Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A k-mer-based GWAS approach empowering gene mining in polyploids

Xingtan Zhang

zhangxingtan@caas.cn

Agricultural Genomics Institute at Shenzhen https://orcid.org/0000-0002-5207-0882

Shuai Chen

FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, National Sugarcane Engineering Technology Research Center, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, F
https://orcid.org/0000-0002-6861-2682

Xinlong Liu

State Key Laboratory for Tropical Crop Breeding, Sugarcane Research Institute, Yunnan Academy of Agricultural Sciences/Yunnan Key Laboratory of Sugarcane Genetic Improvement

Shenyang Qu

Agricultural Genomics Institute at Shenzhen

Yuhan Song

Agricultural Genomics Institute at Shenzhen

Kun Chai

Chinese Academy of Agricultural Sciences

Hongbo Liu

State Key Laboratory for Tropical Crop Breeding, Sugarcane Research Institute, Yunnan Academy of Agricultural Sciences

Yuebin Zhang

State Key Laboratory for Tropical Crop Breeding, Sugarcane Research Institute, Yunnan Academy of Agricultural Sciences

Zhongqiang Xia

Agricultural Genomics Institute at Shenzhen https://orcid.org/0000-0003-1759-4143

Xiaofeng Li

Agricultural Genomics Institute at Shenzhen https://orcid.org/0000-0001-6033-9979

Jungang Wang

State Key Laboratory of Tropical Crop Breeding, Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences

Muqing Zhang

Guangxi University https://orcid.org/0000-0003-3138-3422

Hongbo Li

College of Horticulture Science and Engineering, Shandong Agricultural University
https://orcid.org/0000-0003-1579-4600

Guo-Bo Chen
People's Hospital of Hangzhou Medical College https://orcid.org/0000-0001-5475-8237

Chris Maliepaard
Wageningen University and Research Centre https://orcid.org/0000-0002-7319-5270

Article

Keywords: GWAS, Polyploid, K-mer, Sugarcane

Posted Date: November 6th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7347406/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. A patent on the KMERIA method has been filed by Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences with S.C., S.Q. X.Z. and M.Z. as inventors. The remaining authors declare no competing interests.

Version of Record: A version of this preprint was published at Nature Genetics on June 12th, 2026. See the published version at https://doi.org/10.1038/s41588-026-02641-8.

A k-mer-based GWAS approach empowering gene mining in polyploids

Shuai Chen[1,#], Xinlong Liu[2,#], Shenyang Qu[1], Yuhan Song[1], Kun Chai[1], Hongbo Liu[2], Yuebin Zhang[2], Zhongqiang Xia[1], Xiaofeng Li[1], Jungang Wang[3], Muqing Zhang[4], Hongbo Li[5], Guo-Bo Chen[6], Chris Maliepaard[7], Xingtan Zhang[1,*]

[1]State 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

[2]State Key Laboratory for Tropical Crop Breeding, Sugarcane Research Institute, Yunnan Academy of Agricultural Sciences/Yunnan Key Laboratory of Sugarcane Genetic Improvement, Kaiyuan 661699, China

[3]State Key Laboratory of Tropical Crop Breeding, Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Sanya, Hainan, 571101, China

[4]Guangxi Key Laboratory for Sugarcane Biology, Guangxi University, Nanning, 530005, Guangxi, China

[5]College of Horticulture Science and Engineering, Shandong Agricultural University, Tai'an 271018, Shandong, China

[6]Center 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.

[7]Plant Breeding, Wageningen University and Research, Wageningen, The Netherlands

[#]These authors contribute equally

[*]Corresponding authors: Xingtan Zhang, **zhangxingtan@caas.cn;**

ABSTRACT

Genome-wide association studies (GWAS) serve as a cornerstone for deciphering the genetic architecture of complex traits. However, conventional GWAS tools, predominantly optimized for diploid species, encounter substantial limitations when applied to complex polyploids due to challenges such as genotyping complexity, multi-allelic variant interpretation, and allele dosage ambiguity. Here, we present KMERIA, a k-mer-based framework specifically engineered 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 both genotyping accuracy and statistical power. To demonstrate its utility in high-ploidy systems, we deployed KMERIA in an auto-polyploid natural population of 290 wild sugarcane accessions (*Saccharum spontaneum)* spanning diverse ploidy levels. To assess biases inherent to linear reference genomes in capturing allelic diversity and structural variations, we constructed a graph-based pan-genome integrating structural variations and haplotype diversity across 15 *S. spontaneum* accessions. Integrating KMERIA with a graph pangenome revealed novel sucrose biosynthesis (*SsMGT*) and tillering regulators (*SsERF14*, *SsNGA5*, *SsNAC*, *SsARF8*, *SsLOG*, *SsSCR*) in *S. spontaneum*, including functionally validated *SsNGA5*. These discoveries not only elucidate the genetic basis of *S. spontaneum* for its yield potential but also provide actionable targets for sugarcane breeding. Collectively, KMERIA bridges a critical methodological gap in polyploid genomics, while the integration of graph pan-genomes provides a robust framework for deciphering genotype-phenotype relationships in crops with complex genome architectures.

Keywords: GWAS, Polyploid, K-mer, Sugarcane

INTRODUCTION

Polyploidy, the presence of multiple sets of chromosomes, is a defining feature of many cultivated crops. Remarkably, approximately 75% of domesticated plants—including staples such as wheat, potato, cotton, and sugarcane—are polyploids[1,2]. This genetic phenomenon has played a pivotal role in plant evolution, driving the development of traits that confer significant advantages over their diploid counterparts[3–5]. Polyploidy often leads to increased vigor, larger plant organs, and enhanced genetic diversity, which collectively contribute to improved resilience against environmental stresses and deleterious mutations[6,7]. These attributes have made polyploid crops indispensable in agriculture, enabling the development of high-yielding, stress-tolerant, and quality-enhanced cultivars[8,9]. As such, understanding the genetic architecture of polyploids is critical for advancing crop improvement and ensuring global food security.

Despite their agricultural significance, the genetic complexity of polyploids poses substantial challenges for traditional genetic and genomic analyses, particularly for GWAS. GWAS has emerged as a powerful tool for identifying genetic variants associated with traits using natural mapping population. However, conventional GWAS methods, primarily developed for diploid organisms, often fail to adequately address the redundancy and allelic diversity characteristic of polyploid genomes. These limitations hinder the detection of critical genetic mechanisms, such as non-additive effects and dosage-dependent phenotypic contributions, leading to the loss of valuable genetic information. Although a few studies have implemented allele dosage-based GWAS approaches using GWASpoly in polyploids (e.g., tetraploids and hexaploids)[10–12], these efforts often come at the expense of substantial computational resources and time costs on variant calling and genotyping. Consequently, there is a pressing need for innovative approaches specifically designed to unravel the unique genetic architecture of polyploids.

GWAS can leverage k-mers derived from whole genome sequencing (WGS) data as genetic markers. Unlike traditional markers such as single nucleotide polymorphisms (SNPs), k-mers capture a broader spectrum of genomic variations, including long INDELs and SVs. This enables a more comprehensive exploration of genetic diversity. Furthermore, k-mer-based GWAS operates independently of a reference genome, mitigating biases inherent in genome comparisons. For example, in a study involving 242 accessions of the diploid wheat progenitor *Aegilops tauschii*, k-mer-based analysis pinpointed discrete disease resistance regions that were successfully transferred into hexaploid wheat[14]. In addition, application of the k-mer-based GWAS in hexaploid wheat (*Triticum aestivum*) identified 25% more loci associated with powdery mildew resistance compared to SNP-based methods, while achieving higher mapping precision[13].

Collectively, these findings highlight the potential of k-mer-based GWAS to address the "missing heritability" problem by uncovering genetic factors that conventional GWAS methods may overlook, offering new insights into the genetic basis of complex traits[15].

Despite its promise, the application of k-mer-based GWAS to polyploid species remains limited. Current implementations fail to account for allele dosage effects, which are critical for understanding the phenotypic variation in polyploid organisms[16,17]. In this context, variations in k-mer copy numbers across populations provide a robust alternative, offering a more precise means of assessing allele dosage. This makes k-mer-based GWAS particularly advantageous for polyploids, where complex patterns of genetic variation and dosage-dependent effects play a significant role in shaping phenotypes.

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 adaptability[18]. Its genome is characterized by high ploidies, ranging from 4 to 16, and extreme complexity, including extensive homologous regions and frequent chromosomal rearrangements[19], 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 potential genetic loci by correlating variations in k-mer copy numbers with phenotypic differences, offering a robust approach to decipher the genetic architecture of complex polyploid genomes. We also constructed a graph-based pan-genome map by integrating PacBio HiFi sequencing technology from 15 *S. spontaneum* accessions and re-sequenced 290 accessions. Using KMERIA, we identified several important regulatory genes linked to critical agronomic traits, including tillering number (TN) and apparent purity (AP). By leveraging the graph-based pan-genome framework, we achieved precision mapping of genes associated with targeted agronomic traits in *S. spontaneum*. Furthermore, we evaluated the performance of KMERIA using simulated data and real mapping populations across crops with varying ploidy levels, including rice, potato, alfalfa, and sweet potato, demonstrating its broad applicability in polyploid crop improvement. Our study introduces an enhanced GWAS approach tailored for polyploids, offering a robust approach to accelerate molecular breeding initiatives and address the challenges of polyploid genome analysis.

RESULTS

Challenges in applying GWAS to polyploids

GWAS hold immense promise for dissecting the complex genetics of polyploid crops, but substantial challenges remain, which stem primarily from the inherent difficulties in analyzing highly similar, repetitive sequences within polyploid genomes.

One of the major challenges is the accuracy of short-read alignment. Polyploid genomes are characterized by high sequence similarity among homologous chromosomes, which complicates the alignment of resequencing data. This similarity often leads to multi-mapping of reads, where a single read aligns to multiple homologous regions. Such ambiguities result in inaccurate alignments and biased read distribution, particularly in repetitive regions. Furthermore, the reliance on a single reference genome exacerbates these issues. Non-reference alleles may be underrepresented or misaligned, introducing reference bias and limiting the detection of true genetic variation (**Figure 1a**). These alignment challenges underscore the need for advanced methods that can accurately resolve homologous regions and account for allelic diversity in polyploid genomes. To illustrate these alignment challenges, we simulated 200 million paired-end reads based on the genome of *Saccharum hybrid* cultivar POJ2878, a recently characterized auto-allopolyploid species with a genome size of 10.37 Gb. This genome was fully resolved by our team into 118 chromosomes, organized into 10 homo(eo)logous groups (HGs). Alignment results revealed that only 43.08% of reads were high-quality mapped reads ($MAPQ \geq 1$). Of these high-quality reads, 49.49% aligned correctly to their original genomic loci, while 50.51% were misaligned to other regions. This misalignment propagates errors through downstream analyses, such as variant calling, complicating the genotyping process (**Figure S1** and **Table S1**). These findings highlight the urgent need for advanced alignment methods capable of resolving homologous regions and accounting for allelic diversity in polyploid genomes.

Genotype calling in polyploids presents another significant hurdle. Tools such as updog[20], SuperMASSA[21], and fitPoly[22] have been developed to estimate genotypes by modeling read depth, preferential pairing, and technical artifacts, often incorporating uncertainty metrics. However, these methods typically require pre-identified single nucleotide polymorphisms (SNPs), which are themselves difficult to detect accurately in polyploids. Moreover, widely-used tools like the Genome Analysis Toolkit (GATK), effective in diploids, struggle in polyploids due to ambiguous alignments and frequent chromosomal rearrangements. GATK's reliance on high-quality reference alignments for variant calling is often unmet in polyploids, leading to errors in identifying SNPs, small insertions/deletions (indels), and structural variations.

Additionally, GATK employs a Bayesian framework to infer genotypes by evaluating all possible allele combinations. While this approach is effective for diploids, it becomes computationally unwieldy in polyploids due to the exponential increase in potential genotype combinations. For each potential genotype $g' \in \mathcal{G}$, GATK must compute the likelihood $Pr(D|g)$. As $|\mathcal{G}|$ grows exponentially with P and A, the computational burden becomes unsustainable for high-ploidy species (**Figure 1b-c**, **Supplementary Note 1, Formula 1 and 2**). This computational burden is further compounded by the complexities of allele dosage in polyploids. Unlike diploids, where allele ratios are straightforward, polyploids exhibit a wider range of allele ratios depending on ploidy and genotype. Insufficient sequencing depth can obscure these ratios, leading to misclassification of genotypes or failure to detect low-frequency alleles[23].

Algorithm overview of KMERIA and validation datasets

To overcome these challenges, we propose a k-mer-based computational method (KMERIA) specifically designed for the multidimensional complexity of polyploid genomes[24]. This method enables comprehensive detection of trait-associated loci while maintaining biological interpretability in polyploid. This KMERIA workflow implements five integrated steps (**Figure 2**): *Counting*, *Matrixing*, *Filtering and correcting*, *K-mer testing*, and *post-GWAS*.

Counting*:** The initial processing converts preprocessed resequencing data into whole-genome k-mer profiles that capture the full spectrum of genetic variations (SNPs, InDels, SVs) without prior variant calling. ***Matrixing*:** A computationally efficient strategy was employed to construct a population-level k-mer genotype matrices based on the k-mer abundance while preserving allelic dosage information. ***Filtering and correcting: A dual-filtering strategy systematically removes low-confidence k-mers through (i) abundance thresholding (eliminating sequencing artifacts) and (ii) repeat masking (excluding uninformative repetitive elements). Ploidy-aware normalization then adjusts copy number estimates, followed by quantile-based dosage recalibration to generate ploidy-specific allele dosage matrices that represent polyploid genetic architectures. ***K-mer testing*:** A mixed linear model (MLM) incorporates population structure covariates and kinship to address polyploid-specific confounding factors while analyzing dosage-dependent trait associations. ***Post-GWAS*:** A dual-correction strategy, incorporating BH-FDR and modified Bonferroni correction, identifies significant associations. Subsequently, graph-based genome alignment maps significant k-mers or k-mer-associated reads to pangenome reference coordinates. This structural-aware mapping facilitates precise localization of causal elements in rearrangement-prone polyploid genomes, enabling three-

dimensional trait dissection through haplotype-resolved gene networks and cis-regulatory landscape analysis.

To validate the performance of KMERIA, we conducted phenotypic simulations using established polyploid populations. Additionally, to assess its performance across diverse scenarios, we also constructed multiple autopolyploid population datasets. These datasets were generated by creating 'synthetic' autopolyploid genomes and collecting real datasets with varying ploidy levels. (**Table S2**). To reflect the characteristics of real autopolyploid genomes in our simulation data, we constructed several 'synthetic' tetraploid genomes by integrating the chromosome-level haploid assemblies of *Arabidopsis* Col-0 with simulated genetic variations[25]. These synthetic genomes were used to generate subsequent resequencing datasets. These 'synthetic' genomes served as the basis for generating subsequent resequencing datasets. Following the establishment of these 'synthetic' genomes, we also simulated genetic mapping populations for polyploids. This involved generating genotypes for multiple tetraploid populations by variable heritability of the simulated traits and different numbers of quantitative trait nucleotides (QTLs) for various traits (**Figure S2**). This approach allowed us to generate resequencing datasets at different population levels, facilitating to assess the performance of KMERIA. To further demonstrate the power of KMERIA, we also applied the algorithm to several public polyploid datasets, including hexaploidy sweet potato[10], tetraploid potato[11], and tetraploid alfalfa[26], to discover potential phenotypic association loci. Additionally, we extended the application of KMERIA to diploid crops such as rice[27] and the complex genome of cultivated sugarcane population, one of the most intricate aneuploidy genomes[28], highlighting the broad applicability of KMERIA across diverse genetic characteristics of species.

Statistical power under different levels of Type I error rate (α)

We evaluated its statistical power and false discovery rate (FDR) under varying heritability (h^2) levels and QTLs using a simulated dataset of 294 hexaploidy sweet potato accessions. This 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. Genetic markers were classified as residing within QTL regions or non-QTL regions for this assessment. We first characterized the genetic features of population via k-mer analysis. This yielded 450 million non-redundant k-mers across all individuals. Only 100 million (22.2%) of these k-mers were present in all of the genomes (**Figure 3a**), indicating substantial genetic variations within this population. The distribution of total k-mer occurrence counts (KOCs) revealed that the vast majority of k-mers had low

KOCs (1-6), consistent with the sweet potato genome structure (**Figure 3b**). This demonstrates that k-mer KOCs effectively represent allelic dosage in polyploids. Subsequently, PCA performed on the KOC matrix produced population clusters consistent with published studies (**Figures 3c-d**), validating k-mer KOCs as an alternative for population structure study.

KMERIA was compared against three widely used GWAS methods: GEMMA[29], a standard GWAS approach for diploids, a k-mer based GWAS approach kmersGWAS[16], and GWASpoly[12], an algorithm specifically designed for autopolyploids. For traits influenced by a small number of major-effect genes, we simulated 10 QTLs with effects following a gaussian distribution (**Figure S3**). 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 (**Figure 3e and Figure 3g**). These results demonstrate the higher power of KMERIA in detecting major-effect QTLs. In scenarios involving polygenic traits influenced by 100 QTLs (**Figure S3**), KMERIA again exhibited higher statistical power compared to diploid GWAS approach (GEMMA), kmersGWAS, and GWASpoly, regardless of heritability levels (**Figure 3b and Figure 3h**).

To further assess statistical performance, we also compared FDR between KMERIA and established GWAS approach by recording false positive QTLs across 10 repeated tests. KMERIA consistently demonstrated lower FDR than GEMMA, kmersGWAS, and GWASpoly across most scenarios (**Figures 3i-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. Across the simulated datasets, KMERIA showed a strong correlation ($R = 0.54$, $P < 0.001$) with GWASpoly in identifying significant association loci (**Figure 3m**). Importantly, 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: (i) the proportion of k-mers used to infer population structure in the LMM; (ii) the optimal k-mer size for GWAS analysis; (iii) the effect of missing k-mer markers on polyploid GWAS outcomes; and (iv) the impact of population size on polyploid GWAS results.

KMERIA employs 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% of k-mer tags from the tetraploid potato genotype matrices to perform principal component analysis (PCA) 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, $P < 2.2\times10^{-26}$) and aligned well with the known population structure[11] (**Figure S4**). Notably, even with just 0.1% of the 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 lengths 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, k is the k-mer size). Among the tested values, 31-mer exhibited the lowest FDR, indicating that larger k-mer sizes enhance signal specificity in polyploid GWAS (**Figure 3n** and **Figure S5**), 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 (**Figure S6**), 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 simulated auto-tetraploid populations with sample sizes of 200, 400, 600, and 800 individuals. Using a fixed k value of 31, we conducted 100 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 (**Figure 3o**). This trend aligns with findings from diploid GWAS studies[30], reaffirming that increasing sample size is an effective strategy to enhance power in polyploid GWAS.

Computational Efficiency

In addition to enhancing statistical power, KMERIA demonstrates impressive computational efficiency. We conducted a theoretical analysis of the time complexity for KMERIA genotyping in polyploid GWAS and compared its performance to that of GATK (with ploidy mode) using real datasets (**Supplementary Note 1 and Figure S7**). The complexity of polyploid genotypes leads to a rapid increase in the number of possible genotype combinations at each variant site, which substantially elevates the computational demands for genotype inference using the Bayesian-based method (i.e., GATK). This results in considerable time and resource costs. Furthermore, for high-ploidy species, GATK (-ploidy mode) often struggles to reliably determine genotypes, limiting its utility in polyploid research[23].

To directly compare the efficiency of KMERIA and GATK (-ploidy mode), we tested both methods on a real dataset from a tetraploid potato population with 100 individuals. The genotype calling was performed on a single-node Linux platform equipped with 28 cores and 196 GB RAM. The results revealed that GATK (-ploidy mode) required approximately 35,413 real-time hours to complete the genotyping progress, while KMERIA accomplished the same task in only 38.9 to 82.3 hours - approximately 1/430 of the time taken by GATK (**Table S3 and Figure S8**). The peak of memory usage by KMERIA was comparable to that of GATK (-ploidy) and remained within acceptable limits (**Figure S8**).

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 rice[27], tetraploid alfalfa[26] and potato[11], hexaploid sweetpotato[10], and highly polyploid cultivated sugarcane, focusing on key agronomic traits.

We first tested KMERIA in a diploid rice population comprising 400 accessions[27], focusing on amylose content (AC), a critical determinant of grain quality. KMERIA successfully re-identified *Waxy*[31] gene, a well-known major locus controlling AC. Compared to the conventional GWAS method, KMERIA exhibits significantly enhanced association signals for the trait with superior statistical validation (KMERIA: $P = 2.89\times10^{-29}$; Wei et al study[27]: $P = 4.51\times10^{-17}$) (**Figure 4a**).

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 (SLR),

a key trait influencing forage quality and plant morphology. KMERIA identified two novel QTLs on chromosomes 3 and 5, containing *MsTUB5*[32], *KINESIN-like*[33], and *MsKEG*[34] (*KEEP ON GOING*), genes known to regulate plant morphogenesis (**Figure 4b** and **Figure S9**). In potato, we analyzed 137 accessions with an average sequencing depth of 67×, focusing on tuber flesh color, an essential quality trait in potato breeding. KMERIA identified significant loci, including known QTLs on Chr03 [35] and Chr06 [11]. Additionally, it uncovered novel QTLs on Chr04, and Chr09, harboring two functional genes, *StAUR*[36], and *St3GT*[37] involved in the root development and flavonoid biosynthesis pathway (**Figure 4c** and **Figure S10**). Notably, SNP-based GWAS failed to detect novel several loci identified by KMERIA in both crops, further highlighting KMERIA's superior detection power in 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 monoploid genome[10]. We identified 472 million non-redundant k-mers present in at least 50 accessions. Among the significant QTLs detected ($P < 0.05$, BH-FDR multiple-testing correction), we successfully re-identified *IbFbox*, a previously reported gene associated with leaf shape[10,38] (**Figure 4d and Figure S11**), and a novel leaf shape-associated gene encoding *CUP SHAPED COTYLEDON* (*IbCUC3*), which is a cell growth inhibitor[39] (**Figure 4a**). Notably, the leading k-mer identified by KMERIA exhibited lower *P*-value than those obtained using the GWASpoly[12] (**Figure S11**). Similar superiority was observed for two additional traits (**Figure S11**), highlighting the ability of KMERIA to detect both known and novel associations in hexaploidy genomes.

To further demonstrate the effectiveness of KMERIA in highly complex genomes, we tested it on hybrid sugarcane, which has an estimated genome size of 10 Gb and exhibits aneuploidy levels ranging from 9- to 10-fold across each homoeologous chromosome group. This population consisted of 603 individuals with an average sequencing depth of 118×. While SNP-GWAS failed to detect any leading signals exceeding the significant threshold, KMERIA identified 12 QTLs linked to key agronomic traits, including tiller angle, plant height, soil-plant analysis development (SPAD) and tiller number (**Figure S12**). This striking contrast highlights the ability of KMERIA to overcome the limitations of traditional SNP-based methods.

Construction of *S. spontaneum* graph pangenome

S. spontaneum has played a pivotal role in modern sugarcane breeding, contributing essential genetic factors for enhanced tillering, improved sucrose synthesis, and greater photosynthetic efficiency—traits that collectively boost yield, sugar accumulation, and biomass production[18]. To maximize the improvement of

genetic diversity in *S. spontaneum*, we selected 14 representative accessions from a pool of 290 *S. spontaneum* accessions based on k-mer-based phylogenetic relationships (**Figure 5a and Figure S13**). These accessions, spanning 11 geographic regions, reflect broad genetic diversity (**Table S4**).

To account for the genomic ploidy variation among *S. spontaneum* accessions, we estimated ploidy levels using re-sequencing reads. The selected accessions exhibited ploidy levels ranging from 8 to 11, with decaploids comprising 11 of the 14 samples (**Table S4**). Notably, decaploids also dominated the broader pool of 290 accessions, representing 68.9% (200/290) of the samples. High-depth sequencing generated an average of 183 Gb HiFi reads per accession, corresponding to ~225× genome coverage based on the *S. spontaneum* monoploid genome size (~800 Mb) (**Table S5**).

These HiFi reads were assembled into 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 Table S6**). Quality assessments confirmed the robustness of these assemblies: BUSCO analysis revealed an average completeness score of 99.4%, while Merqury QV values averaged 64.3, indicating high base accuracy and assembly completeness. Additionally, the LTR Assembly Index (LAI)[40] ranged from 14.47 to 16.43, confirming that the assemblies achieved reference-grade quality (**Figure 5b, Table 1**). Gene prediction across the assemblies identified between 231,354 (Ss09, ploidy: 8) and 332,535 (Ss02, ploidy: 11) protein-coding genes (including allelic genes) through integrating *ab initio*-, homology-based and RNA-seq prediction (**Table S7 and Table S8**). BUSCO[41] evaluation consistently yielded scores above 98%, affirming the completeness and representation of core genes (**Table S7**).

Due to the absence of Hi-C data, genome de-redundancy was performed for each *S. spontaneum* assembly to meet graph pangenome construction requirements. The deduplicated genomes were scaffolded using the Sspon82-114 telomere-to-telomere monoploid (T2T-mono) assembly (reported in our companion study **Table S9**), resulting in 14 monoploid assemblies ranging from 760 Mb to 772 Mb in size. BUSCO completeness assessments averaged 93.5%, confirming the suitability of these assemblies for graph pangenome construction (**Table S9**).

We subsequently built a graph-based pangenome incorporating these 15 *S. spontaneum* monoploid genomes, including Sspon82-114 T2T-mono genome. The resulting graph spanned ~3 Gb, comprising 142,571,025 nodes (sequence fragments) connected by 193,033,606 edges (connections between nodes). Non-reference nodes accounted for 2,029 Mb of the graph (**Figures 5c-d**). Structural variants (SVs $\geqslant$ 50 bp) were identified using a bubble-popping algorithm, yielding 277,133 SVs, with 70.1% < 1,000 bp, 21.5%

< 10 kb, and 8.4% > 10 kb. These SVs were genotyped against the *S. spontaneum* T2T-mono reference genome or at least one other genome (**Figure 5c and Table S10**). Of the identified SVs, 42.7% were biallelic (insertions, deletions, or divergent alleles), while 57.3% were multiallelic, spanning 2.51 Gb and containing more than one non-reference path (**Figure 5d and Table S11**). Among reference genes, 118,552 biallelic SVs were located in potential regulatory regions or coding sequences (CDS), while multiallelic SVs impacted 63.1% of reference genes (**Figure 5d and Table S11**).

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 methods[19], limiting the application of GWAS. To identify potential functional genetic variants associated with important agronomic traits in sugarcane, we employed KMERIA for GWAS on two crucial traits—apparent purity (AP) and tiller number (TN) (**Figures 6a-c, and Table S12**).

Using k-mer genotype matrices constructed from 290 *S. spontaneum accessions*, we identified 12.4 billion non-redundant k-mers present in at least 30% of individuals. For AP, a significant association signal was detected at Chr06: 8.1 Mb on the Sspon82-114 T2T-mono genome (**Figure 6d**). This locus contains a gene encoding a MAGNESIUM TRANSPORTER (*SsMGT*), which likely regulates sucrose synthesis by modulating Mg^{2+} transport, influencing sugarcane quality and yield[42,43]. The top associated k-mer locating at the upstream of *SsMGT* exhibited a significant negative correlation with AP (R = -0.3, $P = 2.0\times10^{-7}$) (**Figure 6e**). Further analysis of the *S. spontaneum* pangenome revealed a 7-bp insertion near the transcription start site (TSS) of *SsMGT* in a distinct haplotype (**Figure 6f**). This insertion may disrupt transcriptional initiation, impair 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. Our analysis based on KMERIA GWAS identified 598 significantly associated k-mers ($P < 2.0\times10^{-7}$; **Figure 6g**), with the strongest signal mapped to a 18.9 Mb locus on Chr01. This genomic region harbors an *ETHYLENE-RESPONSIVE FACTOR* (*SsERF*) gene orthologous to *OsEATB* (*OsERF*)[44], 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 tiller number, a critical yield determinant, where dwarf cultivars frequently demonstrate higher productivity (**Figure 6g**). The dosage effect of the top-associated k-mer

exhibited a significant positive correlation with TN ($R = 0.38$, $P = 4.6\times10^{-11}$) (**Figure 6h**). Additional tiller-regulatory genes were identified, including *NGATHA5* (*SsNGA5*)[45], *NAC DOMAIN TRANSCRIPTION FACTOR* (*SsNAC20*)[46], *AUXIN RESPONSE FACTOR* (*SsARF8*)[47], *LONELY GUY* (*SsLOG*)[48], and *SCARECROW* (*SsSCL*)[49]. Orthologs of these genes in rice and related species are well-established regulators of tiller development. 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 (**Figure 6i**). Notably, overexpression of the *SsNGA5* gene in rice (ZH11, *Oryza sativa* ssp. *japonica*) resulted in a significantly decreased tiller number ($P = 0.025$, Student's t-test; **Figures S14a-c**). These findings provide potential molecular targets for breeding programs aimed at improving sugarcane yield-related traits.

The application of KMERIA not only facilitated the identification of key genetic determinants of agronomic traits but also demonstrated the critical utility of the *S. spontaneum* graph-based pangenome in resolving k-mer localization challenges inherent to autopolyploid genomes. For TN trait, 35.5% (40,074/112,807) of trait-associated k-mers ($P < 1\times10^{-4}$) aligned to coordinates in the linear reference genome (Sspon82-114 T2T-mono). However, 62.6% of k-mers absent from the reference were successfully localized using structural annotations from the graph pangenome. Crucially, comparative analysis revealed that the graph pangenome approach enhanced associated k-mer mappability by 59.5% compared to linear reference-based alignment (**Table S13**), enabling recovery of previously inaccessible loci in non-reference genomic regions.

Our study demonstrates the efficacy of KMERIA in addressing the inherent challenges of GWAS implementation in complex autopolyploid genomes, establishing a robust framework for precise identification of key genetic loci governing agronomic traits in autopolyploid crops.

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 genetics. A key innovation of KMERIA lies in its ability to circumvent the constraints of conventional genotyping pipelines. For instance, while widely used tools like GATK[50]—originally designed for diploid species—struggle with allele combination complexity and multi-locus genotyping in polyploids, KMERIA eliminates the need for error-prone read

alignments and reference-dependent genotyping. This approach not only enhances genotyping efficiency but also minimizes reference bias, a critical advantage in species lacking high-quality reference genomes.

A pivotal strength of KMERIA is its capacity to quantify allele dosage, a parameter with profound phenotypic implications in polyploid systems. Unlike diploid k-mer GWAS that employ binary presence-absence (0 or 1 for PA k-mers[16]) encoding, 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 facilitates the detection of genotypes, including structural variants and complex allele interactions that are frequently undetected by conventional SNP-centric approaches. By providing a holistic view of genetic diversity, KMERIA reveals associations that would otherwise remain obscured, particularly in polyploid species where structural rearrangements are evolutionarily prevalent.

Despite these advancements, certain limitations warrant consideration. Although k-mer abundance profiles reflect population-level polymorphisms, resolving the causal genetic variants driving phenotypic variation remains challenging in polyploids. This ambiguity arises because k-mers may map to multiple homologous loci in complex polyploid genomes, complicating precise variant localization. While integrating graph-based pangenomes could mitigate reference bias and accelerate variant identification as demonstrated in our preliminary analysis—such approaches are currently constrained by the limited availability of polyploid pangenome resources. Additionally, despite implementing rigorous pre-processing steps (e.g., depth-based k-mer filtering), residual sequencing artifacts may persist, potentially introducing noise into downstream analyses.

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 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 autopolyploid organisms.

FIGURES AND LEGENDS

Figure 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 $\mathcal{G}$ **c.** A diagram depicting the complexity and computational time required for genotyping genomes of species with varying ploidy levels using GATK**. d.** A simplified schematic showing how k-mer abundance variation across individuals can be utilized to detect genetic variants, which can subsequently be linked to phenotypic traits.

Figure 2. Overview of the KMERIA pipeline. KMERIA workflow includes five steps: *Counting*, *Matrixing*, *Filtering and correcting*, *K-mer testing*, and *Post-GWAS*. Following quality control 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) via 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.

Figure 3. Performance of KMERIA in simulated polyploid datasets under different scenarios. a. The distribution of filtered k-mer numbers in 294 sweet potato accessions. **b**. The distribution of total normalized k-mer occurrence count per k-mer summed from 294 sweet potato accessions. **c-d**. Principal component analysis (PCA) of sweet potato accessions. **c** PC1 versus PC2, **d** PC1 versus PC3. **e-h**. Power at various levels of Type I error. 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 employed to test these populations: GEMMA, GWASpoly, kmersGWAS and KMERIA. **i-l**. The bar plot illustrates the false discovery rate (FDR) of four methods under Type I error significance thresholds. The ‘synthetic’ population consisted of autotetraploid individuals, with a total of 200

individuals, and population structure information was incorporated. The population structure is visualized through the first two principal components (PCs) in the PCA scatter plot shown in **Figure S2. m**. Correlation of *P*-values between the top k-mers and SNPs. **n**. The point plot is used to assess the effect of k-mer size (ranging from 17 to 31) on the false discovery rate. **o**. The bar plot indicates the statistical power in simulated autotetraploid populations with varying population sizes (100, 200, 400, 800).

Figure 4. 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** show the GWAS association results for various traits: **a**. Amylose content of diploid rice. **b**. Stem-leaf ratio of tetraploid alfalfa. **c**. Flesh color of tetraploid potato. **d.** Leaf shape of hexaploid sweet potato. The left panel shows 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 novel QTLs identified in this study. Genes significantly associated with the traits are highlighted in red within the QTLs. GWAS significant threshold was determined use modified Bonferroni or BH-FDR corrections based on the number of the k-mers. The right panel shows the QQ plot for each GWAS result.

Figure 5. Construction of graph-based pan-genome map from the 15 *S. spontaneum.* **a**. The 290 re-sequenced *S. spontaneum* genomes were subjected to principal component analysis (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*. PC, principal component. **b**. Assembly quality assessment of 14 *S. spontaneum* species was performed using the Long Terminal Repeat (LTR) Assembly Index (LAI) score and QV, with the K-mer-based strategy implemented in Merqury (i.e., quality consensus). The upper and lower edges of the box plots represent the first and third quartiles, respectively, the central line denotes the median, and the whiskers extend to 1.5× the interquartile range (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 structural variation (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. M represents megabase pairs, and G represents gigabase pairs. **f**. Total length of the core genome with varying numbers of individuals.

Figure 6. 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 Apparent Purity (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 arrows and dashed lines. **e.** Correlation analysis between the top k-mer dosage of the major-effect QTL on chromosome 06 and the AP trait. **f**. A 7-bp InDel (InDel_6_8306558) highlighted by the red triangle, located upstream of *SsMGT* (*Ss82114mono.06G0005950*). **g.** 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**. h.** Correlation analysis between the top k-mer dosage of the major-effect QTL on chromosome 01 and the TN trait. **i.** Expression heatmap of TN trait candidate genes across different tissues (leaf, stem, root, tiller-1, tiller-2, tiller-3) in *S. spontaneum* accession 82-114.

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 (**Figure 2)** comprises the following steps:

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 KMC[51] is also supported. We employed 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 prior to association testing, as these likely represent sequencing errors.

Matrixing: We developed a computationally efficient strategy to build population-level k-mer genotype matrices from individual k-mer count files using 'kmeria kctm2'. This strategy preserves allelic dosage information critical for polyploid analysis.

Filtering and correcting: Invalid k-mers (low-abundance k-mers likely representing errors and highly repetitive k-mers typically from non-informative genomic regions) were filtered based on their frequency distribution within individual genomes to reduce noise and computational burden. Additionally, we corrected for the influence of variable sequencing depth on k-mer counts. The average haploid depth d was estimated by identifying the main peak in the k-mer count frequency distribution. This ploidy-aware correction adjusts k-mer copy number statistics, ensuring accurate representation in the genotype matrix. Finally, k-mers with a missing rate exceeding 80% across the population were removed based on genome ploidy and association validity considerations. The calculation formula is:

$$d = C/(g \times k)$$

where C represents the k-mer count, g denotes the haploid genome size, and k is the k-mer length. Subsequently, the corrected k-mer count $N_{corrected}$ is calculated using the formula:

$$N_{corrected} = C/d$$

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 5th 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:

$$y = \begin{cases} 0, & x \le x_{0.95} \\ \dfrac{2 \times (x - x_{0.05})}{x_{0.95} - x_{0.05}}, & x_{0.05} < x < x_{0.95} \\ 2, & x \ge x_{0.95} \end{cases}$$

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. A 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, 3 PCs were used for population stratification correction. Kinship relatedness was calculated using GEMMA (https://github.com/genetics-statistics/GEMMA) with default parameters. The KMERIA association study model using a mixed linear model represented by the equation:

$$\mathbf{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot \boldsymbol{kmer} + \sum_{k=1}^{m} \boldsymbol{\gamma}_k \cdot \boldsymbol{PC}_k + \mathbf{u} + \boldsymbol{\varepsilon}$$

Where $\mathbf{y}$ represents the vector of observed phenotypic values (quantitative trait measurements) for all individuals in the study population;

$\boldsymbol{\beta_0}$ is fixed effect, representing the overall mean of the phenotype when all other variables are zero;

$\boldsymbol{\beta_1}$ indicates the effect size (fixed effect coefficient) of the k-mer on the phenotype - the primary target of inference in GWAS;

$\boldsymbol{kmer}$ is the genotype vector for a k-mer across individuals, coded as scaled allele dosage;

$\boldsymbol{\gamma}_k$ represents the fixed effect coefficient for the k-th principal component (PC) of population structure, correcting for population stratification;

$\boldsymbol{PC}_k$ is the principal component 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 $\boldsymbol{N}(\mathbf{0}, \boldsymbol{K}\sigma_g^2)$ where K is the kinship matrix;

$\boldsymbol{\varepsilon}$ is residual error vector representing environmental noise and model misspecification, distributed as $N(0, \mathbf{I}\sigma_e^2)$.

Post-GWAS. To control false discoveries while maximizing detection sensitivity, we implemented a dual-correction strategy for association testing. First, Benjamini-Hochberg false discovery rate (BH-FDR) correction (adjusted-$P < 0.05$) 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. Significant k-mers and their linked sequencing reads were mapped to our haplotype-resolved graph pangenome reference using Minigraph (v0.21)[52] with the following parameters: 'minigraph -x sr'.

Simulation of autopolyploid genomes and mapping populations

Simulation of mapping population. We simulated auto-polyploid mapping populations using the R package AlphaSimR[53] with the *Arabidopsis thaliana TAIR10* haploid genome as reference to enable precise variant tracking. Following the workflow in **Figure S2**, we generated mapping populations across varied ploidy levels by introducing mutations at 0.01 per base pair, yielding 2,000 segregating sites per chromosome (10,000 genome-wide). From these sites, we designated 100 genome-wide quantitative trait loci (QTLs; 10 per chromosome) and 10,000 genetic markers (2,000 per chromosome). Founder populations of 200 individuals (effective population size = 200) underwent 100 generations of selection with random mating (each male × two females; two offspring per female), maintaining 100 individuals per generation. Population sizes were systematically varied (100, 200, 400, 600, 800) to assess scaling effects. All simulations considered strictly additive genetic effects, with additive QTL effects scaled to achieve target genetic variances for two heritability (h^2) scenarios: (1) tetraploid (4x), h^2 = 50%; (2) tetraploid (4x), h^2 = 80%. The additive effects were then scaled to achieve a desired genetic variance. Ten independent replicates were generated per parameter combination.

The simulation of polyploid genomes were performed by introducing 7000 single base changes, 750 indels of random lengths less than 10 bp and 750 indels of random lengths between 10 and 1000 bp at random locations. Then different number of reads were generated using wgsim (https://github.com/lh3/wgsim) introducing more mutations with default parameters and paired end reads of length 150 (~10× genome coverage) were generated with sequencing error rate of 0.01.

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:

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

$$\boldsymbol{\alpha_i} \sim N(0, \sigma^2)$$

$$\boldsymbol{\sigma_e^2} = \frac{\sigma_\alpha^2}{h^2} - \sigma_g^2$$

$$\boldsymbol{e} \sim N(0, \mathbf{I}\sigma_e^2)$$

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. According to the equation, we simulated phenotypic values of tetraploid for different combinations of population sizes (100, 200, 400, 800), numbers of QTLs (10, 100), heritability (50%, 80%).

Power evaluation across different Type I Error levels.

The statistical power, Type I error (α), and false discovery rate (FDR) were evaluated in association tests conducted on simulated datasets, which included pre-defined QTLs. The evaluation utilized the method[54], alongside two approaches from our previous research[55,56]. 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 QTLs, 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 non-independent 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 Benjamini-Hochberg FDR (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 $\frac{k-1}{k}$ nucleotide overlap, violating independence assumptions underlying traditional corrections. We therefore calculated the effective number of independent tests. The formula as follow:

$$M_{eff} = \frac{N_k}{k}$$

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 length (e.g., 31 bp). The genome-wide significance threshold was then defined as:

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

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 *Saccharum spontaneum* accessions provided by the Yunnan Academy of Agricultural Sciences. The plant materials were grown at a subtropical plateau climate at Kaiyuan City, Yunnan Province (23°30′–23°58′ N, 103°04′–103°43′ E, altitude: 1051 m). Prior to 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 Mashtree[57].

HiFi sequencing. High quality genomic DNA of 14 *S. spontaneum* was extracted from fresh leaves using the DNAsecure Plant Kit (TIANGEN). A total of 28 Single Molecule Real Time (SMRT) bell libraries with ~20 kb DNA insert fragments were sequenced with the circular consensus sequencing (CCS) mode on a PacBio REVIO platform at ANNOROAD (Beijing, China). 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 employed a de Bruijn graph-based method PloidyFrost[58], 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 (CDBG) Construction: A CDBG is generated from single or multiple samples using the Bifrost API, ensuring efficient and accurate graph representation. (3) Graph analysis and variant extraction: Three sub-algorithms 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 using hifiasm

(https://github.com/chhylp123/hifiasm; version 0.19.8-r603)[59] 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 RepeatModeler[60] (v2.0.3) to create a custom repeat library. The repeat library was then combined with Viridiplantae repeats from RepBase[61] (v20150807) to generate the final repeat library. Repeat masking of each of *S. spontaneum* genome assembly was performed using the final CRL and RepeatMasker[62] (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)[63] to cut adapter and remove low quality reads and then aligned to the *S. spontaneum* genome assembly using HISAT2 (v2.2.1)[64]. The aligned RNA-seq reads were each assembled using StringTie (v2.2.1)[65]. Initial gene models were generated using BRAKER (v3.1)[66] 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 benchmarking universal single-copy orthologs (BUSCOs) (v5.4.1)[41] using 1,614 single-copy orthologs from the 'embryophyta_odb10' database. Assembly contiguity was assessed through the LTR Assembly Index (LAI)[40], calculated using LTR_retriever (v2.9.6)[67] with default parameters. Briefly, the completeness of long terminal repeat (LTR) retrotransposons was analyzed using LTR_FINDER_parallel (v1.2)[68] and LTR_HARVEST_parallel (v1.0)[69]. These tools were used to identify LTR elements for each genome, which were then compiled into a non-redundant LTR library using LTR_retriever (v2.9.6) for calculating the LAI. For accuracy assessment, the base-level quality value (QV) was computed using Merqury (v1.4.1)[70], based on k-mer analysis.

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

GWAS were performed using SNP dosage data, employing a linear mixed model to account for population structure and relative kinship. The analysis was carried out using the R package GWASpoly(v2.10)[12]. Initially, the covariance matrix for the polygenic effect was calculated using the set.K function. Following this, the GWASpoly function was utilized 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 , where $\frac{\alpha}{M}$ is 0.05 and M is the number of

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)[71] with default parameters. To ensure the accuracy and reliability of the graph structure, a combination of read mapping and assembled alignments was employed to validate the paths, edges, and nodes within the *S. spontaneum* pangenome.

Statistical analysis

Details on all statistical tests used in this study are indicated in the corresponding figure legends. Statistical analyses were performed using R software (v.4.3.1).

Acknowledgements

We thank Xiangbo Zhang (Guangdong Academy of Agricultural Sciences) for providing the SNP-based genotype files and the corresponding reference genome for the sweet potato. This work was funded by National Natural Science Foundation of China grant (No. 32222019 and No. 32370714 to X.Z.) and 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.).

Author Contributions

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

Data availability

14 PacBio HiFi sequencing data and 290 resequencing reads of *S. spontaneum* have been deposited in NCBI database (BioProject accession number: PRJNA1303125). Any additional re-sequencing data reported in

this paper data were downloaded from NCBI (BioProject accession numbers: Potato (PRJNA944441, PRJNA63145, PRJNA766763, and PRJNA378971), Sweet potato (PRJNA1089346), Alfalfa (PRJNA995892), and Rice (PRJNA623686).

Code availability

The KMERIA program and some analysis scripts related to this work are publicly available at (https://github.com/Sh1ne111/KMERIA, https://github.com/Sh1ne111/KMERIA/wiki).

Competing interests

A patent on the KMERIA method has been filed by Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences with S.C., S.Q. X.Z. and M.Z. as inventors. The remaining authors declare no competing interests.

References

1. Adams, K. L. & Wendel, J. F. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141 (2005).

2. Kuckuck, H., Kobabe, G. & Wenzel, G. *Fundamentals of Plant Breeding*. (Springer Berlin Heidelberg, 1991).

3. Van De Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).

4. 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).

5. Otto, S. P. The Evolutionary Consequences of Polyploidy. *Cell* **131**, 452–462 (2007).

6. Booker, W. W. & Schrider, D. R. The genetic consequences of range expansion and its influence on diploidization in polyploids. Preprint at https://doi.org/10.1101/2023.10.18.562992 (2023).

7. Orr-Weaver, T. L. When bigger is better: the role of polyploidy in organogenesis. *Trends Genet.* **31**, 307–315 (2015).

8. Trojak-Goluch, A., Kawka-Lipińska, M., Wielgusz, K. & Praczyk, M. Polyploidy in Industrial Crops: Applications and Perspectives in Plant Breeding. *Agronomy* **11**, 2574 (2021).

9. Van De Peer, Y., Ashman, T.-L., Soltis, P. S. & Soltis, D. E. Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* **33**, 11–26 (2021).

10. Zhang, X. *et al.* Refining polyploid breeding in sweet potato through allele dosage enhancement. *Nat. Plants* (2024) doi:10.1038/s41477-024-01873-y.

11. Li, H. *et al.* Genomic basis of divergence of modern cultivated potatoes. Preprint at https://doi.org/10.21203/rs.3.rs-3968149/v1 (2024).

12. 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* **9**, plantgenome2015.08.0073 (2016).

13. Jaegle, B. *et al.* k-mer-based GWAS in a wheat collection reveals novel and diverse sources of powdery mildew resistance. *Genome Biol.* **26**, (2025).

14. Gaurav, K. *et al.* Population genomic analysis of Aegilops tauschii identifies targets for bread wheat improvement. *Nat. Biotechnol.* **40**, 422–431 (2022).

15. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).

16. Voichek, Y. & Weigel, D. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nat. Genet.* **52**, 534–540 (2020).

17. Gaurav, K. *et al.* Population genomic analysis of Aegilops tauschii identifies targets for bread wheat improvement. *Nat. Biotechnol.* **40**, 422–431 (2022).

18. Jackson, P. A. Breeding for improved sugar content in sugarcane. *Field Crops Res.* **92**, 277–290 (2005).

19. Liu, X. *et al.* Phylogenetic Analysis of Different Ploidy Saccharum spontaneum Based on rDNA-ITS Sequences. *PloS One* **11**, e0151524 (2016).

20. Gerard, D., Ferrão, L. F. V., Garcia, A. A. F. & Stephens, M. Genotyping Polyploids from Messy Sequencing Data. *Genetics* **210**, 789–807 (2018).

21. 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).

22. Voorrips, R. E., Gort, G. & Vosman, B. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* **12**, (2011).

23. Cooke, D. P., Wedge, D. C. & Lunter, G. Benchmarking small-variant genotyping in polyploids. *Genome Res.* **32**, 403–408 (2022).

24. Leal, J. L. *et al.* Complex Polyploids: Origins, Genomic Composition, and Role of Introgressed Alleles. *Syst. Biol.* **73**, 392–418 (2024).

25. Swarbreck, D. *et al.* The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–D1014 (2007).

26. Zhang, F. *et al.* Evolutionary genomics of climatic adaptation and resilience to climate change in alfalfa. *Mol. Plant* **17**, 867–883 (2024).

27. Wei, X. *et al.* A quantitative genomics map of rice provides genetic insights and guides breeding. *Nat. Genet.* **53**, 243–253 (2021).

28. Healey, A. L. *et al.* The complex polyploid genome architecture of sugarcane. *Nature* **628**, 804–810 (2024).

29. 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).

30. Hong, E. P. & Park, J. W. Sample size and statistical power calculation in genetic association studies. *Genomics Inform.* **10**, 117–122 (2012).

31. 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).

32. 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).

33. Li, J., Xu, Y. & Chong, K. The novel functions of kinesin motor proteins in plants. *Protoplasma* **249 Suppl 2**, S95-100 (2012).

34. 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).

35. 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).

36. Yin, H. *et al.* SAUR15 Promotes Lateral and Adventitious Root Development via Activating H^+ -ATPases and Auxin Biosynthesis. *Plant Physiol.* **184**, 837–851 (2020).

37. Liu, W. *et al.* The Flavonoid Biosynthesis Network in Plants. *Int. J. Mol. Sci.* **22**, 12824 (2021).

38. Chen, M. *et al.* Genome-wide identification of agronomically important genes in outcrossing crops using OutcrossSeq. *Mol. Plant* **14**, 556–570 (2021).

39. Serra, L. & Perrot-Rechenmann, C. Spatiotemporal control of cell growth by CUC3 shapes leaf margins. *Development* **147**, dev183277 (2020).

40. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* (2018) doi:10.1093/nar/gky730.

41. 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).

42. Wang, Y. *et al.* Comparative genomics revealed the gene evolution and functional divergence of magnesium transporter families in Saccharum. *BMC Genomics* **20**, 83 (2019).

43. 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).

44. 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).

45. 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).

46. 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).

47. 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.* S2090123225001249 (2025) doi:10.1016/j.jare.2025.02.026.

48. 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).

49. 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).

50. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

51. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating *k* -mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).

52. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, (2020).

53. Gaynor, R. C., Gorjanc, G. & Hickey, J. M. AlphaSimR: an R package for breeding program simulations. *G3 GenesGenomesGenetics* **11**, jkaa017 (2021).

54. 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).

55. 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).

56. Li, M. *et al.* Enrichment of statistical power for genome-wide association studies. *BMC Biol.* **12**, 73 (2014).

57. Katz, L. *et al.* Mashtree: a rapid comparison of whole genome sequence files. *J. Open Source Softw.* **4**, 1762 (2019).

58. 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).

59. 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).

60. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).

61. Bao, Z. & Eddy, S. R. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Res.* **12**, 1269–1276 (2002).

62. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* **25**, (2009).

63. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

64. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, (2015).

65. 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**, (2019).

66. 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 Genomics Bioinforma.* **3**, lqaa108 (2021).

67. 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).

68. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 48 (2019).

69. David, Ellinghaus, Stefan, KurtzUte, & Willhoeft. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* (2008).

70. 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).

71. Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).

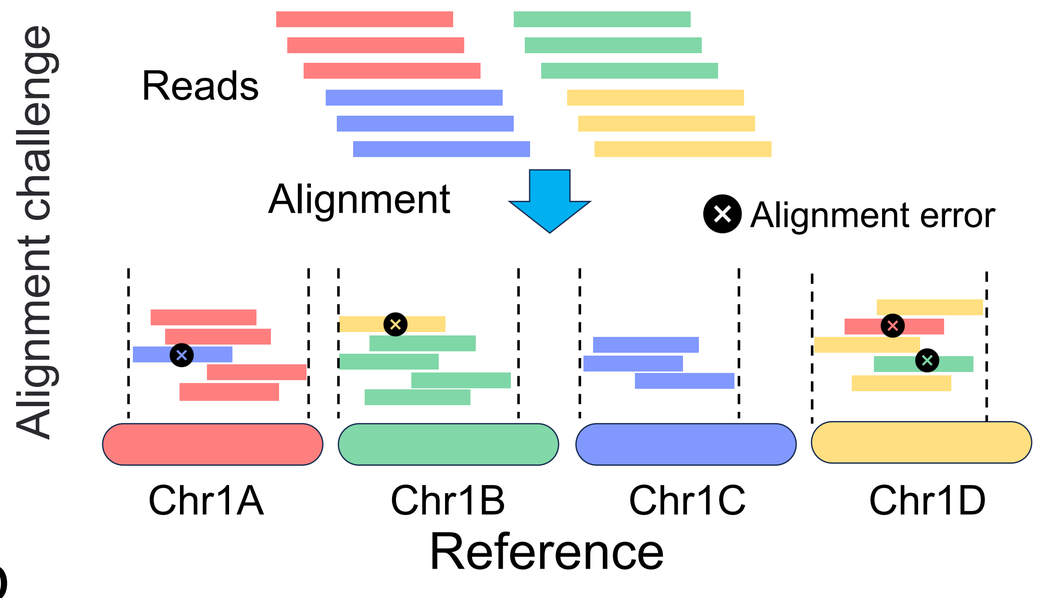
867 **Table 1. Summary of *S. spontaneum* genome assemblies and gene annotations**

Accessions	Accession name	Species	Estimated ploidy	Genome Length (bp)	N50 (bp)	Assemblies of BUSCO (%)	No. of Genes	Annotation of BUSCO (%)	LAI	QV
Ss01	fujian88-1-4	*Saccharum spontaneum*	10	8,368,781,140	2,484,927	99.4	296879	97.9	15.42	64.5482
Ss02	fujian89-1-7	*Saccharum spontaneum*	11	9,324,246,601	1,791,182	99.4	332535	98.3	15.49	65.1491
Ss03	guangdong8hao	*Saccharum spontaneum*	10	7,939,614,880	1,519,160	99.3	284471	98.7	14.53	64.5211
Ss04	guangxi86-8	*Saccharum spontaneum*	10	7,813,376,143	1,802,156	99.3	282072	98.4	14.47	64.2456
Ss05	guangxi87-51	*Saccharum spontaneum*	9	7,440,054,966	2,319,968	99.4	271315	98.6	15.86	65.914
Ss06	hainan92-5	*Saccharum spontaneum*	10	7,603,310,039	1,834,047	99.2	282165	98.2	14.9	64.7646
Ss07	sichuan79-1-19	*Saccharum spontaneum*	10	8,146,733,135	997,139	99.5	284006	98.3	15.38	62.2023
Ss08	sichuan88-22	*Saccharum spontaneum*	10	8,502,472,194	733,173	99.3	286578	98.1	15.18	64.1903
Ss09	yunnan82-149	*Saccharum spontaneum*	8	6,518,384,450	1,048,187	99.7	231354	98.8	15.66	62.9316
Ss10	yunnan82-132	*Saccharum spontaneum*	10	7,640,365,253	1,694,461	99.4	274973	98.8	16.43	63.1965
Ss11	yunnan82-48	*Saccharum spontaneum*	10	7,952,106,007	1,729,590	99.5	283350	98.9	15.54	62.9776
Ss12	yunnan82-84	*Saccharum spontaneum*	10	8,760,448,060	1,741,544	99.5	307442	98.6	15.77	65.2838
Ss13	yunnanheqinggeshoumi	*Saccharum spontaneum*	10	7,900,885,789	1,668,255	99.3	280316	99.1	15.05	65.0709
Ss14	yunnankaiyuangeshoumi	*Saccharum spontaneum*	10	7,763,673,639	1,388,620	99.5	283978	98.8	14.84	65.3216

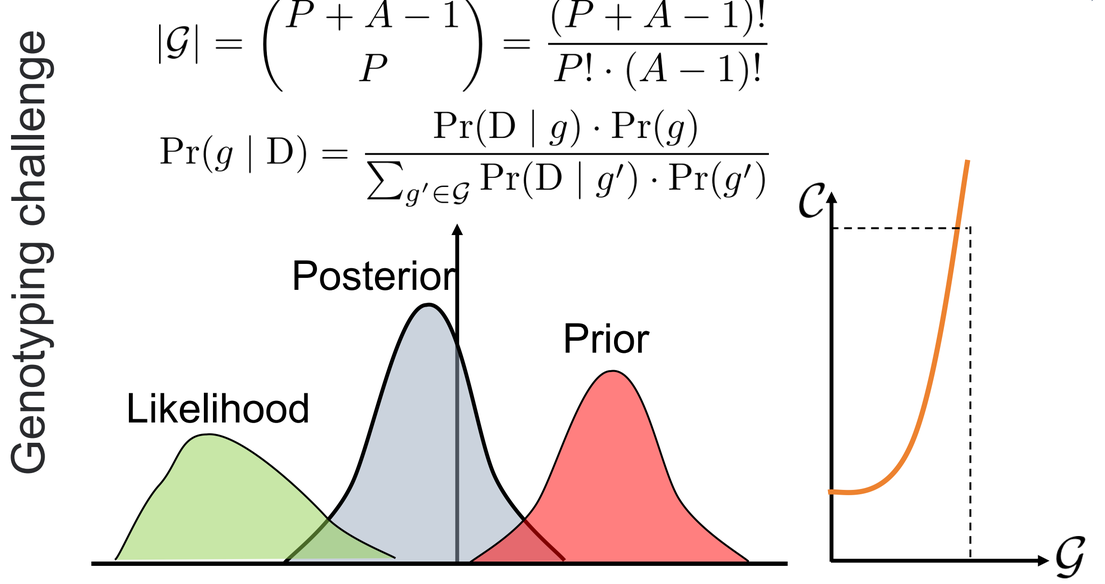
868

GATK bayesian genotyper

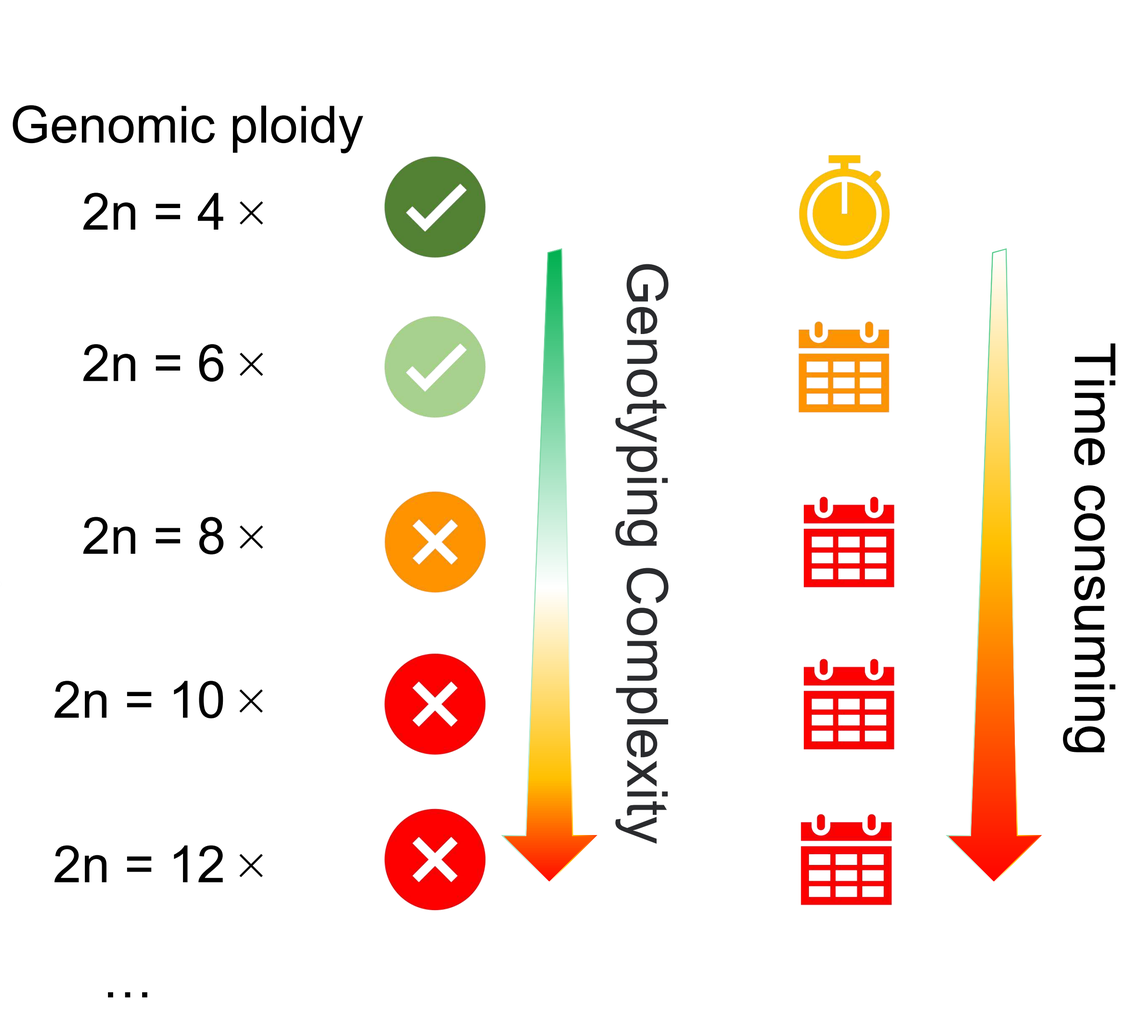
a



b

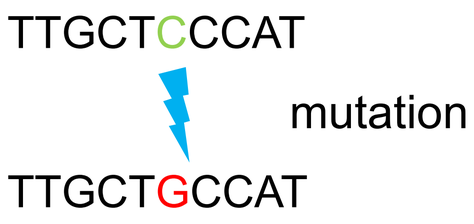


c



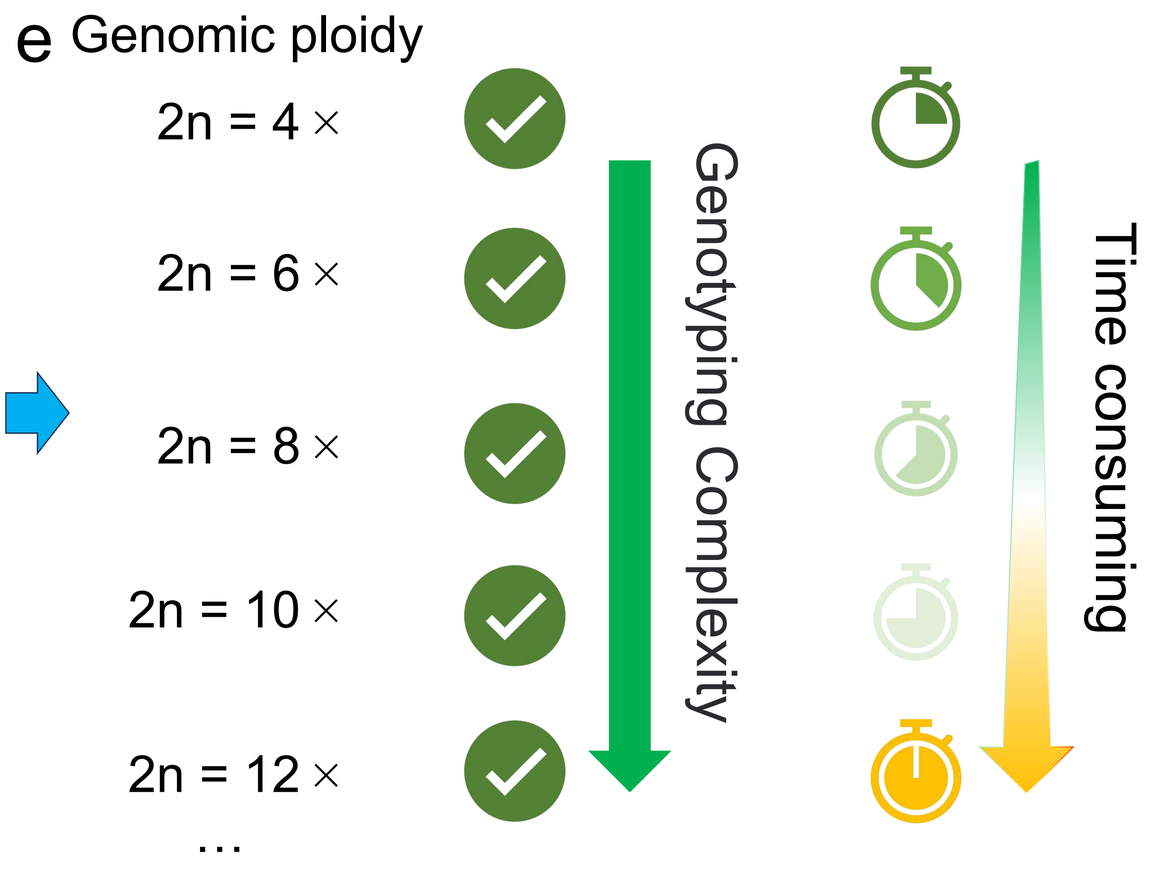
K-mer-based genotyper

d



2-mers	Count before mutation	Count after mutation
TT	1	1
TG	1	2
GC	1	2
CT	1	1
TC	1	0
CC	2	1
CA	1	1
AT	1	1

e



1 Counting

Sample N
Sample 3
Sample 2
Sample 1

k-mer counting

QC & trimming

AAAAAA 100
AAAAAA 100
AAAAAA 100
AAAAAA 100
AAAACG 14
……
ACCCCT 12

2 Matrixing

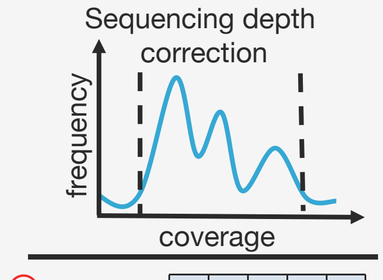
koc matrix building

individuals

k-mers

0	0	60	38	0	10	0
0	20	0	0	0	58	80
12	92	0	48	0	44	22
33	0	0	28	63	0	0
0	0	32	0	26	0	33
10	0	33	17	92	14	0
98	91	0	0	68	0	0

3 Filtering and Correcting



AAAAAAA	100	100	100	100	…
AAAAAAT	30	0	0	0	…
CCGTAAT	10	0	0	0	…
TTGTCAA	0	0	50	0	…

K-mer filtering at population level

Correction & Filtering

individuals

k-mers

0	0	6	4	0	1	0
0	2	0	0	0	6	8
1	9	0	5	0	4	2
3	0	0	3	6	0	0
0	0	3	0	3	0	3
1	0	3	2	9	1	0
10	9	0	0	7	0	0

4 K-mer testing

Allelic dosage recoding

individuals

k-mers

0	0	2	1.3	0	0.2	0
0	2	0	0	0	1.2	1.6
1	2	0	1	0	1	0.1
1	0	0	1	2	0	0
0	0	2	0	2	0	2
0.1	0	0.2	2	2	0.1	0
2	1.8	0	0	1.3	0	0

Chunks 1 … N

PCA

Kinship

Calculating Q & K from sampling k-mers

$$Y = K\text{-}mer + Q(PCs) + K + e$$

5 Post-GWAS

Genome 1
Genome 2
Genome 3
Genome N
Linear genomes

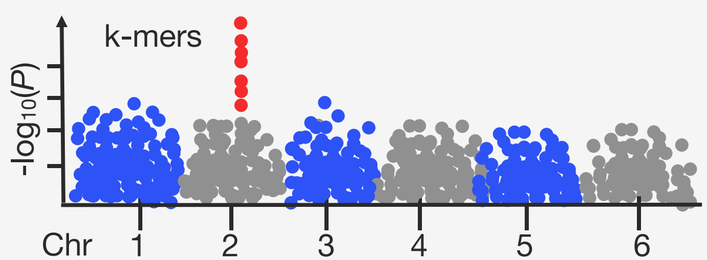
Graph pangenome

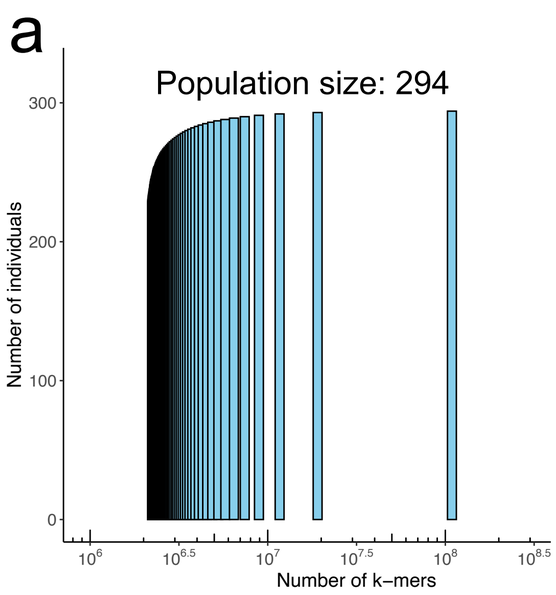
Association with Linear Mixed Model (LMM)

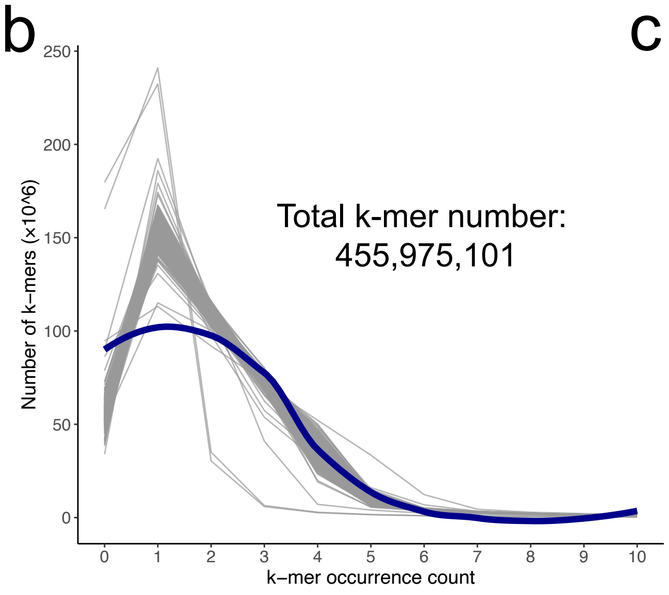
TTCATT
AATTGC
Associated k-mers

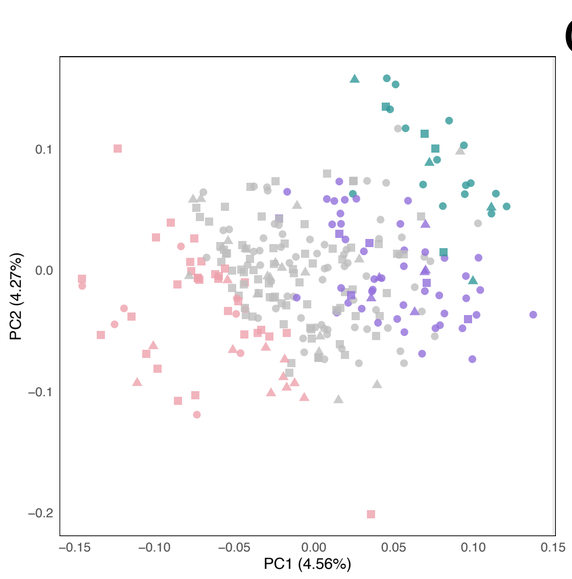
or

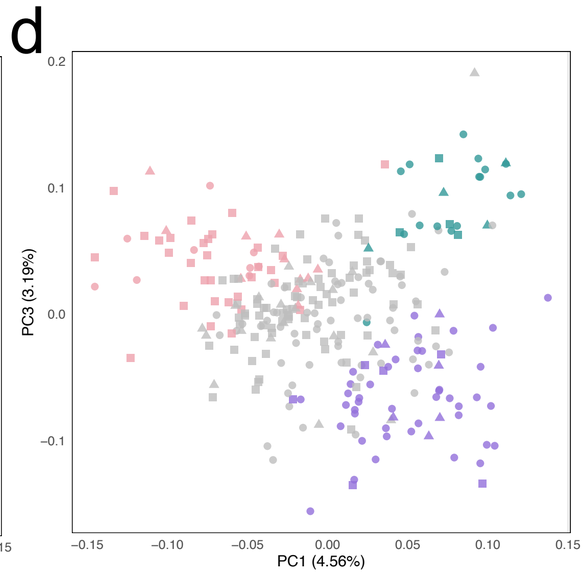
CGCGTT
TGGCAT
TTCATT
Associated k-mer reads

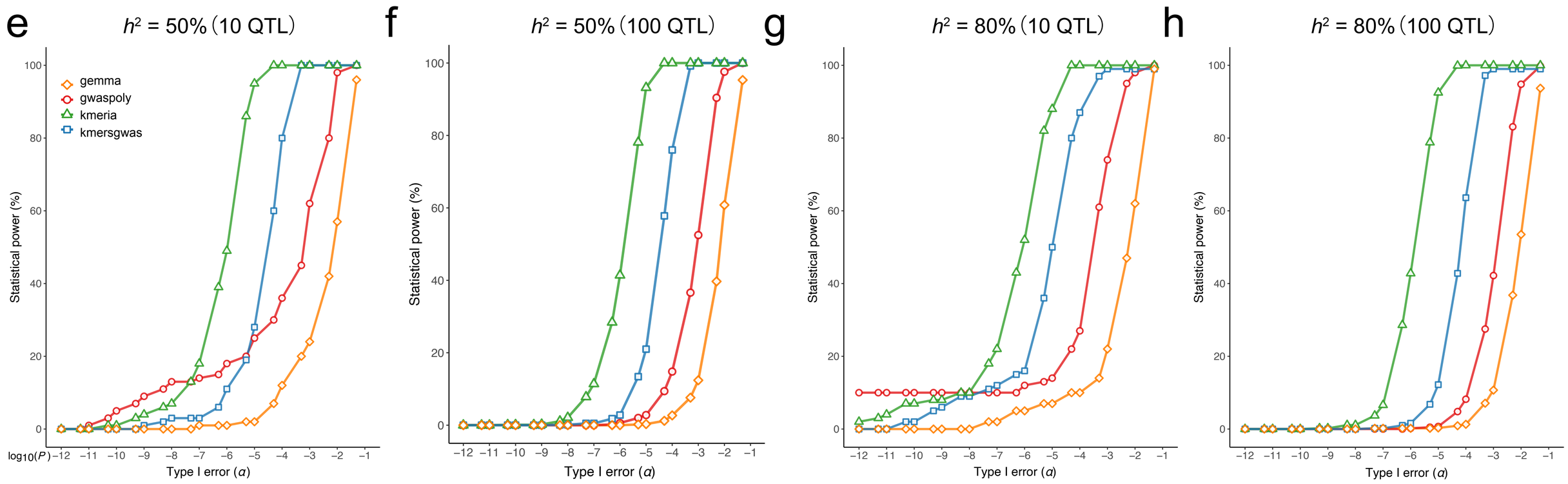


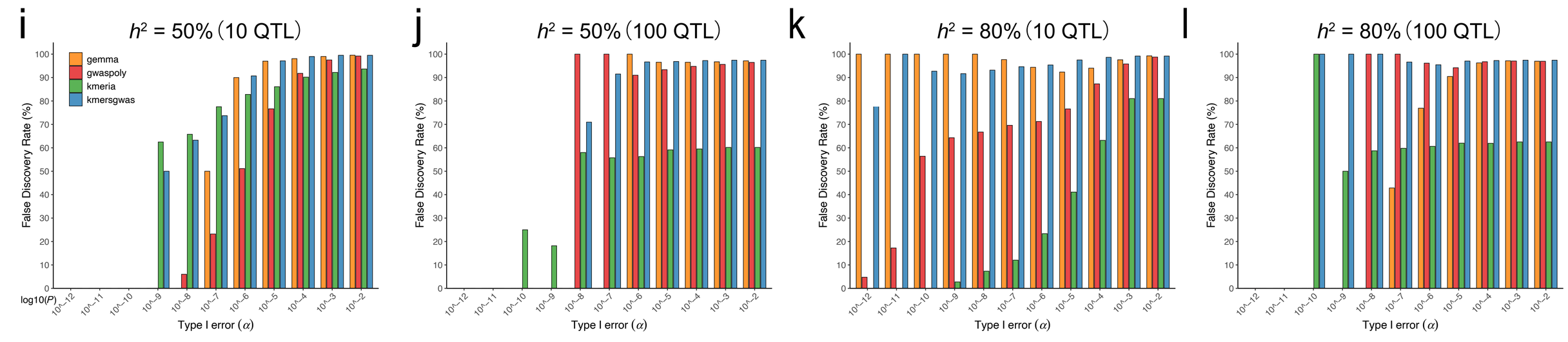

a
Population size: 294
Number of individuals
Number of k−mers

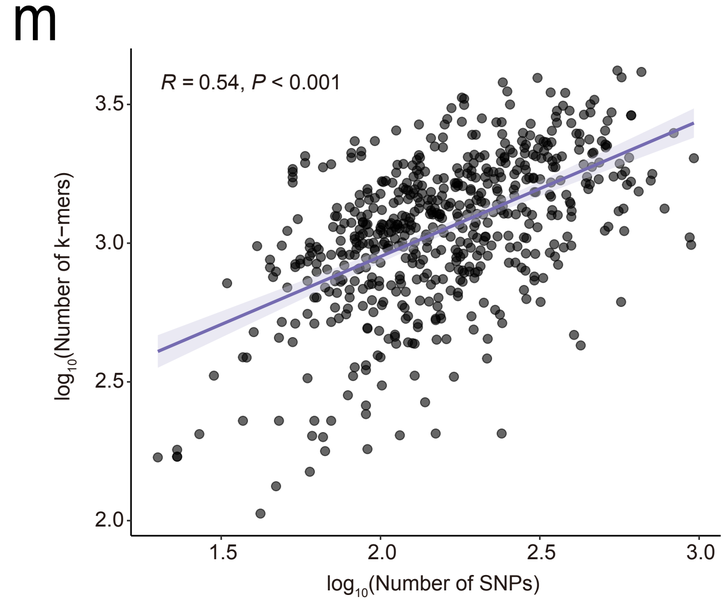

b
Total k-mer number:
455,975,101
Number of k−mers (×10^6)
k−mer occurrence count

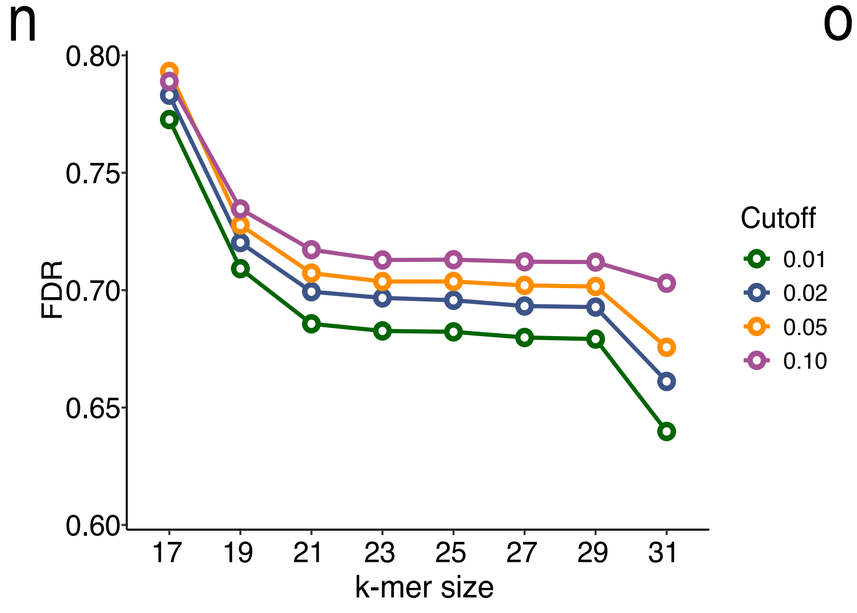

c
PC2 (4.27%)
PC1 (4.56%)

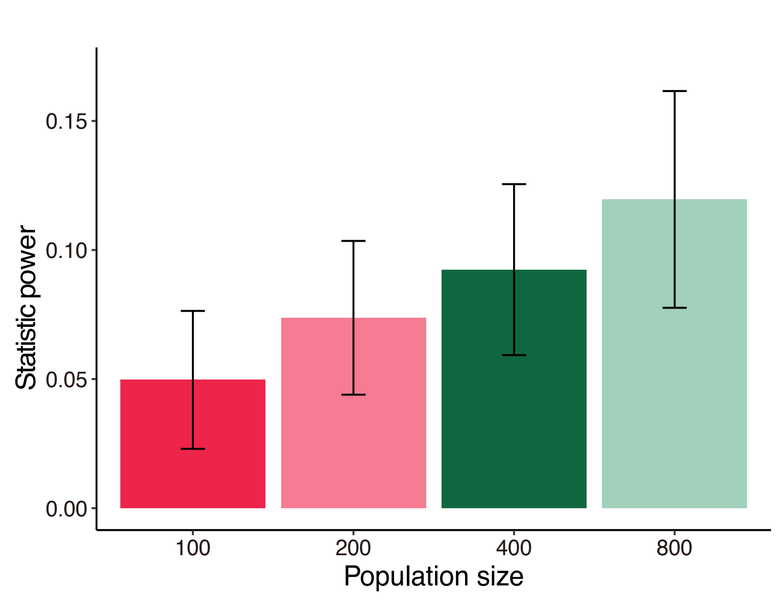

d
PC3 (3.19%)
PC1 (4.56%)
China cultivar
Introduced
Landrace
Colombia−Venezuela
Mesoamerica
Mix
Peru−Ecuador


e
h2 = 50% (10 QTL)
f
h2 = 50% (100 QTL)
g
h2 = 80% (10 QTL)
h
h2 = 80% (100 QTL)
gemma
gwaspoly
kmeria
kmersgwas
Statistical power (%)
log10(P)
Type I error (α)


i
h2 = 50% (10 QTL)
j
h2 = 50% (100 QTL)
k
h2 = 80% (10 QTL)
l
h2 = 80% (100 QTL)
gemma
gwaspoly
kmeria
kmersgwas
False Discovery Rate (%)
log10(P)
Type I error (α)

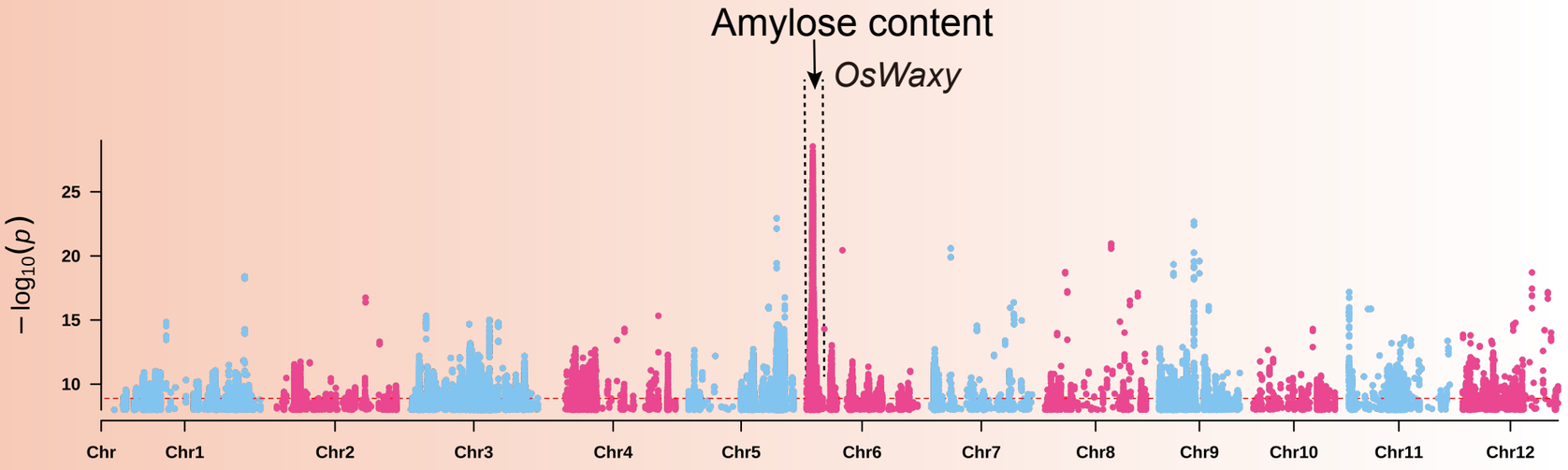

m
R = 0.54, P < 0.001
log10(Number of k−mers)
log10(Number of SNPs)

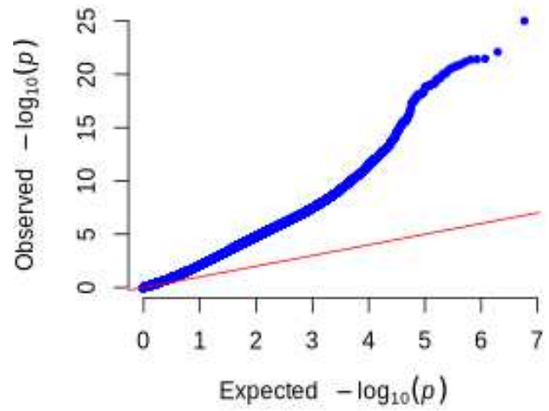

n
FDR
k-mer size
Cutoff
0.01
0.02
0.05
0.10


o
Statistic power
Population size

a

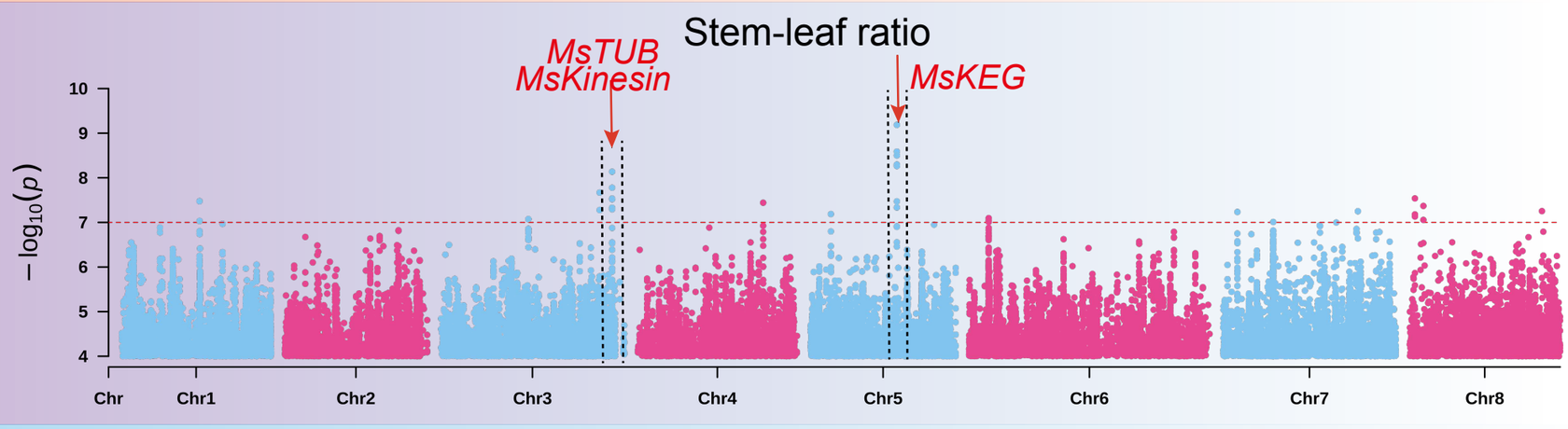
Rice
(2n = 2x)

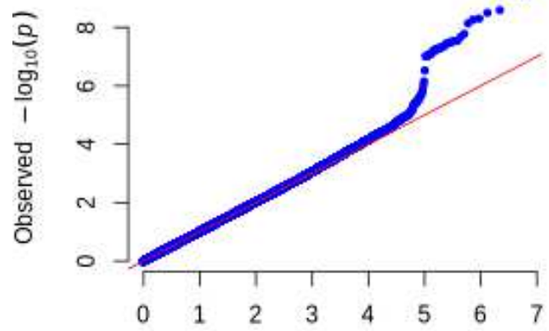




b

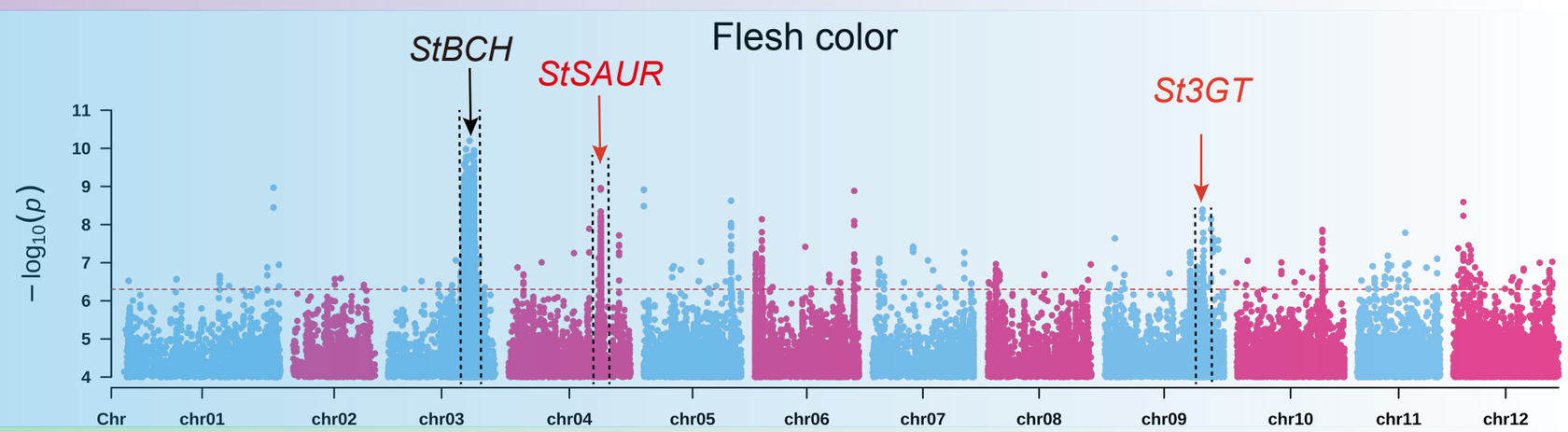
Alfalfa
(2n = 4×)

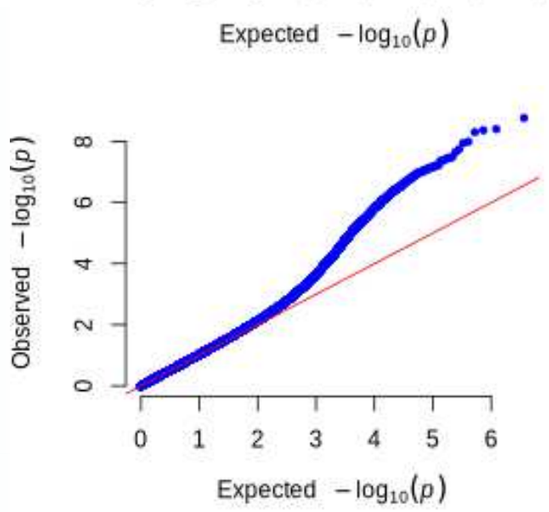




c

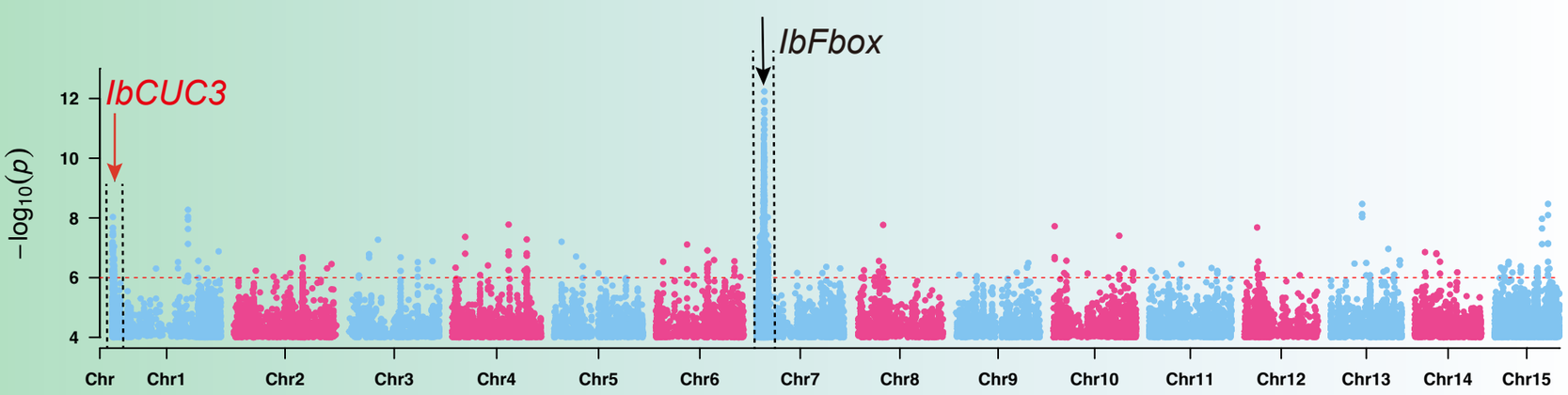
Potato
(2n = 4×)

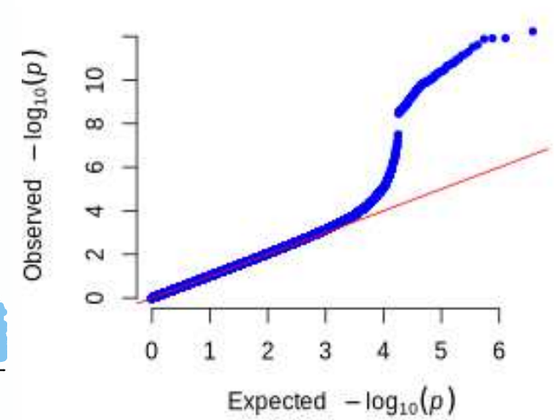


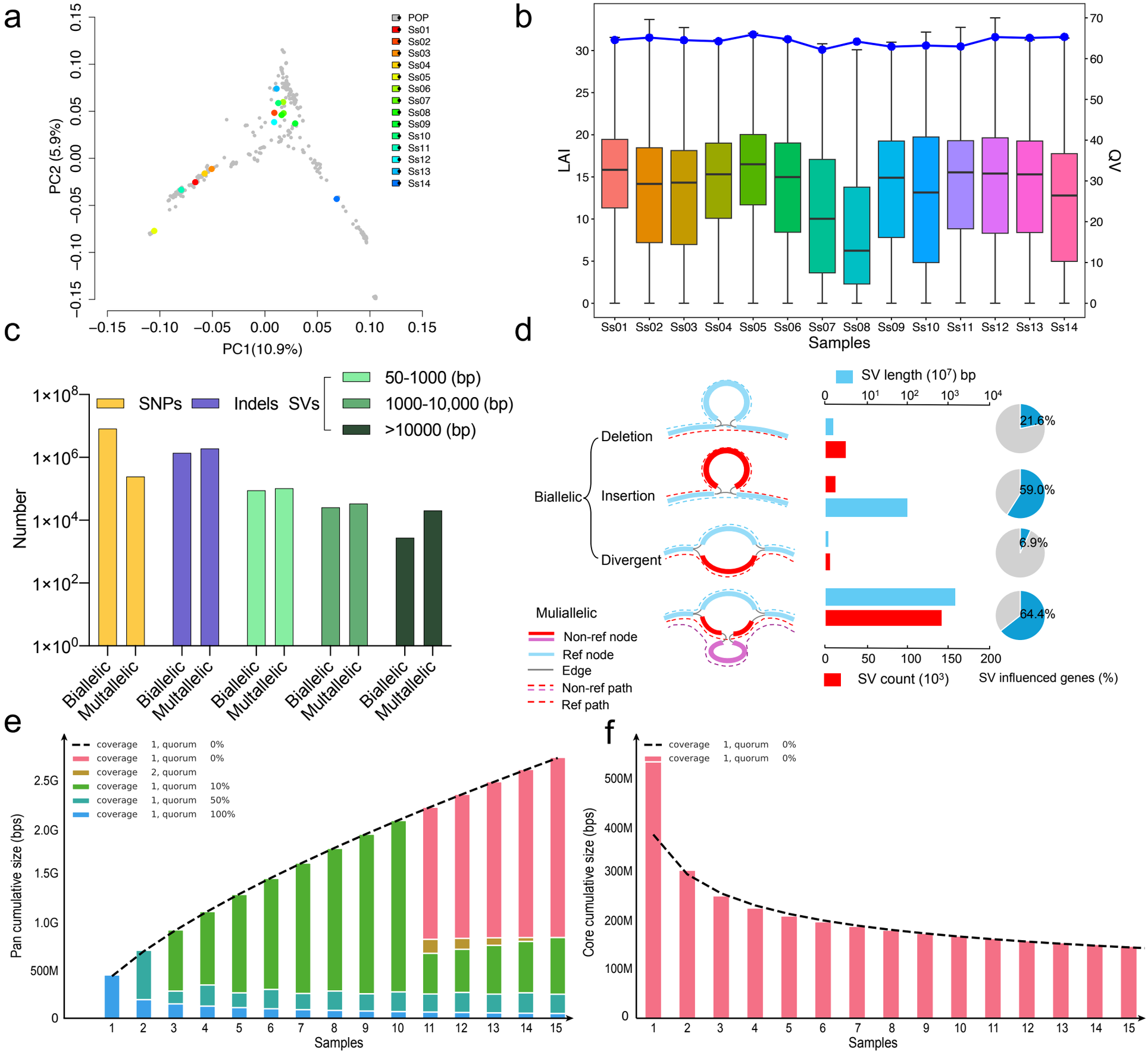


d

Sweetpotato
(2n = 6×)






a
POP
Ss01
Ss02
Ss03
Ss04
Ss05
Ss06
Ss07
Ss08
Ss09
Ss10
Ss11
Ss12
Ss13
Ss14
PC2 (5.9%)
PC1(10.9%)
b
LAI
QV
Samples
c
SNPs
Indels
SVs
50-1000 (bp)
1000-10,000 (bp)
>10000 (bp)
Number
Biallelic
Multallelic
d
Deletion
Insertion
Divergent
Biallelic
Muliallelic
Non-ref node
Ref node
Edge
Non-ref path
Ref path
SV length (10^7) bp
SV count (10^3)
SV influenced genes (%)
21.6%
59.0%
6.9%
64.4%
e
Pan cumulative size (bps)
Samples
f
Core cumulative size (bps)
Samples

a

Saccharum Spontaneum

Photosynthesis

CO_2

CO_2

O_2

O_2

Tillering

b

Apparent purity (AP)

Density

0.00

0.02

0.04

0 10 20 30 40 50 60

c

Tiller number (TN)

Density

0.00

0.04

0.08

0.12

0 5 10 15 20 25

d

Apparent purity (AP)

SsMGT

$-\log_{10}(p)$

4 5 6 7 8 9 10

Chr Chr1 Chr2 Chr3 Chr4 Chr5 Chr6 Chr7 Chr8

e

R = –0.3, *P* = 2e–07

N = 284

Apparent purity (%)

0 10 20 30 40 50 60 70

1 2 3 4 5 6 7 8 9 10 11 12 13 14

Kmer dosage

f

8.15 8.25 8.35 8.45 (Mbp) Chr06

Hap1

SsMGT (*Ss82114mono.06G0005950*)

SsMGT (*Ss05.g8893*)

Hap2

Indel_6_8306558 (7 bp)

g

Tiller number (TN)

SsNGA5

SsERF14

SsNAC

SsLOG

SsARF8

SsSCR

$-\log_{10}(p)$

4 5 6 7 8 9

Chr Chr1 Chr2 Chr3 Chr4 Chr5 Chr6 Chr7 Chr8

h

R = 0.38, *P* = 4.6e–11

N = 286

Tiller number

0 5 10 15 20 25 30

2 3 4 5 6 7 8 9 10 11

Kmer dosage

i

SsNGA5 (Ss82114mono.01G0027240)

SsERF14 (Ss82114mono.01G0013880)

SsNAC (Ss82114mono.03G0016070)

SsARF8 (Ss82114mono.04G0016010)

SsLOG (Ss82114mono.04G0016050)

SsSCR (Ss82114mono.07G0013360)

Leaf Stem Root Tiller-1 Tiller-2 Tiller-3

$Log_2(TPM+1)$

0 2 4

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 2.SupplementaryNotesTablesandFigures.pdf