nature genetics

Article

https://doi.org/10.1038/s41588-025-02401-0

Genetic basis of flavor complexity in sweet corn

Received: 1 February 2023

Accepted: 9 October 2025

Published online: 7 November 2025

Check for updates

Kun Li[1,2,8], Yongtao Yu[1,8], Shijuan Yan[3,8], Wenqiang Li[2,4,8], Jieting Xu[2,5], Gaoke Li[1], Wu Li[1], Jianhua Liu[1], Xitao Qi[1], Wenjie Huang[3], Qunjie Zhang[3], Qian Kong[3], Yingni Xiao[1], Nan Zhang[1], Jingyun Luo[2,4], Lu Chen[2], Liying Feng[6], Wenguang Zhu[1], Tianxiang Wen[1], Lihua Xie[1], Yuliang Li[1], Wenjia Lu[1], Chunyan Li[1], Songtao Gui[2,4], Yingjie Xiao[2,4], Ning Yang[2,4], Lin Zhuo[2,4], Alisdair R. Fernie[7], Hai-Jun Liu[6] ✉, Jianguang Hu[1] ✉ & Jianbing Yan[2,4] ✉

Sweet corn is an important vegetable crop consumed globally. However, the genetic differentiation between field corn and sweet corn, and the impact of breeding on the metabolite composition and flavor (other than sweetness) of sweet corn, remain poorly understood. Here we assembled a cultivated sweet-corn genome de novo and re-sequenced 295 diverse sweet-corn inbred lines. We examined the genetic architecture of sweet-corn kernel quality by combining genetic, metabolite and expression profiling methodologies. New genes (for example, *ZmAPS1*, *ZmSK1* and *ZmCRR5*) and metabolites associated with flavor and consumer preference were identified, highlighting important target flavor metabolites, including sugars, acids and volatiles. These findings provide valuable knowledge and targets for future genetic breeding of sweet-corn flavor, and to balance grain yield and quality and contribute to our broader understanding of crop diversification.

Sweet corn is a corn variety containing defective alleles of starch synthesis genes, such as shrunken2 (*sh2*) and sugary1 (*su1*), and has become an important vegetable and fruit crop globally[1]. The sweet-corn industry has achieved enormous economic value, for example, generating over US$774 million in the United States in 2021[1]. Sweet corn probably originated from a spontaneous mutation in an ancient Peruvian corn, which was preserved by Native American tribes. The first historical reference to sweet corn was to an event at which the Iroquois gave the sweet-corn 'Papoon' to European settlers in 1779[2]. Since then, sweet corn has undergone marked improvement via selective breeding and has been illuminated at the genomic level[3–5].

Over the past decades, at least eight genes have been used in sweet-corn breeding programs, with the *sh2* allele being the most successful, followed by the combination of *su1* and sugary enhancer1 (*se1*)[3]. The defective alleles in starch synthesis genes cause sweet corn to lose 50–70% of the starch in the endosperm, which is critical for germination and seedling development[6,7]. In addition to mutations affecting sweetness, some genes influencing other flavor qualities and economic traits, such as volatile emissions and pericarp thickness and texture, have been identified[7]. Simultaneous selection across breeding programs has made sweet-corn breeding quite distinct from that of field corn[8–10]. However, the genetics underlying quality traits like flavor remain inadequately understood[11].

Corn is widely recognized as an important model crop for studying cereal evolution, metabolic pathways and quality improvement[12–16]. However, the germplasm pool for sweet corn remains largely unexplored. Extensive knowledge in genomics, transcriptomics and metabolomics of field corn would not only have laid the technical foundation

[1]Crop Research Institute, Guangdong Academy of Agricultural Sciences, Guangdong Key Laboratory of Crop Genetic Improvement, Guangzhou, China. [2]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China. [3]Guangdong Key Laboratory for Crop Germplasm Resources Preservation and Utilization, Agro-biological Gene Research Center, Guangdong Academy of Agricultural Sciences, Guangzhou, China. [4]Hubei Hongshan Laboratory, Wuhan, China. [5]WIMI Biotechnology Co. Ltd, Qingdao, China. [6]Yazhouwan National Laboratory, Sanya, China. [7]Department of Molecular Physiology, Max-Planck-Institute of Molecular Plant Physiology, Potsdam-Golm, Germany. [8]These authors contributed equally: Kun Li, Yongtao Yu, Shijuan Yan, Wenqiang Li. ✉e-mail: liuhaijun@yzwlab.cn; hujianguang@gdaas.cn; yjianbing@mail.hzau.edu.cn

but also have provided a reference for the dissection of sweet-corn improvement[17–20]. In the current work, we generated a large dataset, including a high-quality, de novo, sweet-corn genome assembly, and transcriptomic and metabolomic evaluations of kernel tissues at two distinct harvest times from a population of 295 diverse sweet-corn inbred lines. Furthermore, we have assessed the flavor quality of these lines at the fresh stage. Using a metabolite genome-wide association study (mGWAS), we proposed candidate genes for flavor-associated metabolites and validated five of them via clustered regularly interspaced short palindromic repeats (CRISPR)–Cas9. This study provides major insights into sweet-corn genome features, kernel quality formation mechanisms and the balance of consumer quality and grain yield. In doing so, our study has provided valuable resources for sweet-corn biology, as well as presenting a model for repurposing major crops.

Results

Population genomics of sweet corn

The *sh2-R* allele is widely used in modern super-sweet-corn commercial breeding programs. Its structural variation was initially documented in a recent report profiting from the high-quality genome assembly of Ia453 (ref. 5). To have a better understanding of the origin and formation of this structural variation and further sweet-corn genome architecture, we sequenced a white super-sweet-corn genome, named RC, which was widely used in Chinese sweet-corn breeding programs. The genome size of RC was assembled at 2,171.95 Mb with the N50 values of contig and scaffold as 51.84 Mb and 220.16 Mb, respectively (Supplementary Tables 1 and 2). A total of 2,113.72 Mb (97.32%) of the assembled genome sequence was anchored on to 10 chromosomes. Using Benchmarking Universal Single-Copy Orthologs (BUSCO), we assessed genome completeness and found that 98.7% of conserved BUSCO genes aligned to the RC genome. Furthermore, quality metrics indicated a consensus quality value (QV) of 29.21 and a mean long terminal repeat (LTR) assembly index (LAI) score of 26.27 (Supplementary Tables 3). Employing homology prediction from 5 closely related species together with transcriptome data derived from 10 tissues, we have annotated a total of 43,023 gene models (Supplementary Tables 4). Furthermore, RC-specific genome sequences of 23.1–32.2 Mb were observed compared with B73 and other sweet-corn genomes (Extended Data Fig. 1a–h).

To further study the population genomics of sweet corn, we collected a total of 295 sweet-corn accessions from Asia, America and Africa (Supplementary Table 5). These included 213 super-sweet corn (*sh2*-R), 17 ordinary sweet corn (*su1*), 9 strengthened sweet corn (*su1-se1*) and 56 double-recessive sweet-waxy corn (*wx1-sh2*). These accessions represent the major races of sweet-corn germplasm and consumption types (Fig. 1a). We obtained 7.3 TB of whole-genome sequencing (WGS) data from this population, with an average depth of 11.8-fold (ranging from 10-fold to 12-fold) genome coverage and heterozygosity ranging from 3.65% to 5.93% (average 5.26%; Supplementary Table 5). By mapping the WGS to our RC genome, we generated a variant set with 9.9 million high-quality single nucleotide polymorphisms (SNPs), averaging 4.6 SNPs per kb.

For a comprehensive comparison, we reanalyzed our WGS data of 507 diverse field-corn inbreds from temperate, subtropical and tropical regions[21–23], using the same pipeline. Clear group differences between the sweet-corn and field-corn populations were observed (Fig. 1b–d and Extended Data Fig. 1i). Although the genetic diversity was comparable among sweet, temperate and tropical populations, sweet corn showed greater population differentiation compared to that between temperate and tropical populations. In addition, sweet corn exhibited slower linkage disequilibrium (LD) decay than field maize, which may be derived from population history; for example, it could be caused by a stronger bottleneck (Fig. 1c). This is also reflected in the higher number of large LD blocks (>900 kb) in sweet corn (SC, 245) compared to field corn (164), with 85 (34.7%) unique to sweet corn. These characteristics collectively suggest that sweet corn has undergone a unique breeding selection process.

Taking the distinctive differences in temperate and tropical germplasm into account, we conducted crosspopulation composite likelihood ratio (XP-CLR) and crosspopulation extended haplotype homozygosity (XP-EHH) analyses to identify regions potentially under positive selection in the sweet-corn population, comparing sweet corn to temperate (TEM) field corn (SC versus TEM) and sweet corn versus tropical and subtropical (TST) field corn (SC versus TST) (Fig. 1e–h). We identified 6,098 regions of selection in the top 5% of the XP-CLR values from the two analyses and 3,126 regions from 2 XP-EHH analyses (the adjacent 2 were merged; Supplementary Tables 6 and 7). These regions cover approximately 10.3% of the maize genome, with a mean length of ~30.3 kb. Strikingly, 75% (64 put of 85) of the large LD blocks unique to sweet corn overlapped with these selection signals, a significantly higher proportion than would be expected by chance ($P < 0.01$, Fisher's exact test). A total of 6,975 genes in these regions were considered as candidate genes, 5.7% (400) of which were identified in both XP-CLR and XP-EHH analyses. The differences between sweet corn and field corn are evident not only in carbohydrate metabolism (for example, carbohydrate metabolic process) but also in nucleotide metabolism regulation (false discovery rate (FDR) <0.05; Supplementary Table 8).

The role of *Sh2* and *Su1* in shaping sweet corn

Our newly assembled genome includes the complete sequence of *sh2-R*, which shows 95% identity with the Ia453 sequence (another recently assembled genome)[5] in the *a1-sh2* region. All super-sweet-corn lines in our current study harbor the *sh2-R* allele. We next compared our RC genome with 38 published regular field maize and teosinte genomes (https://maizegdb.org/) and discovered an intermediate inversion in the *Sh2* region, shared among all super-sweet corns and present only in the genome of field corn Tx303 (Extended Data Fig. 2a). The gene model of the *sh2-R*$_{RC}$ allele likely derives from the *Sh2*$_{Tx303}$ allele, which in turn probably originated from the *Sh2*$_{TIL11}$ allele, consistent with the above finding of the inversion (Extended Data Fig. 2a). The unique inversion appears to have become fixed relatively recently, as supported by the likely single origin for the *sh2-R*$_{RC}$ allele, the stronger LD block and reduced π value observed in this region (Extended Data Fig. 2b–f).

Another important gene for sweet corn is *Su1*, which encodes a starch debranching enzyme that is essential for normal starch granule production and has rarely been studied previously. The defective allele of *Su1* results in decreased starch and increased sugar and water-soluble polysaccharides, producing a creamy palatability[7]. Multiple defective *su1* alleles (*su1-ref*, *su1-sw*, *su1-nc* and so on) were independently isolated by indigenous peoples[4,24]. Recent haplotype analysis demonstrated two radically distinct SNP patterns between sweet-corn (*su1*) and field-corn (*Su1*) populations[5]. In our sweet-corn population, a higher π value was observed than in field corn at the *su1* region, identifying five different alleles based on deep-sequencing data. The two dominant alleles, *Su1-RC* (*Su1* allele from RC lines with four nonsynonymous substitutions) and *Su1-KE* (with a Lys-to-Glu substitution at position 707), were found among these (Extended Data Fig. 2g–k). These five alleles cluster into three groups based on the SNPs in the *Su1* region (Extended Data Fig. 2l).

As expected, the knockout lines of *Sh2* and *Su1* displayed shrunken and well-marked phenotypes similar to those of *sh2-R* and *su1* types, respectively (Extended Data Fig. 3a–c). Combined -omics analyses revealed common effects of *sh2* and *su1* knockouts on the transcriptome and metabolome (Extended Data Fig. 3d–h). Specifically, they increased levels of several sugars (for example, sucrose and rhamnose) and decreased levels of certain amino acids (for example, asparagine and ornithine). RNA sequencing (RNA-seq) analysis revealed 171 differentially expressed genes (DEGs) common to both knockouts,

https://doi.org/10.1038/s41588-025-02401-0

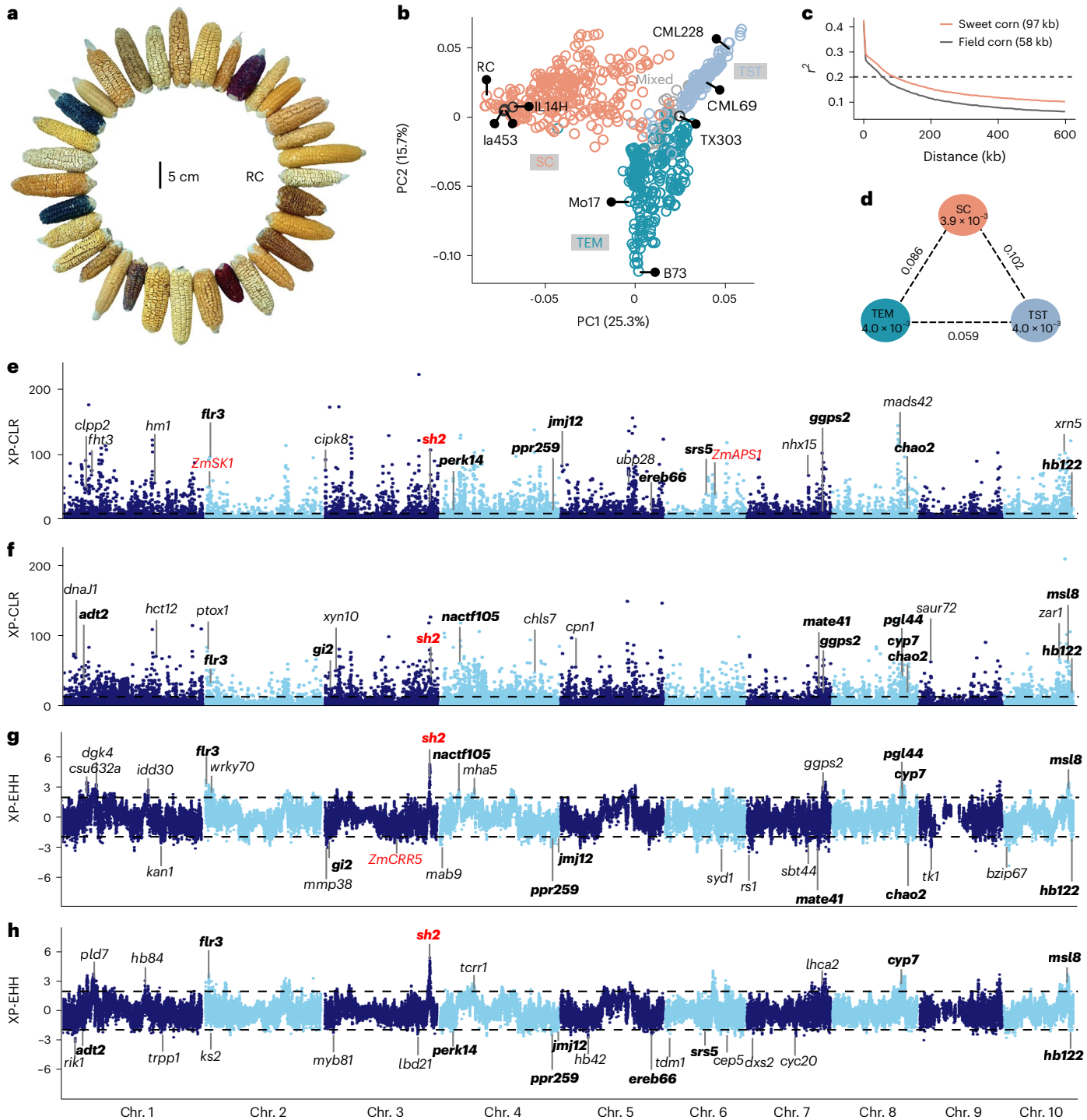


Fig. 1 | Genetic diversity between sweet-corn and field-corn populations. **a**, Dry ears from a subset of the sweet-corn population showing rich diversity. **b**, Principal component analysis (PCA) of sweet corn and field corn, revealing three identifiable subpopulations: TEM ($n = 247$) field corn, TST ($n = 221$) field corn and SC ($n = 295$). RC is a sweet-corn line with a high-quality genome assembled in the present study. Several maize lines with genome assembled are indicated by labels. The mixed group ($n = 39$) has unclear tropical and temperate assignments. **c**, Genome-wide averaged distance where LD decays to $r^2 = 0.2$ for sweet-corn and field-corn populations. **d**, Values in each circle representing nucleotide diversity (π) for each group. The values on each line represent pairwise population divergence (F_{st}) between groups. **e**,**f**, Detection of selection signatures (by XP-CLR) between the sweet-corn and the TEM population (**e**) and between the sweet corn and the TST population (**f**). **g**,**h**, Detection of selection signatures (measured by XP-EHH) between sweet corn and the TEM (**g**) and TST (**h**) populations, respectively. Genes in bold are identified by both XP-CLR and XP-EHH analyses and those in red represent the key genes highlighted in this study.

which were enriched in response to abiotic stimuli and hormones (Supplementary Tables 9 and 10), together with differentially abundant metabolites, reflecting the response of the plant to an obstruction of starch synthesis. Moreover, 205 DEGs directly affected by *sh2* or *su1* were also discovered in the selection analyses (XP-CLR and/or XP-EHH), significantly more overrepresented than expected (permutation test, $P < 0.05$). These findings suggest that the two mutations have multifaceted influence beyond sweetness, providing a model to study how complex interplay between mutated genes impacts downstream regulatory networks and the overall biochemical landscape.

https://doi.org/10.1038/s41588-025-02401-0

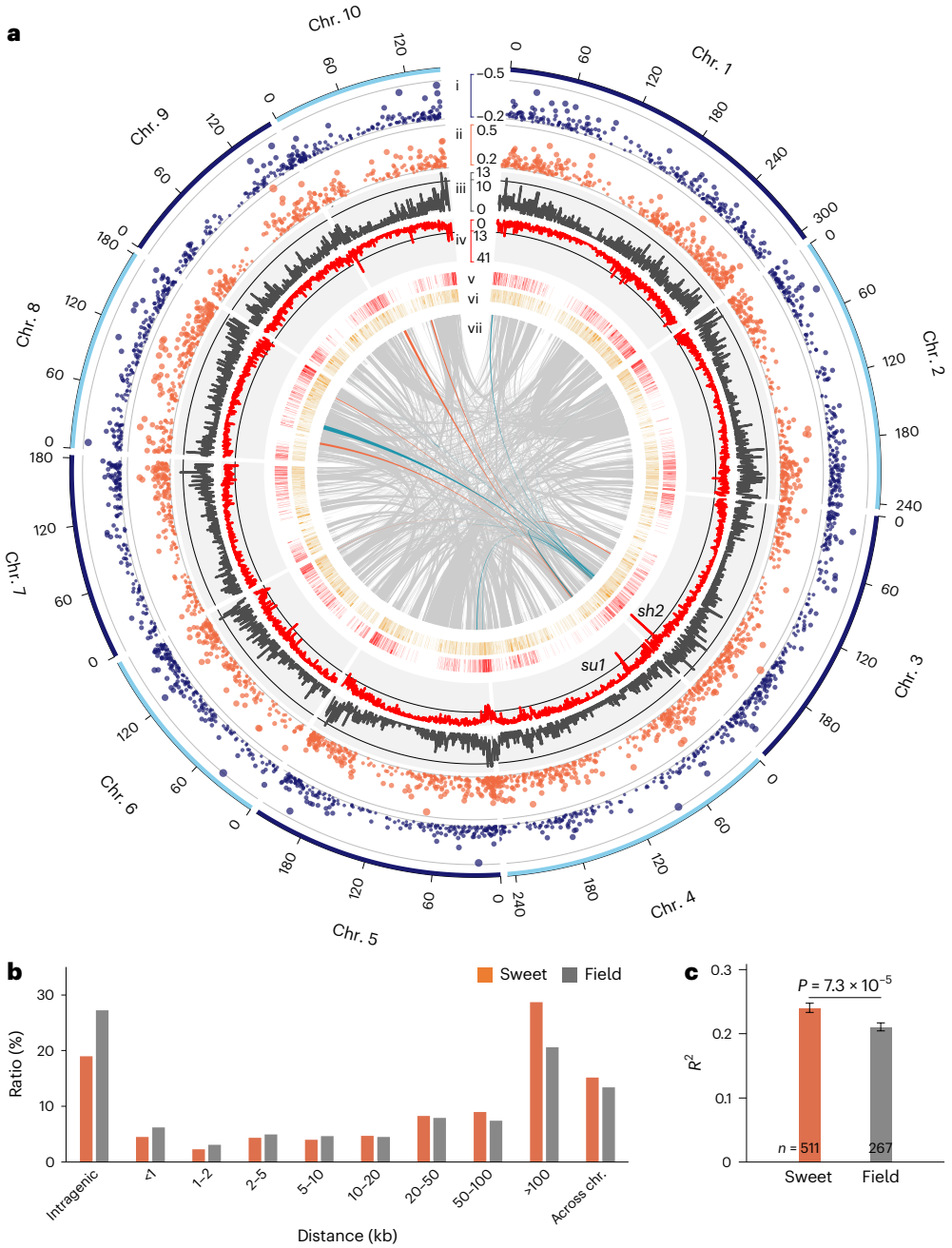


Fig. 2 | Whole-genome expression analyses. a, (i) Gene expressions negatively correlated with flavor ratings; (ii) gene expressions positively correlated with flavor ratings; (iii) eGWAS distribution for the field-corn population; (iv) eGWAS distribution for the sweet-corn population; (v, vi) genes with *cis*-eQTLs (iv) and *trans*-eQTLs (vi) across the maize genome (an eQTL was classified as *cis* here if the lead SNP was located within 2 kb of the gene that it regulates; otherwise, it was classified as *trans*); and (vii) genes with *trans*-eQTLs between chromosomes. **b**, Summary of distances between eQTL lead SNPs and corresponding genes in field-corn and sweet-corn populations. **c**, Comparison of eQTL effects on DEGs in field-corn and sweet-corn populations. R^2 indicates the phenotypic variation explained by the lead SNP of each eQTL. The differences between groups were assessed using a two-tailed, unpaired Student's *t*-test.

Rewired transcriptome regulation in sweet corn

To clarify the transcriptome characteristics of sweet corn, we performed RNA-seq analysis on the immature kernels (at 15 d after pollination (DAP)) of 280 sweet-corn lines. We identified 20,073 genes as expressed and used them for subsequent analyses. In total, 13,080 significant expression quantitative trait loci (eQTLs) involving 10,712 genes were detected (Fig. 2a, Extended Data Fig. 4 and Supplementary Table 11). Comparing this eQTL dataset with a reanalysis of previous RNA-seq data from the aforementioned field-corn population[22] under the same workflow, we found that the sweet-corn analysis unveiled nearly all of the previously reported eQTLs plus an additional 845 unique to sweet corn. Although analyzed in a smaller population size (280 in sweet corn versus 342 in field corn), sweet corn has more expressed genes (18,972 in field corn) and stronger transcriptional regulatory effects (an average eQTL effect of 0.25 versus 0.22 in field corn) in immature kernel tissue.

It is interesting that our analysis of the sweet-corn population revealed a greater prevalence of *trans*-eQTLs, suggesting that gene expression in sweet corn is predominantly influenced by distantly located regulatory elements, including those residing >100 kb away or on different chromosomes (Fig. 2b). Why this happens can be explained by the observation, from the *sh2* and *su1* knockout lines, that most DEGs are upregulated and more eQTLs were identified and

https://doi.org/10.1038/s41588-025-02401-0

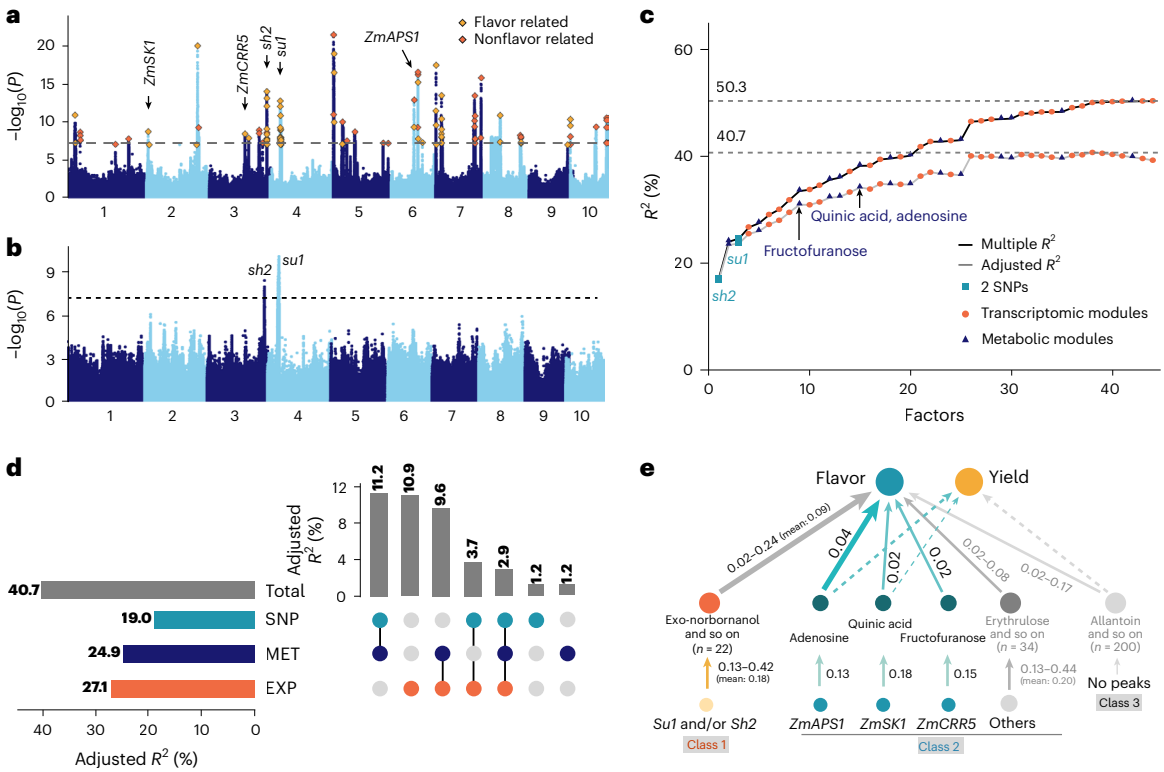


Fig. 3 | Genetics basis of sweet-corn flavor. a, Association for structurally annotated metabolites from GC–MS or LC–MS. Those metabolites showing significant correlation with flavor-rating traits are named as flavor related. **b**, GWAS for flavor rating. **c**, Regression analyses correlating flavor ratings with genomic (SNP), transcriptomic and metabolomic factors. R^2 indicates phenotypic variation, calculated using linear regression analysis. **d**, UpSet plot showing flavor-rating variation (adjusted R^2) by 44 identified factors from **c**. MET, metabolome; EXP, transcriptome. **e**, Summary of genetic and metabolomic factors affecting sweet-corn flavor and yield. The numbers linking genetic factors to metabolites indicate the variation explained by lead SNPs of the corresponding candidate genes. The numbers connecting metabolites to flavor and yield are variations explained by linear regression, with line width indicating significance as detailed in Supplementary Table 18. Class 1 includes metabolites uniquely associated with *Sh2* and/or *Su1* and class 3 comprises metabolites with no significant peaks identified. Although class 2 consists of metabolites correlated with flavor or yield that have new peaks, cyan highlights the three genes and their corresponding metabolites chosen for in-depth study.

together explained higher gene expression variation for these DEGs in the sweet-corn population (Fig. 2c). Specifically, *cis*-eQTLs exhibited larger effects (coefficient of determination, R^2) than the *trans*-eQTLs (phenotypic variation explained, 0.27 versus 0.24 on average, $P < 0.001$; Supplementary Table 11), consistent with the study in field corn[22].

The identified eQTLs showed a significant uneven distribution across the genome, with 41 hotspots in sweet corn compared to 12 in field corn ($P < 0.05$ by permutation test; Fig. 2a and Extended Data Fig. 4). The eQTLs of DEGs from *sh2* and *su1* knockout lines cover 35 of the 41 eQTL hotspots, predominantly *trans*-regulatory. These 35 eQTL hotspots are specific to the sweet-corn population and absent in field corn. Gene ontology (GO) analysis revealed that genes in these eQTL hotspots were significantly enriched in chloroplast components and oxidoreduction metabolic processes (false discovery rate (FDR) <0.05). This suggests that *sh2* and *su1* mutations significantly influence the expression of many other genes across different chromosomes.

Genetic landscape of metabolite regulation in sweet corn

To investigate the biochemical basis of sweet-corn flavor, we performed metabolomic analyses of 20-DAP kernels using gas chromatography–mass spectrometry (GC–MS) and liquid chromatography–MS (LC–MS) techniques. We identified 234 metabolites from GC–MS and 260 from LC–MS with known molecular structures (Supplementary Table 12).

A total of 101 metabolite (m)QTLs ($P \leq 3.95 \times 10^{-8}$) were identified with high mapping resolution (Fig. 3a, Extended Data Fig. 5a,b and Supplementary Table 13). Most of the mQTLs had major effects, with 68.3% showing $R^2 > 0.15$. Such mQTLs were also involved in multiple levels of sweet-corn growth and development. Major GWAS signals for metabolites were involved in glycometabolism (including erythrose, rhamnose, maltose, raffinose and xylose) and amino acid metabolism pathways (including those for glutamate and phenylalanine).

In sweet corn, defective alleles of genes associated with starch synthesis result in excess sugar accumulation in kernels, which may in turn influence other metabolic pathways. The two key genes, *sh2* and *su1*, crucial for sweet corn formation, were located in metabolite (m) GWAS hotspots (Fig. 3a). In particular, 22 metabolites were significantly associated with SNPs near *sh2* and/or *su1*, with 10 metabolites shared (Supplementary Table 13). Furthermore, conditional association analysis revealed that the association of *sh2* (lead SNP: S3_224136540) with these ten metabolites was dependent on their association with *su1* (lead SNP: S4_43073379; LD $r^2 = 0.89$). This suggests limited genetic exchange between super-sweet and ordinary sweet corn and, previously unnoticed, that *sh2* and *su1* share similar biological effects on the metabolome, as illustrated by metabolic changes in *sh2* and *su1* knockout lines (Supplementary Table 14).

Multifactorial determinants of flavor perception in sweet corn

Sweet-corn flavor is a complex trait consisting of taste, mouthfeel and aroma. To have a relatively objective evaluation of this comprehensive metric, we further quantified each accession for flavor attributes including sweetness (taste), brittleness and pericarp thickness (mouthfeel) and volatiles (aroma) of the sweet-corn population. We organized two experiments assessing these quantitative metrics in 2017 and 2019 (Methods and Supplementary Table 15).

https://doi.org/10.1038/s41588-025-02401-0

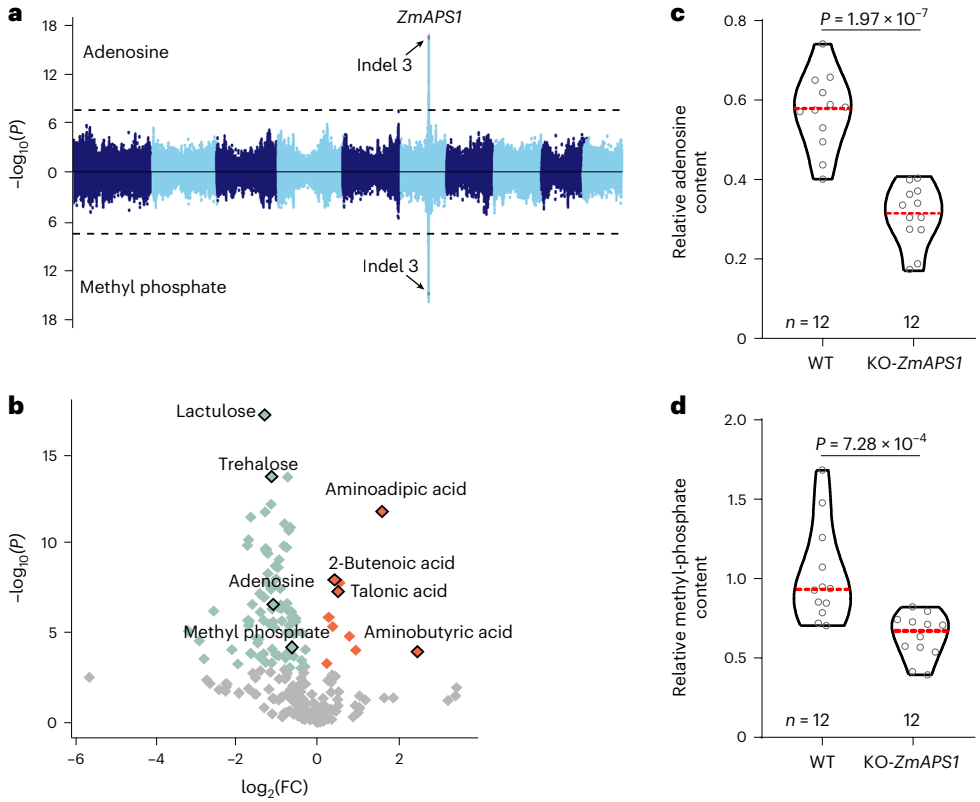


Fig. 4 | *ZmAPS1* functional validation. a, Manhattan plot of GWASs for adenosine and methyl-phosphate content. **b**, Comparison of the metabolome between mutant *ZmAPS1* and WT plants ($n = 12$ of 12 for knockout (KO) and WT). **c**,**d**, Violin plots of adenosine (**c**) and methyl-phosphate (**d**) levels in WT and KO *ZmAPS1* lines (two-tailed, unpaired Student's *t*-test was used). FC, fold-change.

A high correlation was observed between the two experiments (Extended Data Fig. 5c,d) with a broad-sense heritability (H^2) of 0.67. Furthermore, in 2019, we invited 95 ordinary consumers to score the 18 typical lines chosen by the expert flavor experiments (Methods). The correlation coefficient between the average consumer score and the best linear unbiased prediction (BLUP) values from 2017 and 2019 was 0.81 ($P < 4.6 \times 10^{-5}$). These results confirm that our flavor evaluation of the sweet-corn population is reliable and representative of the general public.

GWASs for flavor ratings and individual components (including brittleness, sweetness, pericarp thickness and volatiles) identified *sh2* and *su1* as the two strongest association signals (Fig. 3b and Supplementary Table 16). The colocalization between these two signals and mQTLs from metabolite mapping provides strong evidence for dissecting the genetic basis of sweet-corn flavor quality (Extended Data Fig. 5a,b). It is interesting that the *Su1-RC* allele, despite having lower expression and conferring lower sugar content (Extended Data Figs. 2j and 6a–d), was associated with higher flavor ratings than the *Su1-B73* allele (Extended Data Fig. 6e,f). Perhaps surprisingly, the discrepancy in sweetness between the two allele types was imperceptible to flavor-rating evaluators, with thinner pericarp thickness being the greatest contributor to flavor-quality variation instead (Extended Data Figs. 2j,k and 6g,h). These statistical analyses suggest that this locus is not under positive selection, but may represent a potential candidate for future breeding. As expected, knockout lines of *su1* and *sh2* had higher pericarp thickness values (approximately 5%) compared to wild-types (Extended Data Fig. 3i), confirming the association between sweetness and pericarp thickness. This suggests that, in the *sh2*-type sweet-corn background, modulating *Su1* expression can improve texture by reducing pericarp thickness without sacrificing perceived sweetness.

To model the complex architecture of flavor, we performed a multi-omics regression analysis (Fig. 3c, Supplementary Table 17 and Methods). First, we identified two representative SNPs within the confidence interval of flavor-score GWAS results across the genome (Fig. 3c). Correlation analyses revealed the expression of 4,711 genes and quantification of 139 metabolites to be significantly correlated with flavor ratings (FDR < 0.05; Supplementary Table 18). Positive correlations included sucrose, exo-norbornanol, heptanoic acid, sorbopyranose and quinolinecarboxaldehyde, whereas negative correlations included maltose, xylose, allantoin, glyoxime and glutamine. Second, we conducted genome-wide (9.93 million SNPs), transcriptome-wide (of 20,073 genes) and metabolome-wide (494 metabolites) analyses to predict flavor ratings. Genome-wide SNPs explained 42.0–73.7% of flavor-rating variance (using the GCTA[25] GREML module; Supplementary Table 17), whereas the top two GWAS SNPs alone explained 19.0% (adjusted R^2; Fig. 3d). Transcriptome and metabolite variations explained 27.1% and 24.9% of flavor variation, respectively, and combined could explain up to 39.5% (Fig. 3d and Methods). The total -omics variations explained as much as 40.7% of total flavor-score variation, including the top 2 SNPs, 28 gene expression modules and 14 metabolite clusters. This highlights the multifaceted genetic and biochemical regulation of this critical trait. In addition, we identified 4,711 genes directly or indirectly relevant to sweet-corn flavor, with 26.0% (1,227 genes) located in selection regions between sweet corn and field corn, which is significantly higher than expected (permutation test, $P < 1.0 \times 10^{-5}$).

Balancing sweet-corn flavor and yield

Although understanding that the genetic architecture of complex flavor-related traits is crucial, cloning the key underlying genes allows for more efficient and precise targeted improvements. Many

https://doi.org/10.1038/s41588-025-02401-0

metabolites (22) that are highly correlated with flavor traits are either only (except rhamnose) influenced by known genes (for example, *sh2*, *su1*; class 1 in Fig. 3e; Supplementary Table 18) or largely polygenic or highly affected by environmental factors with no significant loci identified in our association analysis (class 3 in Fig. 3e; Extended Data Fig. 5b). However, multi-omics data generated in this study provided numerous leads for further gene discovery (Extended Data Fig. 5b). We prioritized the investigation of metabolites identified with new loci, especially those with potential links to yield or other developmental traits, even if their flavor associations are moderate (class 2 in Fig. 3e).

The gene *ZmAPS1* (acid phosphatase 1; Zm00001eb277460) was identified based on association analysis of adenosine and methyl-phosphate content (Fig. 4a). Adenosine was additionally found to be significantly correlated with flavor ratings (FDR < 0.05; Supplementary Table 18). We identified an insertion and/or deletion (indel) (−/ACC) in the 3′-UTR of *ZmAPS1* that was significantly associated with both adenosine and methyl-phosphate content (Extended Data Fig. 7a–c). To verify the function of the *ZmAPS1*, we created mutants in the first exon of this gene using CRISPR–Cas9 (Extended Data Fig. 7a). Knockout lines of *ZmAPS1* exhibited significantly lower adenosine and methyl-phosphate content in the kernel (Fig. 4b–d), directly validating its function.

Metabolites are crucial for plant composition, growth and development. Consequently, variation in flavor-related metabolites may also impact the yield of sweet corn. Approximately 7.5% (37) of detected metabolites were significantly correlated with grain yield and half of these (18, including quinic acid, cellobiose, fumaric acid and sorbitol) were also significantly correlated with flavor scores (Supplementary Table 18). Quinic acid, derived from the shikimate pathway, is a common substrate in the biosynthesis of the three essential aromatic amino acids: tryptophan, phenylalanine and tyrosine. It has been revealed as an important organic acid in sweet corn, with an antioxidant activity, and as a precursor to aromatic amino acids[26]. Our analysis showed that quinic acid is slightly positively associated with pericarp thickness ($r = 0.13$, $P = 3.6 \times 10^{-2}$; Supplementary Table 18), negatively affecting flavor.

A major mQTL for quinic acid was identified on chromosome 2, with *ZmSK1* (Zm00001eb069000) the putative underlying gene (Fig. 5a,b), which encodes shikimate kinase 1 and catalyzes the conversion of shikimate to shikimate 3-phosphate. An indel (TTATTGCC/−) near the lead SNP (S2_6412295) is significantly associated with both the quinic acid content and *ZmSK1* expression (Fig. 5c,d). Quinic acid content negatively correlates with *ZmSK1* expression ($P = 1.04 \times 10^{-2}$) and positively correlates with yield-related traits (for example, ear weight and hundred-kernel weight, $P = 3.38 \times 10^{-3}$ and 8.08×10^{-5}, respectively; Fig. 5e and Extended Data Fig. 8a,b). However, higher hundred-kernel weight is linked to a thicker pericarp ($P = 3.61 \times 10^{-4}$; Extended Data Fig. 8c) that may diminish flavor perception, highlighting the trade-off between yield and quality. Supporting these, metabolite profiles of *ZmSK1* knockout mutants displayed more increased quinic acid levels than wild-types (WTs), although the overall metabolome remained unchanged (Fig. 5f and Extended Data Fig. 8d). Furthermore, higher quinic acid content in *ZmSK1* knockout lines was correlated with increased grain yield (Fig. 5g and Extended Data Fig. 8e). Co-expressed genes of *ZmSK1* (both positively and negatively correlated, $n = 142$ or 59, FDR < 0.001; Supplementary Table 19) in immature kernel tissue were enriched in ribosome-related, xylan metabolic and cell-wall polysaccharide biosynthetic processes (GO, FDR < 0.05; Supplementary Table 20) and potentially enriched in carbon fixation in photosynthesis (Kyoto Encyclopedia of Genes and Genomes (KEGG), FDR < 0.05; Supplementary Table 21). This reflects that *ZmSK1* mediates the close interrelationship of quinic acid, pericarp thickness and grain yield.

Fructose, including fructofuranose and fructopyranose, is a major soluble monosaccharide in plants, with approximately 93% present

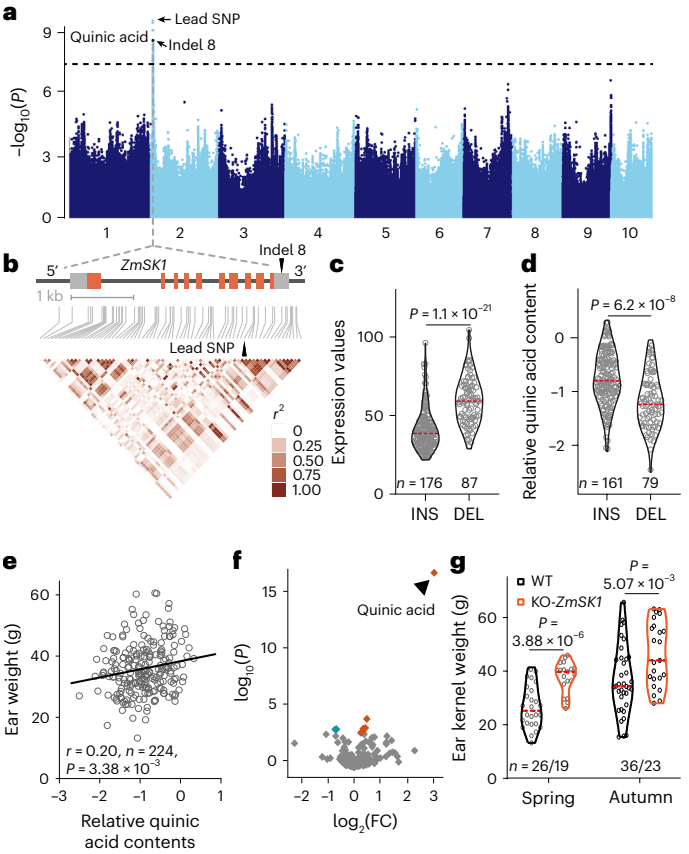


Fig. 5 | *ZmSK1* functional validation. a, Manhattan plot of the GWAS on quinic acid content. **b**, Gene model of *ZmSK1* (Zm00001eb069000), with exons and UTRs represented by filled orange and gray boxes, respectively. Pairwise LD r^2 values among polymorphic sites near *ZmSK1* were shown. **c**,**d**, Violin plot of *ZmSK1* expression values (**c**) and quinic acid content (**d**), plotted against genotypes at indel 8 (TTATTGCC/−). **e**, Correlation analysis between quinic acid content and ear weight. Pearson's correlation coefficient (r) and its associated P value were calculated using a two-tailed test. **f**, Metabolomic comparison between *ZmSK1* mutant and WT plants ($n = 9$ and 12 for KO and WT, respectively). **g**, Violin plots comparing ear kernel weight between WT and *ZmSK1* KO lines across two growing seasons. Group comparisons were performed using two-tailed, unpaired Student's t-tests. DEL, deletion; INS, insertion.

as fructofuranose in sweet-corn kernels (Supplementary Table 12). It plays a critical role in signal transduction and plant growth, as seen in *Arabidopsis* root development[27]. Our study found that fructofuranose was moderately correlated with flavor ratings (such as volatiles) at 20 DAP (Supplementary Table 18). A GWAS identified a peak (SNP S3_153509969; $P = 8.57 \times 10^{-9}$) within *ZmCRR5* (cytokinin response regulator 5; Zm00001d042066) (Fig. 6a,b), encoding a glycoside hydrolase, with a strong correlation ($r = 0.30$; $P = 1.17 \times 10^{-6}$) between its expression and fructofuranose content (Fig. 6c).

In *ZmCRR5* knockout lines, fructofuranose and 5 other metabolites significantly decreased (Fig. 6d,e and Extended Data Fig. 9a), whereas 37 metabolites (including monosaccharides, oligosaccharides, organic acids and amino acids) increased (Fig. 6e). Of these, 30 were also correlated with flavor ratings: 14 enhanced sweet-corn flavor, whereas the rest were unfavorable, highlighting *ZmCRR5*'s key role in flavor formation (Fig. 6e and Supplementary Table 18). Surprisingly, *ZmCRR5* knockout lines displayed no significant changes in above-ground plant architecture (Fig. 6f,g), but affected root and yield traits. Root number decreased by approximately 31% at the mature stage (Extended Data Fig. 9b,c), and a strong correlation was observed between lateral roots and fructofuranose content at the seedling stage ($P = 1.72 \times 10^{-3}$; Extended Data Fig. 9d).

https://doi.org/10.1038/s41588-025-02401-0

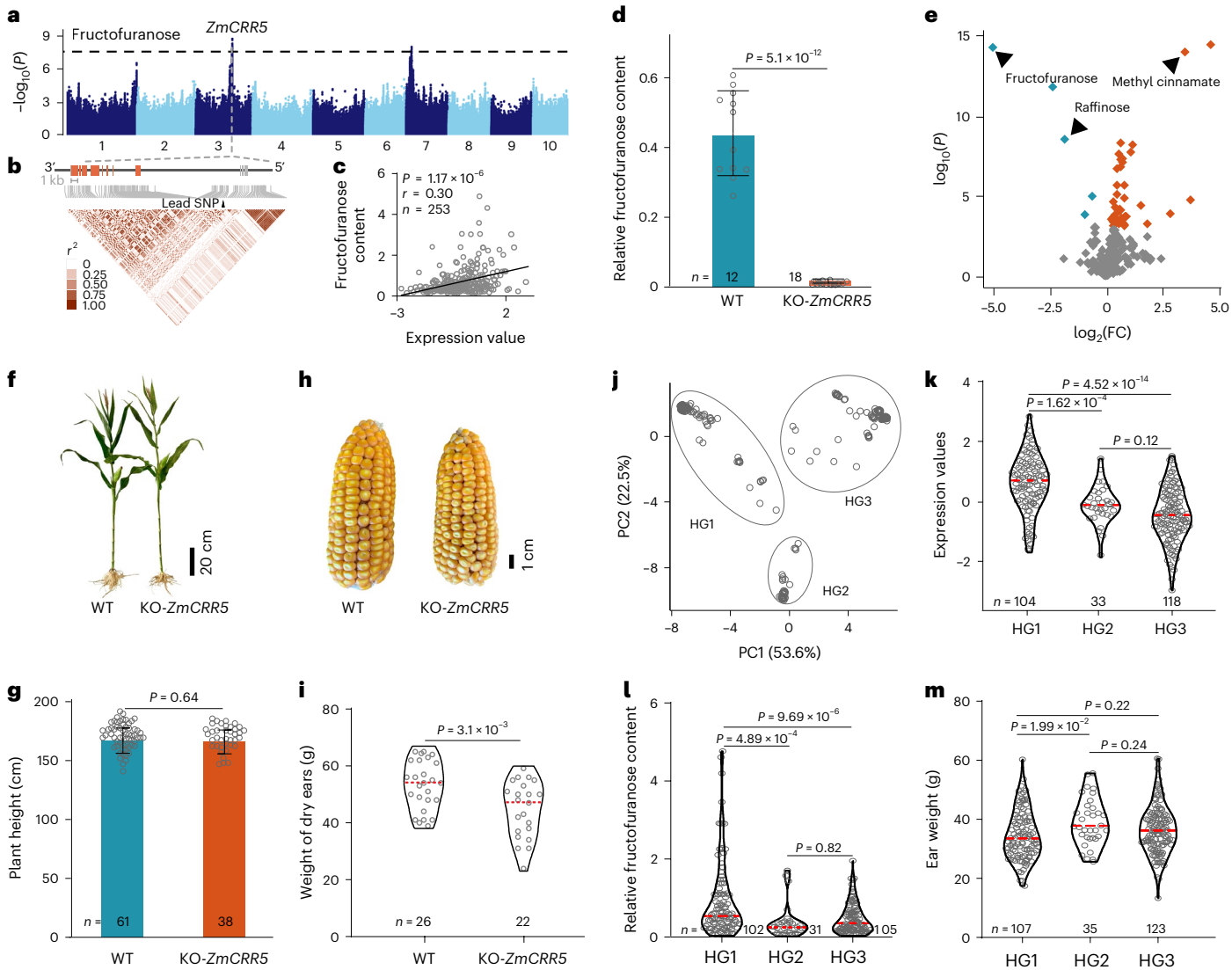


Fig. 6 | Functional validation of *ZmCRR5*. **a**, Manhattan plot of GWAS results on fructofuranose content. **b**, Gene model of *ZmCRR5* (Zm00001d042066), with exons and UTRs represented by filled orange and gray boxes, respectively. Pairwise LD r^2 values among polymorphic sites near *ZmCRR5* are shown. **c**, Correlation between fructofuranose content and *ZmCRR5* expression. **d**, Comparison of fructofuranose content between *zmcrr5* and WT plants. For mutants in the first and/or second exons, the sample size (*n*) is indicated. **e**, Metabolomic comparison between *zmcrr5* and WT plants ($n = 18$ and 12 for KO and WT, respectively). **f**, Representative images of WT and KO *ZmCRR5* maize plants. **g**, Comparison of plant heights between WT and KO *ZmCRR5* lines. **h**, Representative images of maize ears from WT and KO *ZmCRR5* lines. **i**, Dry weight of maize ears of WT and KO *ZmCRR5* lines. **j**, Haplotype groups (HG1–3) defined by PCA of 271 SNPs within the *ZmCRR5* region. **k**–**m**, Violin plots of *ZmCRR5* expression (**k**), relative fructofuranose content (**l**) and ear weight (**m**) across the three haplotype groups. Individual data points are shown in **d** and **g**, with error bars representing the mean ± s.d. Differences among groups (**k**–**m**) were determined by a one-way analysis of variance followed by Tukey's multiple-comparison test.

The kernel yield of *ZmCRR5* knockout lines decreased by 16%, with cytokinin components decreased by 25.2–54.9% (on average, 42.5%) in immature kernels compared to WT (Fig. 6h,i, Extended Data Fig. 9e and Supplementary Table 22).

As *ZmCRR5* knockout causes severe phenotypes, we explored natural allelic variations to improve flavor quality without adverse agronomic traits. In the sweet corn, 3 *ZmCRR5* haplotype groups (HG1–3) were observed in 269 lines within the *sh2-R* background (Fig. 6j and Extended Data Fig. 10a). HG2, with moderate *ZmCRR5* expression, exhibited the lowest fructofuranose content, thinner pericarp and highest yield (Fig. 6k–m and Extended Data Fig. 10b,c), suggesting that fine-tuning *ZmCRR5* can balance pericarp thickness and yield. It is interesting that further analysis revealed that two CRR homologs were upregulated in the shoot apical meristem, but not in immature kernels (*ZmCRR1* and *ZmCRR2*; Extended Data Fig. 9f–h). This gene family thus provides valuable targets for the precise design of high-yield and high-quality sweet corn.

Knockout of the above three genes (*ZmAPS1*, *ZmSK1* and *ZmCRR5*) successfully altered corresponding metabolite levels, thereby validating the approach to functional gene prioritization based on public information and multi-omics data generated in this study (Extended Data Fig. 5b). However, this was not always the case. For instance, two additional candidate genes (Zm00001eb211960 and Zm00001eb405310) proposed for erythrose and DL-2-amino-octanoic acid, selected based on the same principles, did not exhibit the anticipated phenotypic changes on knockout (Extended Data Fig. 7d–g).

Discussion

Our multi-omics dataset revealed a surprising genetic divergence between sweet corn and field corn, greater than that between tropical

and temperate maize[28]. This deep divergence is the result of a distinct evolutionary history[4,29] and strong positive selection of genes crucial for carbohydrate metabolism, as well as light and hormone responses. Although sweet corn's precise origins remain unresolved, our findings confirm that it has been on a separate trajectory, developing unique transcriptional networks to adapt to the high-sugar, low-starch environment created by mutations like *sh2* and *su1*.

A primary achievement of this study is the creation of a predictive framework for sweet-corn flavor. By integrating large-scale sensory evaluations with multi-omics data, we have moved beyond analyzing single components to modeling a complex, consumer-relevant trait[30–32]. This framework identified numerous loci that make flavor breeding more predictable and uncovered key relationships, such as the inverse correlation between sweetness and pericarp thickness. Our analysis shows that improving mouthfeel by reducing pericarp thickness can enhance overall flavor perception even without increasing sugar content, offering a new strategy for breeding.

This work also directly confronts the classic trade-off between flavor and yield. We demonstrated that these traits are often negatively correlated, exemplified by the gene *ZmSK1*, where alleles that increase yield also increase quinic acid and pericarp thickness, thereby diminishing flavor. However, our discovery of natural variation in genes like *ZmAPS1* and *ZmCRR5* shows that this trade-off is not absolute. Specific alleles of *ZmCRR5*, for instance, can balance high yield with desirable flavor attributes. These findings provide concrete genetic targets for uncoupling unfavorable linkages, enabling the creation of tailored alleles through gene editing to simultaneously improve both quality and agronomic performance.

Although our multi-omics workflow successfully identified and validated the function of three new genes (*ZmAPS1*, *ZmSK1* and *ZmCRR5*), it is not infallible. The failure to confirm phenotypes for two other candidates highlights that a robust QTL-to-gene pipeline remains elusive. The complexity of metabolic networks and the influence of genetic background can mask the effects of single-gene knockouts, a necessary caution for future functional studies.

In summary, by integrating consumer preferences with deep genetic and metabolic data, our research provides more than just a valuable community resource. It delivers a strategic roadmap for the future of sweet-corn improvement, establishing a data-driven foundation to precisely design elite varieties that possess both superior flavor and high yield.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-025-02401-0.

References

1. *Quick Stats* (United States Department of Agriculture (USDA), National Agricultural Statistics Service (NASS), 2022); https://quickstats.nass.usda.gov/results/1AB625B8-19CC-312D-ACF6-58D2EA98E806?pivot=short_desc
2. Singh, I., Langyan, S. & Yadava, P. Sweet corn and corn-based sweeteners. *Sugar Tech* **16**, 144–149 (2014).
3. Tracy, W. F., Shuler, S. L. & Dodson-Swenson, H. The use of endosperm genes for sweet corn improvement: a review of developments in endosperm genes in sweet corn since the seminal publication in *Plant Breeding Reviews*, Volume 1, by Charles Boyer and Jack Shannon (1984). *Plant Breed. Rev.* **43**, 215–241 (2019).
4. Tracy, W. F., Whitt, S. R. & Buckler, E. S. Recurrent mutation and genome evolution: example of Sugary1 and the origin of sweet maize. *Crop Sci.* **46**, S49–S54 (2006).
5. Hu, Y. et al. Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. *Nat. Commun.* **12**, 1227 (2021).
6. De Vries, B. D. & Tracy, W. F. Characterization of endosperm carbohydrates in isa2-339 maize and interactions with su1-ref. *Crop Sci.* **56**, 2277–2286 (2016).
7. Dodson-Swenson, H. G. & Tracy, W. F. Endosperm carbohydrate composition and kernel characteristics of shrunken2-intermediate (sh2-i/sh2-i Su1/Su1) and shrunken2-intermediate-sugar y1-reference (sh2-i/sh2-i su1-r/su1-r) in sweet corn. *Crop Sci.* **55**, 2647–2656 (2015).
8. Da Fonseca, R. R. et al. The origin and evolution of maize in the Southwestern United States. *Nat. Plants* **1**, 14003 (2015).
9. Allam, M. et al. Identification of QTLs involved in cold tolerance in sweet x field corn. *Euphytica* **208**, 353–365 (2016).
10. Song, J. F. et al. Carotenoid composition and changes in sweet and field corn (*Zea mays*) during kernel development. *Cereal Chem.* **93**, 409–413 (2016).
11. Szymanek, M., Tanas, W. & Kassar, F. H. Kernel carbohydrates concentration in sugary-1, sugary enhanced and shrunken sweet corn kernels. *Agric. Agric. Sci. Proc.* **7**, 260–264 (2015).
12. Hufford, M. B. et al. Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
13. Zhang, X. et al. Maize sugary enhancer1 (*se1*) is a gene affecting endosperm starch metabolism. *Proc. Natl Acad. Sci. USA* **116**, 20776–20785 (2019).
14. Baseggio, M. et al. Genome-wide association and genomic prediction models of tocochromanols in fresh sweet corn kernels. *Plant Genome* **12**, 180038 (2019).
15. Li, K. et al. Large-scale metabolite quantitative trait locus analysis provides new insights for high-quality maize improvement. *Plant J.* **99**, 216–230 (2019).
16. Chen, W. K. et al. Convergent selection of a WD40 protein that enhances grain yield in maize and rice. *Science* **375**, eabg7985 (2022).
17. Hufford, M. B. et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).
18. Yang, N. et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat. Genet.* **51**, 1052–1059 (2019).
19. Fu, J. et al. RNA sequencing reveals the complex regulatory network in the maize kernel. *Nat. Commun.* **4**, 2832 (2013).
20. Wen, W. et al. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat. Commun.* **5**, 3438 (2014).
21. Yang, X. H. et al. Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. *Mol. Breed.* **28**, 511–526 (2011).
22. Liu, H. et al. Distant eQTLs and non-coding sequences play critical roles in regulating gene expression and quantitative trait variation in maize. *Mol. Plant* **10**, 414–426 (2017).
23. Chen, L. et al. Genome sequencing reveals evidence of adaptive variation in the genus *Zea*. *Nat. Genet.* **54**, 1736–1745 (2022).
24. Trimble, L., Shuler, S. & Tracy, W. F. Characterization of five naturally occurring alleles at the sugary1 locus for seed composition, seedling emergence, and isoamylase1 activity. *Crop Sci.* **56**, 1927–1939 (2016).
25. Yang, J. et al. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
26. Yang, R. et al. Widely targeted metabolomics analysis reveals key quality-related metabolites in kernels of sweet corn. *Int. J. Genom.* **1**, 2654546 (2021).
27. Li, P. et al. Fructose sensitivity is suppressed in *Arabidopsis* by the transcription factor ANAC089 lacking the membrane-bound domain. *Proc. Natl Acad. Sci. USA* **108**, 3436–3441 (2011).

28. Liu, H. et al. Genomic, transcriptomic, and phenomic variation reveals the complex adaptation of modern maize breeding. *Mol. Plant* **8**, 871–884 (2015).
29. Zhang, X. et al. The tin1 gene retains the function of promoting tillering in maize. *Nat. Commun.* **10**, 5608 (2019).
30. Zhu, G. et al. Rewiring of the fruit metabolome in tomato breeding. *Cell* **172**, 249–261.e12 (2018).
31. Colantonio, V. et al. Metabolomic selection for enhanced fruit flavor. *Proc. Natl Acad. Sci. USA* **119**, e2115865119 (2022).
32. Fernie, A. R. & Alseekh, S. Metabolomic selection-based machine learning improves fruit taste prediction. *Proc. Natl Acad. Sci. USA* **119**, e2201078119 (2022).

Methods

Sweet-corn germplasm planting and sampling

A total of 295 sweet-corn accessions were collected from various regions worldwide, including China (165), America (43), Thailand (22), Japan (4), Canada (1) and Argentina (1) and 59 accessions with no clear geographic record. All sweet corns were planted in one-row plots using a randomized block design at the field experiment station (113°224′E, 23°093′N) of Guangdong Academy of Agricultural Sciences. Detailed information about these accessions, such as traditional names, pedigree source descriptions and germplasm types, is available in Supplementary Table 5.

Leaf samples from each line were harvested at seedling stage (50 d after sowing) for DNA extraction. All maize plants were self-pollinated and five immature seeds were collected at 15 d and 20 d after pollination from three ears in each block. These seeds were bulked for total RNA and metabolite extraction. All samples for DNA extraction, RNA isolation and metabolite profiling were immediately stored at −80 °C after sampling for further analysis. In total, 295 leaf samples at the seedling stage were obtained and used for WGS and 280 immature kernel tissue samples were harvested and successfully processed for RNA-seq and metabolite profiling.

Genome assembly

The genome of RC was assembled using 247.21 Gb of clean data (reads with a length of <1,000 bp or bases with a quality score <7 were filtered out) obtained from the Nanopore GridION sequencing platform. The assembly pipeline integrated Necat[33] (v0.01) and Pilon[34] (v1.22). In brief, Necat was used for Nanopore read correction, initial contig assembly and contig correction. The assembly results were then corrected for indel and SNP errors using Pilon based on next-generation sequencing data. A total of 891 contigs was obtained, with the shortest contig exceeding 2,000 bp.

Leaf tissues from RC lines at the seedling stage were fixed with formaldehyde to maintain the three-dimensional (3D) structure of the DNA, then digested with restriction endonuclease. Biotin-labeled bases were introduced to DNA sequences using the DNA terminal repair mechanism. The DNA was then fragmented and 300-bp to 700-bp fragments were recovered. High-quality Hi-C libraries were sequenced on the BGI DNBSEQ platform. In total, 877.40 million reads (263.2 Gb clean data; 121.3 × genome coverage) were obtained from the Hi-C library. Of these, 65.79% of the read pairs were uniquely mapped to the assembled genome and 47.96% were valid interaction pairs used for Hi-C assembly.

The software fastp[35] (v0.23.2) was used to filter raw reads with the parameters '--average_qual 15 -l 150'. Juicer[36] (v1.6) was used to align the clean pair-end reads to the assembled genome to obtain the uniquely mapped read pairs. Software 3D-DNA[37] (v170123) was used to cluster, order and orient the genome contigs on to the pseudochromosomes. Finally, the RC assemblies were divided into 500-kb bins and the interaction signals generated by the read pairs between each bin were visualized in the heatmap.

Genome evaluation and annotation

The completeness of the genome was assessed using BUSCO[38] (v3) with the dataset of embryophyta_odb10. Repetitive sequences, including tandem repeats and transposable elements, were annotated with de novo predictions (RepeatModeler v4.1.0 and LTR_FINDER v1.06) and homolog searches (RepeatMasker v4 and RepeatProteinMask v4.0.7) based on the database of RepBase[39–41] (v21.12). Using Meryl[42] (v1.4.1), a *k*-mer database was constructed with a *k*-mer size of 20. The consensus *k*-mer QV score was then computed using Merqury[42] (v1.3) with its default parameters. LTRs were predicted using EDTA[43] (v2.2) with its default parameters. The LAI was computed utilizing the LTR_retriever[44,45] (v3.0.1) pipeline to assess the continuity of assembled repetitive sequences. Gene annotation utilized an evidence-based prediction pipeline. For homolog analysis, five model species (*Setaria italica*, *Oryza sativa*, *Brachypodium distachyon*, *Sorghum bicolor* and *Arabidopsis thaliana*) and three maize genomes (B73, Mo17 and SK) were selected[18,46–52]. A set of 3,000 well-constructed genes were randomly selected for Augustus as a training dataset for the de novo prediction. Finally, Maker[53] (v2.31.8) was used for gene annotation, integrating RNA-seq data from ten tissues (germ, radicle, male spikelet, embryo, unpollinated silks, immature ear, leaf, stem, root and endosperm) of the RC inbred line.

Resequencing and variant calling

Genomic DNA from sweet-corn seedling leaf tissues was isolated and sequencing libraries were prepared following instructions for the Illumina platform. These libraries underwent pair-end sequencing on the Illumina Hiseq2500 platform. To investigate the population affinities of sweet corn, we analyzed them alongside a published dataset of field corn, which contains 507 inbred lines with diverse genetic backgrounds[21,22]. These field corn inbred lines were collected globally, encompassing temperate, tropical and/or subtropical inbred lines and some landraces. Detailed information on these field corns has been described in a previous study[21]. The field corn were re-sequenced using an Illumina NovaSeq6000 platform[23], with sequencing depth varying between 15.4× and 34.6× and an average genome coverage of 22.5×.

For sweet corn, heterozygosity was estimated with GenomeScope 2.0 (ref. 54), which analyzes the *k*-mer count distribution from Jellyfish[55] (v1.1.11). Trimmomatic[56] (v0.33) was employed to remove low-quality bases and reads. Retained read pairs were then mapped against our chromosome-scale RC genome using the Burrows–Wheeler Aligner[57] (v0.7.17) for each sample, respectively. SAMtools[58] (v1.9) was used to filter reads with a mapping quality (MAPQ) <30. SNP calling was performed using GATK[59] (v4.1.3), following the best-practice pipeline. Read depth and coverage were determined using BEDtools[60] (v2.25.0).

Population stratification

An unsupervised ancestral component analysis for sweet corn and field corn was performed using ADMIXTURE[61] (v1.3.0). A subset of SNP dataset (167,000 SNPs; filter criterion: minor allele frequency (MAF) >0.05, missing ratio <10% and pairwise LD $r^2 < 0.2$ within 100 kb) were selected for admixture analyses. To determine the number of ancestral components (*K*) of the inbred lines, a tenfold crossvalidation approach was implemented for each *K*, $K = 1$ to $K = 10$. We chose $K = 3$ as the number of ancestries for these lines, because the crossvalidation error sharply converged at this value.

Statistical analyses of nucleotide diversity (π) and population divergence (F_{st}) were conducted using VCFtools[62] (v0.1.16) with a 1,000-bp sliding window and 100-bp steps. PCA was carried out using PLINK[63] with the '-pca' option. SNPs were filtered with an MAF of 0.05 and a missing ratio of 10. High LD SNPs were also filtered using pairwise LD values calculated and processed using an R script. Haplotype networks were constructed using the R package pegas[64].

LD decay

PopLDdecay[65] (v3.40) software was used to calculate LD values based on the r^2 values and the corresponding distance between given SNPs within 600 kb. The parameters were set as following: '-MaxDist 600 -MAF 0.05'. The distance of LD decay was obtained when r^2 dropped to 0.2.

Genome-wide selective sweep scan

Two approaches were used to identify genomic regions with positive selection signals. We compared the SNP dataset of sweet corn (test population) with that of field corn (reference population). XP-CLR (v1.0) values for each window were calculated with the published script[66]. Parameters were: sliding window size 0.5 cM, grid size 20 kb, maximum number of SNPs within a window 200 and a correlation level cutoff of 0.70. Regions with the top 5% highest XP-CLR values were identified as candidate selective sweeps, where adjacent intervals without gaps are merged to form continuous regions.

XP-EHH was measured using the software selscan[67] (v1.3.0). The genome was divided into consecutive, nonoverlapping 20-kb windows and XP-EHH values were standardized using the 'norm --xpehh' function in selscan. Regions assigned as candidates for significant selection featured a $P < 0.05$, with adjacent and nongapped intervals merged directly to form continuous regions.

GC–MS and LC–MS analysis

Kernels of sweet-corn accessions at 20 DAP were harvested in two biological replicates and stored at −80°C before metabolic analyses. The kernels, pre-cooled in liquid nitrogen, were ground using a Mixer mill (Retsch, cat. no. MM400) for 30 s at 30 Hz. Then 50 mg of each sample powder was extracted following the procedures described in previous studies[68,69].

For primary metabolite profiling, dried treatments were derivatized with *N*-methyl-*N*-(trimethylsilyl) trifluoroacetamide[70] and analyzed using GC–MS (Agilent, cat. no. 7890A-5975C); 1 µl of the liquid mixture from each sample was injected into the GC–MS at 270 °C in split mode (50:1) with helium carrier gas (>99.999% purity) flow set to 1 ml min^{-1} and separated by a DB-35MS UI (30 m × 0.25 mm, 0.25-µm) capillary column.

For secondary metabolite profiling, another dried treatment was resuspended in 150 µl of ultra-performance LC-grade methanol:water (1:1, v:v). Samples were then subjected to MS analysis using an Orbitrap fusion (Thermo Fisher Scientific) equipped with a reversed-phase LC system (Dionex, Thermo Fisher Scientific) in heated electrospray ionization mode. First, 10 µl of each sample was eluted using a TSS T3 column (100 mm × 2.1 mm containing 1.8-µm diameter particles, Waters) with a 0.4 ml min^{-1} flow rate. Mobile phase A was water with 0.1% formic acid and mobile phase B was acetonitrile with 0.1% formic acid.

Metabolomic data were analyzed according to the protocol described in our previous study[71]. The Agilent MassHunter Quantitative Analysis software (vB.07.01) was used for GC–MS data analyses. An NIST library and in-house database established using authentic standards were used together for metabolite identification.

LC–MS-based metabolomics were first analyzed using Xcalibur software (v4.2; Thermo Fisher Scientific). Compound Discovery (v3.1; Thermo Fisher Scientific) and Trace Finder (v3.3; Thermo Fisher Scientific) were used for qualitative and quantitative analysis of the secondary metabolome. Secondary metabolite identification was supported by in-house databases and online databases, including mzCloud, Chemspider, Human Metabolome Database (HMDB), KEGG and BioCyc. Both GC–MS and LC–MS metabolites were reported following the latest reporting standards[71].

Metabolite GWAS

We used a mixed linear model to evaluate the association between SNPs and metabolic traits with Tassel[72] (v3.0), integrating both population structure and kinship matrix. The top five principal components (PCs) mentioned above were used as population structure and the kinship was estimated using the Tassel program.

Heterozygous genotypes in the SNP dataset were replaced with missing ones. The SNP dataset was filtered with a 10% cutoff for missing data and markers with MAF > 0.05 remained for GWASs. A uniform significance threshold ($P = 0.05/n$, where $n = 1{,}267{,}142$, the effective number of independent SNPs used in the GWASs) was used for all metabolic traits[73,74]. The *P*-value threshold for significance in the present sweet-corn population was approximately $P = 3.95 \times 10^{-8}$.

The SNP with the lowest *P* value in each locus was treated as the lead SNP and genes within a 50-kb region (downstream and upstream) of the lead SNP were selected as candidate genes. The R software package 'coloc' was used to perform the colocalization analyses for two GWAS results using the corresponding summary data[75]. The candidate genes for further validation were selected based on multiple rules (for example, statistical analyses, gene expression, correlation to metabolites and flavor traits and prior biological knowledge) (Extended Data Fig. 5b).

RNA-seq and eQTL analysis

All samples for RNA-seq were collected from immature kernel tissues at 15 DAP. Total RNA was extracted using a Quick RNA Isolation Kit according to the manufacturer's instructions. The quality of the extracted RNA was evaluated using BioRad Experion. Illumina stranded messenger RNA libraries with an insert size of 300–500 bp were constructed using a Truseq stranded mRNA sample preparation kit (Illumina). Paired-end 150-bp sequencing was conducted on the Illumina Hiseq2500 platform, yielding an average of 31.27 million raw reads per sample. After removing sequencing adapters and low-quality reads with Trimmomatic, the trimmed reads were then mapped to the RC genome using Hisat2 (ref. 76) (v2.1.0). An average of 24.76 million reads per sample with high mapping quality (MAPQ > 30) were used for expression abundance evaluation. BAM files were sorted and organized using SAMtools. StringTie[76] (v2.20) was used to assemble transcripts and estimate their expression abundances. Ballgown[76] package (v3.6.0) in R was used to extract fragments per kilobase per million read (FPKM) values for all genes in all samples.

To detect eQTLs through a mixed linear model, the expression value of each gene was normalized to a normal distribution using the normal quantile transformation function (qqnorm) in R to meet the assumptions of the GWAS mixed linear model statistical method. A total of 20,073 genes (with median expression of FPKM > 1) was obtained to conduct downstream analyses. Association analyses followed the same procedures and significance threshold criterion as described for mGWASs.

For the field-corn population, all raw sequencing data were obtained from our previous studies[19,22]. We reanalyzed the eQTLs using the same pipelines and parameters. In brief, 8,407,536 high-quality SNP loci were identified within the field-corn panel ($n = 342$), with 1,412,504 independent SNPs obtained. A total of 18,972 genes (with median expression of FPKM > 1) were regarded as expressed genes and used for subsequent eGWASs. The *P*-value threshold for significance in the present field corn population was approximately $P = 3.54 \times 10^{-8}$ ($P = 0.05/n$).

The eQTLs were classified into *cis*-eQTLs and *trans*-QTLs according to the distance between the lead SNP and the corresponding gene. For each eQTL, if the lead SNP was located within the gene or within 2 kb upstream of the gene, it was regarded as a *cis*-eQTL; otherwise, it was considered to be a *trans*-eQTL.

Flavor test

To observe the difference in sweet-corn quality among 295 accessions, we conducted tasting tests with both sweet-corn breeders and ordinary consumers. Briefly, fresh sweet-corn ears were harvested at the optimal eating period (20 DAP). To reduce the influence of human-induced factors, two independent experiments were organized.

In 2017 and 2019, 5 and 7 experts from a high-quality sweet-corn breeding program scored 232 and 228 cooked corn samples, respectively. Participants rated four properties (sweetness, brittleness, pericarp thickness and volatiles) on a scale from 1 to 10 (10 being most favorable and 1 least favorable), with a medium performance line as a control. Given the inverse correlation between original pericarp thickness scores and actual physical thickness—with 10 representing the thinnest pericarp (most favorable) and 1 the thickest (least favorable)—we applied a data transformation: subtracting each score from 11. This transformation ensured that, in subsequent analyses, higher numerical values directly corresponded to thicker pericarps, which is more consistent with common sense.

In addition, in 2019, 89 samples from different individuals were scored twice to evaluate the reliability of the scores. To evaluate the representative nature of the expert data, 95 untrained testers (33 men, 62 women) scored 18 typical lines from the sweet-corn population using the same procedure and criteria as the expert tests.

The tasting data from 2017 and 2019 were disposed as independent environments and merged using the BLUP method. These merged data

were then used for association and correlation analyses with metabolite traits and gene expression values.

Knockout of candidate genes using the CRISPR–Cas9 technique

For each candidate gene (*Sh2*, *Su1*, *ZmAPS1*, *ZmSK1* and *ZmCRR5*), two small guide RNAs (Supplementary Table 23) were designed using CRISPR-P[77] (v2.0). Vectors carrying the small guide RNAs were imported into the *Agrobacterium* strain and then used to transform the immature embryos of the maize inbred line KN5585 through *Agrobacterium*-mediated transformation[78]. All edited seedlings, together with control plants, were planted in randomized block design plots at Sanya, Hainan province (109°51′E, 18°25′N) in 2020. Mutations were identified using Sanger sequencing. All kernel tissue samples were harvested at 20 DAP. The procedure for immature kernel transcriptome and metabolome profiling was the same as that used for the sweet-corn population. All the relevant materials generated here are available to be shared on request from the corresponding authors.

Statistics

Broad-sense heritability (H^2) was estimated using the variance component method as the ratio of genotypic to the total phenotypic variance. $H^2 = \sigma_g^2/(\sigma_g^2 + \sigma_y^2/n_y)$, where σ_g^2 and σ_y^2 are the variance components for genotypes and years, respectively, and n_y represents the number of years. Broad-sense heritability was estimated using the lme4 (ref. 79) package in R. SNP heritability (variation in flavor ratings explained by all SNPs across the genome) was estimated using GCTA[25,80] (v1.94.1). Genotype data were converted to PLINK format using VCFtools[62] to meet GCTA requirements. The statistical significance of the deviation of observed data distribution (for example, eQTL hotspot identification) from uniform distribution were assessed using permutation tests implemented in R scripts. GO and KEGG analyses were performed using the clusterProfiler[81] package in R.

A three-step regression approach was adopted to evaluate the factors (including mutants, gene expressions and metabolites) involved in sweet-corn flavor quality. Briefly, two mutants with significant associations to flavor scores (that is, *sh2* and *su1*) were included in the regression analyses. For the 494 metabolites and 20,073 gene expressions, they were grouped into different modules using the 'kmeans' method[82] and 28 and 14 modules were determined using the 'wss' method in the R package of factoextra (http://www.sthda.com/english/rpkgs/factoextra). Mean values of each normalized potential factor were additionally incorporated into the regression models. We next performed the least absolute shrinkage and selection operator[83] regression analysis to evaluate the priorities of these factors. The proportion of variance was adjusted with collinear effects excluded in the multiple linear regression model used for the perception of sweet-corn flavor. First, a linear regression mode was used for collinearity screening with a variance inflation factor <4. Second, the least absolute shrinkage and selection operator was adopted for the evaluation of the relative weight and to assign priority to each factor simultaneously. Last, adjusted R^2 values were estimated from linear regression models with factors introduced based on the priority obtained above. All these steps were implemented in R. The specific statistical tests used were indicated in the corresponding figure legends or corresponding sections in Methods.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets supporting the findings of this study have been deposited into the CNGB Sequence Archive of the China National GeneBank DataBase (https://db.cngb.org/) under the following accession nos.: de novo assembled genomes and raw data at CNP0004684, CNP0003283 and CNP0003295; raw WGS data for the 295 sweet-corn accessions at CNP0003213; RNA-seq data for the 280 sweet-corn accessions at CNP0003294; and RNA-seq for knockout and wild-type lines at CNP0003291 and CNP0004707. Source data are provided with this paper.

Code availability

No customized code was generated for this study. All analyses were performed using publicly available software with parameters detailed in Methods.

References

33. Vaser, R. et al. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
34. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
35. Chen, S. et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
36. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
37. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
38. Simão, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
39. Jurka, J. et al. Repbase update: a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
40. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152–i158 (2005).
41. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
42. Rhie, A. et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
43. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
44. Ou, S. & Ning, J. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
45. Ou, S. et al. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
46. Zhang, G. et al. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* **30**, 549–554 (2012).
47. Yu, J. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).
48. International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
49. Paterson, A. et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
50. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
51. Jiao, Y. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
52. Sun, S. et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat. Genet.* **50**, 1289–1295 (2018).
53. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **12**, 491 (2011).

54. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
55. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
56. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
58. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
59. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
60. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
61. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
62. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
63. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
64. Paradis, E. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26**, 419–420 (2010).
65. Zhang, C. et al. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
66. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
67. Szpiech, Z. A. selscan 2.0: scanning for sweeps in unphased data. *Bioinformatics* **40**, btae006 (2024).
68. Salem, M. A. et al. Protocol: a fast, comprehensive and reproducible one-step extraction method for the rapid preparation of polar and semi-polar metabolites, lipids, proteins, starch and cell wall polymers from a single sample. *Plant Methods* **12**, 45 (2016).
69. Wang, H. et al. A subsidiary cell-localized glucose transporter promotes stomatal conductance and photosynthesis. *Plant Cell* **31**, 1328–1343 (2019).
70. Yan, S. et al. Comparative metabolomic analysis of seed metabolites associated with seed storability in rice (*Oryza sativa L.*) during natural aging. *Plant Physiol. Biochem.* **127**, 590–598 (2018).
71. Alseekh, S. et al. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat. Methods* **18**, 747–756 (2021).
72. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
73. Li, M. X. et al. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**, 747–756 (2012).
74. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**, 241–247 (1995).
75. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
76. Pertea, M. et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
77. Liu, H. et al. CRISPR-P 2.0: an improved CRISPR-Cas9 tool for genome editing in plants. *Mol. Plant* **10**, 530–532 (2017).
78. Liu, H. J. et al. High-throughput CRISPR/Cas9 mutagenesis streamlines trait gene identification in maize. *Plant Cell* **32**, 1397–1413 (2020).
79. Bates, D. et al. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
80. Sun, G. et al. Variation explained in mixed-model association mapping. *Heredity* **105**, 333–340 (2010).
81. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
82. Hartigan, J. & Wong, M. A K-means clustering algorithm. *J. R. Stat. Soc. C* **28**, 100–108 (1979).
83. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Ann. Stat.* **32**, 407–499 (2004).

Acknowledgements

The work was supported by National Natural Science Foundation of China (grant nos. 32321005 to J.Y., U1901201 to J.Y. and 32001563 to K.L.), the National Key Research and Development Program of China (grant no. 2022YFD1201502 to G.L.), the Earmarked Fund for CARS (grant no. CARS-02-85 to G.L.), the Guangdong S&T Program (grant no. 2022B0202060003 to Y.Y.), the Science and Technology Major Program of Hubei Province (grant no. 2021ABA011 to J.Y.), the Agricultural Competitive Industry Discipline Team Building Project of Guangdong Academy of Agricultural Sciences (grant no. 202115TD to G.L.), the Science and Technology Program of Guangzhou (grant no. 202102021015 to K.L.), the Project of Collaborative Innovation Center of Guangdong Academy of Agricultural Sciences (grant no. XTXM202203 to S.Y.), the Guangdong Provincial Science and Technology Plan Project (grant no. 2023B1212060038 to G.L.) and the Special Fund for Scientific Innovation Strategy-construction of High-Level Academy of Agriculture Science (grant nos. R2019YJ-YB1002 to K.L. and R2020PY-JX019 to S.Y.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We are grateful to Maize Research Institute of Shandong Academy of Agricultural Sciences for help in collecting sweet corn resources. Computation resources were provided by the high-throughput computing platform of the National Key Laboratory of Crop Genetic Improvement at Huazhong Agricultural University and supported by Hao Liu.

Author contributions

J.Y. and J.H. conceived and designed the research. Y.Y., W.Q.L., G.L., W.L., Y.N.X., N.Z., L.Z. and K.L. managed the project. K.L., Y.Y., Q.Z., J.Y.L., L.C., Y.J.X., N.Y., H.-J.L., L.F. and S.G. performed the genome sequencing and bioinformatics. Y.Y., J.H.L., L.X., X.Q., C.L., W.J.L., Y.L. and Y.X. prepared the samples for resequencing and transcriptome profile of the sweet-corn population and contributed to data analysis. W.Q.L. was responsible for the filed corn planting and data collection. S.Y., W.H. and Q.K. performed metabolome analyses using GC–MS and LC–MS platforms. J.X., K.L., W.Z. and T.W. managed the knockout lines and performed the molecular experiments. K.L., H.-J.L., S.Y., A.R.F., J.H. and J.Y. wrote and revised the manuscript.

Competing interests

J.X. is an employee of WIMI Biotechnology Co., Ltd. The other authors declare no competing interests.

Additional information

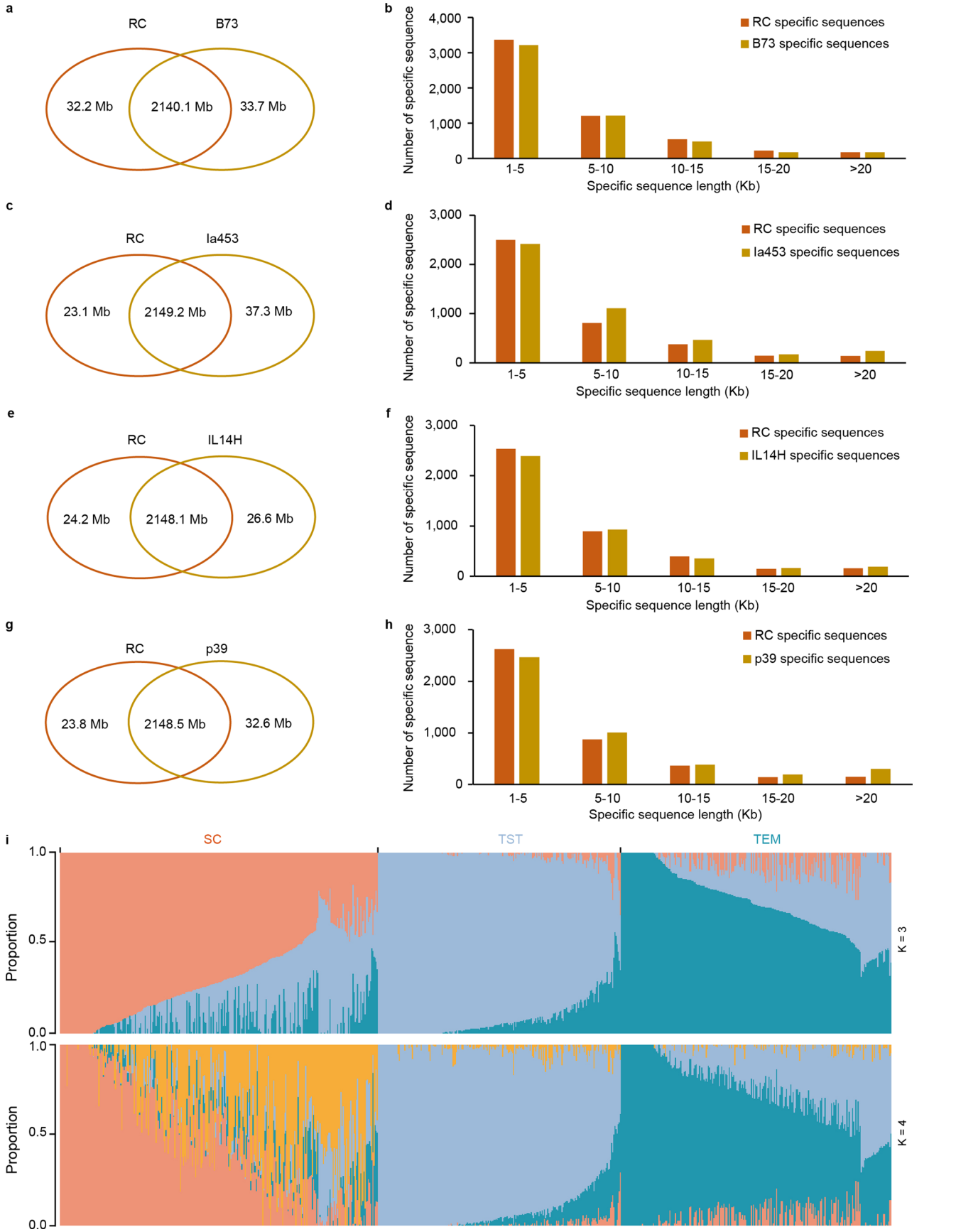
Extended data is available for this paper at https://doi.org/10.1038/s41588-025-02401-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02401-0.

Correspondence and requests for materials should be addressed to Hai-Jun Liu, Jianguang Hu or Jianbing Yan.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-025-02401-0

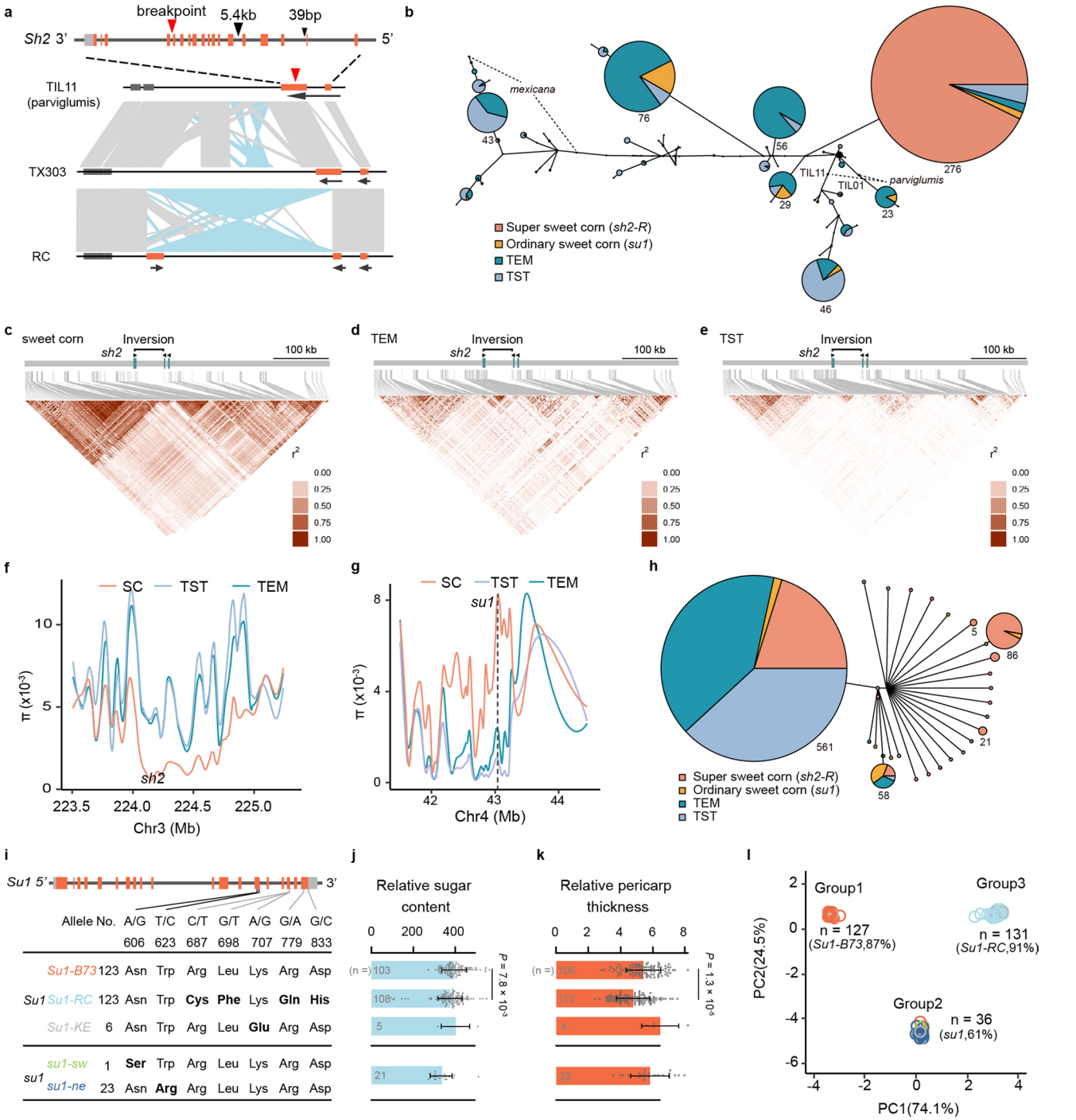


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41588-025-02401-0

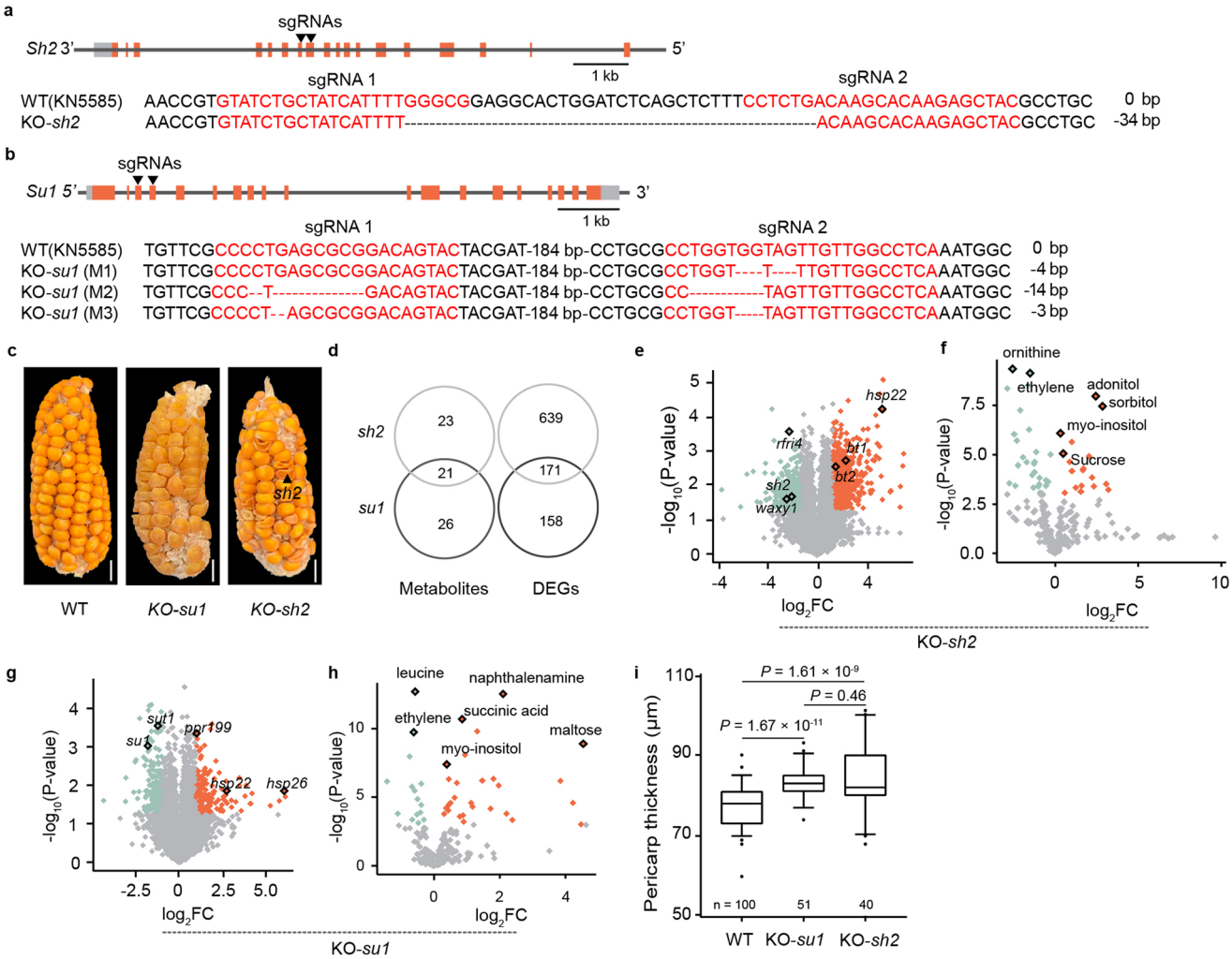
Extended Data Fig. 1 | Genetic differences of sweet and field corn. a, Length of specific sequences in the RC genome compared to B73. **b**, Distribution of specific sequences in the RC genome compared to B73. **c**, Length of specific sequences in the RC genome compared to Ia453. **d**, Distribution of specific sequences in the RC genome compared to Ia453. **e**, Length of specific sequences in the RC genome compared to IL14H. **f**, Distribution of specific sequences in the RC genome compared to IL14H. **g**, Length of specific sequences in the RC genome compared to p39. **h**, Distribution of specific sequences in the RC genome compared to p39. **i**, Population structure of sweet and field corn populations, inferred using the maximum-likelihood method with three and four ancestral components (K = 3, 4).

https://doi.org/10.1038/s41588-025-02401-0



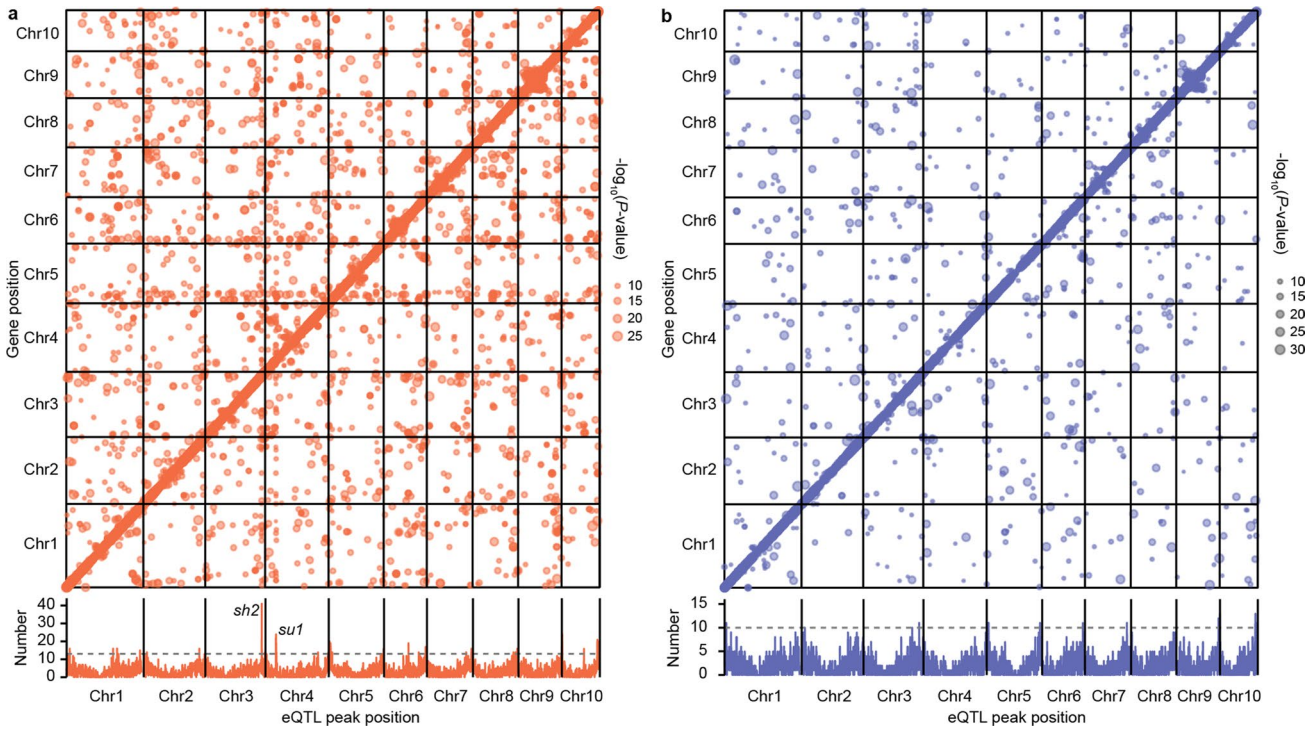
Extended Data Fig. 2 | Genetic variation in *sh2* and *su1*. **a**, Synteny map highlighting an inversion in *sh2* in RC genome compared to the field corn genome TX303 and *parviglumis* genome TIL11. **b**, Haplotype network analysis of 264 SNPs in the *sh2-RC* region, based on the RC gene model that includes the identified inversion. (**c**) LD plot (r^2 values) for *sh2* and flanking regions (< 500 kb) in sweet corn population, and in kernel corn from (**d**) TEM population and (**e**) TST population. Nucleotide diversity (π) around (**f**) *sh2* and (**g**) *su1* in both sweet and field corn populations. **h**, Haplotype network analysis of SNPs in the *su1* region. **i**, Structural analysis identifying 5 alleles in the *Su1* gene within sweet corn population. **j**, Comparison of relative sugar content (sum of sucrose, maltose, glucose, and fructose) among different *su1* alleles. **k**, Comparison of relative pericarp thickness in immature kernels across different alleles of *su1*. The values to the left of each bar represent the number of sweet corn lines analyzed in (**j**) and (**k**). Data points show individual measurements; bar heights represent mean values and error bars represent the mean values ± s.d. Differences between groups were assessed using a two-tailed unpaired t-test. **l**, Principal component analysis of 63 SNPs discovered in *su1* region.

https://doi.org/10.1038/s41588-025-02401-0



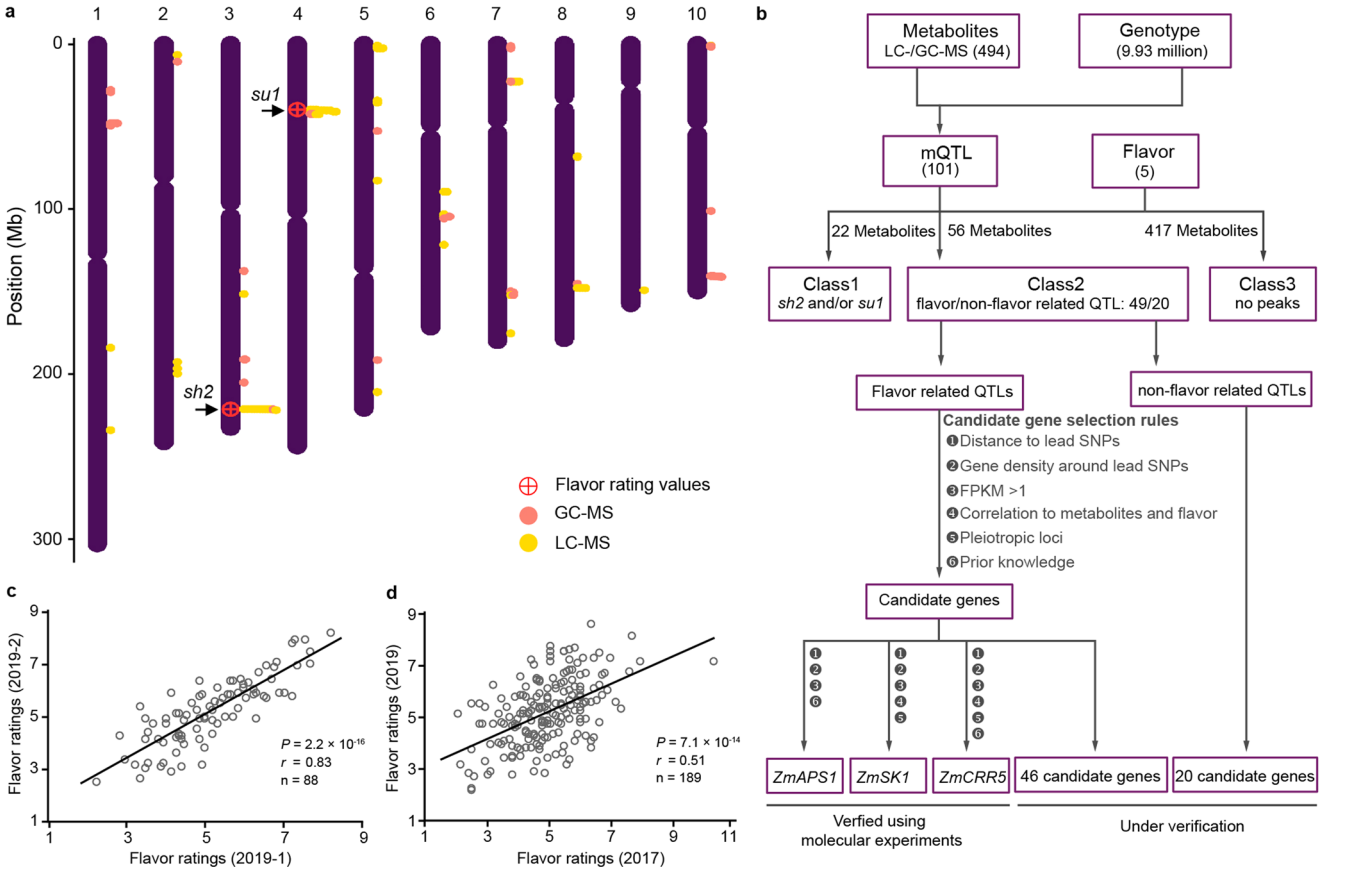
Extended Data Fig. 3 | Knockout of *Sh2 and Su1* by CRISPR-Cas9. Two sgRNAs designed for gene editing on *Sh2* (**a**) *and Su1*(**b**), respectively, shown in red. Mutations and deletions are indicated. **c**, Phenotypes of maize ears from *su1* and *sh2* CRISPR-knockout line. Scale bars, 1 cm. **d**, Metabolites identified with significant changes and differentially expressed genes (DEGs) in *Sh2* and *Su1* mutants compared to wild types. **e**, DEGs between *sh2* and WT in kernel tissues at 20 DAP. **f**, Metabolomic comparison of *sh2* and WT in kernel tissues at 20 DAP; fold change calculated using mean values (WT, n = 12; Mutant, n = 12). **g**, DEGs between *su1* and WT in kernel tissues at 20 DAP. **h**, Comparison of metabolome of *su1* and WT in kernel tissues at 20 DAP (WT, n = 12; Mutant, n = 12). **i**, Pericarp thickness comparisons between *Su1* and *Sh2* CRISPR-knockout lines and wild type kernels. Box plots are defined by the median (centre line), the 5th and 95th percentiles (box limits), and the whiskers extend to the minimum and maximum values. Individual data points are overlaid. *P* values (**i**) were calculated by one-way ANOVA followed by Tukey's multiple comparisons test.

https://doi.org/10.1038/s41588-025-02401-0



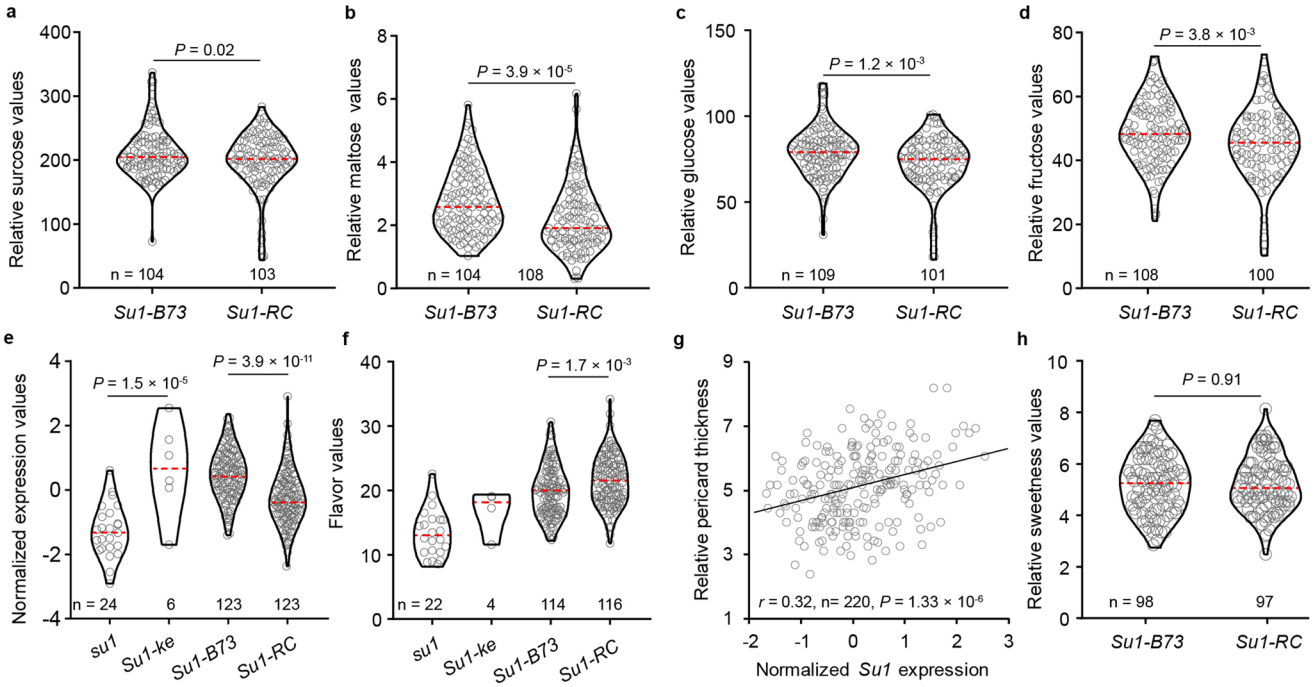
Extended Data Fig. 4 | Distribution of eGWAS signals in sweet and field corn populations. Scatter plots show the genomic positions of genes and corresponding eQTL signals in sweet (**a**) and field (**b**) corn populations (upper panels). Distribution of eGWAS signal counts per 500 kb window across the genome in sweet (**a**) and field (**b**) corn populations (lower panels). The horizontal dashed line indicates the threshold for signal hotspots (permutation test $P < 0.05$).

https://doi.org/10.1038/s41588-025-02401-0



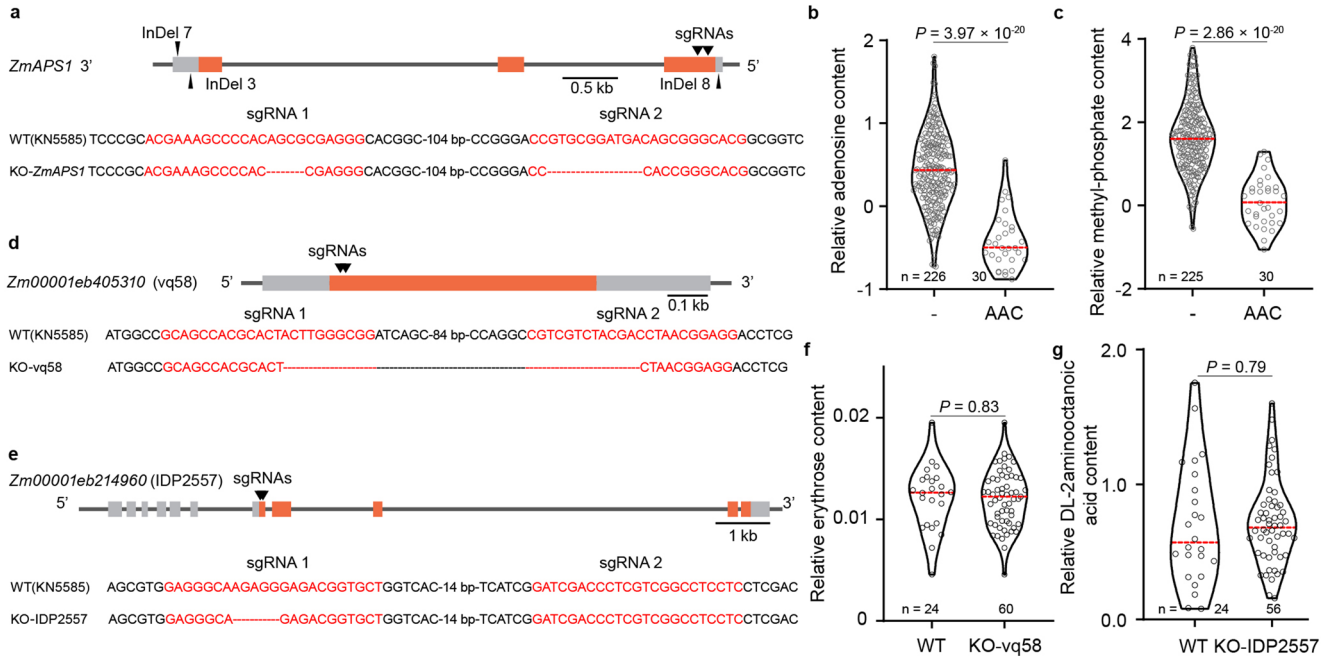
Extended Data Fig. 5 | Mapping results for flavor ratings and metabolites. **a**, Distribution of mQTL and flavor rating-related QTL across the maize genome. The arrow shows loci where mQTL and flavor rating-related QTL are colocalized. **b**, Flowchart illustrating the rationale and statistical framework used in the present study. **c**, Correlation analysis of sweet corn flavor ratings between two biological replicates in 2019. **d**, Correlation analysis of sweet corn flavor ratings between experiments conducted in 2017 and 2019.

https://doi.org/10.1038/s41588-025-02401-0



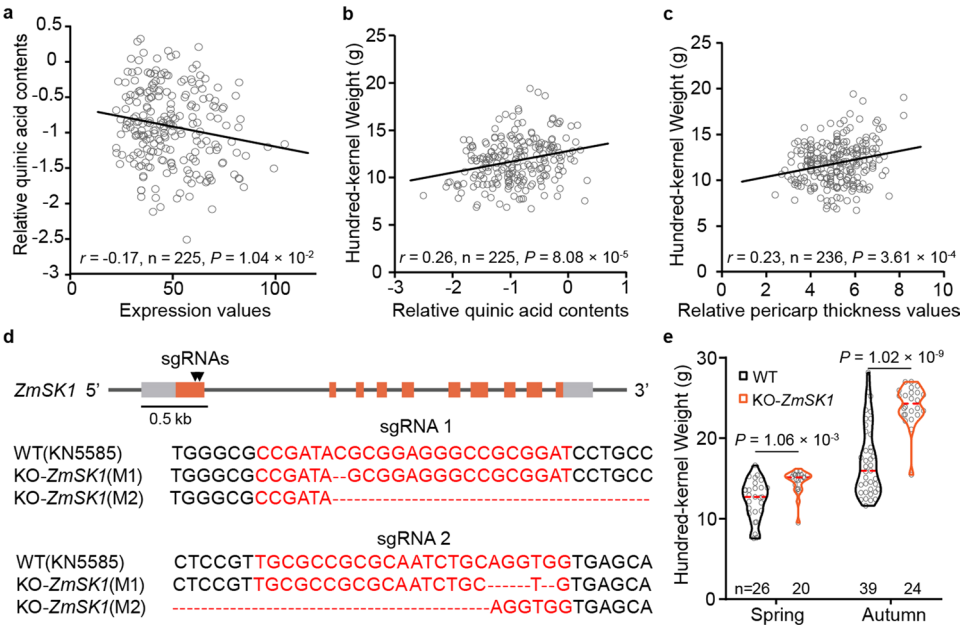
Extended Data Fig. 6 | Functional analysis of *su1* alleles. Significant differences between *Su1-RC* and *Su1-B73* alleles across the sweet corn population in sucrose (**a**), maltose (**b**), glucose (**c**), and fructose (**d**), respectively. **e**, Significant differences in expression levels among four *su1* alleles. **f**, Significant differences in flavor values among four alleles of *su1*. **g**, Correlation analysis between pericarp thickness from taste rating experiments and *su1* expression levels. **h**, Sweetness from taste rating experiments, showing no difference between *Su1-B73* and *Su1-RC* types. Differences between groups were assessed using a two-tailed unpaired t-test.

https://doi.org/10.1038/s41588-025-02401-0



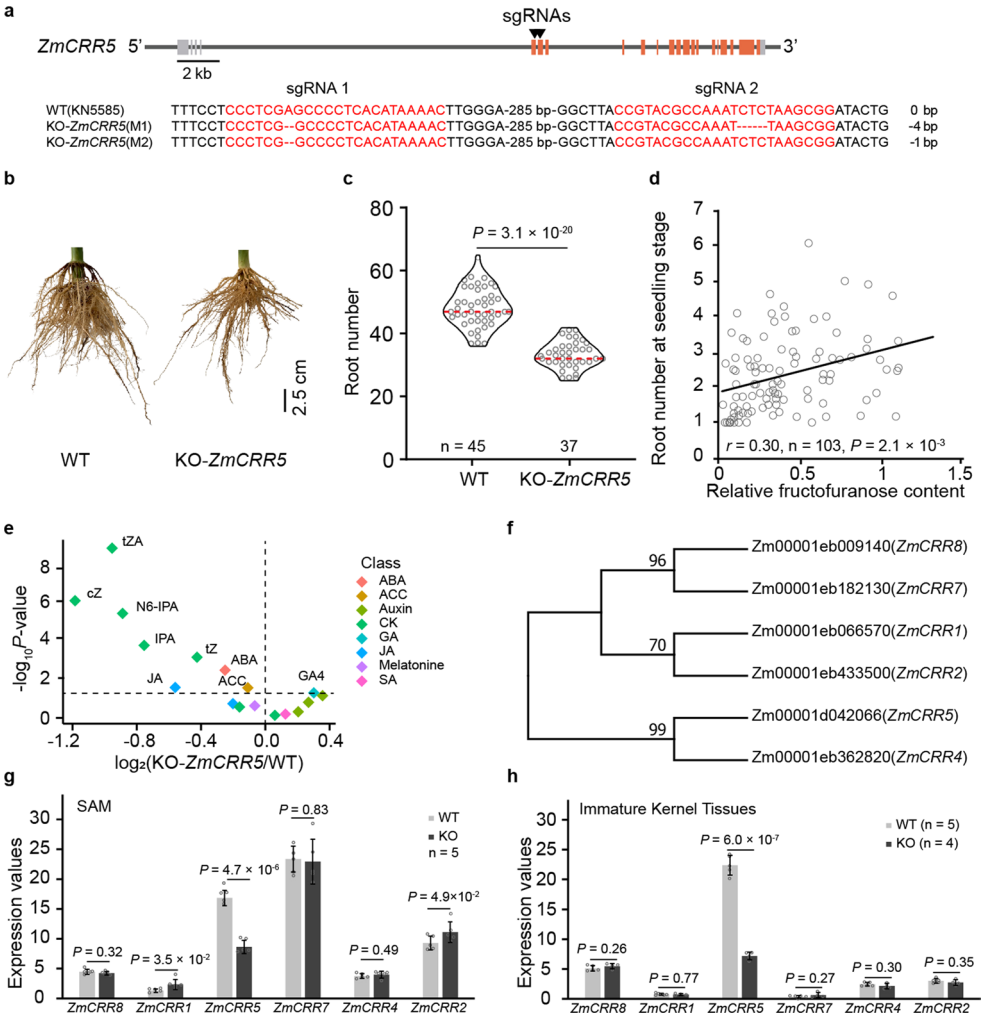
Extended Data Fig. 7 | Functional identification of candidate genes. a, Gene model of *ZmAPS1*. CRISPR-Cas9 generated mutants in *ZmAPS1* are shown, with sgRNAs in red and deletions shown as dashes. Violin plots showing differences of adenosine (**b**) and methyl-phosphate (**c**) content between genotypes at the InDel 3 (-/AAC). Knockout of Zm0001eb405310 (**d**) and Zm0001eb214960 (**e**) by CRISPR-Cas9. Violin plots of relative erythrose (**f**) and DL-2aminooctanoic acid (**g**) content between wild-type and knockout lines. Group comparisons were performed using two-tailed unpaired t-tests.

https://doi.org/10.1038/s41588-025-02401-0



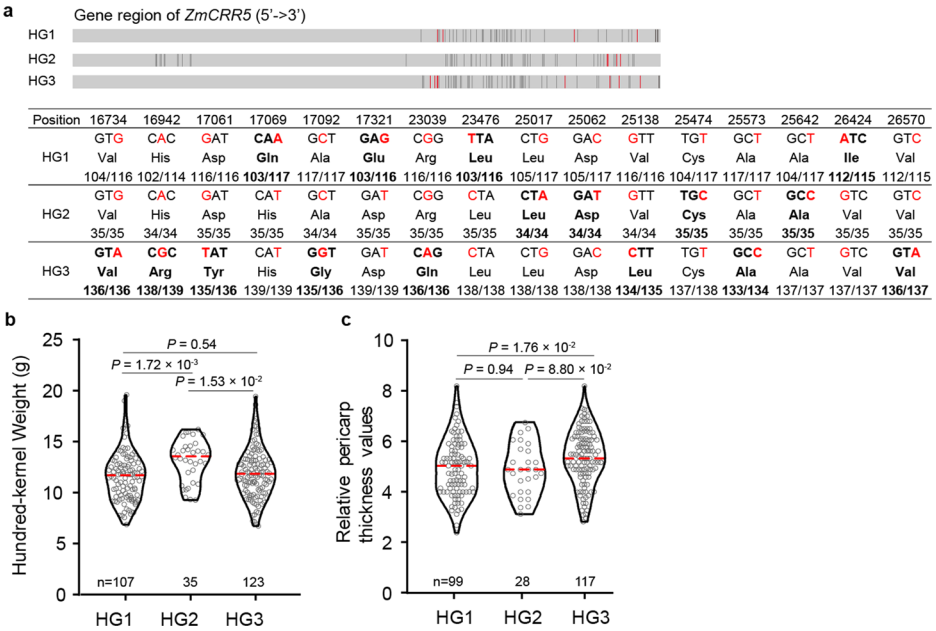
Extended Data Fig. 8 | Functional identification of *ZmSK1*. **a**, Correlation analysis between *ZmSK1* expression and quinic acid content. **b**, Correlation analysis between quinic acid content and hundred-kernel weight. **c**, Correlation analysis between pericarp thickness and hundred-kernel weight. **d**, Knockout of *ZmSK1* by CRISPR-Cas9. Mutants of *ZmSK1* are shown. **e**, Violin plots of hundred-kernel weight between wild type and *ZmSK1* knockout lines in 2022. Differences between groups were assessed using a two-tailed unpaired t-test.

https://doi.org/10.1038/s41588-025-02401-0



Extended Data Fig. 9 | Functional identification of *ZmCRR5*. **a**, Knockout of *ZmCRR5* by CRISPR-Cas9. Representative images of maize roots (**b**) and statistical comparison of root numbers (**c**) between wild type and KO-*ZmCRR5* lines. **d**, Correlation between root numbers at seedling stage and normalized fructofuranose content. **e**, Hormones levels in immature kernel of *ZmCRR5* knockouts and wild types. *P*-values were analyzed using two-tailed unpaired t-test (n = 12/12). **f**, Inferred phylogenetic analysis of six CRR genes across maize genome. **g**, Expression of six CRR genes in short apical meristem tissues of *ZmCRR5* knockout and wild types. **h**, Expression of six CRR genes in immature kernel (20 DAP) of *ZmCRR5* knockouts and wild types. Data points (**g** and **h**) show individual measurements; bar heights represent mean values and error bars represent the mean values ± s.d. Differences between groups were assessed using two-tailed unpaired t-test.

https://doi.org/10.1038/s41588-025-02401-0



Position	16734	16942	17061	17069	17092	17321	23039	23476	25017	25062	25138	25474	25573	25642	26424	26570
HG1	GTG	CAC	GAT	**CAA**	GCT	**GAG**	CGG	**TTA**	CTG	GAC	GTT	TGT	GCT	GCT	**ATC**	GTC
	Val	His	Asp	**Gln**	Ala	**Glu**	Arg	**Leu**	Leu	Asp	Val	Cys	Ala	Ala	**Ile**	Val
	104/116	102/114	116/116	**103/117**	117/117	**103/116**	116/116	**103/116**	105/117	105/117	116/116	104/117	117/117	104/117	**112/115**	112/115
HG2	GTG	CAC	GAT	CAT	GCT	GAT	CGG	CTA	**CTA**	**GAT**	GTT	**TGC**	GCT	**GCC**	GTC	GTC
	Val	His	Asp	His	Ala	Asp	Arg	Leu	**Leu**	**Asp**	Val	**Cys**	Ala	**Ala**	Val	Val
	35/35	34/34	35/35	35/35	34/34	35/35	35/35	35/35	**34/34**	**34/34**	34/34	**35/35**	35/35	**35/35**	35/35	35/35
HG3	**GTA**	**CGC**	**TAT**	CAT	**GGT**	GAT	**CAG**	CTA	CTG	GAC	**CTT**	TGT	**GCC**	GCT	GTC	**GTA**
	Val	**Arg**	**Tyr**	His	**Gly**	Asp	**Gln**	Leu	Leu	Asp	**Leu**	Cys	**Ala**	Ala	Val	**Val**
	136/136	**138/139**	**135/136**	139/139	**135/136**	139/139	**136/136**	138/138	138/138	138/138	**134/135**	137/138	**133/134**	137/137	137/137	**136/137**

Extended Data Fig. 10 | Haplotype analysis of *ZmCRR5*. **a**, Identification of haplotype-specific (minimum allele frequency < 15% in one haplotype and corresponding allele frequency > 85% in other two haplotype groups) SNPs in *ZmCRR5*, shown as vertical lines in gray (introns), red (coding regions), and black (untranslated regions). were presented with gray, red and black vertical lines located in introns, coding regions and untranslated regions, respectively (upper). Below, details of 16 SNPs in coding regions, with allele frequencies for each haplotype group. Haplotype-specific SNPs are highlighted in bold. Violin plots displaying hundred-kernel weight (**b**) and pericarp thickness (**c**) across the three haplotype groups. *P*-values were determined by one-way ANOVA followed by Tukey's multiple comparisons test.

nature portfolio

Corresponding author(s): Jianbing Yan, Jianguang Hu and Hai-Jun Liu

Last updated by author(s): Oct 1, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Software used in this study: Necat (v0.01), Pilon (v1.22), BUSCO (v3), RepeatModeler (v4.1.0), LTR_FINDER (v1.06), RepeatMasker (v4), RepeatProteinMask (v4.0.7), RepBase (v21.12), Maker (v2.31.8), Trimmomatic (v0.33), BWA (v0.7.17), SAMtools (v1.9), GATK (v4.1.3), BEDtools (v2.25.0), MassHunter Quantitative Analysis (vB.07.01), Xcalibur (v4.2), Compound Discovery (v3.1), Trace Finder (v3.3), Plink (v1.90), Tassel (v3.0), Hisat2 (2.1.0), StringTie (v2.20), R (v3.6.0), ADMIXTURE (v1.3.0), VCFtools (v0.1.16), XP-CLR (v1.0), PopLDdecay (v3.40), CRISPR-P (v2.0), fastp (v0.23.2), Juicer (v1.6), 3D-DNA (v170123), GenomeScope (v2.0), Jellyfish (v1.1.11), selscan (v1.3.0), EDTA (version 2.2), GCTA (v1.94.1) and LTR_retriever (version 3.0.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets supporting the findings of this study have been deposited into the CNGB Sequence Archive (CNSA) of the China National GeneBank DataBase (CNGBdb) under the following accession numbers: De novo assembled genomes and raw data: CNP0004684, CNP0003283, and CNP0003295; Raw WGS data for the 295 sweet corn accessions: CNP0003213; RNA-seq data for the 280 sweet corn accessions: CNP0003294; RNA-Seq for knock-out and wild-type lines: CNP0003291 and CNP0004707.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size for corn population genetic analysis is 507 (field corn) and 295 (sweet corn); For eGWAS, the sample size is 368 (field corn) and 280 (sweet corn). 295 sweet corn lines were used for kernel flavor quality estimation. For eGWAS, the sample sizes were 368 for field corn and 280 for sweet corn that had been sequenced by RNA-seq. These sample sizes were the largest possible we could collect to ensure the GWAS had high power, and no other particular statistical methods were used to predetermine the sample sizes. For CRISPR-based validation of each gene, at least 10 individuals were used to ensure reliable and consistent phenotypic effects.
Data exclusions	Raw reads and genotype with low quality were excluded as described in the methods.
Replication	The phenotype of the flavor quality estimation used in this study were validated by at least two years field experiments. The phenotype of the knock out mutants used in this study were validated by at least nine biological replication.
Randomization	The sweet corns were planted in a random order. The sample of sweet corn for whole genome re-sequencing and RNA-seq were randomly collected in one line comprising 12 individuals. Five ears of sweet corn were randomly collected for kernel quality estimation. Population structure was integrated as covariates in all GWAS analyses. For CRISPR validations, all the knock-outs and their corresponding controls were planted, cultivated and analyzed in the same environment and the same batch of sequencing.

Blinding	The investigators are blinded to the group allocation during phenotypic data collection.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging