Scientia Horticulturae 327 (2024) 112838

Contents lists available at ScienceDirect

Scientia Horticulturae

journal homepage: www.elsevier.com/locate/scihorti

Research Paper

Comparing different statistical models for association mapping and genomic prediction of fruit quality traits in tomato

Natakorn Prateep-Na-Thalang [a,b], Pumipat Tongyoo [a,b], Chalermpol Phumichai [c], Janejira Duangjit [b,d,*]

[a] Center for Agricultural Biotechnology, Kasetsart University, Kamphaeng Saen Campus, Nakhon Pathom 73140, Thailand
[b] Center of Excellence on Agricultural Biotechnology: (AG-BIO/MHESI), Bangkok 10900, Thailand
[c] Department of Agronomy, Faculty of Agriculture, Kasetsart University, 50 Ngam Wong Wan Road Lat Yao, Chatuchak, Bangkok 10900, Thailand
[d] Department of Horticulture, Faculty of Agriculture, Kasetsart University, Lat Yao, 50 Ngamwongwan Rd, Chatuchak, Bangkok 10900, Thailand

ARTICLE INFO

Keywords:
Association mapping
Genome-wide association analysis
Quantitative trait loci
Genomic prediction

ABSTRACT

In our study, we analyzed three key fruit quality traits in a diverse panel of 265 tomato accessions. We assessed phenotypic variations and undertook Genome-Wide Association Studies (GWAS) employing eight different statistical models. From the 19,843 single nucleotide polymorphisms (SNPs) obtained by DArTseq method, 4,253 SNPs were retained for genome-wide association studies (GWAS). Comparative analysis revealed the superior performance of the FarmCPU and BLINK models in accurately determining SNPs associated with fruit weight, locule number, and citric acid content, effectively reducing both false positives and false negatives. Other models such as MLM, CMLM, and MLMM struggled to identify relevant SNPs, and GLM was outperformed by BLINK and FarmCPU but was more effective than SUPER and FaST-LMM in reducing spurious SNP associations. Among the SNPs examined, 20 SNPs were significantly associated to the traits, accounting for 2.32 % to 22.14 % of the phenotypic variance. BLAST search yielded seven putative candidate genes, four of which were annotated in NCBI's plant protein reference database. Additionally, the application of genomic prediction with ridge regression Best Linear Unbiased Prediction (rrBLUP) yielded high accuracy for all three examined traits. These findings underscore the efficacy of genomic selection in tomato breeding, particularly for traits with high heritability, and highlight the critical role of selecting appropriate GWAS models to ensure precise association mapping.

1. Introduction

Tomato (*Solanum lycopersicum* L.) holds considerable nutritional and economic importance as a crop species within the Solanaceae family, which also includes potato, pepper, and eggplant. It is cultivated worldwide and ranks among the most widely consumed vegetables. Economically speaking, in 2020, global tomato production exceeded 187 million tons, grown across 5 million hectares (Branthôme, 2022). In Thailand, tomatoes are considered the most important vegetable as their export income amounts to about 1200 million THB. Due to its economic significance, significant efforts have been invested in enhancing horticultural traits and disease resistance through tomato breeding programs.

Tomato exhibits a wide range of genetic variations in fruit traits, such as shape, size, and weight. As a result, extensive QTL mapping has been conducted using bi-parental populations to genetically analyze fruit traits, leading to the identification of several major genes (Celik et al., 2017; Lippman et al., 2001; Muños et al., 2011; Ranc et al., 2012; Rodriguez et al., 2013; Xu et al., 2013). However, QTL detection in structured populations derived from two parents has a drawback, as it results in low mapping resolution due to limited recombination events (Holland, 2007; Pérez-de-Castro et al., 2012).

Genome-wide association study (GWAS) is a highly effective method for mapping complex traits, enabling the identification of strong linkages between markers and quantitative trait loci (QTLs) in unstructured populations, such as collections of germplasm and breeding lines. These GWAS panels exhibit higher recombination rates, leading to improved mapping resolutions compared to bi-parental populations. Moreover, these populations offer the opportunity to explore diverse alleles associated with traits of interest.

Advancements in next-generation sequencing (NGS) technology have greatly contributed to GWAS in crop species by enabling the discovery of genome-wide single nucleotide polymorphisms (SNPs). SNPs,

* Corresponding author.

https://doi.org/10.1016/j.scienta.2023.112838
Received 1 August 2023; Received in revised form 13 December 2023; Accepted 30 December 2023
Available online 13 January 2024

being the most common type of sequence variation, are well-suited for high-throughput genotyping with automation. Therefore, the first high-throughput genotyping array was developed. Furthermore, whole-genome sequencing of diverse tomato accessions across 12 chromosomes led to the identification of a large number of SNPs. For instance, resequencing 96 large-fruit commercial varieties resulted in the detection of 51,912 SNPs with a mean depth of 1.9x, contributing to the development of the Axiom tomato genotyping array (Yamamoto et al., 2016).

Besides employing NGS-based SNP discovery, researchers have developed several statistical models to enhance the accuracy and efficiency of GWAS (Genome-Wide Association Studies) (Segura et al., 2012; Yu and Buckler, 2006; Zhang et al., 2010). These advancements have facilitated successful GWAS investigations into allele variations concerning fruit quality and morphology in tomatoes. In one study, using a global collection of 96 accessions representing landraces, vintage, and modern varieties, numerous marker-trait associations were identified for phenolic compounds, ascorbic acid, β-carotene, trans-lycopene, and titratable acidity (Ruggieri et al., 2014). In another GWAS involving 163 tomato accessions, 44 candidate loci associated with 19 fruit metabolites, including amino acids, sucrose, malate, ascorbate, and citrate, were discovered (Sauvage et al., 2014). Furthermore, extensive genetic dissection of tomato flavor resulted in the identification of numerous significant associations with flavor-related traits (Tieman et al., 2017; Zhang et al., 2015; Zhao et al., 2019). For fruit morphological traits, GWAS studies in tomato collections uncovered several favorable alleles (Lin et al., 2014; Sacco et al., 2015). A recent investigation examined the genetic variations in 192 tomato accessions from a core collection for six fruit traits and identified a total of 54 loci associated with these traits (Phan et al., 2019).

Moreover, a germplasm collection consisting of a broad-based population (163 accessions) was employed for GWAS to identify genomic regions associated with fruit, flower, and vegetative traits, leading to the discovery of 107 marker-trait associations for eight quantitative traits, including fruit weight and locule number (Mata-Nicolás et al., 2020). Additionally, The discovery of SNPs has significantly contributed to the successful implementation of genomic prediction in tomatoes, specifically for assessing fruit quality and morphology (Cappetta et al., 2020; Duangjit et al., 2016; Yamamoto et al., 2016)

In this study, we assessed the effectiveness of GWAS and the precision of genomic prediction by examining a population consisting of 359 accessions. Our main goal was to compared eight alternative association mapping statistical models, spanning from single to multi-locus, in order to identify candidate genomic regions associated with fruit quality traits through GWAS.

Then, with the same genotypic and phenotypic data, we applied the GS model called ridge regression best linear unbiased prediction (rrBLUP) combined with cross-validation approach to explore the predictabilities of genomic selection models in traits of interest.

2. Methodology

2.1. Genetic materials and DNA extraction

At the time of this study, the Tropical Vegetable Research Center has collected 531 accessions of tomatoes, 359 of these accessions were selected for genotyping, and their genomic DNA was extracted with a modified CTAB method based on a procedure by Fulton et al. (1995). For each accession, the DNA was isolated from three-week-old tomatoes. The isolated DNA samples were quantified with NanoDrop 2000 spectrophotometer and diluted to 100 ng/μL in TE buffer. The quality was evaluated by performing agarose gel electrophoresis on undigested DNA samples and those digested with *EcoR*I restriction enzyme.

2.2. High-throughput genotyping and SNP filtering

The isolated DNA was sent to DArT Pty Ltd (Canberra, ACT, Australia) and genotyped with the DArTseq analysis. To produce genomic libraries with reduced complexity, the DNA samples were first digested with *Pst*I and *Mse*I restriction enzymes. The digested DNA fragments were then ligated with restriction site-specific adapters and enriched on Illumina flow cell. The sequence analysis was performed with Illumnina Hiseq 2000 (Illumina Inc., San Diego, CA, USA). DArTsoft software was used to align and detect SNPs between the sequencing reads and output the genotypic data as scores. For this study, SNP markers with missing data >20 % and minor allele frequency (MAF) <0.01 % were removed to filter out low quality and nearly monomorphic markers from subsequent analyses.

2.3. Phenotyping

From 359 of these accessions, two hundred and sixty-five tomato accessions from the Tropical Vegetable Research Center grew and were included in this study. Accessions were planted in a randomized complete block design with six replications at the Faculty of Agriculture at Kamphaeng Saen, Kasetsart University in the summer of 2018 and 2020. In each tomato plant, at least ten independent fruits were collected and measured for fruit weight, locule number, citric acid. This is performing to make sure that the total number of phenotypic observations per genotype reflects the complexity of the genetic architecture and the heritability of the trait. Phenotypic means for each trait in every accession were calculated using the least-squares method. This was achieved by applying a linear mixed-effects model of the given format. $Y_i = m + genotype_i + year_i + block(year)_{jk} + e_{ijk}$, where Y_i is the phenotype for accession i, m is the grand mean, $genotype_i$ is the fixed effect (genotypic value) of accession i, $year_j$ and $block(year)_{jk}$ are the random effects of year j and block (nested within year) k, and e_{ijk} is the error. It was assumed that random effects were independent and had identical distributions. The model was fitted using restricted maximum likelihood (REML) through the R package lme4 (Bates et al., 2014).

2.4. GWAS analysis

GenoGenome-wide association study (GWAS) was performed using genomic association and prediction integration tool, GAPIT version 3, (Wang and Zhang, 2021) in R (R Core Team, 2023). The eight association mapping models that were evaluated ranged in complexity from simple to complex and included

(i) general linear model (GLM) with PCA (principle component analysis) (Price et al., 2006)
(ii) mixed linear model (MLM) with PCA + *K* (Kinship matrix for family relatedness estimates) (Yu et al., 2006)
(iii) compressed MLM (Zhang et al., 2010)
(iv) settlement of MLMs under progressively exclusive relationship (SUPER) (Wang et al., 2014b)
(v) multi-locus test methods, including multiple loci mixed model (MLMM) (Segura et al., 2012)
(vi) factored spectrally transformed linear mixed models (FaST-LMM) (Lippert et al., 2011)
(vii) fixed and random model circulating probability unification (Farm CPU) (Liu et al., 2016)
(viii) Bayesian-information and linkage-disequilibrium iteratively nested keyway (BLINK) (Huang et al., 2019)

2.5. Model evaluation

Q-Q plots are commonly used to assess false positives and false negatives in models (Kristensen et al., 2018; Riedelsheimer et al., 2012 Stich and Melchinger, 2009; Würschum et al., 2012). These plots

compare the expected and observed negative-log association probabilities. A straight line close to the 1:1 line indicates a uniform distribution, supporting the null hypothesis of no significant association. Deviations from the straight line indicate the presence of significant associations.

A lack of a straight line and a tail suggests false positives (inflating line) and false negatives (deflating line). Conversely, a straight line near 1:1 with an upward deviated tail indicates controlled false positives and false negatives, with true associations present. While most P-values follow a uniform distribution, a few in linkage disequilibrium with a causal polymorphism produce significant values in the "tail" of the plot.

2.6. Gene ontology mapping and annotation of identified SNPs

For the significant SNPs, allele trimmed sequences from DarTSeq were extended 50 bp upstream and downstream using tomato reference genome SL3.0. The nucleotide sequences are then used as a query, which is first translated in all six reading frames in BlastX using all plant proteins from NCBI as a reference. Blasted sequences were imported into OmicsBox® (Biobam Bioinformatics S.L.) for gene ontology mapping and gene ontology annotation.

2.7. Genomic selection analysis

Genetic prediction was assessed by comparing GEBV with the actual phenotyping values. From 265 accessions that phenotype information is available. Seventy-five percent of the total number of accessions were used to fit the model, in the training validation set. The remaining accessions were used as validation set. In this study, the genotype information from the training set was utilized to fit a model denoted as $Y = 1\mu + Xg + e$, where Y represents a vector of observed phenotypic values, μ denotes the overall mean of the training set, fitted as covariates based on the genotype, and e accounts for the residual effect. Subsequently, the SNP effects were calculated using the mixed-model solver function (mixed.solve). Using the genotype information and SNP effects, the Genomic Estimated Breeding Values (GEBV) of the validation set were predicted. To achieve precision, researchers employed the ridge regression best linear unbiased prediction (rrBLUP) statistical model, as introduced by Endelman (2011). This model, which is akin to the best linear unbiased prediction (BLUP), offers rapid processing due to its utilization of a mixed model algorithm with a single variance component, apart from the residual error, as described by Kang et al. (2008). Additionally, this model is conveniently accessible as an R package, accessible at http://cran.r-project.org/web/packages/rrBLUP/index.html. Prediction accuracies were tested by the correlation between predicted breeding values and evaluated phenotype (true breeding values) for 1000 iterations.

3. Results

3.1. Phenotypic variations of fruit traits in the tomato collection

The collection of 265 tomato accessions demonstrated significant phenotypic diversity for three evaluated fruit characteristics, namely, fruit weight, number of locules, and citric acid content, as illustrated in Table 1 and Fig. 1. The fruit weight exhibited a substantial range, varying from 5.49 g to a maximum of 93.33 g, with an average of 20.89 g across the accessions. The number of locules per fruit also displayed variability, with a range of 0.31 to 0.87. As for the citric acid content, it spanned from 0.25 % to 0.68 %, averaging at 0.42 %. A PCA plot, derived from genomic relationship matrices, distinguished the cultivars into three distinct clusters as depicted in PCA plots (Figs. 2A, 2C) and scree plot (Fig. 2D). Collectively, these results demonstrate the considerable diversity encapsulated within this tomato germplasm for traits associated with fruit quality. Heritability estimates for these traits showed a broad sense heritability of 61 % to 80 %, as detailed in Table 1.

3.2. GWAS model comparison of 3 fruit quality traits in tomato

A total of 19,843 SNPs were initially obtained from the DarTseq analysis. After filtering with the criteria of PIC $\geq$ 0.1 and call rate $\geq$80 %, 4253 SNPs were selected for further genetic diversity and GWAS of three traits in this study. A comparison made between eight distinct association mapping models, varying in complexity from simple to complex. These models identified varying numbers of significant SNP markers at the same significance threshold (Fig. 3). We set the significance threshold at $-\text{Log10 (P)} > 4.93$, applying a Bonferroni adjustment, to gain significant association SNPs. For fruit weight, we identified 10 SNPs from GLM, 73 SNPs from SUPER, 73 SNPs from FaST-LMM, 3 SNPs from FarmCPU, and 3 SNPs from BLINK; and no association was identified by MLM, CMLM, and MLMM models (Table 2).

In terms of locule number, GLM detected the largest number of SNPs, totaling 31. SUPER and FaST-LMM, on the other hand, identified 18 SNPs each. FarmCPU and BLINK, in contrast, revealed 6 and 7 associated SNPs, respectively. Notably, five of these SNPs were detected by both FarmCPU and BLINK, indicating an overlap in their findings (Tables 2 and 3).

In total, the eight models jointly detected 43 SNPs associated with citric acid content in tomatoes. Specifically, GLM pinpointed 5 SNPs, while both SUPER and FaST-LMM independently revealed 30 SNPs each. FarmCPU and BLINK identified four SNPs each, with SNP_06051 on chromosome 3 being a notable standout, as it was identified by four of the models (GLM, SUPER, FarmCPU, and BLINK).

We observed both SUPER and FaST-LMM identified multiple significant SNP marker associations in close physical proximity on the chromosome. These substantial peaks resulted from one SNP within each peak having the highest significant association with the traits, while the other markers in the same peak were in high linkage disequilibrium (LD) with this most significant marker.

3.3. Identified SNPs associated with fruit quality traits in tomato

For the fruit weight and locule number, SUPER and FasT-LMM models had a large number of false positives as indicated by substantial inflation of p-values (Fig. 3). Q-Q plots of MLM, CMLM, MLMM models had a straight line with a slightly deviated tail, but associated SNPs were not identified. In contrast, the FarmCPU and BLINK models followed a straight line close to 1:1, with a sharp upward deviated tail (Fig. 3A, B). For the percentage of citric acid, the FarmCPU showed a uniform distribution close to the 1:1 line. However, slightly inflated straight line was observed with upward deviated tail in BLINK model (Fig. 3C).

It is noteworthy that despite the numerous significant SNPs identified by the eight GWAS models, only those with at least moderate deflation or inflation of Q-Q plots were considered as genuine associated SNPs, as presented in Table 3.

Manhattan plots for the three traits show varied association strengths

Table 1
Descriptive statistics of tomato phenotypic records.

Traits (transformation)	Unit	heritability	skewness	Range	Mean (SD)	Min	Max
Fruit weight (log10)	g	0.614	0.006	1.226	1.32 (0.25)	0.74	1.97
Locule number (NA)	locule	0.795	1.135	0.558	0.46 (0.13)	0.31	0.87
Citric acid (NA)	%	0.612	0.747	0.427	0.42 (0.08)	0.25	0.68

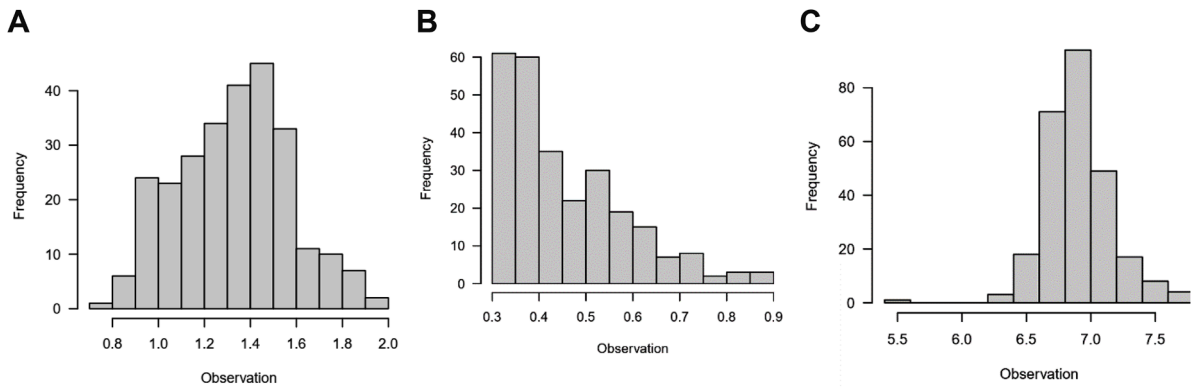

Fig. 1. Phenotypic distribution of three fruit traits in the 265 tomato accessions over three years. Fruit weight (log10) (A) Locule number (B) Citric acid (C).

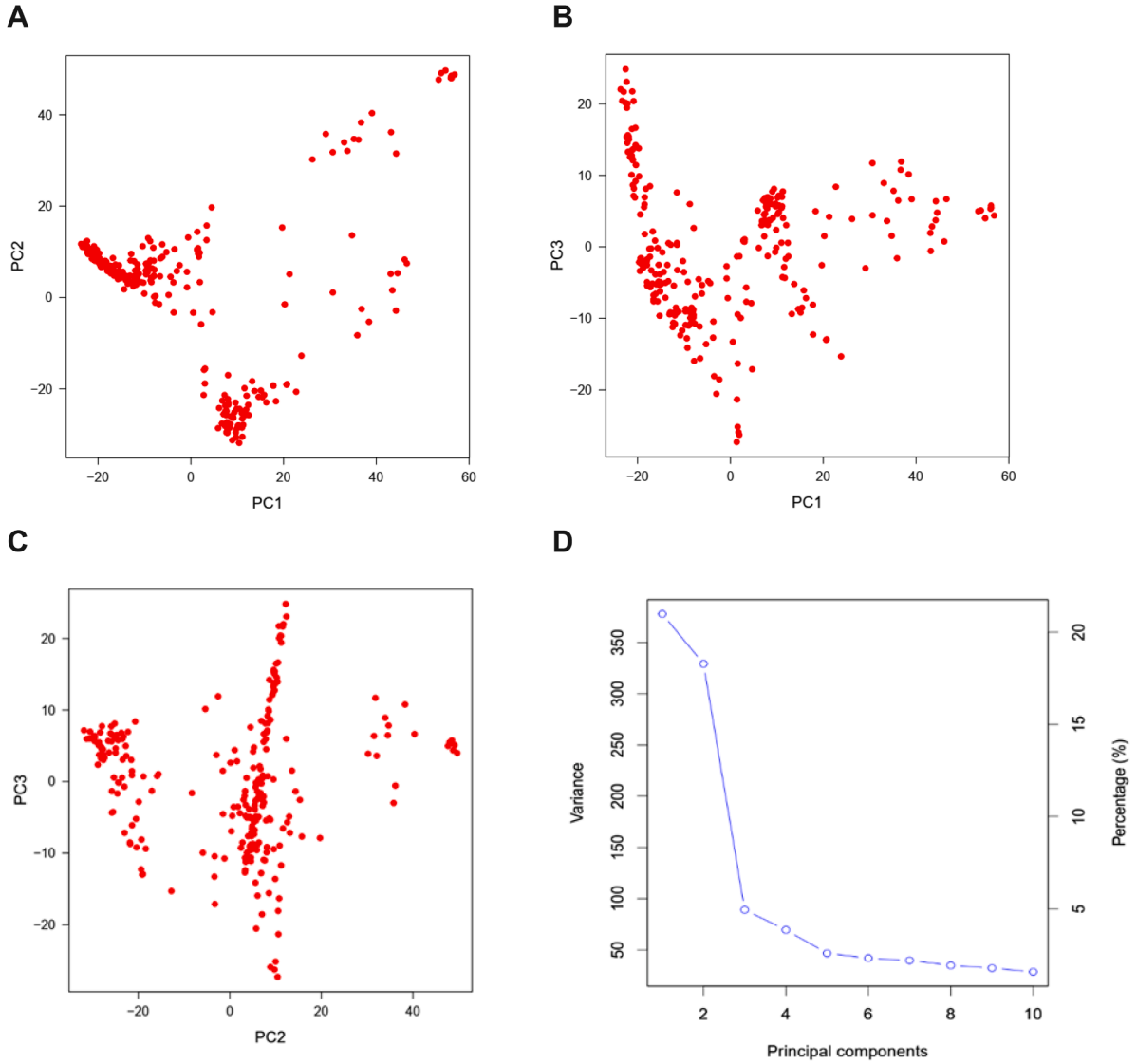

Fig. 2. The PCA plot based on genomic relationship matrices showed that all cultivars were composed of three clusters. Plot between PC1 and PC2 (A), plot between PC1 and PC3 (B), plot between PC2 and PC3 (C), and scree plot representing the proportion of variance accounted for by the principal components (D).

across models, as depicted in Figs. 4 to 6. Notably, chromosome 7 harbors the most significant SNPs for tomato fruit weight, as indicated in Fig. 4. While MLMM analysis's Q-Q plot suggests a generally expected distribution of P-values, there is a slight deviation at the tail end within the 95 % confidence band. No genome-wide significant SNPs for fruit weight were detected (Figs. 3 and 4); however, FarmCPU and BLINK pinpointed a highly significant marker at the same locus on chromosome 7 (SNP_11,268), with attributing 22.14 % (BLINK) and 21.72 %

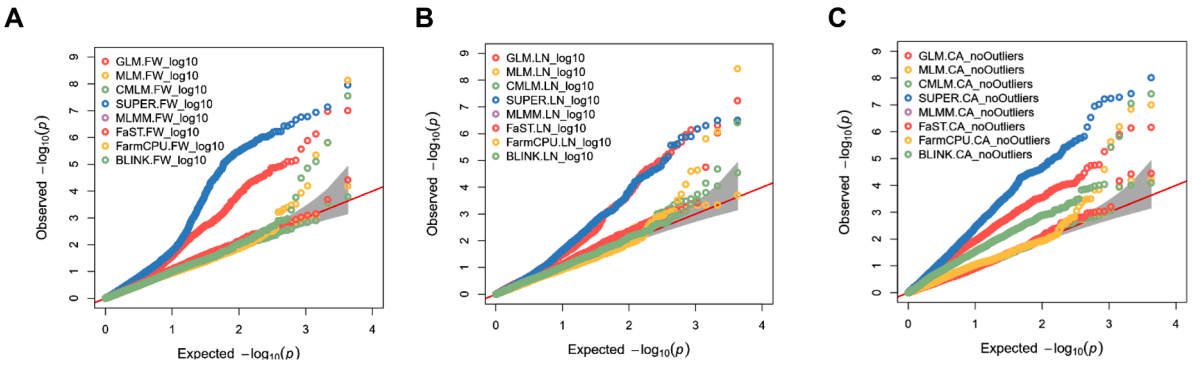

Fig. 3. Quantile-quantile (Q-Q) plots of the eight models including General Linear Model (GLM), Mixed Linear Model (MLM), Compressed MLM (CMLM), multi-locus test methods, including Multiple Loci Mixed Model (MLMM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), Factored Spectrally Transformed Linear Mixed Models (FaST-LMM), Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and linkage-disequilibrium iteratively nested keyway (BLINK) for three fruit quality traits Fruit weight (A), Locule number (B), Citric acid (C). The grey area represents the 95 % concentration band.

Table 2
Comparison of the number of significant markers ($P \leq 0.05$) identified by Bonferroni adjustment.

Models	Fruit weight	Locule number	Citric acid
GLM	10	31	5
MLM	0	0	0
CMLM	0	0	0
MLMM	0	0	0
SUPER	73	18	30
FaST-LMM	73	18	30
FarmCPU	3	6	4
BLINK	3	7	4
total	89	80	43

(FarmCPU) of the phenotypic variance to an allelic effect of −0.089 in both models. Associated markers at chromosomes 3, 4, and 5 were identified in either BLINK model (explaining 3.05 %–8.37 %) or FarmCPU model (explaining 5.93 %–7.19 %) of the phenotypic variation (Table 3 and Fig. 4).

For locule number, FarmCPU and BLINK identified 6 and 7 significant SNPs, respectively, with five concordant markers across chromosomes 1, 2, 3, 7, and 12 (Fig. 5, Table 3). The most significant SNPs, based on BLINK and FarmCPU models, were located on chromosomes 7 (SNP_10,605) and 12 (SNP_17,733). According to BLINK, they account for 8.85 % and 8.56 % of the variance, with allelic effects of −0.0373 and −0.063, respectively. The remaining overlapping SNP markers for locule number collectively explain 3.31 %–6.15 % of the phenotypic variation.

In the context of citric acid content, 43 SNPs demonstrated significance, although only a subset of them is considered robust. The MLMM and FaST-LMM models exhibited the most consistent Q-Q plot uniformity compared to the previous traits (Fig. 6). Notably, FaST-LMM did not identify any SNPs, while one of MLMM's SNPs approached significance with -Log10(P) = 4.91, slightly beneath the cutoff threshold of > 4.93.

One noteworthy SNP, SNP_06051 on chromosome 3, stood out in these models. Although it did not reach significance in other models, it consistently appeared across FarmCPU, and BLINK models, surpassing the significance threshold (Table 3). This particular SNP accounts for 5.82 % of the phenotypic variance, with an allelic effect of 0.026 (Fig. 6). Additionally, the most significant marker associated with citric acid content was identified by BLINK as SNP_01844 on chromosome 1, attributing 20.88 % of the phenotypic variance, with an allelic effect of −0.052.

3.4. Gene ontology mapping and annotation of identified SNPs

Out of the 20 associated SNPs, BlastX analysis successfully identified

Table 3
Genome-wide significant markers.

Trait	Marker	Chromosome	Position	P-value (the lowest)	Phenotypic Variance Explained (%)	Detected model
Fruit weight	SNP_05412	3	61,922,269	7.91E-06	3.05	BLINK
	SNP_07365	4	62,847,790	4.57E-06	5.93	FarmCPU
	SNP_08107	5	28,053,514	1.55E-06	8.37	BLINK
	SNP_08448	5	61,914,120	1.58E-06	7.19	FarmCPU
	SNP_11,268	7	58,111,867	2.82E-08 (BLINK)	22.14	FarmCPU, BLINK
Locule number	SNP_00346	1	4,090,397	4.35E-06 (BLINK)	3.31	FarmCPU, BLINK
	SNP_01821	1	93,785,335	2.03E-07	2.31	BLINK
	SNP_03545	2	46,861,543	5.1E-07	4.32	BLINK
	SNP_03723	2	48,542,528	1.29E-12 (BLINK)	6.54	FarmCPU, BLINK
	SNP_05874	3	68,318,807	8.63E-11 (BLINK)	6.15	FarmCPU, BLINK
	SNP_10,605	7	2,131,386	5.25E-06 (BLINK)	8.85	FarmCPU, BLINK
	SNP_11,754	7	66,929,011	2.49E-07	2.90	FarmCPU
	SNP_17,733	12	6,026,885	7.53E-09 (BLINK)	8.56	FarmCPU, BLINK
Citric acid	SNP_01844	1	94,036,985	3.88E-08 (BLINK)	20.88	BLINK
	SNP_05531	3	63,593,159	3.77E-06	7.07	BLINK
	SNP_06051	3	71,566,916	1.45E-07 (FarmCPU)	7.52	FarmCPU, BLINK
	SNP_13,815	9	38,413,382	1.50E-06	13.24	BLINK
	SNP_13,946	9	52,459,357	2.38E-06 (FarmCPU)	11.13	FarmCPU
	SNP_15,394	10	36,814,495	6.54E-07 (FarmCPU)	4.11	FarmCPU
	SNP_16,862	11	53,240,816	9.99E-08 (FarmCPU)	15.45	FarmCPU

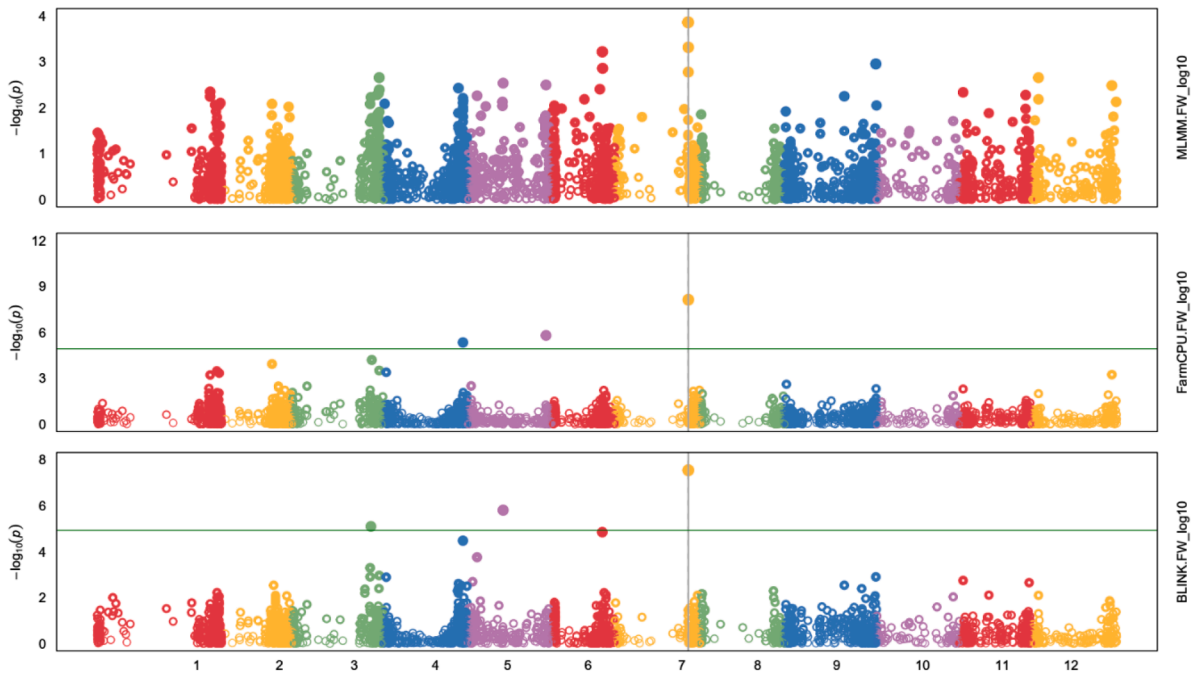

Fig. 4. Manhattan plots of −Log10 (P) vs. chromosomal position of SNP markers associated with qualitative traits, Fruit weight in tomato from three models including, Multiple Loci Mixed linear Model (MLMM), and Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and linkage-disequilibrium iteratively nested keyway (BLINK).

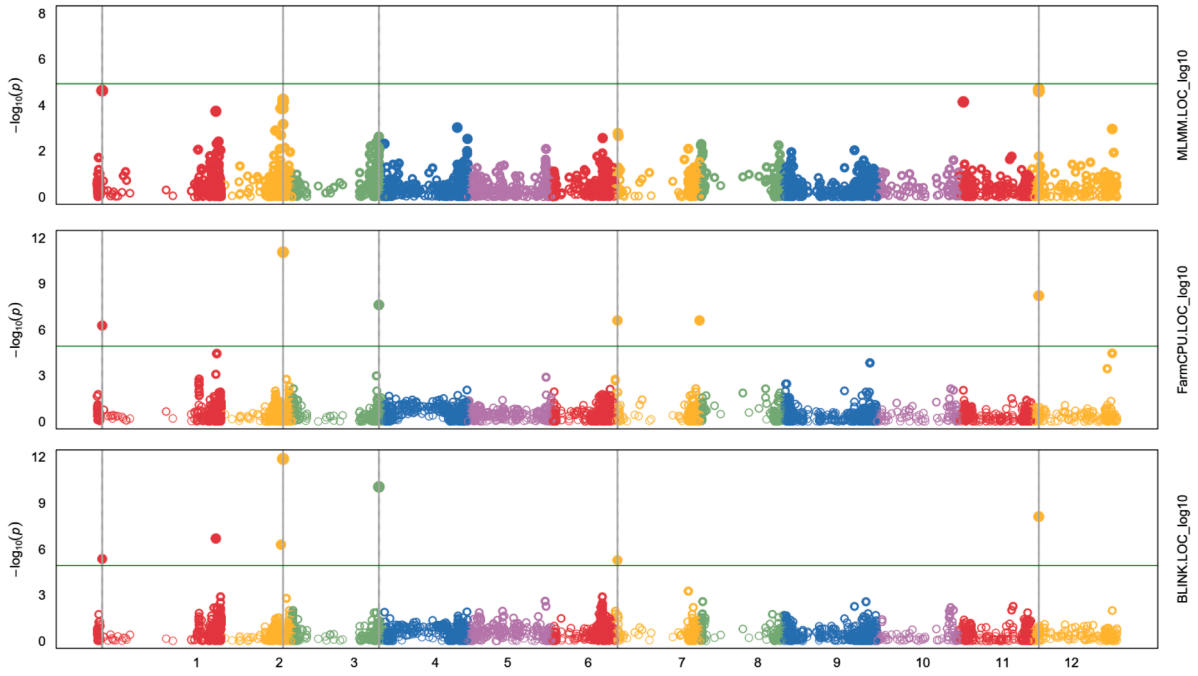

Fig. 5. Manhattan plots of −Log10 (P) vs. chromosomal position of SNP markers associated with qualitative traits, Locule number in tomato from three models including, Multiple Loci Mixed linear Model (MLMM), and Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and linkage-disequilibrium iteratively nested keyway (BLINK).

7 SNPs associated with candidate genes.Subsequently, we utilized OmicsBox® to map and was able to annotate 4 SNPs from 2 phenotypes (SNP_03723 on chromosome 2 and SNP_01821 on chromosome 1 for locule number; SNP_13,946 on chromosome 9 and SNP_16,862 on chromosome 11 for citric acid content). The GO annotations provided offer insight into the biological processes, molecular functions, and cellular components associated with the SNPs (Table 4).

SNP_01844, which exhibited one of the highest percentages of phenotypic variance explained (% PVE) in citric acid levels, was subjected to a BLASTx search. The analysis revealed that this SNP is situated within the coding region of a putative agamous-like MADS-box protein, specifically an AGL21-like isoform X1. However, no additional annotations were found upon further annotation with NCBI's plant protein reference database.

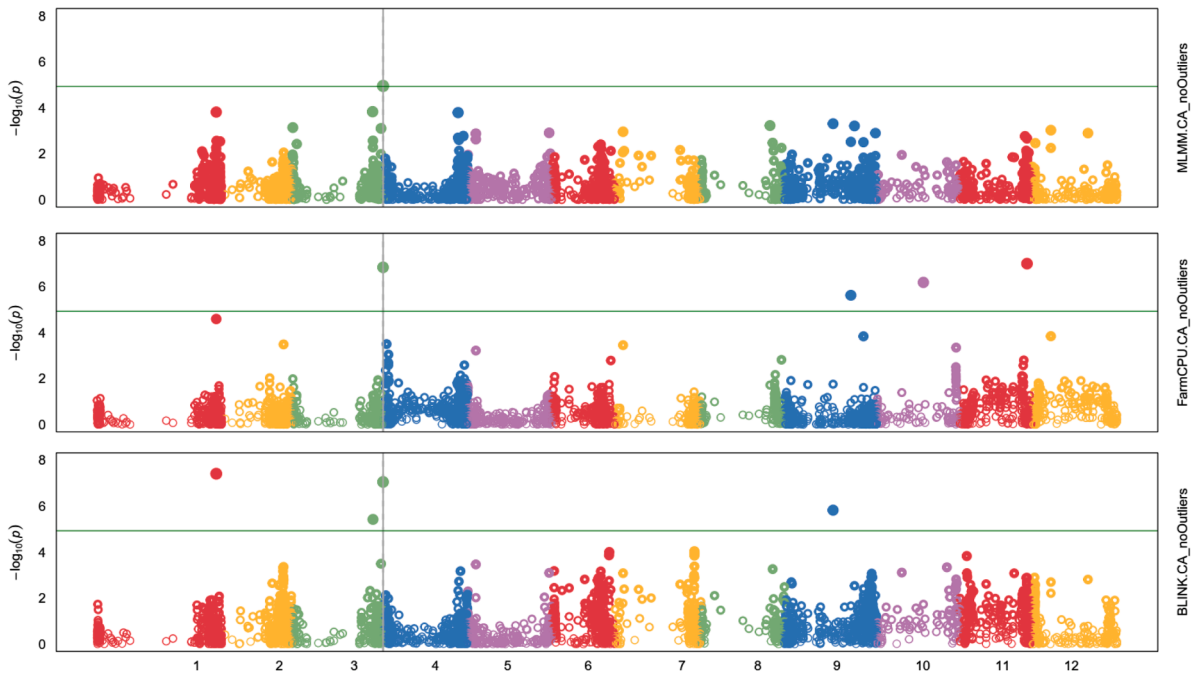

Fig. 6. Manhattan plots of −Log10 (P) vs. chromosomal position of SNP markers associated with qualitative traits, citric acid, in tomato from three models including, Multiple Loci Mixed linear Model (MLMM), and Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and linkage-disequilibrium iteratively nested keyway (BLINK).

Table 4
Significant SNPs (-log10(P) ≥ 4.93), candidate genes, and GO annotation of associated with fruit quality related traits in tomato.

Traits	SNPs	Chro	Position	BlastX description	#GO	GO IDs	GO names
Fruit weight	SNP_08448	5	61,914,120	putative protein TPRXL	–	–	–
Locule number	SNP_00346	1	4,090,397	putative GEM-like protein 5-like	–	–	–
	SNP_03723	2	48,542,528	uncharacterized protein LOC101268823	2	C:GO:0,016,020; C:GO:0,016,021	C:membrane; C:integral component of membrane
	SNP_01821	1	93,785,335	(-)-camphene/tricyclene synthase, chloroplastic		P:GO:0,016,114; F:GO:0,000,287; F:GO:0,050,551; F:GO:0,050,552; F:GO:0,102,701; F:GO:0,102,703; C:GO:0,009,507	P:terpenoid biosynthetic process; F:magnesium ion binding; F:myrcene synthase activity; F: (4S)-limonene synthase activity; F:tricyclene synthase activity; F:camphene synthase activity; C:chloroplast
Citric acid	SNP_01844	1	94,036,985	putative agamous-like MADS-box protein AGL21-like isoform X1	–	–	–
	SNP_13,946	9	52,459,357	N-terminal acetyltransferase B complex auxiliary subunit NAA25	3	P:GO:0,017,196; F:GO:0,004,596; C:GO:0,031,416	P:N-terminal peptidyl-methionine acetylation; F:peptide alpha-N-acetyltransferase activity; C:NatB complex
	SNP_16,862	11	53,240,816	alcohol-forming fatty acyl-CoA reductase-like	6	P:GO:0,006,629; P:GO:0,010,345; P:GO:0,035,336; F:GO:0,080,019; F:GO:0,102,965; C:GO:0,043,231	P:lipid metabolic process; P:suberin biosynthetic process; P:long-chain fatty-acyl-CoA metabolic process; F:fatty-acyl-CoA reductase (alcohol-forming) activity; F:alcohol-forming fatty acyl-CoA reductase activity; C:intracellular membrane-bounded organelle

3.5. Genomic prediction using rrBLUP

When predicting trait outcomes using a training set comprising 75 % of the population (199 accessions) with 4253 markers, we observed variability in accuracy among different traits (as detailed in Table 5, which presents the average, standard deviation, median, minimum, and maximum correlation values between observed phenotypes and predictions). Fruit weight predictions were the most accurate, with a maximum accuracy value of 0.851 (0.851± 0.040), while citric acid content predictions were less accurate, with a minimum accuracy value of 0.633± 0.081). The heritability for these traits was recorded at 0.614 for fruit weight and 0.612 for citric acid content, respectively.

4. Discussion

Genome-Wide Association Studies (GWAS) have markedly enhanced the analysis of tomato fruit quality traits. The selected statistical model heavily dictates the results. For instance, deviations in P-value distribution seen in FaST-LMM and SUPER, as evidenced in Q-Q plots (Fig. 3), corroborate with earlier findings (Kaler et al., 2020), indicating these models are inappropriate for association mapping for the studied traits because they generate spurious marker-trait associations, potentially due to suboptimal correction for complex population structure or stratification (Liu et al., 2016).

False associations may also arise from varying degrees in kinship between individual pairs. This confounding effect can be attenuated by implementing a general linear model (GLM) that quantifies the proportion of genes that are identical by descent for each pair of individuals. Selective exclusion of individuals who exhibit close genetic relatedness can further refine the model, thereby reducing the potential for bias in association estimates (Voight et al., 2005; Li et al., 2014). This is why we observed a presence of moderate rates of spurious associations (Table 2).

In the Q-Q plots for the MLM and CMLM, a pattern was noted that largely followed the expected diagonal but with tails that diverged slightly. This pattern suggests a decrease in false positives with an accompanying increase in false negatives, given that the majority of significant markers clustered near the expected trendline. Such an increase in false negatives could be due to overfitting, wherein the variance attributed to random effects is overestimated, thereby obscuring true genetic associations (Kaler et al., 2020; Zhang et al., 2010). Consistent findings have been reported across various investigations (Li et al., 2018; Tamba et al., 2017; Wen et al., 2015; 2018), indicating a tendency for more sophisticated models to yield an increased incidence of false negatives.

In contrast, the Q-Q plots of the FarmCPU and BLINK, advanced multi-locus GWAS models, demonstrated regulation of both false positives and negatives, as shown by the plots that largely adhered to the expected diagonal with prominent deviations at the tails, with the exception noted for BLINK in citric acid dataset (Fig. 3). However, when it came to determining statistical significance in association mapping, FarmCPU took a notably conservative approach, resulting in lower -log10(P-values) compared to those obtained with BLINK (see Figs. 4–6). Moreover, FarmCPU exhibited the capability to identify specific markers that eluded detection by BLINK, reflecting both FarmCPU's conservative nature and its precision in marker identification. Simultaneously, BLINK was able to uncover certain associated SNPs that were not detected by other methods. These outcomes are consistent with the observations made in the study by Gai et al. (2023) and imply that disparities in statistical power between these models may underlie this variability.

Wang and Zhang (2021) indicate that multi-locus models such as FarmCPU and BLINK outperform less complex models, and BLINK surpasses FarmCPU. FarmCPU, in particular, reduces the risk of overfitting by implementing dual adjustments when assessing markers, according to Liu et al. (2016). Initially, it accounts for population structure, family relatedness, and pseudo-quantitative trait nucleotides. subsequently, it refines the derivation of family relatedness across all markers or selectively factors in or out pseudo-quantitative trait nucleotides, depending on their association with the markers being tested (Liu et al., 2016). BLINK has advanced beyond FarmCPU by discarding the premise that quantitative trait nucleotides (QTNs), must be uniformly spread across the genome, opting instead to employ information on linkage disequilibrium. This advancement allows BLINK to use a more streamlined Fixed Effect Model (FEM) rather than the more resource-intensive Random Effect Model (REM), significantly cutting down on analysis time. Studies have confirmed BLINK's enhanced efficiency in analyzing vast datasets and its ability to bolster statistical power over FarmCPU (Cebeci et al., 2023; Gai et al., 2023; Huang et al., 2019). Further research validates BLINK's effectiveness, particularly its Bayesian framework, assumes that causal genes are distributed variably, thus reducing the rate of false positives—an essential consideration for complex genetic patterns (Kim et al., 2021).

Out of the 20 associated SNPs, BlastX analysis successfully identified 7 SNPs associated with candidate genes. Subsequently, we utilized OmicsBox® to map and annotate these seven SNPs. It's noteworthy that the remaining SNPs lacked annotation in the reference database, possibly due to their lower sequence similarity, which was below the annotation cut-off threshold set at 55. Additionally, shorter sequences may inherently have limited information available for functional annotation, especially if they do not encompass critical functional domains or regions.

Among the annotated SNPs, the discovery of an SNP_01844 within a MADS-box gene holds exciting implications for controlling citric acid content in tomato fruit. MADS-box proteins play crucial roles in plant development, including fruit ripening and floral organ formation. Highlighted MADS-box proteins influence on ethylene production, a key ripening hormone that impacts organic acid metabolism (Garceau et al., 2017; Quinet et al., 2019; Wang et al., 2014a). The AGL21-like isoform X1, associated with this SNP, may affect citric acid regulation. This SNP could be a target for improving tomato taste profiles, though more evidence is needed to confirm causality.

In the practice of genomic selection (GS), breeders target a variety of traits without being constrained by their inheritability or genetic structure. These traits often exhibit commendable heritability, translating into predictable outcomes through GS, as supported by numerous studies (Duangjit et al., 2016; Hayes et al., 2009; Lorenz et al., 2012; Wimmer et al., 2013). High heritability traits are often dictated by a strong genetic influence, making them more amenable to GS predictions (Resende et al., 2012). However, the trait of locule number presents an anomaly with its high heritability (0.795) not correlating with the expected accuracy (0.643), which contrasts with traits like fruit weight that, despite lower heritability, predict more accurately. This pattern is supported by findings from Pascual et al. (2016) and Duangjit et al. (2016). Incorporating quantitative trait loci (QTL) information as part of the GS model for such traits is crucial. Generally, including the effects of QTLs that fall below the usual threshold in GS has been demonstrated to enhance selection efficacy for traits with low heritability (Calus et al., 2008).

The rrBLUP method, in particular, has been lauded for its efficacy (Resende et al., 2012). Incorporating Quantitative Trait Loci (QTL) identified in GWAS as fixed effects enhances the precision of genomic predictions (Spindel et al., 2016), a methodology that could be particularly advantageous for improving complex traits like fruit yield and

Table 5
Result from genomic selection using rrBLUP.

	Fruit weight	Locule number	Citric acid
Mean	0.851	0.643	0.633
SD	0.040	0.068	0.081
Median	0.858	0.646	0.640
Min	0.672	0.428	0.303
Max	0.941	0.826	0.818

quality in tomato. Such advancements in GWAS have the potential to significantly refine trait predictions, offering a promising avenue for tomato breeding programs.

5. Conclusions

This study emphasizes the importance of selecting an appropriate Genome-Wide Association Study (GWAS) model for plant breeding, with a focus on tomato crops where quality trait breeding is complex due to multiple genetic factors. The study tested various models and identified BLINK as the most suitable for its robustness in controlling false positives and negatives, particularly for traits like fruit weight, locule number, and citric acid content in tomato puree. Additionally, the study reported high accuracy in Genomic Selection (GS) using rrBLUP models for these traits in a germplasm population at Kasetsart University, indicating that precise selection of statistical tools is vital for identifying favorable alleles and developing effective breeding strategies to enhance crop quality traits.

CRediT authorship contribution statement

Natakorn Prateep-Na-Thalang: Formal analysis, Methodology, Data curation, Writing – original draft. **Pumipat Tongyoo:** Supervision, Visualization. **Chalermpol Phumichai:** Validation, Visualization. **Janejira Duangjit:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

Author 1, Janejira Duangjit, has received research funding from Research Fund for DPST Graduate with First Placement [Grant no. 019/2016], the Institute for the Promotion of Teaching Science and Technology (IPST), Thailand and the Thailand Plant Genetic Resource Center from KURDI and Center of Excellence on Agricultural Biotechnology, Office of the Permanent Secretary, Ministry of Higher Education, Science, Research and Innovation (AG-BIO/MHESI). Janejira Duangji is currently employed by Kasetsart University. Author 2, Natakorn Prateep-Na-Thalang, is a graduate student under Janejira Duangjit's supervision under the program at Center for Agricultural Biotechnology, Kasetsart University. Author 3, and author 4, Pumipat Tongyoo and Chalermpol Phumichai are employed by Kasetsart University. We affirm that there are no other financial or non-financial interests or relationships that could be perceived as potentially influencing the content or outcomes of this research.

Data availability

Data will be made available on request.

Acknowledgments

This work was financially supported by the Research Fund for DPST Graduate with First Placement [Grant no. 019/2016], the Institute for the Promotion of Teaching Science and Technology (IPST), Thailand, and the Center of Excellence on Agricultural Biotechnology, Office of the Permanent Secretary, Ministry of Higher Education, Science, Research and Innovation. (AG-BIO/MHESI) [AG-BIO/61-001-006].Half of genotyping was funded under research programs Thailand Plant Genetic Resource Center from KURDI and Center of Excellence on Agricultural Biotechnology, Office of the Permanent Secretary, Ministry of Higher Education, Science, Research and Innovation (AG-BIO/MHESI). The authors also acknowledge the Tropical Vegetable Research Center (TVRC), Kasetsart University, Kamphaeng Saen Campus, for providing tomato seeds. We would like to express our sincere gratitude to Assoc. Prof. Dr. Julapark Chunwongse for his invaluable guidance and mentorship throughout the course of this research.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., 2014. Fitting Linear Mixed-Effects Models Using Lme4. Cornell University arXiv:1406.5823. http://arxiv.org/abs/1406.5823 (accessed 1 August 2023).

Branthôme F.X., 2022. Worldwide (total fresh) tomato production exceeds 187 million tonnes in 2020. https://www.tomatonews.com/en/worldwide-total-fresh-tomato-production-exceeds-187-million-tonnes-in-2020_2_1565.html (accessed 1 August 2023).

Calus, M., 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.

Cappetta, E., Andolfo, G., Di Matteo, A., Barone, A., Frusciante, L., Ercolano, M.R., 2020. Accelerating tomato breeding by exploiting genomic selection approaches. Plants 9 (9), 1236. https://doi.org/10.3390/plants9091236.

Cebeci, Z., Bayraktar, M., Gökçe, G., 2023. Comparison of the statistical methods for genome-wide association studies on simulated quantitative traits of domesticated goats (*Capra hircus* L.). Small Rumin. Res. 227, 107053 https://doi.org/10.1016/j.smallrumres.2023.107053.

Celik, I., Gurbuz, N., Uncu, A.T., Frary, A., Doganlar, S., 2017. Genome-wide SNP discovery and QTL mapping for fruit quality traits in inbred backcross lines (IBLs) of *Solanum pimpinellifolium* using genotyping by sequencing. BMC Genom. 18, 1. https://doi.org/10.1186/s12864-016-3406-7.

Duangjit, J., Causse, M., Sauvage, C., 2016. Efficiency of genomic selection for tomato fruit quality. Mol. Breed. 36, 29. https://doi.org/10.1007/S11032-016-0453-3.

Endelman, J., 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4, 250–255. https://doi.org/10.3835/plantgenome2011.08.0024.

Fulton, T., Chunwongse, J., Tanksley, S., 1995. Microprep protocol for extraction of DNA from tomato and other herbaceous plants. Plant Mol. Biol. Report. 13, 207–209. https://doi.org/10.1007/BF02670897.

Gai, W., Yang, F., Yuan, L., ul Haq, S., Wang, Y., Wang, Y., Shang, L., Li, F., Ge, P., Dong, H., Tao, J., Wang, F., Zhang, X., Zhang, Y., 2023. Multiple-model GWAS identifies optimal allelic combinations of quantitative trait loci for malic acid in tomato. Hortic. Res. 10 (4) https://doi.org/10.1093/hr/uhad021 uhad021.

Garceau, D.C., Batson, M.K., Pan, I.L., 2017. Variations on a theme in fruit development: the *PLE* lineage of MADS-box genes in tomato (*TAGL1*) and other species. Planta 246, 313–321. https://doi.org/10.1007/s00425-017-2725-5.

Hayes, B., Bowman, P.J., Chamberlain, A.J., Goddard, M.E., 2009. Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92, 433–443. https://doi.org/10.3168/jds.2008-1646.

Holland, J.B., 2007. Genetic architecture of complex traits in plants. Curr. Opin. Plant. Biol. 10, 156–161. https://doi.org/10.1016/j.pbi.2007.01.003.

Huang, M., Liu, X., Zhou, Y., Summers, R.M., Zhang, Z., 2019. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience 8 (2). https://doi.org/10.1093/gigascience/giy154. Giy154.

Kaler, A.S., Gillman, J.D., Beissinger, T., Purcell, L.C., 2020. Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. Front. Plant Sci. 10, 1794. https://doi.org/10.3389/fpls.2019.01794.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., Eskin, E., 2008. Efficient control of population structure in model organism association mapping. Genetics 178, 1709–1723. https://doi.org/10.1534/genetics.107.080101.

Kim, M., Nguyen, T.T.P., Ahn, J.H., Kim, G.J., Sim, S.C., 2021. Genome-wide association study identifies QTL for eight fruit traits in cultivated tomato (*Solanum lycopersicum* L.). Hortic. Res. 8, 203. https://doi.org/10.1038/s41438-021-00638-4.

Kristensen, P.S., Jahoor, A., Andersen, J.R., Cericola, F., Orabi, J., Janss, L.L., Jensen, J., 2018. Genome-wide association studies and comparison of models and cross-validation strategies for genomic prediction of quality traits in advanced winter wheat breeding lines. Front. Plant Sci. 9 (69), 69. https://doi.org/10.3389/fpls.2018.00069.

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y.M., Todhunter, R.J., Buckler, E.S., Zhang, Z., 2014. Enrichment of statistical power for genome-wide association studies. BMC Biol. 12, 73. https://doi.org/10.1186/s12915-014-0073-5.

Li, C., Huang, Y., Huang, R., Wu, Y., Wang, W., 2018. The genetic architecture of amylose biosynthesis in maize kernel. Plant Biotechnol. J. 16, 688–695. https://doi.org/10.1111/pbi.12821.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., Huang, Z., Li, J., Zhang, C., Wang, T., Zhang, Y., Wang, A., Zhang, Y., Lin, K., Li, C., Xiong, G., Xue, Y., Mazzucato, A., Causse, M., Fei, Z., Giovannoni, J.J., Chetelat, R. T., Zamir, D., Städler, T., Li, J., Ye, Z., Du, Y., Huang, S., 2014. Genomic analyses provide insights into the history of tomato breeding. Nat. Genet. 46 (11), 1220–1226. https://doi.org/10.1038/ng.3117.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., Heckerman, D., 2011. FaST linear mixed models for genome-wide association studies. Nat. Methods 8, 833–835. https://doi.org/10.1038/nmeth.1681.

Lippman, Z., Tanksley, S.D., 2001. Dissecting the genetic pathway to extreme fruit size in tomato using a cross between the small-fruited wild species *Lycopersicon pimpinellifolium* and *L. esculentum* var. Giant Heirloom. Genetics 158, 413–422. https://doi.org/10.1093/genetics/158.1.413.

Liu, X., Huang, M., Fan, B., Buckler, E.S., Zhang, Z., 2016. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PloS Genet. 12 (2), e1005767 https://doi.org/10.1371/journal.pgen.1005767.

Lorenz, A., Smith, K., Jannink, J., 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci. 52, 1609–1621. https://doi.org/10.2135/cropsci2011.09.0503.

Mata-Nicolás, E., Montero-Pau, J., Gimeno-Paez, E., Garcia-Carpintero, V., Ziarsolo, P., Menda, N., Mueller, L.A., Blanca, J., Cañizares, J., van der Knaap, E., Díez, M.J., 2020. Exploiting the diversity of tomato: the development of a phenotypically and genetically detailed germplasm collection. Hortic. Res. 7, 1–14. https://doi.org/10.1038/s41438-020-0291-7.

Muños, S., Ranc, N., Botton, E., Bérard, A., Rolland, S., Duffé, P., Carretero, Y., Le Paslier, M.C., Delalande, C., Bouzayen, M., Brunel, D., Causse, M., 2011. Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near *WUSCHEL*. Plant. Physiol. 156, 2244–2254. https://doi.org/10.1104/pp.111.173997.

Pascual, L., Albert, E., Sauvage, C., Duangjit, J., Bouchet, J., Bitton, F., Desplat, N., Brunel, D., Paslier, M., Ranc, N., Bruguier, L., Chauchard, B., Verschave, P., Causse, M., 2016. Dissecting quantitative trait variation in the resequencing era: complementarity of bi-parental, multi-parental and association panels. Plant Sci. 242, 120–130. https://doi.org/10.1016/j.plantsci.2015.06.017.

Pérez-de-Castro, A.M., Vilanova, S., Cañizares, J., Pascual, L., Blanca, J.M., Díez, M.J., Prohens, J., Picó, B., 2012. Application of genomic tools in plant breeding. Curr. Genom. 13, 179–195. https://doi.org/10.2174/138920212800543084.

Phan, N.T., Trinh, L.T., Rho, M.Y., Park, T.S., Kim, O.R., Zhao, J., Kim, H.M., Sim, S.C., 2019. Identification of loci associated with fruit traits using genome-wide single nucleotide polymorphisms in a core collection of tomato (*Solanum lycopersicum* L.). Sci. Hortic. 243, 567–574. https://doi.org/10.1016/j.scienta.2018.09.003. Amst.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38 (8), 904–909. https://doi.org/10.1038/ng1847.

Quinet, M., Angosto, T., Yuste-Lisbona, F.J., Blanchard-Gros, R., Bigot, S., Martinez, J.P., Lutts, S., 2019. Tomato fruit development and metabolism. Front. Plant. Sci. 10, 1554. https://doi.org/10.3389/fpls.2019.01554.

R Core Team, 2023. R a Language and Environment For Statistical computing. R foundation For Statistical Computing. R Core Team, Vienna, Austria. https://www.R-project.org/. Accessed 1 August 2023.

Ranc, N., Muños, S., Xu, J., Le Paslier, M.C., Chauveau, A., Bounon, R., Rolland, S., Bouchet, J.P., Brunel, D., Causse, M., 2012. Genome-wide association mapping in tomato (*Solanum lycopersicum*) is possible using genome admixture of *Solanum lycopersicum* var *cerasiforme*. G3 2, 853–864. https://doi.org/10.1534/g3.112.002667 (Bethesda).

Resende, J.M., Muñoz, P., Resende, M.D., Garrick, D.J., Fernando, R.L., Davis, J.M., Jokela, E.J., Martin, T.A., Peter, G.F., Kirst, M., 2012. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 190, 1503–1510. https://doi.org/10.1534/genetics.111.137026.

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., Altmann, T., Stitt, M., Willmitzer, L., Melchinger, A.E., 2012. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc. Natl. Acad. Sci. 109 (23), 8872–8877. https://doi.org/10.1073/pnas.1120813109.

Rodríguez, G.R., Kim, H.J., van der Knaap, E., 2013. Mapping of two suppressors of *OVATE* (*sov*) loci in tomato. Heredity 111, 256–264. https://doi.org/10.1038/hdy.2013.45 (Edinb).

Ruggieri, V., Francese, G., Sacco, A., D'Alessandro, A., Rigano, M.M., Parisi, M., Milone, M., Cardi, T., Mennella, G., Barone, A., 2014. An association mapping approach to identify favourable alleles for tomato fruit quality breeding. BMC Plant Biol. 14, 337. https://doi.org/10.1186/s12870-014-0337-9.

Sacco, A., Ruggieri, V., Parisi, M., Festa, G., Rigano, M.M., Picarella, M.E., Mazzucato, A., Barone, A., 2015. Exploring a tomato landraces collection for fruit-related traits by the aid of a high-throughput genomic platform. PLoS One 10, e0137139. https://doi.org/10.1371/journal.pone.0137139.

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P.T., Nikoloski, Z., Fernie, A.R., Causse, M., 2014. Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant Physiol. 165, 1120–1132. https://doi.org/10.1104/pp.114.241521.

Segura, V., Vilhjalmsson, B.J., Platt, A., Korte, A., Seren, U., Long, Q., Nordborg, M., 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44, 825–830. https://doi.org/10.1038/ng.2314.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J.L., McCouch, S.R., 2016. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity 116, 395–408. https://doi.org/10.1038/hdy.2015.113 (Edinb).

Stich, B., Melchinger, A.E., 2009. Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and arabidopsis. BMC Genom. 10, 94. https://doi.org/10.1186/1471-2164-10-94.

Tamba, C.L., Ni, Y.L., Zhang, Y.M., 2017. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. PloS Comput. Biol. 13, e1005357 https://doi.org/10.1371/journal.pcbi.1005357.

Tieman, D., Zhu, G., Resende, M.F.R., Lin, T., Nguyen, C., Bies, D., Rambla, J.L., Beltran, K.S.O., Taylor, M., Zhang, B., Ikeda, H., Liu, Z., Fisher, J., Zemach, I., Monforte, A., Zamir, D., Granell, A., Kirst, M., Huang, S., Klee, H., 2017. A chemical genetic roadmap to improved tomato flavor. Science 355, 391–394. https://doi.org/10.1126/science.aal1556.

Voight, B.F., Pritchard, J.K., 2005. Confounding from cryptic relatedness in case–control association studies. PLoS Genet. 1 https://doi.org/10.1371/journal.pgen.0010032 e32-10.1371.

Wang, S., Lu, G., Hou, Z., Luo, Z., Wang, T., Li, H., Zhang, J., Ye, Z., 2014a. Members of the tomato FRUITFULL MADS-box family regulate style abscission and fruit ripening. J. Exp. Bot. 65, 3005–3014. https://doi.org/10.1093/jxb/eru137.

Wang, Q., Tian, F., Pan, Y., Buckler, E.S., Zhang, Z., Li, Y., 2014b. A super powerful method for genome wide association study. PLoS One 9 (9), e107684. https://doi.org/10.1371/journal.pone.0107684.

Wang, J., Zhang, Z., 2021. GAPIT version 3: boosting power and accuracy for genomic association and prediction. Genom. Proteom. Bioinform. 19 (4), 629–636. https://doi.org/10.1016/j.gpb.2021.07.006.

Wen, Z., Boyse, J.F., Song, Q., Cregan, P.B., Wang, D., 2015. Genomic consequences of selection and genome-wide association mapping in soybean. BMC Genom. 16 (1), 671. https://doi.org/10.1186/s12864-015-1872-y.

Wen, Y.J., Zhang, H., Ni, Y.L., Huang, B., Zhang, J., Feng, J.Y., Wang, S.B., Dunwell, J. M., Zhang, Y.M., Wu, R., 2018. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Brief Bioinform. 19, 700–712. https://doi.org/10.1093/bib/bbw145.

Wimmer, V., Albrecht, T., Auinger, H.J., Schön, C.C., 2013. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195, 573–587. https://doi.org/10.1534/genetics.113.150078.

Würschum, T., Langer, S.M., Longin, C.F.H., Tucker, M.R., Leiser, W.L., 2012. Improved efficiency of doubled haploid generation in hexaploid triticale by *in vitro* chromosome doubling. BMC Plant Biol. 12, 109. https://doi.org/10.1186/1471-2229-12-109.

Xu, J., van Heusden, A.W., Bovy, A., Cao, K., Bai, Y., Visser, R.G.F., Finkers, R., 2013. Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. Theor. Appl. Genet. 126, 567–581. https://doi.org/10.1007/s00122-012-2002-8.

Yamamoto, E., Matsunaga, H., Onogi, A., Kajiya-Kanegae, H., Minamikawa, M., Suzuki, A., Shirasawa, K., Hirakawa, H., Nunome, T., Yamaguchi, H., Miyatake, K., Ohyama, A., Iwata, H., Fukuoka, H., 2016. A simulation-based breeding design that uses whole-genome prediction in tomato. Sci. Rep. 6, 19454. https://doi.org/10.1038/srep19454.

Yu, J., Buckler, E.S., 2006. Genetic association mapping and genome organization of maize. Curr. Opin. Biotechnol. 17, 155–160. https://doi.org/10.1016/j.copbio.2006.02.003.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S., 2006. A unifed mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38, 203–208. https://doi.org/10.1038/ng1702.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., Buckler, E.S., 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42, 355–360. https://doi.org/10.1038/ng.546.

Zhang, J., Zhao, J., Xu, Y., Liang, J., Chang, P., Yan, F., Li, M., Li, J., Jin, W., Lai, J., Zou, Z., 2015. Genome-wide association mapping for tomato volatiles positively contributing to tomato flavor. Front. Plant. Sci. 6, 1042. https://doi.org/10.3389/fpls.2015.01042.

Zhao, J., Sauvage, C., Zhao, J., Bitton, F., Bauchet, G., Liu, D., Huang, S., Tieman, D.M., Klee, H.J., Causse, M., 2019. Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. Nat. Commun. 10, 1534. https://doi.org/10.1038/s41467-019-09462-w.