



## OPEN ACCESS

## EDITED BY

Ainong Shi,  
University of Arkansas, United States

## REVIEWED BY

Chuanliang Deng,  
Henan Normal University, China  
Weidong Ning,  
Chinese Academy of Agricultural Sciences,  
China

## \*CORRESPONDENCE

Zhiyuan Liu

✉ liuzhiyuan01@caas.cn

<sup>†</sup>These authors have contributed equally to this work

RECEIVED 12 December 2025

REVISED 08 January 2026

ACCEPTED 09 January 2026

PUBLISHED 10 February 2026

## CITATION

She H, Wang H, Xu Z, Zhang H and Liu Z (2026) Pan-NLRome of *Spinacia* facilitates the rapid discovery of downy mildew resistance genes. *Front. Plant Sci.* 17:1766206. doi: 10.3389/fpls.2026.1766206

## COPYRIGHT

© 2026 She, Wang, Xu, Zhang and Liu. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Pan-NLRome of *Spinacia* facilitates the rapid discovery of downy mildew resistance genes

Hongbing She<sup>1,2†</sup>, Huiyu Wang<sup>3†</sup>, Zhaosheng Xu<sup>1</sup>, Helong Zhang<sup>1</sup> and Zhiyuan Liu<sup>1,2\*</sup>

<sup>1</sup>State Key Laboratory of Vegetable Biobreeding, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China, <sup>2</sup>Zhongyuan Research Center, Chinese Academy of Agricultural Sciences, Xinxiang, China, <sup>3</sup>Beijing Key Laboratory of New Technique in Agricultural Application, Beijing University of Agriculture, Beijing, China

Plant disease resistance is typically conferred by nucleotide-binding site leucine-rich repeats (NLR) proteins; however, the diversity of NLR genes in spinach has remained largely unexplored. We identified 2,549 NLR genes across 19 *Spinacia* assemblies of cultivated spinach and its two wild species, and constructed a comprehensive pan-NLRome, which was categorized into six subfamilies, and the most frequent NLR class was CC-NBARC-LRR. The pan-NLRome consists of 186 NLR families, comprising 38.7% core, 51.1% dispensable and 10.2% private families. By integrating pan-NLRome with *k*-mer-based genome-wide association studies (GWAS), we developed a novel pipeline for rapid identification of disease resistance genes. Using this approach, we directly pinpointed a candidate gene, *Te17S24XX\_Chr1\_nlr42*, for the *RPF1* locus, which confers resistance to spinach downy mildew pathogen races 1–7, 9, 11, 13, 15, 16, 18, and 20. In contrast, a single-genome-based method identified four candidate genes, which required further analysis confirm the final gene. The *Spinacia* pan-NLRome serves as an invaluable resource for exploring NLR gene evolution and plant disease resistance mechanisms. Our developed pipeline offers a reliable and efficient strategy for cloning resistance genes across multiple crops.

## KEYWORDS

downy mildew, *k*-mer-based GWAS, pan-NLRome, RPF1, spinach

## Introduction

Plants have evolved a two-tiered immune system to defend against diverse pathogenic invasions. The first layer, known as pattern-triggered immunity (PTI), is activated when pattern recognition receptors (PRRs) localized on the cell surface detect pathogen-associated molecular patterns (PAMPs). The second layer of immunity, termed effector-triggered immunity (ETI), is mediated by intracellular nucleotide-binding site leucine-rich repeats (NLRs) (Dangl et al., 2013; Wang et al., 2022). NLRs recognize specific pathogen effectors, thereby leading to a hypersensitive response (HR) to restrict pathogen growth

(Van de Weyer et al., 2019). The N-terminal domain is usually a Toll/interleukin-1 receptor/resistance protein (TIR) or a coiled-coil (CC) (Ameline-Torregrosa et al., 2008; Shao et al., 2016). Numerous studies have demonstrated the crucial role of NLRs in plant resistance, as evidenced in *Brachypodium* (Wu et al., 2022), wheat (Li et al., 2024; Ning et al., 2025), and rice (Wang et al., 2015).

Genome-wide association studies (GWAS) have become a faster approach to identifying statistical associations between single-nucleotide polymorphisms (SNPs) and phenotypic traits across diverse populations (Shang et al., 2023; Liu et al., 2024). However, this approach is limited by its inability to capture the full spectrum of genetic variation, particularly presence-absence variants (PAVs) and copy number variations (CNVs) (Zhou et al., 2022; Meng et al., 2025). This limitation is particularly pronounced in disease resistance research, as resistance (*R*) genes are frequently located in genomic regions rich in structural variation (Van de Weyer et al., 2019). To overcome the limitations, *k*-mer-based GWAS has emerged as a powerful reference-free approach for trait mapping (Voichkek and Weigel, 2020). This method directly processes sequencing reads, comparing the diversity of *k*-mers within the population, thereby avoiding alignment biases introduced by reference genomes. The *k*-mer-based GWAS have proven highly effective in disease resistance, often directly linking loci with phenotypes that standard SNP-based GWAS fail to detect, such as in maize, tomato (Voichkek and Weigel, 2020), and wheat (Arora et al., 2019; Jaegle et al., 2025). Furthermore, compared to a single reference, the pan-genomes more fully represent the entire genetic information of a species (Zhou et al., 2022). Therefore, by utilizing a pan-genome as the foundation for *k*-mer analysis, more novel loci can be captured. For instance, based on the pan-genomes, 93% of powdery mildew resistance-associated *k*-mers were identified in wheat, uncovering more than 25% *k*-mers compared to methods using a single reference genome (Jaegle et al., 2025). Recently, pan-NLRome has been constructed in numerous plant species (Mo et al., 2024; Ning et al., 2024, 2025; Parada-Rojas et al., 2025), laying the foundation for investigating disease resistance mechanisms.

Spinach downy mildew, caused by *Peronospora effusa* (Pe), formerly known as *Peronospora farinosa* f. sp. *spinaciae* (Pfs), is one of the most destructive diseases of spinach worldwide (Qian et al., 2016; Ribera et al., 2020). A total of 20 races were reported since 1824, 16 of which have increased substantially from 1996 to 2024 (Feng et al., 2014, 2018b; Correll et al., 2024). Six spinach downy resistance genes/alleles (*RPF1*–*RPF6*, resistance against *Peronospora Farinose*) have been reported, which could be overcome by specific races of the pathogen (Correll et al., 2011). For example, the *RPF1* locus provides resistance to races 1–7, 9, 11, 13, 15, 16, 18, and 20, while *RPF3* resists races 1, 3, 5, 8, 9, 11, 12, 14, 16, and 19 (Bhattarai et al., 2023). Previous studies have demonstrated that the *RPF1*–*RPF3* locus is located in the 0.34–1.76 Mb on chromosome 3 of the Sp75 genome assembly (Feng et al., 2018a; She et al., 2018; Bhattarai et al., 2020; Gao et al., 2022; Bhattarai et al., 2023). Specifically, the *RPF2* was reported to be the 1.11–1.72 Mb on chromosome 3 of Sp75 assembly, a CC-NBS-LRR domain gene *Spo12821* serving as the potential candidate gene for *RPF2* (Gao et al., 2022; Bhattarai et al., 2023). We previously fine-mapped the *RPF1* locus to 0.89 Mb of Sp75

chromosome 3 between 0.34–1.23 Mb using BC<sub>1</sub> and F<sub>2</sub> population (She et al., 2018), consistent with the 0.84 Mb interval detected using genotyping by sequencing (GBS) based SNP markers (Bhattarai et al., 2020). Although two candidate NLR genes, *Spo12903* and *Spo12784*, for *RPF1* were obtained (She et al., 2018), the key gene has yet to be determined.

Here, we identified comprehensive NLRs in 19 representative spinach assemblies and constructed a pan-NLRome. Based on the pan-NLRome, we optimized the *k*-mer-based GWAS approach to develop a pipeline for rapidly identifying disease-resistant genes/loci. Using this approach, we identified the candidate gene *Te17S24XX\_Chr1\_nlr42* (formerly termed *Spo12903* in Sp75 assembly) for the downy mildew resistance gene *RPF1* in spinach. Together, our work provides a foundation for identifying and functionally investigating disease resistance genes.

## Materials and methods

### NLR identification and classification

To obtain comprehensive NLR from spinach, we identified NLR across 19 representative *Spinacia* assemblies using NLR-Annotator v. 2.1b (Steuernagel et al., 2020) with default parameters. The NLRs in spinach could be further divided into six subfamilies: CC-NBARC-LRR, CC-NBARC, NBARC-LRR, NBARC, TIR-NBARC, TIR-NBARC-LRR. To assess the distribution of NLRs, we visualized them using the telomere-to-telomere genome (Sp\_YY\_v2) as an example.

Based on the result from NLR-Annotator, we also obtained each NLR and its flanking sequences (2 kb) for each genome, generating pan-NLRome sequences.

### Gene prediction

To further determine NLRs and their corresponding protein-coding genes, we performed automatic structural gene annotations for 19 *Spinacia* assemblies using Helixer v0.3.6 (Holst et al., 2023), as previous studies have demonstrated its exceptional accuracy in annotating NLR genes (Belinchon-Moreno et al., 2025). We employed an in-house script to identify NLR and its overlapped protein-coding genes, which were then subjected to further analysis.

### Gene-based pan-NLRome construction

We constructed a pan-NLRome using the NLR protein-coding genes from 19 *Spinacia* assemblies. First, we clustered these genes using OrthoFinder (Emms and Kelly, 2015) with the default parameters, and then divided these gene families into three categories, core, dispensable, and private based on their frequency. We defined the NLR families as core, dispensable, and private if they were present in all 19 accessions, 3–18 accessions, and 1–2 accessions, respectively.

## $K_A/K_S$ analysis of pan-NLRome

For comparison of the  $K_A/K_S$  value between the core and dispensable NLR families, we first selected the longest length NLR protein-coding gene from each accession within both the core and dispensable NLR families as the representative gene. Subsequently, TBtools (Chen et al., 2020) was employed to calculate the  $K_A/K_S$  value for homologous gene pairs between each pair of genomes within both the core and dispensable NLR families.

## K-mer-based GWAS

We utilized 116 spinach accessions (She et al., 2024a), comprising 30 resistant and 86 susceptible to *Peronospora effuse* race 9 (*Pfs9*), to identify causal variants associated with resistance to *Pfs9* based on our optimized *k*-mer-based GWAS. First, we filtered low-quality short reads from 116 accessions using fastp v0.23.4 (Chen et al., 2018) with the parameter “-q 20”. Then, the clean reads were aligned to pan-NLRome sequences using BWA 0.7.17-r1188 (Li, 2013) with default parameters. After removing duplicated reads, we obtained the paired-end (PE) reads that were located at pan-NLRome sequences (termed NLR-reads) using SAMtools (Li et al., 2009) with the parameters “view -F 12 -q 30”.

We created the *k*-mers table as described in Jaegle et al. (2025). Specifically, we firstly extracted *k*-mers (31 bp) from NLR-reads for each accession using KMC (Kokot et al., 2017). Then, we combined and filtered lists of *k*-mers using the script `list_kmers_found_in_multiple_samples` with parameters “-mac 5 -p 0.2”. Last, we obtained the *k*-mers table, containing the presence/absence pattern of each *k*-mers in the 116 spinach accessions. Combining the phenotype and *k*-mer table, we run *k*-mer-based GWAS with a permutation-based threshold for 5% family-wise error rate.

To retrieve the coordinates of the significant *k*-mers, we first extracted the PE reads containing the significant *k*-mers using `fetch_reads_with_kmers` ([https://github.com/voichek/fetch\\_reads\\_with\\_kmers](https://github.com/voichek/fetch_reads_with_kmers)). Subsequently, we quantified the number of PE reads containing significant *k*-mers on pan-NLRome sequences to identify the target NLR or determined the coordinates of the significant *k*-mers by aligning the corresponding PE reads against the target genome using BWA 0.7.17-r1188 (Li, 2013) with default parameters.

## RNA-seq analysis

We aligned the clean reads against the Sp\_YY\_v2 genome using HISAT2 v2.1.0 (Kim et al., 2015), followed by calculating the read count using featureCounts v2.0.1 with the parameters ‘-T 10 -p -t exon -g ID’ (Yang et al., 2014). Gene expression was normalized using the transcripts per million (TPM) method with an in-house script. The expression patterns are shown using R v4.1.1.

## Genome-wide association studies

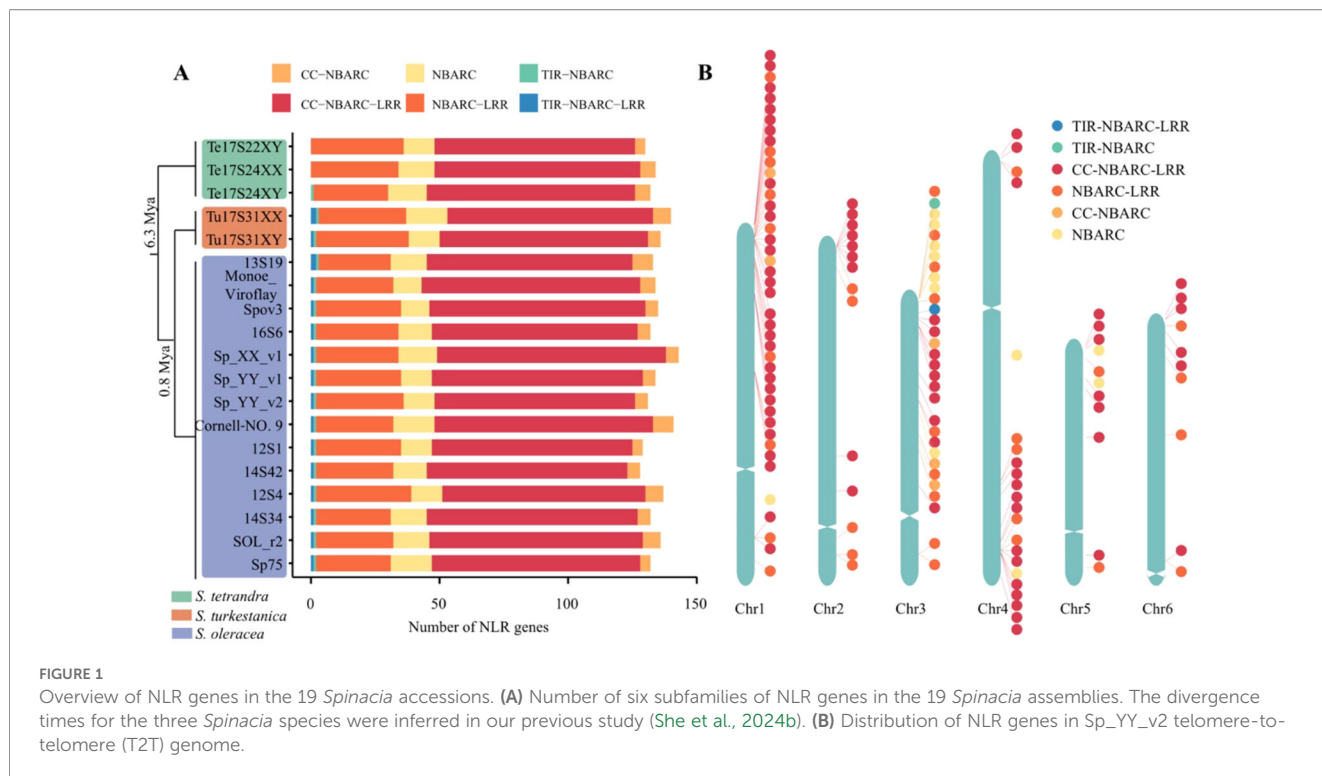
The 116 spinach accessions mentioned above were used for SNP-based GWAS analysis based on the Sp\_YY\_v2 assembly. The clean reads were aligned to Sp\_YY\_v2 assembly using BWA 0.7.17-r1188 (Li, 2013) with default parameters. The SNPs were identified using GATK v4.3 (McKenna et al., 2010), and then filtered using the GATK with “QD < 2.0 || FS > 60.0 || MQ < 40.0 || SOR > 3.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0”, and indels with “QD < 2.0 || FS > 200.0 || SOR > 10.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0; (2)”, and VCFtools v0.1.16 (Danecek et al., 2011) with the parameters “-max-missing 0.85 -mac 4 -minQ 30 -maf 0.05 -min-alleles 2 -max-alleles 2”. A total of 1,087,947 high-quality SNPs were obtained. The GWAS was performed using EMMAX software (Kang et al., 2010).

## Results

### Identification of nucleotide-binding site and leucine-rich repeat genes in 19 *Spinacia* genomes

To determine the dynamics in different *Spinacia* genomes, we predicted NLR genes in 19 *Spinacia* genomes using NLR-annotator (Steuernagel et al., 2020). Fourteen spinach genomes were from our previous studies (She et al., 2023, 2024a, 2024b); the remaining five assemblies (Monoe-Viroflay, SpoV3, Cornell-No.9, SOL\_r2, and Sp75) were used in others (Xu et al., 2017; Cai et al., 2021; Hirakawa et al., 2021; Hulse-Kemp et al., 2021; Ma et al., 2022). These accessions included 14 cultivated spinach species, *S. oleracea*, two of its closest wild relatives, *S. turkestanica* (0.8 million years ago (Mya)), and three more distant relatives, *S. tetrandra* (~6.3 Mya) (Figure 1A; Supplementary Table S1).

In total, 2,549 NLR genes were annotated in the 19 spinach genomes, with a range of 128 to 143 genes per accession (Supplementary Table S2). The 12 cultivated spinach plants and their five wild relatives shared a similar number of NLR genes (Wilcoxon test,  $p = 0.89$ ) (Supplementary Figure S1). Furthermore, these NLR genes were categorized into six subfamilies, and the most frequent NLR class was CC-NBARC-LRR (60.58%), followed by NBARC-LRR (23.90%) and NBARC (9.96%) (Figure 1A, Supplementary Table S2). TIR-NBARC and TIR-NBARC-LRR are rarely detected in *Spinacia* species, particularly in *S. tetrandra* species (Supplementary Table S2), suggesting that the two subfamilies probably evolved after splitting from *S. tetrandra* and *S. turkestanica*/*S. oleracea*. NLRs exhibited an uneven distribution along chromosomes, invariably clustering together and predominantly located near subtelomere regions (Figure 1B). Applying the definition of NLR clusters as genes within 200 kb of each other in the genome (Holub, 2001), 47.73%–57.86% of NLRs in each accession were located in such clusters (Supplementary Table S3, Supplementary Figure S2), consistent with a previous report on *Arabidopsis thaliana* (Van de Weyer et al., 2019).



## Pan-NLRome of *Spinacia*

To understand the variation in NLR content, we inferred 2846 NLR protein-coding genes that overlapped with NLR loci in the corresponding assembly. 2813 (99%) of these genes were annotated and associated with disease resistance using the Non-redundant protein Sequence (NR), Swiss-Prot, Pfam, TrEMBL, and *A. thaliana* databases (Supplementary Table S4). Then, we identified 186 pan-NLRome by clustering NLR protein-coding genes (see Methods) in the 19 *Spinacia* accessions (Supplementary Table S5). The number of NLR families exhibits a positive correlation with genome number, with their abundance stabilizing at a genome number of 12 (Figure 2A), indicating that the 19 *Spinacia* accessions are diverged and that the pan-NLRome closely captures the NLRs of spinach. Furthermore, only one and none additional NLR family was found when the 12<sup>th</sup> and 17<sup>th</sup> accession was added, respectively (Supplementary Figure S3).

Based on the frequency of NLR gene families in the 19 *Spinacia* genomes, we further divided the pan-NLRome into core, dispensable, and private families. In total, we obtained 72 (38.7%) core NLR families present in 19 accessions, 95 (51.1%) dispensable NLR families present in 3–18 accessions, and 19 (10.2%) NLR families present in 1–2 accessions (Figure 2B). The proportion of pan-NLRome gene families across 19 genomes revealed that dispensable NLRs are more likely to be present in *S. oleracea*/*S. turkestanica* compared to *S. tetrandra*, whereas private NLRs showed the opposite trend (Figure 2C), suggesting high diversity between *S. oleracea*/*S. turkestanica* and *S. tetrandra*. Moreover, the 19 spinach accessions shared an average of 58.99% of core NLR genes, exceeding the proportion found in dispensable NLR genes

(39.93%) (Figure 2D, Supplementary Table S6). These dispensable NLR genes were significantly prevalent in *S. oleracea*/*S. turkestanica* (Wilcoxon test,  $p < 0.01$ ); conversely, the private NLR genes (1.48%) were predominantly present in the *S. tetrandra* (Wilcoxon test,  $p < 0.05$ ) (Supplementary Figure S4), further confirming that *S. tetrandra* represents a more distant spinach wild relative.

Among the core, dispensable, and private NLR genes, the most frequent NLR class was CC-NBARC-LRR (52.38%–70.24%), followed by the NBARC-LRR (22.29%–24.16%) and NBARC (3.25–9.52%) (Figure 2E). The proportion of CC-NBARC in private NLR genes (14.29%) was higher than that in core (2.82%) and dispensable (2.29%) NLR genes (Figure 2E; Supplementary Table S7). Moreover, the core NLR genes exhibited longer CDS length (Wilcoxon test,  $p < 2e-5$ ) and lower  $K_A/K_S$  values (Wilcoxon test,  $p < 0.0097$ ) compared to dispensable NLR genes, suggesting that core NLR genes possess more conserved functions, consistent with previous findings in soybean (Liu et al., 2020) and broomcorn millet (Chen et al., 2023).

## A pipeline for identifying resistance genes using pan-NLRome and *k*-mer-based genome-wide association studies

*K*-mer-based GWAS is an efficient approach to identify causal variants associated with phenotypes, particularly disease resistance in plants (Voichek and Weigel, 2020; Jaegle et al., 2025). To enhance the efficiency of disease-resistance gene discovery, we optimized this approach by integrating the pan-NLRome, which primarily consists of three steps: identification of NLR and flanking sequences

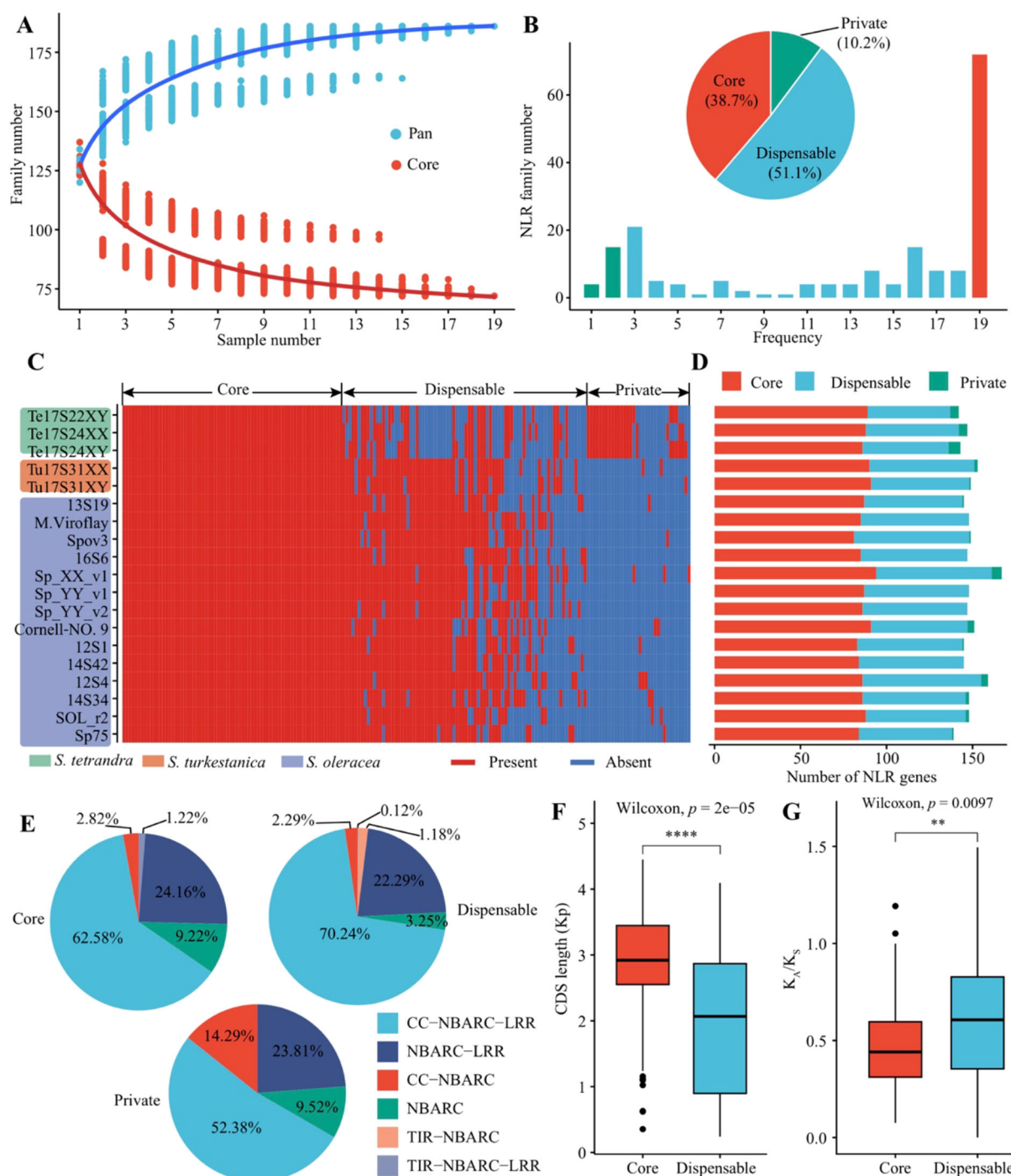


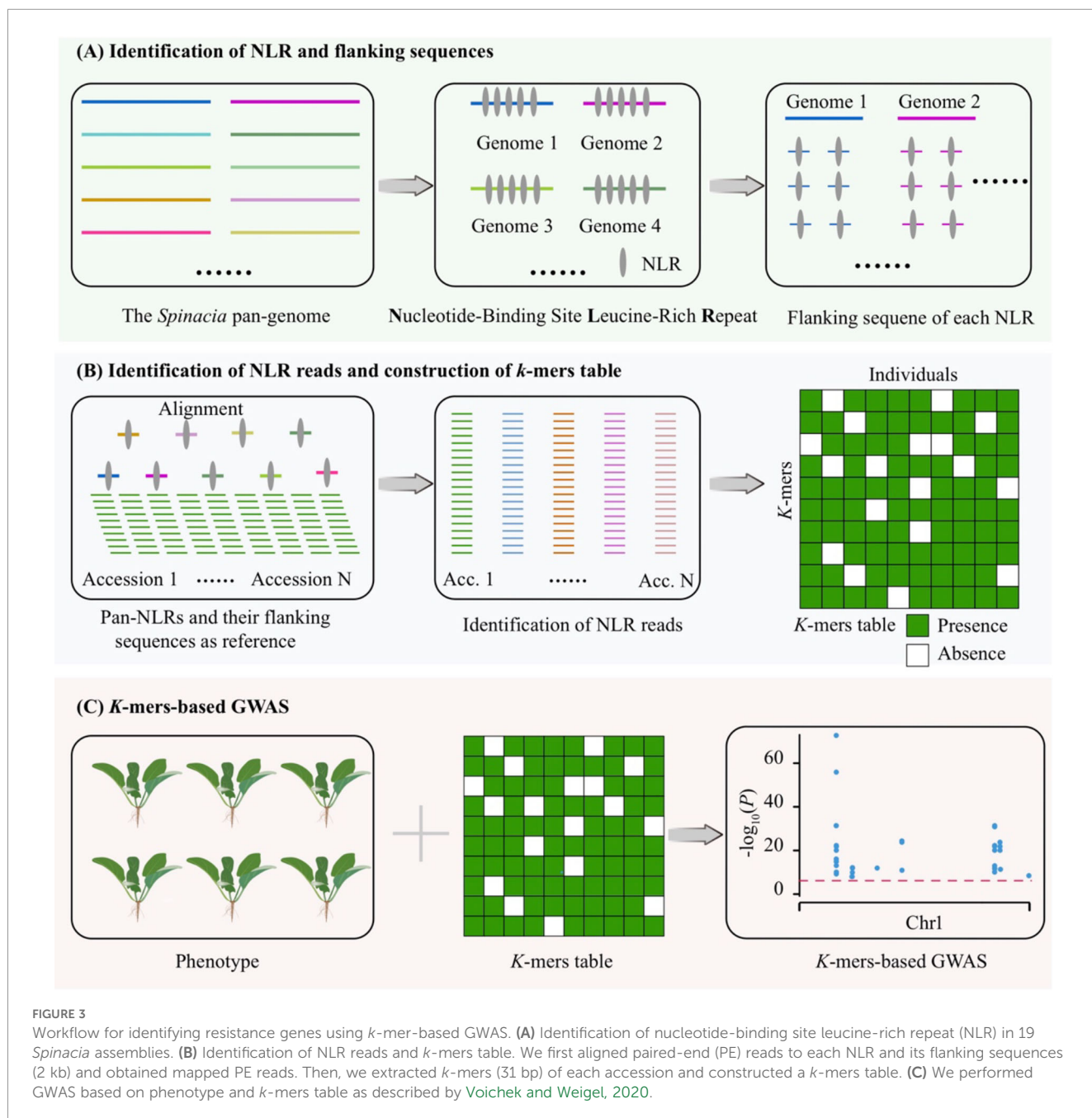
FIGURE 2

Pan-nucleotide-binding site leucine-rich repeat (NLR) analysis. (A) Pan- and core NLR of 19 *Spinacia* accessions. The blue and red curves represent the number of pan- and core NLR families after random combinations for each specific number of accessions. (B) Distribution of NLR families in the pan-NLRome. The histogram shows the number of NLR gene families in the 19 genomes with different frequencies. The pie chart shows the proportion of the NLR gene family in the core (red), dispensable (blue), and private (green) NLR genes. (C) Presence and absence information of pan NLR gene families in the 19 *Spinacia* genomes. (D) The number of core, dispensable, and private NLR genes of each genome. (E) Proportion of different NLR gene types in core, dispensable, and private NLR genes, respectively. Comparison of CDS length (F) and  $K_A/K_S$  (G) between the core and dispensable NLR genes, respectively. Significance was determined using the Kruskal-Wallis test. \*\* $p < 0.01$ , \*\*\*\* $p \leq 0.0001$ . For each boxplot, the box edges represent the interquartile range (IQR), with the centerline indicating the median. The whiskers extend to 1.5x the interquartile range.

(Figure 3A), identification of NLR reads and construction of *k*-mer table (Figure 3B), and conduction of *k*-mer-based GWAS (Figure 3C). First, we obtained the nucleotide-binding site leucine-rich repeat (NLR) and its flanking sequences (2 kb) for each of the 19 *Spinacia* genomes, generating a pan-NLRome sequences. Secondly, we identified NLR reads by aligning short reads of spinach accessions to pan-NLRome sequences, followed by extracting *k*-mers (31 bp) and constructing a *k*-mers table. Last, we performed *k*-mer-based GWAS based on phenotype and *k*-mer table as described by Voichek and Weigel, 2020.

### Screening of downy mildew resistance gene *RPF1* in spinach

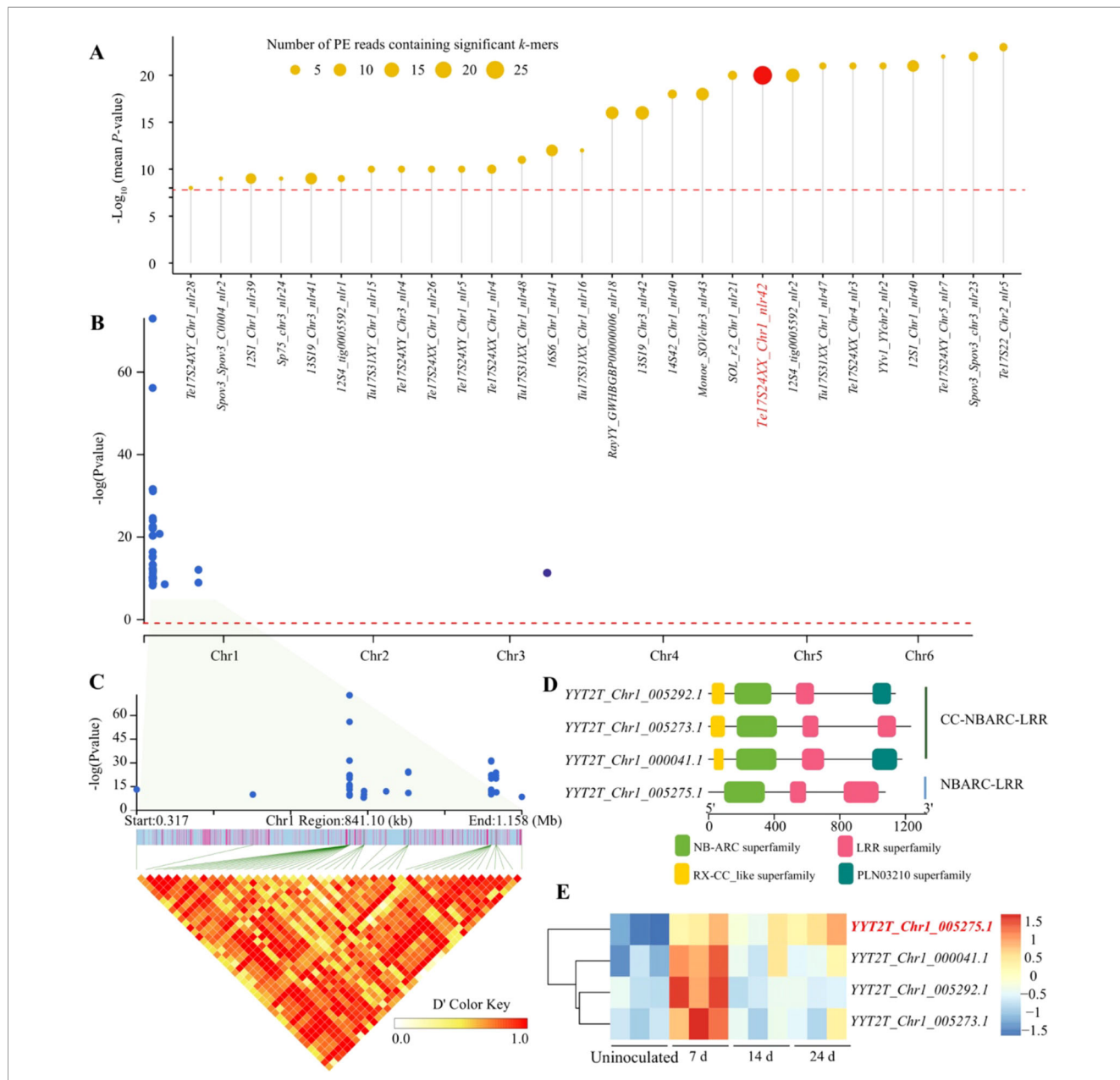
To demonstrate the usefulness of our approach, we utilized 116 spinach accessions (Supplementary Table S8) (She et al., 2024a), comprising 30 resistant and 86 susceptible to *Peronospora effuse* race 9 (*Pfs9*), to identify downy mildew resistance gene *RPF1* loci that resist *Pfs9* (Correll et al., 2011; She et al., 2018). Among these accessions, a total of 111.02 Gb of short reads were obtained from the NLR and flanking sequences, generating a *k*-mer table



comprising 10,001 *k*-mers. In our *k*-mer-based GWAS, we detected 50 significant *k*-mers associated with *RPF1*, with the *Te17S24XX\_Chr1\_nlr42* gene enriching the highest number of *k*-mers and exhibiting high significance. (Figure 4A). Notably, our previous study identified *Te17S24XX\_Chr1\_nlr42* (formerly designated as *Spo12903*) as a candidate gene for *RPF1* using BC<sub>1</sub> and F<sub>2</sub> populations (She et al., 2018).

Otherwise, to assess the effectiveness of *k*-mer-based GWAS in a single genome, we mapped the PE reads containing the significant

*k*-mers against the Sp\_YY\_v2 assembly. The vast majority of these reads were located within the 781 kb–1101 kb interval on chromosome 1 (Figures 4B, C, Supplementary Table S9), consistent with previous observations (She et al., 2018; Bhattarai et al., 2020; Cai et al., 2021). Within the candidate region, we identified four NLR genes, three of which carry the canonical CC-NBARC-LRR architecture, and one gene, *YYT2T\_Chr1\_005275.1*, a homolog of *Te17S24XX\_Chr1\_nlr42* mentioned above, lacks the N-terminal coiled-coil domain (NBARC-LRR only) (Figure 4D). The



**FIGURE 4** Identification of the downy mildew resistance gene *RPF1* in spinach. (A) The  $-\log_{10}(\text{mean } P\text{-value})$  of *k*-mers on pan-NLRome sequences. Dot size represents the number of PE reads containing significant *k*-mers. The red horizontal dashed line indicates the Bonferroni-corrected significance thresholds ( $\alpha = 0.05$ ). (B) Manhattan plot of *k*-mer-based GWAS of downy mildew resistance gene *RPF1*. Only significantly *k*-mers were shown. (C) The expanded version of the significant *k*-mers and LD pattern in Chr1. (D) The conserved domains of four candidate genes of *RPF1*. (E) Expression patterns of the four candidate genes of *RPF1* in uninoculated plants and those inoculated with downy mildew pathogen *Pfs9* at 7, 14, and 24 days post-inoculation. The gene labeled in red indicates that its expression level was significantly higher than that in uninoculated plants across all three post-inoculation periods.

expression analysis indicates that all four genes exhibited elevated expression levels at 7 days post-inoculation, with only the *YYT2T\_Chr1\_005275.1* gene maintaining sustained expression at both 14 and 24 days (Figure 4E). Therefore, we confirmed that *Spo12903* is probably *RPF1*.

Furthermore, we performed the SNP-based GWAS using the Sp\_YY\_v2 reference genome, which successfully detected significant SNPs encompassing the *YYT2T\_Chr1\_005275.1* locus (Supplementary Table S10). However, the SNP-GWAS also identified numerous other significant signals across the genome, making it difficult to unequivocally prioritize this specific gene (Supplementary Figure S5). Taken together, compared to a single genome, pan-NLRome enables faster and more accurate anchoring of candidate genes.

## Discussion

Nucleotide-binding site leucine-rich repeat (NLR) proteins represent one of the most important gene families in plants, as they confer disease resistance by recognizing pathogen proteins (Van de Weyer et al., 2019). Although 139 NBS-LRR genes have been reported previously in Sp75 assembly (Xu et al., 2017), this is far from sufficient. A single genome cannot capture the entirety of spinach genetics, particularly lacking genetic sequences from wild species, which are considered donors of downy mildew resistance loci for cultivated spinach (She et al., 2024b). In this study, we identified 2,549 NLR genes in the 19 spinach genomes, including 14 cultivated spinach, two of its closest wild relatives, *S. turkestanica*, and three more distant relatives, *S. tetrandra* (Figure 1A, Supplementary Table S2). Furthermore, we constructed a pan-NLRome including 186 NLR gene families in *Spinacia*, which were classified into six categories, consistent with the previous findings based on a single genome (Xu et al., 2017), but fewer subfamilies than in grape, which identified two additional subfamilies: TIR-CC-NBAARC-LRR and TIR (Guo et al., 2025).

*K*-mer-based GWAS serves as a powerful tool for identifying disease-resistance-associated genes and has been widely applied in various plants, such as tomato, maize (Voichek and Weigel, 2020), and wheat (Jaegle et al., 2025). Generally speaking, pan-genome captures missing heritability better than a single genome (Zhou et al., 2022). A recent study confirmed that pan-genome-based *k*-mer GWAS approaches can identify 25% more *k*-mers associated with powdery mildew resistance than single-reference methods (Jaegle et al., 2025). Using this reasoning, we developed a pipeline for rapidly identifying resistance-associated genes based on pan-NLRome and *k*-mer-based GWAS (Figure 3). This is because we extracted *k*-mers solely from paired-end (PE) reads within the NLR regions, utilized pan-NLRome to retrieve significantly associated *k*-mers, and ultimately directly obtained associated NLR genes. pan-NLRome *k*-mer-based GWAS rapidly identifies a candidate gene (*Te17S24XX\_Chr1\_nlr42*) for *RPF1*, while a single-genome-based approach detects four candidate genes, which require expression or functional analysis to determine the final gene, although both methods ultimately identify the same gene (Figure 4). Notably,

the candidate gene *Te17S24XX\_Chr1\_nlr42* was identified in the wild relative, *S. tetrandra*, indicating that the downy mildew resistance loci in cultivated spinach were introgressed from wild species, corroborating our earlier conclusions (She et al., 2024b).

As the number of sequenced genomes increases, we foresee that our approach will become highly prevalent for the rapid identification of disease resistance genes, as it directly targets the gene of interest rather than the linked region. However, our pipeline is specifically designed for NLR disease resistance genes; otherwise, candidate genes would be overlooked. For other gene types, we strongly recommend utilizing the pan-genome to retrieve the location of significant *k*-mers, as described by Jaegle et al. (2025). Overall, our study offers a novel approach for rapidly identifying disease-resistance genes, laying the foundation for further elucidating their mechanisms.

## Conclusions

We constructed the comprehensive pan-NLRome for the genus *Spinacia*, revealing an extensive diversity of 2,549 NLR genes. The pan-NLRome provides a crucial framework for understanding the evolution and architecture of the immune system in spinach and its wild relatives. Then, we developed a novel pipeline that integrates the pan-NLRome with *k*-mer-based GWAS. This strategy allows for the rapid and precise identification of resistance genes, overcoming the limitations of traditional single-genome methods that produce ambiguous candidate lists requiring lengthy subsequent validation. The power of this approach is demonstrated by the direct identification of *Te17S24XX\_Chr1\_nlr42* as the candidate for the downy mildew resistance locus *RPF1* in spinach. Our findings provide valuable resources for exploring NLR gene evolution and plant disease resistance mechanisms, while offering a reliable and efficient strategy for cloning resistance genes in a wide range of crops.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

## Author contributions

HS: Funding acquisition, Writing – original draft, Writing – review & editing. HW: Funding acquisition, Software, Supervision, Validation, Visualization, Writing – review & editing, Writing – original draft. ZX: Conceptualization, Resources, Validation, Writing – review & editing. HZ: Conceptualization, Resources, Validation, Writing – review & editing. ZL: Conceptualization, Investigation, Project administration, Resources, Supervision, Validation, Visualization, Writing – review & editing.

## Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by Beijing Natural Science Foundation (6244055), Fund of Beijing Key Laboratory of New Technique in Agricultural Application, Beijing University of Agriculture (KFKT-2025010), and the Scientific Research Team of Zhongyuan Research Center, Chinese Academy of Agricultural Sciences (CAAS-ZRC-ZYZX20230204).

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

## References

- Ameline-Torregrosa, C., Wang, B. B., O'Bleness, M. S., Deshpande, S., Zhu, H., Roe, B., et al. (2008). Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. *Plant Physiol.* 146, 5–21. doi: 10.1104/pp.107.104588
- Arora, S., Steuernagel, B., Gaurav, K., Chandramohan, S., Long, Y., Matny, O., et al. (2019). Resistance gene cloning from a wild crop relative by sequence capture and association genetics. *Nat. Biotechnol.* 37, 139–143. doi: 10.1038/s41587-018-0007-9
- Belinchon-Moreno, J., Berard, A., Canaguier, A., Chovelon, V., Cruaud, C., Engelen, S., et al. (2025). Nanopore adaptive sampling to identify the NLR gene family in melon (*Cucumis melo* L.). *BMC Genomics* 26, 126. doi: 10.1186/s12864-025-11295-5
- Bhattarai, G., Shi, A., Feng, C., Dhillon, B., Mou, B., and Correll, J. C. (2020). Genome wide association studies in multiple spinach breeding populations refine downy mildew race 13 resistance genes. *Front. Plant Sci.* 11, 563187. doi: 10.3389/fpls.2020.563187
- Bhattarai, G., Shi, A., Mou, B., and Correll, J. C. (2023). Skim resequencing finely maps the downy mildew resistance loci *RPF2* and *RPF3* in spinach cultivars whale and Lazio. *Hortic. Res-England* 10, uhad076. doi: 10.1093/hr/uhad076
- Cai, X. F., Sun, X. P., Xu, C. X., Sun, H. H., Wang, X. L., Ge, C. H., et al. (2021). Genomic analyses provide insights into spinach domestication and the genetic basis of agronomic traits. *Nat. Commun.* 12, 7246. doi: 10.1038/s41467-021-27432-z
- Chen, C. J., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y. H., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009
- Chen, J., Liu, Y., Liu, M., Guo, W., Wang, Y., He, Q., et al. (2023). Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. *Nat. Genet.* 55, 2243–2254. doi: 10.1038/s41588-023-01571-z
- Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, 884–890. doi: 10.1093/bioinformatics/bty560
- Correll, J., Bluhm, B., Feng, C., Lamour, K., Du Toit, L., and Koike, S. (2011). Spinach: better management of downy mildew and white rust through genomics. *Eur. J. Plant Pathol.* 129, 193–205. doi: 10.1007/s10658-010-9713-y
- Correll, J., Smilde, D., and Venancio, R. (2024). *Denomination of Pe: 20, a new race of downy mildew in spinach* (ANR Blogs).
- Danecek, P., Auton, A., Abecasis, G., Albers, C., Banks, E., DePristo, M., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330
- Dangl, J. L., Horvath, D. M., and Staskawicz, B. J. (2013). Pivoting the plant immune system from dissection to deployment. *Science* 341, 746–751. doi: 10.1126/science.1236011
- Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157. doi: 10.1186/s13059-015-0721-2
- Feng, C., Bluhm, B., Shi, A., and Correll, J. C. (2018a). Development of molecular markers linked to three spinach downy mildew resistance loci. *Euphytica* 214, 174. doi: 10.1007/s10681-018-2258-4
- Feng, C., Correll, J. C., Kammeijer, K. E., and Koike, S. T. (2014). Identification of new races and deviating strains of the spinach downy mildew pathogen *Peronospora farinosa* f. sp. *spinaciae*. *Plant Dis.* 98, 145–152. doi: 10.1094/PDIS-04-13-0435-RE
- Feng, C., Saito, K., Liu, B., Manley, A., Kammeijer, K., Mauzey, S. J., et al. (2018b). New races and novel strains of the spinach downy mildew pathogen *Peronospora effusa*. *Plant Dis.* 102, 613–618. doi: 10.1094/PDIS-05-17-0781-RE
- Gao, G., Lu, T., She, H., Xu, Z., Zhang, H., Liu, Z., et al. (2022). Fine mapping and identification of a candidate gene of downy mildew resistance, *RPF2*, in spinach (*Spinacia oleracea* L.). *Int. J. Mol. Sci.* 23, 14872. doi: 10.3390/ijms232314872
- Guo, L., Wang, X., Ayhan, D. H., Rhaman, M. S., Yan, M., Jiang, J., et al. (2025). Super pangenome of *Vitis* empowers identification of downy mildew resistance genes for grapevine improvement. *Nat. Genet.* 57, 741–753. doi: 10.1038/s41588-025-02111-7
- Hirakawa, H., Toyoda, A., Itoh, T., Suzuki, Y., Nagano, A. J., Sugiyama, S., et al. (2021). A spinach genome assembly with remarkable completeness, and its use for rapid identification of candidate genes for agronomic traits. *DNA Res.* 28, dsab004. doi: 10.1093/dnares/dsab004
- Holst, F., Bolger, A., Günther, C., Maß, J., Triesch, S., Kindel, F., et al. (2023). Helixer—*de novo* prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *BioRxiv*. doi: 10.1101/2023.02.06.527280
- Holub, E. B. (2001). The arms race is ancient history in *Arabidopsis*, the wildflower. *Nat. Rev. Genet.* 2, 516–527. doi: 10.1038/35080508
- Hulse-Kemp, A. M., Bostan, H., Chen, S. Y., Ashrafi, H., Stoffel, K., Sanseverino, W., et al. (2021). An anchored chromosome-scale genome assembly of spinach improves annotation and reveals extensive gene rearrangements in euasterids. *Plant Genome-Us* 14, e20101. doi: 10.1002/tpg2.20101
- Jaegle, B., Voicheck, Y., Haupt, M., Sotiropoulos, A. G., Gauthier, K., Heuberger, M., et al. (2025). k-mer-based GWAS in a wheat collection reveals novel and diverse sources of powdery mildew resistance. *Genome Biol.* 26, 172. doi: 10.1186/s13059-025-03645-z
- Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2026.1766206/full#supplementary-material>

- Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317
- Kokot, M., Długosz, M., and Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33, 2759–2761. doi: 10.1093/bioinformatics/btx304
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv 1303.3997*. doi: 10.48550/arXiv.1303.3997
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Li, Y., Wei, Z.-Z., Sela, H., Govta, L., Klymiuk, V., Roychowdhury, R., et al. (2024). Dissection of a rapidly evolving wheat resistance gene cluster by long-read genome sequencing accelerated the cloning of Pm69. *Plant Commun.* 5, 100646. doi: 10.1016/j.xplc.2023.100646
- Liu, Y. C., Du, H. L., Li, P. C., Shen, Y. T., Peng, H., Liu, S. L., et al. (2020). Pan-genome of wild and cultivated soybeans. *Cell* 182, 1–15. doi: 10.1016/j.cell.2020.05.023
- Liu, N., Lyu, X., Zhang, X., Zhang, G., Zhang, Z., Guan, X., et al. (2024). Reference genome sequence and population genomic analysis of peas provide insights into the genetic basis of Mendelian and other agronomic traits. *Nat. Genet.* 56, 1964–1974. doi: 10.1038/s41588-024-01867-8
- Ma, X. K., Yu, L. A., Fatima, M., Wadlington, W. H., Hulse-Kemp, A. M., Zhang, X. T., et al. (2022). The spinach YY genome reveals sex chromosome evolution, domestication, and introgression history of the species. *Genome Biol.* 23, 23–75. doi: 10.1186/s13059-022-02633-x
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110
- Meng, Q., Xie, P., Xu, Z., Tang, J., Hui, L., Gu, J., et al. (2025). Pangenome analysis reveals yield- and fiber-related diversity and interspecific gene flow in *Gossypium barbadense* L. *Nat. Commun.* 16, 4995. doi: 10.1038/s41467-025-60254-x
- Mo, C., Wang, H., Wei, M., Zeng, Q., Zhang, X., Fei, Z., et al. (2024). Complete genome assembly provides a high-quality skeleton for pan-NLRome construction in melon. *Plant J.* 118, 2249–2268. doi: 10.1111/tpj.16705
- Ning, W., Wang, W., Liu, Z., Xie, W., Chen, H., Hong, D., et al. (2024). The pan-NLRome analysis based on 23 genomes reveals the diversity of NLRs in *Brassica napus*. *Mol. Breed.* 44, 2. doi: 10.1007/s11032-024-01522-4
- Ning, W., Ye, Z., Ye, S., Xian, W., Zou, T., Liu, Z., et al. (2025). The pan-NLRome and diversity of Watkins wheat provide genetic resources for improving disease resistance and adaptation. *Plant Commun.* 6, 101563. doi: 10.1016/j.xplc.2025.101563
- Parada-Rojas, C., Childs, K., De Soto, M., Salcedo, A., Pecota, K., Yencho, G., et al. (2025). A reference-quality NLRome for the hexaploid sweetpotato and diploid wild relatives. *Mol. Plant-Microbe Interact.* 38, 978–992. doi: 10.1094/MPMI-03-25-0034-R
- Qian, W., Feng, C., Zhang, H., Liu, W., Xu, D., Correll, J., et al. (2016). First report of race diversity of the spinach downy mildew pathogen, *Peronospora effusa*, in China. *Plant Dis.* 100, 1248–1248. doi: 10.1094/PDIS-08-15-0847-PDN
- Ribera, A., Bai, Y., Wolters, A., Treuren, R. V., and Kik, C. J. E. (2020). A review on the genetic resources, domestication and breeding history of spinach (*Spinacia oleracea* L.). *Euphytica* 216, 48. doi: 10.1007/s10681-020-02585-y
- Shang, L., He, W., Wang, T., Yang, Y., Xu, Q., Zhao, X., et al. (2023). A complete assembly of the rice Nipponbare reference genome. *Mol. Plant* 16, 1232–1236. doi: 10.1016/j.molp.2023.08.003
- Shao, Z.-Q., Xue, J.-Y., Wu, P., Zhang, Y.-M., Wu, Y., Hang, Y.-Y., et al. (2016). Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat genes reveal three anciently diverged classes with distinct evolutionary patterns. *Plant Physiol.* 170, 2095–2109. doi: 10.1104/pp.15.01487
- She, H., Liu, Z., Li, S., Xu, Z., Zhang, H., Cheng, F., et al. (2023). Evolution of the spinach sex-linked region within a rarely recombining pericentromeric region. *Plant Physiol.* 193, 1263–1280. doi: 10.1093/plphys/kiad389
- She, H., Liu, Z., Xu, Z., Zhang, H., Wu, J., Cheng, F., et al. (2024a). Pan-genome analysis of 13 *Spinacia* accessions reveals structural variations associated with sex chromosome evolution and domestication traits in spinach. *Plant Biotechnol. J.* 22, 3102–3117. doi: 10.1111/pbi.14433
- She, H. B., Liu, Z. Y., Xu, Z. S., Zhang, H. L., Wu, J., Wang, X. W., et al. (2024b). Insights into spinach domestication from genome sequences of two wild spinach progenitors, *Spinacia turkestanica* and *Spinacia tetrandra*. *New Phytol.* 243, 477–494. doi: 10.1111/nph.19799
- She, H. B., Qian, W., Zhang, H. L., Liu, Z. Y., Wang, X. W., Wu, J., et al. (2018). Fine mapping and candidate gene screening of the downy mildew resistance gene *RPF1* in spinach. *Theor. Appl. Genet.* 131, 2529–2541. doi: 10.1007/s00122-018-3169-4
- Steuernagel, B., Witek, K., Krattinger, S. G., Ramirez-Gonzalez, R. H., Schoonbeek, H.-J., Yu, G., et al. (2020). The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* 183, 468–482. doi: 10.1104/pp.19.01273
- Van de Weyer, A.-L., Monteiro, F., Furzer, O. J., Nishimura, M. T., Cevik, V., Witek, K., et al. (2019). A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. *Cell* 178, 1260–1272. doi: 10.1016/j.cell.2019.07.038
- Voicheck, Y., and Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nat. Genet.* 52, 534–540. doi: 10.1038/s41588-020-0612-7
- Wang, Y., Pruitt, R. N., Nürnberger, T., and Wang, Y. (2022). Evasion of plant immunity by microbial pathogens. *Nat. Rev. Microbiol.* 20, 449–464. doi: 10.1038/s41579-022-00710-3
- Wang, C., Zhang, X., Fan, Y., Gao, Y., Zhu, Q., Zheng, C., et al. (2015). XA23 is an executor R protein and confers broad-spectrum disease resistance in rice. *Mol. Plant* 8, 290–302. doi: 10.1016/j.molp.2014.10.010
- Wu, Q., Cui, Y., Jin, X., Wang, G., Yan, L., Zhong, C., et al. (2022). The CC-NB-LRR protein BSRI from *Brachypodium* confers resistance to Barley stripe mosaic virus in gramineous plants by recognising TGB1 movement protein. *New Phytol.* 236, 2233–2248. doi: 10.1111/nph.18457
- Xu, C., Jiao, C., Sun, H., Cai, X., Wang, X., Ge, C., et al. (2017). Draft genome of spinach and transcriptome diversity of 120 *Spinacia* accessions. *Nat. Commun.* 8, 15275. doi: 10.1038/ncomms15275
- Yang, L., Smyth, G. K., and Wei, S. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656
- Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* 606, 527–534. doi: 10.1038/s41586-022-04808-9