

Association mapping of major economic traits and exploration of elite alleles in *Prunus sibirica*

Jiaxing Wen

Shenyang Agricultural University

Jianhau Chen

Shenyang Agricultural University

Yongqiang Sun

Shenyang Agricultural University

Quangang Liu

Shenyang Agricultural University

Ling Jin

Shenyang Agricultural University

Shengjun Dong (✉ dsj928@163.com)

Shenyang Agricultural University

Research Article

Keywords: *Prunus sibirica*, simple sequence repeats, microsatellite markers, association mapping, phenotypic effect analysis

Posted Date: March 21st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1457343/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Version of Record: A version of this preprint was published at Euphytica on February 27th, 2023. See the published version at <https://doi.org/10.1007/s10681-023-03166-5>.

Abstract

Exploring elite alleles based on association analysis of economic traits will promote the breeding, development, and utilization of *Prunus sibirica* resources. This study used 113 *P. sibirica* clones with diverse provenances to perform association mapping of 20 economic traits based on 149 pairs of highly polymorphic simple sequence repeat (SSR) markers. Population structure analysis split the clones into three subgroups. General linear model (GLM) and mixed linear model (MLM) identified 117 SSR loci associated with 20 *P. sibirica* economic traits, specifically fruit (62 loci), nuclear (69), and kernel (46) traits. Phenotypic variation in these three sets of traits ranged 8.93–82.12% (GLM) or 0.23–84.82% (MLM), 10.96–56.65% (GLM) or 0.17–52.54% (MLM), and 5.17–59.41% (GLM) or 4.86–59.89% (MLM), respectively. Of the 117 marker SSR loci, 35 were simultaneously associated with two or more economic traits. In addition, phenotypic effects analysis identified 1842 elite alleles and seven typical materials related to key economic traits. We then designed elite parental combinations through crossing the typical materials. In conclusion, our study detected alleles and plant materials that allowed the identification of clones expressing desired traits of interest. These findings can be used for improved, targeted breeding of *P. sibirica* through marker-assisted selection.

Introduction

Prunus sibirica (Rosaceae) is a deciduous shrub mainly distributed in Mongolia, eastern Siberia, and northern to northeastern China (Ma et al. 2020). The species has ecological importance, widely used for wind and sand fixation, as well as soil and water conservation (Yin et al. 2019). Furthermore, *P. sibirica* seed kernel is high in oil (50%) that can be used to produce cooking oil, biodiesel fuel, surfactants, almond milk, skin care, and healthcare products (Mai et al. 2020). The seeds are also an ingredient in traditional Chinese medicine (Yan et al. 2018). At present, *P. sibirica* covers approximately 1.7 million ha and yields nearly 192,500 tons of seeds per year in China (Niu et al. 2018).

With rich germplasm resources, *P. sibirica* breeding has the potential for generating improved phenotypes. Current *P. sibirica* yield and quality are significantly reduced by issues such as pistil abortion, vulnerability to freezing injury at the flowering stage, and the bad management. Therefore, the selection and cultivation of appropriate varieties is particularly important. To generate desired cultivars quickly, however, breeding efficiency must be improved, as traditional breeding methods have a very long working cycle and are strongly season-dependent.

Several types of DNA markers are currently available for breeding programs, including RFLP markers (Duval et al. 2001), PCR markers (Ruas et al. 2001; Paz et al. 2005; Feng et al. 2013), and new markers single nucleotide polymorphisms (SNP) (Ravelombola et al. 2017). The most common marker is simple sequence repeats (SSR) or microsatellites, DNA sequences consisting of repeated motifs of 1–6 bp that are widespread distributions biological genomes (Kozlowski et al. 2010). Multiple advantages make SSR markers widespread in research: it is abundant, highly polymorphic, multiallelic, inherited codominant, easy to detect, highly informative, stable, and comparatively inexpensive (Meng et al. 2018). Many SSR markers

have been applied to research on *Prunus* spp., including DNA fingerprinting maps (Li et al. 2019), genetic diversity analysis (Samarina et al. 2022), and linkage analysis (Bhattarai and Mehlenbacher 2018).

With advancements in genomics and improvements to biostatistics software, association mapping has become an important approach for the genetic dissection of quantitative traits (Yeon et al. 2021). This analysis links molecular markers to a target phenotype based on linkage disequilibrium (Zondervan et al. 2004). Additionally, it detects the contributions that different alleles of germplasms to the target trait, facilitating fine mapping of target genes and dramatically improving trait detection efficiency (Xu et al. 2020). For example, Thornsberry et al. (2001) successfully applied association mapping to evaluate markers associated with maize flowering time, and similar studies have followed in corn (Kim et al. 2017), peas (Jha et al. 2020), olives (Ipek et al. 2015), and coconuts (Wei et al. 2014). Several other studies have used association mapping to locate trait-associated loci, then applied phenotypic effect analysis to find favorable alleles and typical materials, including in cotton (Du et al. 2016), rice (Edzesi et al. 2016), coconut (Zhou et al. 2020), poplar (Wei et al. 2010), and China fir (Zhang et al. 2015). However, association mapping and phenotypic effect analysis has not been performed in *P. sibirica*.

Thus, our study applied this combined analysis to identify elite alleles and typical material from 113 *P. sibirica* clones of different regions, using 149 SSR marker pairs. Our results will provide a theoretical basis for molecular breeding studies that aim to manipulate the quality of economic traits in *P. sibirica*, laying the foundation for subsequent development and utilization of plant resources.

Materials And Methods

Plant materials

The study used 113 *P. sibirica* clones supplied by the National Forest Germplasm Resources Preservation Repository for *P. sibirica* in Kazuo, Liaoning Province, P.R. China. Of these, 16 were from Russia and 97 were from China (21 clones from Kazuo, Liaoning; 11 clones from Chaoyang, Liaoning; 19 clones from Beipiao, Liaoning; 12 clones from Aohan Banner, Inner Mongolia; and 34 clones from Zhalantun, Inner Mongolia) (Table 1).

Table 1
Clone number and provenance information of *Prunus sibirica*

Provenances	Clone No.
Kazuo, Liaoning	1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23
Aohan banner, Inner Mongolia	24, 25, 27, 28, 31, 32, 33, 34, 35, 36, 37, 39
Chaoyang, Liaoning	26, 29, 30, 41, 42, 43, 44, 45, 46, 48, 53
Beipiao, Liaoning	47, 49, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 70, 71, 72, 73, 81, 95, 99
Zhalantun, Inner Mongolia	322, 324, 328, 329, 332, 334, 339, 341, 345, 350, 354, 358, 366, 367, 368, 375, 381, 401, 404, 405, 406, 407, 408, 409, 442, 443, 449, 455, 457, 459, 460, 462, 463, 464
Russia	94, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 513, 516, 517, 518

Measurement of economic traits

Twenty economic traits were measured during the 2020 growing seasons: fruit length (FL), fruit width (FW), fruit thickness (FT), flesh thickness (FLT), fruit shape index (FSI), single fruit weight (SFW), fruit yield (FY), nuclear length (NL), nuclear width (NW), nuclear thickness (NT), nuclear shell thickness (NST), nuclear shape index (NSI), single nuclear weight (SNW), nuclear rate (NR), kernel length (KL), kernel width (KW), kernel thickness (KT), kernel shape index (KSI), single kernel weight (SKW), and kernel rate (KR).

DNA amplification and SSR marker genotyping

In 2020, mature and fully expanded leaves were collected and stored at -80°C . Genomic DNA was extracted from leaf tissue with CTAB. The *P. sibirica* genome was used to design 600 SSR primer pairs, which were synthesized by SBS Genetech Co., Ltd. For association mapping and population structure analysis, 149 highly polymorphic primers were selected. The PCR reaction solution (20 μL) comprised: 1 μL of DNA template, 0.15 μL of SSR primers, 2 μL of 10 $\times$ PCR buffer (including Mg^{2+}), 0.4 μL of Taq enzyme, 1.5 μL of dNTP, and 13.1 μL of ddH₂O. Thermocycling conditions were as follows: 94 $^{\circ}\text{C}$ for 5 min; 34 cycles of 94 $^{\circ}\text{C}$ for 30 s, 56 $^{\circ}\text{C}$ for 30 s, 72 $^{\circ}\text{C}$ for 30 s; and extension at 72 $^{\circ}\text{C}$ for 5 min. Amplicons were stored at 4 $^{\circ}\text{C}$, then separated using non-denaturing 12% polyacrylamide gel electrophoresis before silver staining and imaging.

Population structure analysis

Structure 2.3 (Pritchard et al. 2000) with the admixture model was used to analyze population structure and estimate the K value (optimal number of subgroups). Parameter settings were as follows: kinship (K) value between 1 and 10 with a replication of 10, burn-in period of 10 000 for the Markov chain Monte Carlo (MCMC) iteration, followed by 100,000 simulations; all other parameters were set to default. ΔK was calculated based on $\ln P(D)$, and an appropriate K value was selected to obtain the corresponding population structure (Q) matrix.

Association mapping

To reduce the influence of population structure on association mapping, mixed linear model (MLM) and general linear model (GLM) were used to determine associations between genotypes and phenotypes in TASSEL 2.1 (Bradbury et al. 2007). For MLM, multiple levels of population structure (Q-matrix) and relative kinship among individuals (K-matrix) were simultaneously considered. For GLM, the Q matrix was selected as the fixed covariate. The P-value and phenotypic variation explained by a given marker (R^2) for each locus were calculated.

Exploration of elite alleles and germplasm clones

Based on the results of association mapping, alleles of loci significantly associated with target traits were further analyzed using the null allele method proposed (Brescaglio and Sorrells 2006). Average phenotypic values were compared between clones with a given allele and all clones to obtain the phenotypic allele effect (a_i). The calculation formula is as follows:

$$a_i = \sum \frac{x_{ij}}{n_i} - \sum \frac{N_k}{n_k}$$

where a_i is the mean of the i th allelic phenotypic effect value, x_{ij} is the j th material phenotypic value of the i th allele, N_k is the number of materials with the k th allele, x_{kj} is the j th material phenotypic value of all materials, and n_k is the number of materials. An allele is synergistic if $a_i > 0$ and negative if $a_i < 0$.

Results

Population structure analysis

Population structure analysis of the 113 *P. sibirica* clones using 149 polymorphic markers showed that when $K = 3$, ΔK peaked at its maximum value (Fig. 1), splitting the test population into subgroups 1, 2, and 3 (Fig. 2). Subgroup 1 contained 27 clones, with seven from Kazuo, two from Aohan Banner, four from Chaoyang, 11 from Beipiao, and three from Zhalantun. Subgroup 2 contained 43 clones: four from Kazuo, 10 from Aohan Banner, three from Chaoyang, 12 from Beipiao, and 14 from Russia. Subgroup 3 contained 38 clones: eight from Kazuo, four from Chaoyang, five from Beipiao, 19 from Zhalantun, and two from Russia. Based on the results when $K=3$, the corresponding Q matrix was used for marker-trait association mapping.

Marker-trait association mapping

Association mapping of 149 SSR markers and *P. sibirica* economic traits detected 177 significant marker-trait associations ($p < 0.05$), with R^2 ranging from 5.17% to 82.12% (average: 27.93%) for GLM and from 0.17% to 84.82% (average: 21.80%) for MLM. The locus that explained the largest proportion of phenotypic variation was X38 associated with SFW (82.12%, GLM, 84.82%, MLM), followed by X104 (59.41%, GLM) and X34H (59.89%, MLM). The locus that explained the least phenotypic variation was X37 associated with NR (5.17%, GLM, 0.17%, MLM), followed by X67 (8.92%, MLM) and P2 (0.62%, MLM). Among the 117 marker-trait associations, 62 markers were associated with fruit traits, 69 with nuclear traits, and 46 with kernel traits. Their corresponding R^2 ranges from GLM were 8.93%~82.12%, 10.96%~56.65%, 5.17%~59.41%, R^2 ranges from MLM were 0.23%~84.82%, 0.17%~52.54%, 4.86%~59.89%. Among these significant SSR markers, 35 were simultaneously associated with two or more different traits, suggesting tightly linked genes or pleiotropy. We used all 117 marker-trait associations to explore elite alleles (Table 2).

Table 2 SSR loci associated significantly with economic traits

Trait	Number of loci	Marker loci	R ² ranges	
			GLM	MLM
FL	11	X112, X115, X16H, X18, X20, X27, X36, X69, X85, Y22, Y87,	12.54%~43.74%	0.23%~43.86%
FW	8	P2, X112, X16H, X18, X20, X27, X36, X40	16.47%~48.71%	4.88%~46.59%
FT	8	L20H, P2, X112, X16H, X18, X20, X36, X69	19.18%~53.16%	10.40%~49.89%
FLT	10	L2, X112, X115, X15H, X16H, X17H, X18, X20, X27, X40H	14.52%~53.92%	0.91%~37.67%
FSI	7	L34, L49H, X10, X11, X116, X20, X30H	8.93%~34.21%	7.80%~30.81%
SFW	9	L20H, L39, P2, X15H, X18, X20, X36, X38, X69	11.67%~82.12%	11.46%~84.82%
FY	9	L49H, P3, X103, X106, X20H, X40H, X85, X94, X98	13.83%~58.36%	7.46%~39.14%
NL	11	P2, X115, X136, X140, X15H, X16H, X17H, X18, X20, X64, X86	12.88%~45.56%	0.62%~23.17%
NW	12	L39, L75, P3, X136, X17H, X19H, X40, X64, Y26, Y28, Y50, Y86	13.90%~47.16%	5.83%~40.27%
NT	5	L20H, L34, X19H, X74, X98	16.38%~28.61%	14.86%~27.54%
NST	10	L25, L33, L49, X116, X134, X17H, X19H, X36, X37, Y22	13.61%~56.65%	11.58%~52.54%
NSI	19	L20H, L49H, L73H, L75, P56H, X116, X134, X138, X17H, X19H, X27, X37, X38H, X40, X64, X98, Y22, Y50, Y87	10.96%~51.81%	8.32%~45.25%
SNW	5	L68, P2, X20, X69, X8H	18.02%~34.41%	17.93%~24.92%
NR	7	L49, L70H, X11, X18, X27, X37, X76	18.51%~42.33%	0.17%~41.95%
KL	13	L39, L7, P2, X104, X115, X116, X17H, X18, X20, X36, X39, X41, Y87	10.84%~59.41%	7.80%~42.15%
KW	5	X104, X41H, X58, X92, Y87	17.37%~51.14%	17.23%~49.85%
KT	4	L20H, X12, X42H, X93	14.48%~30.05%	9.14%~27.83%
KSI	9	L17, L49H, X10H, X11, X20, X27, X30H, X54H, X67	8.92%~28.39%	7.79%~28.26%
SKW	5	L45H, P2, X20, X34H, Y65	5.17%~57.37%	4.86%~59.89%
KR	10	L28, L75, X101, X20, X29H, X42H, X47, X64, X70, X83	18.12%~33.22%	16.37%~34.46%

Exploration of elite alleles

We investigated 1842 elite alleles, 820 of which were negative and 1022 were synergistic. Phenotypic effects ranged from -245.07 to 803.11. We found 1401 elite alleles that were simultaneously associated with more than two traits. For example, eight elite alleles were associated with 12 traits, 12 with eight traits, nine with seven traits, and 61 with six traits. (All allele effects (ai) and typical materials are listed in S1 Table.) The number of elite alleles for fruit traits FL, FW, FT, FLT, FSI, SFW, and FY were 106, 79, 89, 95, 58, 121, and 70, respectively. The number of elite alleles for nuclear traits NL, NW, NT, NST, NSI, SNW, and NR were 93, 124, 52, 149, 228, 50, and 99, respectively. The number of elite alleles for kernel traits KL, KW, KT, KSI, SKW, and KR were 139, 46, 38, 74, 50, and 82, respectively. Table 3 presents the largest phenotypic effect and most typical material for each trait.

Table 3 Phenotypic effects of alleles of loci associated

Traits	Elite allele	ai	Typical materials (Clone No.)
FL	X18-126	-0.5843	1
FL	X36-144	-0.0010	11,49,95,99,442,443,449,457,511
FL	Y22-111	0.0008	17,24,26,339,341,345,381,404
FL	X36-174	1.3446	94
FW	X18-126	-0.6243	1
FW	X20-140	-0.0010	322,324,334,381
FW	X27-135	0.0211	18,19,21,53,62,71,339,345,350,516,517,518
FW	X27-162	1.3186	464
FT	X69-116	-0.5300	324
FT	X69-144	-0.0067	42,48,375,408,443,455,457,459,460,462,464,511
FT	X18-118	0.0018	59,60,61,407
FT	X36-174	1.4377	94
FLT	X18-126	-1.8357	1
FLT	X27-132	-0.0491	22,328,334
FLT	X27-159	0.0105	14,95,332,367,375,401,405,406
FLT	X40H-138	5.5750	94
FSI	X10-144	-0.1316	464
FSI	X10-165	-0.0016	341
FSI	X30H-147	0.0002	43,46,48,95,322,328,367,368,381,401,404,406,408,442,443

FSI	L49H-129	0.2100	409
SFW	X69-116	-4.3900	324
SFW	X15H-126	-0.0320	4,6,7,20,39,41,43,58,71,81
SFW	X15H-148	0.0458	8,10,14,18,25,27,31,32,322,324,329,334,339,341,345,354, 366,381,405,406,407,408,443,501,505,506
SFW	X36-174	18.8200	94
FY	X20H-124	-245.0905	18,20
FY	X40H-128	-0.7700	24
FY	X98-138	2.9100	509
FY	X98-158	803.1100	95
NL	X18-126	-0.4229	1
NL	X136-144	-0.0071	2,11,13,28,35,49,53,339,341,404,459,463,510,516
NL	X17H-142	0.0026	28,42,53,58,61,71
NL	X115-111	0.4377	43
NW	X19H-148	-0.3890	12,45
NW	X136-156	-0.0008	1,5,17,18,19,21,22,25,27,31,32,33,34,35,36,37,39,41, 46,47,62,71,73,322,324,328,334,345,354,358,381,501, 502,503,504,505,507,509,513,517,518
NW	X17H-134	0.0016	17,30
NW	X40-180	0.3258	43
NT	X98-147	-0.1450	2,55
NT	X74-	-0.0033	19,358

NT	L34-144	0.0021	1,339,345,507,509,518
NT	L20H-158	0.3120	22
NST	X37-146	-0.5929	407
NST	L33-132	-0.0077	2,4,15,17,26,27,32,36,37,42,43,44,58,59,62,95,401,513,517
NST	X19H-150	0.0090	1,4,6,8,10,11,13,14,18,19,20,21,22,23,27,28,31,32,33,34, 35,37,41,42,43,46,49,53,55,62,71,72,81,95,99,322,324,328, 329,332,334,339,341,345,350,354,358,366,375,381,401, 404,405,406,407,408,409,442,443,455,457,459,460,462, 463,464,501,502,503,504,505,506,507,508,509,510,513, 516,517,518
NST	X116-99	1.0168	45
NSI	L20H-158	-0.1580	22
NSI	X138-126	-0.0007	30,61,63
NSI	L20H-132	0.0010	1,6,10,13,21,25,28,31,32,33,34,35,37,72,81,95,366,381,462,463,464
NSI	X38H-162	0.4400	12
NSW	X69-118	-0.5133	25,32
NSW	X69-146	-0.0100	29,366,367
NSW	P2-138	0.0144	4,23,42,328,358,405
NSW	X8H-159	0.4461	5
NR	X11-111	-0.1400	73
NR	X37-136	-0.0007	24,95
NR	L70H-	0.0009	7,13,18,21,22,23,25,26,27,31,32,33,34,35,37,39,41,42,43,

156
 44,45,48,49,53,55,72,94,322,324,329,332,334,339,341,345,
 354,358,375,381,401,406,407,408,442,443,455,457,459,
 460,462,464,501,502,509

NR	L70H-117	0.1400	58,70
KL	X36-104	-0.2603	63
KL	X41-123	-0.0003	5,7,9,12,15,17,20,24,26,29,45,47,63,70,73
KW	X104-208	-0.1280	63
KW	X58-116	-0.0035	23,29,42,55,405,464
KW	X41H-120	0.0040	21,22,23,25,27,28,32,34,35,37,41,42,43,502,506,507
KW	X58-122	0.3632	366
KT	X42H-156	-0.0622	405,409
KT	X12-132	-0.0011	5,7,9,15,17,20,26,29,30,36,45,47,54,56,57,58,59,60,63,70,73,329,368,407
KT	X12-145	0.0002	11,18,27,31,35,39,49,53,503,504,508,510,513
KT	L20H-140	0.1350	61,63
KSI	X10H-120	-0.2339	339
KSI	L17-108	-0.0002	30,36,59,61,63,70,73
KSI	X54H-110	0.0036	17,24,47,70,73,367,368,381
KSI	X11-123	0.4900	409
SKW	X34H-154	-0.0981	39
SKW	Y65-128	-0.0013	15
SKW	Y65-116	0.0002	1,6,8,11,13,18,19,21,28,34,41,42,43,44,46,49,53,55,56,57, 58,59,60,61,62,63,71,95,328,339,345,350,354,358,404,408,

			501,502,503,504,505,508,509,513,516,517,518
SKW	P2-159	0.2422	5
KR	X42H-156	-0.0727	405,409
KR	X42H-122	-0.0047	22,27,33,39,41,43,44,48,62,94
KR	X29H-164	0.0005	1,4,8,11,13,14,18,20,23,25,27,28,31,32,33,34,35,39,42,43,44, 46,49,53,55,71,81,95,99,322,324,328,329,334,339,345,350,354, 358,366,367,368,375,401,404,406,407,408,443,457,459,502,503, 504,505,506,507,508,509,510,511,513,516,517,518
KR	X29H-168	0.1450	6,10

Prediction for novel parental combination

To develop clones with higher economic value, we simulated crossings for molecular breeding. We assumed that clones would have higher economic value with the inclusion of more favorable alleles. Here, we considered synergistic alleles to be favorable for *P. sibirica* breeding programs. Based on the number of favorable alleles and phenotypic effects, we selected seven typical materials as receptor parents to conduct molecular breeding. We then proposed seven optimal cross combinations for improving the economic value of the fruit, nucleus, and kernel (Table 4). Table 5 lists all favorable alleles carried by the parents in these crosses.

Table 4 Cross combinations with excellent alleles for economic traits

Traits	Parental combinations	No. of favorable alleles predicted
Big fruit and High yield	94*95	278
Excellent nuclear	45*43	225
	45*5	221
	45*70	225
Excellent kernel	70*15	209
	15*94	231
	70*94	258

Table 5 Information of parents

Clone No.	Largest phenotypic effect carried	No. of favorable alleles carried
94	FL, FT, FLT, SFW, KR	140
95	FY	138
5	NSW	114
43	NL, NW	118
45	NST	107
70	NR, KT, KSI	118
15	SKW	91

Discussion

The fruit, nucleus, and kernel of *P. sibirica* have a high economic value and varied uses, necessitating the breeding of different seeds for distinct purposes. In this study, we examined more traits than prior research (Du et al. 2016; Edzesi et al. 2016), selecting 20 economic traits (seven fruit, seven nuclear, and six kernel traits). This increase in quantity correspondingly increases the power to evaluate the economic value of clones. We were also able to identify SSR markers associated with these traits, allowing for the use of a molecular breeding approach of marker-assisted selection (MAS) that significantly improves breeding efficiency over traditional methods of based on phenotypic selection.

The population structure is an important factor in association mapping, to avoid false marker trait associations obscuring the loci that truly affect target traits (Ca et al. 2017). Thus, regardless of population type, population structure analysis should be performed prior to association. A relatively complicated population structure occurs when allelic frequency varies within different sub-populations. This unbalanced allelic distribution causes false positive associations between genotypes and phenotypes, complicating association mapping (Cardon et al. 2003). However, based on simple population structure, false positives are minimized and association mapping is highly effective (Cardon et al. 2003). Here, our analysis in Structure clustered the test population of 113 *P. sibirica* clones into three subpopulations, with clones from the same geographic region having a high probability of grouping together. This relatively simple group structure was conducive to the subsequent correlation analysis.

Association mapping has many advantages in improving MAS efficiency. In particular, performing this analysis on a population with numerous genotypes and alleles decreases breeding time because parental crosses are no longer necessary (Xu et al. 2020). In this study, we used both GLM and MLM, the two most prevalent statistical models in association mapping. The MLM (Q + K) model considers both population structure and genetic relationships, reducing the risk of false positives but reducing true positives at the same time. As a result, complex models may miss some important loci (Xie et al. 2017). In contrast, the GLM(Q) model has high detection power and can find more associations (Colonges et al. 2021). Thus,

different association mapping models should be used to make in-depth comparisons, as we have done in this study, to improve the accuracy of the analysis and reduce spurious false positives.

Here, we detected 117 marker-trait associations between 149 SSR and 20 economic traits. Of these associations, the highest phenotypic explanation rates reached 82.12% (GLM) and 84.82% (MLM), while the lowest were 5.17% (GLM) and 0.17% (MLM), respectively. These data show that the tested materials had rich genetic diversity. The larger the phenotypic explanation rate, the stronger the correlation between marker and trait. In this study, the average explained phenotypic variation of the 117 marker-trait associations was 27.93% (GLM) and 21.80% (MLM). These values were higher than those reported for upland cotton (9.5%) (Yuan et al. 2018), wheat (8.7%) (Quan et al. 2019), and poplar (21.4%) (Li et al. 2019).

Most of the biological traits are regulated by multiple loci. The trait with the most association loci (19) was NSI. Additionally, 35 markers were associated with multiple traits. Of these, six loci (X18, X20, X36, X27, X112, and X16H) were related to FL and FW, four (X18, X20, X36, and X69) were related to FL and SFW, and three (P2, X20, and X69) were related to SFW and SNW. These results agree with similar findings in mangoes (Shiv et al. 2017), grapes (Vafaei et al. 2017), and olives (Kaya et al. 2016), suggesting that the traits may exhibit linked inheritance. In terms of breeding, multiple traits can be improved simultaneously by these markers.

After we identified alleles associated with economic traits, we determined their phenotypic effects, and then identified how alleles in different germplasms contributed to the target trait. Through this process, we successfully found resources that carry multiple favorable alleles and bred hybrids with improved target traits by selecting those with the maximum number of favorable alleles. Breseghello and Sorrells et al. (Breseghello et al. 2006) first proposed the use of comparing allelic effects with null alleles to analyze phenotypic effects in 2006, a method that has been widely used to mine elite varieties of coconuts (Zhou et al. 2020), cotton (Wei et al. 2014), and rice (Du et al. 2016). Here, we used the same method to identify 1842 elite alleles of *P. sibirica*, including 1022 with synergistic effects and 820 with negative effects. Except for FY (range – 245.07–803.11), the phenotypic effect of other traits ranged from 10 to -10, similar to cotton (0.0918–2.1461) ((Wei et al. 2014)) and coconuts (-1.02 ~ 3.12) (Zhou et al. 2020). The phenotypic effect value of FY ranged from – 245.07 to 803.11. We speculated that value of FY was large for field measurement, after comparing this value with null allele, the result will also be large.

Peleman and van der Voort first proposed molecular breeding in 2003 (Peleman and van der Voort 2003), This method simulates and optimizes the breeding process before field experiments, thus increasing predictability over conventional breeding. If we want to increase a trait of interest, we should select alleles that have synergistic effects with the elite allele; if we want to reduce a trait, we use alleles with negative effects on the elite allele. For example, to develop stronger salt-tolerant accessions suitable for different saline-alkali soil, Lei et al. (Du et al. 2016) designed a simulated hybridization method to select 13 varieties carrying excellent alleles for molecular breeding. Likewise, Edzesi et al. (Edzesi et al. 2016) selected the five best rice varieties for reducing disease grade and proposed five optimal cross combinations. Our study

identified clones that possess multiple desired elite alleles and considered them core germplasms for breeding. We then designed seven hybrid combinations to improve the economic value of *P. sibirica*.

Conclusions

Association mapping is an efficient tool for unravelling complex quantitative traits in plants and helps identify valuable alleles for breeding through MAS. This study identified 117 stable SSR markers, 1842 elite alleles, and seven typical materials. With these, we generated the seven optimal cross combinations for breeding economically valuable traits. Our findings provide empirical evidence of optimal crosses that can contribute to traditional parental breeding and MAS in the future.

Declarations

Funding This work was supported by National Key R&D Program Project of China (Grant numbers SQ2019YFD100071)

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Jianhua Chen, Yongqiang Sun, Quangang Liu, Ling Jin and Shengjun Dong. The first draft of the manuscript was written by Jiaxing Wen and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability The datasets generated during and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Bhattarai G and Mehlenbacher SA (2018) Discovery, characterization, and linkage mapping of simple sequence repeat markers in hazelnut. *Journal of the American Society for Horticultural Science* 143:347-362
2. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633–2635
3. Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172:1165–1177
4. Cai CP, Zhu GZ, Zhang TZ, Guo WZ (2017) High-density 80 K SNP array is a powerful tool for genotyping *G. hirsutum* accessions and genome analysis. *BMC Genomics* 18:654
5. Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. *The Lancet* 361:598-604
6. Colonges K, Jimenez JC, Saltos A, Seguire E, Solorzano RGL, Fouet O, Argout X, Assemet S, Davrieux F, Cros E, Boulanger R, Lanaud C (2021) Two main biosynthesis pathways involved in the synthesis of

- the floral aroma of the Nacional cocoa variety. *Front Plant Sci* 12:681979
7. Du L, Cai CP, Wu S, Zhang F, Hou S, Guo WZ (2016) Evaluation and Exploration of Favorable QTL Alleles for Salt Stress Related Traits in Cotton Cultivars (*G. hirsutum* L.). *PLoS ONE* 11: e0151076
 8. Duval MF, Noyer JL, Perrier X, Coppens d'Eeckenbrugge & Hamon P (2001) Molecular diversity in pineapple assessed by RFLP markers. *Theor Appl Genet* 102:83–90
 9. Edzesi WM, Dang XJ, Liang LJ, Liu EB, Zaid IU & Hong DL (2016) Genetic diversity and elite allele mining for grain traits in rice (*Oryza sativa* L.) by association mapping. *Frontiers in Plant Science* 7:787
 10. Feng S, Tong H, Chen Y, Wang J, Chen Y, Sun G, He J and Wu Y (2013) Development of pineapple microsatellite markers and germplasm genetic diversity analysis. *Biomed Res Int* 2013:317912
 11. Ipek M, Ipek A, Seker M, Gul M (2015) Association of SSR markers with contents of fatty acids in olive oil and genetic diversity analysis of an olive core collection. *Genetics and molecular research* 14:2241-52
 12. Jha AB, Gali KK, Zhang HX, Purves R W, Tar'an B, Vandenberg A & Warkentin TD (2020) Folate profile diversity and associated SNPs using genome wide association study in pea. *Euphytica* 216: e60-782
 13. Kaya HB, Cetin O, Kaya HS, Sahin M, Sefer F & Tanyolac B, Genetics B (2016) Association Mapping in Turkish Olive Cultivars Revealed Significant Markers Related to Some Important Agronomic Traits. *Biochemical Genetics* 54:506-533
 14. Kim JY, Sa KJ, Ha YJ, Lee JK (2021) Genetic variation and association mapping in the F2 population of the Perilla crop (*Perilla frutescens* L.) using new developed Perilla SSR markers. *Euphytica*. <https://doi.org/10.1007/s10681-021-02867-z>
 15. Kim OG, Sa KJ, Lee JR & Lee JK (2017) Genetic analysis of maize germplasm in the Korean Genebank and association with agronomic traits and simple sequence repeat markers. *Genes & Genomics* 39:843–853
 16. Kozlowski P, de Mezer M, Krzyzosiak WJ (2010) Trinucleotide repeats in human genome and exome. *Nucleic Acids Res* 38:4027-39
 17. Lal S, Singh AK, Singh SK, Srivastav M, Singh BP, Sharma N & Singh NK (2017) Association analysis for pomological traits in mango (*Mangifera indica* L.) by genic-SSR markers. *Trees* 31:1391-1409
 18. Li D, Sun YH, Zhao KQ, Zhang J, Zhang YW, Li XY, Xun SH, Zhao JT, Wang SM & Li Y (2019) Development and application of EST-SSR markers for DNA fingerprinting and genetic diversity analysis of the main cultivars of black locust (*Robinia pseudoacacia* L.) in China. *Forests* 10:644-644
 19. Li XJ, Xu X, Liu WH, Li XQ, Yang XM, Ru ZG, Li LH (2020) Dissection of superior alleles for yield-related traits and their distribution in important cultivars of wheat by association mapping. *Front Plant Sci* 11:175
 20. Ma YX, Wang SX, Liu XJ, Yu HY, Yu D, Li GT, Wang LB (2020) Oil content, fatty acid composition and biodiesel properties among natural provenances of Siberian apricot (*Prunus sibirica* L.) from China. *GCB Bioenergy* 13:112-132

21. Mai YT, Huo KS, Yu HY, Zhou N, Shui LY, Liu Y, Zhang CX, Niu J, Wang LB (2020) Using lipidomics to reveal details of lipid accumulation in developing Siberian apricot (*Prunus sibirica* L.) seed kernels. *GCB Bioenergy* 12:539-552
22. Meng YS, Zhao N, Li Hui, Zhai H, He SZ, Liu QC (2018) SSR fingerprinting of 203 sweetpotato (*Ipomoea batatas* (L.) Lam.) varieties. *Journal of Integrative Agriculture* 17:86-93
23. Niu J, Bi QX, Deng SY, Chen HP, Yu HY, Wang LB, Lin SZ (2018) Identification of AUXIN RESPONSE FACTOR gene family from *Prunus sibirica* and its expression analysis during mesocarp and kernel development. *BMC Plant Biology*. 18:21
24. Paz EY, Gil K, Rebolledo L, Rebolledo A, Uriza D, Martínez O, Isidró M and Simpson J (2005) AFLP characterization of the Mexican pineapple germplasm collection. *Journal of the American Society for Horticultural Science* 130:575–579
25. Peleman JD, van der Voort JR (2003) Breeding by design. *Trends Plant Sci* 8:330–334
26. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
27. Quan MY, Du QZ, L, Lu WJ, Wang LX, Xie JB, Song YP, Xu BH, Zhang DQ (2019) Genetic architecture underlying the lignin biosynthesis pathway involves noncoding RNAs and transcription factors for growth and wood properties in *Populus*. *Plant Biotechnology Journal* 17:302-315
28. Ravelombola W, Qin J, Shi A, Lu WG, Weng YJ, Xiong HZ, Yang W, Bhattarai G, Mahamane S, Payne WA, Miller Jr. JC & Scheuring D (2017) Association mapping revealed SNP markers for adaptation to low phosphorus conditions and rock phosphate response in USDA cowpea (*Vigna unguiculata* (L.) Walp.) germplasm. *Euphytica* 213:183
29. Ruas CF, Ruas PM & Cabral JRS (2001) Assessment of genetic relatedness of the genera *Ananas* and *Pseudananas* confirmed by RAPD markers. *Euphytica* 119:245–252
30. Samarina LS, Matskiv AO, Shkhalakhova RM, Koninskaya NG, Hanke MV, Flachowsky H, Shumeev AN, Manakhova KA, Malyukova LS, Liu SR, Zhu JY, Gvasaliya MV, Malyarovskaya VI, Ryndin AV, Pchikhachev EK and Reim S (2022) Genetic Diversity and Genome Size Variability in the Russian Genebank Collection of Tea Plant [*Camellia sinensis* (L). O. Kuntze]. *Frontiers in Plant Science*. <https://doi.org/10.3389/fpls.2021.800141>
31. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* 28:286–289
32. Vafaei Y, Ghaderi N, Khadivi A (2017) Morphological variation and marker-fruit trait associations in a collection of grape (*Vitis vinifera* L.). *Scientia Horticulturae* 225:771-782
33. Wei ZZ (2021) Evaluation of genetic resources and identification of SSRs associated with important traits in *Populus simonii*. Dissertation, Beijing Forestry University
34. Xia W, Xiao Y, Liu Z, Luo Y, Mason AS, Fan HK, Yang YD, Zhao SL & Peng M (2014) Development of gene-based simple sequence repeat markers for association analysis in *Cocos nucifera*. *Molecular Breeding* 34:525-535

35. Xie D, Dai Z, Yang Z, Sun J, Zhao D, Yang X, Zhang L, Tang Q, Su J (2017) Genome-Wide Association Study Identifying Candidate Genes Influencing Important Agronomic Traits of Flax (*Linum usitatissimum* L.) Using SLAF-seq. *Front Plant Sci* 8:2232
36. Xu L, Jiang S, Wu J, Zou Q (2020) An in silico approach to identification, categorization and prediction of nucleic acid binding proteins. *Briefings in Bioinformatics*. [https://doi, https://doi.org/10.1093/bib/bbaa171](https://doi.org/10.1093/bib/bbaa171)
37. Yin MY, Wuyun T, Jiang ZM & Zeng J (2019) Amino acid profiles and protein quality of Siberian apricot (*Prunus sibirica* L.) kernels from Inner Mongolia. *Journal of Forestry Research* 31:1391–1397
38. Yuan YC, Wang XL, Wang LY, Xing HX, Wang QK, Saeed K, Tao JC, Feng W, Zhang GH, Song XL, Sun XZ (2018) Genome-Wide Association Study Identifies Candidate Genes Related to Seed Oil Composition and Protein Content in *Gossypium hirsutum* L. *Front Plant Sci* 9:1359
39. Zhang SL (2015) Variation regularity analyzation of material properties and association SSR locus mining for *Cunninghamia lanceolata*. Dissertation, Zhejiang Agriculture and Forestry University
40. Zhang Y, Wu YN, Wang W, Yao GD, Jiang XF, Song SJ (2018) Flavones from a natural tea (the leaves of *Armeniaca sibirica* L.) prevent oxidative stress-induced neuronal death. *Journal of Food Biochemistry* 42: e12657.
41. Zhou LX, Yarra RJ, Cao HX (2020) SSR based association mapping analysis for fatty acid content in coconut flesh and exploration of the elite alleles in *Cocos nucifera* L. *Current Plant Biology* 21:2214-6628
42. Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. *Nature Reviews Genetics* 5:89-100

Figures

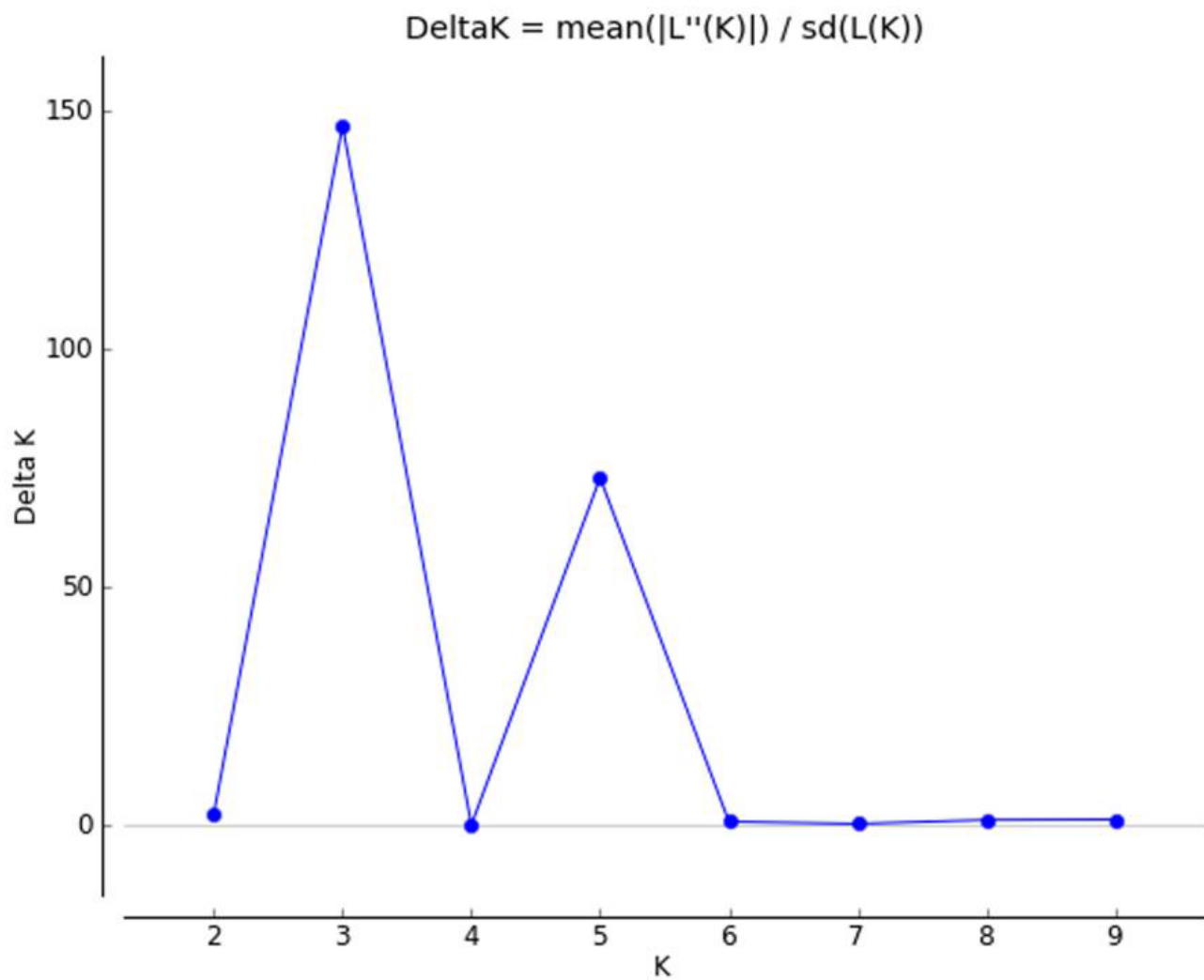


Figure 1

ΔK value with the change of K

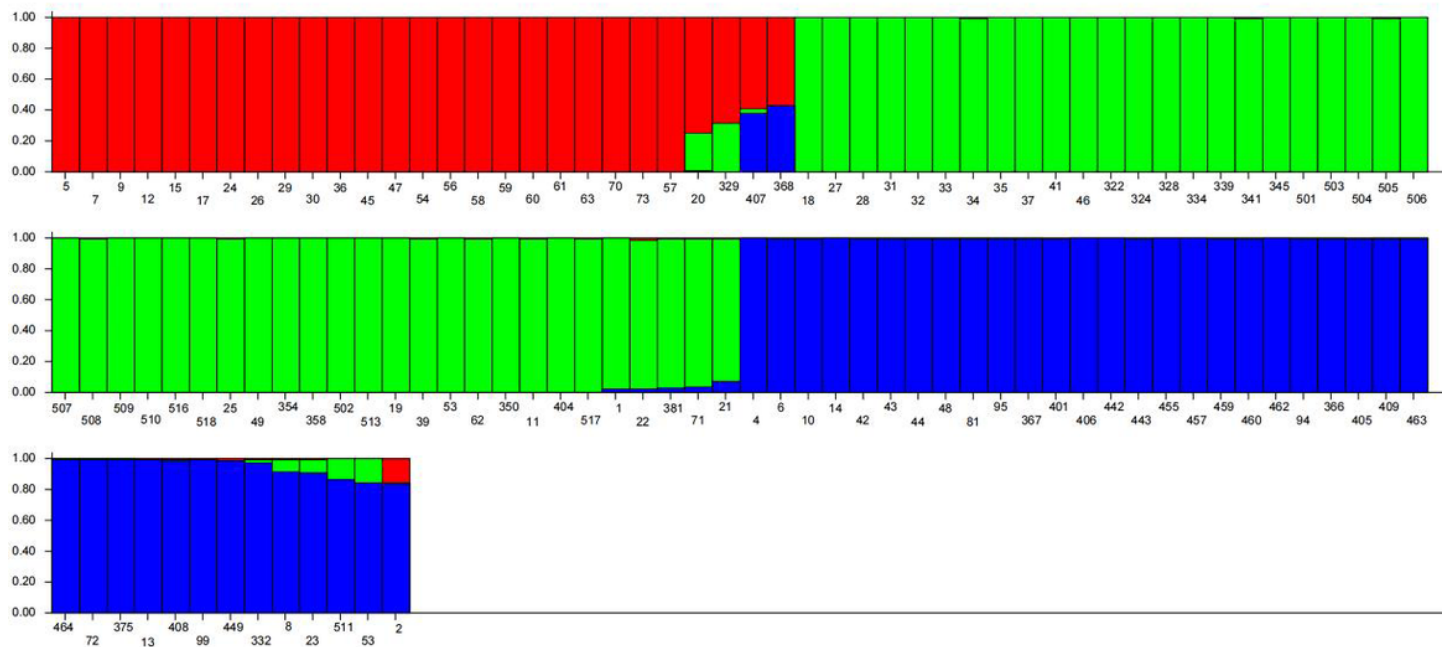


Figure 2

Population structure in 113 *Prunus sibirica*/clones (K = 3)