Journal of Zhejiang University-SCIENCE B in press
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Research Article

https://doi.org/10.1631/jzus.B2600166

Development and application of a 20K liquid-phase SNP panel for genomics-assisted breeding in pea (*Pisum sativum* L.)

Yayuan DENG[3,4], Jiashun MIAO[1], Yan CHUN[1], Yan LI[5], Ting ZHAO[3], Na LIU[1,2]✉

[1]*Xianghu Laboratory, Hangzhou 311231, China*
[2]*State Key Laboratory for Quality and Safety of Agro-products, Ministry of Agriculture and Rural Affairs Key Laboratory of Vegetable Legumes Germplasm Enhancement and Molecular Breeding in Southern China, Institute of Vegetables, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China*
[3]*College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 300058, China*
[4]*Hainan Institute of Zhejiang University, Sanya 572025, China*
[5]*Station of Zhejiang Seed Management, Hangzhou 310020, China*

Abstract: Pea (*Pisum sativum* L., 2*n*=14), as one of the world's most important legume crops, ranks fourth in global legume production. The improvement of complex agronomic traits in peas relies heavily on genetic analysis and the identification and integration of functional loci for targeted breeding. Next-generation sequencing (NGS)-based genotyping is crucial for pea molecular breeding; however, the large and complex pea genome poses significant challenges to high-throughput genotyping. To overcome this difficulty, we integrated data from 314 large-scale pea resequencing projects and incorporated 235 functional loci associated with 57 agronomic traits, thus developed a 20K pea liquid-phase single-nucleotide polymorphism (SNP) panel. From an initial pool of approximately 400 million molecular markers, 21 659 high-quality SNPs were selected and evenly distributed across all chromosomes. Phylogenetic analysis of 44 pea accessions revealed strong concordance between the panel and whole-genome resequencing datasets in pairwise patristic distances (Mantel r = 0.832, P = 0.0001), supporting the panel's utility for germplasm characterization. The effectiveness of the panel was further validated through bulked segregant analysis (BSA) for pod color, which identified two loci on chromosome 3 spanning 225.5-367.1 Mb and 386.7-552.1 Mb. These regions were consistent with loci detected by whole-genome resequencing (WGR), demonstrating the reliability of the panel-based genotyping approach. Four machine learning algorithms were also applied to the panel-based genotypes to predict 30 agronomic traits. Notably, stem length prediction achieved a maximum Pearson correlation coefficient of 0.93. Overall, the 20K liquid-phase SNP panel offers an economical and efficient genotyping platform that has strong potential for marker-assisted selection and genomic prediction in modern pea breeding.

Key words: Pea (*Pisum sativum* L.); Single-nucleotide polymorphisms; Breeding panel; Genomic selection

unedited

1 Introduction

With the continued growth of the global population, food security has become an increasingly urgent issue, making sustained innovation in crop breeding technologies a key driver of agricultural sustainability. Reflecting this demand, breeding strategies for major crops such as rice and maize have evolved through several distinct

✉ Na LIU, naliu@zaas.ac.cn
Na LIU, https://orcid.org/0000-0000-0000-0001
Yayuan DENG, https://orcid.org/0009-0004-9629-3027
Jiashun MIAO, https://orcid.org/0000-0002-5227-5431
Yan CHUN, https://orcid.org/0000-0002-1765-3740
Ting ZHAO, https://orcid.org/0000-0001-5102-0157

Received Mar. 16, 2026; Revision accepted Oct. 26, 2026;
Crosschecked xxx. xx, 20xx

© Zhejiang University Press 2026

stages: early domestication and systematic selection (Breeding 1.0), conventional breeding (2.0), molecular breeding (3.0), and the current era of design-based or intelligent breeding (4.0) (Wallace et al., 2018; He B et al., 2024; Ansari et al., 2025). The ultimate goal of modern breeding is the rational design of an "optimal genome" that relies fundamentally on the availability of high-quality reference genomes and the precise identification of key functional loci. Rapid advances in next-generation sequencing and high-throughput genotyping technologies have enabled the assembly and public release of reference genomes for many major crops (Shang et al., 2023; Zhang et al., 2024; Hu et al., 2025; Wang ZJ et al., 2025). Building on these resources, large-scale resequencing of germplasm collections has expanded substantially. For example, genome sequencing has been completed for over 3000 maize accessions (Lv et al., 2020; Grzybowski et al., 2023), over 5000 rice accessions (Wang et al., 2018; Gutaker et al., 2020; Lv, et al., 2020), and over 6000 cotton accessions (He et al., 2021; Li et al., 2021; Ma et al., 2021). Moreover, population-scale epigenomic and metabolomic datasets have recently become available (Zhao et al., 2024; Wang P et al., 2025). In contrast to these major crops, however, publicly accessible genomic resources for less widely cultivated species such as pea and common bean (*Phaseolus vulgaris* L.) remain limited (Song et al., 2023).

Large genome sizes impose significant constraints on the cost and throughput of whole-genome resequencing (WGR), thereby hindering genome-based design breeding in many plant species. Crops with large genomes, including the faba bean (approximately 12 Gb) (Liu et al., 2025), barley (approximately 4 Gb) (Hu et al., 2023), pea (approximately 4 Gb) (Liu N et al., 2024), and oat (approximately 10 Gb) (He Q et al., 2024), face considerable challenges in large-scale genomic analysis. For instance, for the large genome size of wheat (approximately 15 Gb), resequencing 500 accessions at 10× coverage would require nearly 75 000 Gb of sequencing data. Beyond the sequencing data requirements, the large and repeat-rich genomes of Triticeae species, together with variation in ploidy, also present substantial challenges for de novo sequencing and assembly (Kuang et al., 2023). Pea, the species investigated in this study, possesses a genome of around 4 Gb and thus represent a typical large-genome crop. To overcome the above limitation, SNP genotyping panels have emerged as cost-effective and scalable alternatives to WGR. High-quality SNP panels typically integrate functional variants identified through genome-wide association studies (GWAS) with representative polymorphic markers evenly distributed across the genome, while they exclude SNPs located in regions of high linkage disequilibrium, repetitive sequences, other complex genomic regions. Such designs ensure both accuracy and representativeness in genotyping applications. For example, the Faba_bean_130K SNP array enabled the genotyping of 558 global faba bean accessions, revealing clear genetic structure and geographic differentiation (Liu, et al., 2025). Similarly, high-density SNP-based genomic resources in wheat have supported large-scale association analyses, including the identification of 431 yellow rust resistance loci and key resistance genes such *Yr5x*, *Yr6/Pm5* and *YrKB* (Wu et al., 2025).

Pea (*Pisum sativum* L.), originating from regions including the Mediterranean basin, Transcaucasia, Western Asia, Western Ethiopia, and Asia Minor, has been domesticated and cultivated for more than 6,000 years (Singh et al., 2019). Owing to its high nutritional value and health benefits, pea serves as an important source of human food, animal feed, as well as industrial raw materials, securing a prominent role in global agriculture (Zheng, 1997; Petropoulou et al., 2020). According to FAO statistics for 2022, peas were cultivated in 97 countries, covering 7.16 million hectares with a total production of 14.17 million tons. With the global demand for peas increasing, a central objective in pea breeding has been to achieve a balance between high yield and quality. As a result, significant progress has been made in the genetic improvement of key pea agronomic traits in recent years, including yield, disease resistance and environmental adaptability (Ellis et al., 2021; Sinjushin et al., 2022; Williams et al., 2022; Yang XM et al., 2022; Chen et al., 2024). Nevertheless, the exceptionally large and complex pea genome continues to pose a major obstacle for genetic research and breeding applications. The first chromosome-level pea reference genome, published in 2019, provided a crucial foundation for comparative genomics within legumes (Kreplak et al., 2019). Subsequently, the genome of the cultivar 'Zhongwan 6' was completed in 2022, revealing unique genomic features and evolutionary insights (Yang T et al., 2022). In 2024, a high-quality reference genome of the cultivar 'Zhewan No.1' was released

Uncorrected

alongside a comprehensive variation map derived from 314 accessions (Liu N et al., 2024). The publications of these high-quality reference genomes in recent years represent a transformative advance. More recently, the analysis of nearly 700 globally distributed pea accessions enabled the construction of a high-resolution phenotypic atlas and the dissection of the genetic architecture underlying 72 key agronomic traits, including the seven classical traits described by Mendel (Feng et al., 2025). Together, these advances have generated a robust genomic resource and large-scale resequencing datasets, providing a solid foundation for the development of efficient breeding tools.

Over the years, pea SNP genotyping has progressed from early GoldenGate/VeraCode assays to high-density fixed-content arrays (Deulvot et al., 2010; Sindhu et al., 2014). The GenoPea 13.2K SNP Array contained 13 204 transcript-derived SNPs and supported the construction of a high-density consensus genetic map (Tayeh N et al., 2015). More recently, the Axiom 90K array integrated SNPs from GenoPea, exome-capture data, and WGR, resulting in 84,691 successfully tiled marker features (Ellis et al., 2023). A multispecies Pulses 30K array containing approximately 10 000 field-pea markers has also been applied in pea breeding (Riaz et al., 2026). These arrays provide robust and standardized genotyping for genetic mapping, association analysis and genomic prediction; however, their fixed probe content and platform-specific workflows limit the rapid incorporation of newly discovered pangenome variants and breeding-relevant loci. Therefore, there is a demand for a modular liquid-phase SNP panel that can provide a flexible and complementary resource for pea genomics-assisted breeding. In this study, we developed such liquid-phase SNP panel for pea that not only adheres to established design principles but also integrates experimentally validated functional loci. It enables efficient genotyping, the precise identification of favorable alleles, and the effective utilization of elite genes, thereby enhancing breeding efficiency and accelerating the development of improved pea varieties.

2 Materials and methods

2.1 Selection of SNP loci for panel design

Whole-genome sequencing (WGS) data and genotype file in the variant call format (VCF) were obtained for 314 pea resequencing samples from our previous study (Liu N et al., 2024). This VCF contained genotype calls across accessions, and SNP selection was not restricted to homozygous alleles; both homozygous and heterozygous genotype calls were retained. To identify high-quality and representative markers suitable for panel construction, SNP loci were filtered using the following criteria: (1) loci with a genotype missing rate ≤ 20% were retained (vcftools --max-missing 0.8); (2) loci with minor allele frequency (MAF) < 0.01 were excluded (vcftools --maf 0.01); and (3) loci whose upstream and downstream 100-bp flanking sequences produced additional alignments against the pea reference genome using BWA-MEM (v.0.7.17, -a option) (Li, 2013) with >60% sequence identity over an alignment length of approximately 120 bp were considered duplicated or non-unique loci and excluded, leaving only single-copy loci. Additionally, loci with a GC content between 40% and 60% in the 100-bp flanking regions were selected, while those containing simple sequence repeats (SSR) sequences were excluded. SNPs with insertions and deletions (Indels) larger than 5 bp and loci with 'N' bases in the 100-bp flanking regions were also removed. Furthermore, 235 key SNP loci identified through genome GWAS were included. These loci are significantly associated with 57 important agronomic traits, including yield, quality, and resistance or tolerance traits, and these were derived from previously published datasets (Liu N et al., 2024). For completeness, these functional SNPs are listed in Table S1. After applying all filtering criteria, a total of 21 659 high-quality and representative SNP loci were retained, including the 235 functionally validated SNPs associated with the 57 agronomic traits (Table S2). Capture probes targeting these 21 659 SNP loci were synthesized and used for liquid-phase hybridization capture to construct the 20K SNP panel (Smartgenomics Technology Institute, Tianjin, China).

unedited

2.2 Evaluation of genotyping using the 20K liquid-phase SNP panel

To evaluate the quality of the selected SNPs, 14 representative pea accessions were genotyped using the 20K liquid-phase SNP panel. Genomic DNA was extracted from young leaf tissues using the cetyltrimethylammonium bromide (CTAB) method (Doyle et al., 1987). Library construction and sequencing were performed while following previously described protocols (Guan et al., 2024), with minor modifications. Briefly, DNA was fragmented into 200-300 bp sections and subjected to end repair and 3'-end adenylation, followed by ligation of sequencing adapters to both ends of the DNA fragments. Excess adapters were removed by magnetic bead purification, and pre-PCR amplification was performed to enrich the libraries. For hybridization capture, amplified libraries were purified using magnetic beads at a 1.8× bead ratio. To facilitate target-specific binding, hybridization buffer and biotinylated capture probes were added for overnight hybridization. After hybridization, streptavidin-coated magnetic beads were added followed by incubation at room temperature for 30 min to capture the biotinylated probe-DNA complexes. To remove nonspecifically bound DNA fragments, the beads were washed three times under stringent conditions at 50℃, followed by an 80% ethanol wash to eliminate residual hybridization buffer and impurities. Post-PCR amplification was subsequently performed, and the libraries were purified make them high-quality and sequencing-ready. The enriched target regions were sequenced on the DNBSEQ-T7 platform, generating raw reads in FASTQ format (Chen et al., 2018). To ensure high-quality data for the validation dataset, genotype calls at the targeted panel loci were further quality-filtered for downstream genetic distance and phylogenetic analyses. Raw reads were filtered by removing (1) adapter-contaminated reads, (2) reads with > 10% ambiguous bases ('N'), and (3) reads with > 50% of bases and a quality score below 10. This filtering step was used to generate a reliable genotype matrix for phylogenetic analysis and did not represent an additional marker-selection step for panel design or novel SNP discovery. The clean reads were aligned to the pea reference genome using the Burrows–Wheeler Aligner (BWA) software (Li, 2013). The resulting SAM files were converted to sorted BAM files using Samtools (v.1.9), with a 'q30' parameter to filter alignment quality (Li and Durbin, 2009). Raw SNPs were extracted using the HaplotypeCaller module of the Genome Analysis Toolkit (GATK, v.4.0) (McKenna et al., 2010). Hard filter was applied using the following criteria: SNPs with a mapping quality of ≤ 20.0, minimum sequencing depth ≤ 5, and maximum sequencing depth ≥ 200 were filtered out. Variants from all samples were merged into a single variant call file using GATK (McKenna, et al., 2010) . Further filtering to remove potential false-positive SNPs was conducted using VCFtools (v.0.1.16) (Danecek et al., 2011) according to following criteria: (1) excluding markers missing in over 20% of samples (genotype missing rate ≤ 20%) using --max-missing 0.8, and (2) excluding markers with a minor allele frequency (MAF) < 0.01 using --maf 0.01. This process resulted in high-throughput genotyping of all targeted SNP loci across the tested pea accessions. Finally, based on the filtered SNP dataset, a genetic distance matrix among 14 accessions was calculated using VCF2Dis (v.1.47) (Xu et al., 2025), and a phylogenetic tree was constructed on the basis of distance matrix using FastME (v.2.1.6.3) (Lefort et al., 2015).

2.3 Phylogenetic concordance analysis between the 20K SNP panel and WGR

SNP datasets from 20K SNP and WGR genotyping approaches were independently converted into multiple-sequence alignments using VCF2PHYLIP (Ortiz, 2019). Phylogenetic trees were reconstructed using IQ-TREE v2.0.7 (Minh et al., 2020). The resulting trees were midpoint-rooted and visualized using the R packages 'ape', 'phytools', and 'ggtree' (Revell, 2012; Yu et al., 2017; Paradis et al., 2019). Phylogenetic concordance between the two datasets was further evaluated by comparing pairwise patristic distances calculated using the 'cophenetic.phylo' function in ape. The Pearson's correlation coefficient and Spearman's rank correlation coefficient were calculated to assess linear and rank-based relationships, respectively, between the two sets of phylogenetic distances. The correlation between the corresponding distance matrices was further evaluated using a Mantel test implemented in the R package 'vegan', with 9999 permutations (Mantel, 1967; Oksanen et al., 2025).

Unedited

2.4 Bulked Segregant Analysis (BSA)

To validate the effectiveness of the 20K SNP panel for gene mapping, a combined strategy was employed that integrates Bulked Segregant Analysis (BSA) with panel-based genotyping. The pea accession ZWR12, which exhibits green pods, was used as the female parent, whereas ZWNJ004, which has yellow pods, served as the male parent. An F_2 segregating population was generated from this cross. From the F_2 population, 30 individuals displaying yellow pods and 30 displaying green pods were selected and pooled to form two bulks: the yellow pod bulk (WD_Y1) and the green pod bulk (WD_G1). Both bulks were genotyped using the 20K liquid-phase SNP panel as well as WGR for comparison. Clean reads obtained from both methods were aligned to the PeaZW1 reference genome using BWA software (Li and Durbin, 2009; Liu N et al., 2024). After coordinate sorting with SAMtools (Li H et al., 2009), SNPs calling was performed using HaplotypeCaller module of GATK v4.0 (McKenna, et al., 2010), followed by the variant filtering using predefined criteria. The R package 'QTLseqr' was used to calculate and plot the G-value (Mansfeld et al., 2018).

2.5 Genomic Selection (GS) for agronomic traits

The phenotypic data used for genomic selection were obtained from our previous study (Liu N et al., 2024). Briefly, a diverse pea accession panel was planted at the Yangdu research base of Zhejiang Academy of Agricultural Sciences, China (120.21551° E, 30.25308° N), on November 12 in both 2020 and 2021. Agronomic traits were recorded across these two experimental years, and five independent measurements were obtained for each trait. Mean phenotypic values were calculated and used for downstream analyses. In the present study, 30 agronomic traits with matched phenotypic and 20K panel genotype data were used for genomic selection model evaluation. For the GS analysis, genotype data at SNP loci included in the 20K panel were obtained from the resequencing VCF dataset by extracting the corresponding marker positions. This approach allowed us to evaluate the predictive potential of the selected 20K panel marker set in a population with matched phenotypic and genotypic data. To evaluate the predictive capacity of the SNP loci on the developed panel, four genomic prediction models were separately implemented: linear, ridge regression, Lasso regression, and elastic net regression model. All analyses were carried out in R. Genotype data were imported from the VCF file using the 'vcfR' package (Knaus et al., 2017), and genotype calls were converted to numeric marker dosages, with homozygous reference, heterozygous, and homozygous alternative genotypes coded as 0, 1 and 2, respectively; other genotype formats or missing calls were treated as missing values. For each trait, accessions with missing phenotypic records were removed before model fitting, and SNP markers containing missing genotype values in the retained accessions were excluded. To evaluate the performance of genomic selection, the prediction model was subjected to five-fold cross-validation. The folds were generated using the createFolds function in the 'caret' package (Kuhn, 2008). In each iteration, four folds were used as the training set and the remaining fold was used as the testing set. The linear model was fitted using the lm function. The ridge regression, LASSO and elastic net models were fitted using the 'glmnet' package with alpha values of 0, 1 and 0.5, respectively. For the glmnet-based models, the regularization parameter lambda was selected by internal cross-validation using 'cv.glmnet' (Friedman et al., 2010), and predictions were generated using 'lambda.min'. Model performance was assessed based on the Pearson correlation coefficient (PCC) between observed and predicted phenotypic values.

3 Results

3.1 Selection of SNPs and design of the 20K pea liquid-phase SNP panel

Building on recent advances in pea genomics, Liu et al. (2024) resequenced 237 pea accessions and integrated sequencing data from an additional 77 previously reported accessions (Table S3) (Yang T et al., 2022). This combined dataset comprises both wild and cultivated accessions, ensuring the broad representation of pea genetic diversity. GWAS identified 235 functional SNP loci significantly associated with one or more of

Unedited

57 agronomic traits, including yield, quality and resistance or tolerance traits (Table S1). On the basis of these findings, we developed a targeted strategy to construct a genotyping panel that captures genome-wide genetic variation while prioritizing breeding-relevant functional loci (Fig. 1). Following stringent filtering and selection criteria, a total of 21 659 high-quality SNP loci were retained and incorporated into the final 20K pea liquid-phase SNP panel (Table S2).

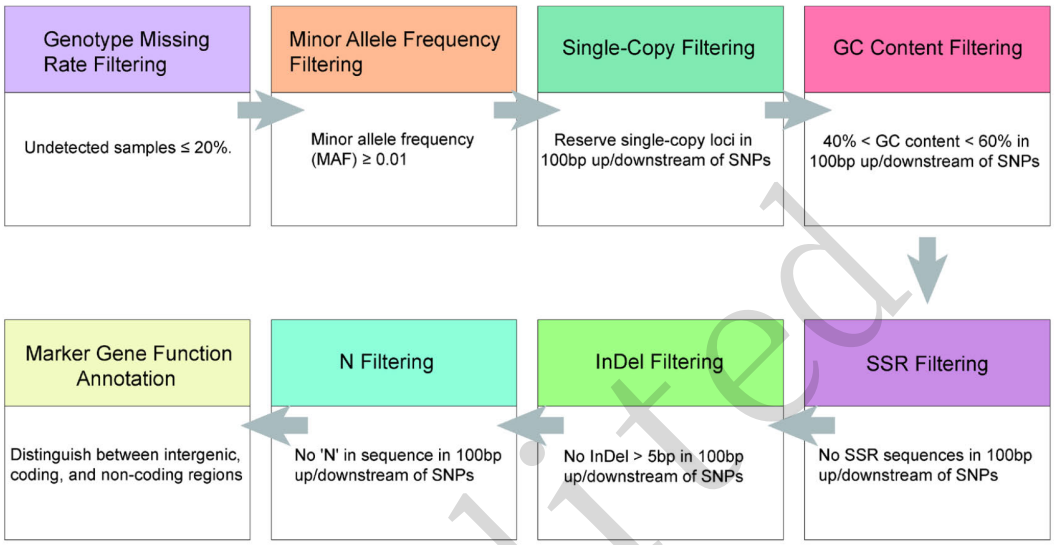


Fig. 1 Workflow for SNP loci selection and construction of the pea liquid-phase SNP panel.

3.2 Characteristics of the 20K pea liquid-phase SNP panel

The 21 659 SNP loci were evenly distributed across the pea genome, providing balanced representation across all chromosomes (Fig. 2a). The number of SNPs loci varied across different chromosomes, ranging from 2626 on chromosome 2 to 3629 on chromosome 5, and was positively correlated with chromosome length. The average inter-locus distance ranged from approximately 170 to 200 kb among chromosomes (Fig. 2b), confirming the uniform genomic coverage achieved by the panel. Regarding genomic annotation, 17 236 loci (79.6%) were located in intergenic regions, whereas 11.7% and 8.7% were positioned in exonic and intronic regions, respectively (Fig. 2c). In addition, the minor allele frequency (MAF) distribution of the selected loci closely mirrored that of the genome-wide SNP dataset, with no substantial deviations observed (Fig. 2d). This indicates that the panel effectively preserves natural allelic diversity present in the pea population. Collectively, these results demonstrate that the 20K pea liquid-phase SNP panel design can provide uniform and representative coverage across both structural and functional regions of the genome, supporting its reliability for downstream applications such as genetic diversity analyses, association mapping and molecular breeding.

Unedited

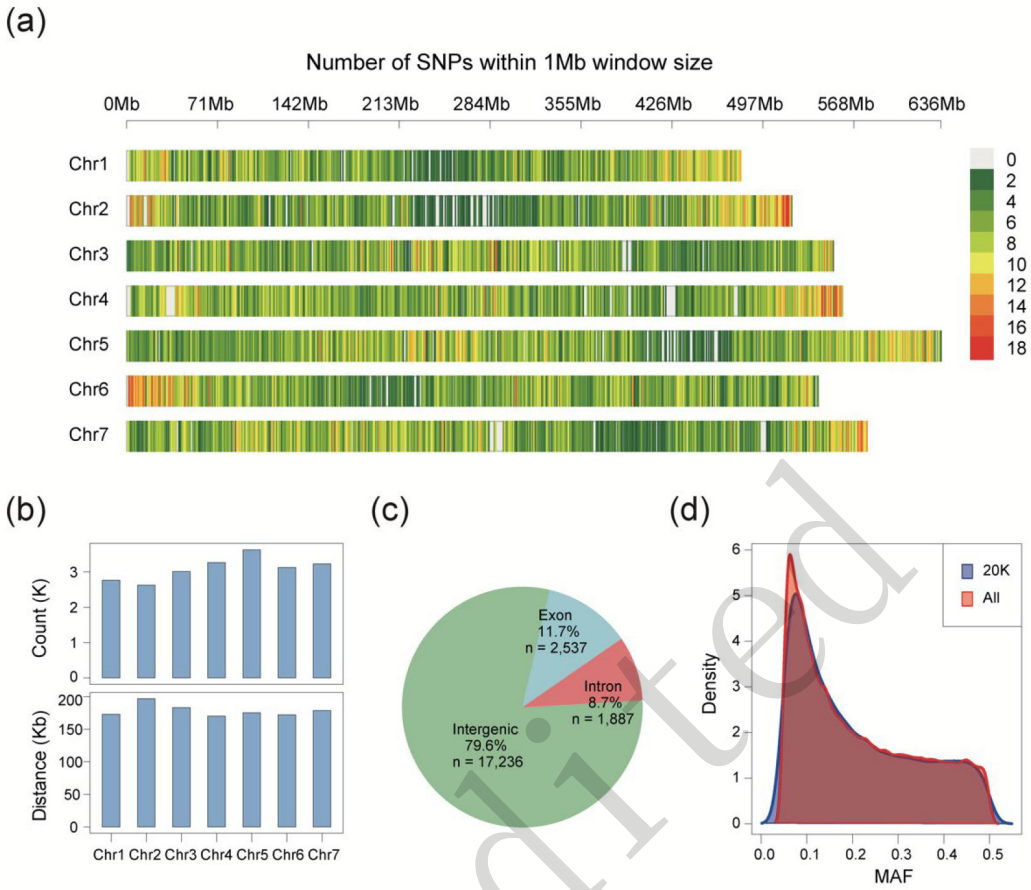


Fig. 2 Characteristics of the 20K pea liquid-phase SNP panel. (a) Distribution of panel SNP loci on the reference genome PeaZW1. (b) Number of panel SNPs on each pea chromosome (top) and the average distance between adjacent SNPs (Kb) (bottom). (c) Distribution ratio of panel SNPs in different genomic regions (intergenic regions, exons, introns). (d) Density plot of MAF values for panel SNPs and total SNPs.

unedited

3.3 Validation of 20K pea liquid-phase probes

To evaluate the performance of the 20K pea liquid-phase SNP panel, we conducted genotyping on 12 representative pea accessions using the designed probe set. Detailed information about these accessions is given in Table S4. Sequencing generated a total of 46.81 Mb of high-quality clean reads, with an average of 3.34 Mb per accession. The overall sequencing quality was exceptionally high, with Q20 values exceeding 98% and Q30 values above 95% across all samples (Table S5), indicating excellent base-calling accuracy.

The percentage of reads mapped to the PeaZW1 reference genome was high across the 14 tested samples with an average of 98.46%, indicating reliable sequencing and alignment quality. We further evaluated the probe capture performance using target-region-based metrics. The average sequencing depth of the target regions was 80.07× (Supplementary Fig. S1). Subsampling analysis showed that the SNP detection rate reached 90% at a mean target-site sequencing depth of 49.3× and 95% at 61.6× (Supplementary Fig. S2). The coverage rate of probe-targeted regions ranged from 91.18% to 98.17%, with an average of 93.86%. The capture efficiency, calculated based on reads covering the probe-targeted regions, ranged from 42.11% to 48.32%, with an average of 44.47% (Table S6). These results indicate that the 20K pea liquid-phase SNP panel achieved stable and effective enrichment of the designed target regions. Genotype concordance between the panel and whole-genome resequencing increased under more stringent filtering criteria and reached approximately 93–98% at DP ≥ 10 and GQ ≥ 20 across accessions with paired data (Supplementary Fig. S3 and Table S6). The

MAF distribution of the detected SNPs revealed that the majority of loci exhibited MAF values between 0.0 and 0.1 (Fig. 3a), consistent with expectation for a diverse set of germplasm accessions. To assess the transferability of the 20K SNP panel across different genetic backgrounds, we further examined the MAF distribution of panel SNPs in the broader resequencing population. When accessions were stratified by geographic origin and cultivar type, the mean MAF values were 0.182 for East improved cultivars, 0.162 for West improved cultivars, and 0.173 for landraces. Similarly, when accessions were stratified by breeding/use classification, the mean MAF values were 0.187, 0.180, 0.132, and 0.149 for the grain pea, vegetable-pod pea, vegetable-seed pea, and *afila* types, respectively. These results show that a substantial proportion of panel SNPs remained polymorphic across different accession groups, supporting the potential cross-population applicability of the panel (Supplementary Fig. S4 and Supplementary Fig. S5). To assess genetic relationships among the genotyped accessions, we constructed a phylogenetic tree using the Neighbor-Joining method (Fig. 3b). The tree exhibited a clustering pattern that was largely consistent with consumption types. Pea accessions cultivated for seed consumption (Zhewan No.2, ZW72, Changshouren, and Zhewan No.1) clustered together; within this group, ZW72 was classified as dry pulse/grain types, whereas the remaining accessions represented vegetable-seed types intended for fresh consumption. Accessions cultivated for shoot consumption, such as Zhushahong and Dawuxu, formed a separate cluster. In addition, accessions cultivated for pod consumption, including Qizhen No.9, 604, Shijiadacaiwan, and ZWR12, grouped together, reflecting their shared breeding usage and genetic background.

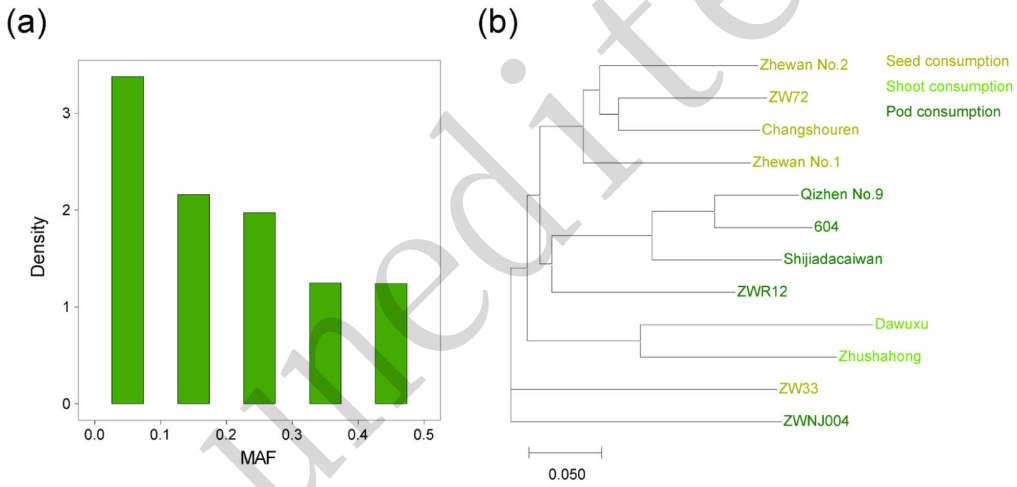


Fig. 3 Validation of the 20K pea liquid-phase SNP panel. (a) Density distribution of minor allele frequency of the detected SNP loci across 14 pea accessions. (b) The Neighbor-Joining phylogenetic tree was reconstructed using the panel genotyping results of the 12 representative pea accessions.

3.4 Phylogenetic concordance between the 20K SNP panel and WGR

To further validate the panel, we extended the analysis to 44 pea accessions with both panel genotyping and whole-genome resequencing (WGR) data. Phylogenetic analyses of the two datasets revealed broadly consistent patterns of genetic relationships among these accessions (Fig. 4a) (Supplementary Fig. S5). The two phylogenetic trees exhibited similar clustering patterns for several groups of closely related accessions. In particular, P035 and P021, P152 and P146, P091 and P090, P234 and P247, P047 and P048, and P042 and P050 exhibited consistent clustering relationships across the two datasets. Nevertheless, some differences in local tree topology were observed, particularly for accessions such as P094, P044 and P167. These observations suggest that the 20K SNP panel largely reproduced the genetic relationships inferred from WGR data, although the resolution of certain phylogenetic relationships differed between the two genotyping approaches.

unedited

To quantitatively assess the concordance between the two datasets, pairwise patristic distances were calculated for all 946 pairs of accessions. A strong positive correlation was observed between the patristic distances inferred from the 20K SNP panel and WGR data (Pearson's r = 0.832; Spearman's ρ = 0.794; Fig. 4b). Mantel analysis further revealed a significant correlation between the two phylogenetic distance matrices (Mantel r = 0.832, *P* = 0.0001). Although some variation in pairwise phylogenetic distances remained, these results demonstrate that the 20K liquid-phase SNP panel effectively captured the major patterns of genetic relationships revealed by WGR, supporting its potential application in genetic diversity assessment and germplasm characterization.

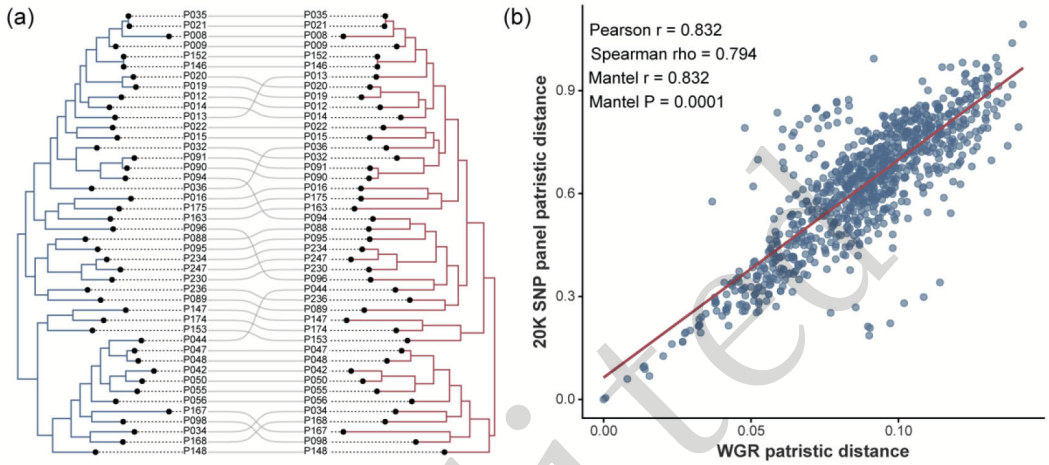
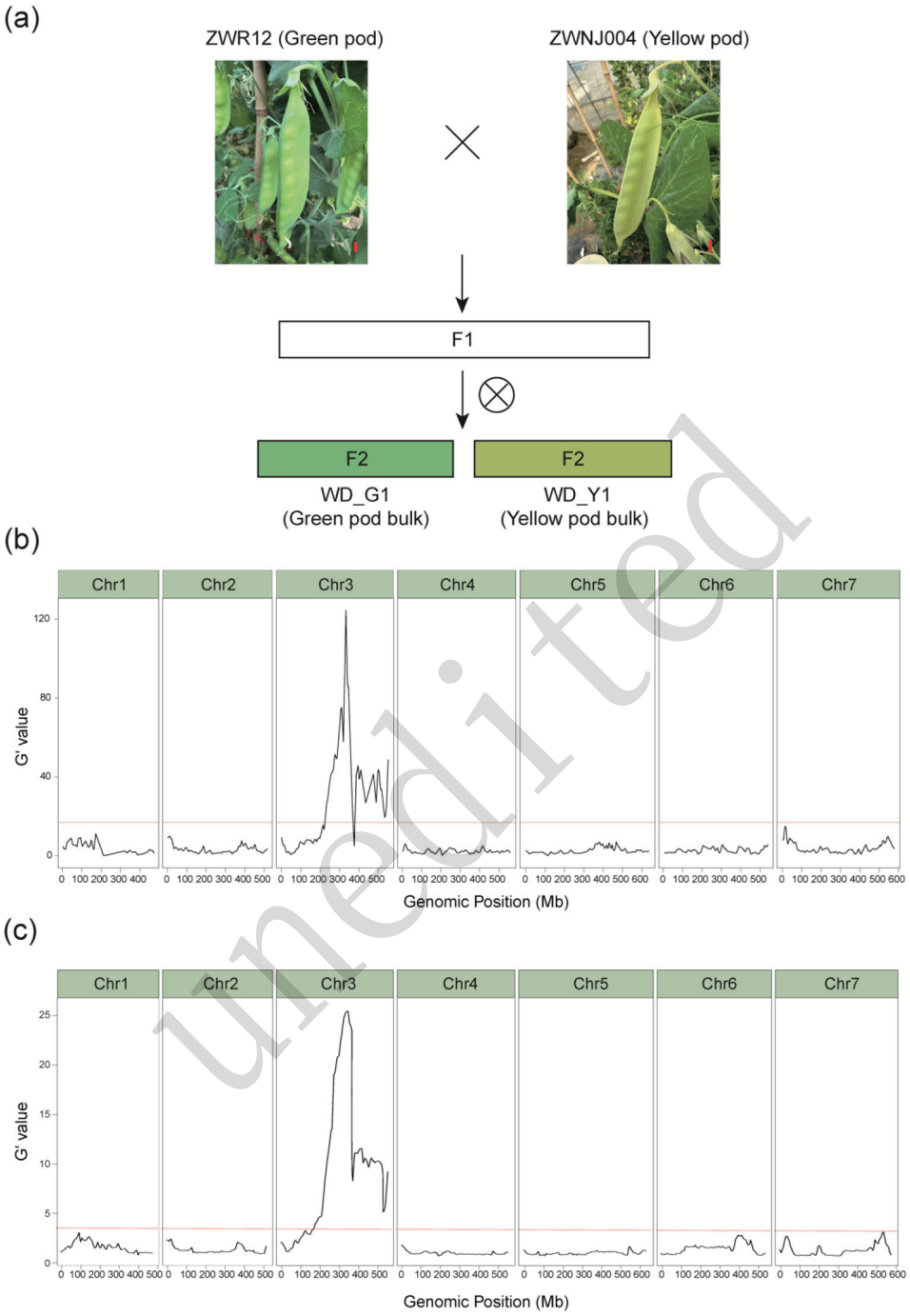


Fig. 4 Phylogenetic concordance between the 20K liquid-phase SNP panel and whole-genome resequencing (WGR) data. (a) Tanglegram comparing phylogenetic trees inferred from the 20K SNP panel (left, blue) and WGR (right, red) datasets. Gray curves connect corresponding accessions. (b) Correlation of pairwise patristic distances between the two datasets. Each point represents an accession pair, and the red line indicates the fitted linear regression. The Pearson's, Spearman's, and Mantel correlation coefficients are shown in the plot.

3.5 Effectiveness of the 20K pea liquid-phase SNP panel in gene mapping

To assess the effectiveness of the 20K pea liquid-phase SNP panel for gene mapping, we evaluated its performance using BSA for pod color. The female parent, ZWR12, exhibited green pods, whereas the male parent, ZWNJ004, displayed yellow pods (Fig. 5a). From the resulting F_2 segregating population, 30 individuals with yellow pods were selected to construct the yellow pod bulk (WD_Y1), and 30 individuals with green pods were chosen to form the green pod bulk (WD_G1). Both bulks were genotyped using the 20K SNP panel. Polymorphic probes from these pools were primarily located on chromosome 3, particularly within the genomic intervals 225 461 652-367 063 164 bp and 386 708 349-552 070 089 bp (Fig. 5b). These intervals corresponded with the candidate genomic coordinates identified by using WGR data of the same populations (Fig. 5c) and included the previously reported *Gp*-associated region. Independent functional evidence has established the chlorophyll synthase gene *ChlG* as the gene corresponding to the classical *Gp* locus (Feng et al., 2025). These results demonstrate the effectiveness of the panel in rapidly identifying genomic regions associated with important agronomic traits and underscore its potential utility for gene discovery and molecular breeding in pea.

unedited



unedited

Fig. 5 Validation of gene mapping using the 20K pea liquid-phase SNP panel in combination with BSA. (a) For the BSA analysis, ZWR12 (green pod, left) was used as the female parent and ZWNJ004 (yellow pod, right) as the male parent. Scale bar = 1 cm. (b)The G-value was calculated using the panel-based genotyping data from two bulks: WD_Y1 (yellow pod) and WD_G1 (green pod). (c) WGR of WD_Y1 and WD_G1 was conducted, and G-values were calculated to validate the accuracy and effectiveness of gene mapping by the developed 20K pea liquid-phase SNP panel.

3.6 Genomic Selection (GS) model based on 20K panels

To assess the predictive performance of the 20K pea panel in genomic selection (GS), we employed four widely used GS models to predict 30 agronomic traits, including 6 qualitative and 24 quantitative traits. Phenotypic data for the genomic selection were derived from our previous two-year field evaluation of pea agronomic traits (Liu N et al., 2024). The models tested included the linear model, ridge regression model, least absolute shrinkage and selection operator (LASSO) regression model, and elastic net regression model.

As shown in Fig. 6, the ridge regression, Lasso regression and elastic net regression models consistently exhibited high predictive accuracy across multiple traits. These models performed particularly well for several qualitative traits, such as the number of seeds per pod (*PCC* = 0.57) and the node position of the first pod (*PCC* = 0.61). Notably, the models achieved the highest predictive accuracy between predicted and observed values for stem length (*PCC* = 0.87) and seedling stem length (*PCC* = 0.93), reflecting the capacity of the panel to capture genetic variance effectively for these growth-related traits. The high prediction accuracy for stem length and seedling stem length likely reflects a comparatively simple genetic architecture for plant height in pea. Stem elongation in pea is a classical Mendelian-derived architectural trait, where the *Le* locus plays a dominant role. *Le* encodes a GA 3β-hydroxylase (*GA3OX*) that controls the formation of bioactive gibberellins and thereby internode elongation (Lester et al., 1997; Martin et al., 1997; Shani et al., 2024); functional variation at this locus produces the well-known plant height phenotype that is a quasi-Mendelian plant height trait. Under such a scenario, a medium–high density marker set can include markers in strong LD with key causal genes, which substantially boosts the performance of linear GS models (Crossa et al., 2017). Consistently, recent genomic analyses of Mendel's pea traits further support the presence of major alleles underlying classical morphological variation (Feng et al., 2025).

unedited

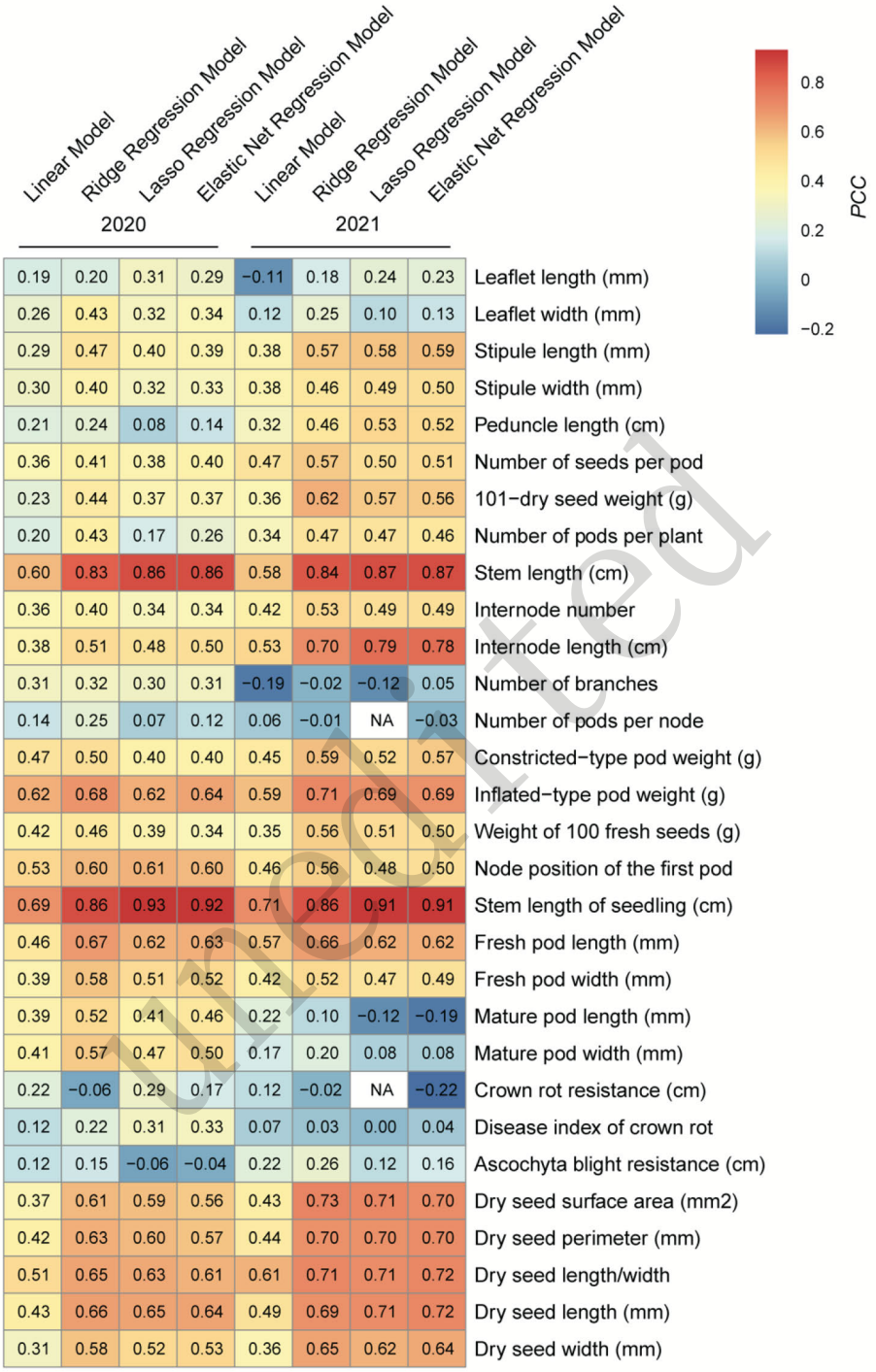


Fig. 6 Pearson correlation coefficient (*PCC*) values for evaluating the performance of four models across 30 agronomic traits. The genotype data was generated based on the 20K panel data. The phenotypic data of these 30 traits were measured in both years 2020 and 2021.

Overall, our results demonstrate that the 20K SNP panel, when integrated with appropriate GS models, enables robust and accurate predictions of both qualitative and quantitative traits. Strategic selection and optimization GS models can therefore substantially improve prediction performance, offering valuable genetic

insights and practical support for breeding programs aimed at developing elite pea cultivars.

4 Discussion

Identifying and efficiently utilizing genetic variation in pea is essential for advancing molecular breeding. Previous studies have reported 17 212 424 high-quality SNPs across 43 pea accessions and one grass pea accession (Kreplak, et al., 2019). A broader survey of 118 pea accessions identified 26 250 039 high-quality SNPs, of which 64.1% were located in intergenic regions and only 2.4% in exonic regions (Yang T et al., 2022). With advances in genome assembly and large-scale resequencing, it is now evident that the pea genome, approximately 4 Gb in size, harbors more than 400 million variants in cultivated populations, corresponding to an average density of 7.35 high-quality SNPs per kilobase (Liu N et al., 2024). While these extensive genomic resources considerably enhance our understanding of pea genetic diversity, they also substantially increase both genotyping and computational costs, highlighting the need for efficient and cost-effective genotyping strategies.

At approximately 20× coverage, the whole-genome sequencing of pea costs approximately 500 to 1000 CNY per sample when both sequencing and computational expenses are considered. In contrast, the liquid-phase SNP panel developed in this study reduces per-sample cost to approximately 100 CNY, providing a substantially more economical solution for large-scale genotyping without compromising accuracy. This high cost-effectiveness is particularly important for large breeding populations.

SNP panels are widely used in genetic diversity analysis, genome-wide association studies, and genomic selection (Gali et al., 2024; Guan et al., 2024; Liu S et al., 2024). Unlike fixed-content solid-phase arrays, solution-based panels allow marker content to be modified and expanded more readily as new loci and breeding objectives emerge. For pea, the GenoPea 13.2K Infinium array established an important foundation for high-throughput genotyping by providing 13,204 transcript-derived SNPs discovered from 16 diverse *Pisum* genotypes, 12 802 of which were incorporated into a high-density consensus map (Tayeh et al., 2015a). The array has also supported genomic-prediction studies in pea (Tayeh et al., 2015b). Building on these resources, the present panel extends marker discovery to resequencing data from 314 accessions (Liu N et al., 2024) and combines genome-wide markers with 235 GWAS-associated SNPs representing 57 agronomic traits (Table S1). Therefore, its broader discovery base, breeding-oriented marker content, and solution-based format provide a complementary and readily updateable platform for current pea genetic research and breeding.

The capture performance of the present pea panel was compared with that of liquid-phase target-capture platforms developed for other crops. For example, the Faba_bean_130K platform reported a mapping rate of 93.14%, an on-target read rate of 53.23%, and 93.6% concordance with Sanger sequencing (Wang et al., 2021). In allotetraploid peanut, a transcript-derived target-enrichment panel achieved an average on-target read rate of 42.47% and an 85.71% validation rate for homozygous SNPs, with off-target reads largely attributed to genomic regions homologous to the intended targets (Peng et al., 2017). The 44.47% on-target rate of the present pea panel therefore falls within the range reported for target-capture genotyping in complex plant genomes. Moreover, the high mapping rate (98.46%), target-region coverage (93.86%), and mean target depth (80.07×) demonstrate stable recovery of the intended loci in the large and repeat-rich pea genome. The wheat GBW16K panel provides a complementary comparison in terms of application: it targeted 14,868 genomic regions, detected 37,669 multi-SNPs in 239 accessions, and successfully supported germplasm analysis, linkage-map construction, and QTL mapping (Liu S et al., 2024). Together, these comparisons support the application of medium-density liquid-phase SNP panels in genome-wide genotyping and preliminary trait mapping in crops with complex genomes.

The experimental validation of probe capture and genotyping performance was conducted using 14 representative accessions, and wild germplasm was not included in the capture-validation experiment. Therefore, although the stratified MAF analysis supports the potential transferability of the panel across different cultivated pea groups, further experimental validation using a broader set of wild and exotic

Under review

germplasm can further assess its applicability across the broader *Pisum* gene pool. Furthermore, validation using 44 pea accessions demonstrated that the panel largely recapitulated the genetic relationships inferred from whole-genome resequencing, with a strong correlation between pairwise patristic distances (Mantel r = 0.832, P = 0.0001). In the BSA validation, the 20K SNP panel and WGR produced broadly concordant signals on chromosome 3 for a pod-color trait whose underlying genetic locus has been previously identified and experimentally validated, supporting the reliability of the panel for rapid locus screening (Feng et al., 2025). However, the candidate intervals obtained in this study were relatively broad, which likely reflects the lower marker density of the 20K panel compared with WGS, the limited number of individuals included in each bulk, and the limited recombination events captured in the F_2 population. Therefore, the present BSA analysis demonstrates the feasibility of using the panel for cost-effective preliminary mapping. The panel is therefore particularly suitable for initial BSA screening and subsequent large-scale genotyping of segregating populations. Further genotyping of 200–300 individual F_2 plants, together with locally developed markers or higher-density genotyping, could help narrow the candidate intervals and strengthen its application in marker-assisted breeding. A comparable strategy was applied in wheat, where the GBW16K panel was used for linkage-map construction and preliminary QTL detection, while higher-density SNP data and KASP markers supported subsequent candidate-region analysis and marker validation (Liu S et al., 2024). Moreover, the flexible liquid-phase format allows future versions of the pea panel to incorporate additional trait-associated markers, rare alleles, and pangenome-derived variants as new genomic resources become available.

Modern pea breeding objectives have expanded to address diverse end uses, including food, vegetable, feed, and green manure, alongside increasing demands for both yield and quality (Piotrowska-Dlugosz et al., 2020; Dhillon et al., 2022; Zhou et al., 2023). In vegetable pea, phenotype-based core collection strategies have been developed to retain representative agronomic diversity while reducing the number of accessions requiring evaluation (Xu et al., 2019). Combining such phenotypic evaluations with genome-wide genotyping could facilitate the characterization and selection of breeding materials. The liquid-phase SNP panel presented here provides a strong technical foundation for the development of elite germplasm and the selection of new pea varieties. By enabling efficient, accurate, and cost-effective genotyping, this resource has the potential to accelerate pea breeding programs, meet market demand for high-quality cultivars, and contribute to sustainable agricultural development.

5 Conclusions

In this study, we developed a 20K liquid-phase SNP panel comprising 21,659 SNPs selected from resequencing data of 314 pea accessions, including 235 trait-associated loci representing 57 agronomic traits. The panel provided genome-wide marker coverage and achieved stable target enrichment across the tested accessions, with a mean reference-genome alignment rate of 98.46%, an on-target read rate of 44.47%, target-region coverage of 93.86%, and a mean target depth of 80.07×. Under stringent filtering criteria, panel-derived genotypes showed approximately 93%–98% concordance with whole-genome resequencing (WGR) data. In 44 accessions, the panel largely reproduced the genetic relationships inferred from WGR, as supported by the strong correlation between pairwise phylogenetic distances (Mantel r = 0.832, P = 0.0001). Panel-based bulked segregant analysis also recovered chromosome 3 signals consistent with those detected by WGR, supporting its application in preliminary locus screening. In addition, genotypes corresponding to the panel loci extracted from existing resequencing data showed predictive potential for 30 agronomic traits, with a maximum cross-validation Pearson correlation coefficient of 0.93.

Together, these findings support the 20K liquid-phase SNP panel as a flexible and economical resource for pea germplasm characterization, preliminary trait mapping, and high-throughput genotyping. Nevertheless, the current experimental validation was based primarily on cultivated accessions, the BSA candidate intervals remained broad, and the genomic prediction results were derived from in silico-extracted genotypes and internal

unedited

cross-validation rather than independent populations genotyped with the physical panel. Future validation using more diverse wild and exotic germplasm, larger segregating populations, and independent genomic prediction populations will be necessary to further establish the panel's utility in marker-assisted and genomics-assisted pea breeding.

Data availability statement

All data supporting the findings of this study are included in the article and Supplementary Materials. The complete list of SNP loci comprising the 20K pea liquid-phase SNP panel is provided in Table S2. Data presented in Tables S1 and S3 were originally published in Liu N et al. (2024), and are reproduced here with appropriate citation. Additional information is available from the corresponding author upon reasonable request.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32522092), Zhejiang Provincial Natural Science Foundation of China (LR25C150001), the Construction Funding of Xianghu Laboratory, Special Cooperation Program between Xianghu Laboratory and Chinese Academy of Agricultural Science (2025X1ZN1Z011), and Zhejiang Province Crop Germplasm Resources Precision Identification and Evaluation Project (2025JZJD003).

Author contributions

Na LIU and Ting ZHAO conceived the project and designed the study. Yayuan DENG, Ting ZHAO, Na LIU, Yan CHUN, Yan LI and Jiashun MIAO collected samples, performed experiments, and data analyses. Yayuan DENG and Ting ZHAO wrote the manuscript. Yayuan DENG, Jiashun MIAO and Na LIU revised the manuscript. All authors read and approved the manuscript.

Compliance with ethics guidelines

The authors declare no competing interests. This article does not contain any studies with human or animal subjects performed by any of the authors.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT (OpenAI) for English-language polishing, and improving formatting and readability. All AI-generated output was critically reviewed, verified, and edited by the authors to ensure accuracy, originality, and compliance with journal standards. The authors take full responsibility for the final content of the manuscript as submitted.

References

Ansari R, Manna A, Hazra S, et al., 2025. Breeding 4.0 vis-à-vis application of artificial intelligence (AI) in crop improvement: an overview. *New Zealand Journal of Crop and Horticultural Science*, 53(4):812-854. https://doi.org/10.1080/01140671.2024.2409775

Chen BZ, Shi Y, Sun YC, et al., 2024. Innovations in functional genomics and molecular breeding of pea: exploring advances and opportunities. *aBIOTECH*, 5(1):71-93. https://doi.org/10.1007/s42994-023-00129-1

Chen SF, Zhou YQ, Chen YR, et al., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):884-890. https://doi.org/10.1093/bioinformatics/bty560

Crossa J, Pérez-Rodríguez P, Cuevas J, et al., 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, 22(11):961-975. https://doi.org/10.1016/j.tplants.2017.08.011

Danecek P, Auton A, Abecasis G, et al., 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Deulvot C, Charrel H, Marty A, et al., 2010. Highly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea. *BMC Genomics*, 11(1):468. https://doi.org/10.1186/1471-2164-11-468

Dhillon LK, Lindsay D, Yang TY, et al., 2022. Biological nitrogen fixation potential of pea lines derived from crosses with nodulation mutants. *Field Crops Research*, 289: 108731. https://doi.org/10.1016/j.fcr.2022.108731

Doyle JJ and Doyle JL, 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical bulletin*, 19(1): 11-15.

Ellis N, Hofer J, Sizer-Coverdale E, et al., 2023. Recombinant inbred lines derived from wide crosses in Pisum. *Scientific Reports*, 13(1):20408.

unedited

https://doi.org/10.1038/s41598-023-47329-9
Ellis THN, Hofer JMI, Vikeli E, et al., 2021. Diversity of Pod Shape in *Pisum*. *Diversity-Basel*, 13(5): 203. https://doi.org/10.3390/d13050203
Feng C, Chen BZ, Hofer J, et al., 2025. Genomic and genetic insights into Mendel's pea genes. *Nature*, 642(8069): 980-989. https://doi.org/10.1038/s41586-025-08891-6
Friedman J, Hastie T and Tibshirani R, 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1-22. https://doi.org/10.18637/jss.v033.i01
Gali KK, Jha A, Tar'an B, et al., 2024. Identification of QTLs associated with seed protein concentration in two diverse recombinant inbred line populations of pea. *Frontiers in Plant Science*, 15: 1359117. https://doi.org/10.3389/fpls.2024.1359117
Grzybowski MW, Mural RV, Xu G, et al., 2023. A common resequencing-based genetic marker data set for global maize diversity. *The Plant Journal*, 113(6):1109-1121. https://doi.org/10.1111/tpj.16123
Guan H, Lu Y, Li X, et al., 2024. Development of a MaizeGerm50K array and application to maize genetic studies and breeding. *The Crop Journal*, 12(6): 1686-1696. https://doi.org/10.1016/j.cj.2024.09.014
Gutaker RM, Groen SC, Bellis ES, et al., 2020. Genomic history and ecology of the geographic spread of rice. *Nature Plants*, 6(5):492-502. https://doi.org/10.1038/s41477-020-0659-6
He B, Pan SS, Zhao JF, et al., 2024. Maize Improvement Based on Modern Breeding Strategies: Progress and Perspective. *Acs Agricultural Science & Technology*, 4(3):274-282. https://doi.org/10.1021/acsagscitech.3c00427
He Q, Li W, Miao YQ, et al., 2024. The near-complete genome assembly of hexaploid wild oat reveals its genome evolution and divergence with cultivated oats. *Nature Plants*, 10(12): 2062-2078. https://doi.org/10.1038/s41477-024-01866-x
He SP, Sun GF, Geng XL, et al., 2021. The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. *Nature Genetics*, 53(6):916-924. https://doi.org/10.1038/s41588-021-00844-9
Hu HF, Wang PH, Angessa TT, et al., 2023. Genomic signatures of barley breeding for environmental adaptation to the new continents. *Plant Biotechnology Journal*, 21(9):1719-1721. https://doi.org/10.1111/pbi.14077
Hu Y, Han JL, Jin SK, et al., 2025. Post-polyploidization centromere evolution in cotton. *Nature Genetics*, 57(4): 1021-1030. https://doi.org/10.1038/s41588-025-02115-3
Knaus BJ and Grünwald NJ, 2017. vcfR: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, 17(1):44-53. https://doi.org/10.1111/1755-0998.12549
Kreplak J, Madoui MA, Cápal P, et al., 2019. A reference genome for pea provides insight into legume genome evolution. *Nature Genetics*, 51(9):1411-1422. https://doi.org/10.1038/s41588-019-0480-1
Kuang L, Li Q and Zhang G, 2023. Advances on whole genome sequencing in Triticeae. *Journal of Zhejiang University (Agriculture & Life Sciences)*, 49(1):1-13. https://doi.org/10.3785/j.issn.1008-9209.2022.05.161
Kuhn M, 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5):1-26. https://doi.org/10.18637/jss.v028.i05
Lefort V, Desper R and Gascuel O, 2015. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Molecular Biology and Evolution*, 32(10):2798-2800. https://doi.org/10.1093/molbev/msv150
Lester DR, Ross JJ, Davies PJ, et al., 1997. Mendel's stem length gene (Le) encodes a gibberellin 3 beta-hydroxylase. *Plant Cell*, 9(8):1435-1443. https://doi.org/10.1105/tpc.9.8.1435
Li H, 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997*. https://doi.org/10.48550/arXiv.1303.3997
Li H and Durbin R, 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754-1760. https://doi.org/10.1093/bioinformatics/btp324

unedited

Li H, Handsaker B, Wysoker A, et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li J, Yuan D, Wang P, et al., 2021. Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biology*, 22(1):119. https://doi.org/10.1186/s13059-021-02351-w

Liu N, Lyu X, Zhang XY, et al., 2024. Reference genome sequence and population genomic analysis of peas provide insights into the genetic basis of Mendelian and other agronomic traits. *Nature Genetics*. https://doi.org/10.1038/s41588-024-01867-8

Liu R, Hu CQ, Gao D, et al., 2025. A special short-wing petal faba genome and genetic dissection of floral and yield-related traits accelerate breeding and improvement of faba bean. *Genome Biology*, 26(1): 62. https://doi.org/10.1186/s13059-025-03532-7

Liu S, Xiang M, Wang X, et al., 2024. Development and application of the GenoBaits WheatSNP16K array to accelerate wheat genetic research and breeding. *Plant communications*, 6(1): 101138. https://doi.org/10.1016/j.xplc.2024.101138

Lv Q, Li W, Sun Z, et al., 2020. Resequencing of 1,143 indica rice accessions reveals important genetic variations and different heterosis patterns. *Nature Communications*, 11(1):4778. https://doi.org/10.1038/s41467-020-18608-0

Ma Z, Zhang Y, Wu L, et al., 2021. High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. *Nature Genetics*, 53(9):1385-1391. https://doi.org/10.1038/s41588-021-00910-2

Mansfeld BN and Grumet R, 2018. QTLseqr: An R Package for Bulk Segregant Analysis with Next-Generation Sequencing. *Plant Genome*, 11(2): 180006. https://doi.org/10.3835/plantgenome2018.01.0006

Mantel N, 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2_Part_1):209-220. https://pubmed.ncbi.nlm.nih.gov/6018555/

Martin DN, Proebsting WM and Hedden P, 1997. Mendel's dwarfing gene: cDNAs from the Le alleles and function of the expressed proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 94(16):8907-8911. https://doi.org/10.1073/pnas.94.16.8907

McKenna A, Hanna M, Banks E, et al., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297-1303. https://doi.org/10.1101/gr.107524.110

Minh BQ, Schmidt HA, Chernomor O, et al., 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5):1530-1534. https://doi.org/10.1093/molbev/msaa015

Oksanen J, Simpson GL and Blanchet FG, 2025. vegan: Community Ecology Package. https://doi.org/10.32614/CRAN.package.vegan

Ortiz EM, 2019. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. *Zenodo*. https://doi.org/10.5281/zenodo.2540861

Paradis E and Schliep K, 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526-528. https://doi.org/10.1093/bioinformatics/bty633

Peng Z, Fan W, Wang L, et al., 2017. Target enrichment sequencing in cultivated peanut (Arachis hypogaea L.) using probes designed from transcript sequences. *Molecular Genetics and Genomics*, 292(5):955-965. https://doi.org/10.1007/s00438-017-1327-z

Petropoulou K, Salt LJ, Edwards CH, et al., 2020. A natural mutation in *Pisum sativum* L. (pea) alters starch assembly and improves glucose homeostasis in humans. *Nature Food*, 1(11):693-704. https://doi.org/10.1038/s43016-020-00159-8

Piotrowska-Dlugosz A and Wilczewski E, 2020. Influence of field pea (*Pisum sativum* L.) as catch crop cultivated for green manure on soil phosphorus and P-cycling enzyme activity. *Archives of Agronomy and Soil Science*, 66(11):1570-1582. https://doi.org/10.1080/03650340.2020.1715950

Revell LJ, 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217-223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Riaz A, Li Y, Pandey BR, et al., 2026. Genomic prediction for grain yield and biotic stress resistance in field pea (Pisum sativum

unedited

L.). *Frontiers in Plant Science*, 17: 1739804.
https://doi.org/10.3389/fpls.2026.1739804

Shang LG, He WC, Wang TY, et al., 2023. A complete assembly of the rice Nipponbare reference genome. *Molecular Plant*, 16(8):1232-1236.
https://doi.org/10.1016/j.molp.2023.08.003

Shani E, Hedden P and Sun TP, 2024. Highlights in gibberellin research: A tale of the dwarf and the slender. *Plant Physiology*, 195(1):111-134.
https://doi.org/10.1093/plphys/kiae044

Sindhu A, Ramsay L, Sanderson L-A, et al., 2014. Gene-based SNP discovery and genetic mapping in pea. *Theoretical and Applied Genetics*, 127(10):2225-2241.
https://doi.org/10.1007/s00122-014-2375-y

Singh S, Singh B, Sharma VR, et al., 2019. Character association and path analysis in diverse genotypes of pea (*Pisum sativum* L.). *Int J Curr Microbiol App Sci*, 8(2):706-713.
https://doi.org/10.20546/ijcmas.2019.802.082

Sinjushin A, Semenova E and Vishnyakova M, 2022. Usage of Morphological Mutations for Improvement of a Garden Pea (*Pisum sativum*): The Experience of Breeding in Russia. *Agronomy-Basel*, 12(3): 544.
https://doi.org/10.3390/agronomy12030544

Song B, Ning WD, Wei D, et al., 2023. Plant genome resequencing and population genomics: Current status and future prospects. *Molecular Plant*, 16(8):1252-1268.
https://doi.org/10.1016/j.molp.2023.07.009

Tayeh N, Aluome C, Falque M, et al., 2015a. Development of two major resources for pea genomics: the GenoPea 13.2K SNP Array and a high-density, high-resolution consensus genetic map. *The Plant Journal*, 84(6):1257-1273. https://doi.org/10.1111/tpj.13070

Tayeh N, Klein A, Le Paslier M-C, et al., 2015b. Genomic Prediction in Pea: Effect of Marker Density and Training Population Size and Composition on Prediction Accuracy. *Frontiers in Plant Science*, 6(941.
https://doi.org/10.3389/fpls.2015.00941

Wallace JG, Rodgers-Melnick E and Buckler ES, 2018. On the Road to Breeding 4.0: Unraveling the Good, the Bad, and the Boring of Crop Quantitative Genomics. *In:* Bonini N M (Ed.), *Annual Review of Genetics*, 52: 421–444. https://doi.org/10.1146/annurev-genet-120116-024846

Wang CY, Liu R, Liu YJ, et al., 2021. Development and application of the Faba_bean_130K targeted next-generation sequencing SNP genotyping platform based on transcriptome sequencing. *Theoretical and Applied Genetics*, 134(10):3195-3207. https://doi.org/10.1007/s00122-021-03885-0

Wang P, Li Z, Liu R, et al., 2025. Identification of genetic basis underlying mature cotton fiber length through metabolite-based genome-wide association analysis. *New Phytologist*, 247(5):2210-2230.
https://doi.org/https://doi.org/10.1111/nph.70265

Wang W, Mauleon R, Hu Z, et al., 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, 557(7703):43-49.
https://doi.org/10.1038/s41586-018-0063-9

Wang ZJ, Miao LF, Tan KW, et al., 2025. Near-complete assembly and comprehensive annotation of the wheat Chinese Spring genome. *Molecular Plant*, 18(5):892-907.
https://doi.org/10.1016/j.molp.2025.02.002

Williams O, Vander Schoor JK, Butler JB, et al., 2022. The genetic architecture of flowering time changes in pea from wild to crop. *Journal of Experimental Botany*, 73(12):3978-3990.
https://doi.org/10.1093/jxb/erac132

Wu J, Ma S, Niu J, et al., 2025. Genomics-driven discovery of superior alleles and genes for yellow rust resistance in wheat. *Nature Genetics*, 57(8):2017-2027.
https://doi.org/10.1038/s41588-025-02259-2

Xu L, He WM, Tai SS, et al., 2025. VCF2Dis: an ultra-fast and efficient tool to calculate pairwise genetic distance and construct population phylogeny from VCF files *GigaScience*, 14: giaf032.
https://doi.org/10.1093/gigascience/giaf032

Xu S, Zheng H, Feng Z, et al., 2019. Key construction strategies for core germplasm collection in vegetable pea (Pisum sativum). *Journal of Zhejiang University (Agriculture & Life Sciences)*, 45(4):401-406. https://doi.org/10.3785/j.issn.1008-9209.2018.06.210

Yang T, Liu R, Luo YF, et al., 2022. Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nature Genetics*, 54(10):1553-1563.
https://doi.org/10.1038/s41588-022-01172-2

Unedited

Yang XM, Gou ZW, Zhu ZD, et al., 2022. Breeding and Evaluation of a New-Bred Semi-Leafless Pea (*Pisum sativum* L.) Cultivar Longwan No. 6. *Agronomy-Basel*, 12(4): 850. https://doi.org/10.3390/agronomy12040850

Yu G, Smith DK, Zhu H, et al., 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28-36. https://doi.org/10.1111/2041-210X.12628

Zhang CY, Shao ZQ, Kong YB, et al., 2024. High-quality genome of a modern soybean cultivar and resequencing of 547 accessions provide insights into the role of structural variation. *Nature Genetics*, 56(10): 2247-2258. https://doi.org/10.1038/s41588-024-01901-9

Zhao T, Guan X, Hu Y, et al., 2024. Population-wide DNA methylation polymorphisms at single-nucleotide resolution in 207 cotton accessions reveal epigenomic contributions to complex traits. *Cell Research*, 34(12):859-872. https://doi.org/10.1038/s41422-024-01027-x

Zheng Z, 1997. Food Legume Crops in China. Beijing: China Agriculture Press.

Zhou JS, Wan ZY, Gali KK, et al., 2023. Quantitative trait loci associated with amino acid concentration and *in vitro* protein digestibility in pea (*Pisum sativum* L.). *Frontiers in Plant Science*, 14: 1083086. https://doi.org/10.3389/fpls.2023.1083086

Supplementary information

Tables S1-S7; Figs. S1-S5

unedited