UNIVERSIDADE FEDERAL DE VIÇOSA

From SNPs to haplotypes: Evaluating Single-SNP and Haplotype-Based approaches for GWAS and genomic prediction in doubled haploid maize

Alice Silva Santana
Doctor Scientiae

VIÇOSA - MINAS GERAIS
2025

ALICE SILVA SANTANA

From SNPs to haplotypes: Evaluating Single-SNP and Haplotype-Based approaches for GWAS and genomic prediction in doubled haploid maize

Thesis submitted to the Genetics and Breeding Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Rodrigo Oliveira de Lima

VIÇOSA - MINAS GERAIS
2025

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
S232f
2025

Santana, Alice Silva, 1997-
From SNPs to haplotypes: evaluating single-SNP and haplotype-based approaches for GWAS and genomic prediction in doubled haploid maize / Alice Silva Santana. – Viçosa, MG, 2025.
1 tese eletrônica (52 f.): il. (algumas color.).

Texto em inglês.
Inclui apêndices.
Orientador: Rodrigo Oliveira de Lima.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Agronomia, 2025.
Referências bibliográficas: f. 26-36.
DOI: https://doi.org/10.47328/ufvbbt.2025.552
Modo de acesso: World Wide Web.

1. Milho - Melhoramento genético. 2. Marcadores genéticos. 3. Alelos múltiplos. I. Lima, Rodrigo Oliveira de, 1982-. II. Universidade Federal de Viçosa. Departamento de Agronomia. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 633.1523

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

ALICE SILVA SANTANA

From SNPs to haplotypes: Evaluating Single-SNP and Haplotype-Based approaches for GWAS and genomic prediction in doubled haploid maize

Thesis submitted to the Genetics and Breeding Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

APPROVED: August 11, 2025.

Assent:

Alice Silva Santana
Author

Rodrigo Oliveira de Lima
Adviser

ACKNOWLEDGMENTS

This work has been sponsored by the following Brazilian research agencies: Coordination for the Improvement of Higher Education Personnel (CAPES; Financing code 001), Minas Gerais State Foundation for Research Aid (FAPEMIG) and National Council of Scientific and Technological Development (CNPq).

ABSTRACT

SANTANA, Alice Silva, D.Sc., Universidade Federal de Viçosa, August, 2025. **From SNPs to haplotypes: Evaluating Single-SNP and Haplotype-Based approaches for GWAS and genomic prediction in doubled haploid maize**. Adviser: Rodrigo Oliveira de Lima.

Single nucleotide polymorphisms (SNPs) are typically considered biallelic for practical purposes, although multiallelic variants can occur. In contrast, haplotypes derived from combinations of SNPs are inherently multiallelic and may capture epistatic interactions among the constituent alleles. In our study, we compared the single-SNP and two haplotype-based approaches for GWAS and genomic prediction models using 487 doubled haploid (DH) maize lines derived from the Iowa Stiff Stalk Synthetic population. The DH lines were evaluated for anthesis-silking interval (ASI), plant (PH, cm) and ear height (EH, cm), flag leaf angle (FLA), number of primary tassel branches (NTB) and tiller rate (TR) across three environments in Iowa, United States. After quality control, we used a total of 13,846 SNP markers to perform genomic analyses. For both single-SNP and haplotype-based approaches, we considered the compressed mixed linear (CMLM) and GBLUP models to run the GWAS and genomic prediction analyses, respectively. Our results showed that genomic regions for FLA were only identified by haplotype approaches, while genomic regions associated with EH were identified by both single-SNP and haplotype-based approaches. Overall, we identified 13 and seven candidate genes within the genomic regions associated with FLA and EH, respectively, whereas no candidate genes were found for NTB and PH. We observed that the use of haplotypes for genomic prediction resulted in slightly higher prediction ability compared to single-SNP based prediction. However, the training set and the trait under consideration played a significant role in the prediction ability. In conclusion, integrating both single-SNP and haplotype-based GWAS approaches enhances the identification of causal variants and allows for selecting the most effective genomic prediction method tailored to each trait’s genetic architecture.

Keywords: Zea mays L.; single nucleotide polymorphism; genome-wide association studies; complex traits; genetic architecture

RESUMO

SANTANA, Alice Silva, D.Sc., Universidade Federal de Viçosa, agosto de 2025. **De SNPs para haplótipos: Avaliando abordagens de marca única e haplótipos para GWAS e predição genômica em milho duplo haploide**. Orientador: Rodrigo Oliveira de Lima.

Polimorfismos de nucleotídeo único (SNPs) são tipicamente considerados bialélicos para fins práticos, embora variantes multialélicas possam ocorrer. Em contraste, haplótipos derivados da combinação de SNPs são inerentemente multialélicos e podem capturar interações epistáticas entre os alelos constituintes. Neste estudo, comparamos abordagens baseadas em SNPs individuais e em haplótipos para análises de GWAS e predição genômica, utilizando 487 linhagens de milho duplo haploides (DH) derivadas da população Iowa Stiff Stalk Synthetic. As linhagens DH foram avaliadas para intervalo entre antese e emissão de estigma (ASI), altura de planta (PH, cm) e espiga (EH, cm), ângulo da folha bandeira (FLA), número de ramos do pendão (NTB) e índice de perfilhamento (TR) em três ambientes no estado de Iowa, Estados Unidos. Após o controle de qualidade, utilizamos um total de 13.846 marcadores SNP nas análises genômicas. Tanto para as abordagens com SNPs individuais quanto com haplótipos, utilizamos os modelos linear misto comprimido (CMLM) e GBLUP para as análises de GWAS e predição genômica, respectivamente. Nossos resultados mostraram que regiões genômicas associadas a FLA foram identificadas apenas pelas abordagens baseadas em haplótipos, enquanto regiões associadas a EH foram detectadas por ambas as abordagens. No total, identificamos 13 e sete genes candidatos nas regiões genômicas associadas a FLA e EH, respectivamente, enquanto nenhum gene candidato foi encontrado para NTB e PH. Observamos que o uso de haplótipos na predição genômica resultou em uma capacidade preditiva ligeiramente superior em comparação à predição baseada em SNPs individuais. No entanto, a população de treinamento e a característica avaliada desempenharam um papel significativo na capacidade de predição. Em conclusão, a integração de abordagens de GWAS baseadas em SNPs individuais e em haplótipos aprimora a identificação de variantes causais e permite a seleção do método de predição genômica mais eficaz, ajustado à arquitetura genética de cada característica.

Palavras-chave: Zea mays L.; polimorfismo de nucleotídeo único; estudos de associação genômica ampla; características complexas; arquitetura genética

LIST OF TABLES

LIST OF FIGURES

SUMMARY

1. Introduction

A haplotype can be defined as a specific combination of multiple adjacent nucleotides or DNA markers from polymorphic sites in the same chromosomal segment (Stephens et al., 2001; International HapMap Consortium, 2005). Three primary methods have been utilized to build haplotypes: (i) fixing a certain number of SNPs per haplotype (Hickey et al., 2013); (ii) establishing a fixed haplotype length (Hess et al., 2017); and (iii) employing LD information (Cuyabano et al., 2014). These methods can be further categorized into non-overlapping segments or distinct windows (Hickey et al., 2013), and overlapping segments or sliding windows (Johansson et al., 2010). Many studies have aimed to determine the optimal method to build haplotypes (Ferdosi et al., 2016; Saad et al., 2018; Bhat et al., 2021; Weber et al., 2023). Generally, haplotypes based on LD information yields better results compared with a fixed length approach (Cuyabano et al., 2014; Maldonado et al., 2019; Bhat et al., 2021), because LD measures the non-random association of alleles at different loci and reflects the historical recombination and evolutionary processes of the target germplasm (Mackay and Powell, 2007). By incorporating LD information, haplotypes may capture more accurately the underlying genetic diversity and structure of the population, addressing the natural patterns of genetic variation. In this context, Gabriel et al. (2002) proposed a method for determining haplotypes based on the confidence interval of D'. D' is a normalized LD measure that accounts for allele frequencies and represents the deviation of actual allelic frequency from linkage equilibrium (Lewontin, 1964). This method has been used in genetic studies to identify haplotypes (Andrade et al., 2019), map phenotypic trait-associated variants in GWAS (Sehgal et al., 2020), asses the prediction ability of different genomic prediction models (Weber et al., 2023), and infer population genetic structure (Shipilina et al., 2023).

The employment of haplotypes in genomic analyses can offer several advantages over single-SNP. Haplotypes provide more informative data for describing recent identity-by-descent (IBD) relationships and may better capture LD with multiallelic quantitative trait loci (QTL) compared with individual SNPs, which are typically biallelic (Broman and Weber, 1999; Browning and Browning, 2013). They also have the potential to capture epistatic interactions within the haplotype alleles (Bhat et al., 2021). Furthermore, haplotypes may be more effective than single-SNP in tracking new mutations especially in the context of SNP arrays, where markers typically represent variants with moderate to high minor allele frequencies due to ascertainment bias (Meuwissen et al., 2014; Bhat et al., 2021). These higher-frequency SNPs generally reflect older

mutations, whereas new mutations initially appear at very low frequencies and may not be directly captured by such arrays.

The fundamental assumption underlying GWAS models based on individual molecular markers is that, given a sufficiently high marker density, at least one marker will be in LD with each QTL (Luan et al., 2012). However, recent studies have favored using haplotypes rather than single molecular markers in GWAS as haplotypes can enhance the detection of associations by capturing the joint effects of adjacent SNPs. This approach not only reduces the dimensionality of the dataset but also allows the identification of novel polymorphisms that are potentially in higher LD with causal variants, thereby improving mapping resolution. Abed and Belzile (2019) compared the performance between single-SNP and haplotype-based GWAS for seven traits in spring barley. They found that the haplotype-based approach unveiled a larger number of associations, some of them were shared with the single-SNP approach. Similarly, Contreras-Soto et al. (2017) concluded that the haplotype-based GWAS provided new insights into the causal variants that were not captured by the single-SNP approach when analyzing complex traits in soybeans. When identifying genes associated with improved yield under drought conditions in Turkish winter wheat germplasm, Sehgal et al. (2024) reported that the haplotype-based GWAS outperformed single-SNP GWAS by identifying a higher number of genomic regions. They observed an average increase of 8% in the percentage of phenotypic variation explained by the significant genomic regions when using the haplotype-based GWAS approach. In maize, Maldonado et al. (2019) stated that approximately 62% of the associated haplotypes did not contain SNPs detected in the association study using single-SNP. According to them, haplotype-based GWAS is more suited for examining genomic regions that control plant lodging and architecture traits than single-SNP approach. In a simulated study, Lorenz et al. (2010) observed that the use of haplotypes provided additional power to detect QTL, especially when QTL architecture was complex.

Haplotypes have also been used in genomic prediction models. The potential stronger LD between haplotypes and QTL in comparison to individual SNPs may yield more accurate genomic estimated breeding values than single-SNP approaches (GEBVs; Calus et al., 2008; Cuyabano et al., 2014; 2015; Matias et al., 2017). Simulated data has reported that prediction models with effects of haplotypes perform better than single-SNP effects (Villumsen et al., 2009). Moreover, empirical studies have also reported a better prediction performance by using haplotypes compared to single-

SNP in sheep (Vahedi et al., 2023), goats (Teissier et al., 2020), pigs (Bian et al., 2021), wheat (He et al., 2019; Sallam et al., 2020) and rice (He et al., 2023). However, the results are contradictory, and while some studies have demonstrated improvement in the prediction ability, others have showed no significant difference or even inferior performance of haplotypes compared to SNP-based approaches (Mucha et al., 2019; Feitosa et al., 2020; Lin et al., 2024). Araujo et al. (2021) hypothesized that haplotype-based approaches could result in more accurate and less biased GEBV prediction when compared to single-SNP models in crossbreeding populations because of their high genetic diversity and more complex haplotype structure than observed in populations with low genetic diversity. However, they concluded that haplotype-based models did not improve the performance of genomic predictions. Thus, while haplotype-based genomic prediction holds great promise in advancing our understanding of genotype-phenotype relationships, ongoing research efforts are imperative to elucidate its true potential and optimize its application across diverse genetic contexts and traits.

In light of these considerations, our objectives were to i) assess the comparative performance of SNP-based GWAS versus two haplotype-based GWAS approaches for identifying QTLs for plant architecture traits in a set of 487 maize DH lines; ii) investigate whether haplotype-based genomic prediction could outperform single-SNP based approach; and iii) assess the potential of genomic prediction on both within the Iowa Stiff Stalk Synthetic (BSSS) subpopulations and across the entire panel.

2. Material and methods

2.1 Plant materials

We used 487 maize DH lines derived from different cycles of recurrent selection of the BSSS, a synthetic population developed in 1933-1934 by Sprague (1946) at USDA Agricultural Research Service. BSSS was formed by intercrossing 16 inbred lines with resistance to stalk lodging, primarily of Reid Yellow Dent heritage, and underwent several cycles of reciprocal recurrent selection (Eberhart et al. 1973; Lamkey et al. 1991; Edwards 2011). The BSSS population has yielded several key founder lines, including B14, B37, and B73 (Troyer 1999; Bornowski et al. 2021). Currently, B73 serves as a benchmark for studying genomic variation within maize populations. The availability of the B73 reference genome has a huge contribution to trait mapping studies and gene discovery efforts in maize (Schnable et al. 2009; Jiao et al. 2017). The DH lines used in our study were developed from the unselected base population, BSSS, the 17th cycle of

reciprocal recurrent selection, BSSS(R)C17, and the cross between them, BSSS/BSSS(R)C17. Random samples (around 1,400 plants) from BSSS, BSSS(R)C17, and BSSS/BSSS(R)C17 were collected to develop DH lines as described by Ledesma et al. (2023b). A total of 135, 194, and 187 DH lines were obtained from the BSSS, BSSS(R)17, and BSSS/BSSS(R)C17 subpopulations, respectively. The primary aim was to investigate traits that have been modified through the recurrent selection cycles and are likely associated with germplasm adaptation to high plant density such as flowering time, plant and ear height, flag leaf angle, tiller rate, tassel size, and number of primary tassel branches (Brekke et al. 2011; Edwards 2011). The panel of DH lines has already been molecularly characterized by Ledesma et al. (2023a), where they reported a lower genetic diversity and a slower LD decay among DH lines from the BSSS(R)C17 compared to BSSS and BSSS/BSSS(R)C17. It has also been phenotypically characterized for plant architecture traits and holds promise as germplasm for association studies (Ledesma et al. 2023b).

2.2. Trial management and experimental design

The 487 maize DH lines were evaluated across three locations in Iowa, USA, during the 2019 summer season: Plant Introduction Station in Ames; Johnson Farm near Kelly; and Burkey at Agronomy Farm near Boone. In each location, the experiments were conducted in a modified split plot design (Bisgaard 2000) with two replications. The whole plot factor was the three subpopulations of DH lines [BSSS, BSSS(R)C17, and BSSS/BSSS(R)C17], while the subplot factor was DH lines. Subplots were 3.80 m long, with single rows spaced 0.76 m apart.

2.3. Trait measurements

We conducted a *per se* evaluation of DH lines for six traits: anthesis-silking interval (ASI, days), plant (PH, cm) and ear height (EH, cm), flag leaf angle (FLA, degrees from vertical), number of primary tassel branches (NTB, unit), and tiller rate (TR, %). The ASI was obtained in days as the difference between female and male flowering (ASI=days to 50% silking-days to 50% anthesis). PH and EH were recorded two weeks after pollination: PH was the height from the soil surface to the flag leaf collar and EH was the height from the soil surface to the stalk node at which the uppermost ear has emerged. FLA was measured using a protractor as follows: the protractor was positioned against the section of the stalk beneath the flag leaf, with its base positioned underneath the midrib of the flag leaf allowing the accurate recording of the flag leaf angle at its insertion point to the stalk. In this context, a smaller angle indicates that the flag leaf is more upright. The NTB was measured as the number of primary tassel branches directly stemming from the main tassel

branch. The TR was obtained as the ratio of plants displaying tillers to the total number of plants per plot. For each plot, measurements of PH, EH, FLA and NTB were taken from three individual plants, and the average was used as the final trait value for subsequent analyses. For ASI and TRA the whole plot was considered for the evaluation. The data were assessed for normality using the Shapiro-Wilk test at a 1% significance level (Shapiro and Wilk (1965). All traits were found to follow a normal distribution ($P > 0.01$). Residual diagnostic plot panel of each trait is provided in the Supplemental Files.

2.4. Phenotypic data analyses

To estimate the variance components, phenotypic values of each trait tested across the three locations were combined and modeled according to the following expression:

$$y_{ijklmn} = \mu + p_i + g_j + e_k + ge_{jk} + r_{l(k)} + p_{m(kl)} + a_{n(kl)} + \varepsilon_{ijklmn}$$

where y_{ijklmn} is the phenotypic value of i^{th} subpopulation and j^{th} DH line at the k^{th} environment in the l^{th} replication, m^{th} pass and n^{th} range; μ is the general mean; p_i is the random effect of the i^{th} subpopulation [BSSS, BSSS(R)17 and BSSS/BSSS(R)C17], with $p \sim N(0, \sigma_p^2)$, where σ_p^2 is the variance component of subpopulation; g_j is the random effect of the j^{th} DH line with $g \sim N(0, \sigma_g^2)$, where σ_g^2 is the variance component of DH line; e_k is the random effect of the k^{th} environment with $e \sim N(0, \sigma_e^2)$, where σ_e^2 is the variance component of environment; ge_{jk} is the random effect of DH line-by-environment interaction with $ge \sim N(0, \sigma_{ge}^2)$, where σ_{ge}^2 is the variance component of line-by-environment; $r_{l(k)}$ is the fixed effect of the l^{th} replication within the k^{th} environment; $p_{m(kl)}$ is the random effect of the m^{th} pass within the k^{th} environment and l^{th} replication with $p \sim N(0, \sigma_p^2)$, where σ_p^2 is the variance component of pass; $a_{n(kl)}$ is the random effect of the n^{th} range within the k^{th} environment and l^{th} replication with $a \sim N(0, \sigma_a^2)$, where σ_a^2 is the variance component of range; and ε_{ijklmn} is the random effect of residual error with $\varepsilon \sim N(0, \sigma^2)$, where σ^2 is the residual variance. All random effects were assumed to be independent and identically distributed. The significance of the random effects was tested by the likelihood ratio test (LRT; Wilks, 1938). For the LRT test, chi-square statistics with 1 degree of freedom and 5% of significance were considered. In the initial stage of the analyses, the DH line effect was treated as random to estimate the variance component. Subsequently, this effect was treated as fixed to calculate the best unbiased linear estimator (BLUE). The BLUEs of DH line effect were used to perform the GWAS and genomic prediction analyses. The heritability on the DH line-mean basis of each trait across environments was

calculated as follows: $H_{\bar{X}}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{ge}^2}{n} + \frac{\sigma^2}{nr}}$, where n and r are the number of environments and replications, respectively. The phenotypic data analyses were conducted using the R package "asreml" (Gilmour et al. 2009).

2.5. Genotypic data

Leaf samples were collected from seedlings of the DH lines, all of which were cultivated in the greenhouse at the Department of Agronomy from Iowa State University. Leaf tissue was harvested individually at the V3–V4 developmental stage. DNA was extracted following the standard CIMMYT protocol (CIMMYT, 2005). Genotyping was performed using the Diversity Arrays Technology platform (DArTseq; Kilian et al. 2012) by the Genetic Analysis Service for Agriculture at CIMMYT. A total of 51,418 SNP markers were initially generated. Of these, 32,929 SNPs were successfully aligned to the B73 RefGen_v4 reference genome (Jiao et al. 2017) and retained for quality control.

Quality control began with the removal of duplicated SNPs (i.e., markers with identical genomic positions) and monomorphic markers (i.e., SNPs with no allelic variation across the 487 DH lines). Next, SNPs with a minor allele frequency (MAF) ≤1% and a call rate below 90% (i.e., >10% missing data) were excluded. These filtering thresholds were applied to reduce noise and ensure that informative, high-confidence markers were used in subsequent analyses, particularly for haplotype assembly.

Missing genotypic data were imputed using the linkage disequilibrium k-nearest neighbor imputation (LDkNNi) algorithm (Money et al. 2016) as implemented in TASSEL v5.2.70 (Bradbury et al. 2007). This method leverages local LD patterns and genetic similarity among samples to predict missing values. LDkNNi is well suited for biparental populations, such as DH lines, where strong LD is expected across large genomic regions. After imputation and filtering, 13,846 high-quality SNP markers remained and were used for haplotype block construction and downstream analyses. To provide an overview of genotyping quality, we included supplementary figures showing the distribution of MAF and call rate by subpopulation (Supplementary Figures S7-S8).

2.6. Haplotype assembly

We employed the SHAPEIT software (Delaneau et al. 2012) to phase the genotypes. The SHAPEIT software estimates individual haplotypes from genotypic data using Hidden Markov

Models. The phasing step was performed within each maize chromosome (Chr) using all available genotypes. In our study, haplotypes were constructed using two different approaches. The first one was referred to as the LD_{hap} approach, where the haplotypes were built based on a fixed LD threshold between each pair of SNPs (Cuyabano et al. 2014; Ferdosi et al. 2016; Teissier et al. 2020). Therefore, a Chr with length equal to *k* was divided into a fixed length equal to *j*. The *j* value was defined as the LD decay distance (kb) within each Chr reported by Ledesma et al. (2023b) when using the entire panel of DH lines. The "fixed" length in our manuscript refers to a window size derived from empirical LD decay, not a fixed physical size or marker count. The second method, referred as LD_{r^2}, followed the confidence interval algorithm proposed by Gabriel et al. (2002) and implemented in the PLINK v1.90 software (Purcell et al. 2007). The 95% confidence intervals between SNPs were calculated and each comparison was categorized as strong LD, inconclusive, or strong recombination. If 95% of the comparisons in one segment were in strong LD, a haplotype was created. To classify two or more SNPs as being in strong LD, the minimum lower and upper confidence interval values were set to 0.70 and 0.95, respectively.

The output files from phasing were used for compatibility with the R package "GHap" (Utsunomiya et al. 2016) that requires phased data. For both haplotype methods, the haplotype alleles were called as described by Teissier et al. (2020). Notably, many haplotypes can be multi-allelic, and several pseudo-SNPs can be created from the multi-allelic haplotypes. At the end, the number of copies of a specific pseudo-SNP allele were counted and coded as 0, 1, or 2 for each genotype. In accordance with the expected homozygosity of DH lines, all heterozygous genotype calls (coded as 1) were considered missing data. The pseudo-SNPs were subjected to the same quality control criteria as the SNPs before their use for GWAS and genomic prediction. Descriptive statistics were used to present the number of haplotypes per Chr and the number of SNPs per haplotype for each haplotype approach. The haplotypes length mean (kb) within each Chr was also computed for the LD_{r^2} approach and the results were plotted in a violin plot using the R package "ggplot2" (Wickham 2016).

2.7. Genome-wide association studies

The BLUEs of ASI, PH, EH, FLA, NTB, and TR were used for performing GWAS in the 487 maize DH lines tested in our study. GWAS was performed for both single-SNP and haplotype-based approaches using R package "ASRgwas" (Galli et al. 2022). For both approaches, the compressed mixed linear model (CMLM) was employed (Zhang et al. 2010). This model groups

individuals and incorporates the genetic values of the groups as random effects. It also presents a complementary approach called 'population parameters previously determined' (P3D), which eliminates the need to re-compute variance components. The implementation of this method markedly reduces computing time and either maintains or improves statistical power compared to traditional MLM methods (Zhang et al. 2010).

The GWAS model used for both SNP and haplotype-based was as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}$$

where y is the vector of BLUEs for each trait; X is the design matrix for fixed effects, including the intercept, the tested SNP or haplotype, and the first three principal components (PCs) derived from PCA to control for population structure (**Q**), which allocated the entire panel into three subpopulations. The random genetic effects u ~N(0, $\mathbf{K}\sigma_g^2$), where **K** is the VanRaden kinship matrix, and ε ~N(0, $\mathbf{I}\sigma_e^2$), is the residual error. We considered two cases for the genomic kinship matrix: $\mathbf{G} = \begin{cases} \frac{\mathbf{ZZ'}}{2\sum_{j=1}^{i} p_j(1-p_j)}, \text{single} - \text{SNP based model} \\ \frac{\mathbf{HH'}}{2\sum_{j=1}^{i} p_j(1-p_j)}, \text{haplotype based model} \end{cases}$

where **Z** is the matrix of SNP markers, p_j is the observed allele frequency at locus *j*, and *i* is the total number of SNP markers. On the other hand, **H** is the haplotype-based kinship matrix that follows the same principles of **Z**, but, instead of SNP markers, the haplotype alleles were considered. Because haplotypes were coded as pseudo-SNPs, the implementation of VanRaden kinship matrix for the haplotype-based GWAS is straightforward (Teissier et al. 2020).

Ledesma et al. (2023) examined the optimal utilization of DH lines for GWAS: either within each one of the three subpopulations used in our study [BSSS, BSSS(R)C17 and BSSS/BSSS(R)C17] or the three sets combined in a unique panel. They found a larger number of SNPs significantly associated with traits when using the entire panel. This was attributed not only to the population structure, but primarily to the larger sample size inherent in using the entire panel for GWAS. Thus, in our study, we performed GWAS for single-SNP and haplotype-based approaches using the entire panel [BSSS, BSSS(R)17 and BSSS/BSSS(R)C17] instead of running it for each subpopulation separately. For multiple testing correction, we used the simple*M* threshold method (Gao et al. 2008, 2010, 2011). By applying it, the following p-values were obtained for each approach: the SNP-based approach produced a p-value of 5.18×10^{-6} with a threshold of 5.28;

the LD_{hap} approach resulted in a p-value of 4.45×10^{-6} with a threshold equal to 5.35; the LD_{r^2} method yielded a p-value of 1.68×10^{-5} with a threshold equal to 4.77.

The available maize genome sequence (B73; RefGen_v5) was used as the reference genome for candidate gene identification. A gene was considered as a candidate if it was located within the range of LD decay observed for each Chr (upstream and downstream) in the single-SNP approach or if it was within the start and end point of the haplotype in the haplotype-based approaches. Candidate genes were identified using the MaizeGDB molecular marker database (http://www.maizegdb.org; Portwood et al. 2019). Functional annotations of candidate genes were predicted in NCBI (http://www.ncbi.nlm.nih.gov/gene) and were also compared to previously published candidate genes.

2.8. Genomic prediction analyses

Genomic prediction was carried for four prediction scenarios: *i*) using BSSS to predict both BSSS(R)17 and BSSS/BSSS(R)C17; *ii*) using BSSS(R)17 to predict both BSSS and BSSS/BSSS(R)C17; *iii*) using BSSS/BSSS(R)C17 to predict BSSS and BSSS(R)17; and *iv*) making predictions across subpopulations through cross-validation. Scenarios *i)* to *iii)* are illustrated in Figure 1. We ran each one of the four prediction scenarios for each one of the three genomic kinship matrices (single-SNP, LD_{hap} and LD_{r^2}). Therefore, in total, we had twelve scenarios for genomic prediction.

We compared the predictive ability of the proposed approaches using the GBLUP method as follows (VanRaden 2008): $\mathbf{y} = \mathbf{Xb} + \mathbf{Wm} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the vector of BLUEs of the genetic effects of DH lines, $\mathbf{b}$ is the intercept, $\mathbf{m}$ is the vector of random effects of genotypes, with $\mathbf{m} \sim N(0, \mathbf{G}\boldsymbol{\sigma}_g^2)$, where $\mathbf{G}$ is the genomic kinship matrix based on single-SNP or haplotypes and $\boldsymbol{\sigma}_g^2$ is the genotypic variance component. The genomic kinship matrix for single-SNP and haplotype-based approaches were estimated following VanRaden (2008) method as explained before. $\mathbf{X}$ and $\mathbf{W}$ are the incidence matrices for $\mathbf{b}$ and $\mathbf{m}$, respectively, while $\boldsymbol{\varepsilon}$ is the vector of random errors, with $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\boldsymbol{\sigma}^2)$, where $\mathbf{I}$ is an identity matrix.

Prediction ability was estimated as the correlation between the BLUE values and the predicted GEBV values in the prediction scenarios from *i* to *iii*. In the complete dataset (scenario number *iv*), we used a fivefold cross-validation to assess the accuracy of GEBVs, where the genotypes were randomly divided in five folds with similar size. In each round, we used four folds

as training set (around 413 lines) to predict one-fold selected as missing phenotype values to be the validation set (around 103 lines). Fivefold cross-validation was repeated 10 times.

3. Results

3.1. Phenotypic results

We observed wide ranges among the BLUEs values of the DH lines, mainly for FLA and TR traits (Table 1). The analysis across the three locations showed that genotypes and subpopulations were highly significant (P<0.001) for all traits (Table 1). The coefficients of variation (CV) were lower than 30% for all traits. NTB had the lowest CV value among all traits (8.28%), while TR trait showed the highest CV value (28.63%). The estimates of heritability ($H^2_{\bar{X}}$) were high, and tested traits exhibited $H^2_{\bar{X}}$ values equal or greater than 0.80.

3.2 Haplotype profiling

The number of SNPs per Chr after quality control ranged from 917 (Chr 10) to 2,174 SNPs (Chr 1; Table 2). The LD_{hap} approach yielded a greater amount of haplotypes within each maize Chr compared to LD_{r^2}. On average, the LD_{hap} exhibited a 64.39% increase in the number of haplotypes within each Chr in comparison to LD_{r^2} approach. In relation to the number of SNPs per haplotype, the LD_{hap} presented an average of 10 SNPs per haplotype across Chrs, whereas LD_{r^2} presented an average of four SNPs per haplotype. Consequently, LD_{hap} not only exhibited a higher quantity of haplotypes but also a greater number of SNPs per haplotype. In both approaches, the minimum number of SNPs per haplotype was two and it remained consistent across all Chrs. However, in LD_{hap}, the maximum number of SNPs per haplotype among Chrs ranged from 20 (Chrs 2, 5, 6 and 10) to 30 (Chr 1), whereas, in LD_{r^2} approach, the maximum number varied from 9 (Chrs 3 and 4) to 13 SNPs (Chr 6). The haplotype length in LD_{hap} for each Chr (Table 2) was fixed for all haplotypes within each Chr and defined by the LD decay profile as outlined by Ledesma et al. (2023b), whereas, in LD_{r^2} approach, the haplotype length varied based on the interval confidence results (Figure 2). With this approach, Chr 6 demonstrated the smallest mean haplotype length (160.95 kb), whereas Chr 8 exhibited the largest one (304.64 kb).

3.3. GWAS based on single-SNP and haplotype-based approaches

The quantile-quantile (Q-Q) plots indicated that population structure and familial relatedness were well controlled in both single-SNP and haplotype-based approaches (Figure 3-5). Based on single-SNP approach, GWAS analysis identified three SNPs significantly (5.18×10^{-6})

associated with EH, NTB and PH traits (Table 3). Regarding haplotype-based approaches, we found two and one haplotype associated with FLA trait using LD_{hap} (4.45×10^{-6}) and LD_{r^2} (1.68×10^{-5}), respectively, and one haplotype associated with EH trait by LD_{hap} approach. In general, the phenotypic variance explained (PVE) by the association ranged from 1.72% (EH) to 3.62% (FLA; Figure 6). Notably, while only the SNP-based approach identified a significant SNP for NTB and PH, significant genomic region for FLA was only identified using the LD_{hap} and LD_{r^2} approaches. Interestingly, we observed genomic regions associated with EH by the two approaches, single-SNP and LD_{hap}. Moreover, although both approaches identified genomic regions associated with FLA on Chr 7, we observed that distinct loci were revealed for each approach. In addition, we observed that selection for all significant QTLs led to a decrease in the mean phenotypic values of FLA and PH, and an increase in the mean for NTB. For EH, divergent effects were observed depending on the approach. QTLs identified by LD_{hap} were associated with increased phenotypic means, whereas those identified by the single-SNP approach were associated with decreased values for EH. CITAR FIGURE 6 AQUI

Concerning the candidate genes, we identified 13 and seven candidate genes associated with FLA and EH, respectively, whereas no candidate genes were found for NTB and PH (Table 4). All candidate genes observed for FLA and EH traits were located on Chr 7 and 9. Additionally, no candidate genes for FLA and EH were identified within the genomic regions identified by the LD_{r^2} and single-SNP approaches. Considering the extensive genomic coverage of the LD_{hap} window, more than one candidate gene was identified within the significant haplotypes detected by the GWAS analysis. For FLA, the gene Zm00001eb317180, located on Chr 7 within the haplotype chr7_B316.39966, appears to encode a *30S ribosomal protein S16*, which is associated with chlorophyll biosynthesis in maize (Li et al. 2023a). Two additional genes within the same haplotype, Zm00001eb317190 and Zm00001eb317200, encode a *putative sarcosine oxidase* and a *calcium-transporting ATPase* (*Ca^{2+}-ATPase*), respectively. For the haplotype chr7_B322.40059, located on Chr 7, the gene Zm00001eb317760, which encodes *Squamosa Promoter-Binding-Like protein 6* (*spl6*), was identified. Nine additional candidate genes were also identified within the same haplotype. Among them, Zm00001eb317730, Zm00001eb317740, Zm00001eb317770, and Zm00001eb317820 encode *defective kernel41* (*dek41*), a *TITAN-like protein*, *developmental protein SEPALLATA 2* (*zmm7*), and *small auxin up RNA63* (*saur63*), respectively. For EH, two haplotypes, chr7_B91.38700 and chr9_B123.48289, were associated with genes related to a *protein*

kinase, *subtilisin-like protease*, plant heme peroxidase, *TLD-domain nucleolar protein*, and *serine carboxypeptidase-like protein*, and to *histone H2AXa* and *fdh2*, respectively.

3.4 Genomic prediction

In the genomic prediction analyses, one of our scenarios (scenario number *iv*) leveraged the entire population to randomly predict a set of DH lines through cross-validation. The genomic prediction using LD_{hap} slightly increased the prediction ability compared to the single-SNP approach for almost all traits, except for PH and TR (Figure 7). For these traits, the single-SNP approach presented a higher prediction ability than both LD_{hap} and LD_{r^2}. The use of LD_{r^2} as a basis for prediction resulted in lower predictive performance relative to the other approaches for all traits. However, it is important to highlight that differences in prediction ability were very small (< 2%), mainly between single-SNP and LD_{hap} approaches.

Overall, the prediction abilities were highly dependent on both the subpopulation used as the training set and the haplotype-based approach employed (Figure 8). Also, we observed that generally the prediction abilities for ASI, NTB and TR using LD_{hap} were higher than those based on single-SNP and LD_{r^2} approaches . The single-SNP approach slightly overpassed the LD_{hap} for the EH, FLA and PH when using BSSS and the interpopulation [BSSS/BSSS(R)C17] as a training set. Particularly for ASI and EH, the prediction ability of BSSS/BSSS(R)C17 was the lowest across the three approaches, whereas for FLA, NTB and PH, BSSS and BSSS/BSSS(R)C17 had the highest prediction abilities for almost all approaches. Regarding FLA, the prediction ability of BSSS/BSSS(R)C17 was much higher than the other prediction scenarios, and the values were equal to 0.67, 0.65, and 0.60 for predictions based on the single-SNP, LD_{hap} and LD_{r^2} approaches, respectively.

4. Discussion

Comparing different approaches to perform GWAS and genomic prediction analyses is very important to help plant breeders to set up breeding strategies most suitable for maintaining and increasing genetic gain in their breeding programs. This comparative approach enables researchers to harness the full potential of genomic technologies for advancing crop improvement efforts (Ibrahim et al. 2020). In our study, we compared the performance of haplotype-based approaches over single-SNP approach using plant architecture traits measured in a large panel of DH lines derived from three subpopulations of BSSS population: BSSS, BSSS(R)C17 and BSSS/BSSS(R)C17.

According to our results, genotype and subpopulation play a significant influence on the tested traits, while moderate to high heritability values were found across traits. Our results complement the findings from Ledesma et al. (2023b) which highlighted significant differences in plant architecture traits across the three subpopulations of DH lines. According to them, the BSSS(R)C17 could be a source of favorable alleles that impact more erect FLA and a smaller NTB, confirming that the 17 cycles of recurrent selection have been effective in improving plant architecture traits. The observed differences in these traits have practical implications for maize breeding strategies. Historically, breeding for higher population density in maize has led to the selection of plants with more erect FLA (Elli et al. 2023). For instance, erect FLA is associated with improved light interception and photosynthetic efficiency, potentially leading to increased yield (Zhang et al. 2021; Tang et al. 2021; Cao et al. 2022a). Similarly, a smaller number of primary tassel branches may contribute to lower photosynthate competition and less shade, thereby improving overall plant performance (Schuetz and Mock 1978; Brewbaker 2015; Li et al. 2023b).

We observed that LD_{hap} approach has the potential to develop a higher number of haplotypes per Chr and a greater number of SNPs per haplotype compared to LD_{r^2}. Both approaches leveraged the LD information. Although the LD_{hap} approach was based on a fixed length within each Chr, the LD information reported by Ledesma et al. (2023b) was implemented to decide its haplotypes length. On the other hand, LD_{r^2} considered the LD confidence interval reported by Gabriel et al. (2002) to build the haplotypes. This helps to explain why LD_{r^2} presented a lower number of haplotypes within each Chr. Each haplotype in LD_{r^2} approach had a different length and only the SNPs that met the criteria of the confidence interval test were allocated into a haplotype, whereas in LD_{hap} approach, all SNPs were included in a haplotype following the fixed length previously stablished. Consequently, a higher number of SNPs per haplotype was observed in the LD_{hap}. In a simulation study, Villumsen et al. (2009) observed that smaller sizes of haplotypes (from 5 to 10 SNPs per haplotype) provided the highest prediction abilities for traits with low heritability values. According to them, the performance of large haplotypes may become poor because the number of haplotype alleles increases quickly with increasing haplotype size, which results in less data available for accurately estimating each allele's effect. On the other hand, Andrade et al. (2019) found that haplotypes with a reduced number of SNPs might not bring a significant advantage for haplotype-based GWAS. Therefore, there is no agreement about the best

size of haplotypes. According to Villumsem et al. (2009), the optimal length of haplotypes will depend on factors such as the LD decay and the population structure and must, therefore, be optimized for each data set. Previous analyses of LD decay in our subpopulations revealed substantial differences in LD extent (Ledesma et al. 2023b). These variations can affect haplotype block structure, as faster LD decay tends to produce smaller and less stable blocks, while slower decay results in longer and potentially more informative haplotype blocks (Gabriel et al. 2002). In popcorn, Andrade et al. (2019) reported that haplotype-based GWAS and genomic selection may offer limited advantages across generations, mainly due to the low SNP density within haplotype blocks. Therefore, conclusions regarding the utility of haplotype-based GWAS should be interpreted in light of the specific LD patterns and genetic architecture of each population and trait under consideration.

We observed trait-specific patterns in the identification of significant genomic regions by GWAS analyses, with different approaches prioritizing distinct sets of genomic regions. Abed and Belzile (2019) compared single-SNP and haplotype-based GWAS in barley and they also observed that each approach uncovered a different subset of QTLs. By applying LD_{r^2} approach, we found genomic regions associated with FLA that were not detected by the single-SNP. Similarly, when examining genomic regions that control plant architecture and lodging in maize plants, Maldonado et al. (2019) concluded that haplotype-based GWAS allowed the identification of genomic regions that were not detected by GWAS based on single-SNP. According to Wray et al. (2013), the major limitations for the use of SNPs include their biallelic nature, the presence of rare alleles, and abundant levels of linkage drag. Therefore, the candidate genomic loci identified by GWAS using single-SNP approach might not always represent the causative locus of the target trait. Instead, they tend to coincide with loci that are in LD with a gene or a regulatory element eventually affecting the trait of interest (Bhat et al. 2021). Khankhanian et al. (2015) concluded that the use of haplotypes can lead to the discovery of new regions of interest, which have not been identified by a single-SNP approach. According to Wu et al. (2021), haplotype-based GWAS is a complementary method that may be beneficial for detecting further genetic variants. Notably, disparities in our GWAS results were evident across different approaches, emphasizing the importance of employing multiple methodologies to capture trait-associated genetic variation comprehensively.

The investigation of candidate genes represents a fundamental objective in plant breeding programs. In our study, no overlap was observed between the candidate genes identified by the LD_{hap}, LD_{r^2} and single-SNP approaches. Candidate genes were found exclusively within genomic regions detected by only one of the methods. This lack of concordance may reflect the distinct sensitivity of each approach in capturing association signals across the genome, suggesting that complementary methods can reveal specific regions linked to the trait of interest. In maize, previous studies have demonstrated that single-SNP and haplotype-based GWAS approaches tend to reveal different yet complementary candidate regions, thereby enhancing QTL detection power and contributing to a more precise localization of causal polymorphisms (Lu et al. 2012; Negro et al. 2019). Moreover, our findings showed that candidate genes could only be identified through the haplotypes defined by the LD_{hap} approach. We compared haplotype-based approaches using LD to group SNPs associated with the identification of candidate genes related to the traits studied. According to Negro et al. (2019), the differences between the approaches can be attributed to variations in local recombination rates and LD patterns. The LD_{r^2} approach considers adjacent SNPs based on LD and is highly sensitive to structural variations (Springer et al. 2019) along with the presence of allelic series with contrasting effects in different populations (Giraud et al. 2017). In contrast, the LD_{hap} approach may overestimate the number of QTLs in high-recombination regions, where SNPs are too far apart genetically to be clustered, or underestimate them by grouping SNPs in regions with low recombination rates (Springer et al. 2019). In our results, these differences had a significant impact on the identification of candidate genes. Furthermore, when comparing the candidate genes reported by Ledesma et al. (2023b), who used the FarmCPU model for GWAS, we identified additional candidate genes through haplotype-based approaches, located in chromosomal regions that were not detected in their study for the FLA and NTB traits. These findings highlight the importance of incorporating haplotype-based approaches in GWAS, as they can uncover additional genomic regions and candidate genes that may be overlooked by single-SNP methods.

For FLA, the *30S ribosomal protein S16* candidate gene (Zm00001eb317180), is associated with chlorophyll biosynthesis in maize (Li et al. 2023a). Its role in plastid translation suggests a potential influence on photosynthetic efficiency and leaf architecture. In both tobacco and *Arabidopsis*, the absence of the S16 protein has been shown to reduce photosynthetic capacity, highlighting its functional importance in chloroplast activity (Fleischmann et al. 2011; Ruf et al.

2025). The gene Zm00001eb317200 encodes a *calcium-transporting ATPase* (*Ca^{2+}-ATPase*) and has been associated with salt tolerance in maize (Hou et al. 2024). Although this gene is not directly involved in FLA regulation like *ZmCLA4* (Zhang et al. 2014) and *ZmDWF1* (Cao et al. 2022b), calcium ions (Ca^{2+}) act as important secondary messengers in the auxin and brassinosteroid signaling pathways, which are key hormonal regulators of leaf angle development (Yan et al. 2015). The Zm00001eb317760 candidate gene has been associated with *Squamosa Promoter-Binding-Like protein 6* (*spl6*). This gene belongs to the SPL transcription factor family, which is involved in the regulation of plant architecture, such as leaf angle and plant and ear height (Wei et al. 2018). The candidate gene *dek41*, associated with FLA, was characterized as a seedling-lethal mutant that causes developmental defects and it is required for correct mitochondrial functioning and also maize kernel development (Zhu et al. 2019). The *saur63*, another candidate gene associated with FLA, seems to play a positive role in cell expansion in Arabidopsis (Chae et al. 2012). The candidate gene *zmm7* (Zm00001eb317770) encodes a developmental protein belonging to the MADS-box transcription factor family, specifically a homolog of *SEPALLATA 2* (*sep2*). In Arabidopsis the MADS-box genes are well known for their roles in floral organ identity and reproductive development (Takagi et al. 2025). In maize, MADS-box genes are involved in inflorescence architecture (Zhang et al. 2012; Shen et al. 2023). On the other hand, for the EH trait, the gene Zm00001eb306210 encodes a *Serine Carboxypeptidase-Like 19*, a member of the SCPL acyltransferase subfamily. In maize, this gene has been implicated in specialized metabolic pathways such as the biosynthesis of *cyanidin 3,7-diglucoside polyacylation* and the *violdelphin biosynthesis*. In wheat, serine carboxypeptidase-like proteins have been characterized as effectors involved in *Fusarium graminearum* infection and stress tolerance (Liu et al. 2024). The gene *fdh2* (Zm00001eb382240) encodes *formate dehydrogenase 2*, an enzyme involved in the oxalate degradation VI pathway. In *Arabidopsis*, the homologous gene is responsible for oxalate degradation (Foster et al., 2012).This result elucidates the consistency of the association, and it might be used as a criterion to prioritize candidate genes for further functional validation and incorporation into breeding programs. Thus, the validation of molecular markers associated with candidate genes found in our study and their incorporation through marker assisted selection in maize breeding programs may help in the development of new varieties more suited to meet the challenges posed by climate change and resource limitations (Ibrahim et al. 2020).

Regarding genomic prediction, our results across the entire panel of DH lines showed that the values of prediction ability based on LD_{hap} approach were slightly higher than based on the other approaches for most traits, although the differences among traits/approaches were very small. This suggests that information carried by single-SNP approach may reflect information contained in the haplotype-based prediction approach. As a result, we may capture a substantial amount of variation when employing single-SNP that is comparable to what is captured by haplotypes. Weber et al. (2023) also observed a few increases in the values of prediction ability with three haplotype-based genomic prediction approaches in canola, maize, wheat, and soybean compared with single-SNP. In a recent simulation study, Araujo et al. (2021) concluded that haplotype-based models did not improve the performance of genomic prediction of breeding values in genetically diverse livestock populations. On the other hand, Sallam et al. (2020) found that the implementation of haplotype-based prediction approach improved the prediction ability of yield in wheat by 14.3%. According to Lin et al. (2024), the extent of improvement with haplotypes compared with single-SNP based predictions can be affected by the genetic structure of the population, genetic architecture of the trait, and the method employed for haplotype construction.

In the scenarios of genomic prediction using different BSSS subpopulations as training set, we observed that for ASI, EH and TR, the BSSS population presented good performance in predicting the other two subpopulations. The superiority of BSSS in the prediction of these three traits may be due to its broader range of alleles and genetic variants. On the other hand, as recurrent selection progressed through cycles, BSSS(R)17 presents reduced diversity compared to BSSS (Ledesma et al. 2023a), potentially limiting the range of genetic variants available for genomic prediction and resulting in lower prediction abilities, especially for NTB and TR. This underscores the importance of choosing training sets that are closely related to the validation set. Additionally, it highlights the importance of periodic updates of prediction models to maintain prediction ability. The selection pressure applied in BSSS(R)C17 has caused changes in allele frequencies and population structure (Ledesma et al. 2023a, b) which make it more distantly related to BSSS and leads to reduction in prediction ability when it is used as training set in the model. It has been shown that prediction ability tends to decrease over successive generations due to reduced genetic relationships between the training and validation sets (Habier et al. 2007; Pszczola et al. 2012). For FLA, NTB and PH traits, the genomic prediction using the BSSS/BSSS(R)17 as training population showed higher prediction abilities than the other strategies. This result was expected considering

that BSSS captures the initial genetic diversity, while BSSS(R)C17 reflects the accumulated effects of selection for these traits. Models trained on the cross between BSSS and BSSS(R)C17 might better calibrate genetic effects and allele frequencies across different stages of selection and, consequently, lead to more accurate predictions compared to models trained solely on BSSS or BSSS(R)C17. This latest strategy can overestimate or underestimate the genetic effects.

In conclusion, incorporating both single-SNP and haplotype-based GWAS approaches can maximize the probability of uncovering causal variants underlying complex traits. This complementary approach increases the robustness of genetic association studies, especially for traits influenced by multiple small-effect variants and complex interactions. Additionally, applying both single-SNP and haplotype-based genomic prediction approaches allows for selecting the one that yields the highest prediction ability for each trait, as different traits may require distinct approaches to accurately capture their genetic architecture. Selecting the most effective genomic prediction approach, whether single-SNP or haplotype-based, should be guided by the trait's genetic architecture and empirical validation of prediction ability. Furthermore, the effectiveness of one BSSS subpopulation to predict the other depends on the trait and the training population used.

5. References

Abed A, Belzile F (2019) Comparing single-SNP, multi-SNP, and haplotype-based approaches in association studies for major traits in barley. Plant Genome 12(3):190036. https://doi.org/10.3835/plantgenome2019.05.0036

Andrade ACB, Viana JMS, Pereira HD, Pinto VB, Silva FF (2019) Linkage disequilibrium and haplotype block patterns in popcorn populations. PLoS ONE 14(9):e0219417. https://doi.org/10.1371/journal.pone.0219417

Araujo AC, Carneiro PLS, Oliveira HR, Schenkel FS, Veroneze R, Lourenco DAL, Brito LF (2021) A comprehensive comparison of haplotype-based single-step genomic predictions in livestock populations with different genetic diversity levels: A simulation study. Front Genet 12:729867. https://doi.org/10.3389/fgene.2021.729867

Bhat JA, Yu D, Bohra A, Ganie SA, Varshney RK (2021) Features and applications of haplotypes in crop breeding. Commun Biol 4(1):1266. https://doi.org/10.1038/s42003-021-02782-y

Bian C, Prakapenka D, Tan C, Yang R, Zhu D, Guo X, Liu D, Cai G, Li Y, Liang Z, Wu Z, Da Y, Hu X (2021) Haplotype genomic prediction of phenotypic values based on chromosome distance and gene boundaries using low-coverage sequencing in Duroc pigs. Genet Sel Evol 53(78):1-19. https://doi.org/10.1186/s12711-021-00661-y

Bisgaard S (2000) The design and analysis of $2^{k-p} \times 2^{q-r}$ split plot experiments. J Qual Technol 32:39. https://doi.org/10.1080/00224065.2000.11979970

Bornowski N, Michel KJ, Hamilton JP et al (2021) Genomic variation within the maize stiff-stalk heterotic germplasm pool. Plant Genome 14(3):e20114. https://doi.org/10.1002/tpg2.20114

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633-2635. https://doi.org/10.1093/bioinformatics/btm308

Brewbaker JL (2015) Diversity and genetics of tassel branch numbers in maize. Crop Sci 55:65-78. https://doi.org/10.2135/cropsci2014.03.0248

Brekke B, Edwards J, Knapp A (2011) Selection and adaptation to high plant density in the Iowa stiff stalk synthetic maize (*Zea mays* L.) population. Crop Sci 51(5):1965-1972. https://doi.org/10.2135/cropsci2010.09.0563

Broman KW, Weber JL (1999) Long homozygous chromosomal segments in reference families from the Centre d'Étude du Polymorphisme Humain. Am J Hum Genet 65:1493-1500. https://doi.org/10.1086/302661

Browning BL, Browning SR (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194(2):459-471. https://doi.org/10.1534/genetics.113.150029

Calus M, Meuwissen T, De Roos A, Veerkamp R (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561. https://doi.org/10.1534/genetics.107.080838.

Cao Y, Zhong Z, Wang H, Shen R (2022a) Leaf angle: a target of genetic improvement in cereal crops tailored for high-density planting. Plant Biotechnol J 20(3):426-436. https://doi.org/10.1111/pbi.13780

Cao Y, Dou D, Zhang D, Zheng Y, Ren Z, Su H, (2022b). ZmDWF1 regulates leaf angle in maize. Plant Science, 325, 111459. https://doi.org/10.1016/j.plantsci.2022.111459

Chae K, Isaacs CG, Reeves PH, Maloney GS, Muday GK, Nagpal P, Reed JW (2012) Arabidopsis SMALL AUXIN UP RNA63 promotes hypocotyl and stamen filament elongation. Plant J 71(4):684-697. https://doi.org/10.1111/j.1365-313X.2012.05024.x

CIMMYT (2005) Laboratory protocols: CIMMYT applied molecular genetics laboratory. 3rd edn. CIMMYT, Mexico.

Contreras-Soto RI, Mora F, Oliveira MAR, Higashi W, Scapim CA, Schuster I (2017) A Genome-wide association study for agronomic traits in soybean using SNP markers and SNP based haplotype analysis. PLoS One 12(2):e0171105. https://doi.org/10.1371/journal.pone.0171105

Cuyabano BC, Su G, Lund MS (2014) Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15:1-11. https://doi.org/10.1186/1471-2164-15-1171

Cuyabano BC, Su G, Lund MS (2015) Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15:1171. https://doi.org/10.1186/1471-2164-15-1171.

Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9(2):179-181. https://doi.org/10.1038/nmeth.1785

Edwards J (2011) Changes in plant morphology in response to recurrent selection in the Iowa Stiff Stalk Synthetic maize population. Crop Sci 51(6):2352-2361. https://doi.org/10.2135/cropsci2010.09.0564

Eberhart SA, Debela S, Hallauer AR (1973) Reciprocal recurrent selection in BSSS and BSCB1 maize populations and half-sib selection in BSSS. Crop Sci 13:451-456. https://doi.org/10.2135/cropsci1973.0011183X001300040017x

Elli EF, Edwards J, Yu J, Trifunovic S, Eudy DM, Kosola KR, Schnable PS, Lamkey KR, Archontouliset SV (2023) Maize leaf angle genetic gain is slowing down in the last decades. Crop Sci 63(6):3520-3533. https://doi.org/10.1002/csc2.21111

Feitosa FLB, Pereira ASC, Amorim ST, Peripolli E, Silva RMDO, Braz CU, Ferrinho AM, Schenkel FS, Brito LF, Espigolan R, Albuquerque LG, Baldi F (2020) Comparison between haplotype-based and individual SNP-based genomic predictions for beef fatty acid profile in Nelore cattle. J Anim Breed Genet 137(5):468-476. https://doi.org/10.1111/jbg.12463

Ferdosi MH, Henshall J, Tier B (2016) Study of the optimum haplotype length to build genomic relationship matrices. Genet Sel Evol 48:75. https://doi.org/10.1186/s12711-016-0253-6

Fleischmann TT, Scharff LB, Alkatib S, Hasdorf S, Schöttler MA, Bock, R. (2011). Nonessential plastid-encoded ribosomal proteins in tobacco: a developmental role for plastid translation and implications for reductive genome evolution. The Plant Cell, 23(9), 3137-3155. https://doi.org/10.1105/tpc.111.088906

Foster J, Kim HU, Nakata PA, Browse J. (2012). A previously unknown oxalyl-CoA synthetase is important for oxalate catabolism in *Arabidopsis*. The Plant Cell, 24(3), 1217-1229. https://doi.org/10.1105/tpc.112.096032

Gabriel SB, Schaffner SF, Nguyen H et al (2002) The structure of haplotype blocks in the human genome. Science 296(5576):2225-2229. https://doi.org/10.1126/science.1069424

Galli G, Gezan SA, Murillo DA, Murray D (2022) ASRgwas: An R package to perform complex Genome-Wide Association Studies (GWAS). Version 1.0.0. VSN International, Hemel Hempstead, United Kingdom.

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

Gao X, Becker LC, Becker DM, Starmer JD, Province MA (2010) Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol 34:100-105. https://doi.org/10.1002/gepi.20430

Gao X (2011) Multiple testing corrections for imputed SNPs. Genet Epidemiol 35(3):154-158. https://doi.org/10.1002/gepi.20563

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Giraud H, Bauland C, Falque M, Madur D, Combes V, Jamin P, et al. (2017). Reciprocal genetics: identifying QTL for general and specific combining abilities in hybrids between multiparental populations from two maize (*Zea mays* L.) heterotic groups. Genetics, 207(3), 1167-1180. https://doi.org/10.1534/genetics.117.300305

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397. https://doi.org/10.1534/genetics.107.081190

He S, Thistlethwaite R, Forrest K, Shi F, Hayden MJ, Trethowan R, Daetwyler HD (2019) Extension of a haplotype-based genomic prediction model to manage multi-environment wheat

data using environmental covariates. Theor Appl Genet 132:3143-3154. https://doi.org/10.1007/s00122-019-03413-1

He S, Liang S, Meng L, Cao L, Ye G (2023) Sparse phenotyping and haplotype-based models for genomic prediction in rice. Rice 16:27. https://doi.org/10.1186/s12284-023-00643-2

Hess M, Druet T, Hess A, Garrick D (2017) Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genet Sel Evol 49(54):1-14. https://doi.org/10.1186/s12711-017-0329-y.

Hickey J, Kinghorn B, Tier B, Clark SA, van der Werf J, Gorjanc G (2013) Genomic evaluations using similarity between haplotypes. J Anim Breed Genet 130:259-269. https://doi.org/10.1111/jbg.12020.

Hou Y, Zeng W, Ao C, Huang, J. (2024). Integrative analysis of the transcriptome and metabolome reveals Bacillus atrophaeus WZYH01-mediated salt stress mechanism in maize (*Zea mays* L.). *Journal of Biotechnology*, *383*, 39-54. https://doi.org/10.1016/j.jbiotec.2024.02.004

Ibrahim AK, Zhang L, Niyitanga S, Afzal MZ, Xu Y, Zhang L, Qi J (2020) Principles and approaches of association mapping in plant breeding. Trop Plant Biol 13:212-224. https://doi.org/10.1007/s12042-020-09261-4

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299-1320. https://doi.org/10.1038/nature04226

Jiao Y, Peluso P, Shi J, et al (2017) Improved maize reference genome with single-molecule technologies. Nature 546:524-527. https://doi.org/10.1038/nature22971

Johansson AM, Pettersson ME, Siegel PB, Carlborg O (2010) Genome-wide effects of long-term divergent selection. PLoS Genet 6:e1001188. https://doi.org/10.1371/journal.pgen.1001188.

Khankhanian P, Gourraud PA, Lizee A, Goodin DS (2015) Haplotype-based approach to known MS-associated regions increases the amount of explained risk. J Med Genet 52(9):587-594. https://doi.org/10.1136/jmedgenet-2015-103071

Kilian A, Wenzl P, Huttner E et al. (2012) Diversity arrays technology: A generic genome profiling technology on open platforms. Methods Mol Biol 888:67-89. https://doi.org/10.1007/978-1-61779-870-2_5

Lamkey KR, Peterson PA, Hallauer AR (1991) Frequency of the transposable element uq in Iowa stiff stalk synthetic maize populations. Genet Res 57:1-9. https://doi.org/10.1017/S0016672300028962

Ledesma A, Ribeiro FAS, Uberti A, Edwards J, Hearne S, Frei U, Lübberstedt T (2023a) Molecular characterization of doubled haploid lines derived from different cycles of the Iowa Stiff Stalk Synthetic (BSSS) maize population. Front Plant Sci 14:1226072. https://doi.org/10.3389/fpls.2023.1226072

Ledesma A, Santana AS, Ribeiro FAS, Aguilar FS, Edwards J, Frei U, Lübberstedt T (2023b) Genome-wide association analysis of plant architecture traits using doubled haploid lines derived from different cycles of the Iowa Stiff Stalk Synthetic maize population. Front Plant Sci 14:1294507. https://doi.org/10.3389/fpls.2023.1294507

Lewontin R (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49-67. https://doi.org/10.1093/genetics/49.1.49

Li Q, Du J, Qiao Z, Pan C, He W, Zhang L, et al. (2023a). White and green striate leaves 1, predicted to encode a 16S rRNA processing protein, plays a critical role in the processing of chloroplast ribosomes in maize (*Zea mays* L.). Molecular Breeding, 43(8), 65. https://doi.org/10.1007/s11032-023-01407-y

Li J, Wang X, Wei J, Miao X, Shang X, Li L (2023b) Genetic mapping and functional analysis of a classical tassel branch number mutant *Tp2* in maize. Front Plant Sci 14:1183697. https://doi.org/10.3389/fpls.2023.1183697

Lin Y-C, Mayer M, Torres DV, Pook T, Hölker AC, Presterl T, Ouzunova M, Schön C-C (2024) Genomic prediction within and across maize landrace derived populations using haplotypes. Front Plant Sci 15:1351466. https://doi.org/10.3389/fpls.2024.1351466

Liu K, Wang X, Qi Y, Li Y, Shi Y, Ren Y, et al. (2024). Effector protein serine carboxypeptidase FgSCP is essential for full virulence in Fusarium graminearum and is involved in modulating plant immune responses. Phytopathology, 114(9), 2131-2142. https://doi.org/10.1094/PHYTO-02-24-0068-R

Lorenz AJ, Hamblin MT, Jannink JL (2010) Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. PloS one 5(11):e14079. https://doi.org/10.1371/journal.pone.0014079

Lu Y, Xu J. Yuan Z, Hao Z, Xie C, Li X, et al. (2012). Comparative LD mapping using single SNPs and haplotypes identifies QTL for plant height and biomass as secondary traits of drought tolerance in maize. Molecular Breeding, 30(1), 407-418. https://doi.org/10.1007/s11032-011-9631-5

Luan T, Woolliams JA, Ødegård J, Dolezal M, Roman-Ponce SI, Bagnato A, Meuwissen THE (2012) The importance of identity-by-state information for the accuracy of genomic selection. Genet Sel Evol 44(28):1-7. https://doi.org/10.1186/1297-9686-44-28

Mackay L, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12(2):57-63. https://doi.org/10.1016/j.tplants.2006.12.001

Maldonado C, Mora F, Scapim CA, Coan M (2019) Genome-wide haplotype-based association analysis of key traits of plant lodging and architecture of maize identifies major determinants for leaf angle: Hap LA4. PloS One, 14(3):e0212925. https://doi.org/10.1371/journal.pone.0212925

Matias FI, Galli G, Granato ISC, Fritsche-Neto R (2017) Genomic prediction of autogamous and allogamous plants by SNPs and haplotypes. Crop Sci 57(6):2951-2958. https://doi.org/10.2135/cropsci2017.01.0022

Meuwissen TH, Odegard J, Andersen-Ranberg I, Grindflek E (2014) On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. Genet Sel Evol 46:1-8. https://doi.org/10.1186/1297-9686-46-49

Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G, Myles S (2016) LinkImpute: fast and accurate genotype imputation for non-model. G3: Genes, Genomes, Genetics 5(11):2383-2390. https://doi.org/10.1534/g3.115.021667

Mucha A, Wierzbicki H, Kamiński S, Oleński K, Hering D (2019) High-frequency marker haplotypes in the genomic selection of dairy cattle. J Appl Genet 60:179-186. https://doi.org/10.1007/s13353-019-00489-9

Negro SS, Millet EJ, Madur D, Bauland C, Combes V, Welcker C, et al. (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. BMC Plant Biology, 19(1), 318. https://doi.org/10.1186/s12870-019-1926-4

Portwood JL, Woodhouse MR, Cannon EK, Gardiner JM, Harper LC, Schaeffer ML, et al. (2019). MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Res 47(D1):D1146-D1154. https://doi.org/10.1093/nar/gky1046

Pszczola M, Strabel T, Mulder A, Calus MPL (2012) Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci 95:389-400. https://doi.org/10.3168/jds.2011-4338

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559-575. https://doi.org/10.1086/519795

Ruf S, Trösch R, Schollbach L, Kroop X, Forner J, Gefen-Treves S, et al. (2025). Reverse genetics in the Arabidopsis chloroplast genome identifies rps16 as a transcribed pseudogene. The Plant Journal, 122(3), e70198. https://doi.org/10.1111/tpj.70198

Saad MN, Mabrouk MS, Eldeib AM, Shaker OG (2018) Comparative study for haplotype block partitioning methods - Evidence from chromosome 6 of the North American Rheumatoid Arthritis Consortium (NARAC) dataset. PloS One 13(12):e0209603. https://doi.org/10.1371/journal.pone.0209603

Sallam AH, Conley E, Prakapenka D, Da Y, Anderson JA (2020) Improving prediction accuracy using multi-allelic haplotype prediction and training population optimization in wheat. G3: Genes, Genomes, Genetics 10(7):2265-2273. https://doi.org/10.1534/g3.120.401165

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. (2009). The B73 maize genome: Complexity, diversity, and dynamics. Science 326:1112-1115. https://doi.org/10.1126/science.1178534

Schuetz SH, Mock JJ (1978) Genetics of tassel branch number in maize and its implications for a selection program for small tassel size. Theor Appl Genet, 53:265-271. https://doi.org/10.1007/BF00280990

Sehgal D, Mondal S, Crespo-Herrera L, Velu G, Juliana P, Huerta-Espino J, Shrestha S, Poland J, Singh R, Dreisigacker S (2020) Haplotype-Based, Genome-Wide Association Study Reveals Stable Genomic Regions for Grain Yield in CIMMYT Spring Bread Wheat. Front Genet 11:589490. https://doi.org/10.3389/fgene.2020.589490.

Sehgal D, Rathan ND, Özdemir F, Keser M, Akin B, Dababat AA, Koc E, Dreisigacker S, Morgounov A (2024) Genomic wide association study and selective sweep analysis identify genes associated with improved yield under drought in Turkish winter wheat germplasm. Sci Rep 14:8431. https://doi.org/10.1038/s41598-024-57469-1

Shapiro SS, Wilk MB (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3-4), 591-611. https://doi.org/10.1093/biomet/52.3-4.591

Shen X, Xiao B, Kaderbek T, Lin Z, Tan K, Wu Q, et al. (2023). Dynamic transcriptome landscape of developing maize ear. The Plant Journal, 116(6), 1856-1870. https://doi.org/10.1111/tpj.16457

Shipilina D, Pal A, Stankowski S, Chan YF, Barton NH (2023) On the origin and structure of haplotype blocks. Mol Ecol 32(6):1441-1457. https://doi.org/10.1007/s12571-011-0140-5

Sprague GF (1946) Early Testing of Inbred Lines of Corn. Agron J. https://doi.org/10.2134/agronj1946.00021962003800020002x

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genetics, 5(11), e1000734. https://doi.org/10.1371/journal.pgen.1000734

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68(4):978-989. https://doi.org/10.1086/319501

Takagi H, Lee N, Hempton AK, Purushwani S, Notaguchi M, Yamauchi K, et al. (2025). Florigen-producing cells express FPF1-LIKE PROTEIN 1 to accelerate flowering and stem growth in *Arabidopsis*. Developmental Cell. https://doi.org/10.1016/j.devcel.2025.02.003

Tang D, Chen Z, Ni J, Jiang Q, Li P, Wang L, Zhou J, Li C, Liu J (2021) Identification of QTL for leaf angle at canopy-wide levels in maize. Euphytica 217:75. https://doi.org/10.1007/s10681-021-02781-4

Teissier M, Larroque H, Brito LF, Rupp R, Schenkel FS, Robert-Granié C (2020) Genomic predictions based on haplotypes fitted as pseudo-SNP for milk production and udder type traits and SCS in French dairy goats. J Dairy Sci, 103(12):11559-11573. https://doi.org/10.3168/jds.2020-18662

Troyer AF (1999) Background of US hybrid corn. Crop Sci 39(3):601-626. https://doi.org/10.2135/cropsci1999.0011183X003900020001xa

Utsunomiya YT, Milanesi M, Utsunomiya AT, Ajmone-Marsan P, Garcia JF (2016) GHap: an R package for genome-wide haplotyping. Bioinformatics 32(18):2861-2862. https://doi.org/10.1093/bioinformatics/btw356

Vahedi SM, Ardetani SS, Brito LF, Karimi K, Afshari KP, Banabazi MH (2023) Expanding the application of haplotype-based genomic predictions to the wild: A case of antibody response against *Teladorsagia circumcincta* in Soay sheep. BMC Genomics 24:335. https://doi.org/10.1186/s12864-023-09407-0

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414-4423. https://doi.org/10.3168/jds.2007-0980

Villumsen TM, Janss L, Lund MS (2009) The importance of haplotype length and heritability using genomic selection in dairy cattle. J Anim Breed Genet 126(1):3-13. https://doi.org/10.1111/j.1439-0388.2008.00747.x

Weber SE, Frisch M, Snowdon RJ, Voss-Fels KP (2023) Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. Front Plant Sci 14:1217589. https://doi.org/10.3389/fpls.2023.1217589

Wei H, Zhao Y, Xie Y, Wang H. (2018). Exploiting SPL genes to improve maize plant architecture tailored for high-density planting. Journal of Experimental Botany, 69(20), 4675-4688. https://doi.org/10.1093/jxb/ery258

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. New York, Springer-Verlag.

Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9(1):60-62.

Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14(7):507-515. https://doi.org/10.1038/nrg3457

Wu P, Wang K, Zhou J, Chen D, Jiang A, Jiang Y, Zhu L, Qiu X, Li X, Tang G (2021) A combined GWAS approach reveals key loci for socially-affected traits in Yorkshire pigs. Commun Biol 4(891). https://doi.org/10.1038/s42003-021-02416-3

Yan J, Guan L, Sun Y, Zhu Y, Liu L, Lu R, et al. (2015). Calcium and ZmCCaMK are involved in brassinosteroid-induced antioxidant defense in maize leaves. Plant and Cell Physiology, 56(5), 883-896. https://doi.org/10.1093/pcp/pcv014

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 4(4):355-360. https://doi.org/10.1038/ng.546

Zhang Z, Li H, Zhang D, Liu Y, Fu J, Shi Y, et al. (2012). Characterization and expression analysis of six MADS-box genes in maize (*Zea mays* L.). Journal of Plant Physiology, 169(8), 797-806. https://doi.org/10.1016/j.jplph.2011.12.020

Zhang J, Ku LX, Han ZP, Guo SL, Liu HJ, Zhang ZZ, et al. (2014). The ZmCLA4 gene in the qLA4-1 QTL controls leaf angle in maize (*Zea mays* L.). Journal of Experimental Botany, 65(17), 5063-5076. https://doi.org/10.1093/jxb/eru271

Zhang M, Jin Y, Ma Y, Zhang Q, Wang Q, Jiang N, Zhao H, Qu J, Guan S, Wanget P (2021) Identification of QTLs and candidate genes associated with leaf angle and leaf orientation value in maize (*Zea mays* L.) based on GBS. Trop Plant Biol 14:34-49. https://doi.org/10.1007/s12042-020-09270-3

Zhu C, Jin G, Fang P, Zhang Y, Feng X, Tang Y, Qi W, Songet R (2019) Maize pentatricopeptide repeat protein DEK41 affects cis-splicing of mitochondrial nad4 intron 3 and seed development. J Exp Bot 70(15):3795-3808. https://doi.org/10.1093/jxb/erz193

TABLES AND FIGURES

Table 1. Ranges and means of the best linear unbiased estimator, variance component estimates due to lines ($\hat{\sigma}_g^2$), subpopulation ($\hat{\sigma}_p^2$), residual variance ($\hat{\sigma}^2$) estimates, estimates of heritability coefficient ($H_{\bar{X}}^2$) and coefficient of variation (CV%) for anthesis-silking interval (ASI, days), plant height (PH, cm), ear height (EH, cm), flag leaf angle (FLA), number of primary tassel branches (NTB) and tiller rate (TR) measured in a set of DH lines derived from the BSSS maize population.

Trait	Min	Mean	Max	$\hat{\sigma}_g^2$	$\hat{\sigma}_p^2$	$\hat{\sigma}^2$	$H_{\bar{X}}^2$	CV%
ASI	-3.50	3.50	9.00	1.32*	0.77*	0.88	0.80	26.92
PH	105.0	170.3	241.0	80.77*	0.24*	83.5	0.83	22.95
EH	0.25	76.4	152.0	65.52*	18.63*	73.9	0.92	21.86
FLA	9.00	37.50	92.50	111.76*	192.22*	78.75	0.88	23.70
NTB	2.00	27.00	47.00	13.67*	16.96*	4.59	0.93	8.28
TR	8.71	32.01	86.26	128.95*	17.49*	84.29	0.89	28.63

* Significant at 5% by the LRT test.

Table 2. Number of SNP markers and haplotypes (nHap) per chromosome (Chr), minimum, mean and maximum number of SNPs per haplotypes within chromosomes for the LD_{hap} and LD_{r^2} approaches, and haplotype length within chromosome for the LD_{hap} approach.

Chr	Number of SNPs	LD_{hap}					LD_{r^2}			
		nHap	Min	Mean	Max	Length (kb)	nHap	Min	Mean	Max
1	2,174	317	2	11.3	30	774	233	2	4.93	12
2	1,639	265	2	9.8	20	529	162	2	4.82	11
3	1,468	206	2	11.9	28	799	133	2	4.1	9
4	1,406	217	2	11.1	26	781	152	2	4.2	9
5	1,542	273	2	9.6	20	386	179	2	4.1	8
6	1,090	207	2	8.5	20	467	104	2	4.2	13
7	1,313	227	2	9.4	22	445	144	2	3.8	10
8	1,146	200	2	10.2	22	597	139	2	4.7	11
9	1,151	218	2	9.5	28	426	110	2	3.8	10
10	917	170	2	9.2	20	348	87	2	4	11

Table 3. Summary of significant genomic regions found by the single-SNP and haplotype-based approaches for flag leaf angle (FLA), ear height (EH), number of primary tassel branches (NTB) and plant height (PH) measured in a set of DH lines derived from the BSSS maize population.

Trait	Approach	Chr	Allele	Start position (kb)	*P*-value	PVE(%)[a]
FLA	LD_{hap}	7	TTTTATTGGG	140175001	2.56×10^{-7}	3.62
		7	AGCC	142845001	1.73×10^{-6}	3.32
	LD_{r^2}	7	AG	143116184	3.72×10^{-6}	3.10
EH	LD_{hap}	7	TTGAC	40050001	1.38×10^{-6}	2.52
		9	CG	51972001	2.41×10^{-6}	1.72
	Single-SNP	7	T/C	152391236	6.38×10^{-5}	2.60
NTB	Single-SNP	4	C/G	22430065	4.20×10^{-6}	3.54
PH	Single-SNP	7	T/C	40387690	4.72×10^{-6}	2.43

[a]PVE(%): proportion of phenotypic variance explained; Chr: chromosome number.

Table 4. Candidate genes associated with flag leaf angle (FLA) and ear height (EH) measured in a set of DH lines derived from the BSSS maize population.

Trait	Approach	Chr	Gene start (kb)	Gene ID maizeGDB	Gene name	Annotation
FLA	LD_{hap}	7	140285989	Zm00001eb317180	-	30S ribosomal protein S16 (Small subunit ribosomal protein 16)
		7	140341852	Zm00001eb317190	*pco143515*	Putative sarcosine oxidase
		7	140572304	Zm00001eb317200	-	Calcium-transporting ATPase (EC 7.2.2.10)
		7	142848972	Zm00001eb317730	*dek41*	Defective kernel41
		7	142862495	Zm00001eb317740	-	TITAN-like protein
		7	142908402	Zm00001eb317760	*spl6*	Squamosa promoter-binding-like protein 6
		7	142963674	Zm00001eb317770	*zmm7*	Developmental protein SEPALLATA 2 (MADS transcription factor)
		7	143066599	Zm00001eb317790	*sacd8*	Stearoyl-acyl-carrier-protein desaturase8
		7	143097907	Zm00001eb317810	*magi99589*	60S ribosomal protein L7a
		7	143104362	Zm00001eb317820	*saur63*	Small auxin up RNA63
		7	143190807	Zm00001eb317830	-	Pollen Ole e 1 allergen and extensin family protein
		7	143196678	Zm00001eb317850	*sweet11b*	sugars will eventually be exported transporter11b
		7	143201415	Zm00001eb317860	*pat31*	Protein S-acyltransferase31
EH	LD_{hap}	7	40067426	Zm00001eb306150	-	Protein kinase domain-containing protein
		7	40258356	Zm00001eb306170	*GRMZM2G143335*	Subtilisin-like protease 3
		7	40264468	Zm00001eb306180	-	Plant heme peroxidase family profile domain-containing protein
		7	40332323	Zm00001eb306200	-	TLD-domain containing nucleolar protein
		7	40372224	Zm00001eb306210	-	Serine carboxypeptidase-like 19
		9	52192999	Zm00001eb382230	*GRMZM2G086641*	Probable histone H2AXa
		9	52240187	Zm00001eb382240	*fdh2*	Formate dehydrogenase2

i)	BSSS (135)	BSSS(R)17 (194)	BSSS/BSSS(R)C17 (187)	Training set
ii)	BSSS (135)	BSSS(R)17 (194)	BSSS/BSSS(R)C17 (187)	
iii)	BSSS (135)	BSSS(R)17 (194)	BSSS/BSSS(R)C17 (187)	Testing set

Figure 1. Genomic prediction scenarios. Numbers in parentheses represent the number of DH lines in each set.

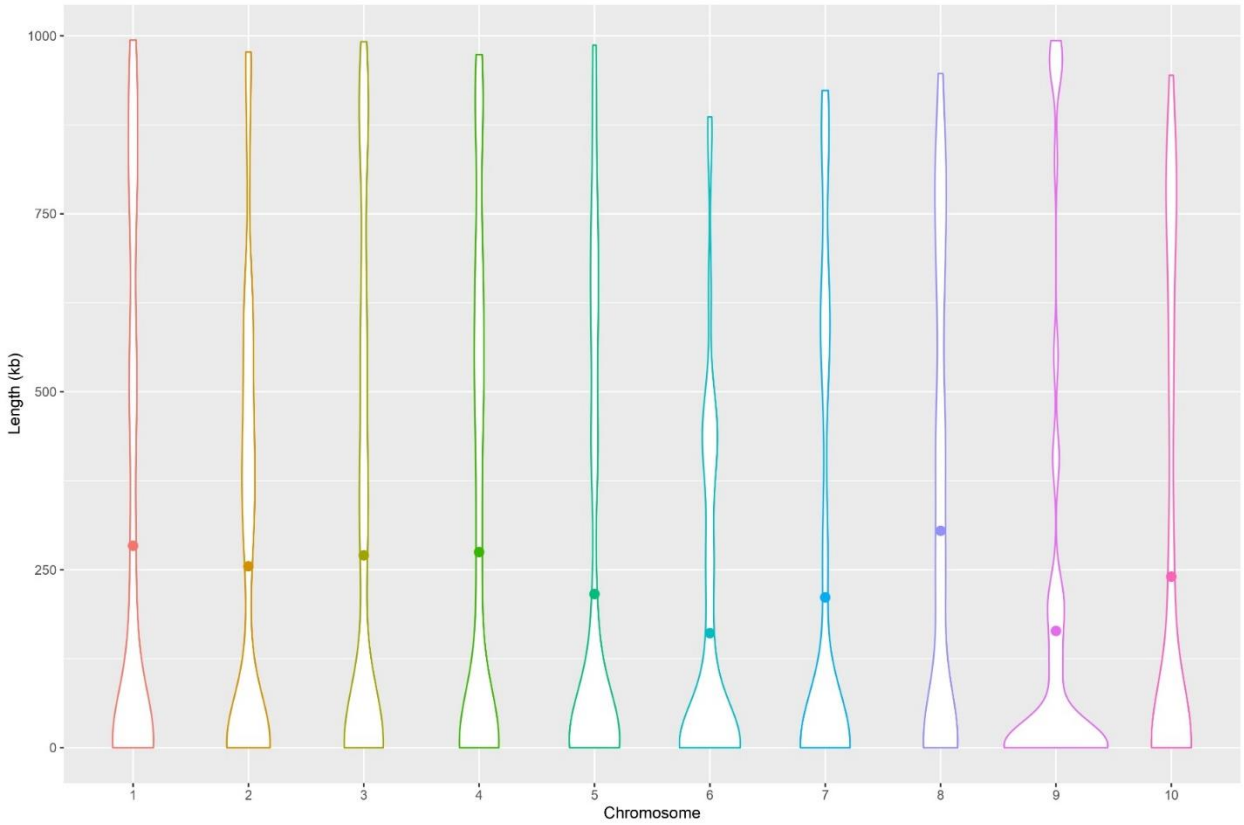


Figure 2. Violin plot of haplotype length (kb) across maize chromosomes using LD_{r^2} approach. Dots on the violin plot show the average haplotype length within each chromosome.

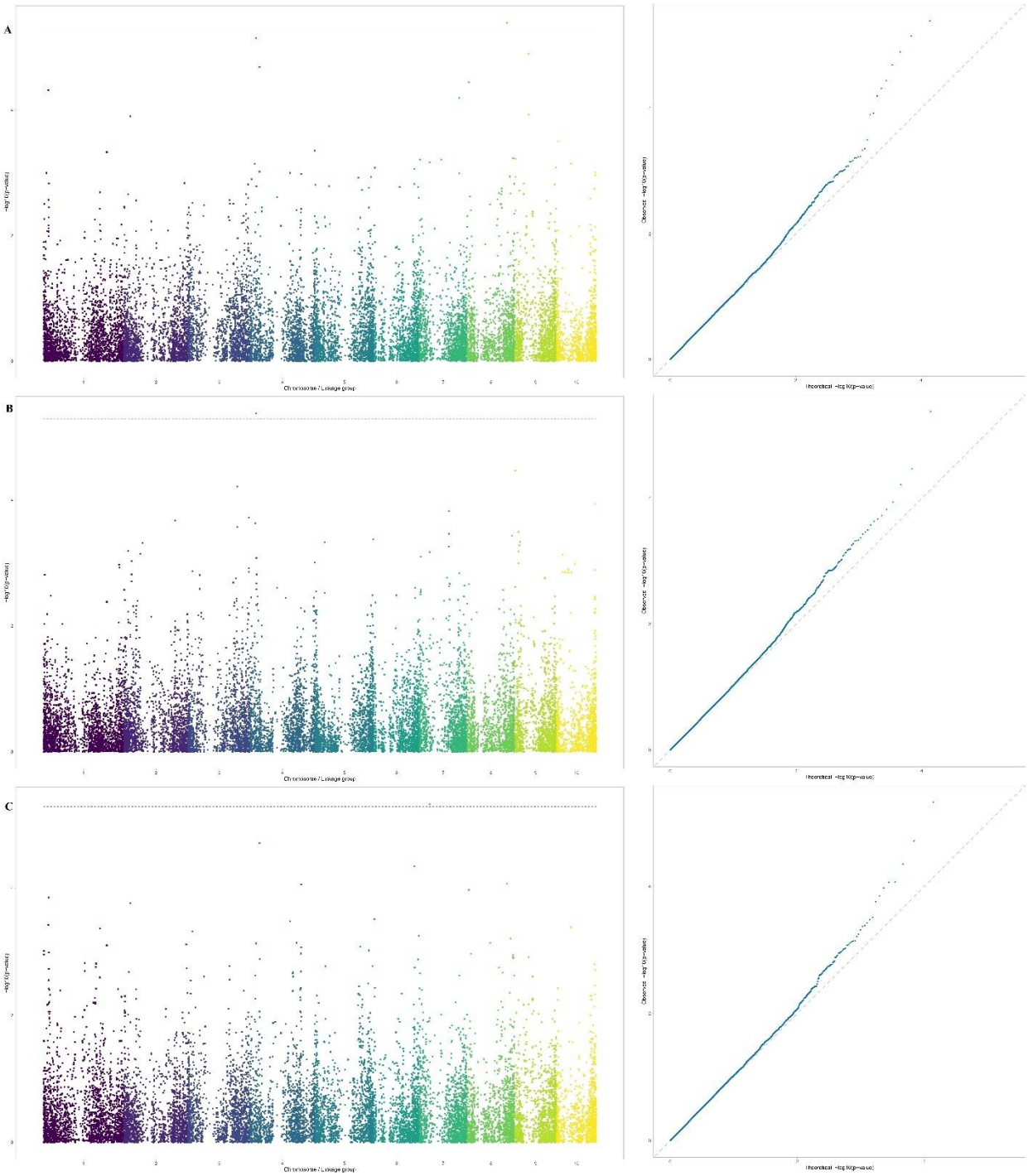


Figure 3. Manhattan plots and Q-Q plots for ear height (a), number of primary tassel branches (b) and plant height (c) for single-SNP approach.

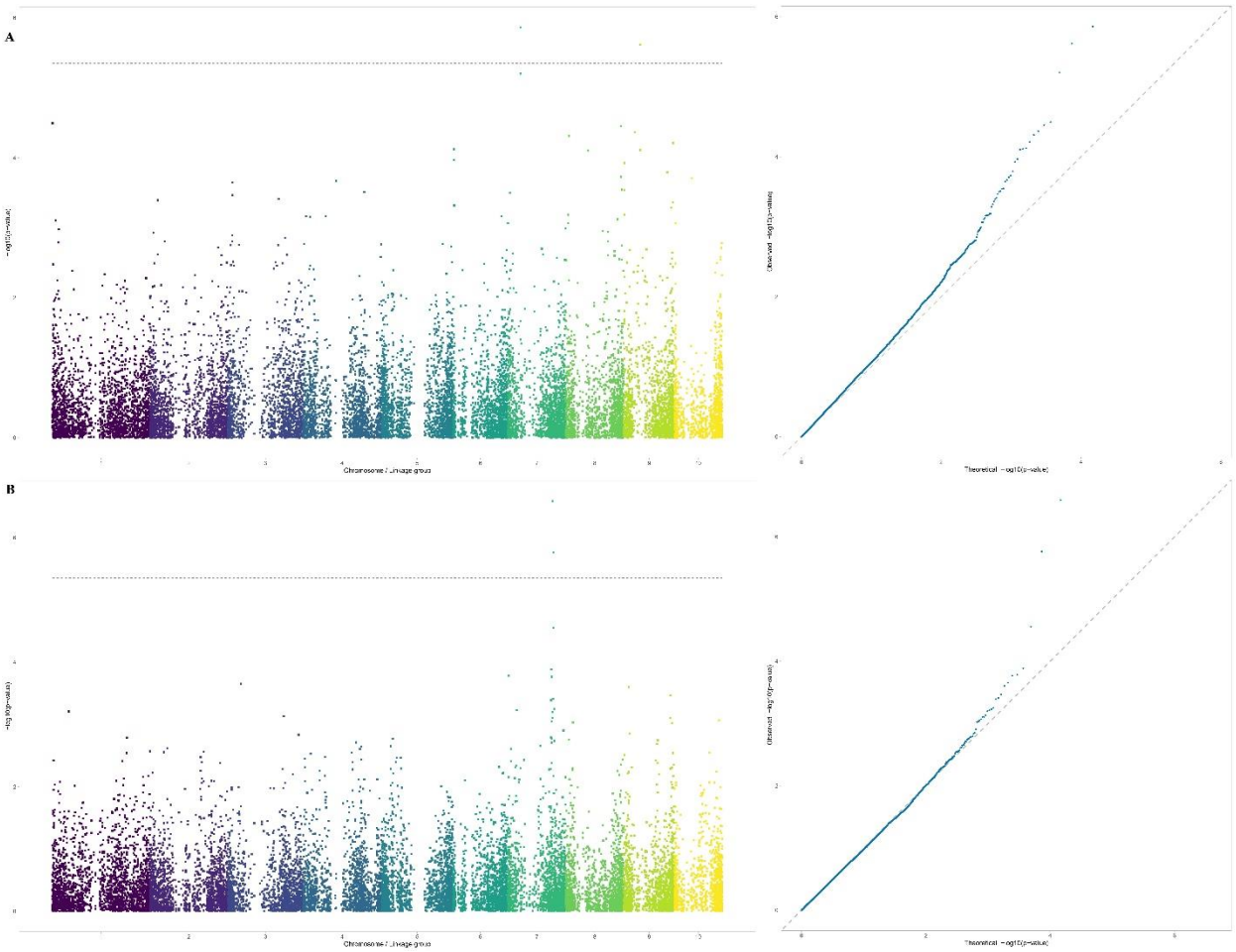


Figure 4. Manhattan plots and Q-Q plots for ear height (a) and flag leaf angle (b) in the LD_{hap} and approach.

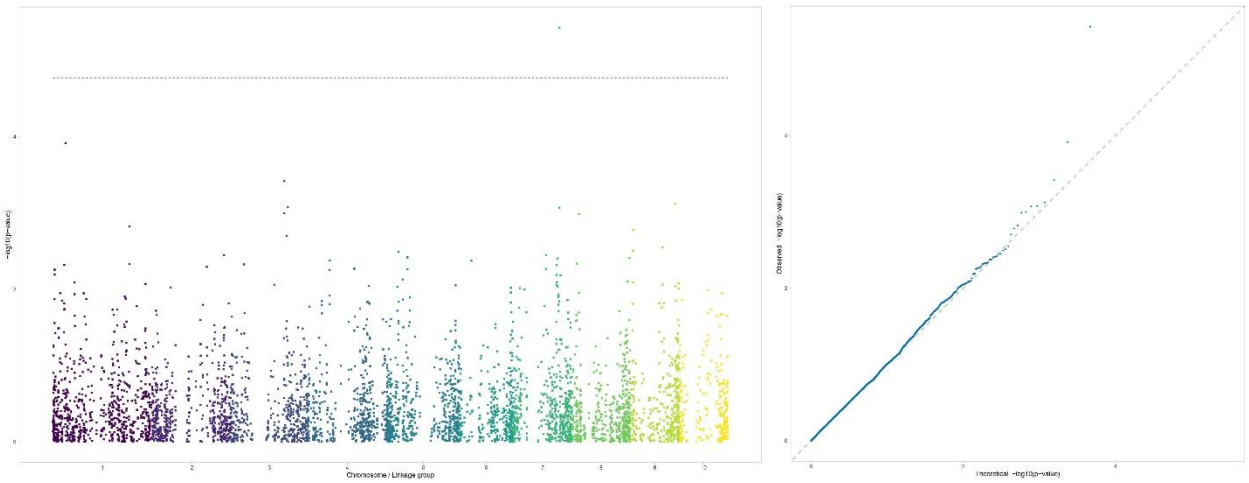


Figure 5. Manhattan plot and Q-Q plot of flag leaf angle in the LD_{r^2} approach.

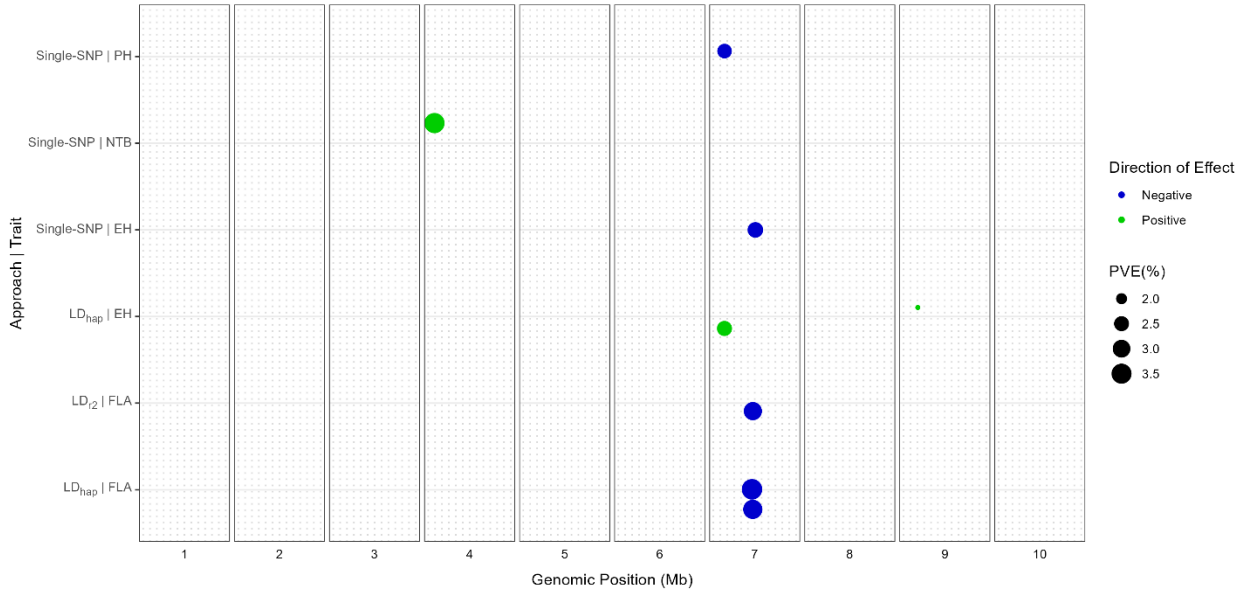


Figure 6. Final set of significant QTLs for flag leaf angle (FLA), ear height (EH), number of primary tassel branches (NTB), and plant height (PH) identified by the LD_{r^2}, LD_{hap}, and single-SNP approaches. Circle diameters are proportional to the phenotypic variance explained (PVE). Colors indicate the direction of effect: green and blue represent the reference allele increasing and decreasing the phenotypic value of the trait, respectively. Physical positions of the markers are based on the B73 reference genome RefGen_v5.

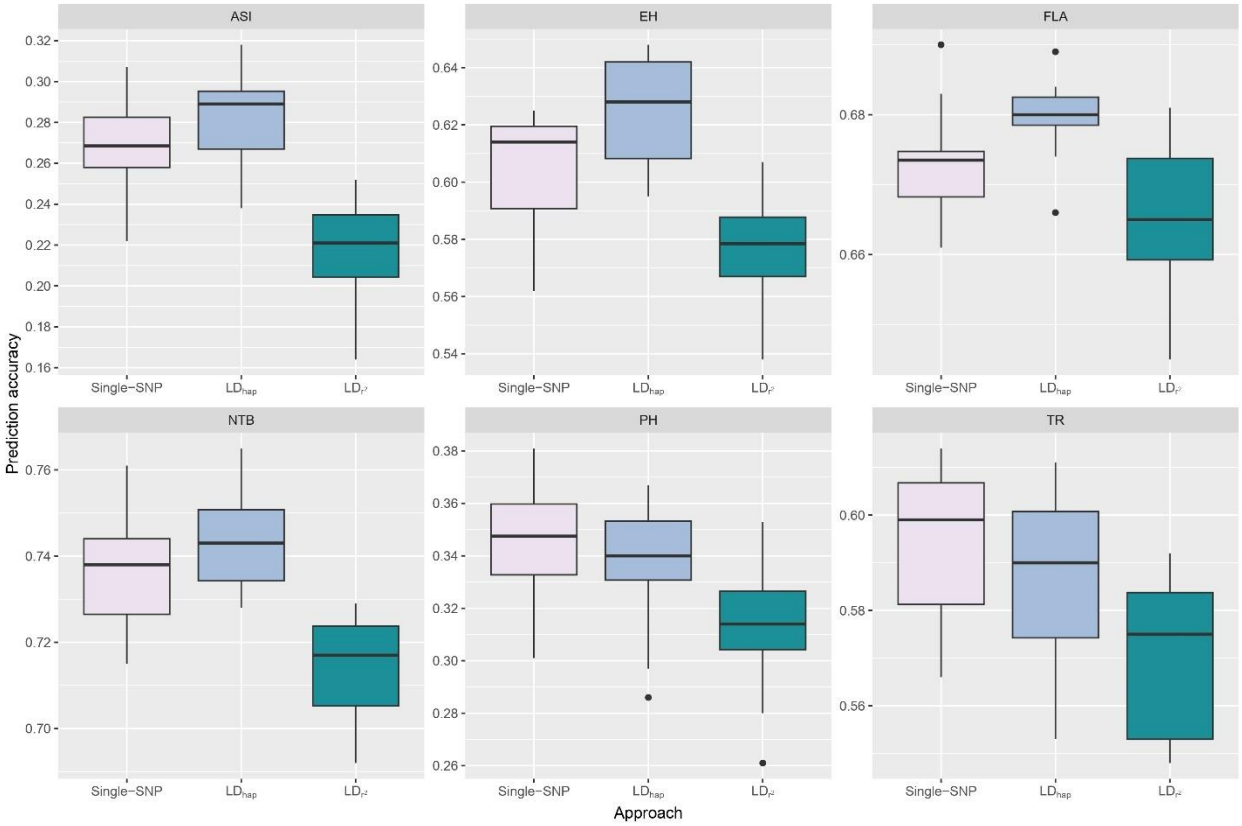


Figure 7. Box-plot of cross-validation prediction ability for anthesis-silking interval (ASI), ear height (EH, cm), flag leaf angle (FLA), number of primary tassel branches (NTB) plant height (PH, cm) and tiller rate (TR) using single-SNP, LD_{hap} and LD_{r^2} approaches and the entire population to randomly predict a set of DH lines in the BSSS maize population.

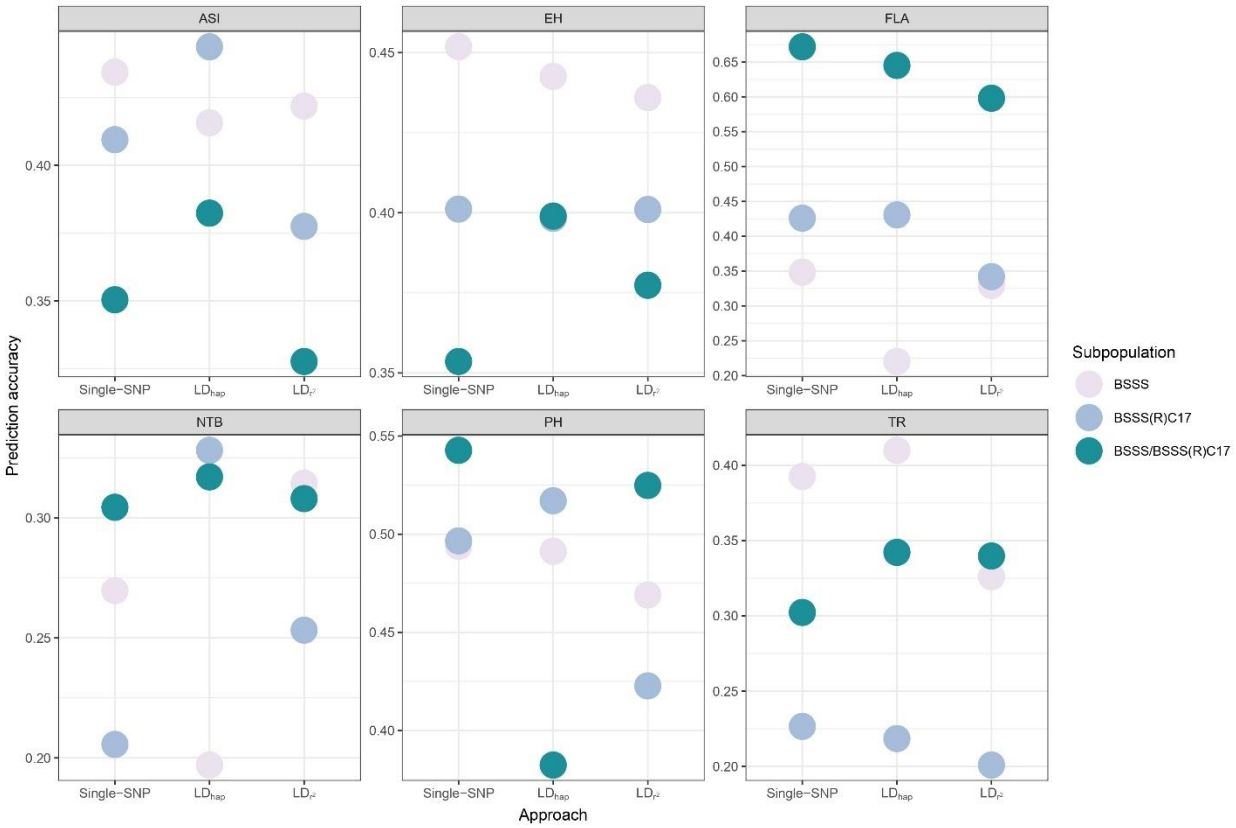
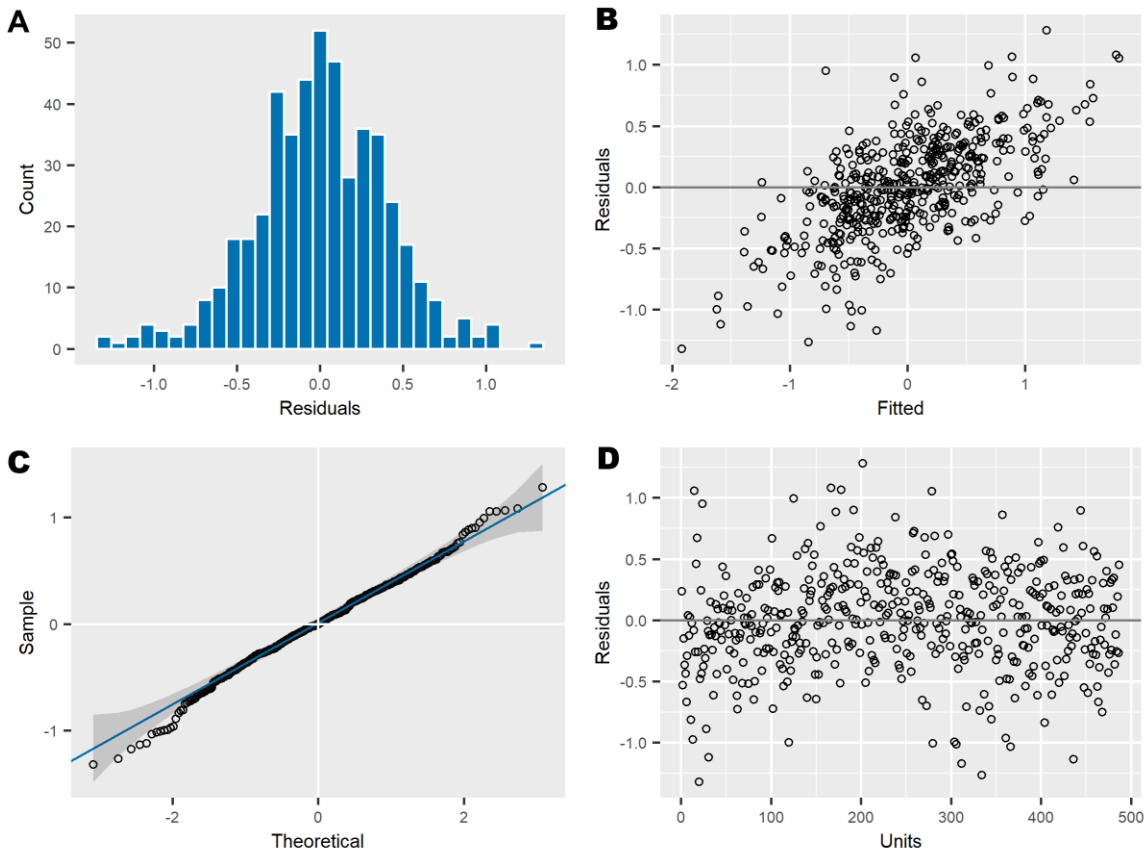
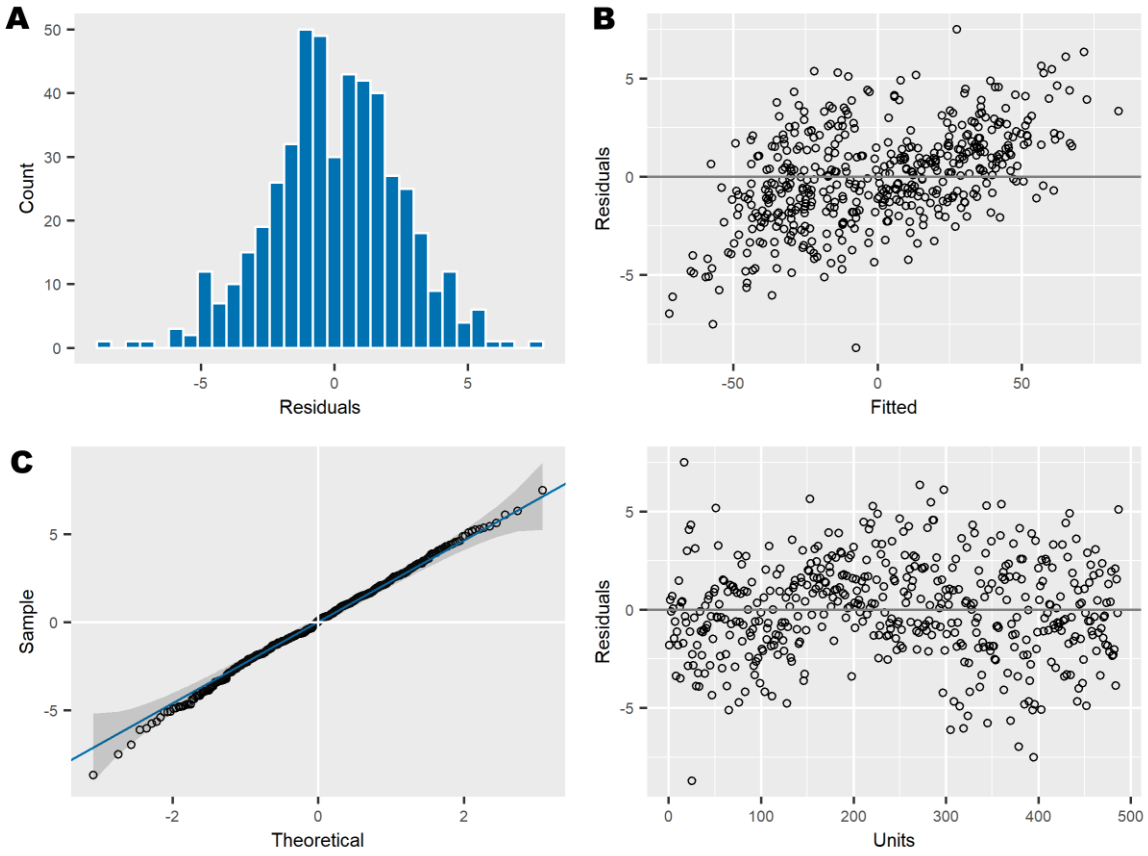


Figure 8. Plots on the prediction ability of anthesis-silking interval (ASI), ear height (EH, cm), flag leaf angle (FLA), number of primary tassel branches (NTB) plant height (PH, cm) and tiller rate (TR) evaluated in three different approaches to build the genomic kinship matrices (single-SNP, LD_{hap} and LD_{r^2}) and using different recurrent selection cycles of the BSSS maize population as training population.

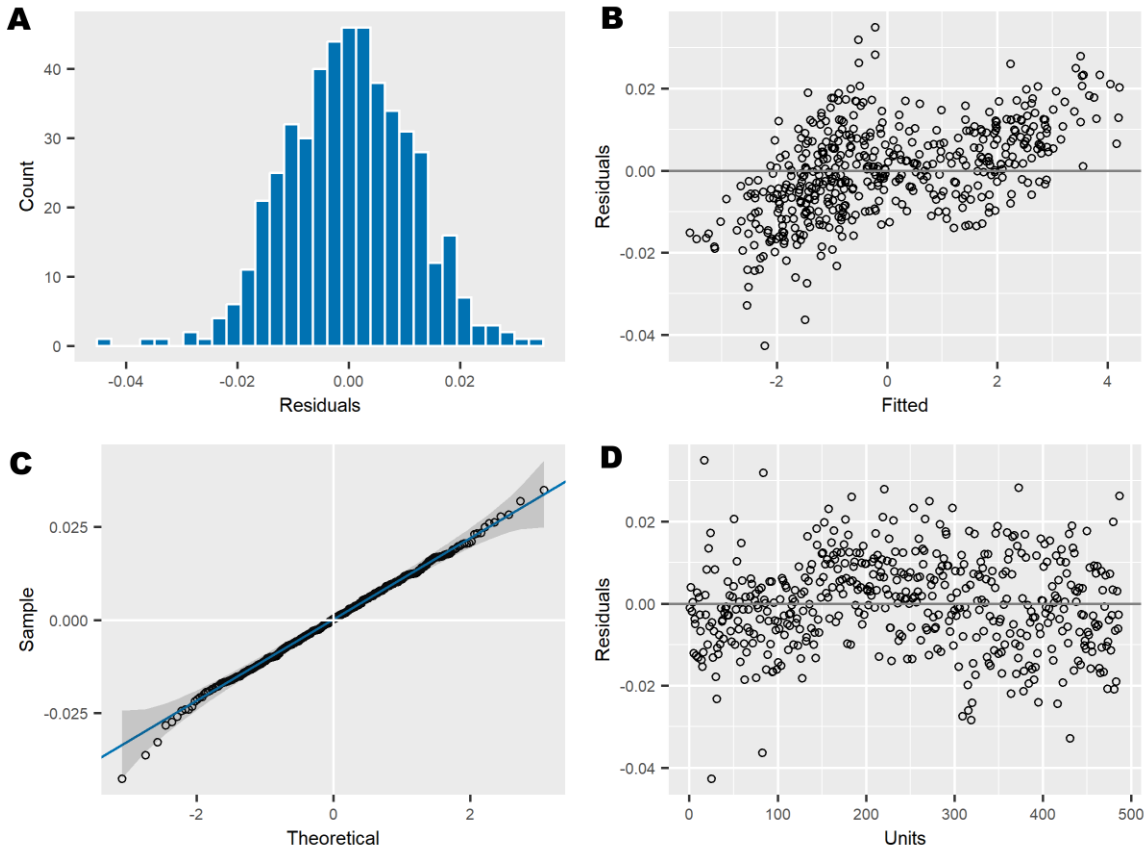
SUPPLEMENTAL FILES



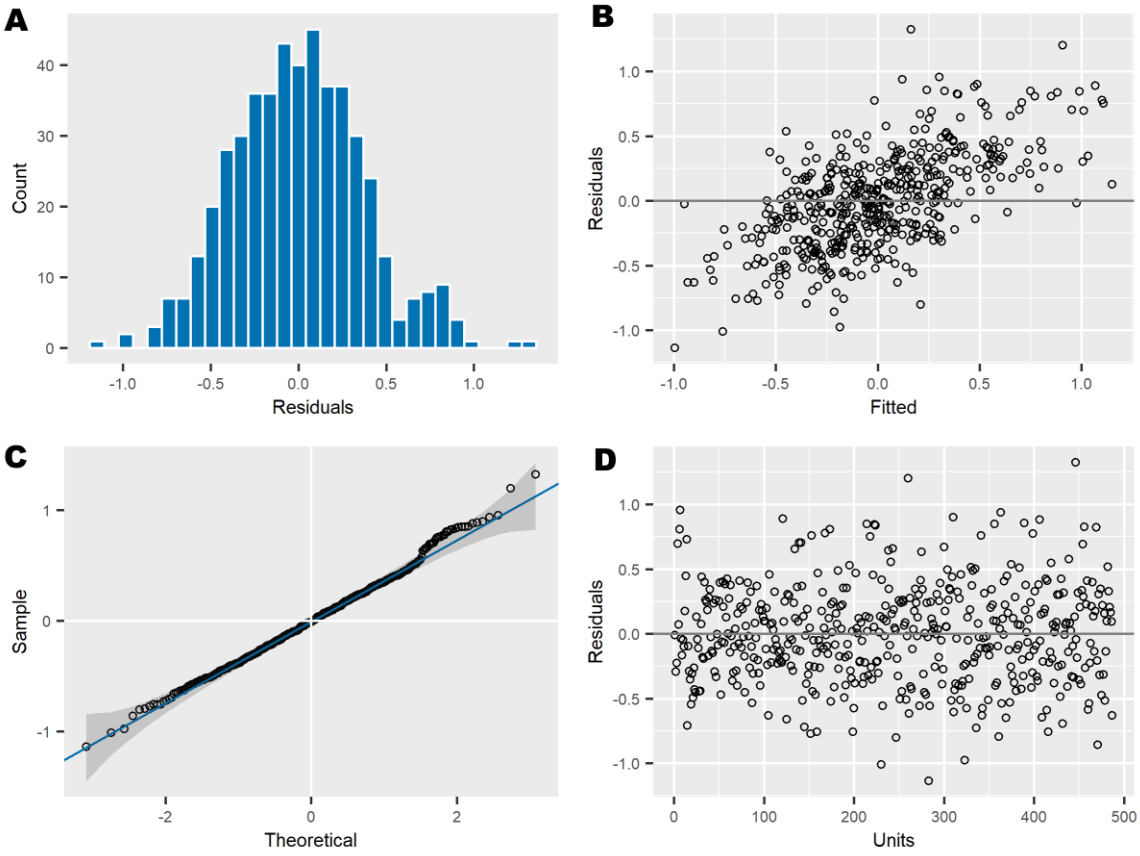
Supplemental Figure 1. Residual diagnostic plots for the ASI trait. (a) Histogram of residuals showing approximate normality. (b) Residuals vs. fitted values plot to assess linearity and homoscedasticity. (c) Q-Q plot comparing the distribution of residuals to a theoretical normal distribution, indicating minor deviations at the tails. (d) Residuals vs. observation index, showing no evident autocorrelation or time-based patterns.



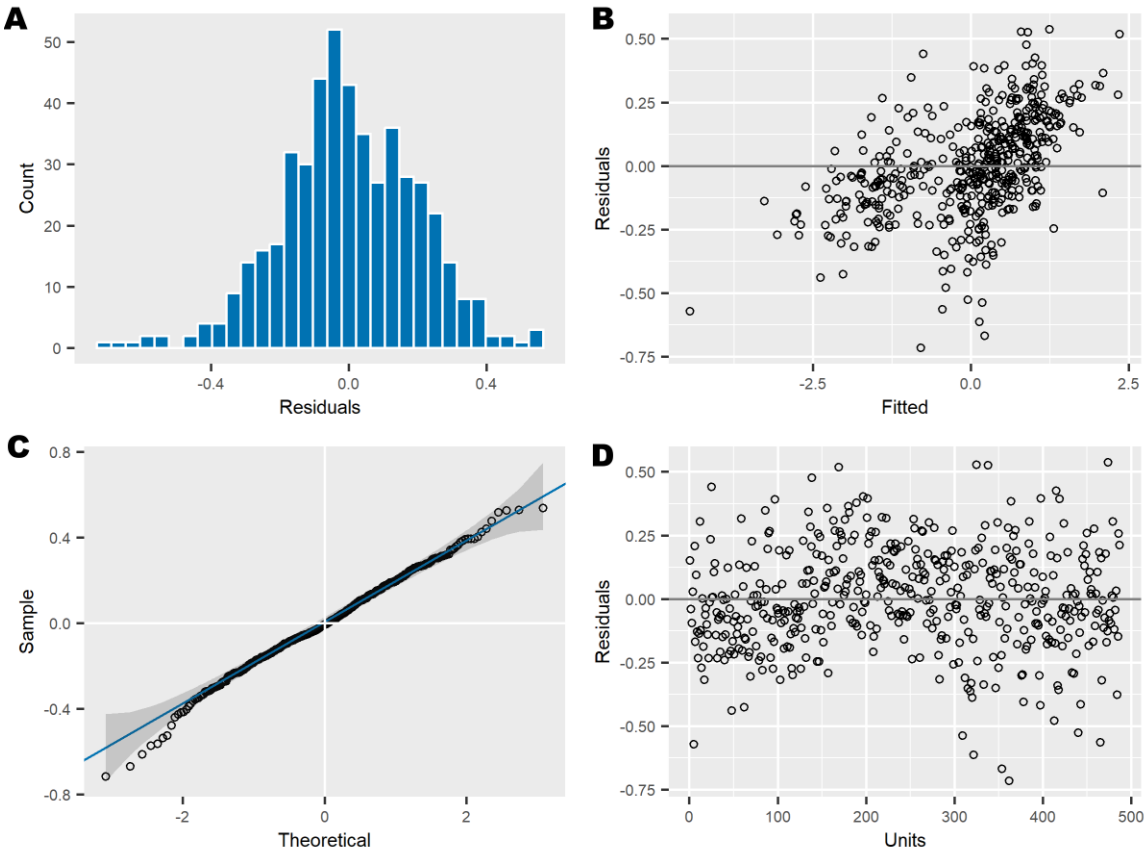
Supplemental Figure 2. Residual diagnostic plots for the PH trait. (a) Histogram of residuals showing approximate normality. (b) Residuals vs. fitted values plot to assess linearity and homoscedasticity. (c) Q-Q plot comparing the distribution of residuals to a theoretical normal distribution, indicating minor deviations at the tails. (d) Residuals vs. observation index, showing no evident autocorrelation or time-based patterns.



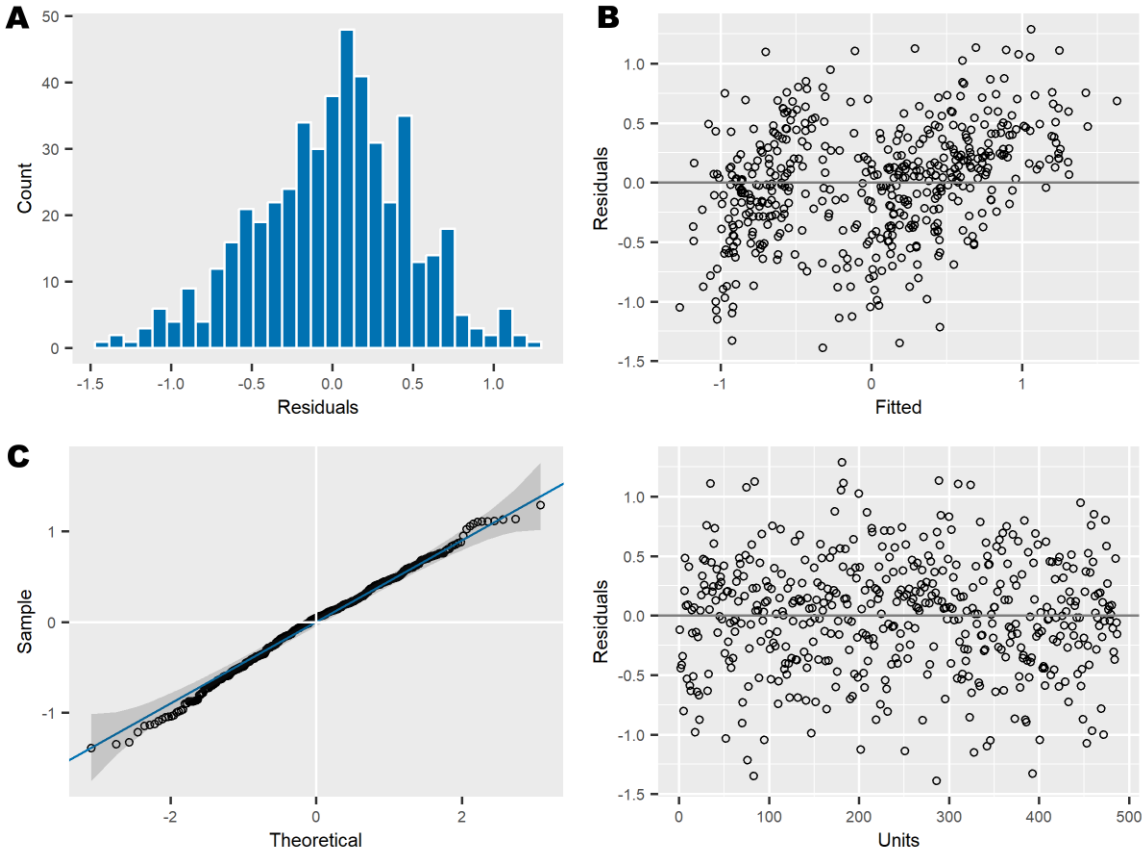
Supplemental Figure 3. Residual diagnostic plots for the EH trait. (a) Histogram of residuals showing approximate normality. (b) Residuals vs. fitted values plot to assess linearity and homoscedasticity. (c) Q-Q plot comparing the distribution of residuals to a theoretical normal distribution, indicating minor deviations at the tails. (d) Residuals vs. observation index, showing no evident autocorrelation or time-based patterns.



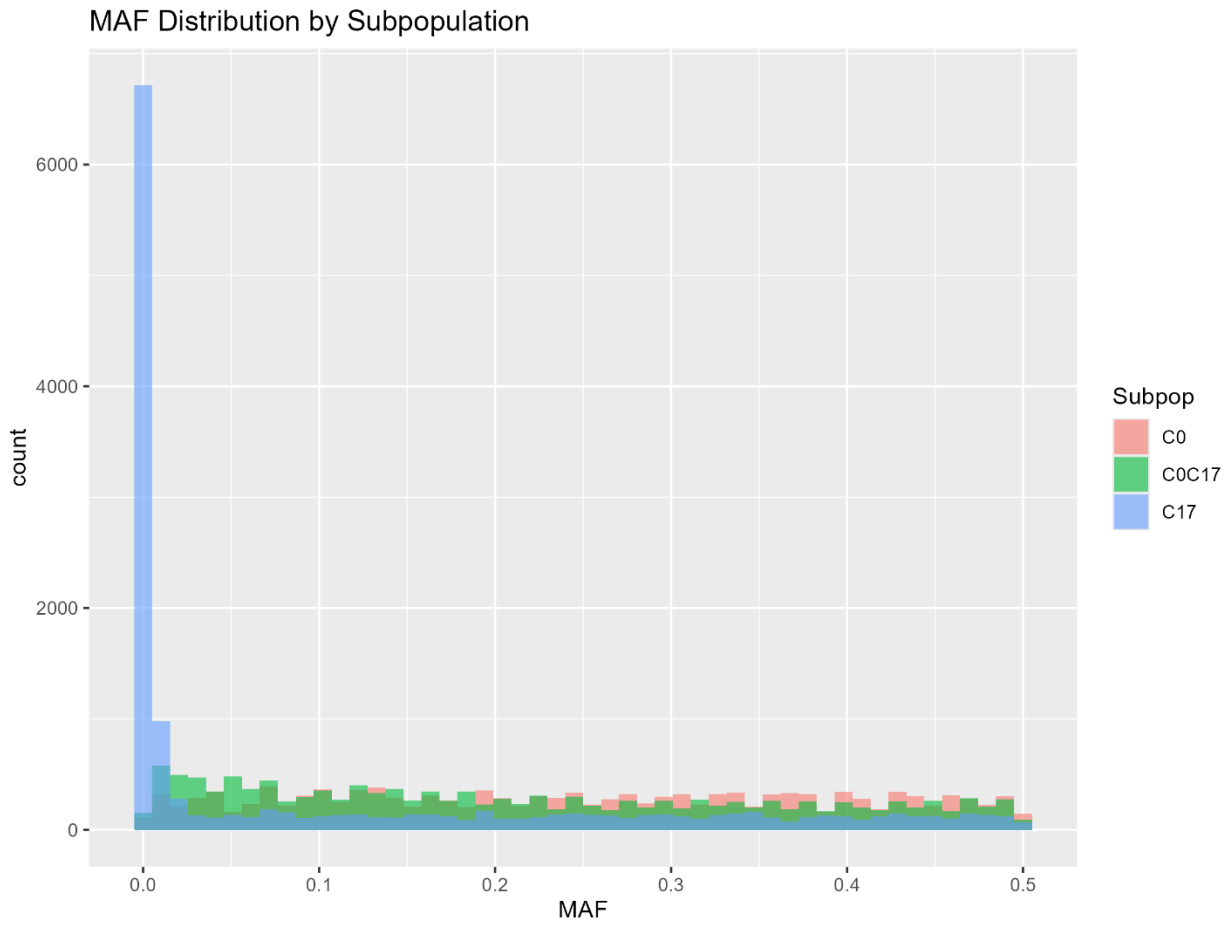
Supplemental Figure 4. Residual diagnostic plots for the FLA trait. (a) Histogram of residuals showing approximate normality. (b) Residuals vs. fitted values plot to assess linearity and homoscedasticity. (c) Q-Q plot comparing the distribution of residuals to a theoretical normal distribution, indicating minor deviations at the tails. (d) Residuals vs. observation index, showing no evident autocorrelation or time-based patterns.



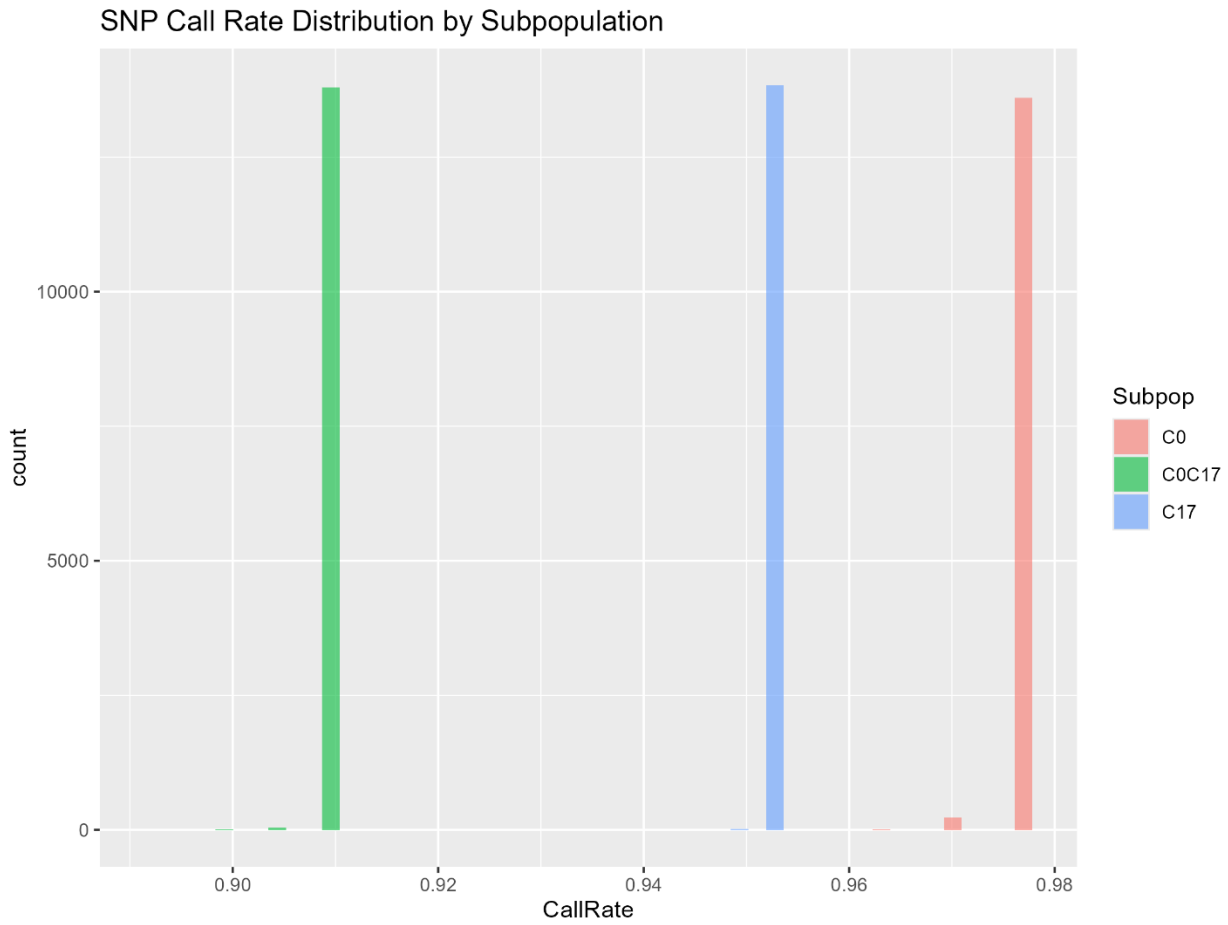
Supplemental Figure 5. Residual diagnostic plots for the NTB trait. (a) Histogram of residuals showing approximate normality. (b) Residuals vs. fitted values plot to assess linearity and homoscedasticity. (c) Q-Q plot comparing the distribution of residuals to a theoretical normal distribution, indicating minor deviations at the tails. (d) Residuals vs. observation index, showing no evident autocorrelation or time-based patterns.



Supplemental Figure 6. Residual diagnostic plots for the TR trait. (a) Histogram of residuals showing approximate normality. (b) Residuals vs. fitted values plot to assess linearity and homoscedasticity. (c) Q-Q plot comparing the distribution of residuals to a theoretical normal distribution, indicating minor deviations at the tails. (d) Residuals vs. observation index, showing no evident autocorrelation or time-based patterns.



Supplementary Figure S7. Distribution of minor allele frequencies (MAF) across subpopulations (C0, C0C17, and C17) before quality control was applied. Over 17 cycles of recurrent selection, C17 has had more time for mutations or fixed alleles to accumulate that may not have been present (or were lost) in earlier cycles like C0.



Supplementary Figure S8. SNP call rate distribution across the three subpopulations (C0, C0C17, and C17). Each bar shows the number of SNPs within a given call rate bin, colored by subpopulation. Most SNPs had high call rates (>90%) across all groups, indicating limited missing data after quality control.