihot esculenta* v6.1 reference genome and gene annotations are available on Phytozome
653 at https://phytozome-next.jgi.doe.gov/info/Mesculenta_v6_1.
654
655 **Acknowledgements**
656 We are grateful to the Cassava Breeding Unit team at the International Institute of Tropical
657 Agriculture, Ibadan, Nigeria, especially including Lynda Adaobi Nnadi, Adenike Damilola Ige, Peter
658 Iluebby, Patrick Akpotuzor, Victor Monday Chukwuyem, Toye Ayankanmi, and Kehinde Nafiu for
659 their assistance with tissue sampling and field trials; Prasad Peteti, Ekanem Ukoabasi, and Mahmud
660 Kehinde Dhikrullah for their assistance with data curation; and Peter Kulakow, Peace Ikeanyi, and
661 Richard Ofei for their administrative support. The development of the breeding populations used in
662 this study was sponsored by HarvestPlus and the NextGen Cassava project. We are grateful to Kelly
663 Robbins and to the Cornell University BRC Bioinformatics Core Facility (RRID:SCR_021757) for
664 providing access to computational resources, Michael Gore for insightful discussions, and Salvador
665 Gezan at VSN International for providing ASReml-R technical support. We thank Andrew Siefert for
666 statistical consultation. The authors thank the UK's Foreign, Commonwealth & Development Office
667 (FCDO) and the Bill & Melinda Gates Foundation (Grant INV 007637; www.gatesfoundation.org) for
668 their financial support. SSV is supported by the USDA National Institute of Food and Agriculture
669 AFRI Predoctoral Fellowship project accession no. 1030847.

References

Abincha W, Dzidzienyo DK, Wesonga LN, Mwimali M, Ozimati A, Kayondo IS, Ofori K, Tongoona P, Kivuva BM. 2024. Enhancing Nutrition: A Review of Provitamin A Carotenoid Cassava Breeding Initiatives in East Africa. XIAO X, editor. Advances in Agriculture. 2024(1). doi:10.1155/aia/9937876. https://onlinelibrary.wiley.com/doi/10.1155/aia/9937876.

Afolami I, Mwangi MN, Samuel F, Boy E, Ilona P, Talsma EF, Feskens E, Melse-Boonstra A. 2021. Daily consumption of pro-vitamin A biofortified (yellow) cassava improves serum retinol concentrations in preschool children in Nigeria: A randomized controlled trial. American Journal of Clinical Nutrition. 113(1):221–231. doi:10.1093/ajcn/nqaa290.

Akinwale MG, Aladesanwa RD, Akinyele BO, Dixon AGO, Odiyi AC. 2010. Inheritance of β-carotene in cassava (Manihot esculenta crantza). Int J Genet Mol Biol. 2(10):198–201.

Asare-Marfo D, Birol E, Gonzalez C, Moursi M, Perez S, Schwarz J, Zeller M. 2013. Prioritizing Countries for Biofortification Interventions Using Country-Level Data. [accessed 2023 Nov 14]. http://ebrary.ifpri.org/cdm/ref/collection/p15738coll2/id/127837.

Atser G. 2014. Nigeria releases more cassava with higher pro-vitamin A to fight micronutrient deficiency. [accessed 2021 Dec 9]. International Institute of Tropical Agriculture, www.reliefweb.int.

Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bye-A-Jee H, Cukura A, et al. 2023. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 51(D1):D523–D531. doi:10.1093/nar/gkac1052.

Beyene G, Solomon FR, Chauhan RD, Gaitán-Solis E, Narayanan N, Gehan J, Siritunga D, Stevens RL, Jifon J, Van Eck J, et al. 2018. Provitamin A biofortification of cassava enhances shelf life but reduces dry matter content of storage roots due to altered carbon partitioning into starch. Plant Biotechnol J. 16(6):1186–1200. doi:10.1111/pbi.12862.

Box GEP, Cox DR. 1964. An Analysis of Transformations Author (s): G . E . P . Box and D . R . Cox Published by : Wiley for the Royal Statistical Society Stable URL : http://www.jstor.org/stable/2984418 Accessed : 08-03-2016 06 : 53 UTC Your use of the JSTOR archive indicates. Journal, Source Statistical, Royal Series, Society. 26(2):211–252.

Bredeson J V., Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, Grimwood J, Schmutz J, Rabbi IY, Egesi C, et al. 2016. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nat Biotechnol. 34(5):562–570. doi:10.1038/nbt.3535.

Browning BL, Tian X, Zhou Y, Browning SR. 2021. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet. 108(10):1880–1890.

Cao H, Wang J, Dong X, Han Y, Ma Q, Ding Y, Zhao F, Zhang J, Chen H, Xu Q, et al. 2015. Carotenoid accumulation affects redox status, starch metabolism, and flavonoid/anthocyanin accumulation in citrus. BMC Plant Biol. 15(1):1–16. doi:10.1186/s12870-015-0426-4.

Caspi R, Billington R, Keseler I, Kothari A, Krummenacker M, Midford P, Ong W, Paley S, Subhraveti P, Karp P. 2020. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res. 48(D1):D445–D453. doi:10.1093/nar/gkz862.

708 Ceballos H, Morante N, Sánchez T, Ortiz D, Aragón I, Chávez AL, Pizarro M, Calle F, Dufour D. 2013. Rapid
709 cycling recurrent selection for increased carotenoids content in cassava roots. Crop Sci. 53(6):2342–
710 2351. doi:10.2135/cropsci2013.02.0123.

711 Cervantes-Flores JC, Sosinski B, Pecota K V., Mwanga ROM, Catignani GL, Truong VD, Watkins RH, Ulmer
712 MR, Yencho GC. 2011. Identification of quantitative trait loci for dry-matter, starch, and β-carotene
713 content in sweetpotato. Molecular Breeding. 28(2):201–216. doi:10.1007/s11032-010-9474-5.

714 Chan AW, Villwock SS, Williams AL, Jannink J-L. 2022. Sexual dimorphism and the effect of wild
715 introgressions on recombination in cassava (Manihot esculenta Crantz) breeding germplasm . G3
716 Genes|Genomes|Genetics. 12(1). doi:10.1093/g3journal/jkab372.

717 Chen K, Li GJ, Bressan RA, Song CP, Zhu JK, Zhao Y. 2020. Abscisic acid dynamics, signaling, and functions
718 in plants. J Integr Plant Biol. 62(1):25–54. doi:10.1111/jipb.12899.

719 Decourcelle M, Perez-Fons L, Baulande S, Steiger S, Couvelard L, Hem S, Zhu C, Capell T, Christou P,
720 Fraser P, et al. 2015. Combined transcript, proteome, and metabolite analysis of transgenic maize seeds
721 engineered for enhanced carotenoid synthesis reveals pleotropic effects in core metabolism. J Exp Bot.
722 66(11):3141–3150. doi:10.1093/jxb/erv120.

723 Diretto G, Al-Babili S, Tavazza R, Scossa F, Papacchioli V, Migliore M, Beyer P, Giuliano G. 2010.
724 Transcriptional-metabolic networks in β-carotene-enriched potato tubers: The long and winding road to
725 the golden phenotype. Plant Physiol. 154(2):899–912. doi:10.1104/pp.110.159368.

726 Endelman JB. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP.
727 Plant Genome. 4(3):250–255. doi:10.3835/plantgenome2011.08.0024.

728 Esuma W, Herselman L, Labuschagne MT, Ramu P, Lu F, Baguma Y, Buckler ES, Kawuki RS. 2016.
729 Genome-wide association mapping of provitamin A carotenoid content in cassava. Euphytica. 212(1):97–
730 110. doi:10.1007/s10681-016-1772-5.

731 Esuma W, Kawuki RS, Herselman L, Labuschagne MT. 2016. Diallel analysis of provitamin A carotenoid
732 and dry matter content in cassava (Manihot esculenta Crantz). Breed Sci. 66(4):627–635.
733 doi:10.1270/jsbbs.15159.

734 Esuma W, Ozimati A, Kulakow P, Gore MA, Wolfe MD, Nuwamanya E, Egesi C, Kawuki RS. 2021.
735 Effectiveness of genomic selection for improving provitamin A carotenoid content and associated traits
736 in cassava. G3: Genes, Genomes, Genetics. 11(9). doi:10.1093/g3journal/jkab160.

737 FAO. 2021. FAOSTAT. License: CC BY-NC-SA 30 IGO. [accessed 2024 Apr 3].
738 https://www.fao.org/faostat/en/#data/SCL.

739 Francischini CW, Quaggio RB. 2009. Molecular characterization of Arabidopsis thaliana PUF proteins -
740 Binding specificity and target candidates. FEBS Journal. 276(19):5456–5470. doi:10.1111/j.1742-
741 4658.2009.07230.x.

742 Galesloot TE, Van Steen K, Kiemeney LALM, Janss LL, Vermeulen SH. 2014. A comparison of multivariate
743 genome-wide association methods. PLoS One. 9(4):1–8. doi:10.1371/journal.pone.0095923.

Gao X, Starmer J, Martin ER. 2008. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol. 32(4):361–369. doi:10.1002/gepi.20310.

Garrick DJ, Taylor JF, Fernando RL. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution. 41(1):1–8. doi:10.1186/1297-9686-41-55.

Gemenet DC, da Silva Pereira G, De Boeck B, Wood JC, Mollinari M, Olukolu BA, Diaz F, Mosquera V, Ssali RT, David M, et al. 2020. Quantitative trait loci and differential gene expression analyses reveal the genetic basis for negatively associated β-carotene and starch content in hexaploid sweetpotato [Ipomoea batatas (L.) Lam.]. Theoretical and Applied Genetics. 133(1):23–36. doi:10.1007/s00122-019-03437-7. https://doi.org/10.1007/s00122-019-03437-7.

Giuliano G. 2017. Provitamin A biofortification of crop plants: a gold rush with many miners. Curr Opin Biotechnol. 44(Figure 1):169–180. doi:10.1016/j.copbio.2017.02.001. http://dx.doi.org/10.1016/j.copbio.2017.02.001.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40(D1):D1178–D1186.

Gutschker S, Ruescher D, Rabbi IY, Rosado-Souza L, Pommerrenig B, Pauly M, Robertz S, van Doorn AM, Schlereth A, Neuhaus HE, et al. 2024. Carbon usage in yellow-fleshed Manihot esculenta storage roots shifts from starch biosynthesis to cell wall and raffinose biosynthesis via the myo-inositol pathway. Plant Journal.:1–18. doi:10.1111/tpj.16909.

Holland JB, Nyquist WE, Cervantes-Martínez CT. 2002. Estimating and Interpreting Heritability for Plant Breeding: An Update. In: Janick J, editor. Plant Breeding Reviews. Vol. 22. John Wiley & Sons, Inc. p. 9–112.

Huh SU. 2021. The role of pumilio rna binding protein in plants. Biomolecules. 11(12). doi:10.3390/biom11121851.

Ikeogu UN, Akdemir D, Wolfe MD, Okeke UG, Chinedozi A, Jannink JL, Egesi CN. 2019. Genetic Correlation, Genome-Wide Association and Genomic Prediction of Portable NIRS Predicted Carotenoids in Cassava Roots. Front Plant Sci. 10(December):1–11. doi:10.3389/fpls.2019.01570.

Jaramillo AM, Londoño LF, Orozco JC, Patiño G, Belalcazar J, Davrieux F, Talsma EF. 2018. A comparison study of five different methods to measure carotenoids in biofortified yellow cassava (Manihot esculenta). PLoS One. 13(12):1–14. doi:10.1371/journal.pone.0209702.

Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 44(9):1066–1071. doi:10.1038/ng.2376.

Kutner MH, Nachtsheim CJ, Neter J, Li W. 2005. Identifying Outlying Y Observations - Studentized Deleted Residuals. In: Gordon B, Hercher Jr. RT, Stone L, editors. Applied linear statistical models. 5th ed. McGraw-Hill/Irwin. p. 390–398.

782 Li L, Yang Y, Xu Q, Owsiany K, Welsch R, Chitchumroonchokchai C, Lu S, Van Eck J, Deng XX, Failla M, et
783 al. 2012. The Or gene enhances carotenoid accumulation and stability during post-harvest storage of
784 potato tubers. Mol Plant. 5(2):339–352. doi:10.1093/mp/ssr099.

785 Mackay TFC, Anholt RRH. 2024. Pleiotropy, epistasis and the genetic architecture of quantitative traits.
786 Nat Rev Genet. 25(September 2024):639–657. doi:10.1038/s41576-024-00711-3.
787 http://dx.doi.org/10.1038/s41576-024-00711-3.

788 Mao Z, Sun W. 2015. Arabidopsis seed-specific vacuolar aquaporins are involved in maintaining seed
789 longevity under the control of ABSCISIC ACID INSENSITIVE 3. J Exp Bot. 66(15):4781–4794.
790 doi:10.1093/jxb/erv244.

791 Njoku DN, Gracen VE, Offei SK, Asante IK, Egesi CN, Kulakow P, Ceballos H. 2015. Parent-offspring
792 regression analysis for total carotenoids and some agronomic traits in cassava. Euphytica. 206(3):657–
793 666. doi:10.1007/s10681-015-1482-4.

794 Olayide P, Alexandersson E, Tzfadia O, Lenman M, Gisel A, Stavolone L. 2023. Transcriptome and
795 metabolome profiling identify factors potentially involved in pro-vitamin A accumulation in cassava
796 landraces. Plant Physiology and Biochemistry. 199. doi:10.1016/j.plaphy.2023.107713.

797 Parkes E, Aina O, Kingsley A, Iluebbey P, Bakare M, Agbona A, Akpotuzor P, Labuschagne M, Kulakow P.
798 2020. Combining ability and genetic components of yield characteristics, dry matter content, and total
799 carotenoids in provitamin a cassava f1 cross-progeny. Agronomy. 10(12).
800 doi:10.3390/agronomy10121850.

801 Perveen S, Zafar S, Iqbal N, Riaz M. 2021. 26. Provitamin A Carotenoids. In: Zia-Ul-Haq M, Dewanjee S,
802 Riaz M, editors. Carotenoids: Structure and Function in the Human Body. Springer Nature Switzerland
803 AG. p. 775–797.

804 Rabbi IY, Kayondo SI, Bauchet G, Yusuf M, Aghogho CI, Ogunpaimo K, Uwugiaren R, Smith IA, Peteti P,
805 Agbona A, et al. 2022. Genome-wide association analysis reveals new insights into the genetic
806 architecture of defensive, agro-morphological and quality-related traits in cassava. Plant Mol Biol.
807 109:195–213. doi:10.1007/s11103-020-01038-3. https://doi.org/10.1007/s11103-020-01038-3.

808 Rabbi IY, Udoh LI, Wolfe M, Parkes EY, Gedil MA, Dixon A, Ramu P, Jannink J, Kulakow P. 2017. Genome-
809 Wide Association Mapping of Correlated Traits in Cassava: Dry Matter and Total Carotenoid Content.
810 Plant Genome. 10(3):1–14. doi:10.3835/plantgenome2016.09.0094.

811 Rockman M V. 2008. Reverse engineering the genotype-phenotype map with natural genetic variation.
812 Nature. 456(7223):738–744. doi:10.1038/nature07633.

813 Rodriguez-Amaya DB, Kimura M. 2004. HarvestPlus Handbook for Carotenoid Analysis. HarvestPlus
814 Technical Monographs.:59.

815 Sánchez T, Chávez AL, Ceballos H, Rodriguez-Amaya DB, Nestel P, Ishitani M. 2006. Reduction or delay of
816 post-harvest physiological deterioration in cassava roots with higher carotenoid content. J Sci Food
817 Agric. 86(4):634–639. doi:10.1002/jsfa.2371.

818 Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. 2013. Pleiotropy in complex traits : challenges
819 and strategies. Nature Publishing Group. 14(July). doi:10.1038/nrg3461.

820 Stanley LE, Ding B, Sun W, Mou F, Hill C, Chen S, Yuan YW. 2020. A tetratricopeptide repeat protein
821 regulates carotenoid biosynthesis and chromoplast development in monkeyflowers (Mimulus). Plant
822 Cell. 32(5):1536–1555. doi:10.1105/tpc.19.00755.

823 Stephens M. 2013. A Unified Framework for Association Analysis with Multiple Related Phenotypes.
824 PLoS One. 8(7). doi:10.1371/journal.pone.0065245.

825 Stevens GA, Bennett JE, Hennocq Q, Lu Y, De-Regil LM, Rogers L, Danaei G, Li G, White RA, Flaxman SR,
826 et al. 2015. Trends and mortality effects of vitamin A deficiency in children in 138 low-income and
827 middle-income countries between 1991 and 2013: A pooled analysis of population-based surveys.
828 Lancet Glob Health. 3(9):e528–e536. doi:10.1016/S2214-109X(15)00039-X.

829 Tahirou A, Bamire AS, Oparinde A, Akinola AA. 2015. Determinants of Adoption of Improved Cassava
830 Varieties among Farming Households in Oyo, Benue, and Akwa Ibom States of Nigeria. HarverstPlus
831 Working Paper. 22(December):21.

832 Taraszka K, Zaitlen N, Eskin E. 2022. Leveraging pleiotropy for joint analysis of genome-wide association
833 studies with per trait interpretations. PLoS Genet. 18(11):1–24. doi:10.1371/journal.pgen.1010447.
834 http://dx.doi.org/10.1371/journal.pgen.1010447.

835 The VSNi Team. 2023. asreml: Fits Linear Mixed Models using REML.

836 Udoh LI, Gedil M, Parkes EY, Kulakow P, Adesoye A, Nwuba C, Rabbi IY. 2017. Candidate gene
837 sequencing and validation of SNP markers linked to carotenoid content in cassava (Manihot esculenta
838 Crantz). Molecular Breeding. 37(10). doi:10.1007/s11032-017-0718-5.

839 Villwock SS, Li L, Jannink J. 2024. Carotenoid-carbohydrate crosstalk: evidence for genetic and
840 physiological interactions in storage tissues across crop species. New Phytologist. In Press.

841 Welsch R, Arango J, Bär C, Salazar B, Al-Babili S, Beltrán J, Chavarriaga P, Ceballos H, Tohme J, Beyera P.
842 2010. Provitamin A accumulation in cassava (Manihot esculenta) roots driven by a single nucleotide
843 polymorphism in a phytoene synthase gene. Plant Cell. 22(10):3348–3356. doi:10.1105/tpc.110.077560.

844 Wiseman EM, Bar-El Dadon S, Reifen R. 2017. The vicious cycle of vitamin A deficiency: A review. Crit
845 Rev Food Sci Nutr. 57(17):3703–3714. doi:10.1080/10408398.2016.1160362.
846 https://doi.org/10.1080/10408398.2016.1160362.

847 Wolfe M. 2022. genomicMateSelectR: Genomic Mate Selection. R package version 0.2.0.
848 https://github.com/wolfemd/genomicMateSelectR.

849 Wolfe MD, Bauchet GJ, Chan AW, Lozano R, Ramu P, Egesi C, Kawuk R, Kulakow P, Rabbi I, Jannink JL.
850 2019. Historical introgressions from a wild relative of modern cassava improved important traits and
851 may be under balancing selection. Genetics. 213(4):1237–1253. doi:10.1534/genetics.119.302757.

852 Xu C, Luo F, Hochholdinger F. 2016. LOB Domain Proteins: Beyond Lateral Organ Boundaries. Trends
853 Plant Sci. 21(2):159–167. doi:10.1016/j.tplants.2015.10.010.

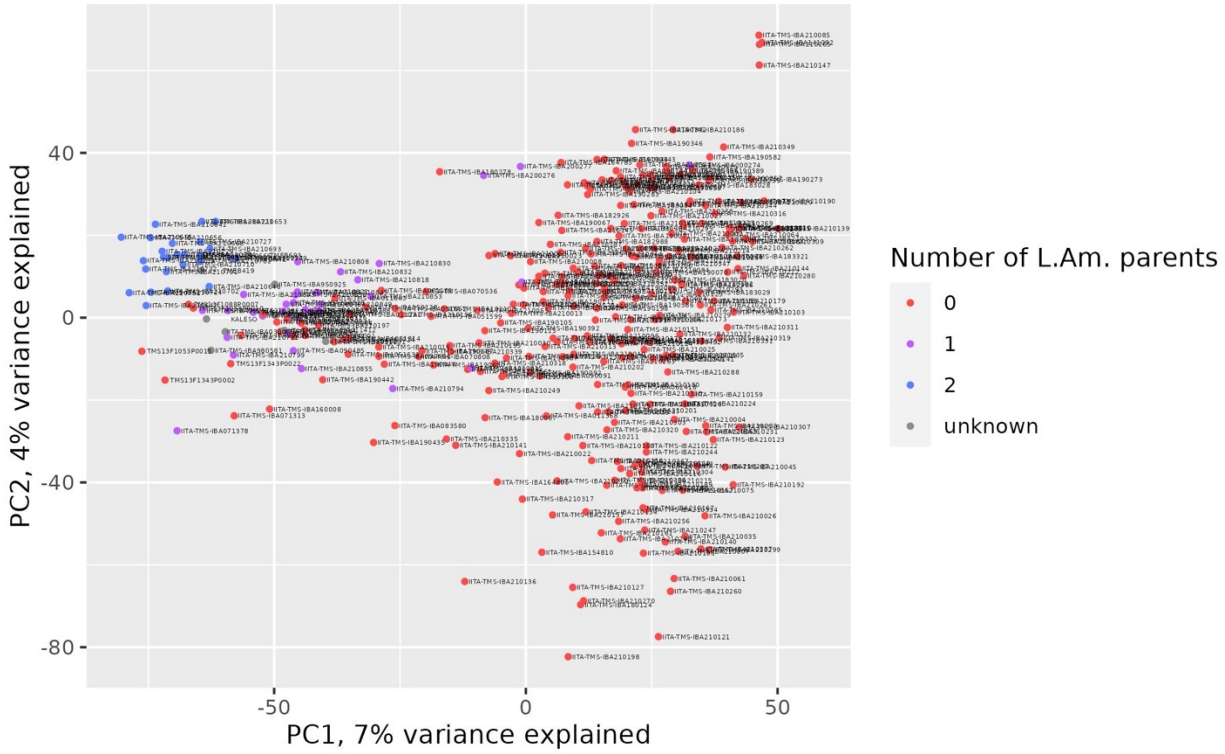
854 Zhou X, Stephens M. 2014. Efficient multivariate linear mixed model algorithms for genome-wide
855 association studies. Nat Methods. 11(4):407–409. doi:10.1038/nmeth.2848.

856

857

858 **Supplementary Information**

859



860

861 **Figure S1.** Principal component analysis of genotyped accessions shows that the first principal
862 component (PC1) separates a cluster of accessions with recent Latin American ancestry. Color
863 represents the number of Latin American parents listed in an accession's pedigree.

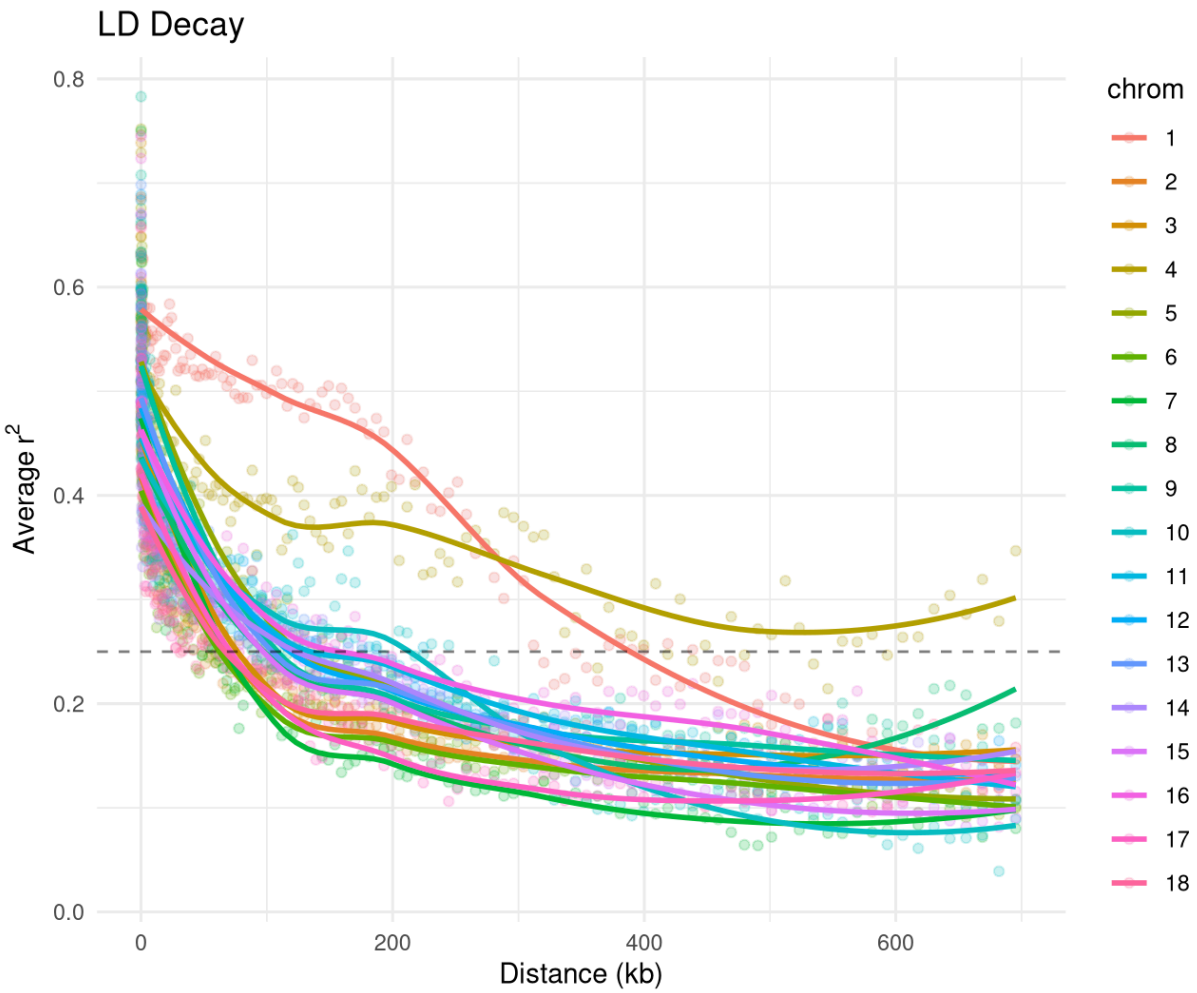


Figure S2. Linkage disequilibrium (LD) decay measured pairwise r^2 between SNPs. Points represent average pairwise r^2 per chromosome within 200 physical distance intervals.

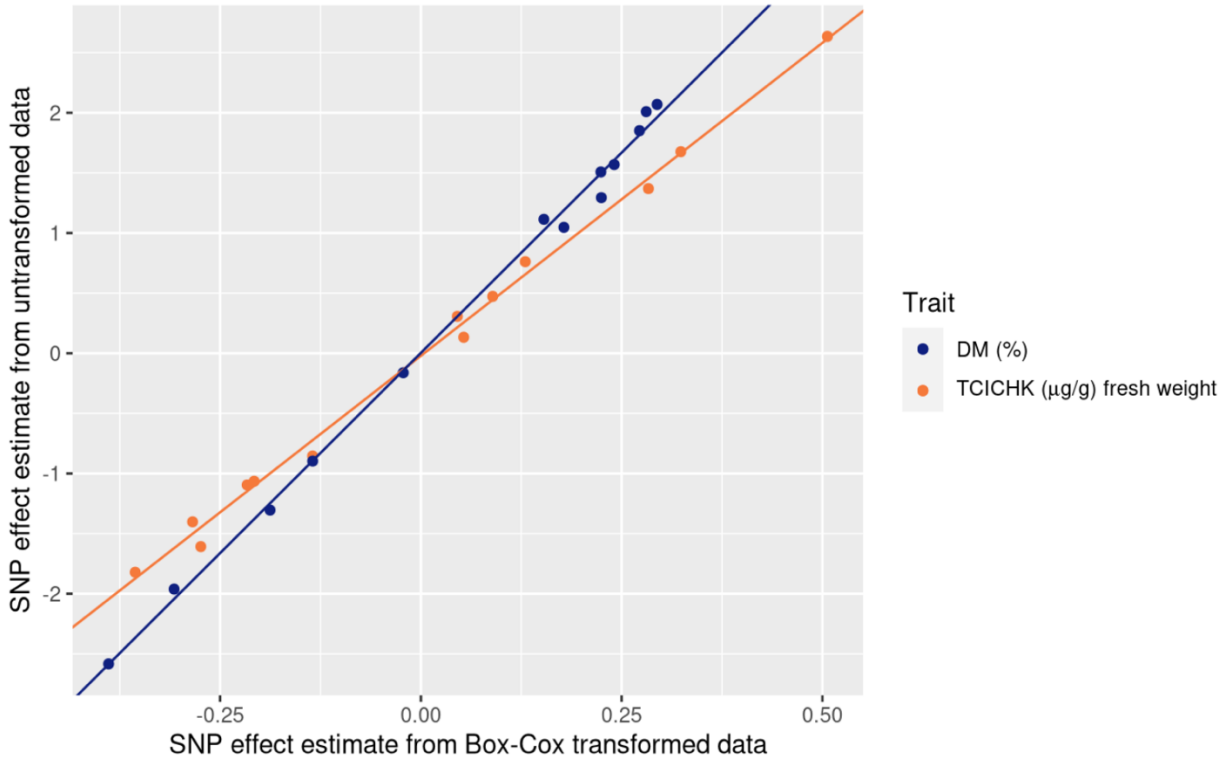
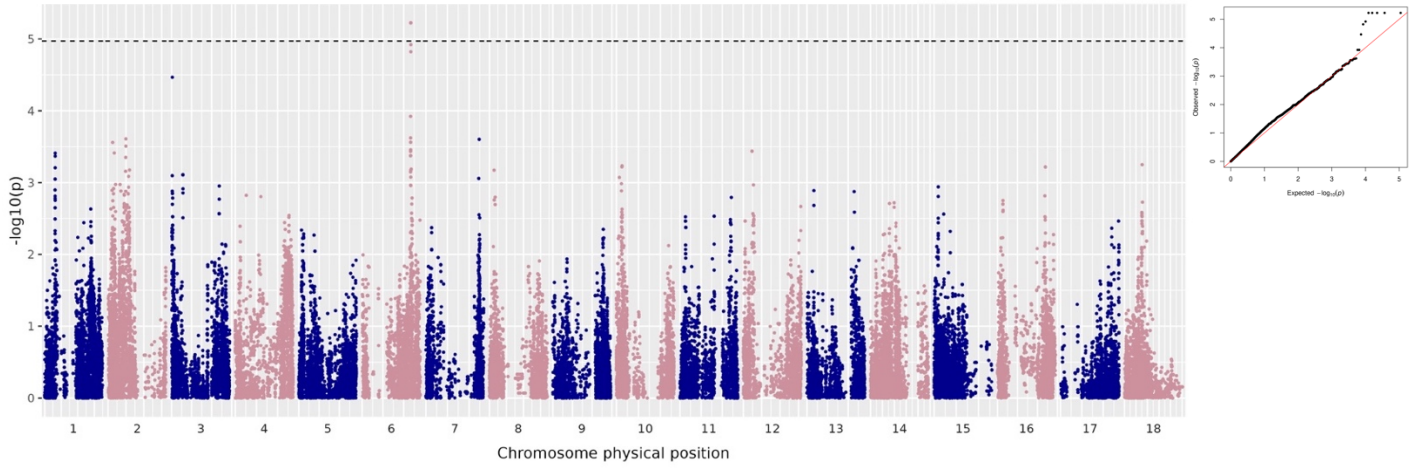


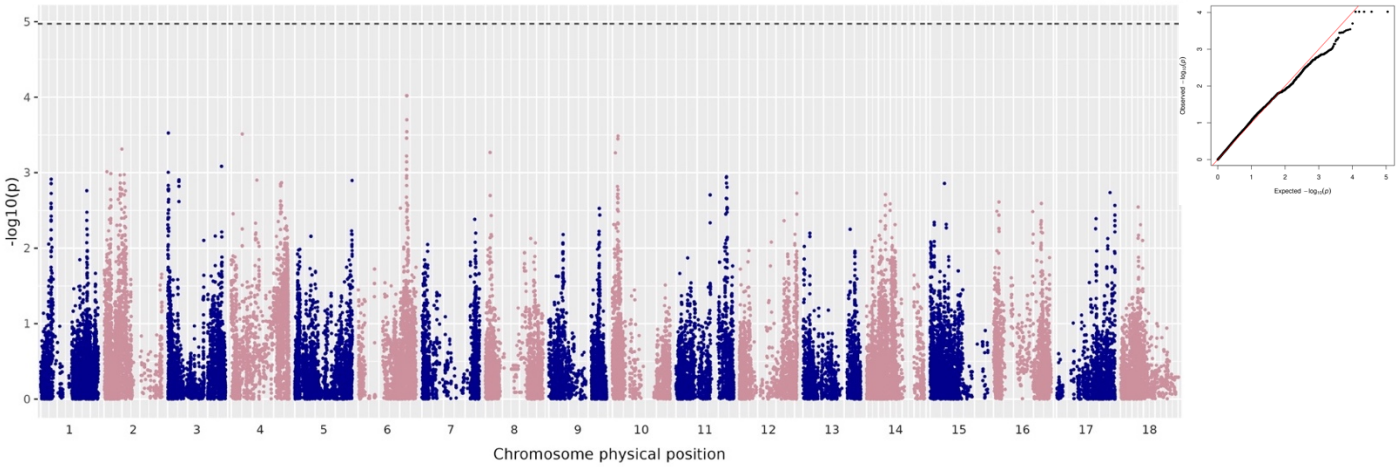
Figure S3. Comparison of peak SNP effect estimates from Box-Cox normalized data and from untransformed data show a linear relationship. SNP effect estimates from the untransformed data correspond to the traits' original units.

A. DRYLD



875

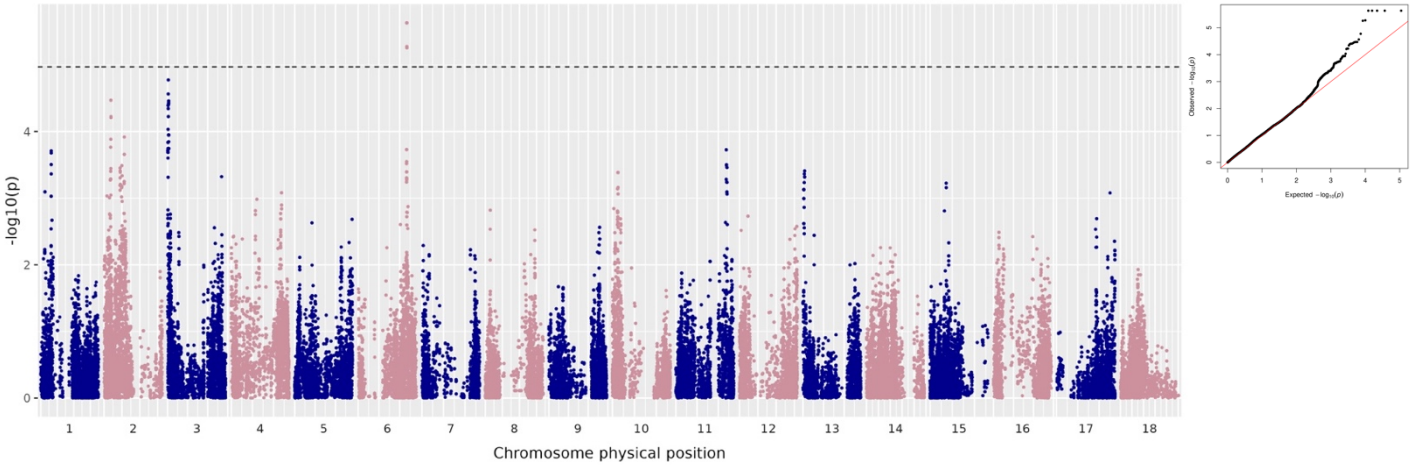
B. FRYLD



876

877

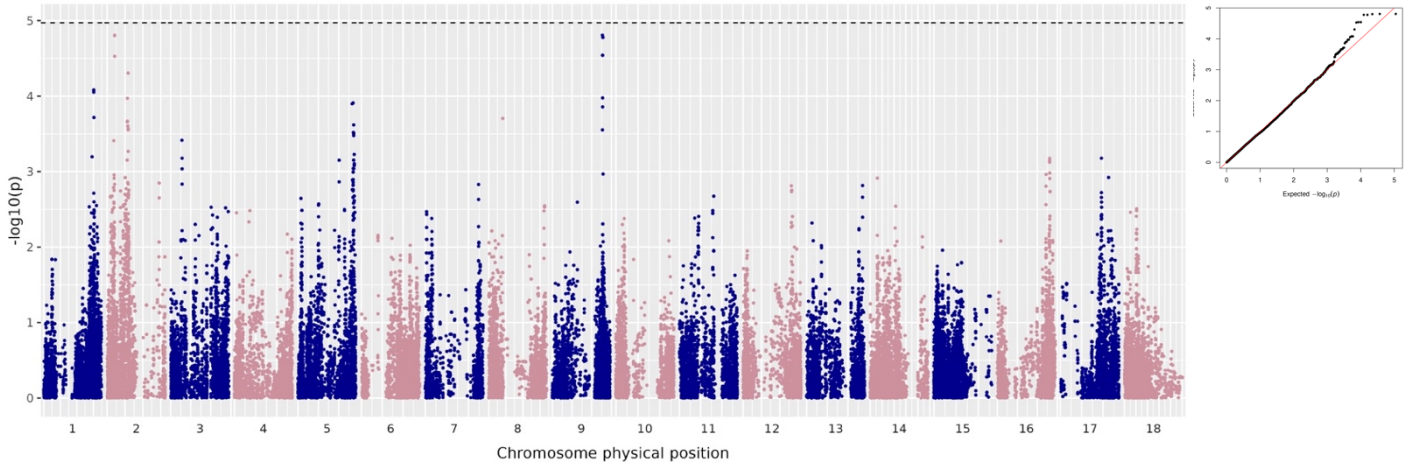
C. RTWT



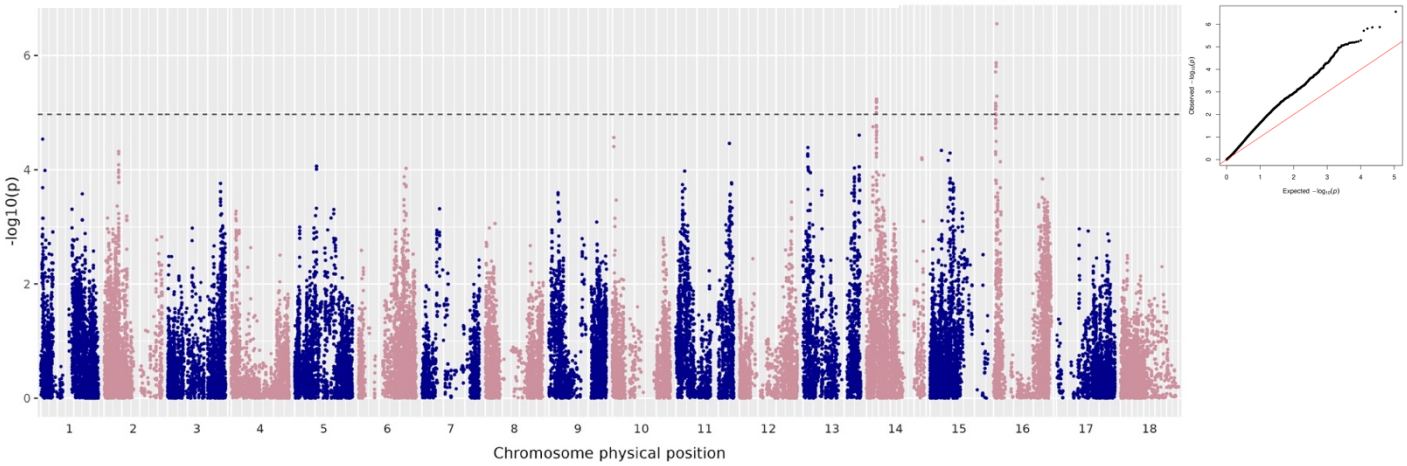
878

879

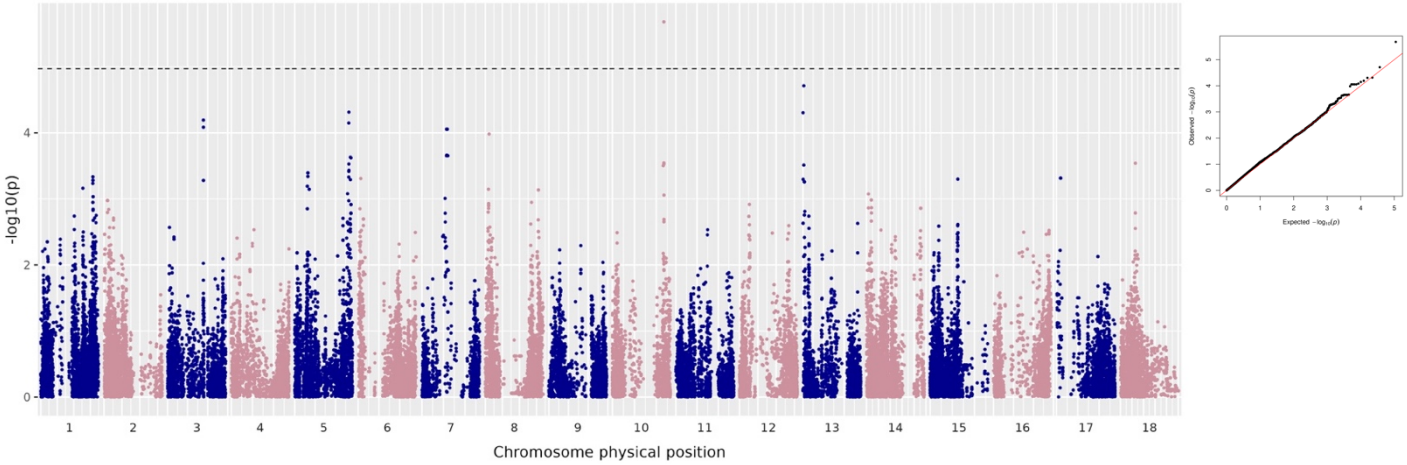
D. HI



E. SPRTPERC



F. CMD3I



G. CMD3S

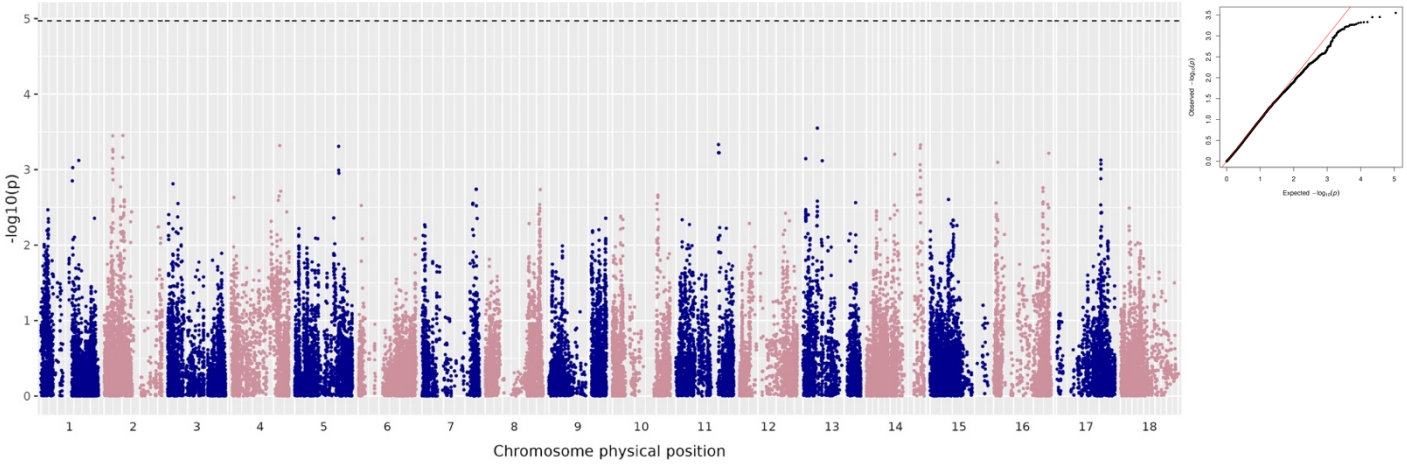


Figure S4. Manhattan plots for univariate GWAS for analyzed agronomic traits: DRYLD (A), FRYLD

886 (B), RTWT (C), HI (D), SPRTPERC (E), CMD3I (F), and CMD3S (G). The dashed black line indicates
887 the modified Bonferroni significance threshold at α = 0.05. Q-Q plots for each analysis are at right.
888
889
890 **Supplementary Tables**
891 *Attached file.*
892 **Table S1.** Metadata for field trials included in this analysis.
893 **Table S2**. *A priori* candidate genes with annotations relating to carotenoid, apocarotenoid, and
894 starch synthesis and degradation pathways.
895 **Table S3.** Summary statistics for peak SNPs identified in univariate associations with agronomic
896 traits at the Bonferroni-corrected significance threshold.
897 **Table S4.** Candidate genes within 100kb of peak SNPs significantly associated with TCICHK and DM
898 at the modified Bonferroni significance threshold from the bivariate GWAS.
899 **Table S5.** Candidate genes within 100kb of peak SNPs identified in univariate associations with
900 agronomic traits at the Bonferroni-corrected significance threshold.