Classification of fruit shape based on decision tree model and identification of QTLs and genes controlling fruit shape in eggplant[1]

Qiang Li[1*#], Shuangxia Luo[1*], Huimin Du[1*], Chive Paradowski[2*], Jingjian Ma[1], Liying Zhang[1], Xupeng Jia[1], Ruoxuan Zhao[1], Dongfang Zhang[1], Wei Yan[3], Jianan Liu[3], Lijun Song[1], Esther van der Knaap[4], Sofia Visa[2#], Xueping Chen[1#]

[1] *College of Horticulture/Key Laboratory of Vegetable Germplasm Innovation and Utilization of Hebei, Hebei Agricultural University, Baoding 071000, China*

[2] *Department of Mathematical & Computational Sciences, The College of Wooster, Wooster, OH 44691, USA*

[3] *College of Life Sciences/Institute of Life Science and Green Development, Hebei University, Baoding 071000, China*

[4] *Center for Applied Genetic Technologies/Department of Horticulture, University of Georgia, Athens, GA 30602, USA*

Highlights

• Thirteen shape categories and ten key attributes determining fruit shape were identified based on decision tree model in eggplant

•Classification rules achieving an accuracy of 92.59% were generated.

•Four QTLs controlling FSI and PAMi were detected using GWAS and QTL-seq.

•Overexpression of the candidate gene *SmFSI3.1*/*SmFL* resulted in the production of elongated tomato fruits

Abstract

Eggplant (*Solanum melongena* L.) shows remarkable diversity in fruit shape, making it an excellent model for studying shape variation. Eggplant fruit shape influences consumer preference and plays an important role in the classification of commercial varieties and germplasm. Despite its importance, existing classification systems are limited to description without quantitative criteria, differ by country or region and fail to fully capture the diversity of eggplant fruit shapes. In the present study, thirteen shape categories were identified using a decision tree model with Gini index-based variable selection. Ten key attributes that largely determine fruit shape were identified and high accuracy (92.59%) classification rules were generated. Five other methods, including random forest, XGBoost, SVM, K-means and GMM, were also applied to fruit shape classification, but they proved less robust for classification compared to the decision tree. The shape modeling informed the key attribute selection for the QTL-seq and GWAS analyses. Four QTLs controlling Fruit Shape Index (FSI) and Proximal Angle Micro (PAMi) were detected using GWAS and QTL-seq. The candidate gene *SmFSI3.1*/*SmFL*, a member of the SUN/IQD family, was over-expressed in tomato and resulted in elongated fruits, indicating the positive roles of this gene in regulating fruit elongation in eggplant. In summary, we developed an accurate and reproducible model for classifying eggplant fruit shapes, which is of significance for eggplant breeding and variety classification. Moreover, we verified the function of the causal gene responsible for *fsi3.1*/*fl3.1* locus, providing a foundation for understanding the genetic regulation of fruit shape in eggplant.

Keywords: Eggplant, Decision tree, Fruit shape, Classification system, IQD

1. Introduction

Eggplant (*Solanum melongena* L.) is one of the most important horticultural crops in the world. The scarlet eggplant (*S. aethiopicum* L.) and the gboma eggplant (*S. macrocarpon* L.) are two species that are mainly cultivated in Africa (Taher *et al*. 2017). Global production of eggplant was 60.8 million tons in 2023, of which China accounted for 64.6%, ranking the country as the largest eggplant producer in the world (FAO, 2023).

[1] [#]Correspondence Qiang Li, E-mail: yylq@hebau.edu.cn; Xueping Chen, E-mail: chenxueping@hebau.edu.cn; Sofia Visa, E-mail: svisa@wooster.edu

[*] These authors contributed equally to this study.

Fruit shape is one of the important domesticated traits for fruit bearing crops (Rodriguez *et al*. 2011). For example, while the direct wild ancestor *S. insanum* of eggplant bears small and round fruits, cultivated eggplants are characterized by diverse fruit shapes (Kaushik *et al*. 2016; Liu *et al*. 2019). Fruit shape is also an important agronomic trait that influences both the use of crops and consumer preferences, which varies across different regions and countries. For example, while round and flat tomatoes are commonly used in the fresh market, varieties bearing elongated and blocky fruits are more suitable for machine harvesting and producing tomato sauce (Li *et al*. 2023; Zhu *et al*. 2023). In China, round eggplant fruits are mainly consumed in north, whereas the long fruits are more popular in the south.

To date, existing classification systems for eggplant fruit shape have limitations. IBPGR (International Board for Plant Genetic Resources) and UPOV (The International Union for the Protection of New Varieties of Plants) are two common classification systems of fruit shape. In IBPGR, the attributes of eggplant fruit shape are length, width and the fruit length to width ratio (IBPGR, 1990). Using these measurements, eggplants can be classified into six shape categories, i.e. broader than long; as long as broad; slightly longer than broad; twice as long as broad; three times as long as broad; several times as long as broad (IBPGR, 1990). In UPOV, fruit length, width, and shape index (Length/Width) are the three attributes of eggplant fruit shape, which are further used to describe eight shape categories, including Flattened globular, Globular, Ovoid, Obovate, Pear shaped, Club shaped, Ellipsoid and Cylindrical (UPOV, 2012). In China, a descriptive norm classified the eggplant fruit shape into 11 categories, including Flattened globular, Globular, Long globular, Ovoid, Long ovoid, Short cylindrical, Long cylindrical, Strip shapped, Extremely long, Short arietiform, Long arietiform (Li and Zhu 2006). Guidelines for the conduct of tests for distinctness, uniformity and stability of eggplant (GB/T 19557.20-2017) showed the same eight fruit shapes as UPOV. In addition, eggplant varieties are often roughly classified into three types, i.e. round, oval and long (Liu *et al*. 2019). Obviously, these categories only show a fraction of shapes in eggplant, since a huge variation of fruits have been observed with widely varying morphologies ranging from flat to extremely long. Importantly, these classifications are mainly focused on description of the shapes but lacking quantitative standards.

Fruit shape is a complex, multi-dimensional quantitative trait influenced by both genetic and environmental factors (Feldmann *et al*. 2020; Li *et al*. 2023). It is also a necessary characteristic for describing germplasm resources and registering new varieties. Currently, fruit shape was often characterized by linear attributes (e.g., fruit length, width and shape index), since fruit length and width could be measured easily. In eggplant, numerous QTLs regulating fruit length, width, and shape index have been characterized. For instance, two major QTLs influencing fruit length variation, *flE03* and *flE11*, were mapped using an F_2 population derived from the cross between '305E40' and '67/3' at two locations (Portis *et al*. 2014). Two QTLs for fruit width, accounting for 38.2% and 30.2% of the phenotypic variance (PV), respectively, were also detected at both sites (Portis *et al*. 2014). In addition, three QTLs for fruit shape index (*fsE01*, *fsE03a* and *fsE07*) were identified in the same environments (Portis *et al*. 2014). A stable QTL on Chr03 was reported for three FSI traits (F_Shape_E_I, F_Shape_E_II, and C_F_Shape) in two introgression lines (ILs) (Mangino *et al*. 2021). Using three F_2 populations, Pang *et al*. (2024) detected a major-effect QTL for fruit length, *fl3.1*, and proposed *Smechr0302217* (*SmeFL*) as the most likely candidate gene underlying *fl3.1*. Furthermore, Yu *et al*. (2024a) identified a QTL for fruit shape index in segregating populations derived from a cross between the oval-fruited line 'E421' and the round-fruited line '145'. This QTL was associated with a 91-bp indel in the promoter region of *EGP11251*, which results in the up-regulation of its expression.

Evaluation fruit shape only with the linear attributes leads to the loss of important information and imprecise assessment of fruit shape (Mangino *et al*. 2021), especially when differences between fruits are very subtle. Tomato Analyzer (TA) has been shown to be an accurate and high-throughput tool for precisely characterizing fruit shape in many fruit-bearing crops, such as tomato (Rodríguez *et al*. 2010; Nankar *et al*. 2020; Quispe-Choque *et al*. 2022), eggplant (Hurtado *et al*. 2013; Mangino *et al*. 2021) and pepper (Tripodi and Greco, 2018; Pereira-Dias *et al*. 2020). TA allows decomposing fruit shape into up to 37 quantitative attributes (Gonzalo *et al*. 2009), which may empower the identification of underlying QTLs or genes through quantitative genetic approaches, including genome-wide association studies (GWAS).

Many studies have demonstrated the effectiveness of morphometrics and machine learning to automate classification in modeling and classifying fruit shapes, contributing to advancements in agricultural automation and quality control. Zakeri *et al*. (2021) proposed a computer vision-based method for grading jujube fruits using decision trees. By extracting visual

features such as color, shape, size, texture, defection and wrinkle features, and applying feature selection algorithms, the study achieved a classification accuracy of 98.8% using decision trees (Zakeri *et al*. 2021). Kaur *et al*. (2024) proposed an automated fruit classification system utilizing both K-Nearest Neighbors and decision tree algorithms, which demonstrated the applicability of decision trees in industrial settings (Kaur *et al*. 2024). We have previously applied Fourier attributes and Bayesian classification to classify tomato fruit into nine distinct shape categories (Visa *et al*. 2014). In addition, machine learning was used to automatically sort fruit into shape categories in both strawberry (Feldmann *et al*. 2020). Despite the significance of fruit shape in eggplant, the modeling and QTLs of fruit shape are still lacking.

In this study, we used TA to measure fruit shape attributes in a natural population of 285 accessions grown in 2021 and 2023. Several machine learning algorithms including decision tree, random forest, XG Boost (Extreme Gradient Boosting), Support Vector Machines (SVM), K-Means, Gaussian Mixture Models (GMM) were applied to classify the fruit shapes. Among these, decision trees performed best, both in modeling the dataset and in generating an interpretable rule set for categorization. Using the decision tree modeling, we identified the most significant descriptors of eggplant fruit shape, which were used to define 13 distinct fruit shape categories. Moreover, four QTLs responsible for fruit shape variation were detected using GWAS and QTL-seq. The function of *SmFSI3.1* was confirmed to regulate fruit shape through transgenic over-expression the gene in tomato.

2. Material and methods

2.1 Plant material and growing conditions

A collection of 285 eggplant accessions was used in the present study (Table S1). The collection was comprised by 244 *S. melongena*, 27 *S. aethiopicum*, 11 *S. macrocarpon* and three *S. integrifolium* accessions (Table S1). The experiment was carried out during the spring-summer season (April to August) of the years 2021 and 2023 at the greenhouse of the Hebei Agricultural University (E 115° 42′, N38° 81′). For each accession, three to four plants were grown with local management standards.

2.2 Measurements of fruit shape attributes and data analyses

In each year, at the physiologically ripe stage, two to three fruits per plant were harvested. Abnormally shaped fruits were removed, resulting in 1,315 fruits and 1,344 fruits in 2021and 2023, respectively to evaluate fruit shape attributes. The fruits were sliced longitudinally and scanned with LA-S scanner (Wanshen, Hangzhou, China) at 300dpi resolution. The resulted images were analyzed with Tomato Analyzer version 4.0 (Rodríguez *et al*. 2010). A total of 22 attributes were measured, with detailed description of the 22 attributes available in the Tomato Analyzer version 4.0 manual (https://vanderknaaplab.uga.edu/tomato-analyzer/). Measurements such as height and width are reported in centimeters. It is worth emphasizing that in the Basic Measurements section of the manual there are two attributes for fruit width and three attributes for fruit height, and in the Fruit Shape Index section there are three attributes of fruit shape index. Given the diverse fruit shapes in our collection, especially curved fruits, we selected only one attribute each for fruit height, width and shape index. For example, attributes like Maximum Height and Height Mid-width are meaningless for curved fruits, so only Curved Height was used for measuring fruit height. For globular fruits, only Maximum Height was kept representing the fruit height, as Height Mid-width, Maximum Height and Curved Height are similar. The Fruit Shape Index was calculated as the ratio of Height to Width. Additionally, morphometric data consisting of 200 two-dimensional boundary points were obtained as previously described (Visa *et al*. 2014).

R software was used to calculate the basic descriptive statistics, including mean, maximum and minimum values, standard deviation (SD), coefficient of variation (CV) and genetic diversity index (H') of the 22 attributes. Pearson correlation coefficient was analyzed using R package 'corrplot'. Two-way analysis of variance (ANOVA) was conducted in R software to determine the influence of genotype (G), environment (years, E) and the interaction between genotype and environment (G × E) on the attributes separately. To determine the contribution of G, E and G × E on the variance of the attributes, the total sum of squares was partitioned into sums of squares for G, E, G × E and residual effect, and the contributions were shown in percentage over total sums of squares (Zeng *et al*. 2017).

2.3 Fruit shape modeling

The process began with data preprocessing, during which malformed fruit images were removed. We then labeled 1,344 fruit images into the 13 predefined shape categories. For each of the 13 categories, accurate shape prototypes were generated using the 200 two-dimensional morphometric boundary points as previously described by (Visa *et al*. 2014).

Using Python scripts and scikit-learn open-source library (Pedregosa *et al*. 2011), we evaluated several models for classifying fruit shapes in the natural population, namely decision tree, random forest, XG Boost, Support Vector Machines (SVM), K-Means, Gaussian Mixture Models (GMM). Each model was evaluated using 5-fold cross-validation, and we have experimented with various parameter configurations. As explained in the results section, decision tree models performed best, and were therefore selected to derive a ruleset describing the 13 shape categories.

To classify fruit shapes, we applied a decision tree model using the Gini impurity splitting criterion. We used an 80/20 train-test split and a maximum of 30 classification rules. This rule limitation allows for one or more rules per class while preventing overfitting and excessive tree complexity. The tree is constructed iteratively by selecting the most informative shape attribute at each node. Once the tree is constructed, each path from root to leaf is converted into an if-then rules. To improve generalization, we applied rule-post pruning, simplifying the model while retaining predictive accuracy. Pruned rules also reduce overfitting and enhance performance on unseen data. If pruning a rule improves generalization, it is kept in its simplified form. Finally, the pruned rule set is used to classify fruit images into shape category. As discussed in the results section, this model performed comparably with more sophisticated approaches such as random forest and XGBoost, while outperforming all other algorithms. Importantly, transparent and interpretable rules based on geometric fruit properties are preferable to black-box predictors, making decision tree the better choice for this application.

For comparison, we also implemented other models and tested them in a 5-fold cross validation setting. The XGBoost algorithm, which combines many small decision trees to improve predictive performance, and random forest models with a maximum of 100 tree estimators gave similar performance to the decision tree. For the XGBoost, we allowed up to 50 boosting iterations (i.e., up to 50 trees added sequentially, each correcting errors of the previous trees), with an early stopping criterion of 10 rounds to prevent overfitting. The SVM model was evaluated with linear, polynomial, and rgb kernels. Because SVMs are naturally binary classifiers, we have modeled a one-vs-one SVM, training a separate SVM for every pair of classes and using a voting system to predict a final category for a test data. In addition, we tested unsupervised clustering methods, including k-means clustering with k values ranging from 7 to 16 and with multiple distance measures, as well as Gaussian Mixture Models (GMMs) with 7 to 16 components. None of these models reached the predictive power of decision trees.

F1 measure is the harmonic mean of precision and recall, requiring both high precision and high recall, giving a better overall picture than accuracy alone. In addition to accuracy, which can be misleading when having imbalanced classes, we also computed the F1 score. For each of the thirteen shape classes C_i with i = 1~13, the F1 score is defined as:

F1(C_i) = 2 * (Precision(C_i) * Recall(C_i)) / (Precision(C_i) + Recall(C_i))

Precision measures how many instances predicted as class C_i are actually in class C_i:

Precision(C_i) = TP(C_i) / (TP(C_i) + FP(C_i))

where TP(C_i) (true positives) is the number of examples correctly classified as C_i, and FP(C_i) (false positives) is the number of examples incorrectly classified as C_i.

Recall, also known as sensitivity, measures how many actual C_i examples are correctly identified:

Recall(C_i) = TP(C_i) / (TP(C_i) + FN(C_i))

Where FN(C_i) (false negatives) is the number of examples belonging to class (C_i) that are incorrectly classified as another class.

2.4 QTL-seq Mapping Approach

To verify the fruit shape modeling and identify the QTLs responsible for the FSI, an F_2 population was generated by crossing '14-345' bearing flattened globular fruits with '16-562' bearing ellipsoid fruits. Note, the decision tree classifies the fruit of these genotypes as C1 Flattened Globular and C7 Ellipsoid, respectively. A total of 211 F_2 plants were phenotyped. A total of 414 fruits were used for fruit shape classification in the modeling analysis (two fruits were collected from each of 203 plants, and one fruit was collected from each of eight plants). For QTL-seq, four DNA pools were prepared: two parental pools ('14-345' and '16-562') and two segregating F_2 bulks. The 'FSI_max pool' contained equal amounts of DNA from 30 F_2 plants with a long-fruit phenotype (FSI > 2), whereas the 'FSI_min pool' contained equal amounts of DNA from 30 F_2 plants with a round-fruit phenotype (FSI <1.3). The four bulked DNA pools were sequenced using the Illumina HiSeq^{TM} 2500 platform at GENOSEQ Science and Technology Ltd (Wuhan, China). After filtering the adapter and low-quality reads, the clean reads were aligned to the 'HQ-1315' eggplant genome V2.0 (http://47.92.172.28:12068/Eggplant/home/index) (Wei *et al*. 2025) using the Burrow-Wheeler Aligner (BWA) (Li and Durbin, 2010). SNPs and Indels were identified using the GATK software (McKenna *et al*. 2010). The SNP and ΔSNP indices were calculated and analyzed by QTLseqr R package (Mansfeld and Grumet, 2018) to identify genomic regions associated with FSI.

2.5 Genome-wide association analysis

The 285 accessions were re-sequenced, and the resulting reads were aligned to the reference genome of the inbred line '14-345' using BWA software to obtain BAM files. PCR duplicates in the BAM files were removed using SAMtools. SNP variations of the 285 accessions were identified with the Genome Analysis Toolkit (GATK) (McKenna *et al*. 2010). For the association analysis, SNPs with a minor allele frequency (MAF) of ≥ 0.2% and a missing rate of ≤ 10% were used. GWAS was performed using the rMVP (R package) with the Mixed Linear Model (MLM) and Fixed and Random Model Circulating Probability Unification (FarmCPU). LDBlockShow (https://github.com/BGI-shenzhen/LDBlockShow/) was used for the linkage disequilibrium (LD) analysis.

2.6 Vector construction and transformation of tomato

The full-length coding DNA sequence of Sme03G2052 was amplified from ovaries at 0DPA of the '14-345'. The amplified product was inserted between XbaI and KpnI in the pBI121 vector driven by Cauliflower mosaic virus (CaMV) 35S promoter. The over-expression vector was then transformed into Agrobacterium tumefaciens GV3101. The method of genetic transformation of tomato mediated by A. tumefaciens was performed as described previously (Li *et al*. 2022). To analysis the expression of *SmeFL*, the ovaries at anthesis were used for RNA isolation (Cat# LS1040; Promega), and complementary DNA from 1 μg of total RNA was synthesized using reverse transcriptase kit (Cat# R2020L; Singabio). Real-time quantitative RT-PCR (qRT-PCR) was performed on LightCycler 96 (Roche, Switzerland) using FastReal qPCR PreMix Ⅱ (TIANGEN, China). Tomato Clathrin adaptor complexes medium subunit (CAC) gene served as the internal control for normalization. The relative expression level was calculated using the $2^{-\Delta\Delta C_T}$ method. The primers used in this study are listed in Table S2.

3. Results

3.1. Diversity of fruit shape attributes in the natural population of eggplant

A total of 22 fruit shape attributes were evaluated in the natural population during 2021 and 2023. Descriptive statistics revealed a wide range of variation, with coefficients of variation (CV) ranging from low (CV < 10%) to very high (CV> 1000%), and average CVs of 103.5% in 2021 and 110.3% in 2023 (Table 1). The highest CV was observed for Distal End Protrusion (DEP) at 870% in 2021 and 1060% in 2023, followed by Shoulder Height (SH) (228% in 2021 and 194% in 2023). The lowest CVs were recorded for Proximal Fruit Blockiness (PFB) (7% in 2021 and 6% in 2023), followed by Distal Fruit Blockiness (DFB) (9% in 2021 and 2023) and Fruit Shape Triangle (FST) (12% in 2021 and 2023) (Table 1). The genetic diversity index (H') of the attributes ranged from 1.41 to 5.66 in 2021 and from 1.01 to 5.65 in 2023, with mean values of 5.16 in both 2021 and 2023 (Table 1). The highest H' values were recorded for of PFB, DFB, FST and Rectangular,

while DEP had the lowest H' values in both 2021 and 2023 (Table 1). These results indicate the high morphological diversity in fruit shape within the natural eggplant population.

3.2. Correlation and variation of fruit shape attributes across years

Pearson correlation coefficient (PCC) analysis revealed significant positive correlations (r = 0.44~0.96, p<0.001) between all 22 attributes in 2021 and 2023 (Table 2). All attributes showed moderate to strong correlation (r ≥ 0.61, p<0.001), with the exception of DEP showing a lower correlation of 0.44. Notably, FSI showed the strongest correlation (r = 0.96, p<0.001), followed by Proximal Angle Micro (PAMi) (r=0.94), Proximal Angle Macro (PAMa) (r=0.93), Distal Angle Macro (DAMa) (r = 0.93), Circular (r = 0.93) and Distal Angle Micro (DAMi) (r = 0.90) (Table 2).

Two-way ANOVA was performed to further assess the effects of genotype, environment and their interaction (G×E) on fruit shape attributes. The results revealed that all 22 attributes were significantly affected by genotype, environment, and G×E, except for environment effects on Ellipsoid, PAMi, DEP and Width Widest Position (WWP) (Table 2). Genotype contributed the most to the total variation of all 22 attributes, accounting for 38.97%~94.98% of total variance (Table 2). The lowest genotype contribution was for DEP (bellow 60%), indicating that this attribute is under substantial environmental and/or developmental control. On the other hand, PAMa (94.98%), DAMa (93.07%), FSI (92.04%) and Circular (91.82%) each exceeded 90% (Table 2). Environment had a significant influence on all the attributes, except for Ellipsoid, PAMi, DEP and WWP. Environmental effects contributed 0%~19.18%, while the genotype by environment interaction contributed 2.14% to 26.38% of the total variance. Residual effect had a contribution to the total variance ranging from 2.43% to 34.58% (Table 2). These results indicate that genotype is the predominant factor influencing the attributes and hence the fruit shape, which explains the high consistency of attribute values across years and environments. Based on this, data from 2023 were used for further analyses.

Pairwise correlation analysis among the 22 fruit shape attributes revealed numerous significant correlations. High positive correlations were observed among the attributes of proximal and distal ends (r = 0.76~0.96), including PIA, SH, DAMa, DAMi, PAMa and PAMi (Fig. 1). That is, SH and PIA increase with larger angle of proximal and distal ends. High positive correlations were also observed in the following comparisons: PFB, FST, and Ovoid (r = 0.83~0.95); DFB, Obovoid, and WWP (r = 0.89~0.90); Circular, Ellipsoid, FSI, and Height (r = 0.68~0.93) (Fig. 1). Interestingly, FST showed high negative correlations with DFB (r= -0.91), Obovoid (r = -0.94), and WWP (r = -0.94). Moreover, FSI showed high negative correlation with the proximal and distal angles (r = -0.96~-0.98) (Fig. 1), i.e. long fruits with larger FSI values have small angles of proximal and distal ends.

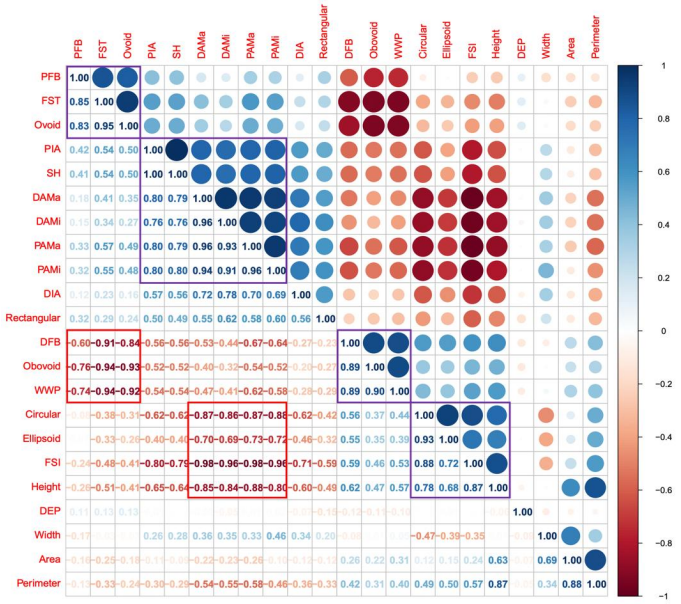

Fig. 1 Correlations of the 22 attributes.

Table 1 Summary of descriptive statistics of the 22 attributes of the 285 accessions in 2021 and 2023

Attribute		2021						2023					
Full name	Abbreviation	Mean	Min	Max	SD	CV	Shannon-H	Mean	Min	Max	SD	CV	Shannon-H
Fruit Area (cm^2)	Area	49.09	2.76	160.27	30.02	0.61	5.46	91.50	6.41	215.69	45.55	0.50	5.51
Fruit Perimeter (cm)	Perimeter	27.85	6.21	58.08	10.39	0.37	5.58	39.44	9.55	68.59	13.14	0.33	5.59
Fruit Height (cm)	Height	10.04	1.80	26.09	4.82	0.48	5.53	14.29	2.65	30.84	6.29	0.44	5.54
Fruit Width (cm)	Width	5.75	1.90	12.56	2.02	0.35	5.61	7.91	2.64	17.22	3.22	0.41	5.57
Fruit Shape Index	FSI	1.91	0.59	10.09	1.22	0.64	5.50	2.08	0.56	8.33	1.37	0.66	5.46
Distal Fruit Blockiness	DFB	0.95	0.76	1.19	0.08	0.09	5.66	0.93	0.71	1.22	0.08	0.09	5.65
Proximal Fruit Blockiness	PFB	0.90	0.75	1.15	0.06	0.07	5.65	0.89	0.73	1.04	0.05	0.06	5.65
Fruit Shape Triangle	FST	0.97	0.68	1.41	0.12	0.12	5.66	0.98	0.69	1.44	0.12	0.12	5.65
Fruit Ellipsoid	Ellipsoid	0.06	0.02	0.20	0.03	0.44	5.57	0.06	0.02	0.18	0.03	0.41	5.57
Fruit Circular	Circular	0.18	0.03	0.43	0.10	0.54	5.50	0.20	0.02	0.44	0.11	0.53	5.50
Fruit Rectangular	Rectangular	0.51	0.27	0.61	0.05	0.10	5.65	0.52	0.25	0.63	0.05	0.09	5.65
Shoulder Height	SH	0.01	0.00	0.18	0.03	2.28	4.01	0.02	0.00	0.25	0.04	1.94	4.27
Proximal Angle Macro (°)	PAMa	76.20	3.05	185.07	51.08	0.67	5.43	68.64	2.86	191.65	56.60	0.82	5.3
Proximal Angle Micro (°)	PAMi	132.78	11.57	255.34	55.68	0.42	5.56	133.08	13.23	274.73	66.23	0.50	5.52
Proximal Indentation Area	PIA	0.07	0.00	0.90	0.15	1.92	4.04	0.09	0.00	0.88	0.18	1.90	4.28
Distal Angle Macro (°)	DAMa	77.98	7.70	174.97	40.21	0.52	5.51	71.07	5.85	173.81	41.92	0.59	5.48
Distal Angle Micro (°)	DAMi	124.92	12.53	241.93	44.89	0.36	5.60	113.86	9.70	229.02	46.86	0.41	5.56
Distal Indentation Area	DIA	0.02	0.00	0.23	0.03	1.62	4.72	0.01	0.00	0.22	0.02	1.76	4.70
Distal End Protrusion	DEP	0.00	0.00	0.04	0.00	8.70	1.41	0.00	0.00	0.13	0.01	10.60	1.15
Fruit Obovoid	Obovoid	0.13	0.00	0.43	0.10	0.80	5.30	0.11	0.00	0.39	0.10	0.97	5.13
Fruit Ovoid	Ovoid	0.06	0.00	0.40	0.07	1.15	5.00	0.07	0.00	0.29	0.07	0.97	5.11
Width Widest Position	WWP	0.54	0.24	0.78	0.09	0.18	5.64	0.54	0.28	0.78	0.09	0.16	5.64

Table 2 Correlation of the 22 attributes between two years and two-way ANOVA analysis of variance of the 22 attributes from 285 accessions in the two environments

Attribute	Pearson correlation coefficient between 2021 and 2023	Contribution to the total of variance (%)			
		Genotype	Environment	Genotype: Environment	Residuals
Area	0.70***	60.84***	19.18***	12.15***	7.83
Perimeter	0.73***	68.74***	16.44***	8.52***	6.30
Height	0.86***	76.50***	10.90***	7.12***	5.48
Width	0.84***	74.38***	11.10***	9.72***	4.81
FSI	0.96***	92.04***	0.36***	3.53***	4.07
DFB	0.79***	76.65***	1.17***	7.77***	14.41
PFB	0.65***	66.71***	0.75***	12.98***	19.56
FST	0.80***	78.18***	0.12**	7.97***	13.73
Ellipsoid	0.87***	83.73***	0.01	5.39***	10.87
Circular	0.93***	91.82***	0.88***	3.04***	4.26
Rectangular	0.61***	63.06***	0.30***	9.73**	26.92
SH	0.80***	68.32***	0.32***	10.54***	20.79
PAMa	0.93***	94.98***	0.45***	2.14***	2.43
PAMi	0.94***	64.78***	0.00	2.33***	32.89
PIA	0.80***	84.36***	0.42***	6.43***	8.79
DAMa	0.93***	93.07***	0.66***	2.72***	3.55
DAMi	0.90***	87.97***	1.19***	4.18***	6.67
DIA	0.66***	60.37***	0.32***	11.29***	28.02
DEP	0.44***	38.97***	0.08	26.38***	34.58
Obovoid	0.73***	71.42***	1.08***	8.57***	18.94
Ovoid	0.64***	61.48***	0.62***	15.49***	22.40
WWP	0.73***	69.26***	0.00	11.61***	19.14
Average	0.78	73.98	3.01	8.62	14.38

3.3. Modeling fruit shapes in eggplant using machine learning

A total of 1,344 fruits were manually classified into 13 distinct shape categories (Fig. 2). The first eight categories (C1-C8) correspond to the UPOV shape standards (UPOV, 2012), while the remaining five categories (C9-C13) represent newly identified shapes in eggplant (Fig. 2A and B). Fig. 2A shows the 13 prototypes, each plotted within a 1,600-pixel box (Table S3). According to Table 1, the minimum and maximum fruit height in the dataset ranges from 2.65 cm to 30.84 cm, while fruit widths range from 2.64 cm to 17.22 cm. We applied a decision tree model with Gini impurity splitting criterion to classify the fruits based on 22 shape attributes. The model selected 10 of these attributes (Fig. 2C) to generate classification rules. These attributes are iteratively selected by the Gini criterion which splits the data to minimize the impurity (or misclassification). The lower the Gini impurity, the better the split and hence the most important the attribute in classification. The top two most significant attributes were FSI and Circular, which together accounted for 66% of the classification importance (Fig. 2C). Additional important attributes included FST, Obovoid, WWP, and DFB, bringing the cumulative importance of these top six attributes to 92% (Fig. 2C). These results align with the correlation analysis, where FSI showed strong association with Circular (r = 0.88). The decision tree also selected four additional attributes (PAMi, PAMa, DAMa and DAMi) that are highly correlated with shapes (Fig. 1), confirming their relevance in distinguishing shape categories.

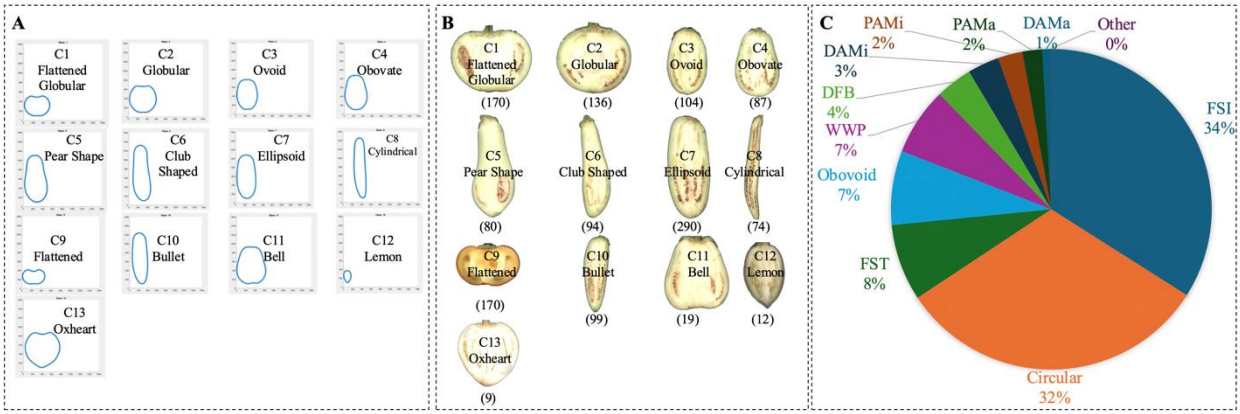

Fig. 2 The 13 fruit shape categories and the 10 significant attributes of fruit shape in eggplant. (A) Category prototypes of the 13 categories, with the axes for each prototype set to 1,600 pixels; (B) Representative fruits of the 13 categories. The number under the fruit represents the fruit number of each category; (C) The top 10 attributes and their significance identified using the decision tree.

Table 3 The confusion matrix for all data run through the pruned tree

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13
C1	160	1	0	0	0	0	0	0	8	0	0	0	1
C2	1	132	1	0	0	0	0	0	0	0	1	0	1
C3	0	0	100	2	0	0	0	0	0	0	1	1	0
C4	0	1	3	76	1	0	5	0	0	0	1	0	0
C5	0	0	0	8	64	1	5	0	0	2	0	0	0
C6	0	0	0	0	2	87	4	1	0	0	0	0	0
C7	0	0	0	4	2	5	276	0	0	3	0	0	0
C8	0	0	0	0	0	0	0	74	0	0	0	0	0
C9	6	0	0	0	0	0	0	0	164	0	0	0	0
C10	0	0	0	0	0	0	1	1	0	97	0	0	0
C11	1	0	0	1	0	0	0	0	0	0	17	0	0
C12	0	0	1	0	0	0	1	0	0	0	0	10	0
C13	0	0	0	0	0	0	0	0	0	0	0	0	9

Note: C1-C13 denote human labels, and G1-G13 denote the decision tree labels. The numbers on the second diagonal show the number of fruits correctly classified in each category.

The decision tree with Gini impurity splitting produced 25 classification rules with train data accuracy, test data accuracy and total accuracy of 94.79%, 92.59% and 94.35%, respectively (Fig. 3; Table 3; Appendix A). To improve generalization and reduce overfitting, we applied a rule-post pruning technique that simplifies rules without compromising accuracy. This technique removes parts of rule antecedents if accuracy remains the same or is minimally decreased (Appendix B). For example, the original rule for classifying C2 Globular "Width Widest Pos <= 0.5864 and Distal Angle Macro <= 128.6500 and Distal Fruit Blockiness < 0.8068 and Circular <= 0.0771 and Fruit Shape Index External I > 0.8900" is simplified to "Circular <= 0.0771". The pruned rule set achieved a total accuracy of 94.20%, a negligible decrease from the original 94.35%, while improving model simplicity and interpretability. The precision, recall and the F1 score (the harmonic mean of precision and recall) were also analyzed (Table S3). All categories achieve strong F1 scores above 0.85, with eight categories exciding 0.90, indicating both high precision and recall. In particular, categories C2 (Globular), C3 (Ovoid), C8 (Cylindrical), C9 (Flattened), C10 (Bullet) have F1 scores of 0.95 or higher (Table S3).

The test accuracies averaged over 5-fold cross-validation of the other methods were also investigated, including random forest, XG Boost, SVM, K-Means (k=13) and GMM (13 components). None of these methods surpassed the 92.59% test accuracy achieved by the decision tree described above (Fig. 3). Interestingly, the two methods (random forest and XG Boost) with performances closest to our decision tree are other decision tree variants. Therefore, although the decision tree

is a simpler model than any of these five more complex models, it is more practical because it not only produces robust classification, but also highlights which descriptors characterize each shape category.

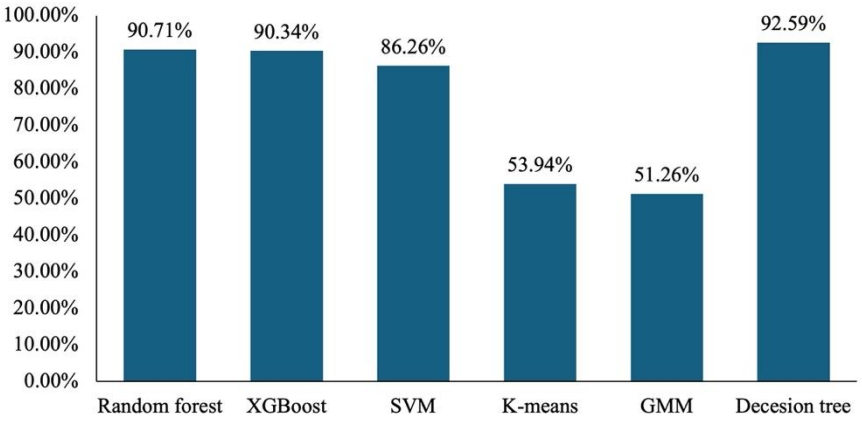

Fig. 3 Average test accuracies of five learning algorithms. The y axis indicates the mean 5-fold cross validation test accuracy

3.4. Comparison of attributes among shape categories

Next, we compared the values of the 10 key attributes across the 13 fruit shape categories. The fruits in C8 (Cylindrical) had the highest FSI, followed by C6 (Club Shaped), C7 (Ellipsoid) and C10 (Bullet), whereas the fruits in C1 (Flattened Globular) and C9 (Flattened) have the lowest FSI (Fig. 2; Fig. 4A). The Circular attribute measures how closely a fruit's shape approximates a circle, and it is calculated as the ratio between best-fit circle area and fruit area (Rodríguez *et al*. 2010). Thus, lower values of Circular indicate that the fruits are more circular. Interestingly, C6, C7, C8 and C10 also showed the highest Circular values (Fig. 4B), consistent with the strong positive correlation between FSI and Circular (r = 0.88) (Fig. 1). As expected, globular fruits in C2 had the lowest Circular values (Fig. 2; Fig. 4B). Fruits in C5 (Pear Shape), C6 and C11 (Bell shaped) have narrower upper and wider lower halves (Fig. 2), resulting in lower FST and higher values of Obovoid, WWP and DFB (Fig. 4C-F). This is consistent with the strong negative correlations between FST and Obovoid (r = -0.94), WWP (r = -0.94), and DFB (r = -0.91) (Fig. 1). Additionally, proximal or distal end angle attributes such as PAMi, PAMa, DAMi and DAMa, show strong negative correlations (r < -0.96) with FSI (Fig. 1). Therefore, fruits in C8 (highest FSI) had lowest values of PAMi, PAMa, DAMi and DAMa, while fruits in C9 (lowest FSI) show highest values of PAMi, PAMa, DAMi and DAMa (Fig. 4G-J).

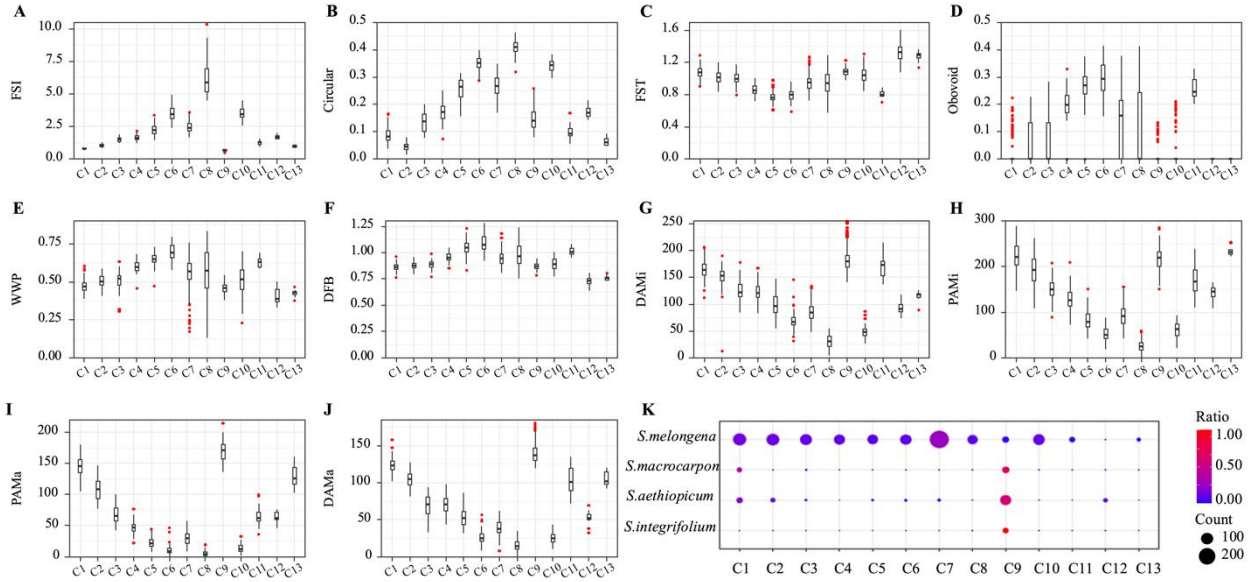

Fig. 4 Comparisons of the 10 attributes among the 13 categories and the distributions of the 13 categories in different species

3.5. Fruit shape categories in the three cultivated eggplant species

We also analyzed the fruit shape categories among the four eggplant species in the natural population. *S. melongena* the most common eggplant worldwide, exhibited the highest diversity, containing all categories except C12 (Lemon) (Fig. 4K). *S. aethiopicum* and *S. macrocarpon*, which are primarily cultivated in Africa (Taher *et al*. 2017), showed more limited shape diversity. *S. aethiopicum* included eight shape categories, with C9 (Flattened) being the most common (Fig. 4K). *S. macrocarpon* only produces Flattened Globular (C1) and Flattened (C9) fruits, and only flattened fruits (C9) were found in *S. integrifolium* (Fig. 4K).

3.6. Application of fruit shape classification in an F_2 population

An F_2 population was constructed by crossing inbred line '14-345' bearing flattened globular fruits (C1) with '16-562' which has ellipsoid fruits (C7) (Fig. 5A). The F_1 progeny exhibited ovoid fruits which were classified as Ovoid (C3) (Fig. 5A). A total of 414 fruits from 211 F_2 individuals were analyzed and grouped into seven categories using the shape classification model (Fig. 5B). The four most frequent classes were C3 (207 fruits, 50.0%), C7 (75 fruits, ~18.1%) and C2 (65 fruits, ~15.7%), and C4 (49 fruits, ~11.8%). Only five, six and seven fruits were classified into C1, C5 and C13, respectively (Fig. 5B). When compared to the human expert's labeling, the model achieved 95.17% classification accuracy (only misclassifications were one fruit in C3 Ovoid, four fruits in C4 Obovate, and five fruits in C7 Ellipsoid). These results further indicated the effectiveness of the fruit shape classification model for eggplant.

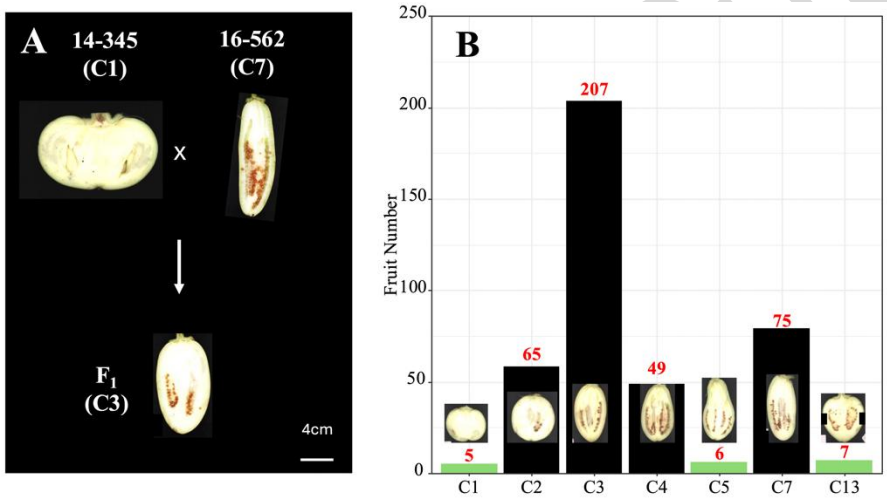

Fig. 5 Fruit shape classifications of '14-345', '16-562', and their F_1 and F_2 progenies

3.7. Fruit shape QTL and gene identification using GWAS and QTL-seq

Considering that the FSI is the most significant attribute for eggplant fruit shape (Fig. 2C), we conducted GWAS using FSI as the phenotypic data to identify QTLs in the population. The association results obtained using both MLM and FarmCPU models are shown in Manhattan plots (Fig. 6A-D). Both models consistently identified two significant loci at the distal ends of Chr03 and Chr09, labeled as *fsi3.1* and *fsi9.1*, respectively. The leading SNPs, Chr03: 86421155 and Chr09: 91217790, were selected by LD decay analysis (Table 4). The candidate regions associated with FSI span 3 Mb (86.1-89.1Mb) on Chr03 and 0.1Mb (91.1-91.2Mb) on Chr09 (Table 4; Table S4), encompassing 242 and four candidate genes, respectively (Table 4; Table S4). To further support these findings, we performed GWAS using PAMi, an attribute highly negatively correlated with FSI (Fig. 1). As expected, both MLM and FarmCPU identified a significant signal on Chr09 (Fig. 6E-H). Interestingly, the QTL of PAMi showed the same lead SNP and candidate region as those found for FSI on Chr09, supporting the reliability of our QTL mapping results.

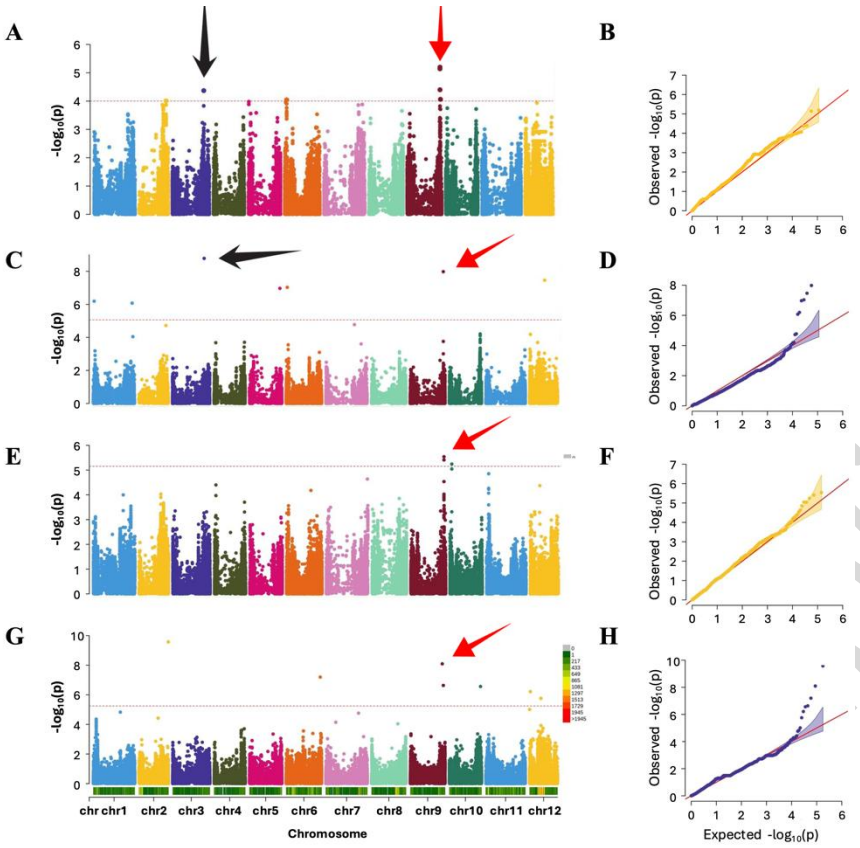

Fig. 6 GWAS for FSI and PAMi of eggplant. (A-B), Manhattan plot (A) and Q-Q plot (B) of GWAS of FSI using MLM model; (C-D) Manhattan plot (C) and Q-Q plot (D) of GWAS of FSI using FarmCPU model; (E-F), Manhattan plot (E) and Q-Q plot (F) of GWAS of PAMi using MLM model; (G-H), Manhattan plot (G) and Q-Q plot (H) of GWAS of FSI using FarmCPU model. Dashed horizontal line for each population indicates the suggestive threshold. The black and red arrows indicate the *fsi3.1* and *fsi9.1*, respectively.

In addition to GWAS, we used the F_2 population (Fig. 5) derived from '14-345' and '16-562' to identify the QTLs controlling the FSI. A total of 77 188 434, 81 701 802, 129 031 524 and 119 161 455 clean reads were generated from '14-345', '16-562', 'FSI_max' and 'FSI_min', respectively. On average, 96.60% of the clean reads from both parent lines and bulk samples were successfully mapped to the eggplant reference genome, covering 95.51% of the genome. A total of 1 017 127 high-quality SNPs were used for SNP-index analysis. The Δ (SNP-index) values revealed one major locus on Chr03 and two additional loci on Chr12 (Fig. 7; Table 4). The region on Chr03 spans 75.80-92.67Mb and overlaps with the identified *fsi3.1* region from the GWAS (Fig. 6). The two QTLs on Chr12 are located in the region of 5.56-11.10Mb

and 71.02-77.34Mb (Fig. 7; Table 4). Notably, the QTL on Chr03 was also identified as *fl3.1* in a previous study (Pang *et al.* 2024).

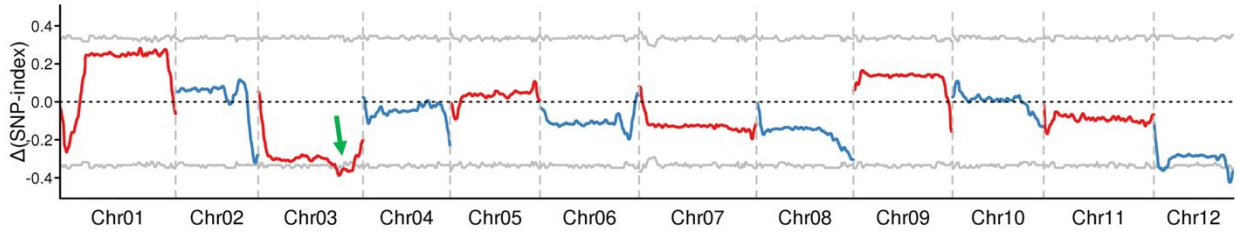

Fig. 7 QTL-seq showing the ΔSNP-index across all twelve chromosomes. Dashed gray lines indicate 95% confidence interval cut-offs of ΔSNP-index. The green arrow indicates the position of the QTL *fsi3.1*.

Table 4 Putative QTLs for FSI and PAMi and their position and related genes in HQ-V2.0 eggplant genome.

Trait	Method	Chromosome	Position	Lead SNP	Ref/Alt	Gene number	Cloned gene
FSI	QTL-seq	Chr03	75.80-92.67Mb	-	-	969	Sme03G2052
FSI	GWAS	Chr03	86.1-89.1Mb	86421155	G/T	242	Sme03G2052
FSI	GWAS	Chr09	91.1-91.2Mb	91217790	C/A	4	-
FSI	QTL-seq	Chr12	5.56-11.10Mb	-	-	183	-
FSI	QTL-seq	Chr12	71.02-77.34Mb	-	-	467	-
PAMi	GWAS	Chr09	91.1-91.2Mb	91217790	C/A	4	-

3.8. Overexpression of *SmeFL* induces elongated fruits in tomato

SmeFL, identified as the best candidate gene of *fl3.1*, belongs to the IQ67 DOMAIN (IQD) family which is widely known for regulating organ shape (Li *et al*. 2023; Xiao *et al*. 2008). However, the specific function of *SmeFL* remains to be validated. To investigate its role, we constructed an *SmFL* overexpression vector (Fig. 8A) and transformed it into tomato through Agrobacterium tumefaciens-mediated genetic transformation. Four independent *SmeFL* over-expression lines (OE2, OE4, OE6 and OE7) were obtained (Fig. 8B). These lines exhibited significantly elevated *SmeFL* expression compared to Microtom (MT) and the negative control (NC) (Fig. 8C-D). As expected, the four over-expression lines produced slender fruits with significantly increased fruit shape index compared to MT and NC (Fig. 8E). These results suggest that *SmeFL* plays a key role in regulating fruit shape in eggplant.

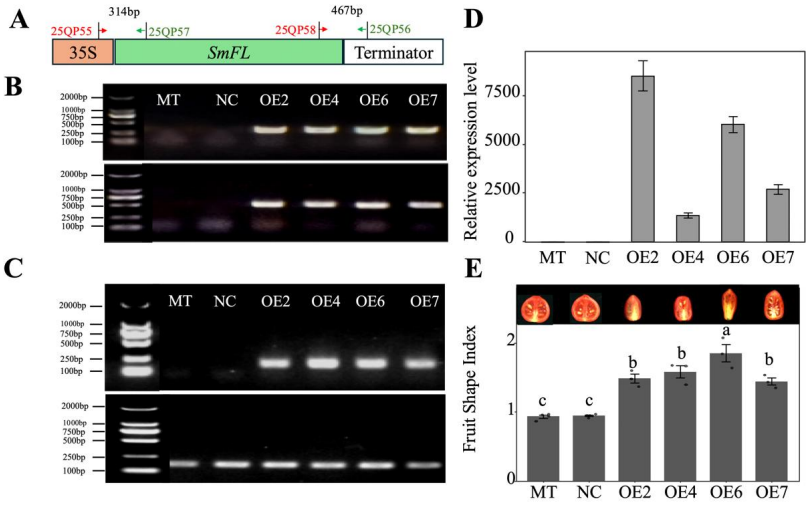

Fig. 8 Over-expression of *SmFL* produces elongated fruits in tomato. (A) A schematic diagram of the vector overexpressing *SmFL*. 25QP55/25QP57 and 25QP58/25QP56 are two primer pairs used for PCR verification of the vector; (B) PCR verification of transgenic lines using 25QP55/25QP57 (Upper panel) and 25QP58/25QP56 (Lower panel); (C) Expression analysis of *SmFL* using semi-qRT-PCR. Upper panel, amplification of *SmFL* using primers

25QP191/25QP192 (Table S2); Lower panel, *CAC* amplification using 24QP96/24QP97 (Table S2); (D) Relative transcript abundances of of *SmFL* examined by qRT-PCR; (E) Comparisons of fruit shape index among over-expression lines, MT and NC, the representative fruits were also shown above the corresponding bars. MT, Microtom; NC, Negative control; OE, over-expression.

4. Discussion

Fruit shape is a domesticated trait in many crops including eggplant that plays a key role in consumer preference, which varies by region and country. Among cultivated species, *S. melongena* shows high variation in fruit shape. Compared to tomato and strawberry, the accurate description and classification of eggplant fruit shapes remain underdeveloped. Previous studies mainly used interspecific and intraspecific F_2 populations for fruit shape measurement and genetic analysis (Mangino *et al*. 2021; Pang *et al*. 2024). In contrast, this study used a diverse natural population comprising 285 accessions for fruit shape evaluation using TA software (Table S1). As expected, a high diversity of fruit shape was identified in this natural population (Table 1), validating the suitability of the natural population for fruit shape analyses.

Fruit shape attributes were measured for the 285 accessions in both 2021 and 2023 (Table 1). Most attributes showed strong positive correlations between the two years (Table 2), and two-way ANOVA analysis confirmed that genotype was the primary contributor to shape variation across all 22 attributes (Table 2). These results further confirm that fruit shape is mainly determined by multiple genetic factors. Environmental factors, such as temperature and light, have been shown to influence fruit shape in plants, including tomato, pepper, cucumber, peach and citrus (Wert *et al*. 2007; Rajametov *et al*. 2021; Zhang *et al*. 2022; Gómez-Devia and Nevo, 2024; Yu *et al*. 2024b). In the present study, the environment significantly influenced most attributes, accounting for up to 19.18% (Table 2) of the variation. Obvious changes in fruit shape are common phenomenon in the production of eggplant. For example, in northern China, varieties that typically bear globular fruits often produce oval fruits in the autumn (Fig. S1), suggesting that lower temperatures promote fruits elongation. Further research is needed to elucidate the causal mechanisms linking environmental factors to change in fruit morphology.

Fruit shape is a key trait that influences consumer preference, thus accurate description and classification of fruit shape is of great value for crop breeding. We previously performed fruit shape modeling using elliptic Fourier attributes and Bayesian classification and identified nine distinct shape categories in tomato (Visa *et al*. 2014). In strawberry, a quantitative method for evaluation of strawberry fruit shape was established using Elliptic Fourier attributes (EFDs) (Nagamatsu *et al*. 2021). In contrast, the assessment and classification of eggplant fruit shapes has mainly relied on basic linear measurements such as length and width (Wei *et al*. 2020b). Moreover, existing fruit shape categories for eggplant are insufficient to represent all shape categories, and often these shape categories vary across regions and countries (Liu *et al*. 2019).

Machine learning modeling has been used in fruit classification due to its inherent nonlinear characteristics and flexibility (Mimma *et al*. 2022; Wang *et al*. 2022; Zakeri *et al*. 2021). For example, four principal categories of strawberry were discovered using unsupervised machine learning (Feldmann *et al*. 2020). In this study, we measured 22 shape attributes of 1,344 fruits using TA and applied a decision tree model for classification (Table S1; Fig. 2). Compared to the UPOV shape categories, our model identified five more categories (Fig. 2), capturing greater resolution. The present study is the first to integrate a natural population, an F_2 population, TA-generated attributes, and a decision tree model to enhance the precision of fruit shape classification in eggplant. Five other methods, including random forest, XGBoost, SVM, K-means and GMM, were also applied to fruit shape classification. Random forest and XGBoost achieved average accuracies of about 90.00%, while SVM, K-means and GMM reached 86.26%, 53.94% and 51.26%, respectively. For practical usability, we provide in the appendix the rule set of a simple decision tree of 92.59% accuracy.

The decision tree has many advantages including the ability to model complex patterns and interactions between attributes and to handle well outliers (Sarker 2021). Further, decision trees inherently rank features by importance and deliver if-then decision rules that are easy to understand and interpret. These strengths make them particularly suitable for modeling shapes and complement the other analyses used in this research. Therefore, the decision tree model is a simple and practical method in producing robust classification and determining key attributes characterizing each shape category.

Fruit shape genes have been widely studied in plants and tomato is one of the model plants to study fruit shape regulation (Snouffer *et al*. 2020; Li *et al*. 2023). OVATE Family Protein (OFP), TONNEAU1 Recruiting Motif (TRM) and IQ67

domain (IQD) family genes have been implicated in regulating fruit shape in many plants. In tomato, OVATE, a founding member of OFP family, acts as a repressor of fruit elongation, especially at the proximal end (Liu *et al*. 2002). SlTRM5 and SlTRM19 are a microtubule associated proteins, while null mutant of SlTRM5 produces a slighter flatten fruits (Wu *et al*. 2018), knockout SlTRM19 resulted in elongated fruits (Zhang *et al*. 2023), indicating the contrary effects of SlTRM5 and SlTRM19 in regulating fruit elongation. OFP and TRM proteins, such as SlOVATE and SlTRM5, can physically interact with each other to form OFP-TRM module through the M8 motif of TRM protein. The module controls fruit shape through influencing the microtubule organization and cell division patterns (Zhang *et al*. 2023). Tomato *SUN* is a member of IQD gene family, over-expression of *SlSUN* in tomato resulted in extremely long fruits, indicating the positive roles of *SlSUN* in regulating fruit elongation. Rice *GW5* encodes an IQD protein, and loss-of-function of GW5 significantly increased grain width through affecting cell number (Duan *et al*. 2017; Liu *et al*. 2017). In cucumber, a SNP (G651A) within the IQ domain of *CsSUN* lead to reduced fruit length and increased width (Zhang *et al*. 2024).

Numerous QTLs controlling fruit length, width, and shape index have been identified in eggplant (Portis *et al*. 2014; Liu *et al*. 2019; Wei *et al*. 2020a; Barchi *et al*. 2021; Mangino *et al*. 2021). Notably, a QTL on Chr03 associated with fruit shape has been reported in many studies. For example, *fs3.1*, which accounted for 10.92% of the FSI variation, was identified using an F_2 population derived from a cross between cultivated eggplant '1836' and the wild relative *S. linnaeanum* '1809' (Wei *et al*. 2020a). Recently, Pang *et al*. (2024) identified a fruit length-related locus, *fl3.1*, in three F_2 populations, and proposed *SmeFL*, an IQD family protein, as the best candidate gene. However, until now, none of these candidate genes have been functionally verified through stable genetic transformation. In this study, we identified the QTL on Chr03 using GWAS and QTL-seq, providing strong support for its role in controlling fruit shape. Moreover, over-expression of *SmeFL* in tomato leads to elongated fruits, reinforcing the hypothesis that *SmeFL* (*Sme03G2052*) is the causal gene underlying *fsi3.1*/*fl3.1*. Nevertheless, further functional validation in eggplant such as over-expression and genome editing is necessary to confirm the role of *SmeFL*.

5. Conclusion

In this study, twenty-two fruit shape attributes were evaluated using a natural eggplant population with high morphological diversity. An effective decision tree model was developed for eggplant fruit shape classification, resulting in the identification of thirteen distinct shape categories, ten key attributes that largely determine fruit shape, and highly accurate classification rules (92.59%). Other machine learning models were also tested, but they proved less robust for classification compared to the decision tree. While *S. melongena* showed high variation across 12 fruit shape categories, the other two cultivated species, *S. aethiopicum* and *S. macrocarpon*, predominantly produced fruits in the Flattened Globular (C1) and Flattened (C9) categories. Four QTLs influencing FSI, the most significant attribute determining fruit shape, were detected on Chr03, Chr09, and Chr12 through GWAS and QTL-seq analyses. Overexpression of the candidate gene *SmFSI3.1*/*SmFL*, a member of the SUN/IQD family, in tomato resulted in elongated fruits, confirming its regulatory role in fruit shape. Overall, this study not only established a quantitative model for fruit shape classification but also identified candidate QTLs and genes that contribute to fruit shape regulation in eggplant.

Acknowledgments

This work was supported by the Hebei Provincial Natural Science Foundation for Distinguished Young Scholars (C2023204028), Youth Special Program for Basic Research in Biological Breeding of the National Natural Science Foundation of China (32441074), Science Research Project of Hebei Education Department (ZD2022111), Hebei Agriculture Research System (HBCT2023100207) and S&T Program of Hebei (21326309D).

Declaration of generative Al and Al-assisted technologies in the writing process

The authors declare that they did not use AI in the preparation and writing of this manuscript.

Declaration of competing interest

The authors declare that they have no competing interests

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request

Supplementary materials

Fig. S1 Representative fruits of a variety 'Nongda604' in spring (Mid-May) (A) and autumn (Late October) (B and C)

Table S1 List of the 285 accessions and the number of plants and fruits used for fruit shape modeling

Table S2 Primers used in this study

Table S3 Precision, recall, and the harmonic mean of precision and recall for each class

Table S4 Information of genes in the candidate regions of *fsi3.1* and *fsi9.1*

Appendix A If-then rules for each of the thirteen classes

Appendix B Pruned if-then rules for each of the thirteen classes

Appendices associated with this paper are available on http://www.ChinaAgriSci.com/V2/En/appendix.htm

References

Barchi L, Rabanus-Wallace M T, Prohens J, Toppino L, Padmarasu S, Portis E, Rotino G L, Stein N, Lanteri S, Giuliano G. 2021. Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. *The Plant Journal*, **107**, 579-596.

Duan P, Xu J, Zeng D, Zhang B, Geng M, Zhang G, Huang K, Huang L, Xu R, Ge S, Qian Q, Li Y. 2017. Natural variation in the promoter of *GSE5* contributes to grain size diversity in rice. *Molecular Plant*, **10**, 685-694.

FAO. 2023. FAOSTAT. https://www.fao.org/faostat/en/#data.

Feldmann M J, Hardigan M A, Famula R A, Lopez C M, Tabb A, Cole G S, Knapp S J. 2020. Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry. *GigaScience*, **9**, 1-17.

Gómez-Devia L, Nevo O. 2024. Effects of temperature gradient on functional fruit traits: an elevation-for-temperature approach. *BMC Ecology and Evolution*, **24**, 94.

Gonzalo M J, Brewer M T, Anderson C, Sullivan D, Gray S, van der Knaap E. 2009. Tomato fruit shape analysis using morphometric and morphology attributes implemented in Tomato Analyzer software program. *Journal of the American Society for Horticultural Science*, **134**,77-87.

Hurtado M, Vilanova S, Plazas M, Gramazio P, Herraiz F J, Andújar I, Prohens J. 2013. Phenomics of fruit shape in eggplant (*Solanum melongena* L.) using Tomato Analyzer software. *Scientia Horticulturae,***164**, 625-632.

International Board for Plant Genetic Resources (IBPGR). 1990. Descriptors for eggplant. https://alliancebioversityciat.org/publications-data/descriptors-eggplant

Kaur A, Gill K S, Malhotra S, Devliyal S. 2024. Automated fruit classification using KNN and decision trees for enhanced agricultural efficiency and accuracy. 4th Asian Conference on Innovation in Technology (ASIANCON), 1-5.

Kaushik P, Prohens J, Vilanova S, Gramazio P, Plazas M. 2016. Phenotyping of eggplant wild relatives and interspecific hybrids with conventional and phenomics descriptors provides insight for their potential utilization in breeding. *Frontiers in Plant Science*, **7**, 677.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* **26**, 589-595.

Li, Q, Feng Q, Snouffer A, Zhang B, Rodriguez G R, van der Knaap E. 2022. Increasing fruit weight by editing a cis-regulatory element in tomato KLUH promoter using CRISPR/Cas9. *Frontiers in Plant Science*, **13**, 879642.

Li Q, Luo S, Zhang L, Feng Q, Song L, Sapkota M, Xuan S, Wang Y, Zhao J, van der Knaap E, Chen X, Shen S. 2023. Molecular and genetic regulations of fleshy fruit shape and lessons from Arabidopsis and rice. *Horticulture Research*, **10**, uhad108.

Li X, Zhu D. 2006. *Descriptive Specifications and Data Standards for Eggplant Germplasm Resources*. China Agriculture Press, Beijing, China. (in Chinese).

Liu J, Van Eck J, Cong B, Tanksley S D. 2002. A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proceedings of the National Academy of Sciences*, 99, 13302-13306.

Liu J, Chen J, Zheng X, Wu F, Lin Q, Heng Y, Tian P, Cheng Z, Yu X, Zhou K, Zhang X, Guo X, Wang J, Wang H, Wan J. 2017. GW5 acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. *Nature Plants*, **3**, 17043.

Liu W, Qian Z, Zhang J, Yang J, Wu M, Barchi L, Zhao H, Sun H, Cui Y, Wen C. 2019. Impact of fruit shape selection on genetic structure and diversity uncovered from genome-wide perfect SNPs genotyping in eggplant. *Molecular Breeding,* **39**, 140.

Mangino G, Vilanova S, Plazas M, Prohens J, Gramazio P. 2021. Fruit shape morphometric analysis and QTL detection in a set of eggplant introgression lines. *Scientia Horticulturae*, **282**, 110006.

Mansfeld B N, Grumet R. 2018. QTLseqr: an R package for bulk segregant analysis with next‐generation sequencing. *The Plant Genome*, **11**, 180006.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297-1303.

Mimma N E A, Ahmed S, Rahman T, Khan R, Tang Z. 2022. Fruits classification and detection application using deep learning. *Scientific Programming,* **2022,** 1-16.

Nagamatsu S, Tsubone M, Wada T, Oku K, Mori M, Hirata C, Hayashi A, Tanabata T, Isobe S, Takata K, Shimomura K. 2021. Strawberry fruit shape: quantification by image analysis and QTL detection by genome-wide association analysis. *Breeding Science*, **71**, 167-175.

Nankar A N, Tringovska I, Grozeva S, Ganeva D, Kostova D. 2020. Tomato phenotypic diversity determined by combined approaches of conventional and high-throughput tomato analyzer phenotyping. *Plants*, **9**, 197

Pang H, Ai J, Wang W, Hu T, Hu H, Wang J, Yan Y, Wu X, Bao C, Wei Q. 2024. Fine mapping of QTL *fl3.1* reveal *SmeFL* as the candidate gene regulating fruit length in eggplant (*Solanum melongena* L.). *Vegetable Research*, **4**, e028.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*,**12**, 2825-2830.

Pereira-Dias L, Fita A, Vilanova S, Sánchez-López E, Rodríguez-Burruezo A. 2020. Phenomics of elite heirlooms of peppers (*Capsicum annuum* L.) from the Spanish centre of diversity: Conventional and high-throughput digital tools towards varietal typification. *Scientia Horticulturae*, **265**, 109245

Portis E, Barchi L, Toppino L, Lanteri S, Acciarri N, Felicioni N, Fusari F, Barbierato V, Cericola F, Vale G, Rotino G L. 2014. QTL mapping in eggplant reveals clusters of yield-related loci and orthology with the tomato genome. *PLoS One*, **9**, e89499.

Quispe-Choque G, Rojas-Ledezma S, Maydana-Marca A. 2022. Morphological diversity determination of the tomato fruit collection (*Solanum lycopersicum* L.) by phenotyping based on digital images. *Journal of the Selva Andina Research Society,* **13**, 51-68.

Rajametov S N, Lee K, Jeong H B, Cho M C, Nam C W, Yang E Y. 2021. The effect of night low temperature on agronomical traits of thirty-nine pepper accessions (*Capsicum annuum* L.). *Agronomy*, **11**, 1986.

Rodriguez G R, Munos S, Anderson C, Sim S C, Michel A, Causse M, Gardener B B, Francis D, van der Knaap E. 2011. Distribution of *SUN*, *OVATE*, *LC*, and *FAS* in the tomato germplasm and the relationship to fruit shape diversity. *Plant Physiology,* **156**, 275-285.

Rodríguez G R, Moyseenko J B, Robbins M D, Morejón N H, Francis D M, van der Knaap E. 2010. Tomato analyzer: a useful software application to collect accurate and detailed morphological and colorimetric data from two-dimensional objects. *Journal of Visualized Experiments*, **37**, e1856.

Sarker I H. 2021. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science,* **2**, 160.

Snouffer A, Kraus C, van der Knaap E. 2020. The shape of things to come: ovate family proteins regulate plant organ shape. *Current Opinion in Plant Biology*, **53**, 98-105.

Taher D, Solberg S O, Prohens J, Chou Y Y, Rakha M, Wu T H. 2017. World vegetable center eggplant collection: origin, composition, seed dissemination and utilization in breeding. *Frontiers in Plant Science*, **8**, 1484.

Tripodi P, Greco B. 2018. Large scale phenotyping provides insight into the diversity of vegetative and reproductive organs in a wide collection of wild and domesticated peppers (*Capsicum* spp.). *Plants*, **7**, 103

UPOV. Eggplant: Guidelines for the conduct of tests for Distinctness, uniformity and stability, 2012

Visa S, Cao C, Gardener B M,v an der Knaap E. 2014. Modeling of tomato fruits into nine shape categories using elliptic fourier shape modeling and Bayesian classification of contour morphometric data. *Euphytica*, **200**, 429-439.

Wang X, Yan M, Wang X, Wu Z, Zhou J, Wang C, Chen R, Qin X, Yang H, Wei H, Gu W. 2022. The phenotypic diversity of *Schisandra sphenanthera* fruit and SVR model for phenotype forecasting. *Industrial Crops and Products*, **186**, 115162.

Wei Q, Wang W, Hu T, Hu H, Wang J, Bao C. 2020a. Construction of a SNP-based genetic map using SLAF-Seq and QTL analysis of morphological traits in eggplant. *Frontiers in Genetics*, **11**, 178.

Wei Q, Wang J, Wang W, Hu T, Hu H, Bao C. 2020b. A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant. *Horticulture Research*, **7**,153.

Wei Q, Wang W, Wang Y, Ai J, Hu T, Hu H, Wang J, Yan Y, Pang H, Hu N, Bao, C. 2025. A complete telomere-to-telomere genome assembly of Solanum melongena uncovers key regulators in pan-tissue anthocyanin biosynthesis. *Plant Communications*, **6**, 101533.

Wert T W, WilliamsonJ G, Chaparro J X, Miller E P, Rouse R E. 2007. The influence of climate on fruit shape of four low-chill peach cultivars. *HortScience* ,**42**, 1589-1591.

Wu S, Zhang B, Keyhaninejad N, Rodriguez G R, Kim H J, Chakrabarti M, Illa-Berenguer E,Taitano N K , Gonzalo M J, Diaz A, Pan Y, Leisner C P, Halterman D, Buell C R, Weng Y, Jansky S H ,van Eck H, Willemsen J, Monforte A J, Meulia T, van der Knaap E. 2018. A common genetic mechanism underlies morphological diversity in fruits and other plant organs. *Nature Communications*, **9**, 4734.

Xiao H, Jiang N, Schaffner E, Stockinger E J, Van Der Knaap E. 2008. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, **319**, 1527-1530.

Yu C, Yang Q, Li W, Jiang Y, Gan G, Cai L, Li X, Li Z, Li W, Zou M, Yang Y, Wang Y. 2024a. Development of a 50K SNP array for whole-genome analysis and its application in the genetic localization of eggplant (*Solanum melongena* L.) fruit shape. *Frontiers in Plant Science*, **15**,1492242.

Yu X, Du C, Wang X, Gao F, Lu J, Di X, Zhuang X, Cheng C, Yao F. 2024b. Multivariate analysis between environmental factors and fruit quality of citrus at the core navel orange-producing area in China. *Frontiers in Plant Science*, **15**, 1510827.

Zakeri A, Hedayati R, Khedmati M, Taghipour-Gorjikolaie M. 2021. Classification of jujube fruit based on several pricing factors using machine learning methods. *Computer Vision and Pattern Recognition,* arXiv: 2111.00112.

Zeng Y, Shi J, Ji Z, Wen Z, Liang Y, C. Yang. 2017. Genotype by environment interaction: the greatest obstacle in precise determination of rice sheath blight resistance in the field. *Plant Disease*, **101**, 1795-1801.

Zhang B, Li Q, Keyhaninejad N, Taitano N, Sapkota M, Snouffer A, van der Knaap E. 2023. A combinatorial TRM-OFP module bilaterally fine-tunes tomato fruit shape. *New Phytologist*, **238**, 2393-2409.

Zhang T, Hong Y, Zhang X, Yuan X, Chen S. 2022. Relationship between key environmental factors and the architecture of fruit shape and size in near-isogenic lines of cucumber (*Cucumis sativus* L.). *International Journal of Molecular Sciences*, **23**, 14033.

Zhang Z, Zhang H, Liu J, Chen K, Wang Y, Zhang G, Li L, Yue H, Weng Y, Li Y. 2024. The mutation of *CsSUN*, an IQD family protein, is responsible for the *short and fat fruit* (*sff*) in cucumber (*Cucumis sativus* L.). *Plant Science*, **346**, 112177.

Zhu Q, Deng L, Chen J, Rodríguez G R, Sun C, Chang Z, Yang T, Zhai H, Jiang H, Topcu Y, Francis D, Hutton S, Sun L, Li C B, van der Knaap E, Li C. 2023. Redesigning the tomato fruit shape for mechanized production. *Nature Plants*, **9**, 1659-1674.