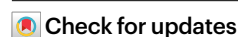


A *k*-mer-based genome-wide association study approach empowering gene mining in polyploids

Received: 11 August 2025

Accepted: 19 May 2026

Published online: 12 June 2026



Shuai Chen^{1,8}, Xinlong Liu^{2,8}, Shenyang Qu¹, Yuhan Song¹, Kun Chai¹, Hongbo Liu², Yuebin Zhang², Zhongqiang Xia¹, Xiaofeng Li¹, Jungang Wang³, Muqing Zhang⁴, Hongbo Li⁵, Guo-Bo Chen⁶, Chris Maliepaard⁷ & Xingtang Zhang¹✉

Genome-wide association studies in complex polyploids are hindered by genotyping ambiguity and allele dosage complexity. Here we present KMERIA, a *k*-mer-based framework specifically designed to address these challenges, enabling efficient genotyping and robust association mapping in complex polyploid genomes. Rigorous benchmarking with simulated and empirical datasets demonstrates that KMERIA surpasses existing methods in accuracy and statistical power. By applying KMERIA to 290 wild sugarcane (*Saccharum spontaneum*) accessions and integrating a 15-accession graph pangenome to capture structural variations, we identified new genes regulating sucrose biosynthesis (*SsMGT*) and tillering (for example, *SsERF14*, *SsNGAS*, *SsNAC*, *SsARF8*, *SsLOG* and *SsSCR*). These findings elucidate the genetic architecture of yield-related traits and provide actionable targets for sugarcane breeding. Collectively, KMERIA bridges a critical methodological gap in polyploid genomics, while our graph-pangenome integration provides a powerful framework for deciphering genotype–phenotype relationships in crops with complex architectures.

Polyploidy, the presence of multiple sets of chromosomes, is a defining feature of many cultivated crops. Remarkably, approximately 75% domesticated plants, including staples such as wheat, potato, cotton and sugarcane, are polyploids^{1,2}. This genetic phenomenon has driven the development of traits that confer advantages over their diploid counterparts^{3–5}, including increased vigor, larger organs and enhanced resilience against environmental stresses^{6,7}. Understanding the genetic architecture of polyploids is therefore critical for advancing crop improvement and ensuring global food security.

Despite their agricultural significance, the genetic complexity of polyploids poses substantial challenges for traditional genome-wide association studies (GWASs). Conventional GWAS methods, primarily developed for diploid organisms, often fail to adequately address the redundancy and allelic diversity of polyploid genomes, hindering detection of nonadditive effects and dosage-dependent phenotypic contributions. Although allele dosage-based GWAS using tools like GWASpoly has been implemented in tetraploids and hexaploids^{8–10}, these approaches demand extensive computational resources for

¹State Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. ²State Key Laboratory for Tropical Crop Breeding, Sugarcane Research Institute, Yunnan Academy of Agricultural Sciences, Kaiyuan, China. ³State Key Laboratory of Tropical Crop Breeding, Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Sanya, China. ⁴Guangxi Key Laboratory for Sugarcane Biology, Guangxi University, Nanning, China. ⁵College of Horticulture Science and Engineering, Shandong Agricultural University, Tai'an, China. ⁶Center for Laboratory Medicine, Department of Genetic and Genomic Medicine, and Clinical Research Institute, Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou, China. ⁷Plant Breeding, Wageningen University and Research, Wageningen, the Netherlands. ⁸These authors contributed equally: Shuai Chen, Xinlong Liu. ✉e-mail: zhangxingtang@caas.cn

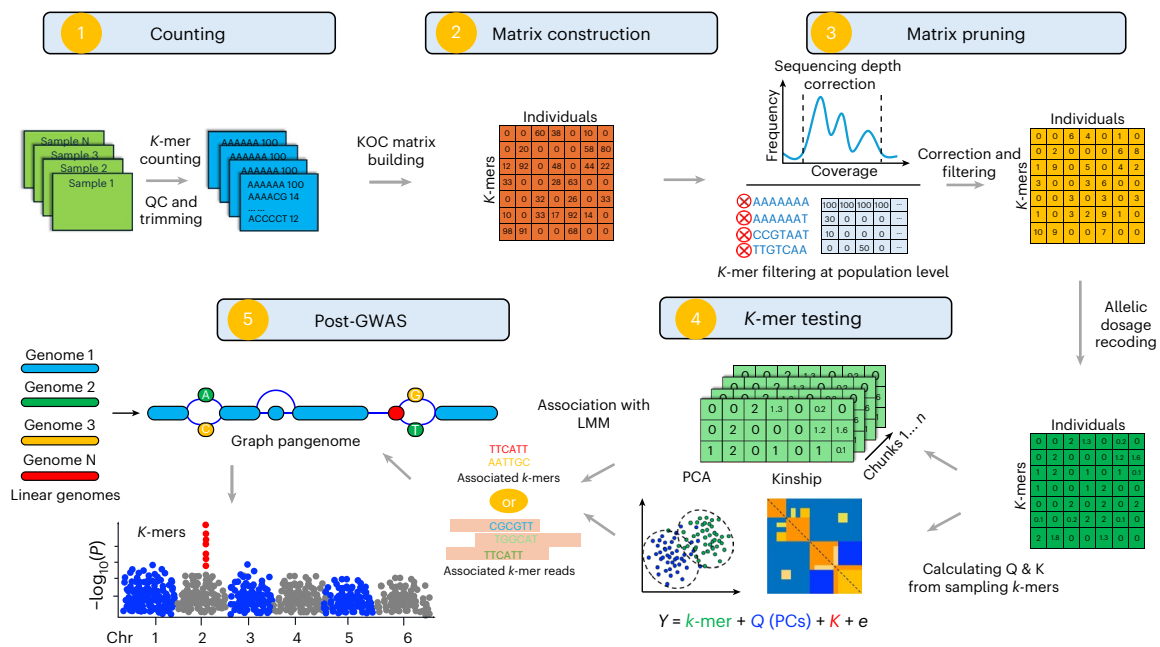


Fig. 1 | Overview of the KMERIA pipeline. KMERIA workflow includes five steps—counting, matrix construction, matrix pruning, *k*-mer testing and post-GWAS. After QC of resequencing samples, *k*-mer abundances were quantified per sample. These counts were compiled into a raw *k*-mer abundance matrix, which underwent ploidy-aware filtering to remove *k*-mers with aberrant depth (outside ploidy-informed thresholds). Sequencing depth correction was then applied to

retained *k*-mers. Additional filtering excluded rare *k*-mers (missing rate >20% across samples). The filtered matrix was recoded to a continuous allelic dosage scale (0–2) through quantile normalization, ensuring compatibility with mixed linear models. Finally, significantly associated *k*-mers or *k*-mer-related reads were mapped to a graph-based reference genome to resolve precise genomic coordinates. QC, quality control; PCA, principal component analysis.

variant calling and genotyping. Consequently, innovative strategies specifically designed for the unique genetic architecture of polyploids are urgently needed.

GWAS can leverage *k*-mers derived from whole-genome sequencing data as genetic markers. Unlike traditional markers, *k*-mers capture single-nucleotide polymorphisms (SNPs), indels and structural variations (SVs), enabling comprehensive genetic analysis independent of a reference genome and reducing alignment biases. This approach can address the ‘missing heritability’ problem by uncovering genetic factors that conventional methods overlook⁴¹. While promising in diploids, *k*-mer-based GWAS in polyploids has been limited, as current implementations fail to account for allele dosage effects^{12,13}. In this context, variations in *k*-mer copy numbers across populations provide a robust means of assessing allele dosage, making *k*-mer-based GWAS particularly advantageous for polyploids.

Saccharum spontaneum, a wild sugarcane species, serves as a valuable germplasm resource for modern commercial hybrid breeding due to its advantageous traits such as high tillering capacity, robust stress tolerance and broad environmental adaptability¹⁴. Its genome is characterized by high ploidy (4–16×), extensive homologous regions, and frequent chromosomal rearrangements¹⁵, making it an ideal model for studying polyploid-specific genomic methods. Here we present KMERIA, a *k*-mer-based GWAS strategy specifically designed for polyploids. KMERIA identifies trait-associated loci by correlating *k*-mer copy number variations with phenotypes. We also constructed a graph-based pangenome from 15 *S. spontaneum* accessions and 290 resequenced individuals, using KMERIA to identify genes linked to tiller number (TN) and apparent purity (AP). We further validated KMERIA across crops with varying ploidy levels, demonstrating its broad applicability in polyploid crop improvement.

Results

Challenges in applying GWAS to polyploids

GWASs hold immense promise for polyploid crops, but substantial challenges remain, particularly in genotyping. A major obstacle is the

accuracy of short-read alignment. High sequence similarity among homologous chromosomes leads to multimapping, resulting in misalignments and reference bias (Extended Data Fig. 1a). Using simulated data from the haplotype-resolved genome of *Saccharum hybrid* POJ2878 ($2n = 118$)¹⁶, we found that only 43.08% reads were high-quality alignments; among them, 50.51% were misaligned to incorrect genomic locations (Extended Data Fig. 2 and Supplementary Table 1). This introduces substantial errors in variant calling.

A second challenge is polyploid genotype calling. Although tools like updog¹⁷, SuperMASSA¹⁸ and fitPoly¹⁹ estimate genotypes from read depth but require pre-identified SNPs, accurate variant discovery itself is difficult in polyploids. Widely used tools such as GATK²⁰ rely on high-quality alignments that are often unattainable, and their Bayesian genotype inference becomes computationally prohibitive as the number of possible allele combinations grows exponentially with ploidy (Extended Data Fig. 1b,c and Supplementary Note). Insufficient sequencing depth further obscures allele dosage ratios, leading to misclassification or missed low-frequency alleles²¹.

Algorithm overview of KMERIA and validation datasets

KMERIA circumvents these limitations through the following five integrated steps (Fig. 1): (1) ‘counting’ converts resequencing data into whole-genome *k*-mer profiles, (2) ‘matrix construction’ builds population-level *k*-mer genotype matrices that preserve dosage information, (3) ‘matrix pruning’ applies abundance thresholding, repeat masking and ploidy-aware normalization to generate high-confidence dosage matrices, (4) ‘*k*-mer testing’ uses a linear mixed model (LMM) incorporating population structure and kinship to identify dosage-dependent associations; and (5) ‘post-GWAS’ applies dual correction (Benjamini–Hochberg false discovery rate (FDR) and modified Bonferroni) and aligns significant *k*-mers to a graph-based pangenome for precise localization in rearrangement-prone genomes. This structural-aware mapping facilitates precise localization of causal elements in rearrangement-prone polyploid genomes, facilitating the

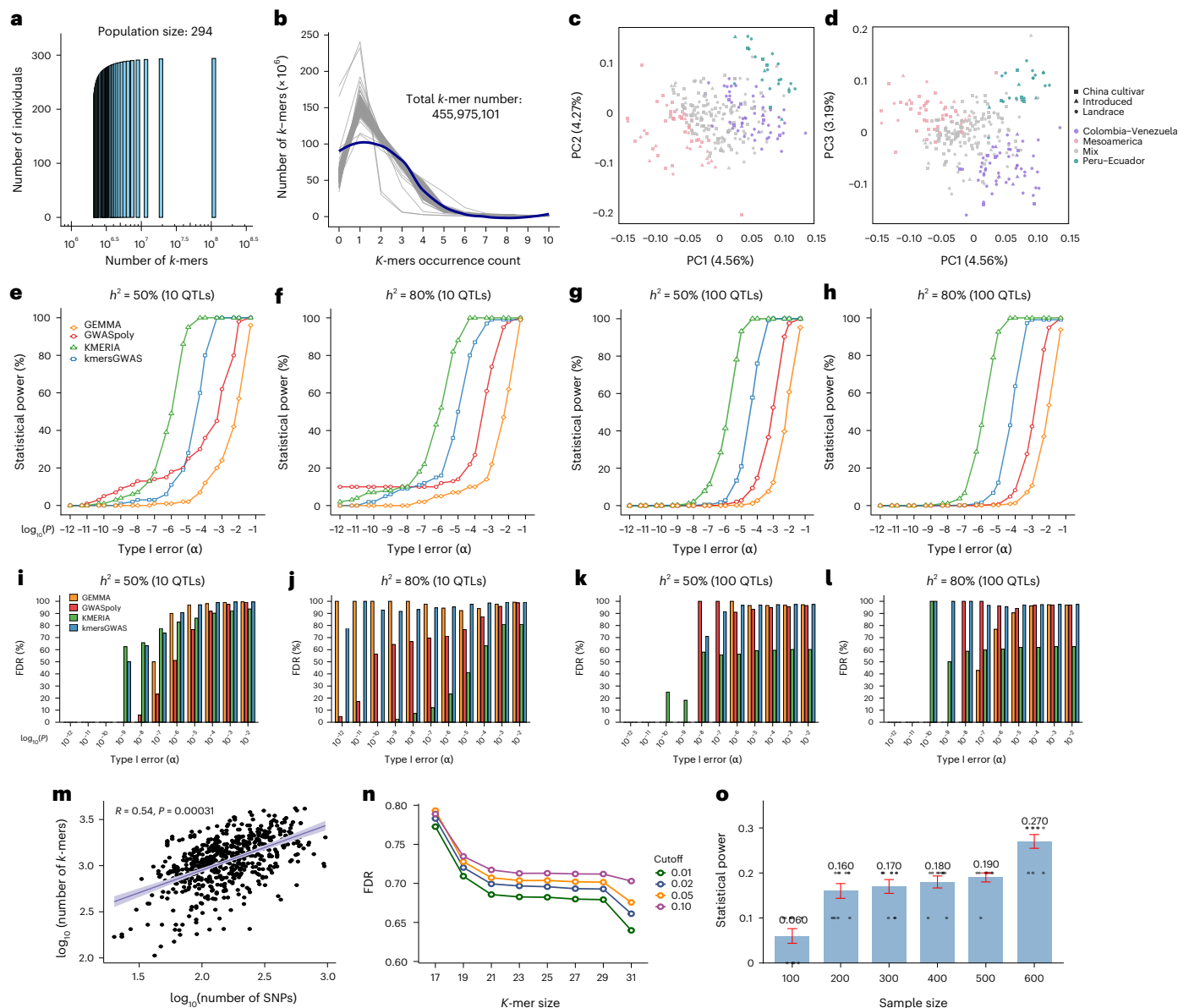


Fig. 2 | Performance of KMERIA in simulated polyploid phenotype datasets under different scenarios. **a**, The distribution of filtered k -mer numbers in 294 sweet potato accessions. **b**, The distribution of total normalized KOC per k -mer summed from 294 sweet potato accessions. **c, d**, PCA of sweet potato accessions. PC1 versus PC2 (**c**) and PC1 versus PC3 (**d**). **e–h**, Statistical power at various levels of Type I error under four simulation scenarios: 10 QTLs with 50% heritability (**e**), 10 QTLs with 80% heritability (**f**), 100 QTLs with 50% heritability (**g**) and 100 QTLs with 80% heritability (**h**). Additive effects were simulated using 10 and 100 QTLs randomly selected from variant sites. Phenotypes with 50% and 80% heritability were generated by adding normally distributed residuals to genetic effects. Statistical power was evaluated across multiple type I error significance thresholds. True positives were defined as SNPs or k -mers within ± 50 kb of a QTL; all other significant SNPs or k -mers were false positives. Four methods were used to test these populations—GEMMA, GWASpoly, kmersGWAS and KMERIA.

i–l, The bar plots illustrate the FDR under the corresponding scenarios: 10 QTLs with 50% heritability (**i**), 10 QTLs with 80% heritability (**j**), 100 QTLs with 50% heritability (**k**) and 100 QTLs with 80% heritability (**l**). **m**, Correlation of P values between the top k -mers and SNPs. Individual points represent individual numbers ($n = 691$). The solid line represents the linear regression fit, and the shaded area indicates the 95% confidence interval. Pearson's correlation coefficient R and P value are shown (two-sided t test). **n**, The point plot is used to assess the effect of k -mer size (ranging from 17 to 31) on the FDR. **o**, The bar plot indicates the statistical power in *Arabidopsis* populations with varying population sizes (100, 200, 300, 400, 500, 600). Data are presented as mean $\pm$ s.e.m. The heights of the blue bars represent the mean statistical power, with the exact values indicated above each bar. Individual gray dots represent independent replicates ($n = 10$). Red error bars denote the s.e.m.

integration of haplotype-resolved gene networks and *cis*-regulatory landscape analysis for trait dissection.

To validate the performance of KMERIA, we used simulated phenotypes in established polyploid populations and real datasets from hexaploid sweet potato⁸, tetraploid potato⁹, tetraploid alfalfa²², tetraploid *Arabidopsis*²³, diploid rice²⁴ and cultivated sugarcane¹⁶ (Supplementary Table 2 and Supplementary Fig. 1), highlighting

the broad applicability of KMERIA across diverse genetic characteristics of species.

Statistical power and FDR

We assessed statistical power and FDR using a simulated hexaploid sweet potato population of 294 accessions under varying heritability ($h^2 = 50\%$, 80%) and quantitative trait locus (QTL) numbers

(10 or 100). Genetic markers were classified as residing within QTL regions or non-QTL regions for this assessment. We first characterized the genetic features of population through *k*-mer analysis. This yielded 450 million nonredundant *k*-mers across all individuals. Only 100 million (22.2%) of these *k*-mers were present in all of the genomes (Fig. 2a), indicating substantial interindividual *k*-mer variation. The distribution of total *k*-mer occurrence counts (KOCs) revealed that the vast majority of *k*-mers had low KOCs (1–6; Fig. 2b). This demonstrates that *k*-mer KOCs effectively represent allelic dosage in polyploids. Subsequently, principal component analysis performed on the KOC matrix produced population clusters consistent with published studies (Pearson correlation with PC1– $R = 0.966$, $P = 9.23 \times 10^{-174}$; PC2– $R = 0.735$, $P = 2.73 \times 10^{-51}$ (Fig. 2c,d and Supplementary Fig. 2a,b)), validating *k*-mer KOCs as an alternative for population structure study.

Phenotypic simulated datasets were designed to reflect traits controlled by either a few major-effect genes (10 QTLs) or multiple genes (100 QTLs), with heritability levels of 50% and 80%. Statistical power was defined as the proportion of true QTLs detected at a given type I error rate (α), while FDR was calculated as the proportion of false positives among detected association loci. KMERIA was compared against three widely used GWAS methods—GEMMA²⁵, a standard GWAS approach for diploids, a *k*-mer-based GWAS approach kmersGWAS¹² and GWASpoly¹⁰, an algorithm specifically designed for autopolyploids.

For traits influenced by a small number of major-effect genes, we simulated ten QTLs with effects following a Gaussian distribution (Supplementary Fig. 3a,b). Across heritability levels of 50% and 80%, KMERIA consistently outperformed GEMMA, kmersGWAS and GWASpoly in terms of statistical power while maintaining lower Type I error rates (Fig. 2e,f). These results demonstrate the higher power of KMERIA in detecting major-effect QTLs. In scenarios involving polygenic traits influenced by 100 QTLs (Supplementary Fig. 3c,d), KMERIA again exhibited higher statistical power compared to diploid GWAS approach (GEMMA), kmersGWAS and GWASpoly, regardless of heritability levels (Fig. 2g,h and Extended Data Fig. 3a–d).

To further assess statistical performance, we also compared FDR between KMERIA and established GWAS approach by recording false-positive QTLs across ten repeated tests. KMERIA consistently demonstrated lower FDR than GEMMA, kmersGWAS and GWASpoly across most scenarios (Fig. 2i–l), highlighting its robustness and reliability in reducing spurious associations.

Additionally, we compared the loci detected by KMERIA and GWASpoly to evaluate their overlap and unique contributions. KMERIA showed a strong correlation ($R = 0.54$, $P = 0.00031$) with GWASpoly in identifying significant association loci (Fig. 2m). Notably, KMERIA not only identified shared QTLs but also detected additional association signals that were missed by GWASpoly, suggesting the robustness and superiority of KMERIA.

Factors influencing statistical power of KMERIA

To assess the robustness of KMERIA, we systematically examined several factors that may influence its statistical power—(1) the proportion of *k*-mers used to infer population structure in the LMM, (2) the optimal *k*-mer size for GWAS analysis, (3) the effect of missing *k*-mer markers on polyploid GWAS outcomes and (4) the impact of population size on polyploid GWAS results.

KMERIA uses a random sampling strategy to select a subset of *k*-mer markers from the full *k*-mer genotype matrix for inferring population structure and relatedness in LMM. This approach reduces computational burden and mitigates overfitting by eliminating redundant information. To determine the minimum number of *k*-mers required for population structure calculation, we randomly selected 5%, 1% and 0.1% *k*-mer tags from the tetraploid potato genotype matrices to perform principal component analysis as a measure of population structure. The results showed that the first two principal

components (PC1 and PC2) derived from the three sampling subsets exhibited strong correlations ($R = 0.99$, $P = 6.7 \times 10^{-116}$) and aligned well with the known population structure⁹ (Supplementary Fig. 4a–e). Notably, even with just 0.1% *k*-mer genotypes, the core features of the population structure were preserved. These results demonstrate that a small, randomly sampled subset of *k*-mers is sufficient for population structure analysis, enabling computational efficiency without sacrificing accuracy.

The selection of *k*-mer size *k* is a potential factor influencing GWAS performance. Using a tetraploid potato dataset, we systematically evaluated *k*-mer sizes ranging from 17 to 31 bp, assessing their impact on association mapping. Statistical power and empirical FDR were calculated using modified Bonferroni-corrected significance thresholds ($P < \alpha \times \frac{k}{M}$, where *M* is the total number of *k*-mers). Among the tested values, 31-mer exhibited the lowest FDR, indicating that larger *k*-mer sizes enhance signal specificity in polyploid GWAS (Fig. 2n and Supplementary Fig. 5a–i), suggesting that larger *k*-mers capture functional genetic variation more effectively.

We also assessed the effect of missing *k*-mer genotypes on GWAS performance using tetraploid potato datasets with increasing rates of genotype missingness rates (20%, 50% and 80%) under a fixed *k* value of 31 bp. As the rate of missingness increased, statistical power decreased (Supplementary Fig. 6a–c), primarily due to reduced accuracy in kinship estimation. This finding highlights the importance of genotype completeness in maintaining the reliability of polyploid GWAS.

Population size is a major determinant of GWAS statistical power. To investigate its impact in polyploid GWAS, we used an *Arabidopsis* population with sample sizes of 100, 200, 300, 400, 500 and 600 individuals. With a fixed *k* value of 31, we conducted ten replicate tests for each population size. Results demonstrated a clear positive correlation between population size and statistical power, with larger populations yielding significantly improved detection of QTLs (Fig. 2o). This trend aligns with findings from diploid GWAS studies²⁶, reaffirming that increasing sample size is an effective strategy to enhance power in polyploid GWAS.

KMERIA also demonstrated substantial computational efficiency. On a tetraploid potato population (100 individuals), GATK (-ploidy mode) required ~35,413 h, while KMERIA completed the same task in 38.9–82.3 h (~1/430 of the time), with comparable peak memory usage (Supplementary Table 3 and Supplementary Figs. 7 and 8a,b). The association module bimbamAsso reduced runtime to ~1/10 of GEMMA and ~1/27 of GWASpoly, using only 1/5 and 1/54 of their peak memory, respectively (Supplementary Fig. 8c,d). These results highlight the superior performance of KMERIA in handling large-scale association mapping with minimal computational overhead.

Application and validation of KMERIA in GWAS across crops with different ploidy levels

To demonstrate the versatility and effectiveness of KMERIA across crops with varying polyploid levels, we analyzed publicly available resequencing datasets from five species—diploid rice²⁴, tetraploid alfalfa²² and potato⁹, hexaploid sweet potato⁸ and highly polyploid cultivated sugarcane¹⁶, focusing on key agronomic traits.

We first tested KMERIA in a diploid rice population comprising 400 accessions²⁴, focusing on amylose content, a critical determinant of grain quality. KMERIA successfully re-identified *Waxy*²⁷ gene, a well-known major locus controlling amylose content. Compared to the conventional GWAS method and previous kmersGWAS, KMERIA exhibits significantly enhanced association signals for the trait with superior statistical validation (KMERIA– $P = 6.76 \times 10^{-27}$; ref. 24– $P = 4.51 \times 10^{-17}$; Fig. 3a and Extended Data Fig. 4).

To evaluate the performance of KMERIA in tetraploid crops, we applied it to two economically important species—alfalfa and potato. In alfalfa, we analyzed 176 accessions for the stem-to-leaf ratio, a key trait influencing forage quality and plant morphology. KMERIA identified

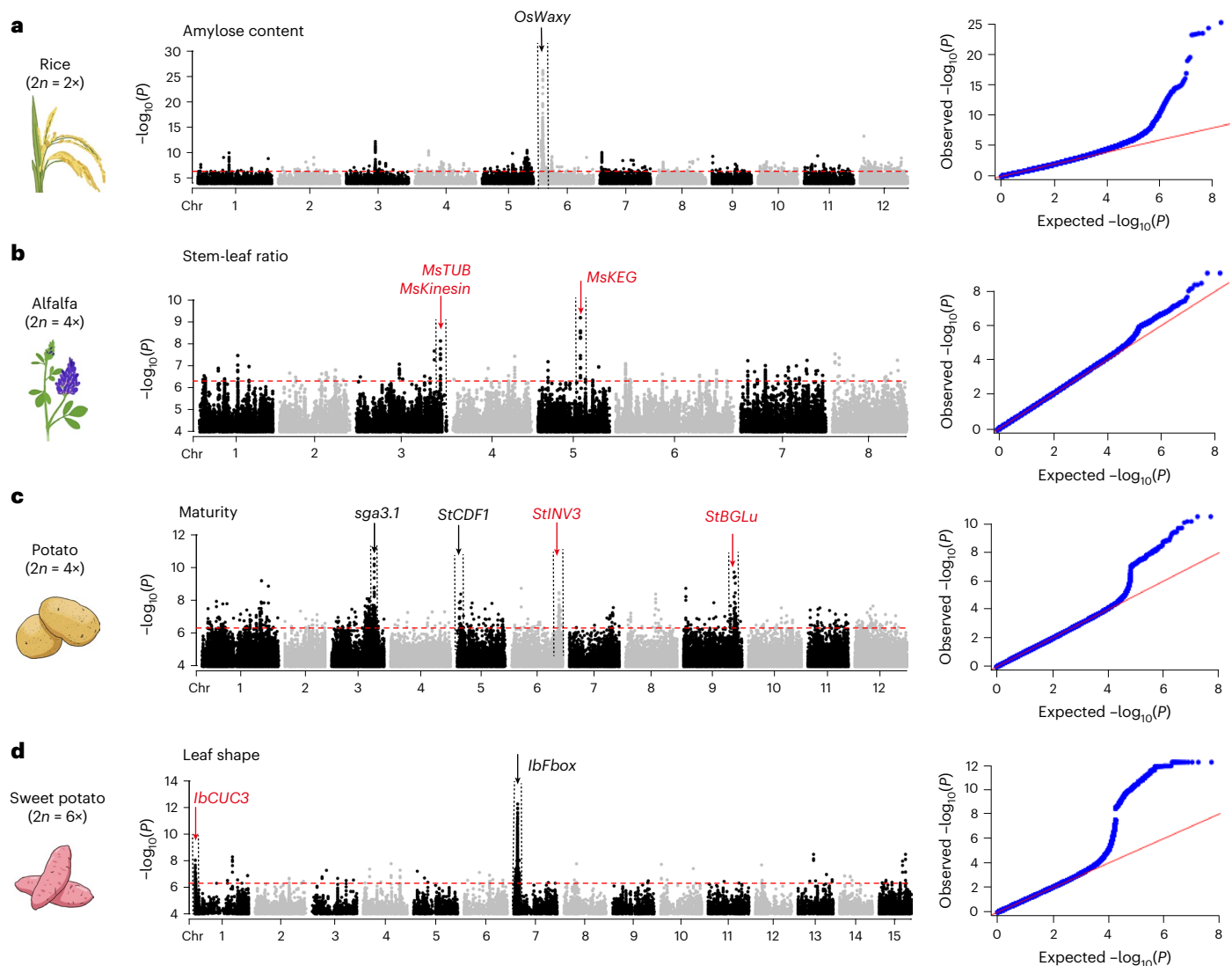


Fig. 3 | Application performance of KMERIA methods across multiple ploidy crop mapping populations. Manhattan plots displaying GWAS results for different ploidy crop populations using KMERIA. **a–d**, The GWAS association results for various traits—amylose content of diploid rice (**a**), stem-leaf ratio of tetraploid alfalfa (**b**), flesh color of tetraploid potato (**c**) and leaf shape of hexaploid sweet potato (**d**). Left, a schematic of each crop and its corresponding genome ploidy, with the Manhattan plot in the center. The regions highlighted by vertical dashed lines indicate significant QTLs, with black arrows representing QTLs previously reported in the literature and red arrows indicating new

QTLs identified in this study. Genes significantly associated with the traits are highlighted in red within the QTLs. *P* values were calculated using a two-sided Wald test. The horizontal dashed lines indicate the GWAS significance threshold, adjusted for multiple comparisons using BH-FDR method. Right, the QQ plot for each GWAS result. Owing to the prohibitively large number of *k*-mers in the full dataset, a random sample of 25,000,000 *k*-mers was used to construct the QQ plot. BH-FDR indicates the false discovery rate controlled using the Benjamini–Hochberg procedure.

two new QTLs on chromosomes 3 and 5, containing *MsTUB5* (ref. 28), *KINESIN*-like²⁹ and *MsKEG*³⁰ (*KEEP ON GOING*) genes known to regulate plant morphogenesis (Fig. 3b and Extended Data Fig. 5a). Notably, *k*-mers associated with these genes exhibited significant differences in dosage across phenotypic groups, demonstrating that KMERIA effectively captures allelic dosage variation in a tetraploid context (Extended Data Fig. 5b,c). In potato, we analyzed 137 accessions with an average sequencing depth of 67×, focusing on several key agronomic traits. For tuber maturity, KMERIA successfully recaptured established QTLs (or genes), such as *sga3.1* (ref. 31) and *StCDF1* (ref. 32). Beyond these known loci, KMERIA identified a series of new genes involved in starch and sucrose metabolism, such as *StINV3* (ref. 33) and *StBGLu*³⁴. Furthermore, we identified several genes associated with tuber skin color (*StUGP2* (ref. 35), *StOPR*³⁶, *StDOX*³⁷) and tuber flesh color—a primary objective in potato quality breeding. For flesh color, KMERIA not only detected previously reported QTLs on

chr3 (ref. 38) and chr6 (ref. 9) but also uncovered new QTLs on chr4 and chr9, harboring functional genes *StAUR*³⁹ and *St3GT*⁴⁰ involved in root development and flavonoid biosynthesis, respectively (Fig. 3c and Extended Data Fig. 6a–c). While traditional SNP-based GWAS failed to detect many of these new loci, the previous kmersGWAS approach yielded several QTLs colocalized with KMERIA. Furthermore, *k*-mer dosage analysis suggested that these specific traits are relatively insensitive to allele dosage (Extended Data Fig. 6d,e). However, for traits significantly influenced by allele dosage, kmersGWAS exhibited a higher false-positive rate (Extended Data Fig. 6b,f). Collectively, these results underscore the superior detection power and robustness of KMERIA in the complexities of autopolyploid genomes.

In the sweet potato population, we applied KMERIA to 294 accessions with published whole-genome resequencing data, achieving an average sequencing coverage of 167× relative to the monoid genome⁸. We identified 472 million nonredundant *k*-mers present

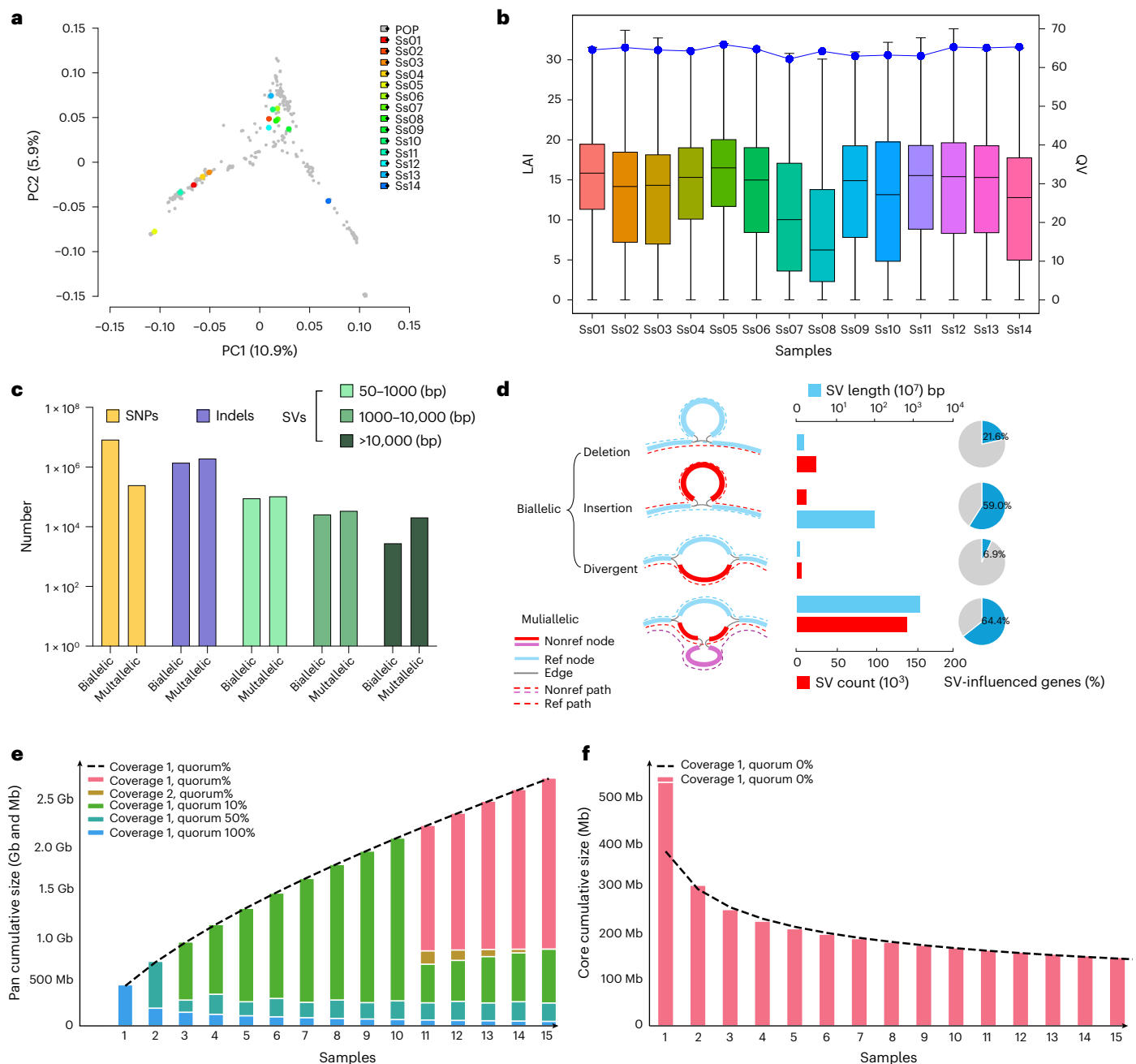


Fig. 4 | Construction of graph-based pangenome map from the 14 *S. spontaneum*.

a, The 290 resequenced *S. spontaneum* genomes were subjected to PCA to assess their genetic diversity. The colored dots represent the cultivars selected for pangenome analysis, which collectively capture the broad genetic diversity of *S. spontaneum*. 'POP' stands for '*S. spontaneum* population'. **b**, Assembly quality assessment of 14 *S. spontaneum* species was performed using the LAI score and QV, with the *k*-mer-based strategy implemented in Merquy (that is, quality consensus). The upper and lower edges of the box plots represent the first and third quartiles, respectively, the centerline denotes the

median and the whiskers extend to 1.5× the IQR. **c**, Bar plots show the number of genetic variants of different types in the *S. spontaneum* genome, with counts for biallelic and multiallelic variants presented separately. **d**, Diagrams of diverse SV types from the bovine graph-based pangenome, based on the *S. spontaneum* monoploid reference genome. The bar chart displays the total number and length of each SV type, while the pie chart illustrates the proportion of genes affected by SV relative to the total number of genes. **e**, Total length of graph-based pangenome sequences with varying numbers of individuals. **f**, Total length of the core genome with varying numbers of individuals. IQR, interquartile range.

in at least 50 accessions. Among the significant QTLs detected, we successfully re-identified *lbFbox*, a previously reported gene associated with leaf shape^{8,41} (Fig. 3d and Extended Data Fig. 7a), and a new leaf shape-associated gene encoding cup-shaped cotyledon (*lbCUC3*), which is a cell growth inhibitor⁴² (Fig. 3d). Notably, the leading *k*-mer identified by KMERIA exhibited lower *P* value than those obtained using the GWASpoly¹⁰ and kmersGWAS (Extended Data Fig. 7a). Similar superiority was observed for two additional traits (Extended Data Fig. 7b–f),

highlighting the ability of KMERIA to detect both known and new associations in hexaploid genomes.

To further validate the utility of KMERIA in highly complex genomes, we tested KMERIA in cultivated sugarcane, one of the most genomically complex crop species due to its high polyploidy and aneuploidy. This population consisted of 603 individuals with an average sequencing depth of 118× (ref. 16). While SNP-GWAS and kmersGWAS failed to detect any leading signals exceeding the threshold, KMERIA

Table 1 | Summary of *S. spontaneum* genome assemblies and gene annotations

Accessions	Accession name	Species	Estimated ploidy	Genome length (bp)	N50 (bp)	Assemblies of BUSCO (%)	Number of genes	Annotation of BUSCO (%)	LAI	QV
Ss01	fujian88-1-4	<i>S. spontaneum</i>	10	8,368,781,140	2,484,927	99.4	296,879	97.9	15.42	64.5482
Ss02	fujian89-1-7	<i>S. spontaneum</i>	11	9,324,246,601	1,791,182	99.4	332,535	98.3	15.49	65.1491
Ss03	guangdong8hao	<i>S. spontaneum</i>	10	7,939,614,880	1,519,160	99.3	284,471	98.7	14.53	64.5211
Ss04	guangxi86-8	<i>S. spontaneum</i>	10	7,813,376,143	1,802,156	99.3	282,072	98.4	14.47	64.2456
Ss05	guangxi87-51	<i>S. spontaneum</i>	9	7,440,054,966	2,319,968	99.4	271,315	98.6	15.86	65.914
Ss06	hainan92-5	<i>S. spontaneum</i>	10	7,603,310,039	1,834,047	99.2	282,165	98.2	14.9	64.7646
Ss07	sichuan79-1-19	<i>S. spontaneum</i>	10	8,146,733,135	997,139	99.5	284,006	98.3	15.38	62.2023
Ss08	sichuan88-22	<i>S. spontaneum</i>	10	8,502,472,194	733,173	99.3	286,578	98.1	15.18	64.1903
Ss09	yunnan82-149	<i>S. spontaneum</i>	8	6,518,384,450	1,048,187	99.7	231,354	98.8	15.66	62.9316
Ss10	yunnan82-132	<i>S. spontaneum</i>	10	7,640,365,253	1,694,461	99.4	274,973	98.8	16.43	63.1965
Ss11	yunnan82-48	<i>S. spontaneum</i>	10	7,952,106,007	1,729,590	99.5	283,350	98.9	15.54	62.9776
Ss12	yunnan82-84	<i>S. spontaneum</i>	10	8,760,448,060	1,741,544	99.5	307,442	98.6	15.77	65.2838
Ss13	yunnanheqinggeshoumi	<i>S. spontaneum</i>	10	7,900,885,789	1,668,255	99.3	280,316	99.1	15.05	65.0709
Ss14	yunnankaiyuangeshoumi	<i>S. spontaneum</i>	10	7,763,673,639	1,388,620	99.5	283,978	98.8	14.84	65.3216

identified multiple QTLs linked to key agronomic traits, including tiller angle, plant height, soil-plant analysis development and TN (Extended Data Fig. 8a–d). This stark contrast highlights the ability of KMERIA to overcome the limitations of traditional SNP-based methods.

Construction of *S. spontaneum* graph pangenome

S. spontaneum has made substantial genetic contributions to modern sugarcane breeding, providing key genetic factors for enhanced tillering, improved sucrose synthesis and higher photosynthetic efficiency, all of which contribute to increased yield, sugar accumulation and biomass production¹⁴. To maximize the improvement of genetic diversity in *S. spontaneum*, we selected 14 representative accessions based on *k*-mer-based phylogenetic relationships of 290 *S. spontaneum* accessions (Fig. 4a and Extended Data Fig. 9). These accessions span across 11 counties and geographical regions (Supplementary Table 4). Given the difference in genomic ploidy among the *S. spontaneum* accessions, we estimated the ploidy levels of these 14 accessions using resequencing reads. The results indicated that ploidy levels ranged from 8 to 11, with decaploid constituting 11 of 14 selected samples (Supplementary Table 4). Notably, decaploids also dominated among the 290 resequenced *S. spontaneum* accessions, accounting for 68.9% (200/290) of these samples. Consequently, we generated an average of 183 Gb HiFi reads and 225× genome coverage on average for the 14 accessions (the *S. spontaneum* monoploid genome size of ~800 Mb; Supplementary Table 5). These HiFi reads were assembled at the contig level with heterozygous regions retained. The assembly size ranged from 6.52 Gb (Ss09) to 9.32 Gb (Ss02), with the average contig N50 of 1.63 Mb (Table 1 and Supplementary Table 6). The completeness of assemblies was supported by using 1,614 benchmarking universal single-copy orthologs (BUSCO)⁴³, with a completeness of 99.4% on average (Table 1). Merquy quality value (QV; 64.3 on average, QV 30 represents 99.9% base accuracy) evaluation also suggested that the base accuracy and completeness of assemblies arrived a high quality. The long terminal repeat (LTR) Assembly Index (LAI)⁴⁴ of the assembled genomes are from 14.47 to 16.43, indicating that these assemblies have reached a reference level (Fig. 4b and Table 1). This suggests high quality of the assembled genomes. We predicted 231,354 (Ss09, ploidy = 8) to 332,535 (Ss02, ploidy = 11) protein-coding genes (including allelic genes) through integrating ab initio-based, homology-based and RNA-seq prediction (Supplementary Tables 7 and 8). BUSCO⁴³ evaluation for gene consistency scored above 98%, indicating a high level of completeness and representation of core genes in these assemblies (Supplementary Table 7).

Due to the absence of Hi-C data, we performed genome deduplication for each *S. spontaneum* assembly to meet the requirements for graph pangenome construction. These deduplicated genomes were scaffolded using the *Sspon82-114* monoploid reference genome, yielding 14 monoploid assemblies with sizes ranging from 760 Mb to 772 Mb. BUSCO completeness assessment revealed an average score of 93.5%, confirming their suitability for graph pangenome construction (Supplementary Table 9). We subsequently built a graph-based pangenome incorporating these 14 monoploid genomes and the telomere-to-telomere *Sspon82-114* monoploid assembly (Supplementary Table 9). This comprised ~3 Gb genome size spread across 142,571,025 nodes and connected by 193,033,606 edges, with nonreference nodes spanning 2,029 Mb (Fig. 4c–f). We detected SVs (≥50 bp) in the graph-based genome using the bubble popping algorithm and retained 277,133 SVs (70.1% < 1 kb, 21.5% < 10 kb and 8.4% > 10 kb) that could be genotyped in the *S. spontaneum* monoploid reference genome or at least one other genome (Fig. 4c and Supplementary Table 10). These SVs were either biallelic (42.7%, insertions, deletions and divergent alleles) or multiallelic (57.3%, spanning 2.51 Gb) with more than one nonreference path (Fig. 4d and Supplementary Table 11). For genes present in the reference genome, we found that 118,552 biallelic SVs are in potential expression regulatory regions or CDS in the *S. spontaneum* genome. We also discovered multiallelic SVs influencing 63.1% reference genes (Fig. 4d and Supplementary Table 11). These results indicate large effects of SVs on the integrity and gene expression.

Identification of genes associated with agronomic traits in wild sugarcane

The complexity of the genome ploidy (4–16×) in *S. spontaneum* poses a challenge for accurate genotyping using traditional methods¹⁵, limiting the application of GWAS. To identify potential functional genetic variants associated with important agronomic traits in sugarcane, we used KMERIA for GWAS of two crucial traits—AP and TN (Fig. 5a–c and Supplementary Table 12). We constructed *k*-mer genotype matrices for 290 *S. spontaneum* accessions, identifying 12.4 billion nonredundant *k*-mers present in at least 30% individuals. A significant association signal for AP was detected at chr6–8.1 Mb on the *S. spontaneum* 82–114 reference genome (Fig. 5d). This locus harbors a gene encoding a *MAGNESIUM TRANSPORTER* (MGT), which is hypothesized to regulate sucrose synthesis by modulating Mg²⁺ transport, thereby influencing sugarcane quality^{45,46}. The top-associated

k-mer located at the upstream of *SsMGT* exhibited a significant negative correlation with AP ($R = -0.3$, $P = 2.0 \times 10^{-7}$; Fig. 5e). Further comparative genomic analysis within the *S. spontaneum* pangenome identified a 7.9-Kb SV and a 7-bp indel near the transcription start site of *SsMGT* in a distinct haplotype (Fig. 5f). Consistent with this, dosage analysis of SV-specific *k*-mers at the pangenome level revealed a negative correlation with AP ($R = -0.483$, $P = 0.0804$; Fig. 5g), suggesting that these SVs may impair *SsMGT* expression by disrupting transcriptional initiation. Such a mechanism could subsequently hinder Mg^{2+} transport efficiency and reduce sucrose accumulation.

The high tillering capacity of *S. spontaneum* has been a critical genetic resource for modern sugarcane breeding programs. To elucidate the genetic architecture underlying this agronomic trait, we also performed GWAS using KMERIA on TN trait (Fig. 5h). Our analysis identified 598 significantly associated *k*-mers ($P < 2.0 \times 10^{-7}$), with the most prominent signal mapped to an 18.9 Mb locus on chr1. This genomic region harbors ethylene-responsive factor (*SsERF*) gene orthologous to *OseATB* (*OseERF*)⁴⁷, which modulates gibberellin metabolism through transcriptional repression of ent-kaurene synthase A during internode elongation. Notably, this regulatory mechanism explains the observed inverse relationship between plant height and TN, a critical yield determinant, where dwarf cultivars frequently demonstrate higher productivity. The dosage effect of the top-associated *k*-mer exhibited a significant positive correlation with TN ($R = 0.38$, $P = 4.6 \times 10^{-11}$; Fig. 5i). Furthermore, we identified several tiller-regulatory genes, including *NGATHA5* (*SsNGAS*)⁴⁸, NAC domain transcription factor (*SsNAC20*)⁴⁹, auxin response factor (*SsARF8*)⁵⁰, lonely guy (*SsLOG*)⁵¹ and scarecrow (*SsSCL*)⁵², of which orthologs in *Oryza sativa* and closely related species have established roles in tiller development. Overexpression of *SsNGAS* in rice resulted in a significant 30.6% decrease in TN ($P = 0.025$; Extended Data Fig. 10a–c). Tissue-specific expression profiling across six *S. spontaneum* organs revealed preferential transcriptional activation of these candidate genes in root and tiller tissues, with comparatively lower expression levels observed in leaf and stem samples (Fig. 5j). These findings provide potential molecular targets for breeding programs aimed at improving sugarcane yield-related traits.

Using KMERIA, we not only identified key genetic determinants of agronomic traits in *S. spontaneum* but also demonstrated the critical utility of the *S. spontaneum* graph-based pangenome in resolving *k*-mer localization challenges inherent to autopolyploid genomes. In the TN trait, 83.4% (94,035/112,807) trait-associated *k*-mers ($P < 1 \times 10^{-4}$) were aligned to coordinate in the linear reference genome (*S. spontaneum* 82–114). However, 16.6% *k*-mers absent from the reference were successfully localized using the graph pangenome. Crucially, comparative analysis revealed that the graph pangenome approach enhanced associated *k*-mer mappability by 11.6 percentage points compared to linear

reference-based alignment (Supplementary Table 13), effectively recovering previously inaccessible loci in nonreference genomic regions.

Our study demonstrates that KMERIA effectively overcomes the inherent challenges of GWAS in complex autopolyploid genomes, offering a robust framework to pinpoint genetic loci underlying agronomic traits.

Discussion

The inherent complexity of polyploid genomes poses substantial challenges for traditional SNP-based GWAS methodologies. To address these limitations, we developed KMERIA, a *k*-mer-based GWAS framework specifically optimized for complex polyploid population. Its key innovation is circumventing conventional genotyping pipelines, where traditional alignment-based approaches²⁰ often struggle with allele combination complexity and multilocus genotyping in polyploids, KMERIA eliminates the need for error-prone read alignments and reference-dependent genotyping. This not only enhances genotyping efficiency but also minimizes reference bias, an advantage that is particularly critical when high-quality reference genomes are absent.

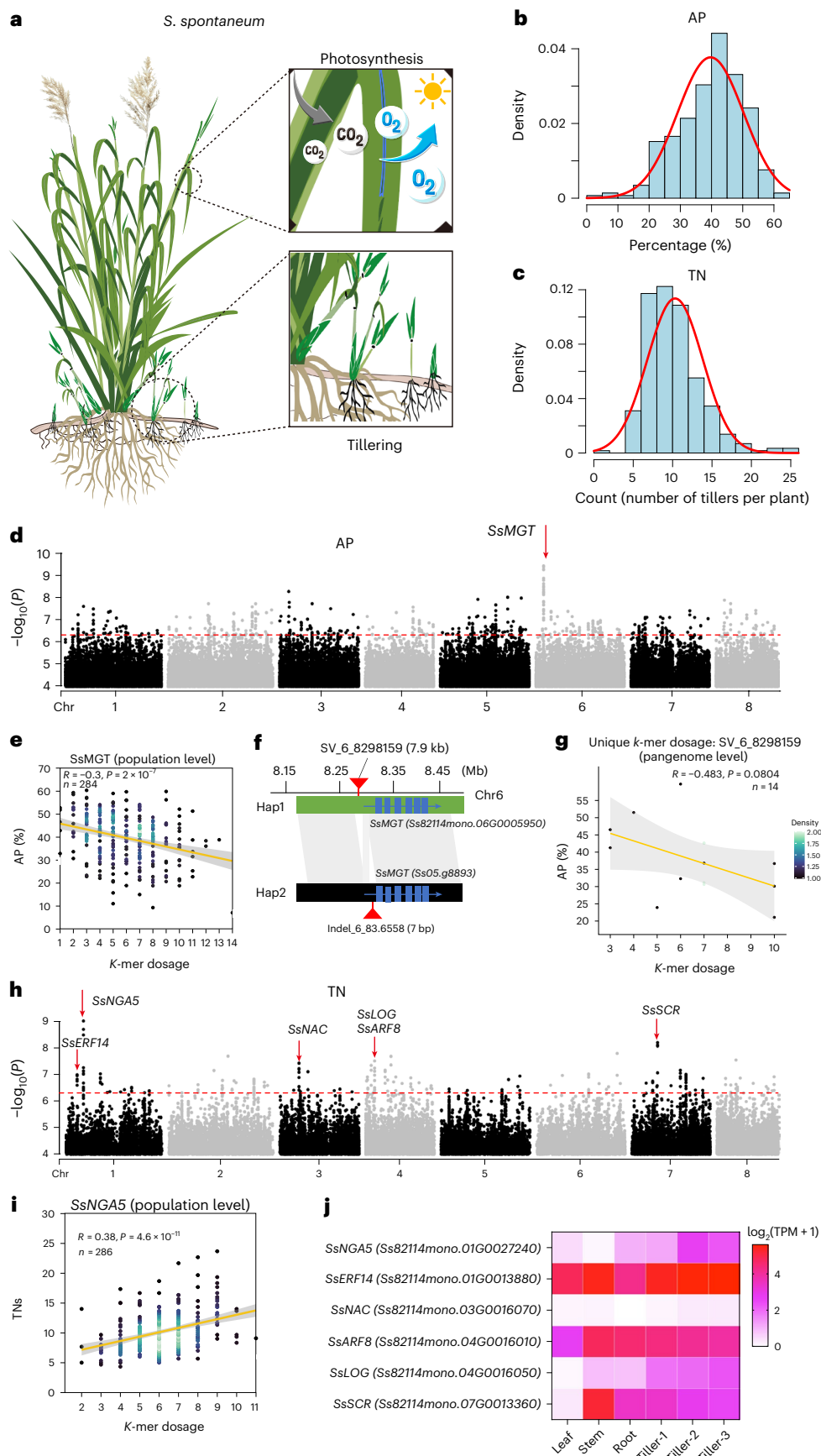
A pivotal strength of KMERIA is its capacity to quantify allele dosage, a parameter with profound phenotypic implications in polyploids. Unlike diploid *k*-mer GWASs that use binary (0 or 1) presence–absence *k*-mers encoding¹², KMERIA leverages *k*-mer abundance variations to resolve dosage states, thereby capturing genetic architecture of autopolyploid traits. This dosage-aware framework enables more precise genotype–phenotype associations compared to traditional SNP-based methods.

Furthermore, KMERIA provides a distinct advantage in genotyping non-SNP variants, including SVs and complex allelic variations, which are frequently overlooked by standard, SNP-centric approaches. By incorporating a graph-based pangenome into its downstream analysis, KMERIA enhances its capability to decipher complex variants. This underscores the importance of transitioning from linear reference genomes to graph-based representations, particularly for uncovering large-scale SVs that remain inaccessible through traditional methods. The ability to anchor previously unmappable *k*-mers, such as the 7.9-kb SV we identified in the promoter of *SsMGT* in sugarcane, enables the detection of genetic signals that would otherwise remain obscured, bridging the gap between association testing and the discovery of functionally relevant structural rearrangements. This capability is especially relevant for polyploid species, where such rearrangements are evolutionarily prevalent.

The pangenome approach further improved mapping ratio, consistent with previous studies^{53,54}. To systematically evaluate *k*-mer origin strategies, we compared single-reference, multireference

Fig. 5 | KMERIA methods accelerate the identification of genes associated with agronomic traits in *S. spontaneum*. **a**, Morphological view of *S. spontaneum* with detailed visualization of leaf photosynthetic characteristics, rhizome structure and tillering traits. **b**, Phenotypic distribution of AP trait across the population. **c**, Phenotypic distribution of TN trait across the population. **d**, Manhattan plot displaying the GWAS results for the AP trait. Red dashed lines indicate the BH-FDR correction significance thresholds for *k*-mer-based GWAS. Associated QTLs and candidate genes are highlighted with red arrow and dashed lines. *P* values were calculated using a two-sided Wald test. The horizontal dashed lines indicate the GWAS significance threshold, adjusted for multiple comparisons using BH-FDR method. **e**, *K*-scatterplot showing the relationship between the top *k*-mer dosage of the major-effect QTL on chromosome 06 and the AP trait ($n = 284$). Individual points represent separate samples, with colors indicating point density. The solid line represents the linear regression fit, and the shaded gray area denotes the 95% confidence interval. Pearson correlation coefficient *R* and the *P* value (two-sided *t* test) are indicated. **f**, Both a 7.9-kb SV (SV_6_8298159) and a 7-bp indel (InDel_6_8306558), highlighted by the red triangle, are located upstream of *SsMGT* (Ss82114mono.06G0005950). **g**, *K*-correlation analysis between the unique *k*-mer dosage of SV_6_8298159

and the AP phenotype. This analysis was performed at the pangenome level ($n = 14$). Individual dots represent sugarcane accessions, with colors indicating point density. The solid line represents the linear regression fit, and the shaded gray area denotes the 95% confidence interval. Pearson correlation coefficient (*R*) and *P* value (two-sided *t* test) are shown. **h**, Manhattan plot displaying the GWAS results for the TN trait. Red dashed lines indicate the BH-FDR correction significance thresholds for *k*-mer-based GWAS. Associated QTLs and candidate genes are highlighted with red arrows and dashed lines. *P* values were calculated using a two-sided Wald test. The horizontal dashed lines indicate the GWAS significance threshold, adjusted for multiple comparisons using BH-FDR method. **i**, *K*-correlation analysis between the top *k*-mer dosage of the major-effect QTL on chromosome 1 and the TN trait. This analysis was conducted at the population level ($n = 286$). The solid line indicates the linear regression fit, and the shaded gray area represents the 95% confidence interval. Individual points are colored by density. Pearson correlation coefficient (*R*) and *P* value (two-sided *t* test) are provided. **j**, Expression heatmap of TN trait candidate genes across different tissues (leaf, stem, root, tiller-1, tiller-2, tiller-3) in *S. spontaneum* accessions 82–114.



and graph-based pangenome mapping using GWAS *k*-mers from our *S. spontaneum* dataset (Supplementary Table 14). Single-reference mapping achieved only 81.2–84.6% anchoring, confirming the inadequacy of any single genome. The graph-based pangenome recovered 95.10% *k*-mers, while the multireference strategy reached 98.33%. The higher anchoring rate of the multireference approach reflects its ability to retain all sequence variations present across the constituent genomes, making it particularly suitable when exhaustive *k*-mer recovery is the primary goal. However, multireference mapping scatters hits across unlinked reference coordinates, preventing the direct assignment of *k*-mers to specific subgenomes, which limits dosage-sensitive interpretation. The graph pangenome, by providing a unified coordinate framework with explicit subgenomic paths, naturally resolves homeolog identity at the cost of a modest reduction in anchoring rate. The optimal choice is therefore goal dependent; multireference mapping excels at maximizing variant discovery, whereas the graph pangenome is preferable when biological interpretation demands precise homeolog resolution. A combined workflow can leverage the strengths of both approaches.

Despite these advancements, pinpointing causal variants remains challenging in polyploids. A key issue is the handling of *k*-mers mapping to multiple locations in pangenome. In polyploids, homeologous multimapping *k*-mers often reflect total allele dosage and should not be discarded. We classify significant *k*-mers into three interpretative categories. Uniquely mapped *k*-mers are treated at single-locus resolution, while those assigned to repetitive elements are excluded from single-locus estimation. Homeologous multimapping *k*-mers, whose aggregate counts capture total subgenome dosage, are retained as dosage signals; thus, this is the category most relevant to polyploid GWAS. Genic multicopy *k*-mers falling outside these classes require manual curation. This classification is best applied within a pangenome context; without subgenome annotations, a mapping-count threshold may serve as a proxy but risks misclassifying gene families. Although graph-based pangenomes mitigate reference bias, their adoption is constrained by limited polyploid resources, and residual artifacts may persist.

In conclusion, KMERIA represents a more efficient approach for polyploid GWAS, addressing longstanding bottlenecks in genotyping accuracy, dosage estimation and variant discovery. Its application to high-ploidy species, as exemplified by our *S. spontaneum* case study, underscores its potential to accelerate genetic research in polyploid crops and wild species. Future efforts to integrate expanding pangenome resources and refine *k*-mer-to-variant mapping algorithms will further enhance the utility of KMERIA in deciphering the genetic basis of complex traits in polyploid organisms.

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/s41588-026-02641-8>.

References

- Adams, K. L. & Wendel, J. F. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141 (2005).
- Kuckuck, H., Kobabe, G. & Wenzel, G. *Fundamentals of Plant Breeding* (Springer, 1991).
- Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
- Fox, D. T., Soltis, D. E., Soltis, P. S., Ashman, T.-L. & Van de Peer, Y. Polyploidy: a biological force from cells to ecosystems. *Trends Cell Biol.* **30**, 688–694 (2020).
- Otto, S. P. The evolutionary consequences of polyploidy. *Cell* **131**, 452–462 (2007).
- Booker, W. W. & Schridder, D. R. The genetic consequences of range expansion and its influence on diploidization in polyploids. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.10.18.562992> (2024).
- Orr-Weaver, T. L. When bigger is better: the role of polyploidy in organogenesis. *Trends Genet.* **31**, 307–315 (2015).
- Zhang, X. et al. Refining polyploid breeding in sweet potato through allele dosage enhancement. *Nat. Plants* **11**, 36–48 (2025).
- Li, H. et al. Genomic basis of divergence of modern cultivated potatoes. Preprint at *Research Square* <https://doi.org/10.21203/rs.3.rs-3968149/v1> (2024).
- Rosyara, U. R., De Jong, W. S., Douches, D. S. & Endelman, J. B. Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome* <https://doi.org/10.3835/plantgenome2015.08.0073> (2016).
- Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
- Voicheck, Y. & Weigel, D. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nat. Genet.* **52**, 534–540 (2020).
- Gaurav, K. et al. Population genomic analysis of *Aegilops tauschii* identifies targets for bread wheat improvement. *Nat. Biotechnol.* **40**, 422–431 (2022).
- Jackson, P. A. Breeding for improved sugar content in sugarcane. *Field Crops Res.* **92**, 277–290 (2005).
- Liu, X. et al. Phylogenetic analysis of different ploidy *Saccharum spontaneum* based on rDNA-ITS sequences. *PLoS ONE* **11**, e0151524 (2016).
- Wang, J. et al. Genetic architecture of sugarcane traits in a polyploid genomics framework. *Nature* <https://doi.org/10.1038/s41586-026-10576-7> (2026).
- Gerard, D., Ferrão, L. F. V., Garcia, A. A. F. & Stephens, M. Genotyping polyploids from messy sequencing data. *Genetics* **210**, 789–807 (2018).
- Serang, O., Mollinari, M. & Garcia, A. A. F. Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PLoS ONE* **7**, e30906 (2012).
- Voorrips, R. E., Gort, G. & Vosman, B. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* **12**, 172 (2011).
- McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
- Cooke, D. P., Wedge, D. C. & Lunter, G. Benchmarking small-variant genotyping in polyploids. *Genome Res.* **32**, 403–408 (2022).
- Zhang, F. et al. Evolutionary genomics of climatic adaptation and resilience to climate change in alfalfa. *Mol. Plant* **17**, 867–883 (2024).
- Bohutínská, M. et al. Polyploids broadly generate novel haplotypes from *trans*-specific variation in *Arabidopsis arenosa* and *Arabidopsis lyrata*. *PLoS Genet.* **20**, e1011521 (2024).
- Wei, X. et al. A quantitative genomics map of rice provides genetic insights and guides breeding. *Nat. Genet.* **53**, 243–253 (2021).
- Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
- Hong, E. P. & Park, J. W. Sample size and statistical power calculation in genetic association studies. *Genomics Inform.* **10**, 117–122 (2012).
- Wang, Z. et al. The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene. *Plant J.* **7**, 613–622 (1995).

28. Gavazzi, F. et al. Evolutionary characterization and transcript profiling of β -tubulin genes in flax (*Linum usitatissimum* L.) during plant development. *BMC Plant Biol.* **17**, 237 (2017).
29. Li, J., Xu, Y. & Chong, K. The novel functions of kinesin motor proteins in plants. *Protoplasma* **249**, S95–S100 (2012).
30. Stone, S. L., Williams, L. A., Farmer, L. M., Vierstra, R. D. & Callis, J. *KEEP ON GOING*, a RING E3 ligase essential for *Arabidopsis* growth and development, is involved in abscisic acid signaling. *Plant Cell* **18**, 3415–3428 (2007).
31. Vos, P. G. et al. GWAS in tetraploid potato: identification and validation of SNP markers associated with glycoalkaloid content. *Mol. Breed.* **42**, 76 (2022).
32. Kloosterman, B. et al. Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature* **495**, 246–250 (2013).
33. Liu, X. et al. Systematic analysis of potato acid invertase genes reveals that a cold-responsive member, *StvacINV1*, regulates cold-induced sweetening of tubers. *Mol. Genet. Genomics* **286**, 109–118 (2011).
34. Ren, R., Li, D., Zhen, C., Chen, D. & Chen, X. Specific roles of *Os4BGlu10*, *Os6BGlu24*, and *Os9BGlu33* in seed germination, root elongation, and drought tolerance in rice. *Planta* **249**, 1851–1861 (2019).
35. Offen, W. et al. Structure of a flavonoid glucosyltransferase reveals the basis for plant natural product modification. *EMBO J.* **25**, 1396–1405 (2006).
36. Strassner, J. et al. Characterization and cDNA-microarray expression analysis of 12-oxophytodienoate reductases reveals differential roles for octadecanoid biosynthesis in the local versus the systemic wound response. *Plant J.* **32**, 585–601 (2002).
37. Campbell, R. et al. The metabolic and developmental roles of carotenoid cleavage dioxygenase4 from potato. *Plant Physiol.* **154**, 656–664 (2010).
38. Kloosterman, B. et al. From QTL to candidate gene: genetical genomics of simple and complex traits in potato using a pooling strategy. *BMC Genomics* **11**, 158 (2010).
39. Yin, H. et al. SAUR15 promotes lateral and adventitious root development via activating H^+ -ATPases and auxin biosynthesis. *Plant Physiol.* **184**, 837–851 (2020).
40. Liu, W. et al. The flavonoid biosynthesis network in plants. *Int. J. Mol. Sci.* **22**, 12824 (2021).
41. Chen, M. et al. Genome-wide identification of agronomically important genes in outcrossing crops using OutcrossSeq. *Mol. Plant* **14**, 556–570 (2021).
42. Serra, L. & Perrot-Rechenmann, C. Spatiotemporal control of cell growth by CUC3 shapes leaf margins. *Development* **147**, dev183277 (2020).
43. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
44. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
45. Wang, Y. et al. Comparative genomics revealed the gene evolution and functional divergence of magnesium transporter families in *Saccharum*. *BMC Genomics* **20**, 83 (2019).
46. Zheng, C., Huang, Y.-X. & Liu, K.-X. Influence of magnesium on photosynthetic characteristics and growth effects of sugarcane at seedling stage. *Am. J. Biochem. Biotechnol.* **15**, 33–38 (2019).
47. Qi, W. et al. Rice ethylene-response AP2/ERF factor *OsEATB* restricts internode elongation by down-regulating a gibberellin biosynthetic gene. *Plant Physiol.* **157**, 216–228 (2011).
48. Kwon, S. H. et al. Overexpression of a *Brassica rapa* *NGATHA* gene in *Arabidopsis thaliana* negatively affects cell proliferation during lateral organ and root growth. *Plant Cell Physiol.* **50**, 2162–2173 (2009).
49. Yarra, R. & Wei, W. The NAC-type transcription factor *GmNAC20* improves cold, salinity tolerance, and lateral root formation in transgenic rice plants. *Funct. Integr. Genomics* **21**, 473–487 (2021).
50. Hu, Y. et al. An auxin response factor regulates tiller angle and shoot gravitropism by directly activating related gene expression in rice. *J. Adv. Res.* **78**, 147–162 (2025).
51. Chen, L., Jameson, G. B., Guo, Y., Song, J. & Jameson, P. E. The *LONELY GUY* gene family: from mosses to wheat, the key to the formation of active cytokinins in plants. *Plant Biotechnol. J.* **20**, 625–645 (2022).
52. Di Laurenzio, L. et al. The *SCARECROW* gene regulates an asymmetric cell division that is essential for generating the radial organization of the *Arabidopsis* root. *Cell* **86**, 423–433 (1996).
53. Wiersma, A. T. et al. K-mer genome-wide association study for anthracnose and BCMV resistance in a *Phaseolus vulgaris* Andean diversity panel. *Plant Genome* **17**, e20523 (2024).
54. Jaegle, B. et al. K-mer-based GWAS in a wheat collection reveals novel and diverse sources of powdery mildew resistance. *Genome Biol.* **26**, 172 (2025).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

Methods

Association mapping with KMERIA on polyploids

We present a method for identifying genetic loci associated with quantitative trait in polyploid species using whole-genome resequencing data, bypassing the need for read alignment to a reference genome. The workflow comprises the steps discussed further below.

Counting. Whole-genome resequencing reads from population samples were processed using our developed tool KMERIA ‘kmeria count’ to calculate k -mer abundances. Alternatively, the hash-based tool KMC3 is also supported. We used a k -mer size of 31, as this represents the maximum length efficiently storable using 64-bit integers. To enhance computational efficiency, k -mers occurring only once across all samples were discarded before association testing, as these likely represent sequencing errors.

Matrix construction. The matrix construction of the k -mer frequencies from individual sample outputs is handled by a highly memory-efficient parallel merge algorithm. KMERIA leverages ‘a k -way merge strategy’ across n sorted k -mer databases, where n is the number of individuals. KMERIA also supports k -mer hashing database formats and processes sorted k -mer streams in a single pass. For each unique k -mer k_j , the matrix entry $M_{j,i}$ is defined as:

$$M_{j,i} = \begin{cases} f_i(k_j) & \text{if } k_j \in D_i \\ 0 & \text{otherwise} \end{cases}$$

while $f_i(k_j)$ represents the frequency count of k -mer k_j in database D_i . At each iteration, KMERIA identifies the minimum k -mer across all active database readers and constructs the corresponding matrix row by collecting frequencies from all databases matching that k -mer.

To efficiently manage the large memory footprint of the initially constructed k -mer matrix, which is often highly sparse, KMERIA also implemented an adaptive compression scheme. This scheme is designed to maximize memory savings by automatically selecting the optimal encoding strategy based on the frequency distribution characteristics of the matrix.

Matrix pruning. Following matrix construction, k -mer frequency data underwent quality control and normalization to account for sequencing depth heterogeneity. The filtering pipeline implements a two-stage approach—ploidy-aware depth correction followed by population-level quality filtering.

Depth correction for ploidy. To account for sequencing depth variation and differences in genome ploidy across samples, raw k -mer counts (KMCs) were normalized to a standardized haploid depth scale.

The average haploid depth d_i^{haploid} was estimated by identifying the main peak in the KMC frequency distribution. This ploidy-aware correction adjusts k -mer copy number statistics, ensuring accurate representation in the genotype matrix. The calculation formula is:

$$d_i^{\text{haploid}} = C/(g \times k)$$

where C represents the KMC, g denotes the haploid genome size and k is the k -mer size.

For each sample i with sequencing depth D_i and ploidy level p_i , the haploid depth was calculated as:

$$d_i^{\text{haploid}} = \frac{D_i}{p_i}$$

The corrected k -mer frequency f'_{ij} for k -mer j in sample i was then computed as:

$$f'_{ij} = \text{round} \left(\frac{f_{ij}}{d_i^{\text{haploid}}} \right)$$

where f_{ij} represents the raw observed count. This normalization enables direct comparison of k -mer frequencies across samples with different sequencing depths.

K -mer quality filtering criteria. The missing data threshold reflects the balance between marker density and data completeness. Extremely rare k -mers (present in <20% samples) may represent the de novo mutations or recent variants with low population frequency. While such k -mers may be biologically informative, their statistical power for GWAS is limited due to small effect-size estimation and multiple testing burden. To mitigate artifacts from repetitive genomic regions, sequencing errors (GC-content bias or PCR amplification), k -mers with extreme coverage were filtered. For each k -mer j , the missing rate m_j was calculated as:

$$m_j = \frac{|(i : f'_{ij} = 0)|}{n}$$

where n is the total number of samples. K -mers with $m_j > \theta_{\text{miss}}$ were excluded.

To eliminate the interference of abnormal k -mer abundance and achieve more accurate allele dosage estimation, while also optimizing the speed of subsequent association analyses, we re-encoded k -mer abundance to better suit an efficient GWAS model. Specifically, we introduced a quantile-based approach to handle extreme k -mer abundance (outliers)— k -mer abundance values below the fifth quantile were converted to 0. K -mer abundance exceeding the 95th quantile was converted to 2. The remaining k -mer abundance values were linearly transformed into a range between 0 and 2. The recoding process can be expressed as:

$$y = \begin{cases} 0, & x \leq x_{0.05} \\ \frac{2 \times (x - x_{0.05})}{x_{0.95} - x_{0.05}}, & x_{0.05} < x < x_{0.95} \\ 2, & x \geq x_{0.95} \end{cases}$$

where x is the corrected k -mer abundance, $x_{0.05}$ and $x_{0.95}$ denote the 5th and 95th quantiles of the k -mer abundance distribution, respectively, and y represents recoded allele dosage value.

K -mer testing. To effectively control for false true positives in GWAS analysis, an LMM-based model was also introduced into KMERIA method. First, we take a random sampled 0.1% total k -mers from the recoded k -mer genotype matrices used for the population structure and kinship relatedness. By default, three PCs were used for population stratification correction. Kinship was calculated using *bimbamKin* (<https://github.com/Sh1ne111/KMERIA/tree/main/bimbamAsso>) or *GEMMA* (<https://github.com/genetics-statistics/GEMMA>) with default parameters, and association tests were conducted through *bimbamAsso* (<https://github.com/Sh1ne111/KMERIA/tree/main/bimbamAsso>). The KMERIA association study model using a mixed linear model represented by the equation:

$$y = \beta_0 + \beta_1 \times k\text{-mer} + \sum_{k=1}^m y_k \times \text{PC}_k + u + \varepsilon$$

where y represents the vector of observed phenotypic values (quantitative trait measurements) for all individuals in the study population; β_0 is fixed effect, representing the overall mean of the phenotype when

all other variables are zero; β_1 indicates the effect size (fixed-effect coefficient) of the k -mer on the phenotype—the primary target of inference in GWAS; k -mer is the genotype vector for a k -mer across individuals, coded as scaled allele dosage; γ_k represents the fixed-effect coefficient for the k -th PC of population structure, correcting for population stratification; PC_k is the PC vector capturing axes of genetic variation across individuals, calculated from sampling k -mers; $\mathbf{u}$ is polygenic random effects vector modeling background genetic relatedness between individuals, distributed as $n(0, K\sigma_g^2)$ where K is the kinship matrix; and ϵ is residual error vector representing environmental noise and model misspecification, distributed as $n(0, I\sigma_e^2)$.

Post-GWAS. To control false discoveries while maximizing detection sensitivity, we implemented a dual-correction strategy for association testing. First, BH-FDR correction ($P_{\text{adj}} < 0.05$) was applied to account for multiple testing across all k -mer associations. We also proposed a modified Bonferroni threshold based on the effective number of independent tests. We derived a biologically informed significance threshold accounting for k -mer interdependency. The sequencing reads linked to significant k -mers were retrieved and mapped to graph pangenome reference using Minigraph (v0.21)⁵⁵ with the following parameters: ‘minigraph -x sr’.

Phenotype simulation. We take ploidy into account to generate the phenotypes in polyploidy organisms, the phenotypes of the polyploids for i_{th} individual $i(y_i)$ (equivalent equation) is simulated from:

$$y_i = \sum_{j=1}^Q k_{ij} \alpha_j + e_i$$

$$\alpha_i \sim n(0, \sigma^2)$$

$$\sigma_e^2 = \frac{\sigma_a^2}{h^2} - \sigma_g^2$$

$$e \sim n(0, I\sigma_e^2)$$

where k_{ij} denotes the number of copies of the alternative allele, α_j is the additive effect of each locus and e_i is the normal residual.

Power evaluation across different type I error levels

The statistical power, type I error (α) and FDR were evaluated in association tests conducted on simulated datasets, which included predefined QTLs. The evaluation used the method⁵⁶, alongside two approaches from our previous research^{57,58}. A QTL was deemed successfully detected if a positive marker fell within a specified genomic distance threshold; this study uses 50 kb as the detected distance. Statistical power was quantified as the proportion of QTLs detected at a given threshold of type I error. In cases where no QTL was present within the interval, markers were used to generate the null distribution for negative control. The null distribution for type I error was based on markers not associated with any QTL, and FDR was calculated as the ratio of non-QTL markers to the total positive markers.

Statistical significance threshold for KMERIA

To resolve the extreme multiple testing burden in KMERIA (typically 10^8 – 10^{10} tests), we implemented a dual-correction strategy. Traditional Bonferroni correction proved overly conservative due to false negatives from nonindependent tests, while standard FDR methods remained insufficient for this scale. This limitation arises because adjacent k -mers exhibit high spatial correlation (nucleotide overlap),

violating statistical independence assumptions. We therefore applied BH-FDR as the primary P -value correction method to control false discoveries while retaining sensitivity. We also proposed a modified Bonferroni threshold based on the effective number of independent tests. We derived a biologically informed significance threshold accounting for k -mer interdependency. Adjacent k -mers exhibit high correlation due to their inherent $k-1$ nucleotide overlap, violating independence assumptions underlying traditional corrections. We therefore calculated the effective number of independent tests. The formula is as follow:

$$M_{\text{eff}} = \frac{n_k}{k}$$

where M_{eff} is effective number of independent tests, n_k represents the total unique k -mers after quality control and k is k -mer size (for example, 31 bp). The genome-wide significance threshold was then defined as:

$$\text{Threshold} = \frac{\alpha}{M_{\text{eff}}} = \alpha \times \frac{k}{n_k}$$

where α donates the nominal significance level (0.05). This biologically informed approach accounts for k -mer redundancy, where overlapping k -mers share bases of sequence information. The M_{eff} adjusted threshold served as a genome-wide significance benchmark complementary to BH-FDR.

Genome sequencing, assembly and annotation

Plant material. This study analyzed 290 *S. spontaneum* accessions provided by the Yunnan Academy of Agricultural Sciences. The plant materials were grown at a subtropical plateau climate at Kaiyuan, Yunnan (23°30′–23°58′N, 103°04′–103°43′E, altitude–1051 m). Before planting, the buds were disinfected, and standardized field management practices, including irrigation, pest control and weed management, ensured uniform germination.

Sample selection. We carefully selected 14 representative *S. spontaneum* accessions based on the phylogenetic relationships among 290 accessions, aiming to capture a diverse spectrum of *S. spontaneum* germplasm resources. To establish the phylogeny of the selected accessions, the phylogenetic tree among *S. spontaneum* was estimated using Mashree⁵⁹.

HiFi sequencing. High-quality genomic DNA of 14 *S. spontaneum* was extracted from fresh leaves using the DNasequre Plant Kit (TIANGEN). A total of 28 Single-Molecule Real-Time bell libraries with ~20 kb DNA insert fragments were sequenced with the circular consensus sequencing mode on a PacBio REVIO platform at ANNOROAD. These yielded HiFi reads of ~180 Gb with 225× coverage on average for *S. spontaneum* haploid genome.

Estimation of genome ploidy. To estimate the genome ploidy of the *S. spontaneum* population data, we used a de Bruijn graph-based method ploidyfrost⁶⁰, which infers ploidy levels directly from whole-genome sequencing datasets without the need for a reference genome. This method involves the following steps: (1) k -mer database construction—for each sample, a k -mer database is built using KMC, which computes allelic k -mer coverage. To ensure a concise and error-free graph, low-coverage k -mers in the reads are masked. (2) Compacted de Bruijn graph construction—a compacted de Bruijn graph is generated from single or multiple samples using the Bifrost API, ensuring efficient and accurate graph representation. (3) Graph analysis and variant extraction—three subalgorithms are applied to analyze the graph structure, with putative variant information being extracted and output for further analysis.

Evaluation of genome size. The Illumina sequencing reads were applied to evaluate genome size based on the *k*-mer distribution using KMC (v3.2.158) and genomescope (v2.0.0; <https://github.com/schatzlab/genomescope/>) with the parameters: -k 21.

Genome assembly. Genome assembly of 14 HiFi-sequenced accessions uses hifiasm (v0.19.8-r603; <https://github.com/chhylp123/hifiasm>)⁶¹ with default parameters. The initial output of hifiasm resulted in the primary assembly (p_utg), representing a mosaic polyploid assembly.

Gene prediction. Repeats were first identified in the *S. spontaneum* genome assembly with RepeatModeler⁶² (v2.0.3) to create a custom repeat library. The repeat library was then combined with Viridiplantae repeats from RepBase⁶³ (v20150807) to generate the final repeat library. Repeat masking of each of *S. spontaneum* genome assembly was performed using the final CRL and RepeatMasker⁶⁴ (v4.1.2) with the parameters '-e ncbi -s -nolow -no_is -gff'. *S. spontaneum* RNA-seq raw reads were processed using Trimmomatic (v1.0)⁶⁵ to cut adaptor and remove low-quality reads and then aligned to the *S. spontaneum* genome assembly using HISAT2 (v2.2.1)⁶⁶. The aligned RNA-seq reads were each assembled using StringTie (v2.2.1)⁶⁷. Initial gene models were generated using BRAKER (v3.1)⁶⁸ using the soft-masked genome assembly and RNA-seq read alignments as hints. High-confidence gene models were identified from the working gene model set by filtering out those without expression evidence, a Pfam domain match or those that were partial gene models or contained an interior stop codon.

Genome quality evaluation. The completeness of conserved genome regions in the assemblies was evaluated by BUSCOs (v5.4.1)⁴³ using 1,614 single-copy orthologs from the 'embryophyta_odb10' database. Assembly contiguity was assessed through the LAI⁴⁴, calculated using LTR retriever (v2.9.6)⁶⁹ with default parameters. Briefly, the completeness of LTR retrotransposons was analyzed using LTR_FINDER_parallel (v1.2)⁷⁰ and LTR_HARVEST_parallel (v1.0)⁷¹. These tools were used to identify LTR elements for each genome, which were then compiled into a nonredundant LTR library using LTR retriever (v2.9.6) for calculating the LAI. For accuracy assessment, the base-level QV was computed using Merquy (v1.4.1)⁷², based on *k*-mer analysis.

SNP-based GWAS in potato and sweet potato populations using GWASpoly

GWASs were performed using SNP dosage data, using an LMM to account for population structure and relative kinship. The analysis was carried out using the R package GWASpoly (v2.10)¹⁰. Initially, the covariance matrix for the polygenic effect was calculated using the set.K function. After this, the GWASpoly function was used to conduct the GWAS under an 'additive' model, where the effect of each SNP was assumed to be proportional to the dosage of the minor allele. Genome-wide significance threshold was calculated based on the Bonferroni correction of $\frac{\alpha}{M}$, where α is 0.05 and M is the number of valid SNPs.

Construction of the *S. spontaneum* graph-based pangenome

To capture the genetic diversity within *S. spontaneum*, a graph pangenome was constructed using a collection of 15 monoploid genome assemblies. The graph pangenome was built using Minigraph-Cactus (v2.6.11)⁷³ with default parameters. To ensure the accuracy and reliability of the graph structure, a combination of read mapping and assembled alignments was used to validate the paths, edges and nodes within the *S. spontaneum* pangenome.

Functional characterization of *SsNGA5*

Vector construction. To generate the overexpression construct, the full-length CDS of *SsNGA5* was amplified through PCR (Supplementary Table 15) using cDNA derived from leaf of

S. spontaneum as a template. The plant expression vector pBWA(V) HS-ccdB-GFP was linearized by digestion with BsaI and Eco31I restriction endonucleases. The amplified *SsNGA5* fragment was then inserted downstream of the CaMV 35S promoter and upstream of the GFP reporter gene using the Biorun 2× EasyClone seamless cloning kit, resulting in the 35S::SsNGA5-GFP recombinant plasmid.

Rice genetic transformation. The recombinant plasmid was transformed into *Agrobacterium tumefaciens* strain EHA105. *Agrobacterium*-mediated transformation of rice (ZH11) was performed using a callus-based system. Briefly, sterilized mature rice seeds were placed on induction medium to generate embryonic calli. After 10–15 min of immersion in the *Agrobacterium* suspension, the calli were transferred to coculture medium for 2 days in the dark. Transgenic calli were subsequently identified through multiple rounds of selection on medium. Actively growing resistant calli were transferred to regeneration and rooting media to obtain *T*₀ plantlets. Positive transformants were further validated by PCR-based screening using genomic DNA extracted from leaves, specifically targeting the *SsNGA5* sequence.

Statistical analysis

Complete details of all statistical analyses for Figs. 2–5, Extended Data Figs. 3–10 and Supplementary Figs. 2–6 are provided in their respective legends and the Methods section. Statistical assessment of replicate data was performed using R software (v4.3.1).

Reporting summary

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

Data availability

The 14 *S. spontaneum* genome assemblies and annotations generated in this study have been deposited in the figshare database under DOIs (<https://doi.org/10.6084/m9.figshare.32251332>)⁷⁴ and (<https://doi.org/10.6084/m9.figshare.32142925>)⁷⁵. The genome assemblies, annotations and raw PacBio HiFi sequencing reads are also accessible through the National Center for Biotechnology Information (NCBI) under BioProject accession PRJNA1393212. The reference genome assembly of *S. spontaneum* 82–114 was obtained from our published study¹⁶. Previously published resequencing datasets analyzed in this study were obtained from NCBI Sequence Read Archive using the following BioProject accessions: potato (PRJNA944441, PRJNA63145, PRJNA766763, PRJNA378971), sweet potato (PRJNA1089346), alfalfa (PRJNA995892), rice (PRJNA623686) and *S. spontaneum*/hybrid (PRJNA1303125).

Code availability

The KMERIA program has been released on Zenodo (<https://doi.org/10.5281/zenodo.20117337>)⁷⁶ and GitHub (<https://github.com/ShIne111/KMERIA>, <https://github.com/ShIne111/KMERIA/wiki>).

References

- Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
- Segura, V. et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44**, 825–830 (2012).
- Wang, Q., Tian, F., Pan, Y., Buckler, E. S. & Zhang, Z. A SUPER powerful method for genome wide association study. *PLoS ONE* **9**, e107684 (2014).
- Li, M. et al. Enrichment of statistical power for genome-wide association studies. *BMC Biol.* **12**, 73 (2014).
- Katz, L. et al. Mashtree: a rapid comparison of whole genome sequence files. *J. Open Source Softw.* **4**, 1762 (2019).

60. Sun, M., Pang, E., Bai, W., Zhang, D. & Lin, K. ploidyfrost: reference-free estimation of ploidy level from whole genome sequencing data based on de Bruijn graphs. *Mol. Ecol. Resour.* **23**, 499–510 (2023).
61. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
62. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
63. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
64. Tarailo-Graovac, M. & Chen, N. RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4.10.1–4.10.14 (2009).
65. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
66. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
67. Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R. & Pertea, M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
68. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **3**, lqaa108 (2021).
69. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
70. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 48 (2019).
71. David, E., Stefan, K. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
72. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
73. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
74. Chen, S. 14 genome assemblies of *Saccharum spontaneum*. *figshare* <https://doi.org/10.6084/M9.FIGSHARE.32251332> (2026).
75. Chen, S. 14 genome annotations of *Saccharum spontaneum*. *figshare* <https://doi.org/10.6084/M9.FIGSHARE.32142925> (2026).
76. Chen, S. A k-mer-based GWAS approach empowering gene mining in polyploids. *Zenodo* <https://doi.org/10.5281/ZENODO.20117337> (2026).

Acknowledgements

We thank X. Zhang (Guangdong Academy of Agricultural Sciences) for providing the SNP-based genotype files and the corresponding reference genome for the sweet potato.

Author contributions

X.Z. and S.C. conceived and designed this study. S.C. developed the KMERIA method. S.C. and S.Q. implemented the simulated analyses. S.C. performed the real data analyses. S.C. and Y.S. assembled the *S. spontaneum* genomes and performed the pangenome analyses. K.C. performed the genetic variant analyses. H. Liu, Y.Z., Z.X., X. Li, X. Liu, J.W. and M.Z. provided the sequencing materials of *S. spontaneum* accessions and collected the phenotypic data. H. Li provided SNP-based genotypes for potato and performed SNP-GWAS analyses using these datasets. G.-B.C. and C.M. performed the population genetic analysis and revised the paper. S.C. and X.Z. wrote the paper with input from all the co-authors.

Funding

This work was supported by National Natural Science Foundation of China (32370714 to X.Z.), Science and Technology Major Project of Guangxi (Gui Ke AA24206004), Basic Research Center for Agricultural Frontiers and Interdisciplinary Sciences (BRC-AFIS), Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-BRC-AFIS-2025-02 to X.Z.), Hainan Provincial Sanya Yazhou Bay Science and Technology Innovation Joint Project (ZDYF2025GXJS139) and the Project of State Key Laboratory for Tropical Crop Breeding (NKLTCBCXTD23).

Competing interests

The authors declare no competing interests.

Additional information

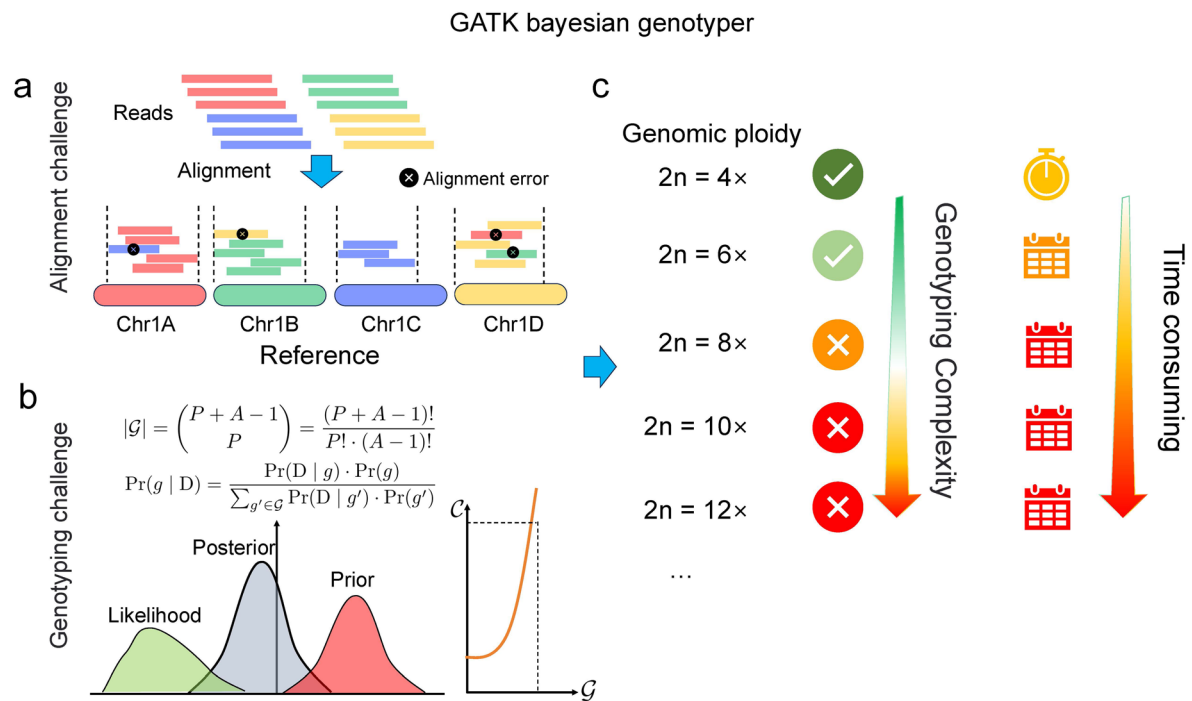
Extended data is available for this paper at <https://doi.org/10.1038/s41588-026-02641-8>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-026-02641-8>.

Correspondence and requests for materials should be addressed to Xingtang Zhang.

Peer review information *Nature Genetics* thanks Benjamin Jaegle and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

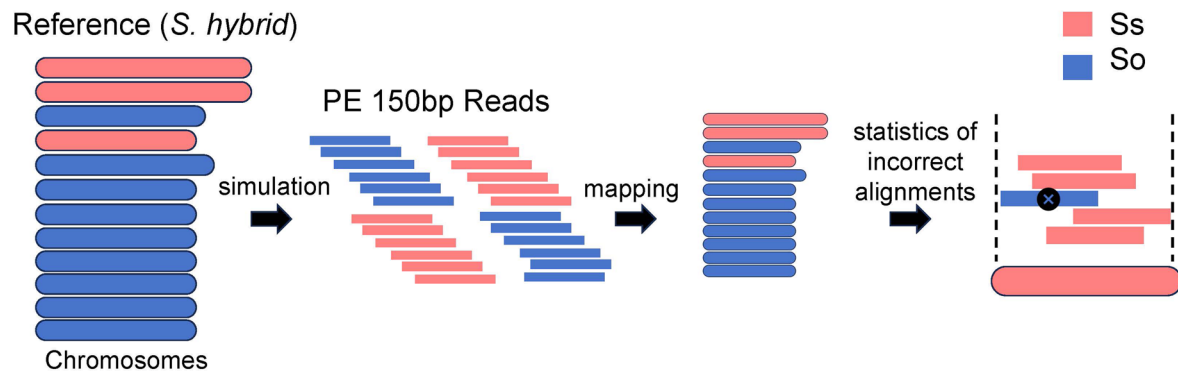
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Comparison of traditional GATK genotyping and k-mer based genotyping strategies in polyploid GWAS. a, A schematic illustrating the challenges associated with genotyping in polyploid GWAS, particularly when aligning resequencing data to a reference genome. **b,** The core principle of the GATK variant caller is the inference of individual genotypes through

Bayesian algorithms. When applied to polyploid genotyping, the computational complexity $\mathcal{C}$ increases exponentially with both genome ploidy and allele count g . **c,** A diagram depicting the complexity and computational time required for genotyping genomes of species with varying ploidy levels using GATK.

Benchmarking Misalignment in Polyploid Genomes Using Short Reads

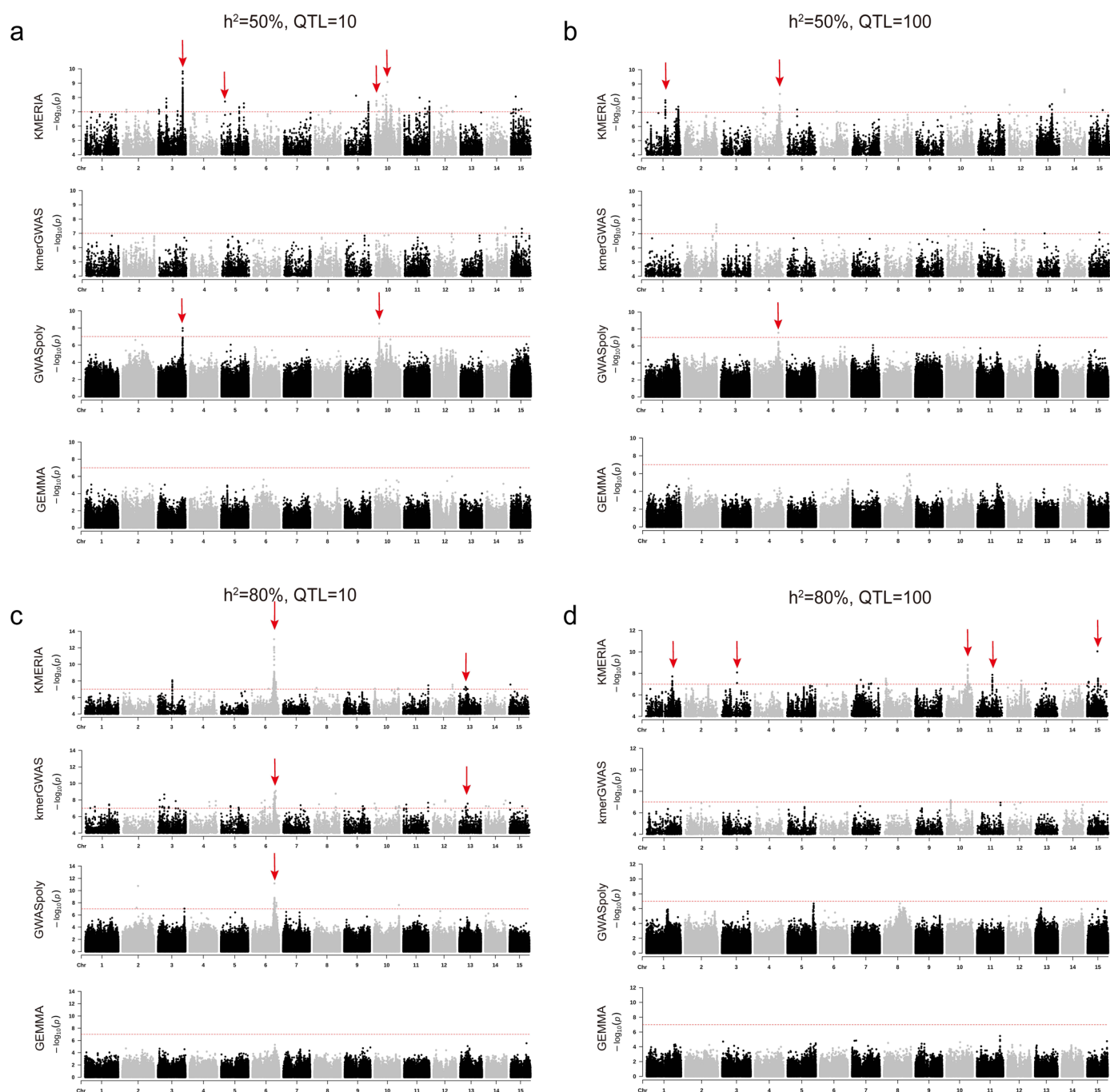


Extended Data Fig. 2 | Benchmarking strategy for assessing misalignment rates using simulated short reads. To evaluate alignment fidelity in polyploids, the *Saccharum hybrid* (*S. hybrid*) was used as a reference template. Paired-end

150 bp reads were simulated from the Ss (red) and So (blue) subgenomes and subsequently mapped back to the reference. The accuracy of the alignment was determined by calculating the statistics of incorrectly aligned reads.

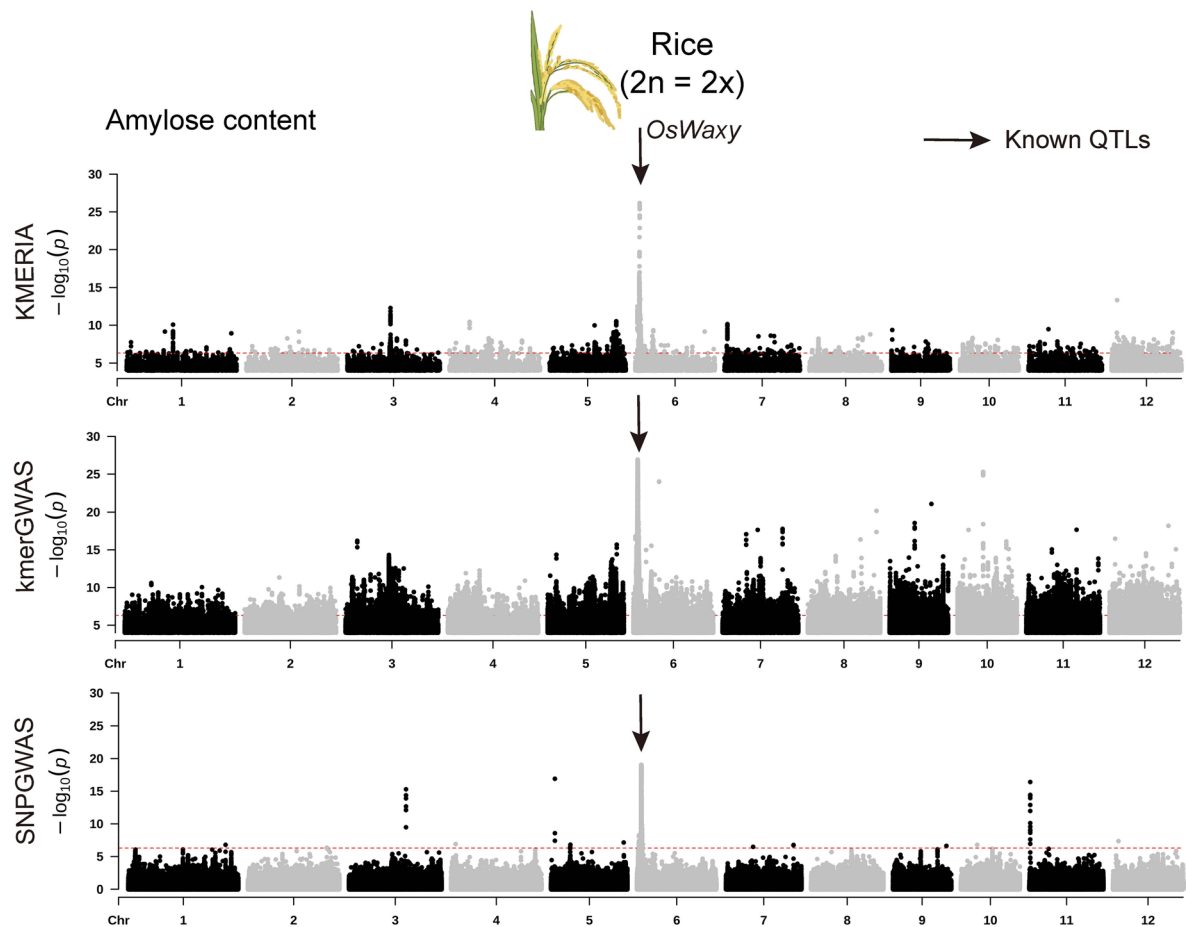
Hexaploid ($2n=6\times$)

→ Simulated QTLs



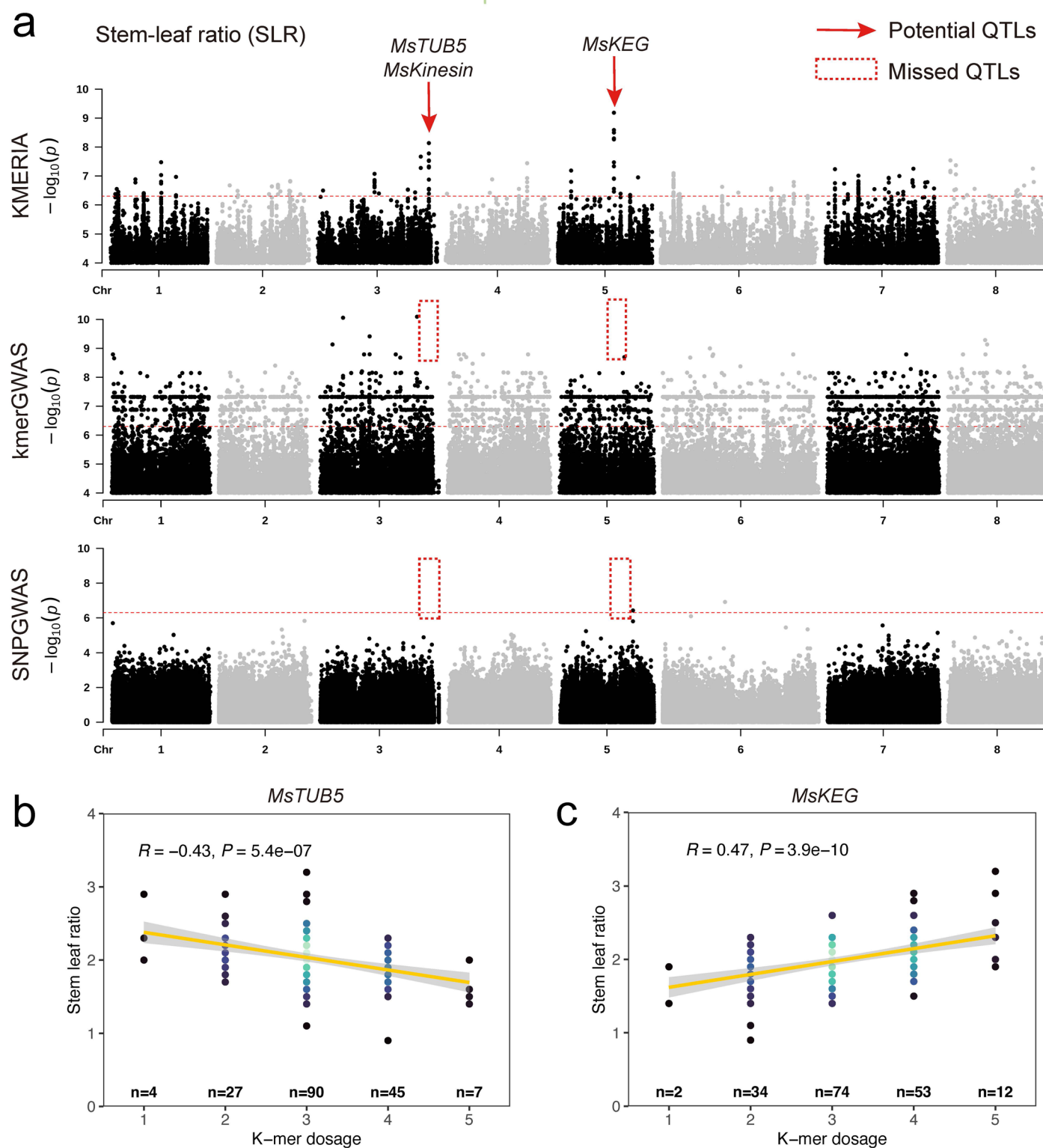
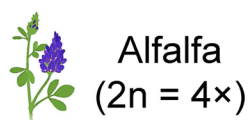
Extended Data Fig. 3 | Comparative evaluation of statistical power for different GWAS methods in hexaploid sweet potato using simulated phenotypes. a–d, Manhattan plots display the association results from KMERIA, kmerGWAS, GWASpoly, and GEMMA applied to a natural population. Phenotypes were simulated under four distinct genetic architectures varying in

heritability (h^2) and the number of QTLs: (a) $h^2 = 50\%$, 10 QTLs; (b) $h^2 = 50\%$, 100 QTLs; (c) $h^2 = 80\%$, 10 QTLs; and (d) $h^2 = 80\%$, 100 QTLs. Red arrows indicate the genomic coordinates of the presimulated (true) QTLs, serving as ground truth to assess detection sensitivity. The horizontal dashed lines represent the genome-wide significance thresholds.



Extended Data Fig. 4 | Manhattan plots of association results for rice amylose content. Association signals were analyzed using KMERIA, kmersGWAS, and the standard SNP-GWAS method. The analysis successfully identified the major gene *OsWaxy* (indicated by the black arrow) on chr6, a known QTL regulating amylose

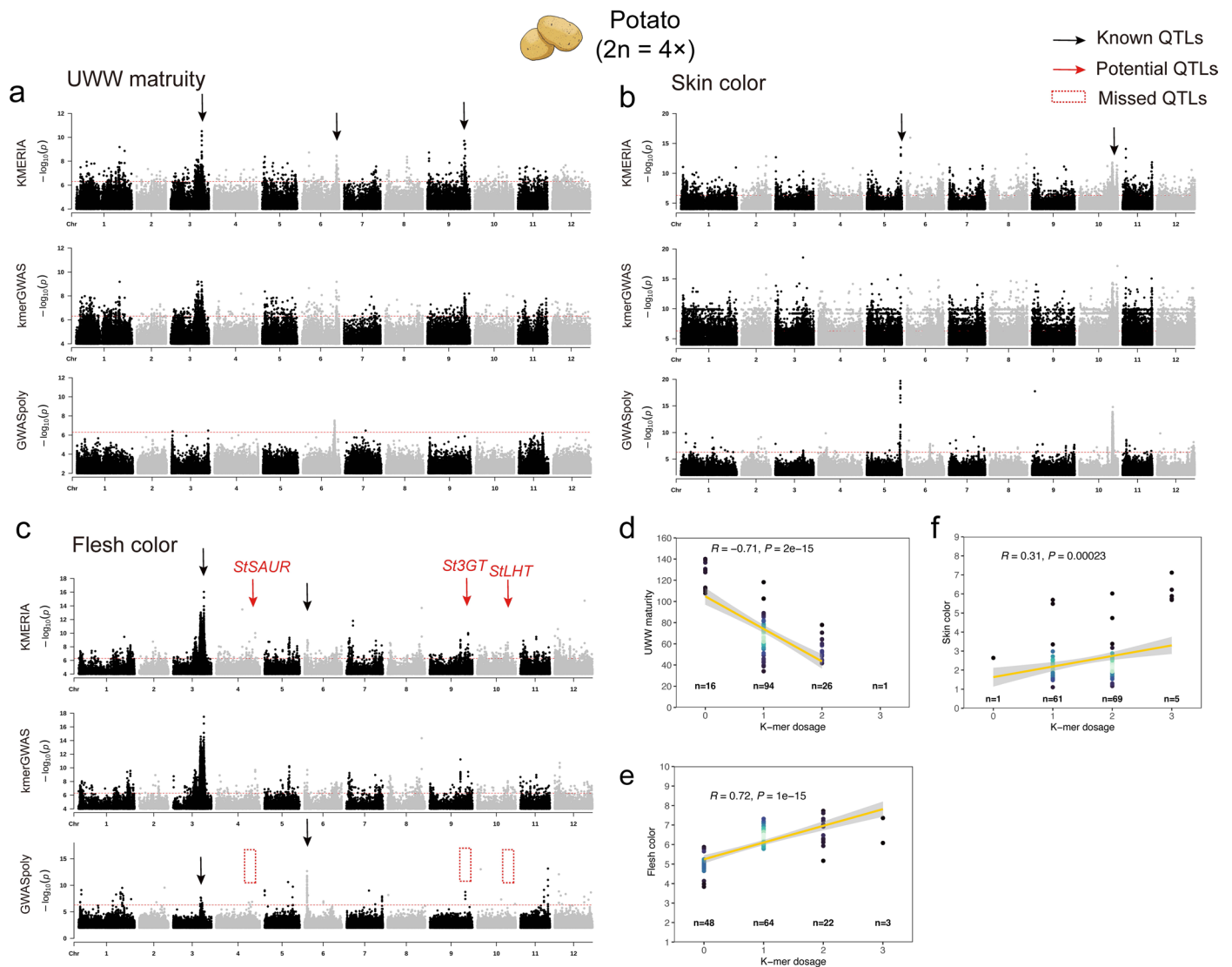
content. *P* values were calculated using a two-sided Wald test. The horizontal dashed lines indicate the GWAS significance threshold, adjusted for multiple comparisons using BH-FDR method.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Comparative GWAS analysis and validation of significant *k*-mer associations for stem-leaf ratio (SLR) in tetraploid alfalfa. **a**, Manhattan plots display the GWAS results for the SLR phenotype in the alfalfa population, using KMERIA, kmerGWAS, and SNP-GWAS method. Red solid arrows indicate potential QTLs identified by KMERIA, corresponding to the known genes *MsTUB5* and *MsKEG*. Red dashed boxes highlight the genomic regions where the known QTLs were missed by kmersGWAS and SNP-GWAS. *P* values were calculated using a two-sided Wald test. The horizontal dashed lines in all Manhattan plots indicate the GWAS significance threshold, adjusted

for multiple comparisons using BH-FDR method. **b,c**, Scatterplots illustrate the correlation between the estimated *k*-mer dosage and the SLR phenotype for the two significant candidate genes identified: **(b)** *MsTUB5* and **(c)** *MsKEG*. The top left text is Pearson correlation coefficient (*R*), with the corresponding *P*-value (two-sided *t*-test) indicating the significance of the association. The number of accessions (*n*) for each *k*-mer dosage level is provided at the bottom of the plot. The solid line indicates the linear regression fit, and the shaded gray area represents the 95% confidence interval. Individual points are colored by density.



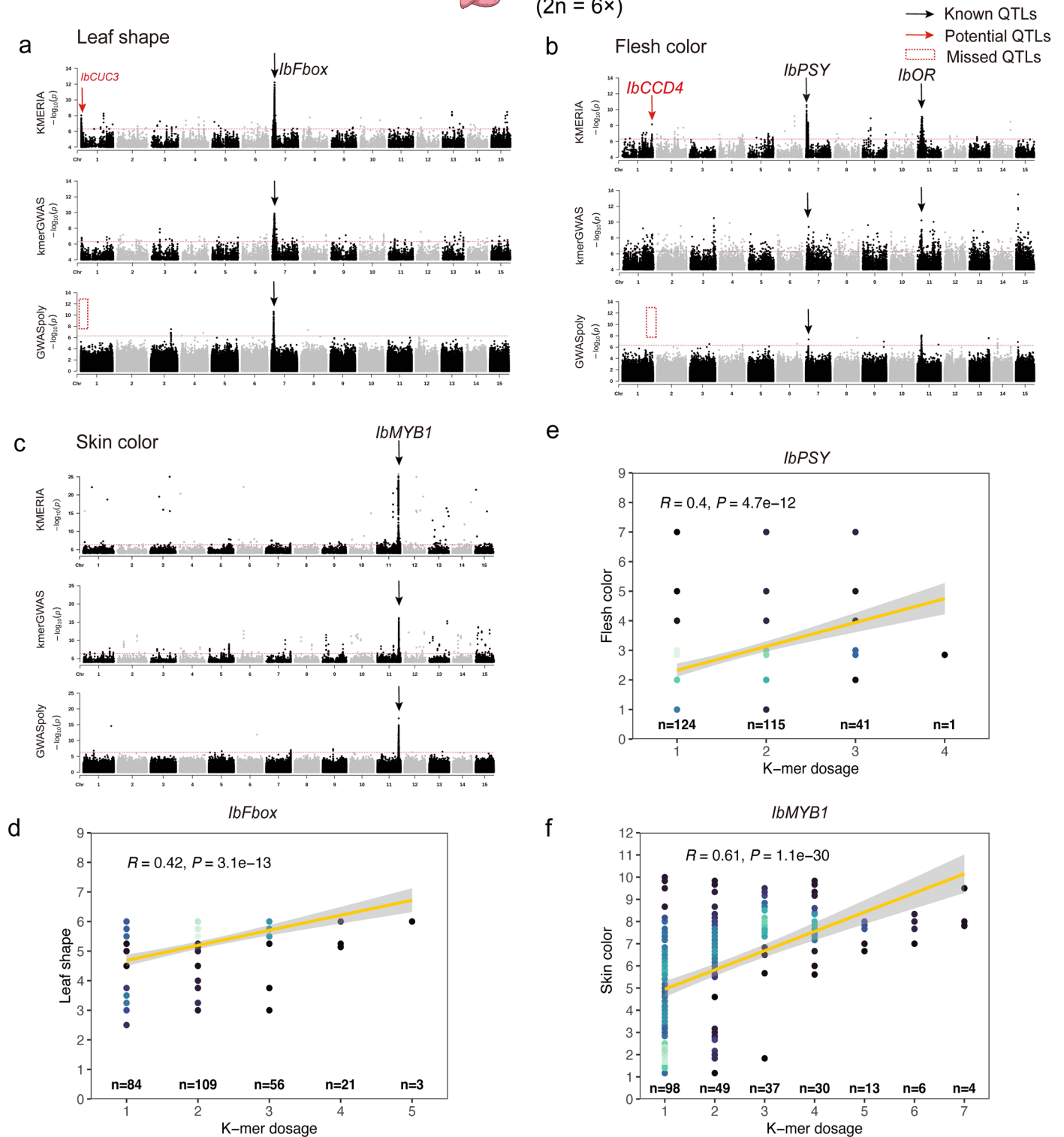
Extended Data Fig. 6 | Comparative GWAS analysis and validation of significant *k*-mer associations for three traits in tetraploid potato.

a–c, Manhattan plots display the GWAS results for three traits in the potato population: (a) UWW maturity, (b) skin color, and (c) flesh color. The results are compared across three GWAS methods: KMERIA, kmersGWAS, and GWASpoly. Black arrows indicate the genomic coordinates of known QTLs previously associated with the respective traits, serving as positive controls. Red solid arrows indicate potential QTLs identified by KMERIA, corresponding to candidate genes such as *StSAUR*, *St3GT*, and *StLHT*. Red dashed boxes highlight the genomic regions where the potential QTLs were missed by GWASpoly. *P* values were calculated using a two-sided Wald test. The horizontal dashed

lines in all Manhattan plots indicate the GWAS significance threshold, adjusted for multiple comparisons using BH-FDR method. **d–f**, Scatterplots illustrate the correlation between the estimated *k*-mer dosage and the respective phenotype for significant candidate genes identified by KMERIA: (d) a major QTL for UWW maturity, (e) a major QTL for skin color, and (f) a major QTL for flesh color. The top left text is Pearson correlation coefficient (*R*), with the corresponding *P*-value (two-sided *t*-test) indicating the significance of the association. The number of accessions (*n*) for each *k*-mer dosage level is provided at the bottom of the plot. The solid line indicates the linear trend, surrounded by gray shading representing the 95% confidence interval. The color intensity of individual markers reflects point density.



Sweetpotato
($2n = 6x$)

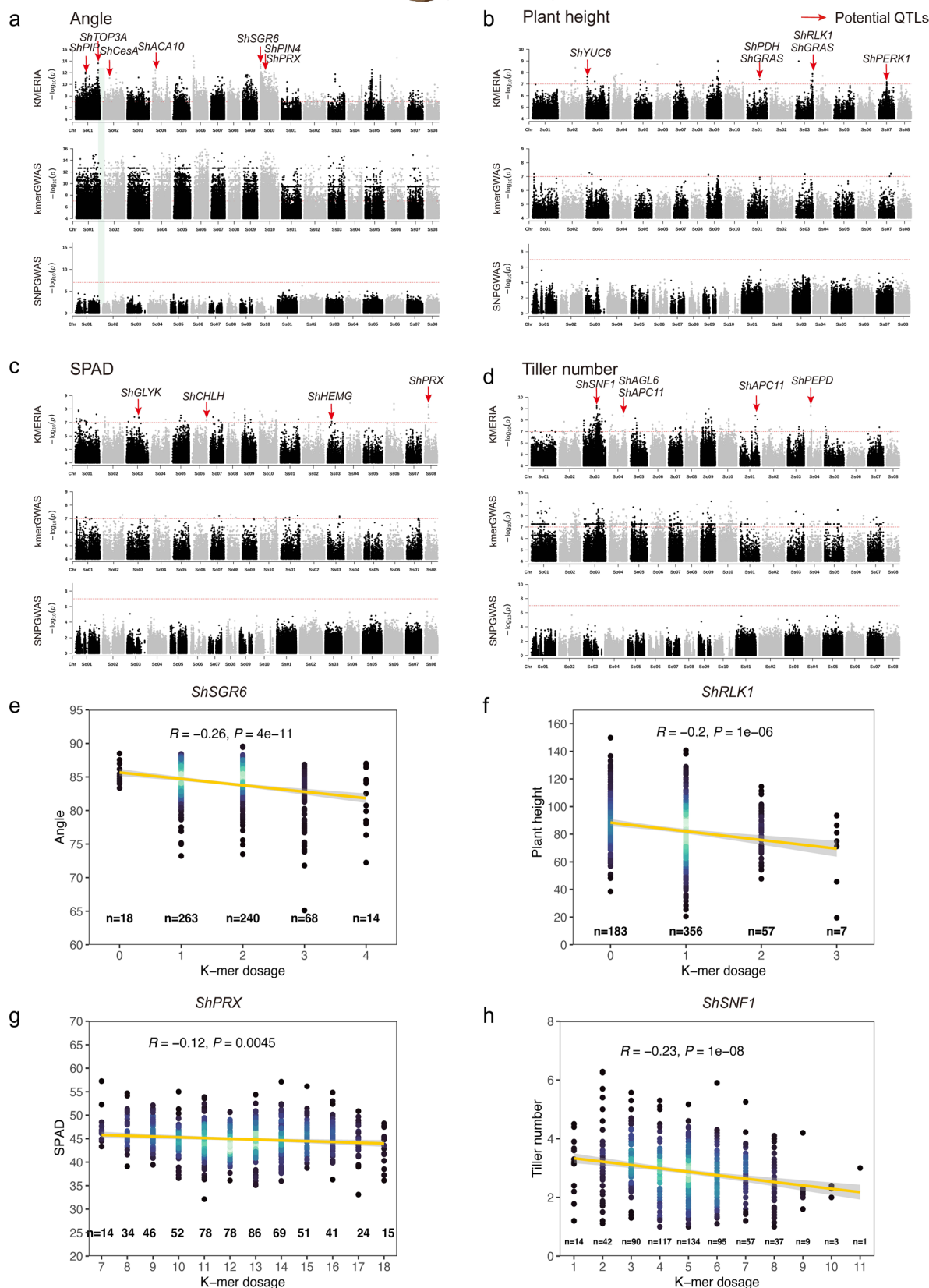


Extended Data Fig. 7 | Comparative GWAS analysis and validation of significant *k*-mer associations for three traits in Hexaploid sweet potato. a–c, Manhattan plots display the GWAS results for three traits in the sweet potato population: (a) leaf shape, (b) flesh color, and (c) skin color. The results are compared across three GWAS methods: KMERIA, kmersGWAS, and GWASpoly. Black arrows indicate the genomic coordinates of known QTLs previously associated with the respective traits, serving as positive controls. Red solid arrows indicate potential QTLs identified by KMERIA. Red dashed boxes highlight the genomic regions where the potential QTLs were missed by GWASpoly. *P* values were calculated using a two-sided Wald test. The horizontal

dashed lines in all Manhattan plots indicate the GWAS significance threshold, adjusted for multiple comparisons using BH-FDR method. **d-f**, Scatterplots illustrate the correlation between the estimated *k*-mer dosage and the respective phenotype for significant candidate genes identified by KMERIA. The top left text represents the Pearson correlation coefficient (*R*), with the corresponding *P*-value (two-sided *t*-test) indicating the significance of the association. The number of accessions (*n*) for each *k*-mer dosage level is provided at the bottom of the plot. The solid line indicates the linear regression fit, and the shaded gray area represents the 95% confidence interval. Individual points are colored by density.



Sugarcane (2n = 10–12×)



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Comparative GWAS analysis and validation of significant *k*-mer associations for cultivated sugarcanes (*S. hybrid*). **a–d**, Manhattan plots display the GWAS results for three traits in the *S. hybrid* population: **(a)** angle, **(b)** plant height, **(c)** SPAD and **(d)** tiller number. The results are compared across three GWAS methods: KMERIA, kmersGWAS, and SNP-GWAS. Red solid arrows indicate potential QTLs identified by KMERIA. *P* values were calculated using a two-sided Wald test. The horizontal dashed lines in all Manhattan plots indicate the GWAS significance threshold, adjusted for multiple comparisons using BH-FDR method. **e–h**, Scatterplots illustrate the correlation

between the estimated *k*-mer dosage and the respective phenotype for significant candidate genes identified by KMERIA. The top left text represents the Pearson correlation coefficient (*R*), with the corresponding *P*-value (two-sided *t*-test) indicating the significance of the association. The number of accessions (*n*) for each *k*-mer dosage level is provided at the bottom of the plot. The solid line represents the linear trend, surrounded by gray shading representing the 95% confidence interval. The color intensity of individual markers reflects point density.



Nature Genetics



Extended Data Fig. 10 | Phenotypic effects of *SsNGA5* overexpression in rice.

a, Comparison of CK-type (n = 11) and *SsNGA5*-OE (n = 9) lines after 3 months of growth under standard laboratory conditions. Data are presented as mean values $\pm$ s.e.m. The heights of the bars represent the mean tiller number for

each group. Individual gray dots represent biological replicates. Statistical significance was determined by a two-sided *t*-test ($P = 0.025$). **b,c**, Representative samples showing tiller number differences between CK-type and *SsNGA5*-OE plants. Arrows indicate tillers. Scale bar, 10 cm.

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 The sequencing reads were undertaken with PacBio REVO and BGI DNBSEQ-T7 platforms.

Data analysis For all analyses, freely available software were used, including: hifiasm(v0.19.8-r603), RepeatModeler (v2.0.3), RepeatMasker62 (v4.1.2), Trimmomatic (v0.39), BWA(v0.7.17), SAMtools(v1.9), minigraph-cactus(v2.6.11), GenomeScope (v2.0), GWASpoly(v2.10), R (v.4.3.1), Merqury (v1.4.1), BUSCO (v5.4.1), LTR_retriever (v2.9.6), LTR_HARVEST_parallel (v1.0), LTR_FINDER_parallel (v1.2), HISAT (v2.2.1), StringTie (v2.2.1), BRAKER (v3.1), KMC (v3.2.158), PloidyFrost (v1.0), Minigraph (v0.21), GATK(v4.3), AlphaSimR (v1.0), GEMMA (v0.98), kmersGWAS (<https://github.com/voichkek/kmersGWAS>) and our developed program KMERIA (<https://github.com/Sh1ne111/KMERIA>).

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](#)

The 14 *S. spontaneum* genome assemblies and annotations generated in this study have been deposited in Figshare database under DOIs (<https://doi.org/10.6084/m9.figshare.32251332>)74 and (<https://doi.org/10.6084/m9.figshare.32142925>)75. The genome assemblies, annotations, and raw PacBio HiFi sequencing reads are also accessible via the National Center for Biotechnology Information (NCBI) under BioProject accession PRJNA1393212. The reference genome assembly of *S. spontaneum*82-114 was obtained from the public repository at the following link (<https://figshare.com/s/e79249d6ce11539fa105>). Previously published resequencing datasets analyzed in this study were obtained from NCBI Sequence Read Archive (SRA) using the following BioProject accessions: potato (PRJNA944441, PRJNA63145, PRJNA766763, PRJNA378971), sweet potato (PRJNA1089346), alfalfa (PRJNA995892), rice (PRJNA623686), and *Saccharum spontaneum*/hybrid (PRJNA1303125).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

NA

Reporting on race, ethnicity, or other socially relevant groupings

NA

Population characteristics

NA

Recruitment

NA

Ethics oversight

NA

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](https://www.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

No statistical methods were used to establish sample size. The 290 *Saccharum spontaneum* sequencing accessions were provided by the Yunnan Academy of Agricultural Sciences. The plant materials were cultivated in a subtropical plateau climate in Kaiyuan City, Yunnan Province

Data exclusions

No data were excluded

Replication

For *S. spontaneum* phenotyping, the experiment was repeated three times. and all attempts at replication were successful.

Randomization

A randomized complete block design was used for phenotyping

Blinding

No analyses required being blind to groups

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	Involvement in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☐	☒ Plants

Methods

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

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	☐ Public health
☒	☐ National security
☐	☒ Crops and/or livestock
☒	☐ Ecosystems
☒	☐ Any other significant area

Hazards

For examples of agents subject to oversight, see the United States Government [Policy for Institutional Oversight of Life Sciences Dual Use Research of Concern](#).

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☒	☐ Demonstrate how to render a vaccine ineffective
☒	☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐ Increase transmissibility of a pathogen
☒	☐ Alter the host range of a pathogen
☒	☐ Enable evasion of diagnostic/detection modalities
☒	☐ Enable the weaponization of a biological agent or toxin
☒	☐ Any other potentially harmful combination of experiments and agents

Precautions and benefits

Biosecurity precautions

Biosecurity oversight

Benefits

Communication benefits

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com