CellPress
Partner Journal

Molecular Plant
Research article

A *k*-mer-based pangenome approach for cataloging seed-storage-protein genes in wheat to facilitate genotype-to-phenotype prediction and improvement of end-use quality

Zhaoheng Zhang[1,3], Dan Liu[1,3], Binyong Li[1], Wenxi Wang[1], Jize Zhang[1], Mingming Xin[1], Zhaorong Hu[1], Jie Liu[1], Jinkun Du[1], Huiru Peng[1], Chenyang Hao[2], Xueyong Zhang[2], Zhongfu Ni[1], Qixin Sun[1], Weilong Guo[1,*] and Yingyin Yao[1,*]

[1]Frontiers Science Center for Molecular Design Breeding, Key Laboratory of Crop Heterosis and Utilization (MOE), and Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing 100193, China

[2]Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing, China

[3]These authors contributed equally to this article.

***Correspondence: Weilong Guo (guoweilong@cau.edu.cn), Yingyin Yao (yingyin@cau.edu.cn)**

https://doi.org/10.1016/j.molp.2024.05.006

ABSTRACT

Wheat is a staple food for more than 35% of the world's population, with wheat flour used to make hundreds of baked goods. Superior end-use quality is a major breeding target; however, improving it is especially time-consuming and expensive. Furthermore, genes encoding seed-storage proteins (SSPs) form multigene families and are repetitive, with gaps commonplace in several genome assemblies. To overcome these barriers and efficiently identify superior wheat SSP alleles, we developed "PanSK" (Pan-SSP *k*-mer) for genotype-to-phenotype prediction based on an SSP-based pangenome resource. PanSK uses 29-mer sequences that represent each SSP gene at the pangenomic level to reveal untapped diversity across landraces and modern cultivars. Genome-wide association studies with *k*-mers identified 23 SSP genes associated with end-use quality that represent novel targets for improvement. We evaluated the effect of rye secalin genes on end-use quality and found that removal of ω-secalins from 1BL/1RS wheat translocation lines is associated with enhanced end-use quality. Finally, using machine-learning-based prediction inspired by PanSK, we predicted the quality phenotypes with high accuracy from genotypes alone. This study provides an effective approach for genome design based on SSP genes, enabling the breeding of wheat varieties with superior processing capabilities and improved end-use quality.

Key Words: wheat, seed-storage protein, end-use quality, *k*-mer, pangenome, genomic prediction

Zhang Z., Liu D., Li B., Wang W., Zhang J., Xin M., Hu Z., Liu J., Du J., Peng H., Hao C., Zhang X., Ni Z., Sun Q., Guo W., and Yao Y. (2024). A *k*-mer-based pangenome approach for cataloging seed-storage-protein genes in wheat to facilitate genotype-to-phenotype prediction and improvement of end-use quality. Mol. Plant. **17**, 1038–1053.

INTRODUCTION

Wheat (*Triticum aestivum*) is the third most highly produced staple crop worldwide, and its flour represents a valuable source of carbohydrates, proteins, dietary fibers, and micronutrients (Appels et al., 2018; Asseng et al., 2020). The global demand for wheat continues to increase to meet the needs of 9.8 billion people by 2050, and high end-use quality of wheat cultivars is a prerequisite for the flour used to make hundreds of unique baked goods (Tilman et al., 2011). End-use quality is closely linked to the taste, texture, appearance, and shelf life of wheat-flour-based foods, and it ultimately determines overall consumer satisfaction. Superior end-use quality therefore remains a major priority for wheat breeders.

Improvement of wheat end-use quality is a considerable challenge, and achievements to date have failed to meet consumer demand (Subedi et al., 2022, 2023) because quality-related traits are

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

quantitative with low heritability and strong genotype–environment interactions. Furthermore, direct measurement of quality traits is time consuming and requires a large quantity of grain, expensive trait-specific instruments, and an expert workforce (Mann et al., 2009; Subedi et al., 2022). Thus, assessment of elite breeding lines is typically delayed until advanced generations, and lines with poor end-use quality are frequently discarded. This situation has favored progress in improving other traits, such as grain yield and disease resistance (Naraghi et al., 2019). Therefore, an approach for highly accurate and efficient genotype-to-phenotype prediction of end-use quality in earlier breeding generations is needed. Such technology could enhance the accuracy of selection and hasten wheat quality improvement.

Marker-assisted selection is a convenient, efficient, and fast way to predict end-use-quality phenotypes in wheat breeding. Some functional markers for genes have been used in early breeding programs, including *Pinb-D1a* and *Pinb-D1a* for grain texture/softness, *Gpc-B1* for protein content, and *Glu-A1* (Ax2, Ax1, and AxNull subunits), *Glu-B1* (Bx7^{OE}), and *Glu-D1* (Dx5+Dy10, Dx2+Dy12) encoding high-molecular-weight glutenins (Subedi et al., 2022). However, there are still few markers with high prediction accuracy that are tightly linked to end-use quality in wheat (Sandhu et al., 2022). To overcome this issue, genomic selection can estimate breeding values using whole-genome-wide markers to facilitate early-generation selection for end-use quality (Battenfield et al., 2016). Michel et al. (2018) used genomic selection to predict baking quality and reported an acceptable prediction accuracy of 0.38–0.63 (Michel et al., 2018). A number of genomic-selection studies for end-use quality traits in wheat have been performed over the last decade, increasing breeders' confidence in this strategy (Plavšin et al., 2021).

Seed-storage proteins (SSPs) in wheat grains, especially high-molecular-weight glutenins (HMW-GS), low-molecular-weight glutenins (LMW-GS), gliadins, and avenin-like proteins (ALPs), play a crucial role in determining end-use quality because they are responsible for dough elasticity and extensibility, essential processing qualities for the production of a wide range of end products (Rasheed et al., 2014; Luo et al., 2021). HMW-GS is encoded by three loci, *Glu-A1*, *Glu-B1*, and *Glu-D1*, which contain two tightly linked genes for "x" and "y" type glutenin subunits at each locus. "Dx5+Dy10," "Ax1," and "Bx7+By8" have positive effects on dough properties, resulting in good bread-making quality (Payne and Lawrence, 1983; Shewry, 2003; Ravel et al., 2006, 2020; Pirozi et al., 2008).

Markers for LMW-GS genes and one *Gli-γ1-I* marker for the elite *γ-gliadin* allele are used in wheat breeding (Ikeda et al., 2006; Wang et al., 2010; Liu et al., 2023). Although DNA-based markers such as SNPs (single-nucleotide polymorphisms) and KASP (Kompetitive allele-specific PCR) markers can distinguish different polymorphism locations (Ravel et al., 2020), it remains challenging to develop gene-specific markers for individual SSP genes because they comprise multi-gene families and contain repetitive domains. In addition, gaps in reference genomes on SSP-encoding loci are widespread and limit the identification of variation. There is thus a need for an SSP-based pangenome resource that fully captures the bulk of SSP genetic diversity in wheat populations. This would enable the identification of superior alleles to improve genotype-to-phenotype prediction and, ultimately, end-use quality.

k-mers are short sequences with a fixed length, *k*, and can be obtained from sequencing data or genome assemblies. They can mark a broad range of polymorphisms independently of a reference genome (Voichek and Weigel, 2020). For example, *k*-mers can be extracted from all sequence reads, and *k*-mer sets from different samples can be compared. Importantly, *k*-mers that are present in some samples but missing from others can assist in identifying a broad range of genetic variants. *k*-mer-based genome-wide association studies (GWASs) have been widely used to identify genetic variants underlying phenotypic variation in plants (Karikari et al., 2023). For example, *k*-mer-based association mapping has been used to identify candidate genes for disease and pest resistance from *Aegilops tauschii*, the diploid wild progenitor of the D subgenome of bread wheat (Gaurav et al., 2022).

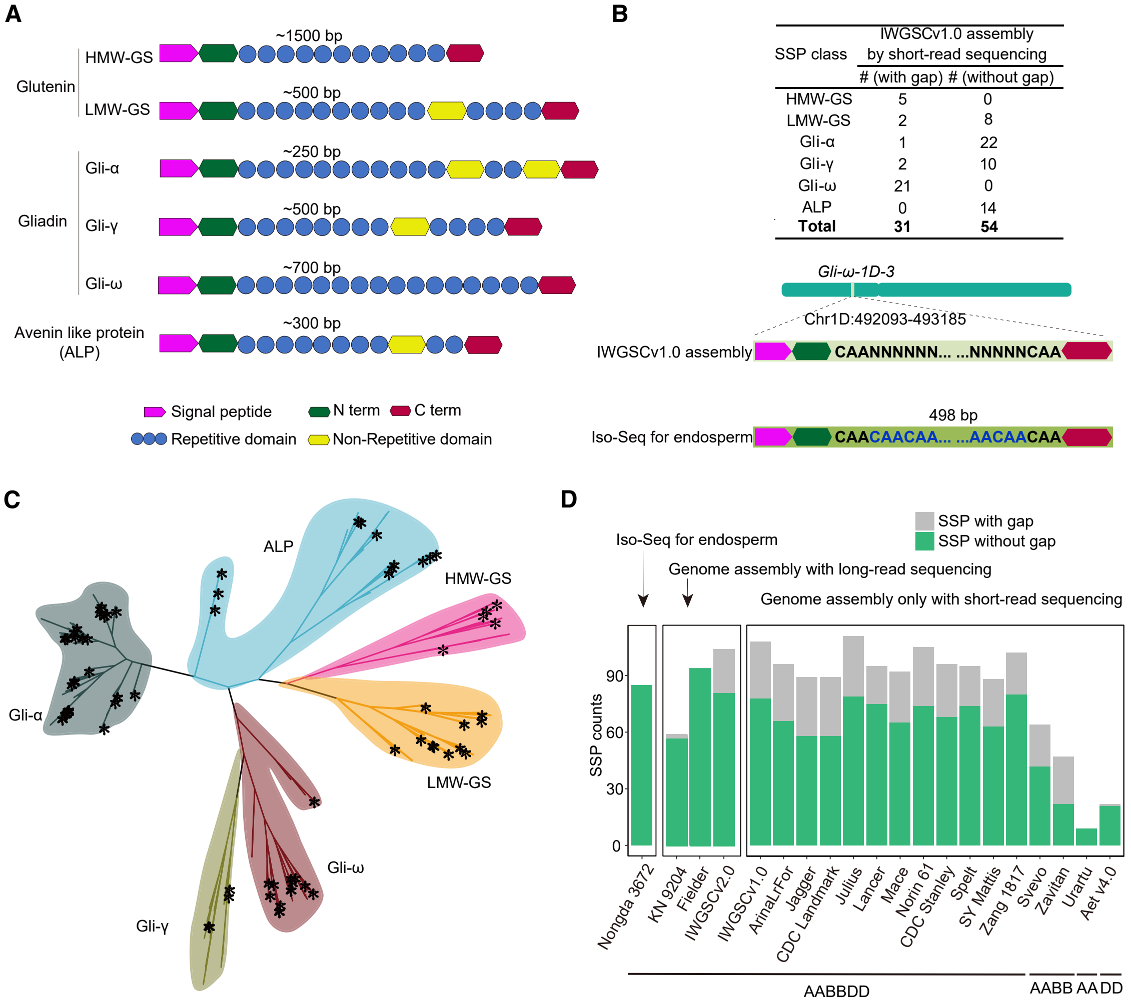
In this study, we cataloged a wide range of genetic variation in wheat SSP genes on the basis of *k*-mer analysis and identified a set of SSP genes associated with variation in end-use quality. We identified unique 29-mer sequences representative of genes encoding wheat SSPs and developed a new analysis workflow, PanSK, which we used to: (1) construct a comprehensive presence/absence-variation map of wheat SSP genes at the pangenome level; (2) perform GWAS analysis to identify candidate SSP loci associated with end-use quality; and (3) exploit a machine-learning approach based on PanSK to predict end-use quality phenotypes with high accuracy. The results demonstrate the power of PanSK to uncover novel targets for genotype-to-phenotype prediction and more quickly advance and rejuvenate interest in the improvement of wheat end-use quality.

RESULTS

Identification of gapless SSP genes by long-read RNA sequencing

Wheat SSPs are encoded by multi-gene families that show high sequence similarity and contain long repetitive-sequence elements (Figure 1A), making it difficult to produce a complete and errorless assembly of SSP genes through short-read sequencing. For example, the Chinese Spring genome assembly IWGSCv1.0 predicts 85 annotated SSP genes, 31 of which contain gaps (Figure 1B and Supplemental Table 1). Such gaps present a barrier to comprehensive functional analysis and the study of wheat end-use quality.

To overcome this barrier and obtain full-length and gapless transcripts of SSP genes, we performed full-length isoform sequencing (Iso-Seq) with PacBio long-read RNA-sequencing technology using RNA extracted from the endosperm of wheat cv. Nongda 3672 at 15 days after pollination (DAP). The average transcript length reached 3 kb, sufficient to cover the entire open reading frames of SSP genes (Figure 1A). We obtained 85 unique full-length transcripts encoding SSPs, including 5 HMW-GS genes, 11 LMW-GS genes, 57 gliadin genes, and 12 ALP genes (Figure 1C and Supplemental Table 2). In genomes assembled by short-read sequencing, such as ArinaLrFor, Jagger, Mace, and Norin 61, most HMW-GS sequences contain gaps (Supplemental Figure 1). However, these gaps can be successfully bridged with Iso-Seq data (Supplemental Figure 1 and Supplemental Table 3). The



SSP class	IWGSCv1.0 assembly by short-read sequencing	
	# (with gap)	# (without gap)
HMW-GS	5	0
LMW-GS	2	8
Gli-α	1	22
Gli-γ	2	10
Gli-ω	21	0
ALP	0	14
Total	**31**	**54**

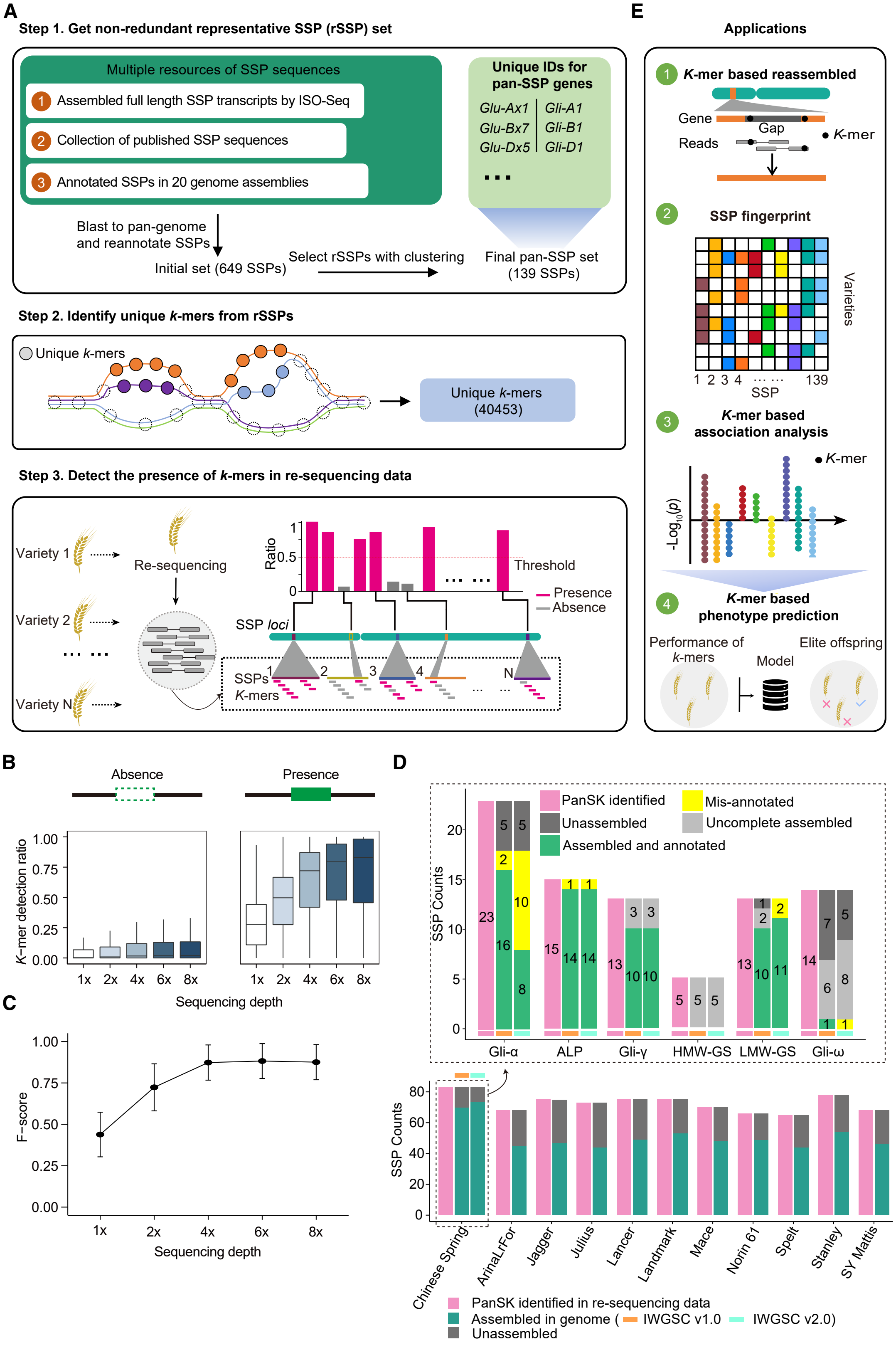
Figure 1. Isoform sequencing (Iso-Seq) fills gaps in transcripts encoding seed-storage proteins.
(A) Schematic of major wheat SSPs and their respective domain structures.
(B) Top: summary of incomplete SSP gene annotations in the cv. Chinese Spring genome version 1 (IWGSCv1.0) assembly obtained by short-read sequencing. Bottom: a partial *Gli-ω-1D-3* sequence, in which "NN" represents gaps in IWGSCv1.0 that were filled by Iso-Seq.
(C) Neighbor-joining tree of annotated SSPs in wheat and rye constructed from 649 sequences. Branch length represents $\log_2$ genetic distance determined by ClustalW. Full-length transcripts identified in Nongda 3672 are marked with asterisks.
(D) SMRT-seq for RNA sequencing (cv. Nogndа3672) can fill in all gaps in SSP transcripts arising from sequencing strategies. Each bar represents a wheat variety used for scanning HMW-GS genes, including hexaploid wheat (AABBDD), tetraploid wheat (AABB), diploid *Triticum urartu* (AA), diploid *Aegilops tauschii* (DD), and diploid *Secale cereale* (RR).

ω-gliadin gene TraesCS1D02G002063, which contains one gap in the IWGSCv1.0 assembly (International Wheat Genome Sequencing Consortium, 2018), was filled here by Iso-Seq data with a length of 498 bp (Figure 1B). Intriguingly, benchmarking confirmed that the Iso-Seq assembly contained more full-length SSP genes than published wheat reference genomes assembled using short-read sequencing, such as cv. ArinaLrFor, Jagger, and Julius, and was comparable to genomes assembled with long-read sequencing, such as cv. Fielder and CS-IWGSCv2.0 (Figure 1D). Thus, integration of transcripts assembled by long-read RNA sequencing has the potential to provide a more accurate and comprehensive catalog of wheat SSPs.

Nevertheless, Iso-Seq remains costly and currently lacks the sensitivity to assemble transcripts with low expression (Figure 1D), which limits its present utility for investigating copy-number variation and sequence polymorphisms at the population level. By integrating full-length SSP sequences obtained by Iso-Seq with sequences from the National Center for Biotechnology Information (NCBI) (Supplemental Table 4), we developed a *k*-mer-based method to establish an SSP-based pangenome in wheat. Such a strategy is effective for discovering *R* genes without mapping (Arora et al., 2019).

Establishment and evaluation of the *k*-mer-based pipeline PanSK

We aimed to identify the *k*-mers unique to each SSP gene, thus enabling fast and accurate determination of genetic variation, including presence/absence variations and nucleotide polymorphisms, by direct scanning of sequencing reads. We developed a *k*-mer-based pipeline, which we named PanSK, for detecting



(legend on next page)

wheat SSPs at the pangenomic level by scanning raw sequencing reads. The PanSK workflow has three main steps (Figure 2A).

In step 1, we collected 649 SSP gene sequences at the pangenomic level from multiple resources: 515 annotated genes from 20 Triticeae genome assemblies, 125 genes derived from Iso-Seq data collected from cv. Nongda 3672 and the cultivated variety Xiaoyan 81 (Wang et al., 2017), and nine published SSP sequences obtained via Sanger sequencing (Supplemental Table 5). All genes with a sequence similarity greater than 99% were grouped together, and the longest sequence from each set was chosen as the non-redundant representative SSP (rSSP) (Supplemental Table 4), for a total of 139 rSSP sets.

In step 2, we performed multiple tests with different *k*-mer sizes (*k* = 17, 19, 21, 23, 25, 27, 29, 31, 33) to determine the optimal *k-mer* size. We generated *k*-mers from each SSP gene to identify unique hits across the entire genome. A value of *k* = 29 was sufficient to identify unique *k*-mers for all 139 SSP genes (Supplemental Figure 2A–2C). Further increases in *k*-mer size did not lead to significant improvements in SSP gene identification but did significantly increase computational costs. Thus, we selected *k* = 29 for *k*-mer size. Finally, a total of 40 453 unique 29-mers uniquely representing all the 139 rSSPs were identified.

In step 3, PanSK was developed to scan raw resequencing reads and infer SSP variants as represented by unique *k*-mers. The presence/absence of SSP genes was estimated by the relative ratio of scanned unique *k*-mers representing each SSP gene. PanSK thus detects presence/absence variation in wheat SSP genes without mapping of reads to a reference genome.

To evaluate the performance of PanSK on calling SSP presence/absence variations, we scanned SSP-specific *k*-mers in resequencing data from 11 wheat varieties. These sources were either randomly sampled sequence data or simulated short-read sequence data from assembled genome sequences, including those of Chinese Spring, ArinaLrFor, Jagger, Lancer, Landmark, Mace, Norin 61, Spelt, Stanley, and SY Mattis. By comparison with the annotated SSPs in these assemblies, the simulated data demonstrate that the *k*-mer detection rate for SSP genes increases with sequencing depth. The distributions of the *k*-mer detection rate for absent and present SSP genes were distinguishable even at a read depth of 1× (Figure 2B), indicating that the detection rate of *k*-mers can serve as a reliable indicator for inferring the presence/absence of these genes. Furthermore, we evaluated PanSK detection power at different sequencing depths and showed that a sequencing depth of 4× is adequate for accurately determining the presence or absence of SSP genes and that further increases in sequencing depth do not significantly improve the *F* score (Figure 2C).

We next evaluated the performance of PanSK in gene assembly from re-sequencing data by comparing its results with assembled genes from the 11 genomes. PanSK identified more SSPs directly from resequencing data compared with genomes assembled from either short reads or long reads (Figure 2D and Supplemental Table 6). For example, PanSK identified more SSP genes using resequencing data from cv. Chinese Spring compared with the number assembled in both the IWGSCv2.0 (long-read) and IWGSCv1.0 (short-read) assemblies. Specifically, 23 predicted α-gliadin genes were identified by PanSK from resequencing data, 16 of which were annotated in IWGSCv1.0 and eight of which were annotated in IWGSCv2.0. For genes encoding ω-gliadins, 14 *Gli-ω* genes were predicted by PanSK, but these loci were poorly assembled in both IWGSCv1.0 and v2.0. In addition, PanSK was highly effective in addressing the issue of misassembled genes. For instance, TraesCS1B02G329711 in IWGSCv1.0 and TraesCS1B03G0904700 in IWGSCv2.0 were annotated as encoding HMW glutenin 1Bx, and both sequences were misassembled relative to the PanSK-assembled sequence, with a gap between 664 bp and 2028 bp (Supplemental Figure 2D). Some SSP genes are partially assembled in both IWGSCv1.0 and v2.0, and the unassembled regions represented by NNNs can be filled using PanSK (Supplemental Figure 2E). These results demonstrate the power and practicality of PanSK for exploration of polymorphisms in SSP genes across wheat varieties.

PanSK is not only capable of identifying SSP genes using resequencing data; the *k*-mer-based strategy can also be used for additional purposes, including (1) reassembly of novel SSP alleles by overlapping corresponding *k*-mers, (2) construction of a presence/absence-variation fingerprint map of SSP genes in large populations, (3) association analysis and discovery of functional SSP alleles associated with wheat end-use quality, and (4) assistance with phenotype prediction (Figure 2E). Therefore, the *k*-mer-based pipeline used by PanSK to identify SSP genes has advantages over traditional assembly methods and could serve as a scalable approach to evaluate wheat SSP genes in large populations.

Fingerprint mapping of SSP genes using PanSK

Wheat SSP composition is commonly characterized by reverse-phase high-performance liquid chromatography or SDS–PAGE,

Figure 2. The *k*-mer-based algorithm PanSK identifies presence/absence variations in SSPs with high sensitivity and without mapping.

(A) Overview of the PanSK pipeline for pan-SSP map construction and its applications. Step 1: SSPs were collected from multiple sources at a pan-genome scale, and 139 non-redundant representative SSPs (rSSPs) were identified and assigned unique IDs. Step 2: *k*-mers (*k* = 29) uniquely representing rSSPs were identified. Step 3: presence/absence variations in SSP genes were inferred from the *k*-mer detection ratio in resequencing reads without read mapping.

(B) Detection ratio of *k*-mers used to infer presence/absence variations in SSP genes from resequencing data. Simulated resequencing data with different coverages from 1× to 8× in two scenarios (SSP absent or SSP present) were used for evaluation.

(C) *F* scores for inference of SSP presence/absence variation from resequencing data at different depths. Eighteen assembled accessions were used for the simulation.

(D) PanSK with resequencing data identifies more SSPs compared with genome assemblies obtained from long reads and short reads. SSP genes identified in cv. Chinese Spring are boxed.

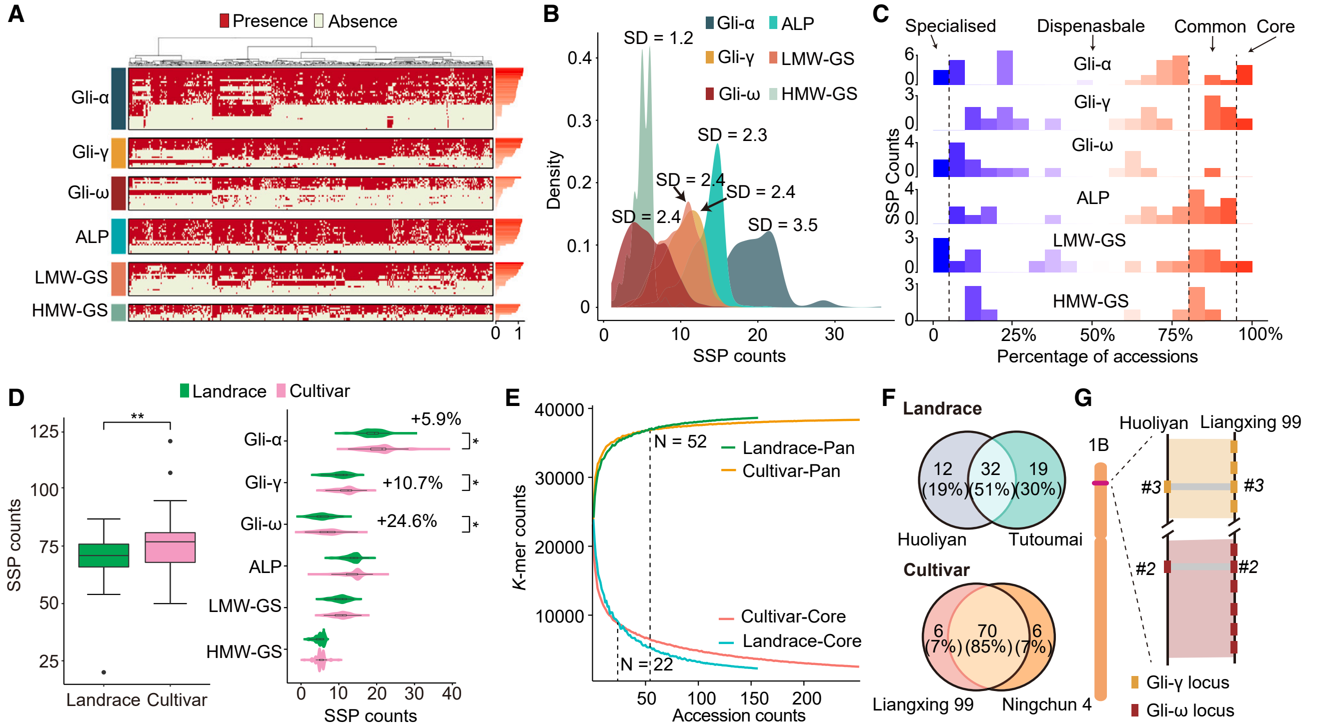


Figure 3. A pan-SSP presence/absence-variation map decodes genetic diversity and selection patterns for wheat SSPs.

(A) Presence/absence-variation profiles for SSPs across 365 wheat accessions. Each row represents an individual SSP gene, and each column represents an accession. Presence frequencies are shown in the rightmost track. SSP genes were ranked in decreasing order of presence frequency within each SSP class.

(B) Distribution of SSP gene counts across 365 accessions. Each SSP class is profiled individually.

(C) Population-scale distribution of presence frequencies. Each SSP class is profiled individually. SSPs with presence frequencies >0.95 were classified as core SSPs, those with presence frequencies ≤0.95 and ≥0.8 as common SSPs, those with presence frequencies <0.05 as specialized SSPs, and the rest as dispensable SSPs.

(D) Comparison of SSP gene presence in landraces versus cultivars. Left: total number of SSP genes in each wheat accession. Right: further subdivision of the SSP genes into Gli-α, Gli-γ, Gli-ω, ALP, LMW-GS, and HMW-GS classes. Statistical significance was assessed by *t*-tests. *$p \leq 0.05$; **$p \leq 0.01$. The proportions of increase in SSP gene presence from landraces to cultivars are labeled.

(E) Pangenome estimation of *k*-mer counts in landraces and cultivars. The pan- and core-curves are shown as the union and intersection of *k*-mers across the varieties in each group. The dashed lines indicate the numbers of SSP genes where the curves cross.

(F) Venn diagram of overlapping SSPs between the landraces Huoliyan and Tutoumai and between the cultivars Liangxing 99 and Ningchun 4.

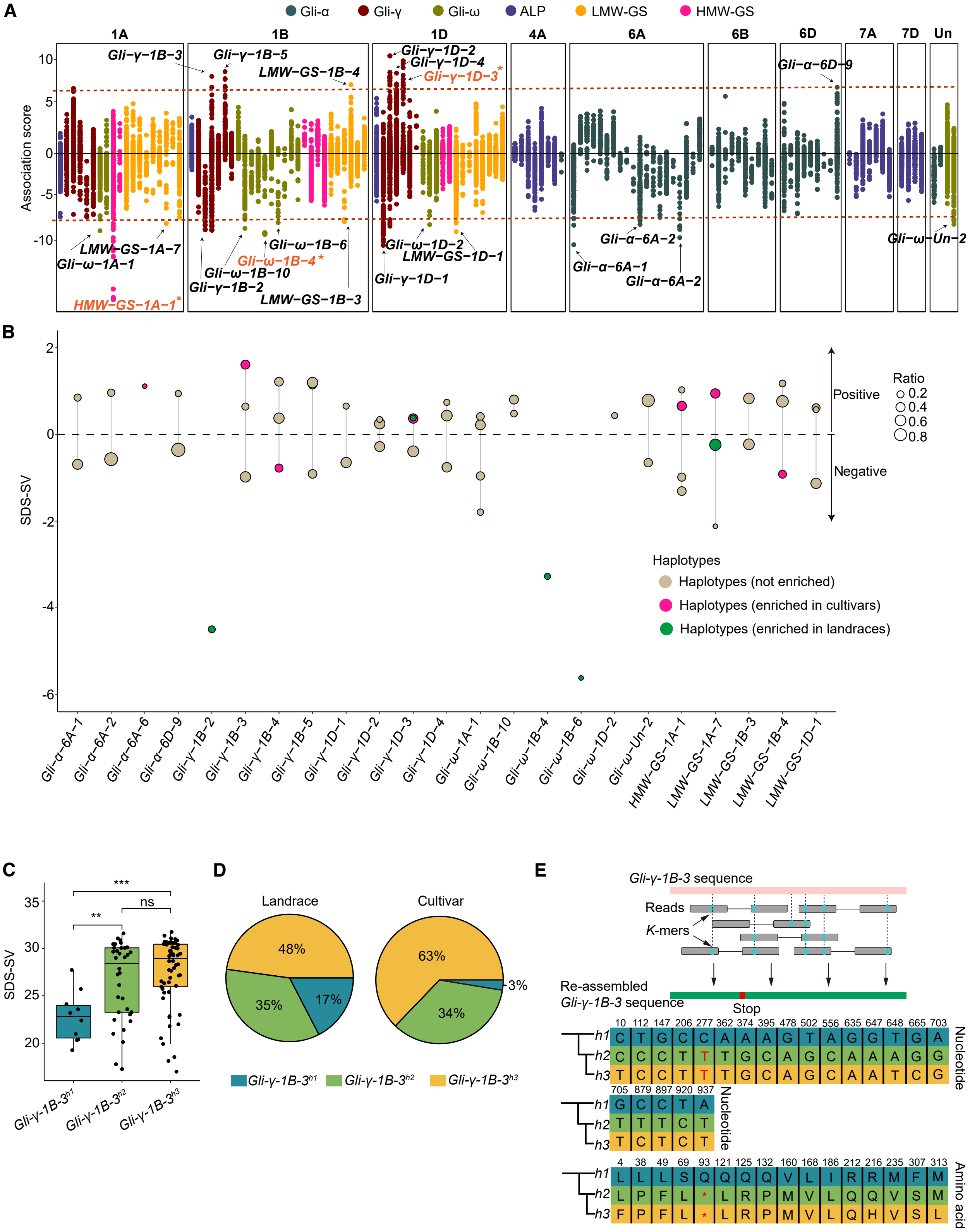
(G) Comparison of Huoliyan and Liangxing 99 in *Gli-γ* and *Gli-ω* loci on chromosome 1B.

but insight into the distribution of individual peaks or bands for particular proteins is still lacking. Thus, it is challenging to identify precise protein markers or patterns that can reliably predict end-use quality. We aimed to establish gene fingerprints for SSP genes in wheat genotypes.

Using PanSK, we built an SSP gene fingerprint map by determining presence/absence variations for 139 genes across a diverse panel of 365 resequenced wheat accessions, including 139 landraces and 226 modern cultivars (combining published [Guo et al., 2020; Hao et al., 2020; Zhou et al., 2020] and unpublished data) (Figure 3A and Supplemental Table 7). The number of predicted SSP genes varied extensively among varieties, with α-gliadins ranging from ≥10 to ≤30 members in individual accessions (Figure 3B).

Using the category nomenclature proposed by Liu et al. (2020), eight of the 139 wheat SSPs were present in 346–365 accessions (>95% of the collection) and were defined as core genes, 26 were present in 292–345 accessions (80%–95% of the collection) and were defined as common genes, 76 were present in 18–291 accessions (5%–80% of the collection) and were defined as dispensable genes, and 29 were present in 0–17 accessions (<5% of the collection) and were defined as specialized genes (Figure 3C and Supplemental Table 4). ω-Gliadin genes had the highest proportions of specialized and dispensable genes (95%), reflecting their variability. Because dispensable genes are typically found as tandem duplicates (Yocca and Edger, 2022), the higher variable counts of gliadin genes at the population level contribute to the underlying genetic plasticity of wheat end-use quality.

Numbers of SSP genes were significantly higher in cultivars (mean = 75.45) than in landraces (mean = 70.61, $p = 3.04 \times 10^{-6}$). This enrichment was mainly due to gliadins, with 5.9%, 10.7%, and 24.6% more α-, γ-, and ω-gliadin genes, respectively, in cultivars than in landraces (Figure 3D). To understand the relationship between diversity and abundance of SSP



	10	112	147	206	277	362	374	395	478	502	556	635	647	648	665	703
h1	C	T	G	C	C	A	A	A	G	T	A	G	G	T	G	A
h2	C	C	C	T	T	T	G	C	A	G	C	A	A	A	G	G
h3	T	C	C	T	T	T	G	C	A	G	C	A	A	T	C	G

	705	879	897	920	937
h1	G	C	C	T	A
h2	T	T	T	C	T
h3	T	C	T	C	T

	4	38	49	69	93	121	125	132	160	168	186	212	216	235	307	313
h1	L	L	L	S	Q	Q	Q	Q	V	L	I	R	R	M	F	M
h2	L	P	F	L	*	L	R	P	M	V	L	Q	Q	V	S	M
h3	F	P	F	L	*	L	R	P	M	V	L	Q	H	V	S	L

Figure 4. *k*-mer-based association analysis identifies beneficial alleles for wheat end-use quality.
(A) *k*-mer-based association analysis with SDS-sedimentation value (SDS-SV) on wheat chromosomes. Each dotted column along the *x* axis represents an SSP gene, and genes are ordered by their positions in the genome. Each dot represents one *k*-mer and its association with SDS-SV. The association score is defined as the negative log of the *p* value obtained from a *t*-test. Values above and below zero on the *y axis* represent positive and negative

(legend continued on next page)

genes during wheat breeding, we generated a saturation curve of the present SSP genes (Supplemental Figure 3) and SSP-specific *k*-mers (Figure 3E) along with the increase in pangenome size for both landrace and cultivar groups. When there were only a few accessions ($n < 52$), the landrace group had fewer pan-SSP *k*-mers than the cultivar group, consistent with the accumulation of SSP genes in modern wheat cultivars (Figure 3E). Similar results were observed for the saturation curve of present pan-SSP genes, with the switching point at $n = 137$ (Supplemental Figure 3). By contrast, when the population size was large enough (i.e., $n > 52$), the landrace group had more pan-SSP *k*-mers than the cultivar group, indicating that wheat landraces have more diverse SSP classes at the population level. Thus, through breeding, SSPs in cultivars have increased in number but decreased in diversity. For example, the landraces Huoliyan and Tutoumai contain 44 and 51 SSP genes, respectively, and only 32 are shared. By contrast, cultivars Liangxing 99 and Ningchun 4 both contain 76 predicted SSP genes, 70 of which are shared (Figure 3F). At the gliadin gene locus on chromosome 1B, Huoliyan contains only one predicted γ-gliadin gene and ω-gliadin gene, whereas Liangxing 99 contains four γ-gliadin and six ω-gliadin genes (Figure 3G).

To understand how breeding has shaped the SSP repertoire at the population level, we identified 29 and 20 SSP genes that were highly enriched in cultivars and landraces, respectively (Supplemental Figure 4 and Supplemental Table 4). For example, the elite HMW-GSs Dx5+Dy10 and Bx14+By15, which contribute to superior end-use-quality (Guo et al., 2020; Zhou et al., 2021), are highly enriched in the cultivar group (Supplemental Figure 4), suggesting that SSP genes have been pyramided in cultivars through artificial selection. This result also highlights the fact that wheat landraces harbor high diversity in SSP genes that remains underexploited in modern cultivars.

k-mer-based genome association reveals elite SSP alleles for wheat end-use-quality improvement

To identify elite SSP genes associated with end-use quality, we performed *k*-mer-based association analysis using a collection of wheat accessions with known SDS-sedimentation values (SDS-SVs), an indicator of end-use quality (Supplemental Table 8). Considering the major impact of the 1Dx+1Dy subunit pair of HMW-GSs and secalin proteins (in wheat-rye 1BL/1RS translocation lines) on quality (Niu et al., 2011; Guo et al., 2013; Chen et al., 2021), we selected 103 non-1BL/1RS translocation wheat varieties carrying identical HMW-GS types (Dx2+Dy12) for association analysis.

A total of 336 *k*-mers from 23 SSP genes were significantly associated with SDS-SV ($|\text{association score}| \geq 7$, $p < 1 \times 10^{-7}$), including one allele encoding HMW-GS 1Ax, four genes each for LMW-GS and α-gliadin, eight γ-gliadin genes, and six ω-gliadin genes (Figure 4A and Supplemental Table 9). Three of these SSP genes are known to play roles in end-use quality, including an HMW-GS gene (TraesCS1A02G317311) (Zhou et al., 2021) and two γ-gliadin genes, *Gli-γ-1D-3* (TraesFLD1D01G005600) (Liu et al., 2023) and *Gli-γ-1B-4* (TraesFLD1B01G010600) (Liu et al., 2023). The remaining 20 SSP genes represent novel candidates for end-use quality.

Next, we queried variation in these 23 SSP genes by distinguishing their haplotypes with *k*-mers. Sixty-three haplotypes, including nucleotide polymorphisms and presence/absence variants, were identified (Figure 4B). Thirty-one haplotypes showed positive effects on SDS-SV, and 23 haplotypes showed negative effects (Figure 4B). To understand the spread of these haplotypes during breeding, we compared the percentage of each haplotype between cultivars and landraces by introducing a breeding-selection score (BSS) for each haplotype. Five haplotypes associated with high SDS-SV were enriched in cultivars, indicating that they had been selected during breeding for end-use-quality improvement (Figure 4B). Four haplotypes associated with low SDS-SV were enriched in landraces, indicating that these genes may have been discarded during breeding (Figure 4B). Another 25 haplotypes associated with high SDS-SV have not been selected during modern breeding (Figure 4B and Supplemental Table 10) and represent novel candidates for future improvement.

Different haplotypes (denoted by the superscript “h”) of the same SSP gene often show opposite effects on end-use quality, highlighting the importance of selection of elite haplotypes using PanSK. For example, *Gli-γ-1B-3* (encoding a γ-gliadin) has three haplotypes: *Gli-γ-1B-3*^{h1}, *Gli-γ-1B-3*^{h2}, and *Gli-γ-1B-3*^{h3}. Accessions with haplotype *Gli-γ-1B-3*^{h2} or *Gli-γ-1B-3*^{h3} have higher SDS-SVs than those with *Gli-γ-1B-3*^{h1}(Figure 4C). The *Gli-γ-1B-3*^{h2} and *Gli-γ-1B-3*^{h3} haplotypes represent the major alleles (34% and 63%) in cultivars, and *Gli-γ-1B-3*^{h3} haplotype frequency is higher in cultivars than in landraces, indicating that *Gli-γ-1B-3*^{h3} has been selected (Figure 4D). *Gli-γ-1B-3*^{h1} haplotype frequency is lower in cultivars than landraces, indicating loss (Figure 4D). To compare sequence variation among the three *Gli-γ-1B-3*

effects. Dashed lines mark thresholds for statistically significant associations. IDs of genes identified with high confidence are shown, and known genes associated with end-use quality are marked with an asterisk. Un: the chromosomes on which the SSP genes are located are unknown (i.e., they are located on unassembled contigs).

(B) Standardized SDS-SV (standardized SDS-SV$_{\text{haplotype } i}$ = the mean SDS-SV of accession with haplotype *i* minus the mean SDS-SV of all accessions) of 23 SSP genes significantly associated with quality as determined from *k*-mer haplotype-partitioning results and breeding-selection signals. Only haplotypes with standardized SDS-SV ≥ 0.2 or ≤ -0.2 are displayed. The enrichment levels of each haplotype in landraces and cultivated varieties were evaluated. Haplotypes with a proportion greater than 75% in cultivated varieties were defined as “enriched in cultivars,” whereas those with a proportion greater than 75% in landraces were defined as “enriched in landraces.”

(C) Comparison of SDS-SVs among *Gli-γ-1B-3* haplotypes. Three biological replicates were quantified for each accession. Student’s *t*-test was used to determine the significance of differences between two groups. $^{**}p \leq 0.01$; $^{***}p \leq 0.001$.

(D) Distribution of the three *Gli-γ-1B-3* haplotypes in landraces and cultivars.

(E) The *Gli-γ-1B-3* coding sequence reassembled by PanSK enables identification of SNPs from three haplotypes. A C→T polymorphism in hap2 and hap3 is highlighted in red.

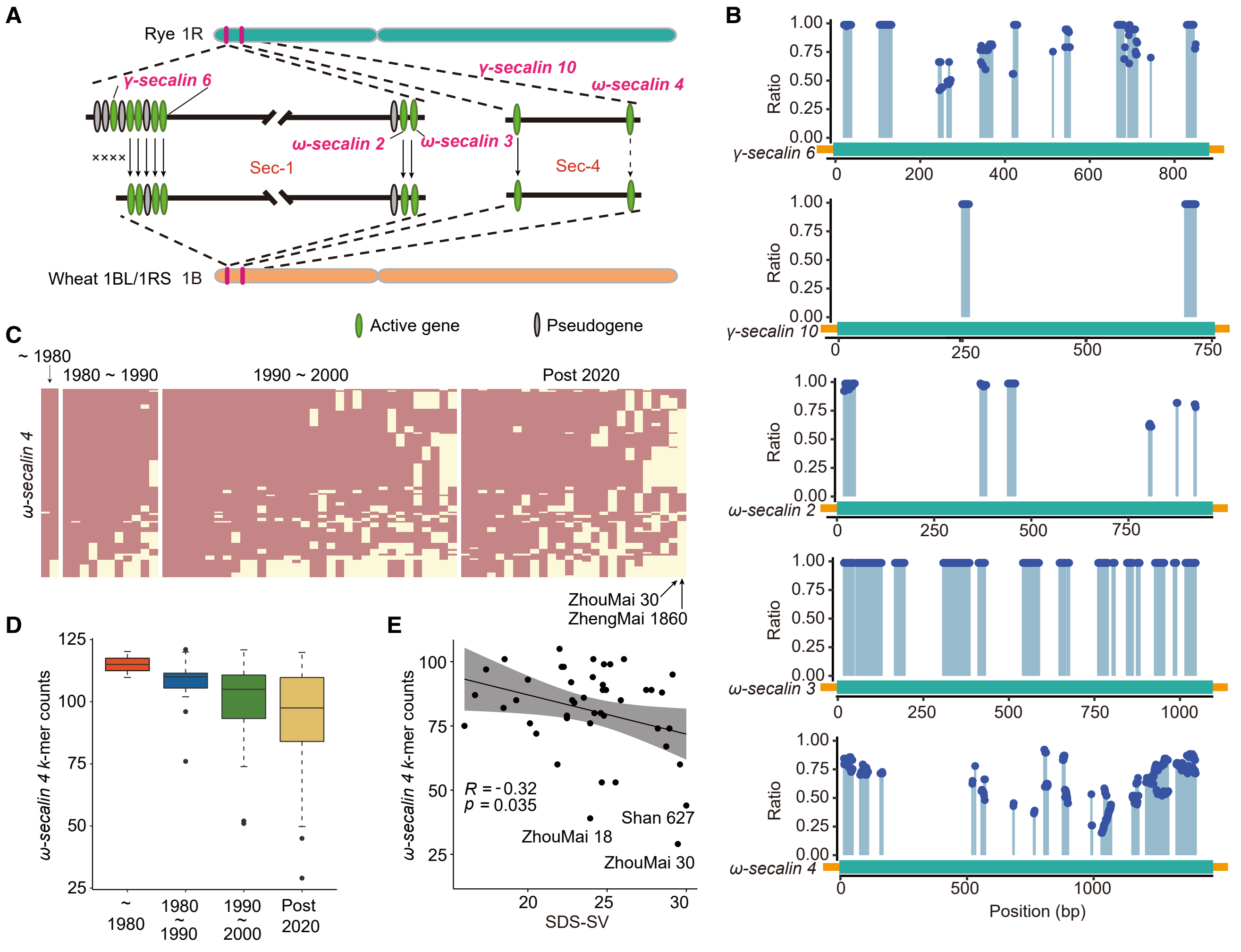


Figure 5. Identification of 1BL/1RS-translocation-specific *k*-mers and secalin genes and their association with end-use quality.

(A) Structure of *Sec-1* and *Sec-4* loci on chromosome 1R and certain secalin genes introduced to wheat chromosome 1B. Known secalin genes associated with the four loci are labeled.

(B) Conservation of secalin *k*-mers in selected secalin genes. Each lollipop signifies a *k*-mer, with the *x* axis indicating the *k*-mer position in each gene and the *y* axis indicating the ratio of *k*-mer presence in all 1BL/1RS translocation varieties.

(C) Secalin *k*-mer loss profile in 1BL/1RS translocation lines. Accessions were ordered by decreasing counts of secalin *k*-mers.

(D) Number of secalin *k*-mers in cultivars that contain the 1R chromosome across different time periods.

(E) Correlation between secalin *k*-mer counts and sedimentation values across 1BL/1RS translocation lines in wheat accessions. The black line denotes the line of best fit.

haplotypes, we reassembled their complete coding sequences using *k*-mers. *Gli-γ-1B-3*^{h1} and *Gli-γ-1B-3*^{h2} differ by 19 SNPs, whereas *Gli-γ-1B-3*^{h2} and *Gli-γ-1B-3*^{h3} differ by only one SNP (Figure 4E). Compared with *Gli-γ-1B-3*^{h1}, both *Gli-γ-1B-3*$^{h2/3}$ have a C→T change 277 bp downstream of the translation start site, resulting in a premature termination codon. These results suggest that loss-of-function alleles of *Gli-γ-1B-3* contribute to improved wheat end-use quality.

Revealing genetic variation in secalins of wheat-rye 1BL/1RS translocation lines using PanSK

In wheat-rye 1BL/1RS translocation lines, the short arm of rye chromosome 1RS replaces the short arm of wheat chromosome 1B (1BS) (Lee et al., 1995), and these lines are used worldwide because of their disease resistance and superior grain yield (Zhao et al., 2012). However, this translocation has deleterious effects on bread-baking quality (Supplemental Figure 5) (Heslop-Harrison et al., 1990; Li et al., 2021) owing to the introduction of rye secalin genes (encoding storage proteins) on 1RS and the loss of gliadins and LMW glutenin genes (Shewry and Bechtel, 2001). Breeding high-yielding, superior-quality elite varieties by regulating secalins in 1BL/1RS to increase grain quality and yield represents a useful strategy. Using PanSK, we detected variation in secalin genes to identify rational targets for improvement of end-use quality.

The assembled 1BL/1RS genome carries the *Sec-1* locus, which contains *γ-secalin 6* (representing *γ-secalin 5*, *γ-secalin 7*, *γ-secalin 8*, and *γ-secalin 9*), *ω-secalin 2*, and *ω-secalin 3*, and the *Sec-4* locus, which contains two active secalin genes, *γ-secalin 10* and *ω-secalin 4* (Figure 5A) (Shi et al., 2022). Using

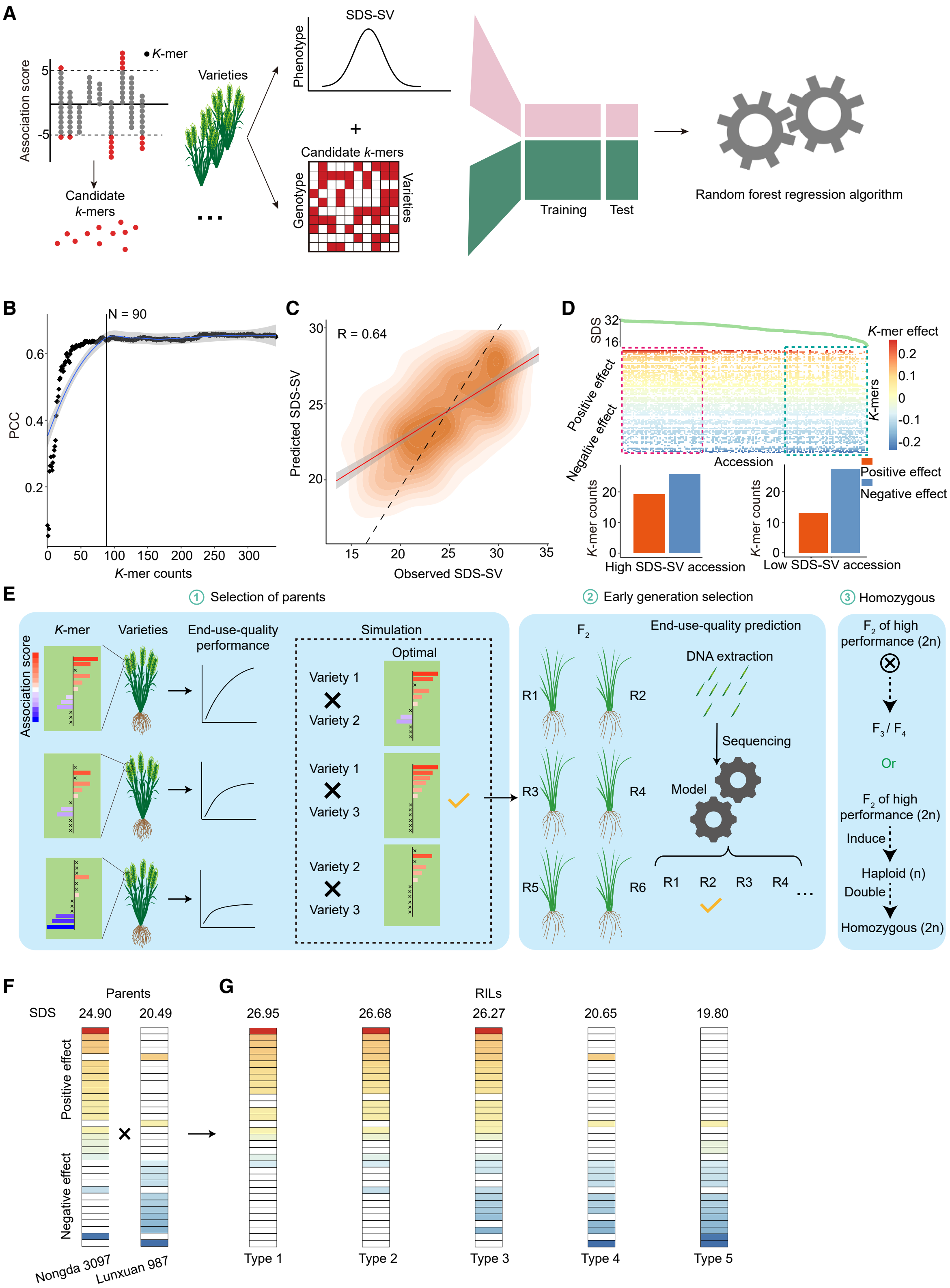


Figure 6. Genotype-to-phenotype prediction with PanSK accelerates selection efficiency for improvement of wheat end-use quality.
(A) Workflow of machine-learning-based prediction of SDS-SVs using *k*-mer presence/absence-variation profiles as an input feature. The random-forest algorithm was used to train and test the models.

(legend continued on next page)

PanSK, we identified *k*-mers unique to these secalins in order to catalog secalin variation in 77 1BL/1RS translocation lines released during the last 40 years (Supplemental Table 11). The representative *k*-mers specific to secalin genes of rye provenance were highly enriched in 1BL/1RS translocation lines and absent in non-1BL/1RS lines (Supplemental Figure 6). The genes *γ-secalin 6*, *γ-secalin 10*, *ω-secalin 2*, and *ω-secalin 3* were highly conserved with low diversity among 1BL/1RS varieties, indicating their narrow origins from rye (Figure 5B). On average, 34.5% of *ω-secalin-4*-specific *k*-mers could not be detected in 1BL/1RS lines, indicating that *ω-secalin 4* may have been mutated or eliminated during breeding.

Presence/absence variation for *ω-secalin-4*-specific *k*-mers indicated that this gene was gradually eliminated during breeding, a result supported by the gradual reduction in these *k*-mers in varieties released from the 1980s to 2020, with superior-quality elite varieties Zhengmai 1860 and Zhoumai 30 containing the fewest *k*-mers (Figure 5C and 5D). There was a significant negative correlation ($r = -0.32$, $p = 0.035$) between *k*-mer counts for *ω-secalin 4* and SDS-SV (Figure 5E), indicating that elimination of *ω-secalin 4* has a positive effect on end-use quality. Together, these results reveal that modern breeding has gradually removed ω-secalins on 1RS to improve end-use quality.

Genotype-to-phenotype prediction by PanSK accelerates selection efficiency for improvement of wheat end-use quality

The *k*-mers identified by PanSK can fully capture presence/absence variations and alleles of SSP genes, which have been challenging to genotype by traditional strategies based on SNP arrays or SNP calling. We therefore aimed to develop a *k*-mer-based prediction model for genotype-to-phenotype prediction and assessed the contribution of each SSP gene to end-use quality. We then proposed an ideal combination of SSP genes that would contribute to end-use-quality-related traits. For a training population, we randomly selected 1000 non-redundant *k*-mers as the initial candidate set. Next, we used a random-forest-based model with presence/absence-variation profiles of *k*-mers as the genotypes and SDS-SV as the phenotype to train the model. We opted for a greedy strategy to prioritize key *k*-mers by selecting candidates from the fewest *k*-mers with the best prediction performance each time, starting from one *k*-mer (Figure 6A). Use of 90 *k*-mers was sufficient to reach a stable Pearson correlation coefficient of 0.64 (Figure 6B). We thus developed an end-use-quality predictor, which we named “KPPer” (*k*-mer-based phenotype predictor), using genetic variation of 90 selected *k*-mers as the genotype (Supplemental Table 12).

To test the predictive power of KPPer, we predicted the end-use-quality performance of 172 wheat varieties and measured their SDS-SVs. Performance was evaluated by adapting a 10-fold cross-validation approach. We obtained a correlation of 0.64 between predicted and observed SDS-SVs (Figure 6C). Accessions with higher SDS-SVs (i.e., the top 30%) carried more *k*-mers that made a positive contribution to end-use quality, whereas accessions with the lowest SDS-SVs (the bottom 30%) carried fewer such *k*-mers (Figure 6D). These results highlight the power of PanSK-derived *k*-mers to accurately predict end-use-quality traits and facilitate selection of breeding lines with high estimated SDS-SVs to promote trait improvement.

KPPer can predict phenotype from genotype, thus mitigating the limitations of traditional phenotyping methods and showing potential for use in breeding. Here, we propose a new breeding strategy that makes use of PanSK for end-use-quality improvement (Figure 6E). First, elite germplasm resources containing elite SSP genes are selected as parental lines. Second, individual elite progeny are selected after genotype-to-phenotype prediction at an early stage (e.g., in the F_2 population). For individual plants in the F_2 population, end-use-quality performance is predicted by KPPer, and elite individuals carrying more SSP genes with positive effects on end-use quality are selected for the next generation. Finally, homozygous lines are selected at the F_3 or F_4 generation. By detecting the composition of different *k*-mers in different varieties, we can select suitable parental combinations that complement each other in order to aggregate *k*-mers with positive effects on quality and exclude *k*-mers with negative effects. This approach is expected to create new high-quality varieties.

(B) Pearson correlation coefficient between predicted and observed phenotypes. Key *k*-mers were prioritized by selecting the candidate *k*-mer with the best predictive performance from the leftmost *k*-mer, starting with 1 *k*-mer. The thin blue line denotes the line of best fit.

(C) Density plots for prediction accuracy of KPPer performance evaluated by adapting a 10-fold cross-validation approach. The correlation coefficient was calculated for each fold, and coefficients were then averaged over 10 folds. Contour lines indicate kernel density. The black line shows the function $y = x$, and the red line indicates the orthogonal distance regression line fitted from contour lines.

(D) Heatmap of presence/absence-variation status of 90 *k*-mers used in this study for SDS-SV prediction across 172 accessions. The color represents the effect of each *k*-mer. Dashed boxes in red and blue indicate accessions showing higher (top 30%) and lower (bottom 30%) SDS-SVs, respectively. Red and blue represent *k*-mers that make positive and negative contributions to SDS-SV, respectively. The bar charts at the bottom show the average number of positive- and negative-effect *k*-mers in the high and low SDS-SV accessions.

(E) Workflow for end-use-quality breeding in wheat. First, elite germplasm resources containing elite SSP genes are selected as parental lines on the basis of KPPer predictions (varieties with more positive-effect *k*-mers have superior quality performance, and those with more negative-effect *k*-mers have inferior quality performance). Second, elite individual plants are selected after genotype-to-phenotype prediction from early-stage offspring of the two parental lines, such as the F_2 population. For individual plants in the F_2 population, end-use-quality performance is predicted by KPPer. DNA extraction and low-throughput resequencing are performed on the offspring population during the juvenile phase, and elite individuals carrying more positive-effect *k*-mers of SSP genes are selected for the next generation. Finally, homozygous lines are selected at the F_3 or F_4 generation. In the future, to achieve high prediction efficiency, it will be necessary to combine double-haploid technology with genomic selection to create genetically uniform lines.

(F) *k*-mer composition and SDS-SVs of the parent strains Nongda 3097 and Lunxuan 987 used for hybridization. Colors represent the effects of each *k*-mer, and cross-hatching represents *k*-mers that differ between Nongda 3097 and Lunxuan 987. *k-mers* specific to Nongda 3097 and Lunxuan 987 are displayed.

(G) *k*-mer composition and SDS-SV performance of five recombinant inbred lines generated from Nongda 3097 and Lunxuan 987.

To validate this strategy, we used recombinant inbred lines (RILs) obtained from a Nongda 3097 × Lunxuan 987 cross to test the genotype-to-phenotype predictions. Nongda 3097 has a higher SDS-SV (observed = 25.0 ml, predicted = 24.90 ml) and carries six special negative and 14 positive *k*-mers. Lunxuan 987 has a lower SDS-SV (observed = 15.4 ml, predicted = 20.49 ml) and carries 11 special *k*-mers with negative effects on SDS-SV and two with positive effects (Figure 6F). In the RIL population, we analyzed *k*-mers in each line to identify the pyramiding of positive *k*-mers. RILs, such as types 1, 2, and 3, which contain 13, 14, and 15 *k*-mers with positive effects and 3, 4, and 10 *k*-mers with negative effects, respectively, have higher SDS-SVs (Figure 6G). By contrast, types 4 and 5, which contain only 2 and 1 *k*-mers with positive effects, have lower SDS-SVs. We obtained 15 lines that pyramided more *k*-mers with positive effects and had higher SDS-SVs (Supplemental Table 13). These results confirm the power of KPPer in genotype-to-phenotype prediction for end-use quality.

DISCUSSION

Improving the quality of wheat is often time consuming, and phenotypic evaluation is expensive, requiring large numbers of grains for an inefficient process (Prasad et al., 2003; Yang et al., 2020; Cao et al., 2024). In contrast to conventional breeding, which relies on phenotype-based evaluation, we propose a rational genome-design approach based on *k*-mers from SSP genes to facilitate the production of wheat varieties with superior end-use quality.

To realize this goal, it is first necessary to evaluate the contribution of each SSP gene to end-use quality, a task that has historically been challenging owing to the repetitive nature of SSP genes. To begin, it is important to differentiate individual SSP genes from among several copies, as they each contribute to functional variation (Niu et al., 2011; Guo et al., 2013; Clavijo et al., 2017; Chen et al., 2021). Our study shows that a *k*-mer approach is effective for analyzing multi-copy SSP gene families in wheat. This new approach overcomes challenges posed by the repetitive nature of SSP genes because it exploits the unique sequences present in repetitive regions. Using *k*-mers, we were able to efficiently identify and quantify the copy-number and sequence variation of multi-copy genes encoding gliadins and HMW and LMW glutenins. We observed that landraces had fewer SSPs but higher SSP diversity, whereas cultivars had more SSP copies but lower diversity (Figure 3D–3G). Landraces, shaped by natural and low-intensity artificial selection, maintain diverse genetic reservoirs. By contrast, cultivated varieties, shaped by intense artificial selection for specific traits, exhibit narrower genetic diversity. In addition, the breeding process often involves repeated crossing of a relatively small number of elite parental lines, which leads to reduced genetic diversity. SSP genes are present in multiple copies, and multi-copy genes are known to play an important role in the adaptability and evolution of organisms. Increased copy numbers of SSP genes may arise through various mechanisms, such as unequal crossing over, retrotransposition, or segmental duplication during breeding. The occurrence of multiple SSP gene copies in cultivars may contribute to a higher gluten content and a diversity of gluten proteins. These features make wheat more suitable for industrial food production.

We showed that a *k*-mer approach is a valuable method for exploring the structural and functional diversity of SSP families at a granular level. It can be particularly useful for investigating multi-copy gene families and their sequence variability, while providing insights into their evolutionary dynamics and functional implications.

k-mer-based GWAS enabled the exploration of associations between *k*-mer patterns and phenotypic traits for end-use quality and thus captured a broader range of variation in SSP genes. Combinations of specific subunits such as Dx5+Dy10, Bx7+By8, Bx14+By15, and the elite γ-gliadin allele *Gli-γ1-I* are associated with improved dough properties (Tanaka et al., 2003; Delorean et al., 2021; Yang et al., 2023). However, the contributions of LMW-GS and other gliadin genes remain less clear. As a result, wheat-cultivar improvement based on LMW-GS and gliadins has enjoyed comparatively limited success to date. After performing a *k*-mer GWAS, we identified 23 SSP genes, including 18 genes encoding gliadins, four encoding LMW glutenins, and one encoding HMW glutenin, as candidates for wheat end-use quality improvement. We also determined the contribution of each gene's genetic variation to end-use quality. This study provides additional avenues for optimizing wheat end-use quality using LMW-GS and gliadins in the future.

Using PanSK, we developed a novel breeding approach that uses genome design to facilitate pyramiding of genes related to end-use quality. In wheat breeding, genotype-to-phenotype prediction for quality improvement can be used to guide selection decisions and improve breeding efficiency (Naraghi et al., 2019; Sandhu et al., 2021). Genomic selection allows for the end-use-quality evaluation of individual plants at an early stage before they express their phenotypes in full (Gill et al., 2023). This accelerates breeding, reduces phenotyping costs, and increases selection accuracy. However, because individuals are heterozygous at early stages of phenotype prediction, the accuracy of prediction is limited. These individuals must be homozygous for an additional generation of self-pollination. In the future, to achieve high prediction efficiency, it will be necessary to combine double-haploid technology with genomic selection to create genetically uniform lines. Because the interaction between genotype and environment plays a substantial and often unpredictable role in wheat breeding, the performance of a genotype can vary widely across different environments (Li and Tao, 2022). Thus, PanSK can be improved by combining predictions from multiple environments, which will require sufficient data from such environments.

Although this study provides an effective approach for genome design based on selection of SSP genes, certain limitations remain. We did not consider non-SSP genes, such as transcription factors that regulate SSP genes, which also inevitably contribute to end-use quality. Pyramiding multiple elite SSP genes requires a large selection population for genotyping by *k*-mers. If more parents for double-crossing are selected, the strategy becomes more complex, and a much larger selection population will be needed. Furthermore, it is essential to balance end-use quality with other traits such as yield and resistance (Song et al., 2022). Our results demonstrate that scanning the PAV status of a limited number of representative *k*-mers can enable high prediction accuracy. It would be practical to design key parent-specific markers for assessing quality potential by scanning F_2 populations in a cost-efficient manner. Moreover, it would be wise to use these key markers to design a low-cost customized breeding chip, such as a liquid chip, that captures low-throughput sequencing data for genotyping SSP genes. We anticipate that this will become a powerful breeding strategy for wheat quality improvement.

METHODS

Plant material and growth conditions

Wheat plants used for end-use quality evaluation were grown in an experimental field of China Agricultural University in Beijing (39°57′N, 116°17′E) during the normal growing seasons from 2021 to 2023. Seeds were distributed in rows that were 2 m long with a 20-cm spacing. Details of the genotypes and germplasm used in this study are listed in Supplemental Table 7. RILs from the parents Nongda 3097 and Lunxuan 987 were generated by eight generations of self-pollination (Supplemental Table 13).

Full-length isoform sequencing and data analysis

Seeds of Nongda 3672 were grown in an open field, and endosperm samples were collected at 15 DAP. RNA was isolated from these samples using a TransZol Plant kit (TransGen Biotech, ET121-01). Purified RNA was dissolved in RNase-free water, and integrity was determined using an Agilent 2100 Bioanalyzer (Agilent Technologies). Only total RNA samples with an RNA integrity number ≥ 8 were used for library construction.

To generate sufficient and accurate read data for SSP annotation, the PacBio Sequel and Illumina NovaSeq 6000 platforms were used for sequencing. Long-read sequencing was performed using CCS mode on the PacBio Sequel platform. HiFi SMRTbell libraries were constructed using the SMARTer PCR cDNA Synthesis Kit according to the manufacturer's protocol. HiSeq libraries were prepared using the Illumina TruSeq RNA Sample Prep Kit. In brief, fragmentation buffer was added to break mRNA into fragments of 250–300 nucleotides. The fragments were used as templates to synthesize first-strand cDNA. After second-strand cDNA synthesis, fragments of suitable size were purified and amplified by PCR. PCR products were sequenced on the Illumina HiSeq 6000 platform.

Effective subreads were obtained using the P_Fetch and P_Filter functions (parameters: miniLength = 50, readScore = 0.75, artifac = −1000) in the SMRT Analysis software suite (http://www.pacificbiosciences.com/devnet/). CCSs were obtained from the P_CCS module using the parameters "MinCompletePasses = 2 and MinPredictedAccuracy = 0." After checking for the poly(A) signal and 5′ and 3′ adaptors, only CCSs with all three signals were considered full-length non-chimeric (FLNC) reads (Dong et al., 2015; Minoche et al., 2015). Unmerged subreads were also examined, and those with the three signals were incorporated into the final FLNC read set. Additional nucleotide errors in FLNC reads were corrected using the Illumina RNA-seq data with the software proovread (Hackl et al., 2014). Finally, 17 791 corrected FLNC reads were obtained, with an average read length of 1845 bp.

Searching full-length transcripts for SSP gene members

The 17 971 corrected FLNC reads were used to search for SSPs in Nongda 3672 endosperm using Pfam (El-Gebali et al., 2019). The conserved gliadin domain (PF13016 and PF00234) and HMW-GS domain (PF03157) were used to search 17 971 corrected FLNCs with Pfam_scan (https://github.com/aziele/pfam_scan) using default parameters. Diamond (Buchfink et al., 2015) was performed to identify SSPs to reduce transcript loss from 17 971 corrected FLNCs, and 42 gliadins in Xiaoyan 81 (Wang et al., 2017) were used as query sequences with the parameters "–sensitive –max-target-seqs 20 –evalue 1e-5." Sequences without signalIP were filtered out, as all SSPs contain signal peptides in their N termini. Signal peptides were predicted with SignalIP 5.0 (https://services.healthtech.dtu.dk/services/SignalP-5.0/). Sequence redundancy was suppressed using CD-HIT (https://github.com/weizhongli/cdhit/) with a 99% sequence-identity threshold, because the accuracy of SMRT sequencing corrected by Illumina sequencing is 99%. Finally, 85 SSPs were annotated in Nongda 3672.

BLASTN was used to identify candidate SSPs in published wheat genome assemblies, including those of Chinese Spring IWGSCv1.0, PI190962 (spelt wheat), Mace, LongReach, Lancer, Julius, ArinaLrFor, CDC Landmark, CDC Stanley, Jagger, Norin 61, SY Mattis, Zang 1817, and Fielder. The sequences of 42 previously characterized gliadins in Xiaoyan 81 and 85 SSPs in Nongda 3672 (Wang et al., 2017) were used as queries. Sequences of all genome assemblies were downloaded from the NCBI Sequence Read Archive or NCBI BioProjects (International Wheat Genome Sequencing Consortium, 2018; Guo et al., 2020; Walkowiak et al., 2020; Sato et al., 2021). Sequences without signalIP were filtered out. A total of 649 SSPs were identified, including 21 ω-gliadins, 89 γ-gliadins, 323 α-gliadins, 97 ALPs, 90 LMW-GS, and 29 HMW-GS (Supplemental Table 5). We then used BLAT (BLAST-Like Alignment Tool) (Kent, 2002) with default parameters for pairwise calculation of sequence similarity and merged sequences with sequence similarity $\geq$99% using MCL-14-137 (Dongen, 2008) with the parameter "-I 2" into one group. The longest sequence in each group was then selected as the representative sequence for subsequent analysis. A final set of 139 non-redundant representative SSP sequences was identified (Supplemental Table 4).

Identified *k*-mers uniquely represent SSP genes

The unique *k*-mers for each non-redundant SSP were identified using JELLYFISH 2.2.10 with the parameters "-t 15 -m k -s 4G -C." Various *k*-mer sizes (*k* = 17, 19, 21, 23, 25, 27, 29, 31, 33) were tested to balance computational complexity with *k*-mer specificity. As *k*-mer length increases, the number of identifiable unique *k*-mers in SSP genes also increases (Supplemental Figure 2) and the computational time and resources required grow exponentially. To find the best option, we compared the number and coverage of identifiable unique *k*-mers at different *k* lengths. Our analysis revealed that a 29-mer was the best choice for identifying sufficient unique *k*-mers for each representative gene while minimizing computational burden; 29 bp was therefore chosen as the optimal *k*-mer length for subsequent analysis. Unique 29-mers were generated and selected using the JELLYFISH subcommand "dump -L 1 -U 1." Finally, a database of 40 453 unique 29-mers was produced to represent the 139 non-redundant SSP genes.

Inference of SSP presence/absence variation in wheat accessions based on *k*-mers

To efficiently infer the presence of *k*-mers by scanning large raw sequence files, we developed a C++-based tool available at https://zhangzhaoheng24.github.io/PanSK/. We detected the presence or absence of 40 453 unique *k*-mers identified in 139 non-redundant representative sequences from the raw fastq sequencing files of each wheat variety. For each SSP gene, if over half of its unique *k*-mers were detected, we concluded that the gene was present in that particular wheat variety.

Reassembling SSP sequences with *k*-mers

To reassemble the sequence of a variety-specific SSP gene, all raw sequencing reads carrying corresponding unique *k*-mers were extracted and aligned to the reference sequence of the SSP gene using BWA (https://github.com/lh3/bwa) with default parameters. SNPs and indels were identified with the HaplotypeCaller module of GATK v3.8 (McKenna et al., 2010). A final sequence was generated by refilling SNPs/indels using the "FastaAlternateReferenceMaker" module of GATK v3.8.

Identification of diversified SSP haplotypes between landrace and cultivar groups

To identify SSP haplotypes that were selected or unselected during breeding, we determined the BSS based on the frequencies of varieties in groups of landraces and cultivars. The BSS for each haplotype was calculated as BSS = (number of cultivars carrying the haplotype/number of all accessions carrying the haplotype) × 2 − 1. Thus, BSS $\geq$ 0.5 indicates enrichment in the cultivar group, and BSS ≤ -0.5 indicates enrichment in the landrace group.

Association analysis

The "presence" or "absence" of *k*-mers was encoded as genotype information for each sequenced accession. For SDS-SV measurement, the sedimentation volume of 2 g of flour was measured for 5 min (Chen et al., 2019). Subsequently, the phenotyped SDS-SVs were associated with all the *k*-mers by performing *t*-tests to compare SDS-SVs between the two genotype groups for each *k*-mer. When *k*-mer present with higher SDS-SV with statistically significant *p*-values, the corresponding haplotypes were considered as a positive correlation, and vice versa. All tests were performed using at least three biological replicates for each sample.

Machine-learning-based prediction of end-use-quality phenotypic traits

To generate a set of non-redundant *k*-mers excluding the linkage for phenotypic prediction, *k*-means were used to cluster *k*-mers on the basis of their presence/absence-variation genotypes into 1000 initial clusters, and one *k*-mer was then randomly selected from each cluster. Thus, 1000 candidate *k*-mers were retained as the non-redundant *k*-mer set.

We then built KPPer to predict SDS-SVs from *k*-mers using a random-forest strategy with the parameters "estimators = 500, random states = 42" using scikit-learn (https://scikit-learn.org/) to train the machine-learning model and perform testing. The presence/absence-variation states of 1000 *k*-mers across 172 wheat varieties were used as genotypes, and corresponding data for SDS-SV, a major trait that contributes to wheat end-use quality, were collected for these varieties and used as phenotype data.

Ten-fold cross-validation was performed by randomly splitting the dataset into a training dataset (90%) and a testing dataset (10%). A greedy strategy was used to prioritize the fewest key *k*-mers for prediction, starting from the first round by selecting one *k*-mer with the highest Pearson's correlation coefficient for prediction. Thereafter, in each round, a new *k*-mer was added to the *k*-mer set to achieve the best prediction performance as a whole. Finally, we showed that 90 selected *k*-mers achieved a stable prediction performance (Pearson's correlation coefficient = 0.64).

DATA AND CODE AVAILABILITY

- The Iso-Seq sequence data generated in this study have been deposited in the National Genomics Data Center under accession number NGDC: PRJCA026228. The re-sequencing data were downloaded from previous studies under Genome Sequence Archive (https://bigd.big.ac.cn/gsa) accessions GSA: CRA001951, CRA001870, and CRA002507.
- The data analysis methods and code were based on previous studies at github (https://zhangzhaoheng24.github.io/PanSK/), and specific data and code are available upon request.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Molecular Plant Online*.

FUNDING

We thank Dr. Yongming Chen, Zihao Wang, Xiaoming Xie, Zhengzhao Yang, and Xiaobo Wang (China Agricultural University) for providing technical assistance. This work was supported by STI 2030-Major Projects (2023ZD04069), the National Natural Science Foundation of China (grant no. 32125030), the Pinduoduo-China Agricultural University Research Fund (PC2023A01003), and the Major Program of the National Agricultural Science and Technology of China (NK20220601).

AUTHOR CONTRIBUTIONS

Y.Y. and W.G. conceived the project. D.L. performed the experiments; Z.Z. performed bioinformatics analysis. B.L., W.W., and J.Z. provided technical assistance. M.X., Z.H., J.L., J.D., H.P., C.H., X.Z., Z.N., and Q.S. provided theoretical contributions to the project. Z.Z., D.L., Y.Y., and W.G. analyzed the data and wrote the article. All authors agreed with the finalized manuscript.

ACKNOWLEDGMENTS

No conflict of interest declared.

Received: March 7, 2024
Revised: May 6, 2024
Accepted: May 23, 2024
Published: May 24, 2024

REFERENCES

Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., Pozniak, C.J., Choulet, F., Distelfeld, A., Poland, J., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science **361**:eaar7191. https://doi.org/10.1126/science.aar7191.

Arora, S., Steuernagel, B., Gaurav, K., Chandramohan, S., Long, Y., Matny, O., Johnson, R., Enk, J., Periyannan, S., Singh, N., et al. (2019). Resistance gene cloning from a wild crop relative by sequence capture and association genetics. Nat. Biotechnol. **37**:139–143. https://doi.org/10.1038/s41587-018-0007-9.

Asseng, S., Guarin, J.R., Raman, M., Monje, O., Kiss, G., Despommier, D.D., Meggers, F.M., and Gauthier, P.P.G. (2020). Wheat yield potential in controlled-environment vertical farms. Proc. Natl. Acad. Sci. USA **117**:19131–19135. https://doi.org/10.1073/pnas.2002655117.

Battenfield, S.D., Guzmán, C., Gaynor, R.C., Singh, R.P., Peña, R.J., Dreisigacker, S., Fritz, A.K., and Poland, J.A. (2016). Genomic Selection for Processing and End-Use Quality Traits in the CIMMYT Spring Bread Wheat Breeding Program. Plant Genome **9**. https://doi.org/10.3835/plantgenome2016.01.0005.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods **12**:59–60. https://doi.org/10.1038/nmeth.3176.

Cao, S., Liu, B., Wang, D., Rasheed, A., Xie, L., Xia, X., and He, Z. (2024). Orchestrating seed storage protein and starch accumulation toward overcoming yield-quality trade-off in cereal crops. J. Integr. Plant Biol. **66**:468–483. https://doi.org/10.1111/jipb.13633.

Chen, H., Li, S., Liu, Y., Liu, J., Ma, X., Du, L., Wang, K., and Ye, X. (2021). Effects of 1Dy12 subunit silencing on seed storage protein accumulation and flour-processing quality in a common wheat somatic variation line. Food Chem. **335**:127663. https://doi.org/10.1016/j.foodchem.2020.127663.

Chen, Q., Zhang, W., Gao, Y., Yang, C., Gao, X., Peng, H., Hu, Z., Xin, M., Ni, Z., Zhang, P., et al. (2019). High Molecular Weight Glutenin Subunits 1Bx7 and 1By9 Encoded by Glu-B1 Locus Affect Wheat Dough Properties and Sponge Cake Quality. J. Agric. Food Chem. **67**:11796–11804. https://doi.org/10.1021/acs.jafc.9b05030.

Clavijo, B.J., Venturini, L., Schudoma, C., Accinelli, G.G., Kaithakottil, G., Wright, J., Borrill, P., Kettleborough, G., Heavens, D., Chapman, H., et al. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res. **27**:885–896. https://doi.org/10.1101/gr.217117.116.

Delorean, E., Gao, L., Lopez, J.F.C., Open Wild Wheat Consortium, Wulff, B.B.H., Ibba, M.I., Poland, J., Wulff, B., Steffenson, B., Steuernagel, B., et al. (2021). High molecular weight glutenin gene diversity in Aegilops tauschii demonstrates unique origin of superior wheat quality. Commun. Biol. **4**:1242. https://doi.org/10.1038/s42003-021-02563-7.

Dong, L., Liu, H., Zhang, J., Yang, S., Kong, G., Chu, J.S.C., Chen, N., and Wang, D. (2015). Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. BMC Genom. **16**:1039. https://doi.org/10.1186/s12864-015-2257-y.

Van Dongen, S. (2008). Graph Clustering Via a Discrete Uncoupling Process. SIAM J. Matrix Anal. Appl. **30**:121–141. https://doi.org/10.1137/040608635.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. **47**:D427–D432. https://doi.org/10.1093/nar/gky995.

Gaurav, K., Arora, S., Silva, P., Sánchez-Martín, J., Horsnell, R., Gao, L., Brar, G.S., Widrig, V., John Raupp, W., Singh, N., et al. (2022). Population genomic analysis of Aegilops tauschii identifies targets for bread wheat improvement. Nat. Biotechnol. **40**:422–431. https://doi.org/10.1038/s41587-021-01058-4.

Gill, H.S., Brar, N., Halder, J., Hall, C., Seabourn, B.W., Chen, Y.R., St Amand, P., Bernardo, A., Bai, G., Glover, K., et al. (2023). Multi-trait genomic selection improves the prediction accuracy of end-use quality traits in hard winter wheat. Plant Genome **16**:e20331. https://doi.org/10.1002/tpg2.20331.

Guo, W., Xin, M., Wang, Z., Yao, Y., Hu, Z., Song, W., Yu, K., Chen, Y., Wang, X., Guan, P., et al. (2020). Origin and adaptation to high altitude of Tibetan semi-wild wheat. Nat. Commun. **11**:5085. https://doi.org/10.1038/s41467-020-18738-5.

Guo, X.-H., Bi, Z.-G., Wu, B.-H., Wang, Z.-Z., Hu, J.-L., Zheng, Y.-L., and Liu, D.-C. (2013). ChAy/Bx, a novel chimeric high-molecular-weight glutenin subunit gene apparently created by homoeologous recombination in Triticum turgidum ssp. dicoccoides. *Triticum turgidum ssp*. *dicoccoides*. Gene **531**:318–325. https://doi.org/10.1016/j.gene.2013.08.073.

Hackl, T., Hedrich, R., Schultz, J., and Förster, F. (2014). proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics **30**:3004–3011. https://doi.org/10.1093/bioinformatics/btu392.

Hao, C., Jiao, C., Hou, J., Li, T., Liu, H., Wang, Y., Zheng, J., Liu, H., Bi, Z., Xu, F., et al. (2020). Resequencing of 145 Landmark Cultivars Reveals Asymmetric Sub-genome Selection and Strong Founder Genotype Effects on Wheat Breeding in China. Mol. Plant **13**:1733–1751. https://doi.org/10.1016/j.molp.2020.09.001.

Heslop-Harrison, J.S., Leitch, A.R., Schwarzacher, T., and Anamthawat-Jónsson, K. (1990). Detection and characterization of 1B/1R translocations in hexaploid wheat. Heredity **65**:385–392. https://doi.org/10.1038/hdy.1990.108.

Ikeda, T.M., Araki, E., Fujita, Y., and Yano, H. (2006). Characterization of low-molecular-weight glutenin subunit genes and their protein products in common wheats. Theor. Appl. Genet. **112**:327–334. https://doi.org/10.1007/s00122-005-0131-z.

International Wheat Genome Sequencing Consortium. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science **361**:eaar7191. https://doi.org/10.1126/science.aar7191.

Karikari, B., Lemay, M.-A., and Belzile, F. (2023). k-mer-Based Genome-Wide Association Studies in Plants: Advances, Challenges, and Perspectives. Genes **14**:1439.

Kent, W.J. (2002). BLAT–the BLAST-like alignment tool. Genome Res. **12**:656–664. https://doi.org/10.1101/gr.229202.

Lee, J.H., Graybosch, R.A., and Peterson, C.J. (1995). Quality and biochemical effects of a IBL/IRS wheat-rye translocation in wheat. Theor. Appl. Genet. **90**:105–112. https://doi.org/10.1007/BF00221002.

Li, G., Wang, L., Yang, J., He, H., Jin, H., Li, X., Ren, T., Ren, Z., Li, F., Han, X., et al. (2021). A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes. Nat. Genet. **53**:574–584. https://doi.org/10.1038/s41588-021-00808-z.

Li, Y., and Tao, F. (2022). Interactions of genotype, environment and management on wheat traits and grain yield variations in different climate zones across China. Agric. Syst. **203**:103521. https://doi.org/10.1016/j.agsy.2022.103521.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M., et al. (2020). Pan-genome of wild and cultivated soybeans. Cell **182**:162–176.e113.

Liu, D., Yang, H., Zhang, Z., Chen, Q., Guo, W., Rossi, V., Xin, M., Du, J., Hu, Z., Liu, J., et al. (2023). An elite γ-gliadin allele improves end-use quality in wheat. New Phytol. **239**:87–101. https://doi.org/10.1111/nph.18722.

Luo, G., Shen, L., Zhao, S., Li, R., Song, Y., Song, S., Yu, K., Yang, W., Li, X., Sun, J., et al. (2021). Genome-wide identification of seed storage protein gene regulators in wheat through coexpression analysis. Plant J. **108**:1704–1720. https://doi.org/10.1111/tpj.15538.

Mann, G., Diffey, S., Cullis, B., Azanza, F., Martin, D., Kelly, A., McIntyre, L., Schmidt, A., Ma, W., Nath, Z., et al. (2009). Genetic control of wheat quality: interactions between chromosomal regions determining protein content and composition, dough rheology, and sponge and dough baking properties. Theor. Appl. Genet. **118**:1519–1537. https://doi.org/10.1007/s00122-009-1000-y.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. **20**:1297–1303. https://doi.org/10.1101/gr.107524.110.

Michel, S., Kummer, C., Gallee, M., Hellinger, J., Ametz, C., Akgöl, B., Epure, D., Güngör, H., Löschenberger, F., and Buerstmayr, H. (2018). Improving the baking quality of bread wheat by genomic selection in early generations. Theor. Appl. Genet. **131**:477–493. https://doi.org/10.1007/s00122-017-2998-x.

Minoche, A.E., Dohm, J.C., Schneider, J., Holtgräwe, D., Viehöver, P., Montfort, M., Sörensen, T.R., Weisshaar, B., and Himmelbauer, H. (2015). Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol. **16**:184. https://doi.org/10.1186/s13059-015-0729-7.

Naraghi, S.M., Simsek, S., Kumar, A., Al Rabbi, S.M.H., Alamri, M.S., Elias, E.M., and Mergoum, M. (2019). Deciphering the Genetics of Major End-Use Quality Traits in Wheat. G3 (Bethesda, Md.) **9**:1405–1427. https://doi.org/10.1534/g3.119.400050.

Niu, Z.X., Klindworth, D.L., Wang, R.R.-C., Jauhar, P.P., Larkin, P.J., and Xu, S.S. (2011). Characterization of HMW Glutenin Subunits in *Thinopyrum intermedium*, *Th. bessarabicum*, *Lophopyrum elongatum*, *Aegilops markgrafii*, and Their Addition Lines in Wheat. Crop Sci. **51**:667–677. https://doi.org/10.2135/cropsci2010.04.0235.

Payne, P.I., and Lawrence, G.J. (1983). Catalogue of alleles for the complex gene loci, Glu-A1, Glu-B1, and Glu-D1 which code for high-molecular-weight subunits of glutenin in hexaploid wheat. Cereal Res. Commun. **11**:29–35.

Pirozi, M.R., Margiotta, B., Lafiandra, D., and MacRitchie, F. (2008). Composition of polymeric proteins and bread-making quality of wheat lines with allelic HMW-GS differing in number of cysteines. J. Cereal. Sci. **48**:117–122. https://doi.org/10.1016/j.jcs.2007.08.011.

Plavšin, I., Gunjača, J., Šatović, Z., Šarčević, H., Ivić, M., Dvojković, K., and Novoselović, D. (2021). An Overview of Key Factors Affecting Genomic Selection for Wheat Quality Traits. Plants **10**:745. https://doi.org/10.3390/plants10040745.

Prasad, M., Kumar, N., Kulwal, P.L., Röder, M.S., Balyan, H.S., Dhaliwal, H.S., and Gupta, P.K. (2003). QTL analysis for grain protein content using SSR markers and validation studies using NILs in bread wheat. Theor. Appl. Genet. **106**:659–667. https://doi.org/10.1007/s00122-002-1114-y.

Rasheed, A., Xia, X., Yan, Y., Appels, R., Mahmood, T., and He, Z. (2014). Wheat seed storage proteins: Advances in molecular genetics, diversity and breeding applications. J. Cereal. Sci. **60**:11–24. https://doi.org/10.1016/j.jcs.2014.01.020.

Ravel, C., Praud, S., Murigneux, A., Linossier, L., Dardevet, M., Balfourier, F., Dufour, P., Brunel, D., and Charmet, G. (2006). Identification of Glu-B1-1 as a candidate gene for the quantity of high-molecular-weight glutenin in bread wheat (Triticum aestivum L.) by means of an association study. Theor. Appl. Genet. **112**:738–743. https://doi.org/10.1007/s00122-005-0178-x.

Ravel, C., Faye, A., Ben-Sadoun, S., Ranoux, M., Dardevet, M., Dupuits, C., Exbrayat, F., Poncet, C., Sourdille, P., and Branlard, G. (2020). SNP markers for early identification of high molecular weight glutenin subunits (HMW-GSs) in bread wheat. Theor. Appl. Genet. **133**:751–770. https://doi.org/10.1007/s00122-019-03505-y.

Sandhu, K.S., Aoun, M., Morris, C.F., and Carter, A.H. (2021). Genomic Selection for End-Use Quality and Processing Traits in Soft White Winter Wheat Breeding Program with Machine and Deep Learning Models. Biology **10**:689.

Sandhu, K.S., Patil, S.S., Aoun, M., and Carter, A.H. (2022). Multi-Trait Multi-Environment Genomic Prediction for End-Use Quality Traits in Winter Wheat. Front. Genet. **13**:831020. 3389/fgene.2022.

Sato, K., Abe, F., Mascher, M., Haberer, G., Gundlach, H., Spannagl, M., Shirasawa, K., and Isobe, S. (2021). Chromosome-scale genome assembly of the transformation-amenable common wheat cultivar 'Fielder. DNA Res. **28**:dsab008. https://doi.org/10.1093/dnares/dsab008.

Shewry, P.R. (2003). Improving wheat quality: the role of biotechnology. In Bread Making: Improving Quality, S.P. Cauvain, ed. (Woodhead Publishing), pp. 168–186. https://doi.org/10.1533/9781855737129.1.168.

Shewry, P.R., and Bechtel, D.B. (2001). Morphology and chemistry of the rye grain. In Rye: Production, Chemistry and Technology, W. Bushuk, ed. (American Association of Cereal Chemists), pp. 69–127.

Shi, X., Cui, F., Han, X., He, Y., Zhao, L., Zhang, N., Zhang, H., Zhu, H., Liu, Z., Ma, B., et al. (2022). Comparative genomic and transcriptomic analyses uncover the molecular basis of high nitrogen-use efficiency in the wheat cultivar Kenong 9204. Mol. Plant **15**:1440–1456. https://doi.org/10.1016/j.molp.2022.07.008.

Song, X., Meng, X., Guo, H., Cheng, Q., Jing, Y., Chen, M., Liu, G., Wang, B., Wang, Y., Li, J., and Yu, H. (2022). Targeting a gene regulatory element enhances rice grain yield by decoupling panicle number and size. Nat. Biotechnol. **40**:1403–1411. https://doi.org/10.1038/s41587-022-01281-7.

Subedi, M., Ghimire, B., Bagwell, J.W., Buck, J.W., and Mergoum, M. (2022). Wheat end-use quality: State of art, genetics, genomics-assisted improvement, future challenges, and opportunities. Front. Genet. **13**:1032601. https://doi.org/10.3389/fgene.2022.1032601.

Subedi, M., Ghimire, B., Bagwell, J.W., Buck, J.W., and Mergoum, M. (2023). Wheat end-use quality: State of art, genetics, genomics-assisted improvement, future challenges, and opportunities. Front. Genet. **13**:1032601. https://doi.org/10.3389/fgene.2022.1032601.

Tanaka, H., Nakata, N., Osawa, M., Tomita, M., Tsujimoto, H., Yasumuro, Y., and Fischbeck, G. (2003). Positive effect of the high-molecular-weight glutenin allele, Glu-D1d, on the bread-making quality of common wheat. Plant Breed. **122**:279–280. https://doi.org/10.1046/j.1439-0523.2003.00847.x.

Tilman, D., Balzer, C., Hill, J., and Befort, B.L. (2011). Global food demand and the sustainable intensification of agriculture. Proc. Natl. Acad. Sci. USA **108**:20260–20264. https://doi.org/10.1073/pnas.1116437108.

Voichek, Y., and Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nat. Genet. **52**:534–540. https://doi.org/10.1038/s41588-020-0612-7.

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T., Brinton, J., Ramirez-Gonzalez, R.H., Kolodziej, M.C., Delorean, E., Thambugala, D., et al. (2020). Multiple wheat genomes reveal global variation in modern breeding. Nature **588**:277–283. https://doi.org/10.1038/s41586-020-2961-x.

Wang, D.-W., Li, D., Wang, J., Zhao, Y., Wang, Z., Yue, G., Liu, X., Qin, H., Zhang, K., Dong, L., and Wang, D. (2017). Genome-wide analysis of complex wheat gliadins, the dominant carriers of celiac disease epitopes. Sci. Rep. **7**:44609. https://doi.org/10.1038/srep44609.

Wang, L., Li, G., Peña, R.J., Xia, X., and He, Z. (2010). Development of STS markers and establishment of multiplex PCR for Glu-A3 alleles in common wheat (Triticum aestivum L.). J. Cereal. Sci. **51**:305–312. https://doi.org/10.1016/j.jcs.2010.01.005.

Yang, C., Chen, Q., Xin, M., Su, Z., Du, J., Guo, W., Hu, Z., Liu, J., Peng, H., Ni, Z., et al. (2023). A highly conserved amino acid in high molecular weight glutenin subunit 1Dy12 contributes to gluten functionality and processing quality in wheat. Journal of genetics and genomics = Yi chuan xue bao **50**:909–912. https://doi.org/10.1016/j.jgg.2022.11.002.

Yang, Y., Chai, Y., Zhang, X., Lu, S., Zhao, Z., Wei, D., Chen, L., and Hu, Y.G. (2020). Multi-Locus GWAS of Quality Traits in Bread Wheat: Mining More Candidate Genes and Possible Regulatory Network. Front. Plant Sci. **11**:1091. https://doi.org/10.3389/fpls.2020.01091.

Yocca, A.E., and Edger, P.P. (2022). Machine learning approaches to identify core and dispensable genes in pangenomes. Plant Genome **15**:e20135. https://doi.org/10.1002/tpg2.20135.

Zhao, C., Cui, F., Wang, X., Shan, S., Li, X., Bao, Y., and Wang, H. (2012). Effects of 1BL/1RS translocation in wheat on agronomic performance and quality characteristics. Field Crops Res. **127**:79–84. https://doi.org/10.1016/j.fcr.2011.11.008.

Zhou, Y., Zhao, X., Li, Y., Xu, J., Bi, A., Kang, L., Xu, D., Chen, H., Wang, Y., Wang, Y.-g., et al. (2020). Triticum population sequencing provides insights into wheat adaptation. Nat. Genet. **52**:1412–1422. https://doi.org/10.1038/s41588-020-00722-w.

Zhou, Z., Zhang, Z., Mason, A.S., Chen, L., Liu, C., Qin, M., Li, W., Tian, B., Wu, Z., Lei, Z., and Hou, J. (2021). Quantitative traits loci mapping and molecular marker development for total glutenin and glutenin fraction contents in wheat. BMC Plant Biol. **21**:455. https://doi.org/10.1186/s12870-021-03221-0.