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. The majority of 17 jointly-associated loci identified at a relaxed significance threshold affected TC and DM in opposite directions, although this pattern did not reach statistical significance in a binomial test. Even after accounting for the effects of these 17 loci as covariates, 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. This population originated from crosses
170 between high carotenoid varieties from Latin America and Africa with high-yielding African
171 varieties. Since the early 2000s, recurrent phenotypic selection and marker-assisted selection have
172 been used to improve yield, disease resistance, nutritional quality, and other agronomic traits in the
173 yellow cassava population, which has led to the release of nine yellow varieties since 2011
174 (HarvestPlus 2024). The population sample used here included recent ancestry from Latin America,
175 with 26 Latin American clones and 34 F1 or F2 hybrids between Latin American and African clones
176 (Fig. S1). 78% of the clones were early-stage selections, while the remainder were more advanced
177 selections. 92% of the clones were classified as "cream" or "yellow" flesh types while the remainder
178 were "white" based on visual color rating. Many accessions of this population have been described
179 and studied previously (Akinwale et al. 2010; Udoh et al. 2017; Ige et al. 2022; Mbanjo et al. 2024;
180 Anokye et al. 2025). Additional accession metadata are available in Table S1.

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

189 To increase mapping resolution, DArTSeqLD SNP data were imputed to a higher density
190 with Beagle5.0 (Browning et al. 2018) using the R package genomicMateSelectR version 0.2.0
191 (Wolfe 2022) with a reference panel of 21,856 accessions available on Cassavabase
192 (https://cassavabase.org/ftp/marnin_datasets/nextgenImputation2019/) as in Nandudu et al.
193 (2024). A principal component analysis (PCA) showed that the study population clustered within
194 the imputation reference panel, indicating the reference panel appropriately captures the genetic
195 diversity in the study population (Fig. S2). Then, 7,096 accessions available on Cassavabase with
196 both DArTSeqLD genotypes and KASP marker genotypes for the *PSY2* causal allele at S1_24155522

197 were joined and phased with Beagle5.0 to create a reference panel for imputing the *PSY2* marker
198 (observed directly in 46 accessions). Following imputation, SNPs were filtered with the following
199 parameters: dosage R^2 (DR2) > 0.75 (with the exception of the *PSY2* marker with DR2 = 0.74),
200 Hardy-Weinberg equilibrium p-value (P_HWE) > 1e-20, and minor allele frequency (MAF) > 0.05
201 for a total of 43,720 biallelic SNPs for 378 unique accessions. Imputed allele dosage expectations
202 were left as unrounded decimals to account for imputation uncertainty (Guan and Stephens 2008;
203 Kutalik et al. 2011).
204 Population structure was visualized with PCA in R using *prcomp()* on the genetic data
205 matrix and *ggplot()* for plotting. The first five principal components were used to account for
206 population structure in the genome-wide associations based on the elbow of the scree plot, where
207 the proportion of variance explained by PCs began to plateau. An additive genomic relationship
208 matrix calculated using the *rrBLUP* package (Endelman 2011) was used to account for relatedness
209 among the genotypes. Linkage disequilibrium (LD) patterns in the data were evaluated by
210 calculating R^2, the squared Pearson correlation coefficient, between each pair of SNPs with *cor().*
211
212 *Phenotypic data*
213 Phenotypic data were downloaded from Cassavabase (https://cassavabase.org/) for field
214 trials that included at least 3 of the genotyped accessions with carotenoid measurements to ensure
215 sufficient data for estimating trial-level effects. A total of 51 trials included 4 diversity collection
216 trials ("Genetic Gain"; GG), 6 Clonal Evaluation Trials (CET), 17 Preliminary Yield Trials (PYT), 12
217 Advanced Yield Trials (AYT), 1 Uniform Yield Trial (UYT), and 11 Nationally Coordinated Research
218 Program variety trials (NCRP), representing different stages of selection in the breeding pipeline
219 and levels of replication (Table S2).
220 Dry matter (DM) content percentage was measured as the ratio of the weight before and
221 after oven drying about 100g of shredded cassava mixed from 8 peeled roots as previously
222 described (Rabbi et al. 2017). Total carotenoids were measured by spectrophotometry using the
223 iCheck™ Carotene device (TCICHK) from BioAnalyt (Teltow, Germany) as previously described
224 (Rodriguez-Amaya and Kimura 2004; Maroya et al. 2012; Udoh et al. 2017; Jaramillo et al. 2018).
225 Briefly, a blended cassava root sample was mixed with an alcohol-based extraction reagent and
226 placed into the iCheck™ spectrophotometer to estimate the concentration of total carotenoids (in
227 μg/g fresh weight (FW)) based on light absorbance. The total carotenoid content in cassava roots is
228 comprised predominantly (~85%) β-carotene with lower levels of α-carotene and xanthophylls
229 such as violaxanthin and lutein (Rodriguez-Amaya and Kimura 2004; Ikeogu et al. 2017). The

230 iCheck™ method has previously shown a high correlation ($r^2 = 0.98$, $p < 0.001$) with HPLC
231 measurement of total carotenoids (Jaramillo et al. 2018). Carotenoid content was also rated by
232 visual color evaluation using a color chart (TCHART) on an ordinal scale from 1 (white) to 7 (deep
233 yellow) (Rabbi et al. 2017). Additional agronomic traits measured for at least 300 accessions were
234 selected for association analyses (Table 1).

235 The phenotypic data were analyzed in a two-step approach to account for environmental
236 effects before running the GWAS. Phenotypic values were transformed using the Box-Cox
237 transformation for normality using the R function *boxcox,* using the optimum λ value for each trait
238 (Box and Cox 1964). The data were then centered and scaled by dividing by the traits' standard
239 deviations. Best linear unbiased predictions (BLUPs) of the effect of genotype ID on each trait and
240 variance components were calculated using the following mixed model fit with ASReml-R version
241 4.1.0.110 (The VSNi Team 2023):

242 Eqn. 1). $$\boldsymbol{y} \sim \mu + X_{l:y}\boldsymbol{\beta_1} + Z_g\boldsymbol{U_1} + Z_{g:l:y}\boldsymbol{U_2} + Z_{t:b}\boldsymbol{U}_3 + \boldsymbol{\epsilon},$$

243 where $\boldsymbol{y}$ is a vector (n x 1) of Box-Cox normalized and scaled trait values for n observations; μ is the
244 overall population mean; $X_{l:y}$ is a design matrix (n x n_{ly}) indicating the year within location for
245 each observation; $\boldsymbol{\beta_1}$ is a vector (n_{ly} x 1) of fixed effect estimates for year nested within location; Z_g
246 is a design matrix (n x n_g) indicating the genotype ID for each observation; $\boldsymbol{U_1}$ is a vector (n_g x 1) of
247 predicted random effects of genotype (BLUPs) for n_g genotypes, assumed to be normally
248 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
249 the genotype, location, and year combination for each observation; $\boldsymbol{U_2}$ is a vector (n_{gly} x 1) of
250 predicted random effects of the interaction of genotype with year/location combination, assumed
251 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
252 design matrix (n x n_{tb}) indicating the block number within trial for each observation; $\boldsymbol{U_3}$ is a vector
253 (n_{tb} x 1) of predicted random effects for block within trial, assumed to be normally
254 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,
255 assumed to be normally distributed $\boldsymbol{\epsilon} \sim N(0, {\sigma_\epsilon}^2)$ with residual variance ${\sigma_\epsilon}^2$. Outliers were
256 identified and removed based on studentized deleted residuals with a Bonferroni corrected α =
257 0.05 (Kutner et al. 2005). Then, Eqn. 1 was re-fitted without the outliers (4 observations excluded
258 for TCICHK; 3 for TCHART; and 6 for DM). The broad-sense heritability (H^2) of each trait was
259 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
260 estimated genetic by environment variance, $\hat{\sigma}_\epsilon^2$ is the estimated residual variance, $\bar{n}_{ly}$ is the

261 harmonic mean number of location/year combinations per accession, and $\bar{n}_r$ is the harmonic mean
262 number of plots per accession across locations, years, and replications (Holland et al. 2002).
263 Finally, BLUPs were de-regressed by dividing by their prediction reliability ($\boldsymbol{r^2} = 1 - \frac{PEV}{\sigma_g^2}$) as
264 described in Garrick et al. (2009). To account for the imbalanced number of observations per
265 genotype, the BLUP prediction reliabilities were carried forward to use as inverse variance weights
266 in the second-stage GWAS model as recommended by Garrick et al. (2009). Trait-specific weights
267 were calculated as $\boldsymbol{w} = \frac{1-h^2}{0.1+\frac{1-r^2}{r^2}h^2}$, where $h^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_{gly}^2 + \hat{\sigma}_\epsilon^2}$.
268
269 *Statistical methods for genome-wide associations*
270 Genome-wide associations for the bivariate phenotype were tested using the following
271 mixed linear model fit for each SNP*:*
272 Eqn. 2). $$(\boldsymbol{y_1}, \boldsymbol{y_2}) \sim \boldsymbol{\mu} + \boldsymbol{X_i}\boldsymbol{\beta_1} + \boldsymbol{P}\boldsymbol{\beta_2} + Q\boldsymbol{\beta_3} + ZU + \epsilon,$$
273 where $(\boldsymbol{y_1}, \boldsymbol{y_2})$ is a matrix (n_g x 2) of de-regressed TCICHK and DM BLUP values for n_g genotypes; $\boldsymbol{\mu}$
274 is a vector (2 x 1) of overall means for each trait; $\boldsymbol{X_i}$ is a vector (n_g x 1) of marker dosages (0, 1, 2)
275 for SNP *i* for n_g genotypes; $\boldsymbol{\beta_1}$ is a vector (1 x 2) of estimates for the additive fixed effect of SNP *i* on
276 each trait*;* $\boldsymbol{P}$ is a vector (n_g x 1) indicating the imputed genotype at the *PSY2* causal SNP,
277 S1_24155522; $\boldsymbol{\beta_2}$ is a vector (1 x 2) of estimates for the additive fixed effect of the *PSY2* SNP on
278 each trait; Q is a matrix (n_g x 5) of the first 5 PCs for the n_g genotypes; $\boldsymbol{\beta_3}$ is a matrix (5 x 2) of fixed
279 effect estimates for the first 5 PCs; Z is an identity matrix (n_g x n_g) identifying the genotype, U is a
280 matrix (n_g x 2) of the genotype random effect on each trait with matrix normal distribution
281 $U \sim MN\left(\begin{smallmatrix}0\\0\end{smallmatrix}, G, F\right)$, with covariance across rows specified by G, the relationship matrix (n_g x n_g) for
282 the genotypes calculated with genome-wide SNPs excluding the *PSY2* SNP, and covariance across
283 columns specified by F, the additive genetic variance-covariance matrix (2 x 2) for the two traits
284 after accounting for the SNP fixed effects; ϵ is a matrix (n_g x 2) of residual effects for each trait, with
285 matrix normal distribution $\epsilon \sim MN\left(\begin{smallmatrix}0\\0\end{smallmatrix}, R, E\right)$ with covariance across rows specified by *R,* a diagonal
286 matrix (n_g x n_g) of inverse BLUP reliability weights $\boldsymbol{w}$*,* and covariance across columns specified by E,
287 the variance-covariance matrix (2 x 2) of the residual effects for the two traits. The averages of the
288 trait-specific BLUP weights for TCICHK and DM were used as accession-level weights. This equation
289 was fit for each SNP in *ASRemlR* with the following statements:
290 fixed = cbind(TCICHK, DM) ~ trait + trait:SNP + trait:PSY2 + PC1 + PC2 + PC3 + PC4 + PC5,
291 random = ~ corgh(trait):vm(genotype, Ginverse),

residual = ~ units:us(trait), weights = w.

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_g x 1) of de-regressed BLUP values for trait *j* for n_g 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_g x 1) of predictions for the genotype random effect, assumed normally distributed $\boldsymbol{U} \sim N(0, G\sigma_G^2)$ with polygenic genetic variance σ_G^2; $\boldsymbol{\epsilon}$ is a vector (n_g x 1) of residual effects, assumed normally distributed $\boldsymbol{\epsilon} \sim N(0, R\sigma_e^2)$ where *R* is a diagonal matrix (n_g x n_g) of inverse BLUP reliability weights $\boldsymbol{w_j}$ with residual variance σ_e^2; 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_j ~ 1 + SNP + PSY2 + PC1 + PC2 + PC3 + PC4 + PC5,

random = ~ vm(genotype, Ginverse),

residual = ~ units, weights = w_j.

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 3657, which was used as the denominator in the Bonferroni calculation to establish a significance threshold at $p < 1.367 \times 10^{-5}$. Manhattan plots were visualized with *ggplot()* and quantile-quantile plots were visualized with the *qqman* package (Turner 2018).

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 enabled 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 (PVE) by each associated SNP *i* for a given trait was estimated according to the formula $PVE_i = \frac{Var(\beta_i x)}{Var(y)} = \frac{\hat{\beta}_i^2 * 2p * (1-p)}{\hat{\sigma}_y^2}$, where $\hat{\beta}_i$ is the estimated additive fixed effect of SNP *i*, *p* is the frequency of the alternate allele, and $\hat{\sigma}_y^2$ is the total genetic variance, the sum of variance components extracted from the reduced model fit on de-regressed BLUPs without the SNP covariate (Eqn. 2). A measure of proportion covariance explained (PCE) was derived from the PVE formula as follows:

Eqn. 4). $$PCE_{y1,y2} = \frac{Cov\left(\beta_{y_1}x, \beta_{y_2}x\right)}{Cov(y_1,y_2)} = \frac{\beta_{y1} * \beta_{y2} * Var(x)}{Cov(y_1,y_2)}; \quad \widehat{PCE}_{y1,y2} = \frac{\hat{\beta}_{y1} * \hat{\beta}_{y2} * 2p * (1-p)}{\hat{\sigma}_{y_1,y_2}},$$

where $\hat{\sigma}_{y_1,y_2}$ is the estimated total genetic covariance for traits y_1 and y_2, extracted from the reduced model without the SNP. Negative PCE values indicate that the SNP contributes to covariance in the opposite direction of the overall genetic covariance.

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

Eqn. 6). $$\boldsymbol{y_2} \sim \mu + \boldsymbol{X_i}\beta_1 + P\boldsymbol{\beta_2} + Q\boldsymbol{\beta_3} + \boldsymbol{y_1}\beta_4 + Z\boldsymbol{U} + \boldsymbol{\epsilon}$$

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

fixed = $\text{trait}_1 \sim 1 + \text{SNP} + \text{PSY2} + \text{trait}_2 + \text{SNP:trait}_2 + \text{PC1} + \text{PC2} + \text{PC3} + \text{PC4} + \text{PC5}$,

random = ~ vm(genotype, Ginverse),

residual = ~ units, weights = w_1.

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. 5) 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. 6). These possible pleiotropic relationships are summarized in Fig. 1.

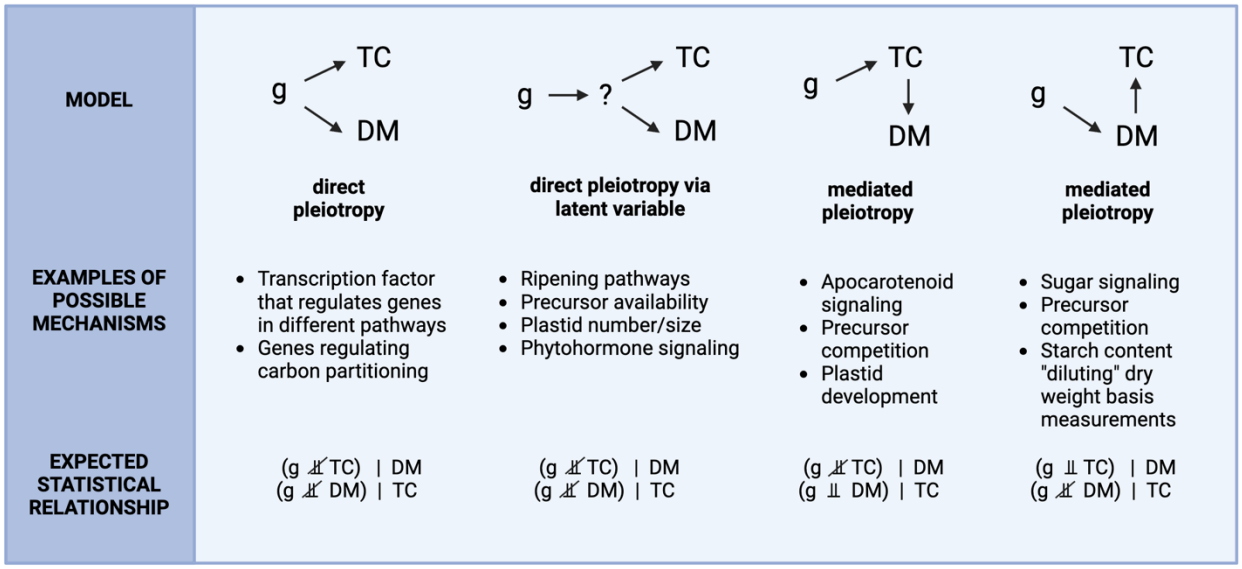


Figure 1. Illustration of direct and mediated pleiotropic relationships and their expected statistical associations, where g ⫫ A | B denotes conditional independence of a genetic variant *g* with trait A given trait B. Hypothesized biological mechanisms are described in detail in Villwock et al. (2024).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

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

The relative contributions of the identified putative QTL to the genetic covariance between TCICHK and DM were assessed by comparing the variance components before and after accounting for the peak SNPs as covariates. A bivariate model as described in Eqn. 2 was fit with all *z* peak SNPs identified at the relaxed significance threshold added as covariates:

Eqn. 6). $$(\boldsymbol{y_1}, \boldsymbol{y_2}) \sim \boldsymbol{\mu} + \sum_{i=1}^{z} \boldsymbol{X_i}\beta_i + \boldsymbol{P\beta_2} + Q\boldsymbol{\beta_3} + ZU + \epsilon$$

With the genotype random effect $U \sim MN\left(\begin{matrix}0\\0\end{matrix}, G, F\right)$, F represents the additive genetic variance-covariance matrix for the two traits remaining after accounting for the fixed effects of the identified QTL. If the identified QTL completely accounted for the negative genetic covariance between traits, then the residual genetic covariance in F would be expected to approach zero. To formally test whether significant residual genetic covariance remained after accounting for the associated SNPs, a likelihood ratio test implemented with *lrt.asreml()* compared models with the residual genetic covariance between traits set to zero or unconstrained.

Candidate gene analysis

M. esculenta reference genome v6.1 gene annotation data were downloaded from the JGI data portal (https://data.jgi.doe.gov/). A list of *a priori* candidate genes involved in the carotenoid synthesis (MetaCyc pathway ontology ID PWY-6475), apocarotenoid synthesis (PWY-6806), starch synthesis (PWY-622), and starch degradation (PWY-6724) pathways annotated in *Manihot esculenta v6.1* on Phytozome was assembled (Table S3) (Goodstein et al. 2012; Bredeson et al. 2016; Caspi et al. 2020). *A priori* candidate genes were screened for LD and physical distance to significantly associated SNPs. Then, all annotated genes within 100kb of the peak SNPs were extracted from the genome annotation file. Genes with missing ontology annotations were supplemented with annotations from the UniProt Knowledgebase where possible (Bateman et al. 2023).

Results

Trait distributions and relationships

The phenotypic distributions of TCICHK and DM observations of genotyped accessions before transformation and statistical analysis are shown in Fig. S3. TCICHK ranged from 0 to 30.8 µg/g FW and DM ranged from 9.1 to 60.0%. TCICHK and DM had moderately high broad-sense heritability of 0.49 and 0.50, respectively. The two traits were negatively genetically correlated with $r = -0.34$ and $p = 1.84 \times 10^{-11}$ (Fig. 2) and phenotypically correlated with $r = -0.15$ and $p = 4.48 \times 10^{-22}$. TCICHK was also negatively genetically correlated with several other agronomic traits: dry yield (DYLD; $r = -0.17$, $p = 1.66 \times 10^{-3}$), root weight ($r = -0.18$, $p = 5.98 \times 10^{-4}$), fresh root yield (FRYLD; $r = -0.15$, $p = 8.43 \times 10^{-3}$), and sprouting percentage (SPRTPERC; $r = -0.25$, $p = 1.26 \times 10^{-6}$) (Fig. 3). The broad-sense heritabilities of all traits are listed in Table 1.

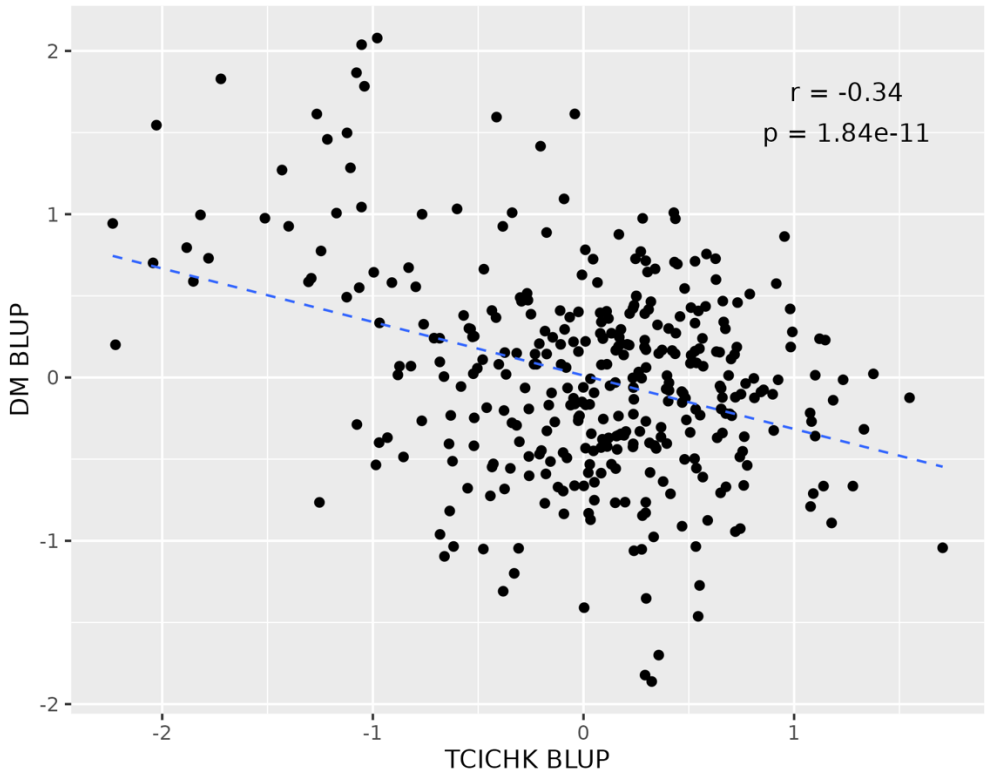


Figure 2. Negative genetic relationship between total carotenoid content as measured by iCheck™ Carotene (TCICHK) and dry matter content (DM) BLUPs among the 378 genotyped accessions after normalization of phenotypic values with Box-Cox transformation. Pearson correlation coefficient (r) and p-value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

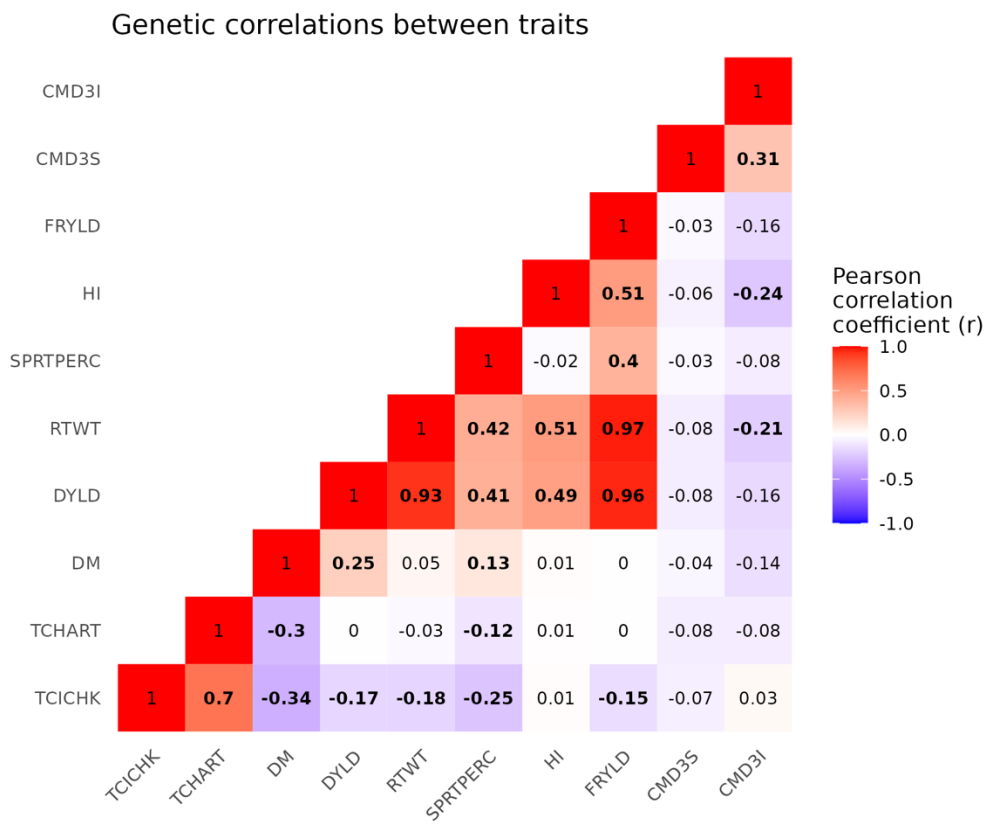


Figure 3. Pairwise genetic correlations between all analyzed traits. Each cell is labeled with the Pearson correlation coefficient (r) between the traits listed on the x and y axis. Significant correlations ($p < 0.05$) are bolded.

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	1221	0.49
Total carotenoids by color chart (1-9) method	TCHART	CO_334.0000161	374	1750	0.73
Dry matter content percentage	DM	CO_334.0000092	370	1714	0.50
Dry yield	DYLD	CO_334.0000014	332	1249	0.21
Fresh storage root weight per plot	RTWT	CO_334.0000012	373	1748	0.26
Harvest index	HI	CO_334.0000015	368	1408	0.43
Sprouting proportion	SPRTPERC	CO_334.0000008	375	1697	0.24
Fresh root yield	FRYLD	CO_334.0000013	334	1333	0.23

Cassava mosaic disease severity 3-month evaluation	CMD3S	CO_334.0000192	376	1792	0.60
Cassava mosaic disease incidence 3-month evaluation	CMD3I	CO_334.0000196	376	1756	0.49

411

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

413 The Manhattan plot from the bivariate associations revealed 17 peaks that passed the
414 relaxed threshold ($p < 0.001$), including two QTL peaks that passed the modified Bonferroni
415 significance threshold for joint association with both TCICHK and DM after controlling for the effect
416 of *PSY2* (Fig. 4). The significant joint associations were located on chromosomes 1 and 10. The top
417 SNP on chromosome 1 is located at 23.484417 Mbp, which is 0.67 Mbp upstream from *PSY2* and
418 was in moderate LD ($R^2 = 0.21$) with the *PSY2* causal variant at 24.155522 Mbp. Fig. 5 illustrates the
419 results from bivariate and univariate associations in this QTL region in higher resolution. The top
420 jointly-associated SNP on chromosome 1 in the bivariate associations was significantly associated
421 with TCICHK but not with DM in univariate analyses. The proportion of genetic variance explained
422 by the two QTL were relatively small ($PVE_{TCICHK} = 3.4\%$ and $PVE_{DM} = 1.1\%$ for the chromosome 1
423 peak; $PVE_{TCICHK} = 2.9\%$ and $PVE_{DM} = 0.7\%$ for the chromosome 10 peak). The *PSY2* SNP explained
424 about 2.4% of the genetic variation in TCICHK and 5.0% of the genetic variation in DM. The
425 chromosome 1 and 10 QTL explained about 3% and 2% of the genetic covariance between traits,
426 respectively, while the *PSY2* SNP explained 5.6%. Summary statistics for all associated SNPs are
427 listed in Table 2.

428 Of the 17 associated loci identified at the relaxed significance threshold, 11 peak SNPs had
429 opposite effects on TC and DM, while 6 SNPs had effects on TC and DM in the same direction. A
430 binomial test with these counts resulted in a p-value of 0.166, which was insufficient to reject the
431 null hypothesis that associated SNPs are equally likely to affect TC and DM in same or opposite
432 directions. A likelihood ratio test indicated there was significant negative genetic covariance
433 between TCICHK and DM remaining after accounting for all 17 of these SNPs and *PSY2* as covariates
434 in a linear model (LR-statistic = 9.976, $p = 0.0025$). Across all SNP association models fit in the
435 bivariate GWAS, the residual genetic covariance was consistently negative, ranging from -0.34 to -
436 0.19. Together, the peak SNPs accounted for 8.6% and the *PSY2* SNP accounted for 5.6% of the
437 negative covariance between traits.

438 Since the phenotypes were normalized with a Box-Cox transformation before analysis,
439 bivariate association models (Eqn. 2) were re-fit with untransformed, de-regressed BLUPs for the

17 peak SNPs to translate their effect size estimates to the traits' original units. For both TCICHK and DM, effect estimates from untransformed data were linearly related ($r^2 > 0.99$) to effect estimates from Box-Cox-transformed data with slopes of 5.8 and 6.7, respectively (Fig. S4). SNP effect estimates from the untransformed BLUPs ranged from -2.42 to 2.56 μg/g FW for TCICHK and -1.91 to 1.72% for DM.

Table 2. Summary statistics of SNPs that are jointly-associated with TCICHK and DM. The SNP with the lowest p-value was selected from each group of associated SNPs. 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. SNP effect estimates (additive effect of the ALT allele) and standard errors (SE) are with respect to the transformed, de-regressed BLUP values. PVE and PCE (proportion variance explained and proportion covariance explained, respectively) are with respect to total genetic variance/covariance. Negative PCE values indicate the SNP weakens the overall negative genetic covariance between traits by contributing to covariance the opposite direction.

Peak SNP (chr#_position)	Bivariate association p-value	TC effect	TC effect SE	DM effect	DM effect SE	Effect direction (TC/DM)	REF allele	ALT allele	Allele freq. (ALT)	TC PVE	DM PVE	TC/DM PCE
S1_23484417	8.81e-06 **	0.424	0.093	-0.217	0.101	+/-	C	G	0.92	0.061	0.018	0.031
S10_2963170	1.17e-05 **	-0.316	0.069	0.145	0.076	-/+	A	C	0.13	0.051	0.012	0.023
S3_3292443	4.76e-05 *	-0.022	0.072	-0.310	0.071	-/-	A	T	0.16	<0.001	0.067	-0.004
S6_3256937	6.38e-05 *	0.370	0.084	-0.056	0.088	+/-	G	C	0.06	0.035	0.001	0.005
S13_25853831	2.00e-04 *	-0.283	0.069	0.067	0.071	-/+	C	T	0.19	0.056	0.004	0.013
S17_20165942	2.66e-04 *	0.249	0.076	0.143	0.080	+/+	C	A	0.12	0.030	0.011	-0.017
S13_23763339	2.82e-04 *	-0.217	0.058	0.116	0.060	-/+	A	G	0.29	0.044	0.014	0.023
S5_22756786	2.84e-04 *	0.105	0.046	-0.169	0.047	+/-	A	T	0.76	0.009	0.027	0.015
S3_294409	3.12e-04 *	-0.254	0.083	0.266	0.091	-/+	T	A	0.08	0.021	0.025	0.021
S2_2073445	3.43e-04 *	0.249	0.064	0.027	0.071	+/+	C	A	0.82	0.043	0.001	-0.005
S6_23101017	3.79e-04 *	-0.368	0.094	0.004	0.102	-/+	T	G	0.06	0.035	<0.001	<0.001
S17_17210448	5.47e-04 *	-0.411	0.106	0.073	0.108	-/+	C	A	0.06	0.047	0.002	0.008
S2_8411339	5.65e-04 *	-0.065	0.062	-0.224	0.064	-/-	A	G	0.21	0.003	0.043	-0.011
S10_20300591	5.80e-04 *	0.164	0.050	-0.127	0.051	+/-	T	A	0.85	0.016	0.010	0.012
S12_5428221	6.55e-04 *	-0.013	0.053	0.205	0.054	-/+	C	A	0.38	<0.001	0.051	0.003
S18_6776129	7.33e-04 *	0.131	0.073	0.219	0.074	+/+	A	C	0.15	0.010	0.030	-0.016
S5_1950967	8.67e-04 *	-0.248	0.090	-0.199	0.102	-/-	T	G	0.08	0.020	0.015	-0.016

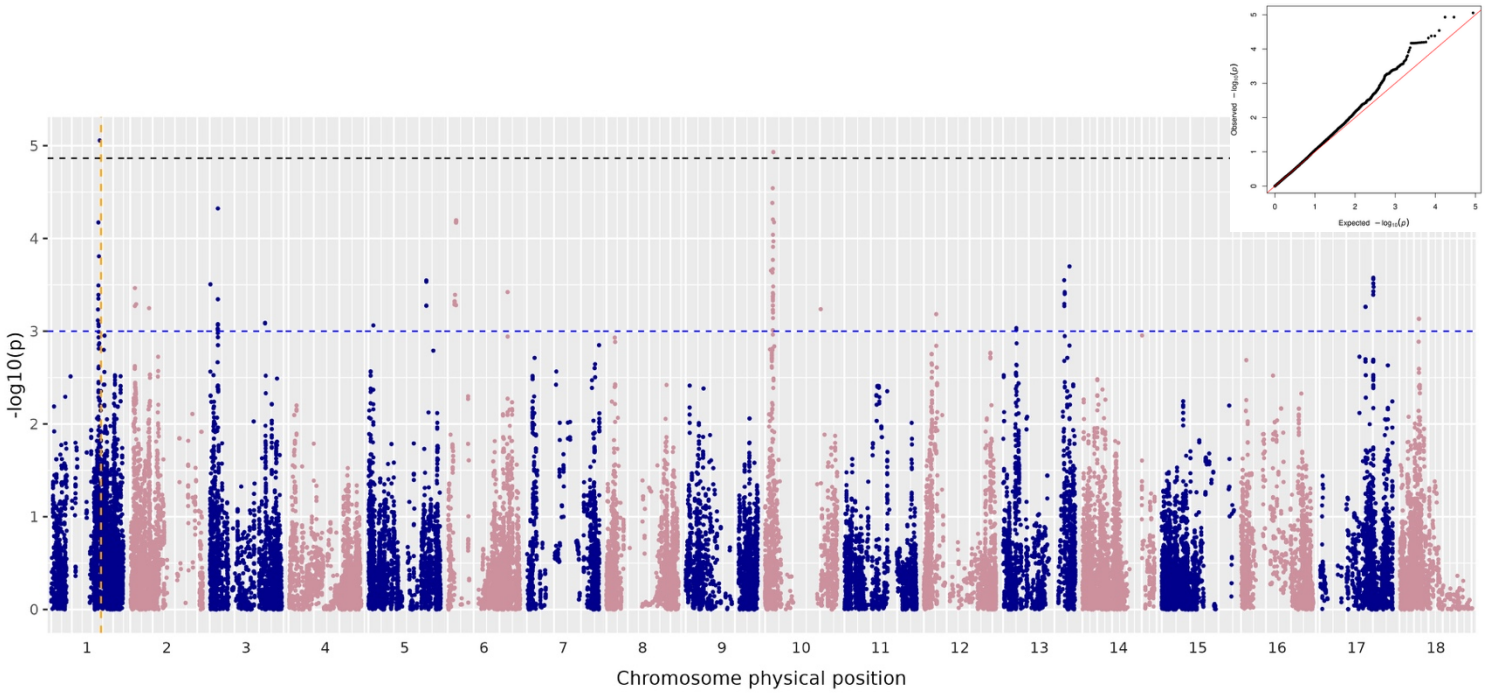
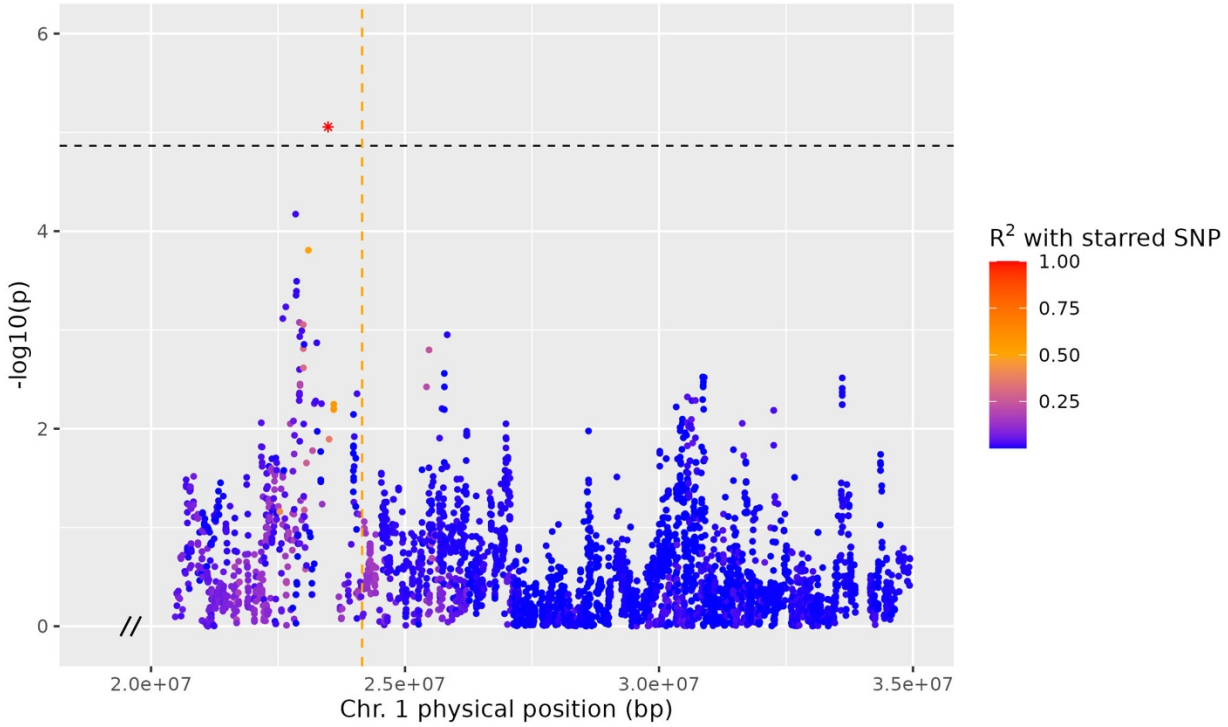
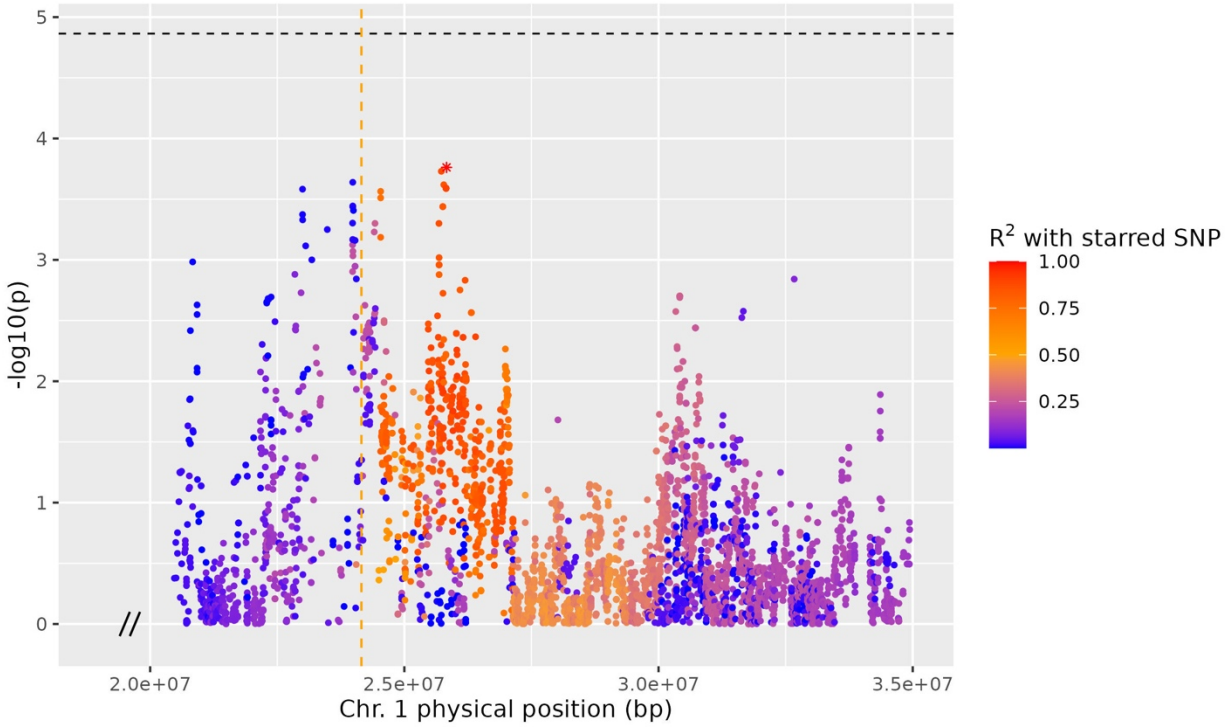


Figure 4. Manhattan plot showing p-values from bivariate Wald tests of SNP effects on TCICHK and DM, accounting for genomic relatedness and the *PSY2* causal SNP. The vertical orange dashed line indicates the position of the *PSY2* causal SNP. The black dashed line indicates the modified Bonferroni significance 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

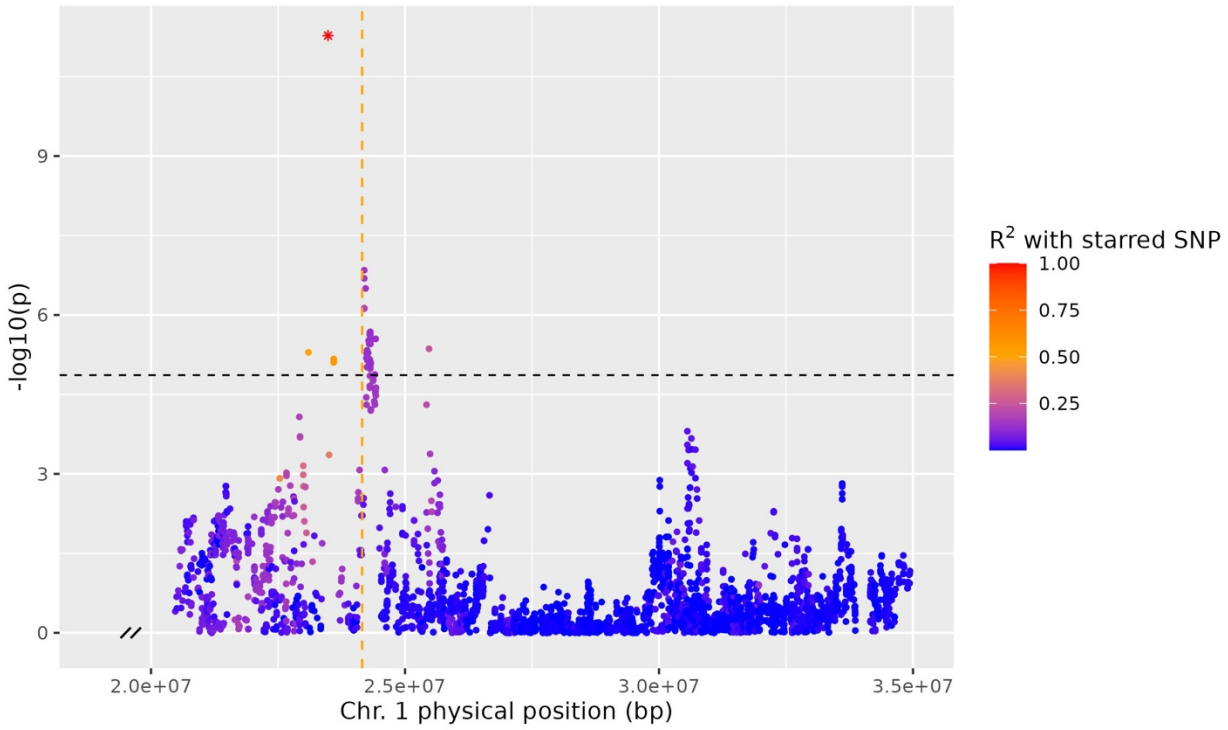


Figure 5. Manhattan plots for A) bivariate TCICHK and DM, B) univariate DM, and C) univariate TCICHK associations on chromosome 1 near the *PSY2* QTL region from 20–35 Mbp. The vertical orange line indicates the position of the *PSY2* causal SNP. Points are colored by their LD (pairwise R^2) with the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

469 *Model comparison*

470 The bivariate association model was compared with univariate models with one trait as the
471 response variable and the other as a covariate to determine whether the associations were direct or
472 indirect. For each of the top SNPs from 17 peaks that passed the relaxed significance threshold, the
473 bivariate model had the lowest AIC compared to the univariate models (Table 3). For 13 of the 17
474 peak SNPs, the SNP effect retained its significance at the $p < 0.001$ threshold in a univariate model
475 for one trait with the other as a covariate. For nine of these SNPs, the TCICHK univariate model
476 retained its significance, indicating support for a direct association with TC and indirect association
477 with DM. The remaining four SNPs retained their significance in the DM univariate model,
478 suggesting a direct association with DM and an indirect association with TC. In no cases did the SNP
479 retain significance in both univariate models with the other trait as a covariate.

480

481 **Table 3.** Comparison of bivariate and reciprocal univariate association models for the 17 peak SNPs
482 identified as jointly associated at the relaxed threshold ($p < 0.001$). "TCICHK model" refers to an
483 association between a SNP and TCICHK with DM as a covariate (Eqn. 4), while the "DM model" is
484 the reverse (Eqn. 5). P-values are from univariate or bivariate Wald tests of SNP effects. P-values
485 are annotated with two asterisks if they pass the modified Bonferroni significance threshold and a
486 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
1_23484417	-47.28	3.07	0.56	8.81e-06 **	3.85e-09 **	6.25E-03	Direct TC, indirect DM
10_2963170	-46.83	4.37	2.79	1.17e-05 **	3.86e-05 *	2.36E-01	Direct TC, indirect DM
3_3292443	-44.93	22.62	-12.52	4.76e-05 *	7.81E-02	8.81e-06 **	Direct DM, indirect TC
6_3256937	-45.12	2.62	3.47	6.38e-05 *	2.12e-06 **	9.44E-01	Direct TC, indirect DM
13_25853831	-42.37	-0.53	3.48	2.00e-04 *	1.24e-06 **	7.54E-01	Direct TC, indirect DM
17_20165942	-42.81	5.79	-0.05	2.66e-04 *	1.19E-03	3.76E-03	NA
13_23763339	-41.57	10.15	6.37	2.82e-04 *	1.01e-04 *	1.56E-01	Direct TC, indirect DM
5_22756786	-39.40	12.37	-0.16	2.84e-04 *	6.98E-02	3.25E-03	NA
3_294409	-42.28	13.24	-4.89	3.12e-04 *	2.69E-03	1.66E-02	NA
2_2073445	-41.11	2.00	6.21	3.43e-04 *	1.11e-04 *	2.80E-01	Direct TC, indirect DM
6_23101017	-42.63	5.03	5.28	3.79e-04 *	3.13e-04 *	3.41E-01	Direct TC, indirect DM
17_17210448	-42.49	1.55	4.88	5.47e-04 *	3.13e-05 *	9.20E-01	Direct TC, indirect DM
2_8411339	-41.16	21.43	-6.49	5.65e-04 *	1.90E-02	7.69e-04 *	Direct DM, indirect TC
10_20300591	-38.65	12.62	4.68	5.80e-04 *	4.35E-03	1.46E-01	NA
12_5428221	-39.56	23.14	-11.18	6.55e-04 *	6.57E-01	5.23e-05 *	Direct DM, indirect TC
18_6776129	-41.06	18.37	-4.00	7.33e-04 *	4.75E-02	9.09e-05 *	Direct DM, indirect TC
5_1950967	-39.50	13.10	-1.26	8.67e-04 *	5.27e-06 **	1.24E-01	Direct TC, indirect DM

Univariate associations for all traits

Several peaks passing the modified Bonferroni significance threshold were identified in univariate analyses: TCICHK (7), DM (2), TCHART (4) (Fig. 6). TCICHK and TCHART QTL overlapped on chromosomes 1 and 13. 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 S4. Among other agronomic traits analyzed (DYLD, FRYLD, RTWT, HI, SPRTPERC, CMD3I, and CMD3S), 2 QTL for SPRTPERC and 1 for CMD3I were identified at the Bonferroni significance level (Fig. S5).

Candidate gene analysis

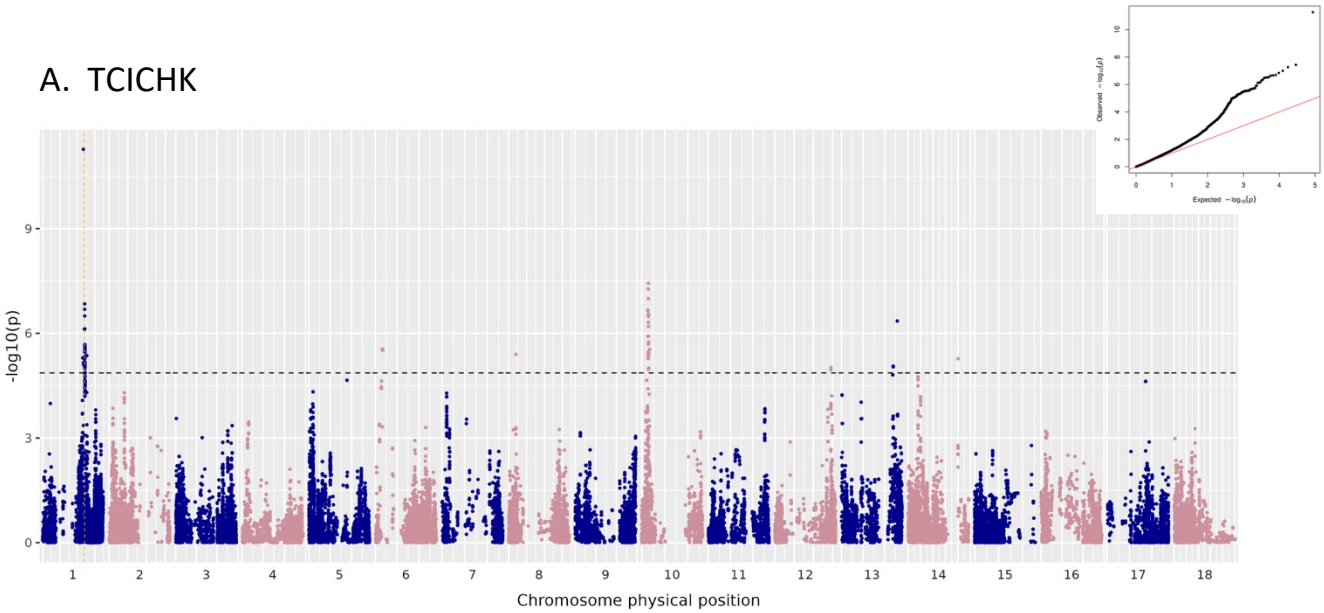
Annotated genes within a window of 100kb from the bivariate peak were identified as candidates based on the rate of LD decay. On most chromosomes, LD decayed to an average R^2 of 0.25 at approximately 100kb physical distance, although chromosomes 1 and 4 had more extensive LD, likely due to the presence of large *M. glaziovii* introgressions on these chromosomes (Bredeson et al. 2016; Wolfe et al. 2019; Chan et al. 2022) (Fig. S6).

Candidate genes within 100kb of bivariate peak SNPs are listed in Table S5. None of the *a priori* candidate genes for the carotenoid or starch pathways were identified in this list. The peak

506 SNP on chromosome 1 fell within the gene Manes.01G115400, annotated as a MYB-like DNA
507 binding protein. The peak SNP on chromosome 10 fell within the coding region of
508 Manes.10G034900, *ATP-binding cassette transporter B family member 20-related.* Candidate genes
509 for other agronomic traits (DYLD, FRYLD, RTWT, HI, SPRTPERC, CMD3I, and CMD3S) from
510 univariate analyses are listed in Table S6.

511

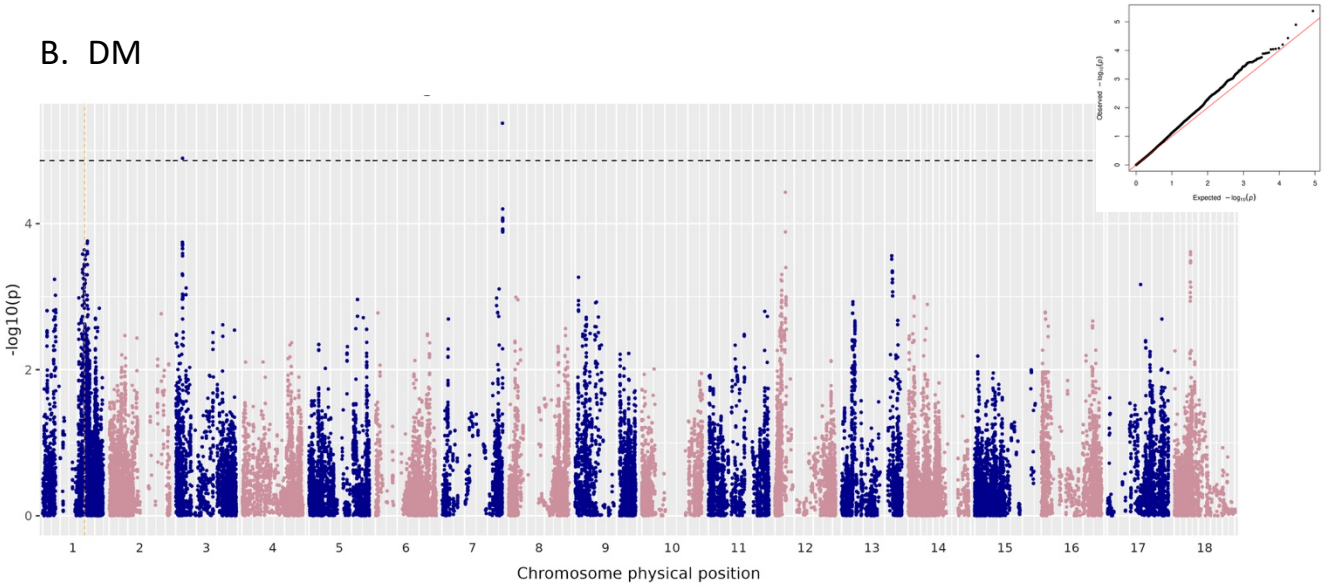
A. TCICHK



512

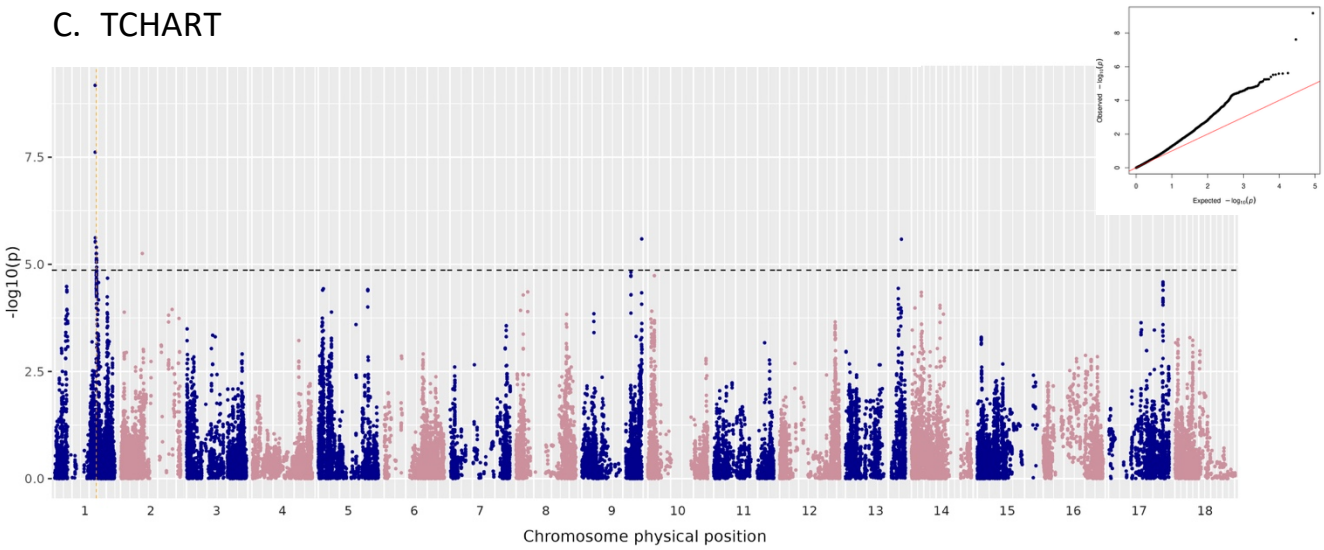
513

B. DM



514

C. TCHART



515 **Figure 6.** Manhattan plots for univariate GWAS for TCICHK (A), DM (B), and TCHART (C). The
516 dashed black line indicates the modified Bonferroni significance threshold at α = 0.05. The vertical
517 dashed orange line indicates the position of the *PSY2* causal locus used as a covariate. Q-Q plots for
518 each analysis are at right.
519
520 **Discussion**
521 The main objective of this study was to test the hypothesis of mediated pleiotropy between
522 TC and DM in cassava by examining genome-wide contributions to the negative covariance between
523 traits. Bivariate mixed linear models were used to estimate the additive effects of each SNP on TC
524 and DM while controlling for the effects of *PSY2* and genome-wide population structure. Four main
525 points of evidence for mediated pleiotropy were considered: whether there are multiple QTL jointly
526 associated with both traits, the extent of residual genetic covariance between traits after accounting
527 for identified QTL, whether QTL tend to affect the traits in opposite directions, and whether QTL
528 affect the traits in a direct or indirect manner.
529 The bivariate genome-wide associations identified two QTL passing the modified
530 Bonferroni significance threshold at α = 0.05 on chromosomes 1 (S1_23484417) and 10
531 (S10_2963170) for joint association with TCICHK and DM. To our knowledge, the QTL on
532 chromosome 10 jointly associated with TC and DM has not previously been reported. Rabbi et al.
533 (2017) did show a single significant SNP associated only with TCHART on chromosome 10, but it
534 does not appear to be in the same location as the QTL peak identified here. These associations
535 therefore represent new observations of apparently pleiotropic QTL.
536 The incorporation of more diverse, early-stage selections within the yellow cassava
537 breeding population likely facilitated the identification of these new loci. None of the TCICHK or
538 TCHART univariate association peak SNPs overlapped with the DM univariate association peaks on
539 chromosomes 3 and 7, despite the presence of the two significant bivariate TCICHK and DM peaks.
540 This observation suggests that the increased power of the bivariate association model facilitated
541 the identification of these joint associations which would have been missed if the two univariate
542 associations were simply compared.
543 The estimated effect sizes of the chromosome 1 and 10 loci correspond to +2.28 μg/g FW
544 TC / -1.42% DM and -1.67 μg/g FW TC / 0.92% DM, respectively (Fig. S3). While these effect sizes
545 are large enough to potentially warrant marker-assisted selection strategies for individual traits,
546 since these loci affect the traits in opposite directions, a genomics-assisted breeding method
547 incorporating information from many markers would likely be more useful. In addition, the

548 frequencies of the favorable TC allele at both loci are already quite high, consistent with the history
549 of selection for yellow color in this breeding population (Fig. S7).
550 An additional 15 loci passed the relaxed significance threshold of $p < 0.001$ for joint
551 association with TCICHK and DM. After accounting for all 17 of the peak SNPs at these loci and *PSY2*
552 as covariates in a linear model, there was still statistically significant negative genetic covariance
553 remaining between TCICHK and DM. With the putative QTL and *PSY2* accounting for 14.2% of the
554 genetic covariance in total, the polygenic background still explained the majority of the covariance.
555 Therefore, we conclude that there are significant polygenic contributions to the genetic correlation
556 between TCICHK and DM, disrupting the hypothesis that *PSY2* is the main driver of the relationship.
557 We then tested whether QTL are more likely to contribute negatively to the covariance
558 between traits than would be expected by chance. Under the hypothesis of a trade-off for carbon
559 precursors between starch and carotenoid synthesis, any variants upregulating metabolic flux
560 towards one pathway would be expected take away from the other pathway. Though the 17 peak
561 SNPs used in the binomial test are not exactly independent (all SNPs were identified in only one
562 dataset), we treated them as such given that they were mostly on different chromosomes and had
563 low R^2 with each other ($R^2 \leq 0.095$). In setting a more relaxed threshold for analyses on effect
564 direction, we assumed that any false-positive peaks would be equally likely to affect the traits in the
565 same or in opposite directions and hypothesized that an inherent biologically-driven trade-off
566 would result in mostly opposite-direction effects.
567 While most peak SNPs (11 of 17) affected the traits in opposite directions, a binomial test of
568 effect directions indicated that this pattern was not statistically significant. Therefore, while the
569 overall genetic covariance is strongly negative, we cannot conclude an inherent underlying trade-
570 off drives consistent opposite-direction effects across all QTL. The identification of several loci that
571 affect the traits positively in the same direction, and variation among the magnitude of opposite-
572 direction effect sizes such that not all effects were equal and opposite, suggests potential for making
573 genetic gain on both traits. However, since none of the peaks affecting traits in the same direction
574 passed the high-confidence significance threshold, we cannot recommend them as targets for
575 marker-assisted selection unless further research can validate their effects. Rather, genomic
576 selection leveraging many small effects throughout the genome may be a more effective approach.
577 To infer whether these joint associations were direct or mediated, we used model
578 comparisons to test whether jointly-associated SNPs retained their associations with each
579 individual trait after accounting for the other trait as a covariate. Both the chromosome 1 and 10
580 QTL peak SNPs retained significance in the univariate model for TCICHK after accounting for DM,

581 but not in the univariate model for DM accounting for TCICHK. This result suggests that the QTL are
582 more likely to be directly associated with TCICHK and indirectly associated with DM. Among all the
583 suggestively-associated SNPs, the model comparison analysis supported a direct relationship with
584 TCICHK for 9 SNPs and a direct relationship with DM for 4 SNPs, while no SNPs seemed to be
585 directly associated with both traits simultaneously. These results support the hypothesis of
586 mediated rather than direct pleiotropy (Fig. 1). Though the directionality of the mediating
587 mechanism remains unclear, there is stronger evidence for TCICHK being the mediating trait since
588 the majority of peak SNPs support that model. In addition, these data would support the hypothesis
589 of a pleiotropy mechanism mediated by a latent variable such as a phytohormone signal, in which
590 case it would be difficult to distinguish between direct/indirect and indirect/direct statistical
591 associations (Fig. 1) (Stephens 2013).
592 Similarly, the lack of *a priori* candidate genes identified among the significant bivariate
593 GWAS hits supports the hypothesis that the interaction between carotenoids and starch is mediated
594 indirectly through regulatory pathways rather than direct pleiotropic action of genes in the
595 biosynthetic pathways or direct competition for precursors (Villwock et al. 2024). The two
596 previously identified candidate genes for DM on chromosome 1, *sucrose synthase* and *ADP-glc*
597 *PPase***,** did not show up as GWAS hits in this study, either because they were not causally
598 polymorphic or because their effects were statistically eliminated by the *PSY2* covariate used in the
599 association models.
600 Of the identified candidate genes (Table S5), several have potential biological relevance to
601 the TCICHK and DM relationship. The gene containing the chromosome 1 peak SNP at
602 S1_23484417, Manes.01G115400, is annotated as homologous to *Arabidopsis MYB4* (Table S5). In
603 *Arabidopsis, MYB4* has been shown to affect leaf carotenoid content via transcriptional regulation of
604 photooxidative stress response (Agarwal et al. 2020; Banerjee et al. 2024). While it is unclear
605 whether Manes.01G115400 has an analogous function to *Arabidopsis MYB4,* other MYB-family
606 transcription factors play roles in regulating carotenoid content as well (Sagawa et al. 2016; Zhu et
607 al. 2017; Song et al. 2023).
608 At the chromosome 10 peak, the *ABC transporter* candidate gene is relevant to the
609 hypothesis that ATP import into the amyloplast is a limiting factor for both starch and carotenoid
610 synthesis (Villwock et al. 2024). Upstream 9.7kb from the chromosome 10 peak is *PsbP-like protein*
611 *2*, a chloroplastic protein involved in alternative electron transport to protect against
612 photooxidative stress (Ishihara et al. 2007). In addition, an *AQUAPORIN/tonoplast intrinsic protein*
613 (Manes.10G035000; *MeTIP1;4*) 21kb from the chromosome 10 peak SNP, is involved in vacuolar

614 water and small solute transport (Zou and Yang 2019). Water content is an indirect component of
615 both TC and DM measured per unit fresh weight. In *Arabidopsis, TIP3* expression is induced by an
616 ABA-responsive transcription factor (Mao and Sun 2015). Therefore, *MeTIP1;4* could hypothetically
617 be involved in increasing water uptake in response to higher ABA synthesis downstream of the
618 carotenoid pathway, resulting in lower dry matter percentage. Candidate gene validation is needed
619 to test these hypotheses in cassava and other starchy crops like sweetpotato exhibiting similar
620 negative genetic relationships between TC and DM (Villwock et al. 2024). Future research will
621 examine the co-expression patterns of candidate genes in cassava roots.
622 While the chromosome 1 QTL at 1_23484417 is relatively near the *PSY2* causal variant at
623 1_24155522, it appears to be a distinct QTL. Since the *PSY2* causal SNP (1_24155522) was imputed
624 with a DR2 value of 0.74, it is possible there were imputation errors such that this covariate did not
625 fully account for the *PSY2* effect. The *PSY2* SNP also explained a smaller proportion of the genetic
626 variation in TCICHK (2.4%) than expected based on previous studies. Other methods of accounting
627 for the *PSY2* effect in the association models were tested, including by fitting the closest non-
628 imputed SNP to 1_24155522 as a fixed effect covariate, using the *PSY2* region (within 20kb of the
629 coding region) haplotype as a random effect factor covariate, and using the first three PCs from a
630 PCA of SNPs in the *PSY2* region as fixed effects. None of these methods significantly affected the
631 results of an additional peak on chromosome 1 (data not shown). In addition, the 1_23484417 SNP
632 was in low LD with SNPs in the *PSY2* region (Fig. 5). Therefore, we assume that the effect of the
633 1_24155522 variant in *PSY2* was effectively captured in the model. Since the new chromosome 1
634 QTL is 690kb upstream of the *PSY2* gene, it is possible that this QTL is capturing a *cis*-regulatory
635 element of *PSY2.*
636 While imputing low-density genotype data to higher SNP density increases genetic mapping
637 resolution, it also introduces imputation error. Using imputed genotype data has become standard
638 practice in many genomic prediction and selection pipelines, in which low-cost genotyping assays
639 are needed to screen large breeding populations, and is becoming more common in GWAS as well
640 (Niehoff et al. 2022; Zhao et al. 2025). Using imputed genotype data has been shown to increase
641 statistical power despite the addition of noise due to imputation error (Guan and Stephens 2008).
642 To mitigate loss of statistical power due to imputation error, we retained the allele dosage
643 estimates as decimals rather than rounded integers to help account for the uncertainty due to
644 imputation error (Kutalik et al. 2011; Auer et al. 2023). Nevertheless, individual imputed marker-
645 trait associations should be interpreted with caution and genotyped directly for QTL validation in
646 future studies.

647 Q-Q plots from bivariate genome-wide associations with TCICHK and DM showed
648 acceptable control of inflation from the genomic relationship matrix and the *PSY2* covariate. In the
649 univariate analyses, there was considerable Q-Q plot inflation for TCICHK and TCHART. Although
650 population structure and the *PSY2* were accounted for in these models, strong selection pressure
651 for root yellowness as the primary trait in the yellow cassava breeding population likely created
652 additional structure around causal loci. In practice, white-fleshed and yellow-fleshed cassava roots
653 are easily distinguished, but quantitative variation among cream-colored roots blurs the line
654 between them. After subsetting the data to retain only the yellow subpopulation, the positions of
655 Manhattan plot peaks were similar, although with generally higher p-values reflecting lower power
656 (data not shown). Therefore, it is unlikely that the major univariate QTL peaks for TCICHK and
657 TCHART were driven by structure between the white and yellow subpopulations.
658 The precision of TCICHK phenotypes did not appear to outweigh the advantage of having
659 many more TCHART observations, since TCHART had a higher H^2 (0.73) compared to that of
660 TCICHK (0.49). (Table 1). However, negative genetic relationships with yield traits (DRYLD, FRYLD,
661 RTWT) were evident with TCICHK but not with TCHART. While TCICHK and colorimetric
662 measurements are both highly correlated with HPLC measurements of TC (Jaramillo et al. 2018),
663 TCHART may be influenced by other factors that affect visual perception of root color, while
664 TCICHK may be more specific to carotenoids. The negative correlation between TCICHK and
665 SPRTPERC (sprouting percentage; $r = -0.25$, $p < 0.05$) in these data is notable. This relationship has
666 been previously observed in the field, though not formally described to our knowledge (P. Kulakow,
667 personal communication, July 2021). We hypothesize that this negative correlation could be driven
668 by the influence of carotenoid synthesis on downstream ABA, which regulates bud dormancy (Chen
669 et al. 2020).
670 In summary, the presence of two new jointly-associated QTL identified at a high confidence
671 level, residual negative genetic covariance after accounting for 17 putatively associated SNPs
672 tending to affect the traits in opposite directions, and a consistent pattern of direct/indirect
673 statistical associations with the two traits supports the hypothesis of polygenic, mediated
674 pleiotropy driving the relationship between TC and DM. While some plausible candidate genes
675 were identified, none were directly in the carotenoid or starch biosynthesis pathways.
676 Furthermore, there was not clear evidence for the direction of the mediation of one trait on the
677 other, and one QTL affecting the two traits in the same direction was also identified. This suggests
678 that the relationship between TC and DM is likely not the result of a simple trade-off or a single
679 causal chain, but more complex regulatory interactions. In addition, new QTL for TC were identified

680 that expand our understanding of the genetic basis of TC accumulation in cassava beyond the *PSY2*
681 locus.
682 This study represents a new application for multivariate GWAS that goes beyond identifying
683 individual variants to test hypotheses about traits' genetic relationships across the genome. Our
684 approach builds on previous multivariate GWAS methods that explicitly estimate SNP effects on
685 each trait, which has a built-in interpretation framework, as opposed to the difficult-to-interpret
686 single p-values given from an overall multivariate association. A downside to our method is that the
687 computational time is high, and it may not be feasible to extend to more than two traits. More
688 efficient but less flexible implementations similar to our multivariate GWAS method are available in
689 the MTMM and GEMMA software (Korte et al. 2012; Zhou and Stephens 2014). Multivariate GEMMA
690 output very similar GWAS results with the data used here and ran much faster, but GEMMA does
691 not offer the same flexibility in specifying and analyzing covariance structures as ASReml-based
692 models do. Future studies could improve on these methods by improving the computational time
693 while retaining the modeling flexibility.
694 Overall, the identification of polygenic mediated pleiotropy in this study demonstrates a
695 multivariate GWAS framework for investigating the genetic architecture of correlated traits,
696 providing insights to guide both applied breeding strategies and future research into underlying
697 biological mechanisms.
698
699 **Data availability**
700 All datasets and R code used for these analyses are available on GitHub
701 (http://www.github.com/serenvillwock/mvGWAS). Raw field trial data and imputed genotype files
702 are additionally available on http://www.cassavabase.org/ under the public list
703 'yellowGWAS23_trials_plusCASS'. The reference panel data used for genotype imputation is
704 available at
705 https://www.cassavabase.org/ftp/marnin_datasets/nextgenImputation2019/ImputationStageIII_7
706 2619/. *Manihot esculenta* v6.1 reference genome and gene annotations are available on Phytozome
707 at https://phytozome-next.jgi.doe.gov/info/Mesculenta_v6_1.
708
709 **Acknowledgements**
710 We are grateful to the Cassava Breeding Unit team at the International Institute of Tropical
711 Agriculture, Ibadan, Nigeria, especially including Lynda Adaobi Nnadi, Adenike Damilola Ige, Peter
712 Iluebby, Patrick Akpotuzor, Victor Monday Chukwuyem, Toye Ayankanmi, and Kehinde Nafiu for

713 their assistance with tissue sampling and field trials; Prasad Peteti, Ekanem Ukoabasi, and Mahmud
714 Kehinde Dhikrullah for their assistance with data curation; and Peter Kulakow, Peace Ikeanyi, and
715 Richard Ofei for their administrative support. The development of the breeding populations used in
716 this study was sponsored by HarvestPlus and the NextGen Cassava project. We are grateful to Kelly
717 Robbins and to the Cornell University BRC Bioinformatics Core Facility (RRID:SCR_021757) for
718 providing access to computational resources, Michael Gore for insightful discussions, and Salvador
719 Gezan at VSN International for providing ASReml-R technical support. We thank Andrew Siefert for
720 statistical consultation. The authors thank the UK's Foreign, Commonwealth & Development Office
721 (FCDO) and the Bill & Melinda Gates Foundation (Grant INV 007637; www.gatesfoundation.org) for
722 their financial support. SSV is supported by the USDA National Institute of Food and Agriculture
723 AFRI Predoctoral Fellowship project accession no. 1030847.

724 **References**

725 Abincha, W., Dzidzienyo, D. K., Wesonga, L. N., Mwimali, M., Ozimati, A., Kayondo, I. S., et al. (2024).
726 Enhancing Nutrition: A Review of Provitamin A Carotenoid Cassava Breeding Initiatives in East
727 Africa. *Advances in Agriculture* 2024. doi: 10.1155/aia/9937876

728 Afolami, I., Mwangi, M. N., Samuel, F., Boy, E., Ilona, P., Talsma, E. F., et al. (2021). Daily consumption of
729 pro-vitamin A biofortified (yellow) cassava improves serum retinol concentrations in preschool
730 children in Nigeria: A randomized controlled trial. *American Journal of Clinical Nutrition* 113, 221–
731 231. doi: 10.1093/ajcn/nqaa290

732 Agarwal, P., Mitra, M., Banerjee, S., and Roy, S. (2020). MYB4 transcription factor, a member of R2R3-
733 subfamily of MYB domain protein, regulates cadmium tolerance via enhanced protection against
734 oxidative damage and increases expression of PCS1 and MT1C in Arabidopsis. *Plant Science* 297.
735 doi: 10.1016/j.plantsci.2020.110501

736 Akinwale, M. G., Aladesanwa, R. D., Akinyele, B. O., Dixon, A. G. O., and Odiyi, A. C. (2010). Inheritance of
737 β-carotene in cassava (Manihot esculenta crantza). *Int J Genet Mol Biol* 2, 198–201.

738 Anokye, B., Potter Jr, B. W., Amoah, P., and Olasanmi, B. (2025). Genetic Diversity Assessment of
739 Hydrocyanic Acid, Total Carotenoid, and Dry Matter Content in Biofortified Cassava Using SNP
740 Markers. doi: 10.20944/preprints202504.1694.v1

741 Asare-Marfo, D., Birol, E., Gonzalez, C., Moursi, M., Perez, S., Schwarz, J., et al. (2013). Prioritizing
742 Countries for Biofortification Interventions Using Country-Level Data. Available at:
743 http://ebrary.ifpri.org/cdm/ref/collection/p15738coll2/id/127837 (Accessed November 14, 2023).

744 Atser, G. (2014). Nigeria releases more cassava with higher pro-vitamin A to fight micronutrient
745 deficiency. Available at: International Institute of Tropical Agriculture, www.reliefweb.int
746 (Accessed December 9, 2021).

747 Auer, P. L., Wang, G., Li, G., DeWan, A. T., and Leal, S. M. (2023). Comparison of multiple imputation and
748 other methods for the analysis of imputed genotypes. *BMC Genomics* 24. doi: 10.1186/s12864-
749 023-09415-0

750 Banerjee, S., Agarwal, P., Choudhury, S. R., and Roy, S. (2024). MYB4, a member of R2R3-subfamily of
751 MYB transcription factor functions as a repressor of key genes involved in flavonoid biosynthesis
752 and repair of UV-B induced DNA double strand breaks in Arabidopsis. *Plant Physiology and*
753 *Biochemistry* 211. doi: 10.1016/j.plaphy.2024.108698

754 Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., et al. (2023). UniProt: the
755 Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 51, D523–D531. doi:
756 10.1093/nar/gkac1052

757 Beyene, G., Solomon, F. R., Chauhan, R. D., Gaitán-Solis, E., Narayanan, N., Gehan, J., et al. (2018).
758 Provitamin A biofortification of cassava enhances shelf life but reduces dry matter content of
759 storage roots due to altered carbon partitioning into starch. *Plant Biotechnol J* 16, 1186–1200. doi:
760 10.1111/pbi.12862

761 Box, G. E. P., and Cox, D. R. (1964). An Analysis of Transformations. *Journal, Source Statistical, Royal*
762 *Series, Society* 26, 211–252.

763 Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-Gonzales, E., et al. (2016).
764 Sequencing wild and cultivated cassava and related species reveals extensive interspecific
765 hybridization and genetic diversity. *Nat Biotechnol* 34, 562–570. doi: 10.1038/nbt.3535

766 Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-
767 Generation Reference Panels. *Am J Hum Genet* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015

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

771 Caspi, R., Billington, R., Keseler, I., Kothari, A., Krummenacker, M., Midford, P., et al. (2020). The
772 MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res* 48,
773 D445–D453. doi: 10.1093/nar/gkz862

774 Ceballos, H., Morante, N., Sánchez, T., Ortiz, D., Aragón, I., Chávez, A. L., et al. (2013). Rapid cycling
775 recurrent selection for increased carotenoids content in cassava roots. *Crop Sci* 53, 2342–2351. doi:
776 10.2135/cropsci2013.02.0123

777 Cervantes-Flores, J. C., Sosinski, B., Pecota, K. V., Mwanga, R. O. M., Catignani, G. L., Truong, V. D., et al.
778 (2011). Identification of quantitative trait loci for dry-matter, starch, and β-carotene content in
779 sweetpotato. *Molecular Breeding* 28, 201–216. doi: 10.1007/s11032-010-9474-5

780 Chan, A. W., Villwock, S. S., Williams, A. L., and Jannink, J.-L. (2022). Sexual dimorphism and the effect of
781 wild introgressions on recombination in cassava (Manihot esculenta *Crantz*) breeding germplasm.
782 *G3 Genes|Genomes|Genetics* 12. doi: 10.1093/g3journal/jkab372

783 Chen, K., Li, G. J., Bressan, R. A., Song, C. P., Zhu, J. K., and Zhao, Y. (2020). Abscisic acid dynamics,
784 signaling, and functions in plants. *J Integr Plant Biol* 62, 25–54. doi: 10.1111/jipb.12899

785 Decourcelle, M., Perez-Fons, L., Baulande, S., Steiger, S., Couvelard, L., Hem, S., et al. (2015). Combined
786 transcript, proteome, and metabolite analysis of transgenic maize seeds engineered for enhanced
787 carotenoid synthesis reveals pleotropic effects in core metabolism. *J Exp Bot* 66, 3141–3150. doi:
788 10.1093/jxb/erv120

789 Diretto, G., Al-Babili, S., Tavazza, R., Scossa, F., Papacchioli, V., Migliore, M., et al. (2010). Transcriptional-
790 metabolic networks in β-carotene-enriched potato tubers: The long and winding road to the
791 golden phenotype. *Plant Physiol* 154, 899–912. doi: 10.1104/pp.110.159368

792 Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package
793 rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

794 Esuma, W., Herselman, L., Labuschagne, M. T., Ramu, P., Lu, F., Baguma, Y., et al. (2016a). Genome-wide
795 association mapping of provitamin A carotenoid content in cassava. *Euphytica* 212, 97–110. doi:
796 10.1007/s10681-016-1772-5

Esuma, W., Kawuki, R. S., Herselman, L., and Labuschagne, M. T. (2016b). Diallel analysis of provitamin A carotenoid and dry matter content in cassava (Manihot esculenta Crantz). *Breed Sci* 66, 627–635. doi: 10.1270/jsbbs.15159

Esuma, W., Ozimati, A., Kulakow, P., Gore, M. A., Wolfe, M. D., Nuwamanya, E., et al. (2021). Effectiveness of genomic selection for improving provitamin A carotenoid content and associated traits in cassava. *G3: Genes, Genomes, Genetics* 11. doi: 10.1093/g3journal/jkab160

FAO (2021). FAOSTAT. *License: CC BY-NC-SA 3.0 IGO*. Available at: https://www.fao.org/faostat/en/#data/SCL (Accessed April 3, 2024).

Galesloot, T. E., Van Steen, K., Kiemeney, L. A. L. M., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PLoS One* 9, 1–8. doi: 10.1371/journal.pone.0095923

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

Garrick, D. J., Taylor, J. F., and Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution* 41, 1–8. doi: 10.1186/1297-9686-41-55

Gemenet, D. C., da Silva Pereira, G., De Boeck, B., Wood, J. C., Mollinari, M., Olukolu, B. A., 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, 23–36. doi: 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, 169–180. doi: 10.1016/j.copbio.2017.02.001

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40, D1178–D1186.

Guan, Y., and Stephens, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genet* 4. doi: 10.1371/journal.pgen.1000279

Gutschker, S., Ruescher, D., Rabbi, I. Y., Rosado-Souza, L., Pommerrenig, B., Pauly, M., 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

HarvestPlus (2024). HarvestPlus Database of Biofortified Crops Released -- Vitamin A Cassava. Available at: https://bcr.harvestplus.org/varieties_released_list?crop=Vitamin_A_Cassava (Accessed June 26, 2024).

Holland, J. B., Nyquist, W. E., and Cervantes-Martínez, C. T. (2002). “Estimating and Interpreting Heritability for Plant Breeding: An Update,” in *Plant Breeding Reviews*, ed. J. Janick (John Wiley & Sons, Inc.), 9–112. doi: 10.1002/9780470650202.ch2

834 Ige, A. D., Olasanmi, B., Bauchet, G. J., Kayondo, I. S., Mbanjo, E. G. N., Uwugiaren, R., et al. (2022).
835 Validation of KASP-SNP markers in cassava germplasm for marker-assisted selection of increased
836 carotenoid content and dry matter content. *Front Plant Sci* 13. doi: 10.3389/fpls.2022.1016170

837 Ikeogu, U. N., Akdemir, D., Wolfe, M. D., Okeke, U. G., Chinedozi, A., Jannink, J. L., et al. (2019). Genetic
838 Correlation, Genome-Wide Association and Genomic Prediction of Portable NIRS Predicted
839 Carotenoids in Cassava Roots. *Front Plant Sci* 10, 1–11. doi: 10.3389/fpls.2019.01570

840 Ikeogu, U. N., Davrieux, F., Dufour, D., Ceballos, H., Egesi, C. N., and Jannink, J. L. (2017). Rapid analyses
841 of dry matter content and carotenoids in fresh cassava roots using a portable visible and near
842 infrared spectrometer (Vis/NIRS). *PLoS One* 12. doi: 10.1371/journal.pone.0188918

843 Ishihara, S., Takabayashi, A., Ido, K., Endo, T., Ifuku, K., and Sato, F. (2007). Distinct functions for the two
844 PsbP-like proteins PPL1 and PPL2 in the chloroplast thylakoid lumen of Arabidopsis. *Plant Physiol*
845 145, 668–679. doi: 10.1104/pp.107.105866

846 Jaramillo, A. M., Londoño, L. F., Orozco, J. C., Patiño, G., Belalcazar, J., Davrieux, F., et al. (2018). A
847 comparison study of five different methods to measure carotenoids in biofortified yellow cassava
848 (Manihot esculenta). *PLoS One* 13, 1–14. doi: 10.1371/journal.pone.0209702

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

852 Kutalik, Z., Johnson, T., Bochud, M., Mooser, V., Vollenweider, P., Waeber, G., et al. (2011). Methods for
853 testing association between uncertain genotypes and quantitative traits. *Biostatistics* 12, 1–17. doi:
854 10.1093/biostatistics/kxq039

855 Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). "Identifying Outlying Y Observations -
856 Studentized Deleted Residuals," in *Applied linear statistical models*, eds. B. Gordon, R. T. Hercher
857 Jr., and L. Stone (McGraw-Hill/Irwin), 390–398.

858 Li, L., Yang, Y., Xu, Q., Owsiany, K., Welsch, R., Chitchumroonchokchai, C., et al. (2012). The Or gene
859 enhances carotenoid accumulation and stability during post-harvest storage of potato tubers. *Mol*
860 *Plant* 5, 339–352. doi: 10.1093/mp/ssr099

861 Mackay, T. F. C., and Anholt, R. R. H. (2024). Pleiotropy, epistasis and the genetic architecture of
862 quantitative traits. *Nat Rev Genet* 25, 639–657. doi: 10.1038/s41576-024-00711-3

863 Mao, Z., and Sun, W. (2015). Arabidopsis seed-specific vacuolar aquaporins are involved in maintaining
864 seed longevity under the control of ABSCISIC ACID INSENSITIVE 3. *J Exp Bot* 66, 4781–4794. doi:
865 10.1093/jxb/erv244

866 Maroya, N. G., Kulakow, P. A., Parkes, E. Y., Friedrich, S. K., Schweigert, F., Alamu, E., et al. (2012).
867 Quantification of total carotene content of yellow root cassava genotypes using iCheckTM
868 Carotene in comparison with the standard spectrophotometer method. doi:
869 10.13140/2.1.4534.4007

Mbanjo, E. G. N., Ogungbesan, A., Agbona, A., Akpotuzor, P., Toyinbo, S., Iluebbey, P., et al. (2024). Validation of SNP Markers for Diversity Analysis, Quality Control, and Trait Selection in a Biofortified Cassava Population. *Plants* 13. doi: 10.3390/plants13162328

Nandudu, L., Strock, C., Ogbonna, A., Kawuki, R., and Jannink, J. L. (2024). Genetic analysis of cassava brown streak disease root necrosis using image analysis and genome-wide association studies. *Front Plant Sci* 15. doi: 10.3389/fpls.2024.1360729

Niehoff, T., Pook, T., Gholami, M., and Beissinger, T. (2022). Imputation of low-density marker chip data in plant breeding: Evaluation of methods based on sugar beet. *Plant Genome* 15. doi: 10.1002/tpg2.20257

Njoku, D. N., Gracen, V. E., Offei, S. K., Asante, I. K., Egesi, C. N., Kulakow, P., et al. (2015). Parent-offspring regression analysis for total carotenoids and some agronomic traits in cassava. *Euphytica* 206, 657–666. doi: 10.1007/s10681-015-1482-4

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

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

Perveen, S., Zafar, S., Iqbal, N., and Riaz, M. (2021). “26. Provitamin A Carotenoids,” in *Carotenoids: Structure and Function in the Human Body*, eds. M. Zia-Ul-Haq, S. Dewanjee, and M. Riaz (Springer Nature Switzerland AG), 775–797. doi: 10.1007/978-3-030-46459-2

Rabbi, I. Y., Kayondo, S. I., Bauchet, G., Yusuf, M., Aghogho, C. I., Ogunpaimo, K., et al. (2022). Genome-wide association analysis reveals new insights into the genetic architecture of defensive, agro-morphological and quality-related traits in cassava. *Plant Mol Biol* 109, 195–213. doi: 10.1007/s11103-020-01038-3

Rabbi, I. Y., Udoh, L. I., Wolfe, M., Parkes, E. Y., Gedil, M. A., Dixon, A., et al. (2017). Genome-Wide Association Mapping of Correlated Traits in Cassava: Dry Matter and Total Carotenoid Content. *Plant Genome* 10, 1–14. doi: 10.3835/plantgenome2016.09.0094

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

Rodriguez-Amaya, D. B., and Kimura, M. (2004). HarvestPlus Handbook for Carotenoid Analysis. *HarvestPlus Technical Monographs*, 59.

Sagawa, J. M., Stanley, L. E., Lafountain, A. M., Frank, H. A., Liu, C., and Yuan, Y. W. (2016). An R2R3-MYB transcription factor regulates carotenoid pigmentation in Mimulus lewisii flowers. *New Phytologist* 209, 1049–1057. doi: 10.1111/nph.13647

Sánchez, T., Chávez, A. L., Ceballos, H., Rodriguez-Amaya, D. B., Nestel, P., and Ishitani, M. (2006). Reduction or delay of post-harvest physiological deterioration in cassava roots with higher carotenoid content. *J Sci Food Agric* 86, 634–639. doi: 10.1002/jsfa.2371

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits : challenges and strategies. *Nature Publishing Group* 14. doi: 10.1038/nrg3461

Song, J., Sun, B., Chen, C., Ning, Z., Zhang, S., Cai, Y., et al. (2023). An R-R-type MYB transcription factor promotes non-climacteric pepper fruit carotenoid pigment biosynthesis. *Plant Journal* 115, 724–741. doi: 10.1111/tpj.16257

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

Stevens, G. A., Bennett, J. E., Hennocq, Q., Lu, Y., De-Regil, L. M., Rogers, L., et al. (2015). Trends and mortality effects of vitamin A deficiency in children in 138 low-income and middle-income countries between 1991 and 2013: A pooled analysis of population-based surveys. *Lancet Glob Health* 3, e528–e536. doi: 10.1016/S2214-109X(15)00039-X

Tahirou, A., Bamire, A. S., Oparinde, A., and Akinola, A. A. (2015). Determinants of Adoption of Improved Cassava Varieties among Farming Households in Oyo, Benue, and Akwa Ibom States of Nigeria. *HarverstPlus Working Paper* 22, 21.

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

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

Turner, S. D. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J Open Source Softw* 3, 731. doi: 10.21105/joss.00731

Udoh, L. I., Gedil, M., Parkes, E. Y., Kulakow, P., Adesoye, A., Nwuba, C., et al. (2017). Candidate gene sequencing and validation of SNP markers linked to carotenoid content in cassava (Manihot esculenta Crantz). *Molecular Breeding* 37. doi: 10.1007/s11032-017-0718-5

Villwock, S. S., Li, L., and Jannink, J. (2024). Carotenoid-carbohydrate crosstalk: evidence for genetic and physiological interactions in storage tissues across crop species. *New Phytologist* 244, 1709–1722. doi: 10.1111/nph.20196

Welsch, R., Arango, J., Bär, C., Salazar, B., Al-Babili, S., Beltrán, J., et al. (2010). Provitamin a accumulation in cassava (Manihot esculenta) roots driven by a single nucleotide polymorphism in a phytoene synthase gene. *Plant Cell* 22, 3348–3356. doi: 10.1105/tpc.110.077560

Wiseman, E. M., Bar-El Dadon, S., and Reifen, R. (2017). The vicious cycle of vitamin A deficiency: A review. *Crit Rev Food Sci Nutr* 57, 3703–3714. doi: 10.1080/10408398.2016.1160362

Wolfe, M. (2022). genomicMateSelectR: Genomic Mate Selection. R package version 0.2.0. Available at: https://github.com/wolfemd/genomicMateSelectR

Wolfe, M. D., Bauchet, G. J., Chan, A. W., Lozano, R., Ramu, P., Egesi, C., et al. (2019). Historical introgressions from a wild relative of modern cassava improved important traits and may be under balancing selection. *Genetics* 213, 1237–1253. doi: 10.1534/genetics.119.302757

944 Zhao, H., MacLeod, I. M., Keeble-Gagnere, G., Barbulescu, D. M., Tibbits, J. F., Kaur, S., et al. (2025).
945 Using genotype imputation to integrate Canola populations for genome-wide association and
946 genomic prediction of blackleg resistance. *BMC Genomics* 26. doi: 10.1186/s12864-025-11250-4

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

949 Zhu, F., Luo, T., Liu, C., Wang, Y., Yang, H., Yang, W., et al. (2017). An R2R3-MYB transcription factor
950 represses the transformation of a-and b-branch carotenoids by negatively regulating expression of
951 CrBCH2 and CrNCED5 in flavedo of Citrus reticulate. *New Phytologist* 216, 178–192. doi:
952 10.1111/nph.14684

953 Zou, Z., and Yang, J. (2019). Genome-wide comparison reveals divergence of cassava and rubber
954 aquaporin family genes after the recent whole-genome duplication. *BMC Genomics* 20. doi:
955 10.1186/s12864-019-5780-4

956

957

Supplementary Information

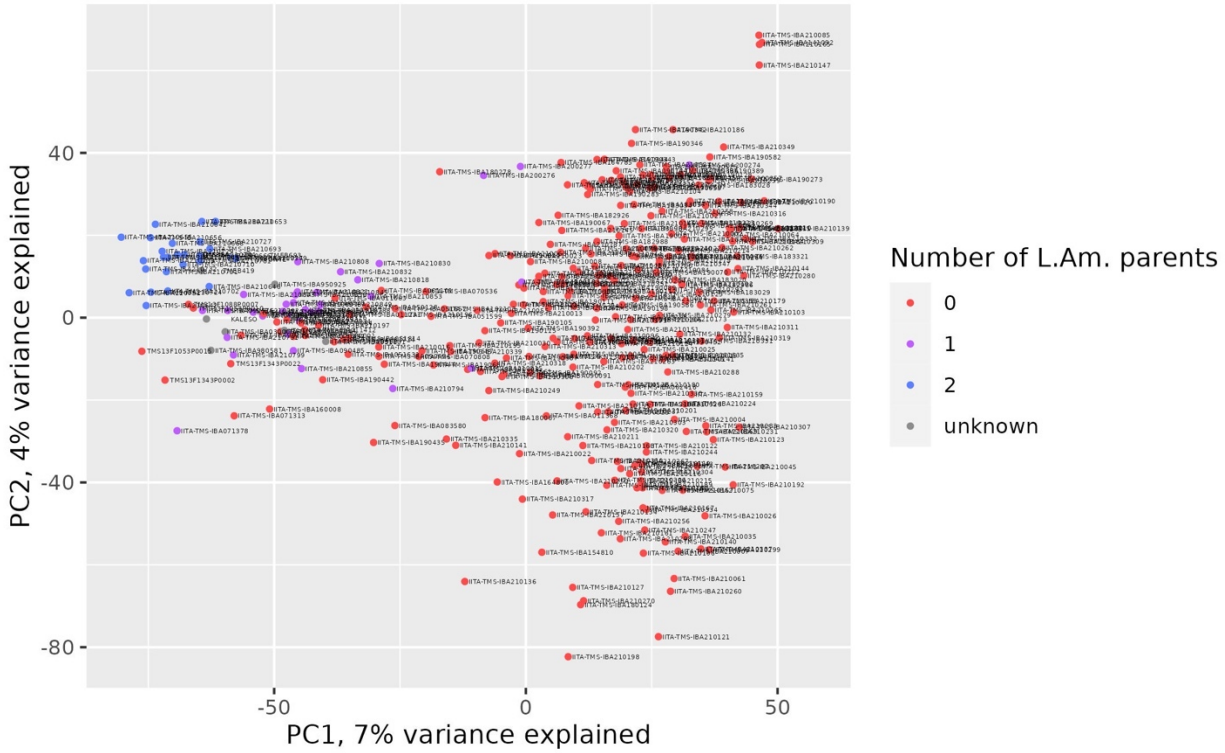
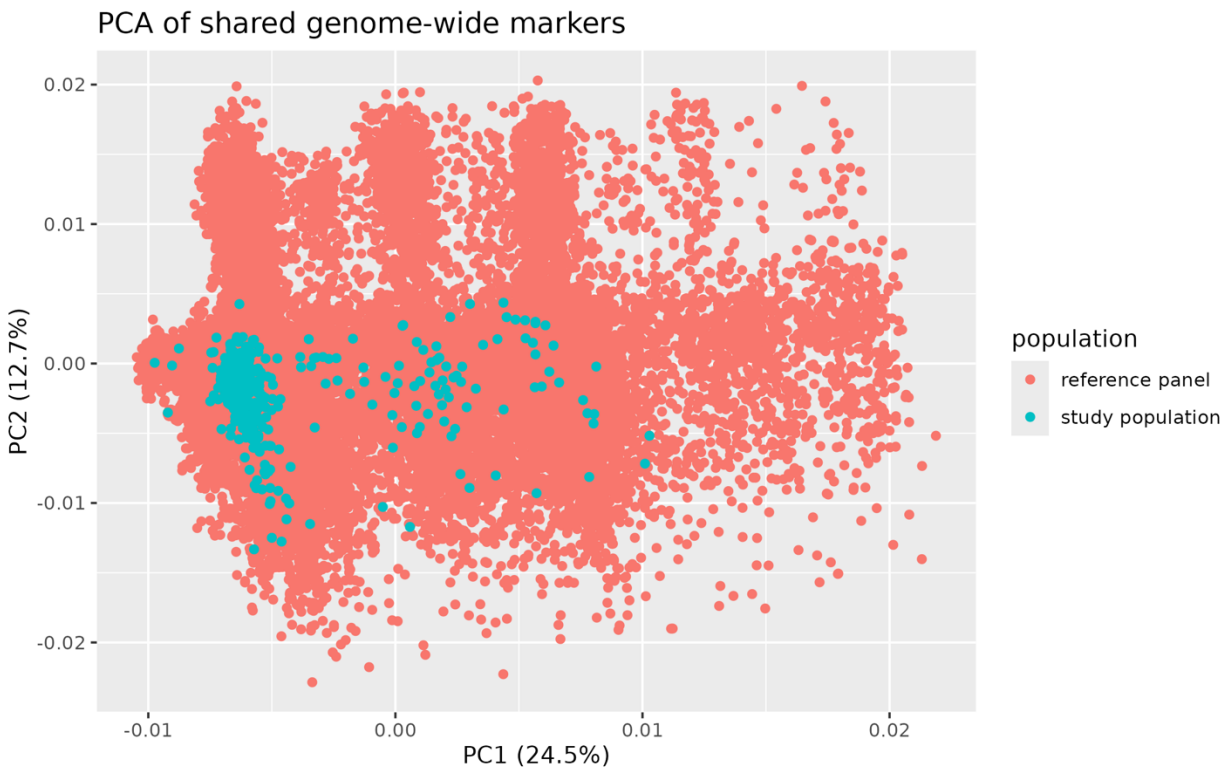


Figure S1. Principal component analysis of genotyped accessions shows that the first principal component (PC1) separates a cluster of accessions with recent Latin American ancestry. Color represents the number of Latin American parents listed in an accession's pedigree.



964

965 **Figure S2.** Principal component analysis of 2,939 genome-wide genetic markers shared across the

966 imputation reference panel and the study population.

967

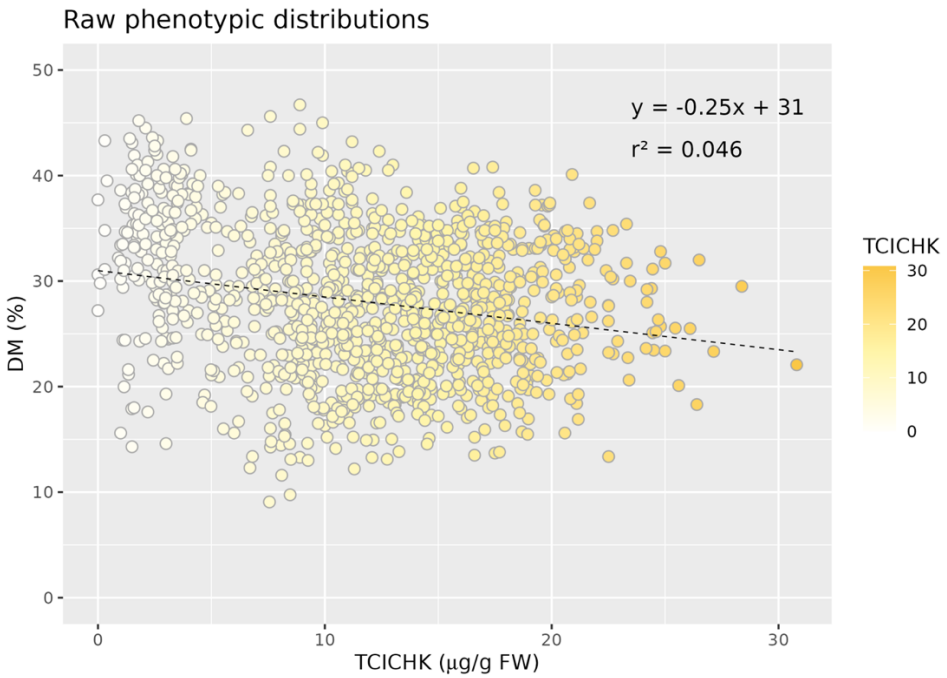


Figure S3. Raw phenotypic distributions of percent dry matter (DM) and total carotenoids measured by iCheck (TCICHK) in µg/g fresh weight. Points represent individual plot-level observations.

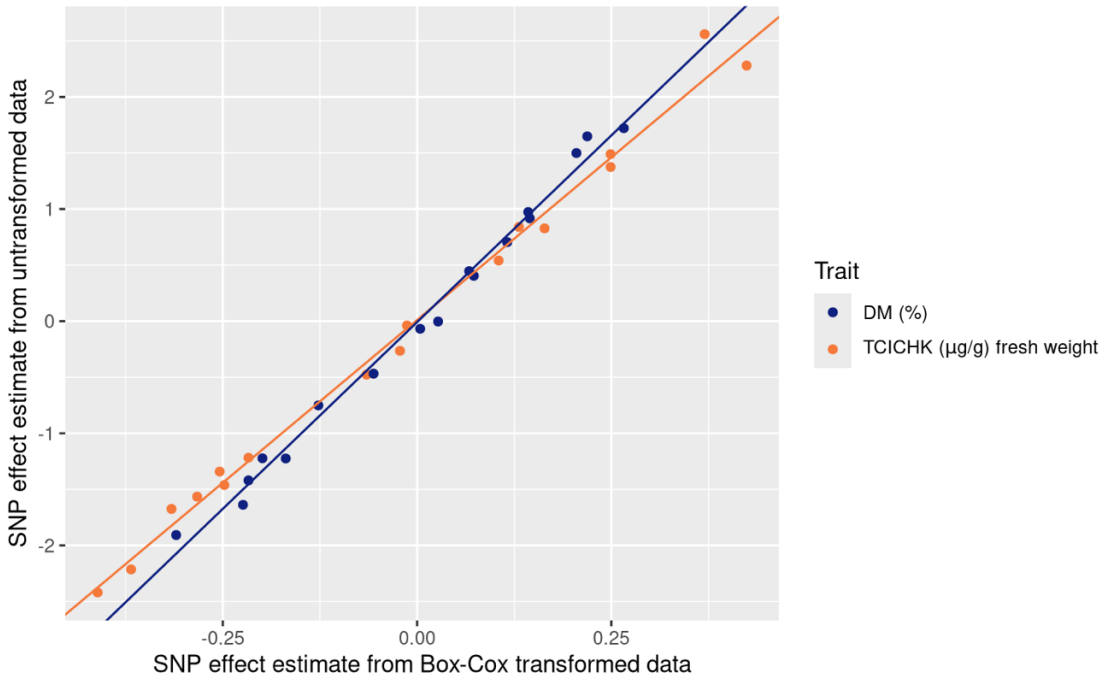
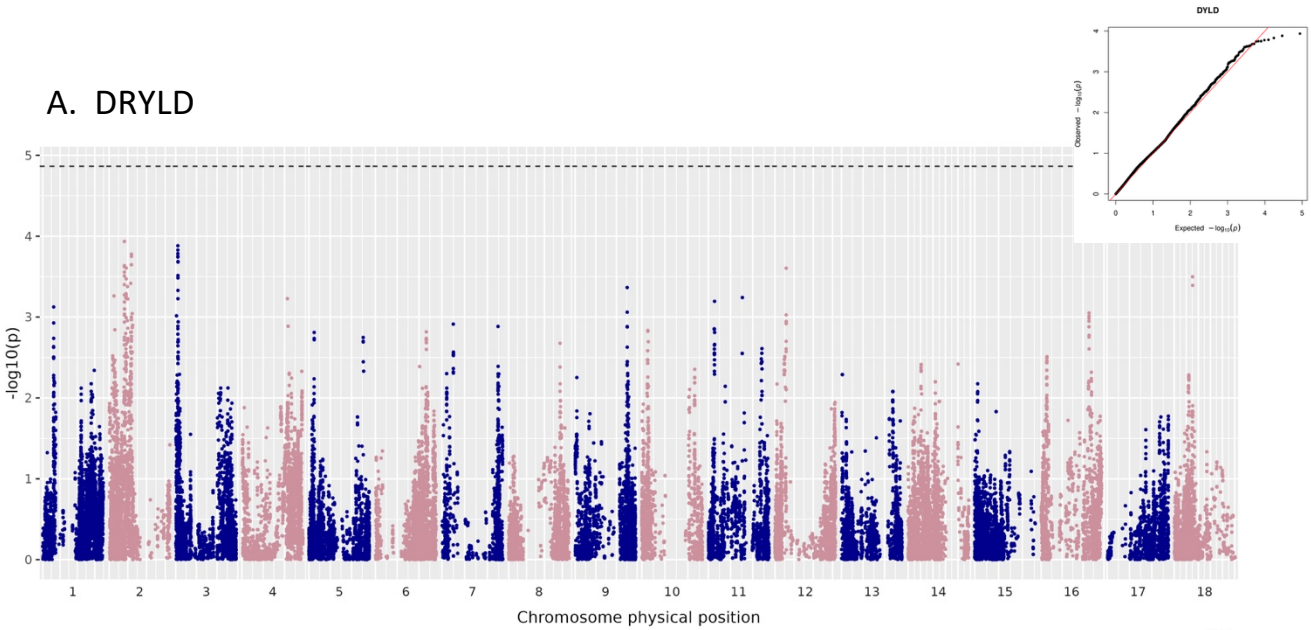


Figure S4. Comparison of peak SNP effect estimates from Box-Cox normalized data and from untransformed data show a linear relationship. SNP effect estimates from the untransformed data correspond to the traits' original units.

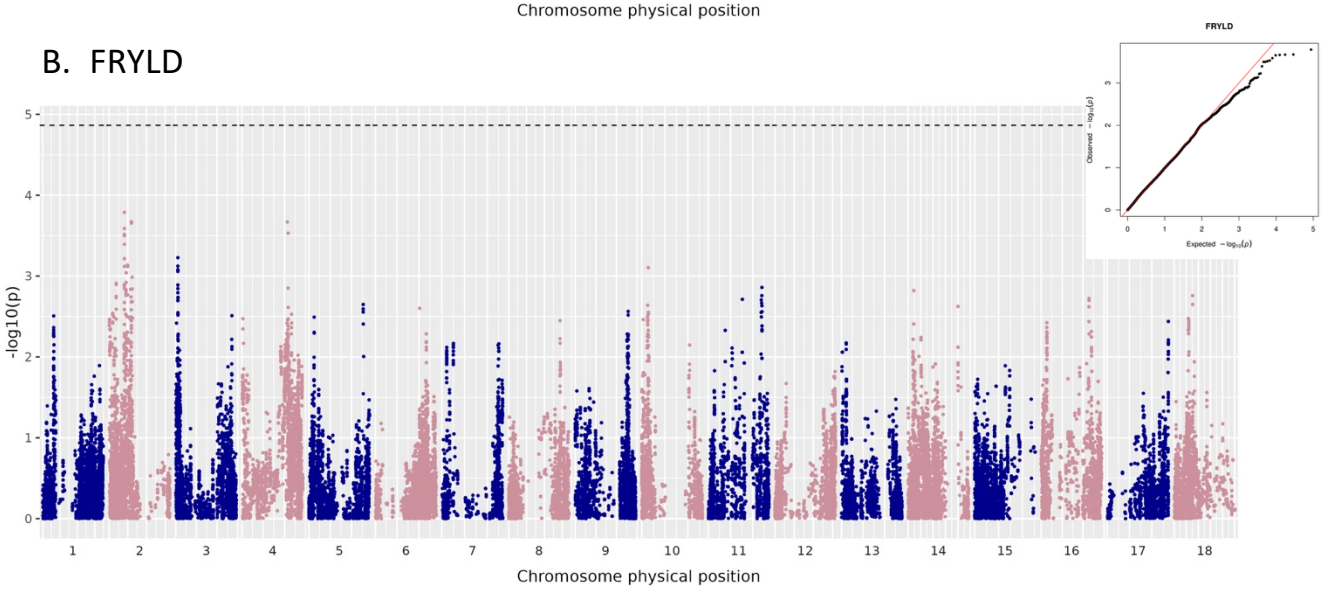
976

A. DRYLD



977

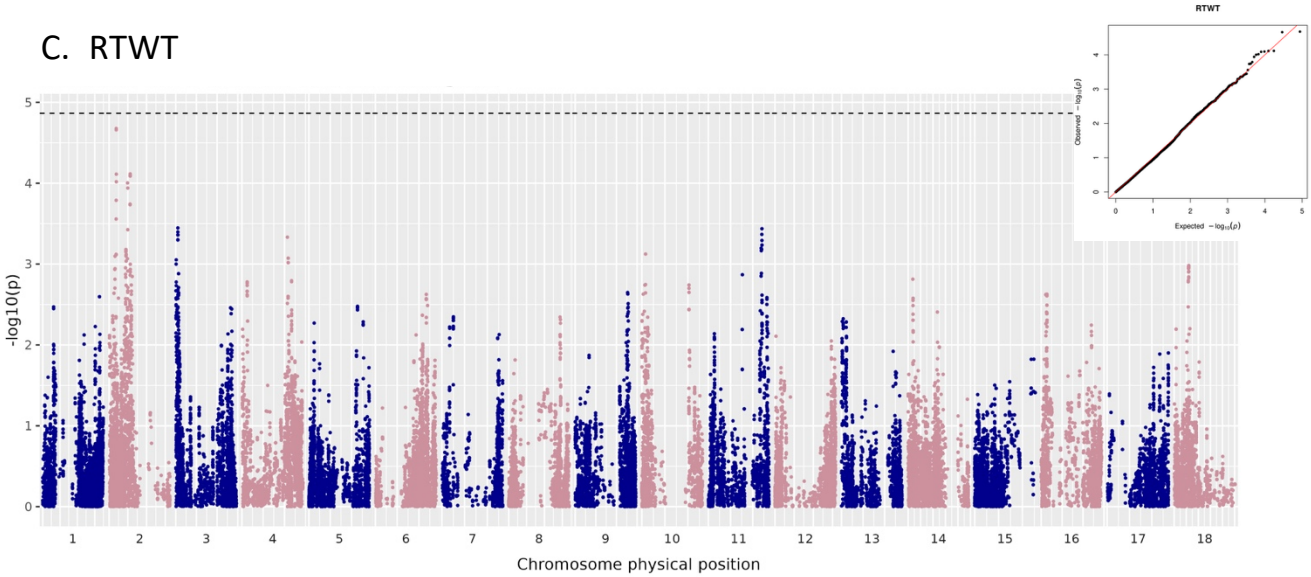
B. FRYLD



978

979

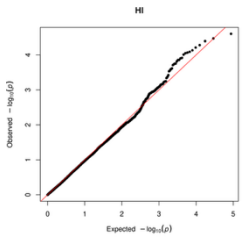
C. RTWT

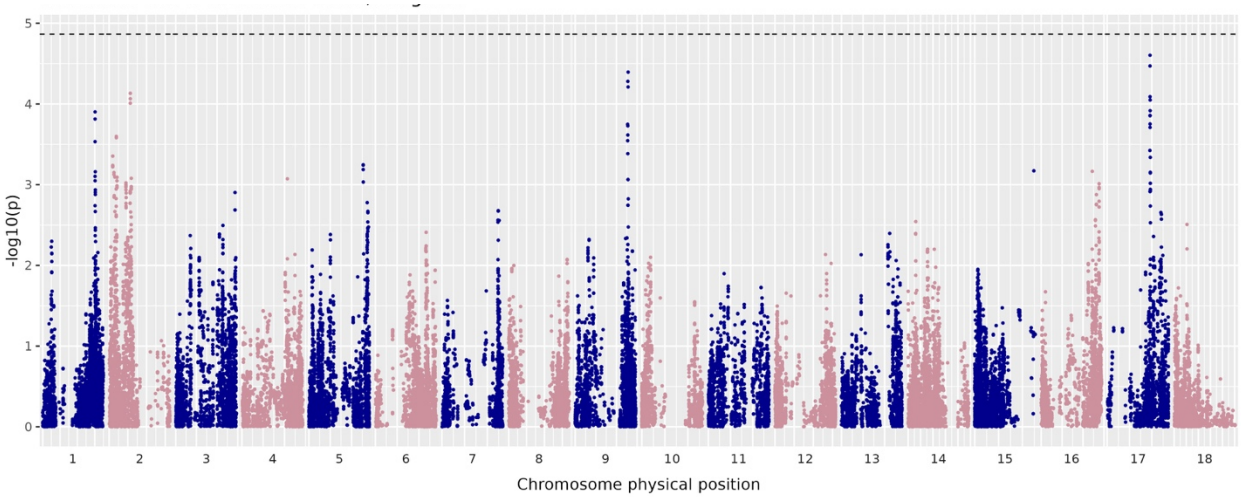


980

981

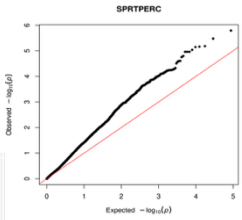
D. HI

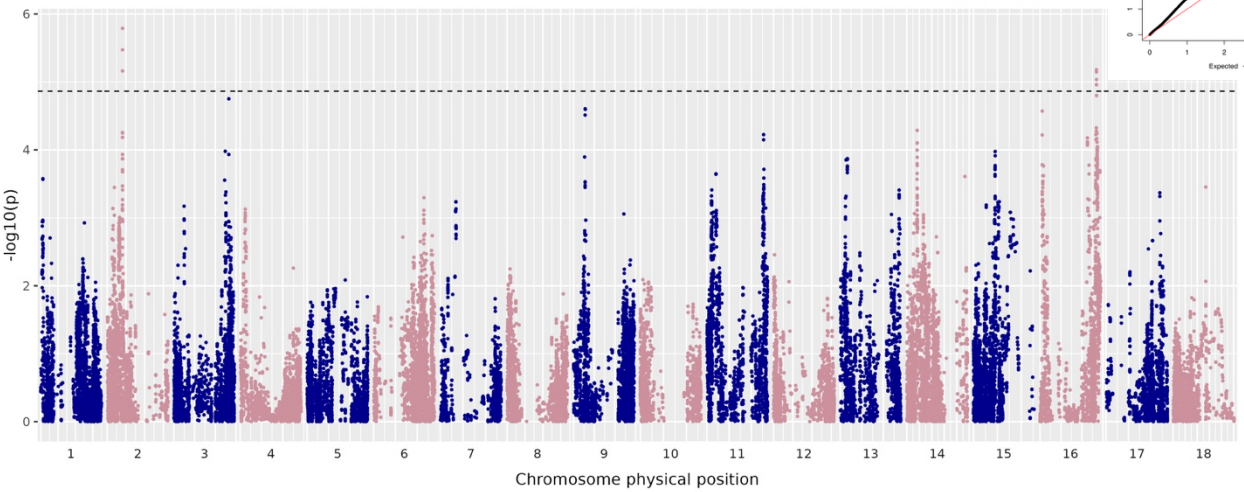




982

E. SPRTPERC





983

984

985

986

987

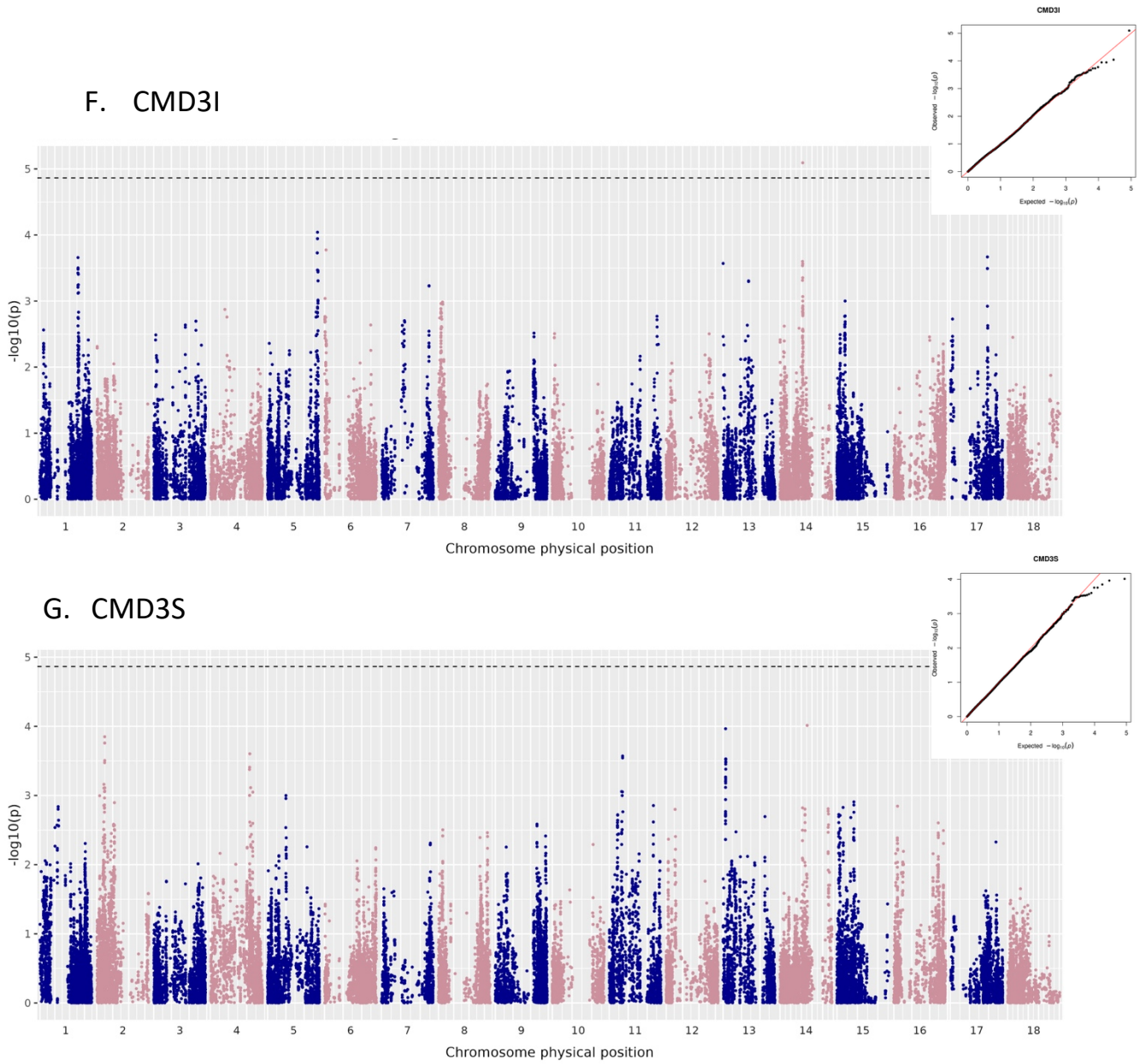


Figure S5. Manhattan plots for univariate GWAS for analyzed agronomic traits: DRYLD (A), FRYLD (B), RTWT (C), HI (D), SPRTPERC (E), CMD3I (F), and CMD3S (G). The dashed black line indicates the modified Bonferroni significance threshold at α = 0.05. Q-Q plots for each analysis are in the upper right corner.

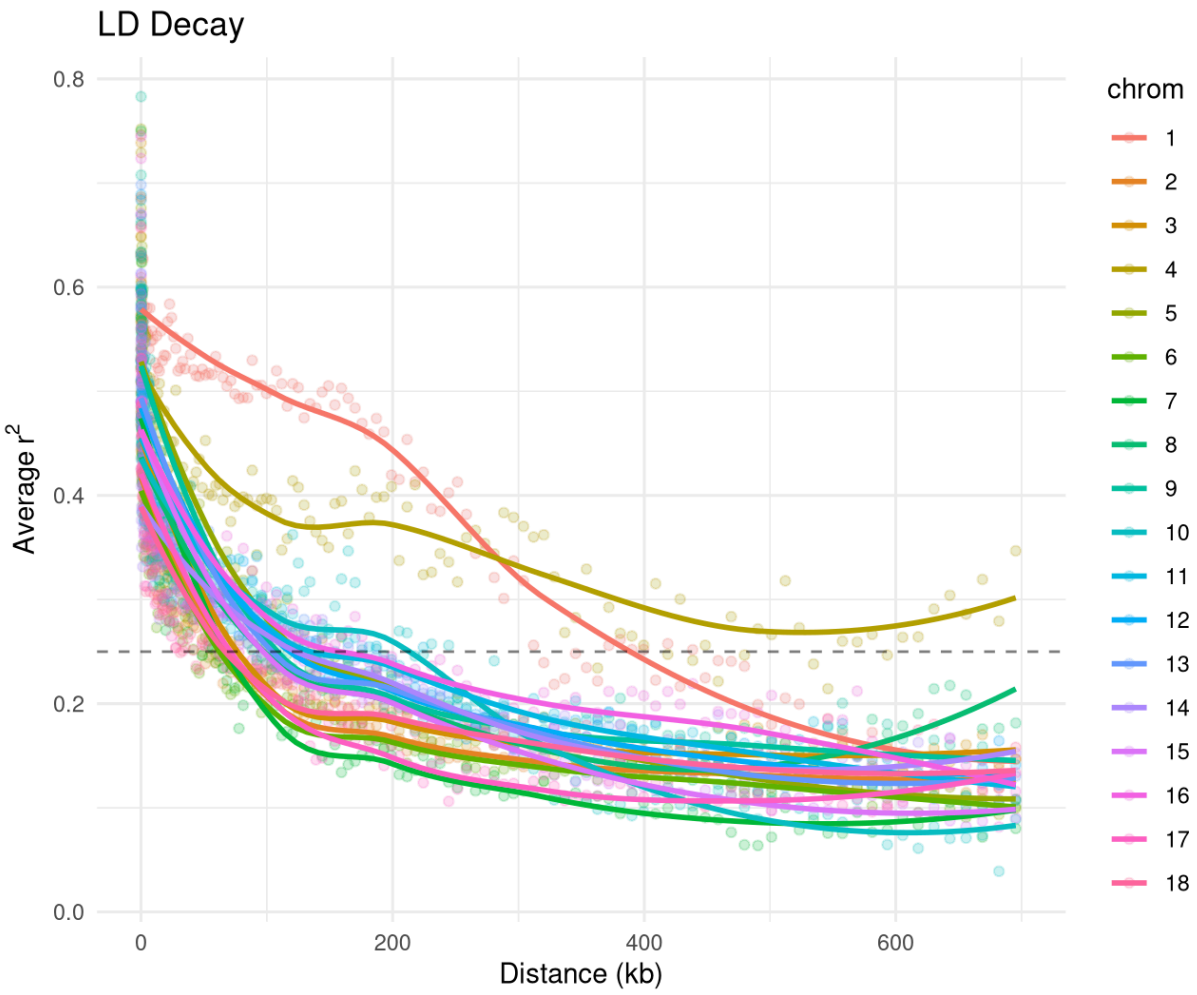
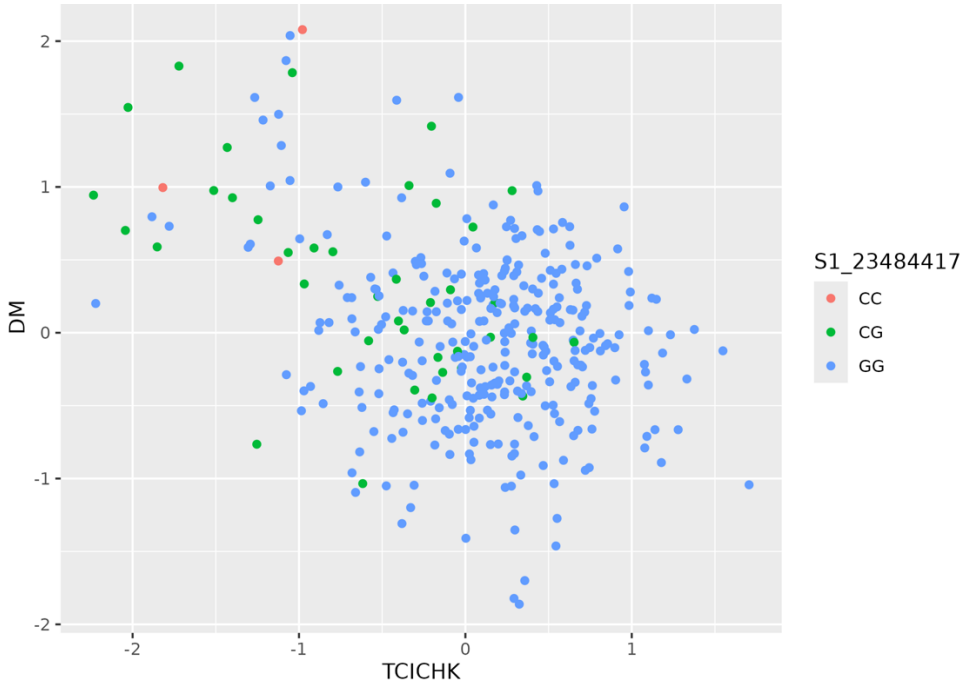


Figure S6. Linkage disequilibrium (LD) decay measured pairwise r^2 between SNPs. Points represent average pairwise r^2 per chromosome within 200 physical distance intervals.

A. S1_23484417



B. S10_2963170

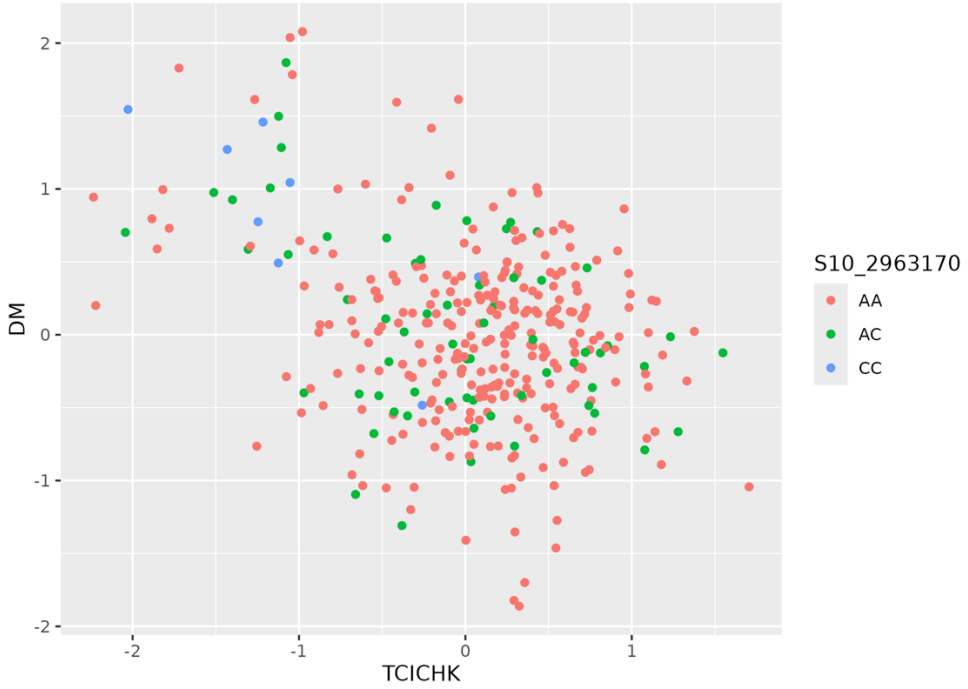


Figure S7. Genotype frequencies and phenotype distributions at the QTL peak SNPs identified on chromosomes 1 (A; S1_23484417) and 10 (B; S10_2963170). The x- and y-axes show the accession

1004 BLUP values for TCICHK and DM, respectively. Points are colored according to the alternate allele
1005 dosage.
1006
1007 **Supplementary Tables**
1008 *Attached file.*
1009 **Table S1.** Metadata for cassava accessions included in this analysis.
1010 **Table S2.** Metadata for field trials included in this analysis.
1011 **Table S3**. *A priori* candidate genes with annotations relating to carotenoid, apocarotenoid, and
1012 starch synthesis and degradation pathways.
1013 **Table S4.** Summary statistics for peak SNPs identified in univariate associations with agronomic
1014 traits at the Bonferroni-corrected significance threshold.
1015 **Table S5.** Candidate genes within 100kb of peak SNPs significantly associated with TCICHK and DM
1016 at the modified Bonferroni significance threshold from the bivariate GWAS.
1017 **Table S6.** Candidate genes within 100kb of peak SNPs identified in univariate associations with
1018 agronomic traits at the Bonferroni-corrected significance threshold.