Journal of Integrative Agriculture 2022, 21(7): 1903–1912

Available online at www.sciencedirect.com

ScienceDirect

RESEARCH ARTICLE

Identification and characterization of long-InDels through whole genome resequencing to facilitate fine-mapping of a QTL for plant height in soybean (*Glycine max* L. Merr.)

LIU Chen[1, 2*], TIAN Yu[2*], LIU Zhang-xiong[2], GU Yong-zhe[2], ZHANG Bo[3], LI Ying-hui[2], NA Jie[1], QIU Li-juan[2]

[1] *School of Life Sciences, Liaoning Normal University, Dalian 116081, P.R.China*

[2] *National Key Facility for Gene Resources and Genetic Improvement/Key Laboratory of Crop Germplasm Utilization, Ministry of Agriculture and Rural Affairs/Institute of Crop Sciences, Chinese Academy of Agricultural Science, Beijing 100081, P.R.China*

[3] *School of Plant and Environmental Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060, USA*

Abstract

With the development of sequencing technology, insertions-deletions (InDels) have been increasingly reported to be involved in the genetic determination of agronomical traits. However, most studies have focused on the identification and application of short-InDels (1–15 bp) for genetic analysis. The objective of this study was to deeply deploy long-InDels (>15 bp) for the genetic analysis of important agronomic traits in soybean. A total of 13 573 polymorphic long-InDels were identified between parents Zhongpin 03-5373 (ZP) and Zhonghuang 13 (ZH), which were unevenly distributed on 20 chromosomes of soybean, varying from 321 in Chr11 to 1 246 in Chr18. Consistent with the distribution pattern of annotated genes, the average density of long-InDels in arm regions was significantly higher than that in pericentromeric regions at the P=0.01 level. A total of 2 704 (19.9% of total) long-InDels were located in genic regions, including 319 large-effect long-InDels, which resulted in truncated or elongated protein sequences. A previously identified QTL (*qPH16*) underlying plant height was further analyzed, and it was found that 26 out of 35 (74.3%) long-InDel markers located in the *qPH16* region showed clear polymorphisms between parents ZP and ZH. Seven markers, including three long-InDels and four previously reported SNP markers, were used to genotype 242 recombinant inbred lines derived from ZP×ZH. As a result, the *qPH16* locus was narrowed from a 960-kb region to a 477.55-kb region, containing 65 annotated genes. Therefore, these long-InDels are a complementary genetic resource of SNPs and short-InDels for plant height and can facilitate genetic studies and molecular assisted selection breeding in soybean.

Keywords: soybean, plant height, whole genome re-sequencing, long-InDels, QTL

Received 25 November, 2020 Accepted 3 March, 2021

LIU Chen, E-mail: 1518753786@qq.com; TIAN Yu, E-mail: 82101181082@caas.cn; Correspondence QIU Li-juan, E-mail: qiulijuan@caas.cn; NA Jie, Tel: +86-411-85827090, E-mail: jiena@lnnu.edu.cn; LI Ying-hui, Tel: +86-10-82105843, E-mail: liyinghui@caas.cn

* These authors contributed equally to this study.

doi: 10.1016/S2095-3119(21)63675-4

1. Introduction

Cultivated soybean (*Glycine max* L. Merr.) is one of the most economical sources supplying plant protein and oil for animal feed and food for human consumption around the world. With increasing demand for soybeans, breeders are expected to develop more high-yielding cultivars. Plant height (PH) is one of the most important

agronomic traits related to seed yield in soybean under the condition of preventing lodging, so ideal PH is crucial for improving soybean yield. As a complex quantitative trait controlled by multiple genes (Liu *et al.* 2013; Lee *et al.* 2015), PH can be improved *via* molecular breeding approaches, such as gene editing, molecular-assisted selection and others, using the information of fine-mapped quantitative trait loci (QTLs) or map-based gene cloning.

During the early stage of QTL mapping, restriction fragment length polymorphisms (RFLP), amplified fragment length polymorphisms (AFLP) and/or simple sequence repeats (SSR) markers were used to construct genetic maps with low to moderate resolution (Mansur *et al.* 1996; Wang *et al.* 2004; Lee *et al.* 2015). The soybean reference genome (var. Williams 82) has facilitated the development of sequence-based molecular markers. With the development of whole genome re-sequencing (WGS), abundant soybean accessions were re-sequenced in recent years (Lam *et al.* 2010; Li *et al.* 2013; Zhou *et al.* 2015; Liu *et al.* 2020). Millions of single nucleotide polymorphisms (SNPs) were detected by WGS, and have been widely used for construction of high-density genetic maps (Liu *et al.* 2017; Ji *et al.* 2018; Yu *et al.* 2021), whole-genome association studies (Zhang *et al.* 2016; Fang *et al.* 2017; Xia *et al.* 2019) and marker-assisted selection (Shi *et al.* 2015; Patil *et al.* 2017). Although many reports of genetic variations focus on SNPs, there is an increasing appreciation for the need to fully capture the different genomic variation types for phenotypic variation (Springer *et al.* 2009).

Insertions-deletions (InDels) are the second most abundant genetic variations after SNPs. InDel variations can be formed by unequal crossover, transposable elements and sequence replication of slippage (Britten *et al.* 2003). Their sizes vary greatly, from 1 bp to several kb. Therefore, most InDel loci could be genotyped using simple and cost-efficient agarose or polyacrylamide gel electrophoresis (PAGE), and they have been widely used for QTL mapping or fine mapping. For example, a crinkly leaf related gene was fine-mapped within a 360-kb region using five InDel markers (Song *et al.* 2015). However, previous studies identified many short-InDels (1–15 bp) in soybean (Li *et al.* 2013; Zhou *et al.* 2015; Yang *et al.* 2020). For example, a total of 876 799 InDels (1–6 bp) were identified by more than 11× average coverage depth of resequencing analysis among 302 wild and cultivated accessions (Zhou *et al.* 2015). Recently, a total of 102 491 InDels (1–15 bp) were identified between Zhongpin 03–5373 (ZP) and Zhonghuang 13 (ZH) (Yang *et al.* 2020), and a large number of structural variants (>50 bp insertion or deletion) were identified by construction of the pan-genome (Liu *et al.* 2020). However, long-InDel variants (>15 bp), instead of structural variation such as presence/absence variants (PAVs) and copy number variants (CNVs), were not adequately identified. Therefore, it is also necessary to explore the characteristics of long-InDels and their relationships with phenotypic variation.

In the current study, a total of 13 573 InDels (>15 bp) were identified by comparisons with the re-sequenced genome of ZP and ZH. They were defined as long-InDels that showed uneven distribution patterns on 20 chromosomes of soybean, and had an average density in arm regions significantly higher than in pericentromeric regions. In order to verify and utilize these long-InDel variants, 26 (74.3%) of the 35 developed long-InDel markers located in the *qPH16* region showed clear polymorphisms as predicted between parents ZP and ZH. Finally, the *qPH16* locus was narrowed from 960 kb to a 477.55-kb region *via* the addition of three long-InDel markers.

2. Materials and methods

2.1. Plant materials

A total of 242 RILs derived from ZP×ZH were used in this study, which were also employed for primary QTL mapping of PH in a previous study (Liu *et al.* 2013). Since *qPH16* was previously identified in Sanya, Hainan, the PH phenotype data in Sanya (18.2°N, 109.5°E) in 2009 (SYa09) and 2010 (SYa10) were used in the current QTL mapping (Liu *et al.* 2013).

2.2. Genome re-sequencing, alignment, and InDel detection

The genomic DNA of RILs derived from ZP×ZH was extracted from young leaves of five seedlings per F_{12} line using the CTAB method (Doyle and Doyle 1990). Parental lines ZP and ZH were re-sequenced using the Illumina Solexa System as described in Yang *et al.* (2020). High-quality paired-end 150-nucleotide reads were mapped to the soybean reference genome Wm82.a2.v1 using Burrows-Wheeler Aligner Software (Li *et al.* 2009). Finally, long-InDel polymorphisms with gaps >15 bp in length were detected between ZP and ZH.

2.3. InDel variation annotation and GO analysis

InDel annotation was analyzed according to the reference genome Williams 82.a2.v1 using the package ANNOVAR (Wang *et al.* 2010). Based on the genome annotation, InDels were categorized into exonic regions, splicing, intronic regions, 5′ and 3′ untranslated regions (UTRs),

upstream and downstream regions (overlapping 1-kb region upstream of transcription start site/downstream of transcription end site) and intergenic regions. The online Software AgriGO (http://bioinfo.cau.edu.cn/agriGO/) was used to analyze GO (Gene Ontology) enrichment.

2.4. The development of InDel markers

InDel primers were designed using Primer5.0 Software. The specific primers were selected by blasting with reference genome Williams 82 using Phytozome (https://phytozome.jgi.doe.gov). PCR reactions (20 μL) contained 2 μL genomic DNA (20 ng μL^{-1}), 2 μL PCR buffer (10×), 1.5 μL dNTPs (2 mmol), 2 μL forward and reverse primers (2 μmol), and 1 U *Taq* polymerase (TransGen Biotech Co., Ltd., Beijing, China). The PCR amplification procedure was performed as follows: a denaturing step at 95°C for 4 min, followed by 34 cycles of denaturation at 95°C for 30 s, annealing at 58–60°C for 30 s, extension at 72°C for 30 s, and a final extension at 72°C for 5 min before cooling to 4°C. The PCR products were detected using 6% polyacrylamide gel (PAGE) and visualized by silver staining.

2.5. Construction of genetic map and QTL mapping

The genetic map was constructed by the Kosambi mapping function of QTL IciMapping (Li *et al.* 2008). Based on the genetic map, inclusive composite interval mapping (ICIM) was used to detect the QTL in SYa09 and SYa10 with 1 cM walk speed, and markers (P<0.001) entering stepwise regression. A LOD score of 2.5 was used as the threshold to define the presence of a QTL.

3. Results

3.1. Identification of long-InDels between ZP and ZH

The parental lines of a RIL population, ZP and ZH, were re-sequenced in a previous study using the Illumina Solexa platform (Yang *et al.* 2020). A total of 12 574 728 reads for ZP and 107 868 279 reads for ZH with high-quality paired-end 150-nucleotides were mapped to the soybean reference genome Williams 82.a2.v1., respectively. This study identified a total of 13 573 long-InDels between ZP and ZH, including 4 748 insertions and 8 825 deletions (Fig. 1-A). A significantly negative relationship was found between the long-InDel lengths and numbers; as the InDel length increased, the number of InDels decreased (P<0.001, R^2=–0.99). The majority of long-InDels (8 418, 62%) were located within the intergenic region, while 2 703 (19.9%) long-InDels were in the genic regions, 301 in exons, 18 in the junction sites between exons and introns resulting in alternative splicing, 1 794 in introns, and 590 in UTRs. The large effect long-InDels which resulted in truncated or elongated protein sequences only accounted for 2.4% of the total of 13 573 identified long-InDels (Fig. 1-B).

3.2. Distribution of long-InDels across the whole genome

The number of long-InDels was largely different among the 20 chromosomes, with a range from 321 for Chr11 to 1 246 for Chr18 (Table 1). Although Chr18 showed the largest number of long-InDels, which was almost four times the smallest number in Chr11, Chr15 exhibited the highest average density of long-InDels with 22.34 InDels/Mb and Chr04 had the lowest density at 7.75 InDels/Mb. Further analyses suggested that the distribution of long-InDels was also uneven on the chromosomes (Fig. 2-A). The average density of long-InDels in arm regions (20.51 InDels/Mb) was significantly higher than in pericentromeric regions (9.07 InDels/Mb) (Fig. 2-B; Table 2).

3.3. GO annotation and enrichment

In order to elucidate the functions of 305 annotated genes containing 319 large-effect long-InDels that encoded the truncated or elongated proteins (Appendix A), GO enrichment analysis was performed using AgriGO (http://bioinfo.cau.edu.cn/agriGO/). The 305 genes were assigned into 112 GO terms, including cellular component (18), molecular function (37) and biological process (57) (Fig. 3). Binding process of molecular function (GO:0005488) was significantly enriched. Especially, ADP binding (GO:0043531, P=6.73E–09) which contained 14 enriched resistant (R) genes with the NB-ARC/TIR-NBS-LRR domain which may be associated with different external stimulants (Appendices B and C). The R genes were co-located with reported QTLs from Soybase (https://www.soybase.org/), which were covered by multiple QTLs related to resistance, such as *SCN4-3*, *11-1*, *40-3*, *44-15*, and *50-3* for *Heterodera glycines* (Concibido *et al.* 1997; Qiu *et al.* 1999; Yue *et al.* 2001; Ferreira *et al.* 2011; Jiao *et al.* 2015), QTL *Sclero5-13*, *9-2* and *3-15* for *Sclerotinia sclerotiorum* (Arahana *et al.* 2001; Guo *et al.* 2008), QTL *Rag3-1* for *Aphis glycines* (Zhang *et al.* 2009), and QTL *Phytoph 6-8*, *8-4* and *14-8* for *Phytophthora sojae* (Li *et al.* 2010; Tucker *et al.* 2010; Lee *et al.* 2013) (Appendix D).

3.4. *qPH16* fine mapping with long-InDel markers

In a previous study, Liu *et al.* (2013) used a lower-

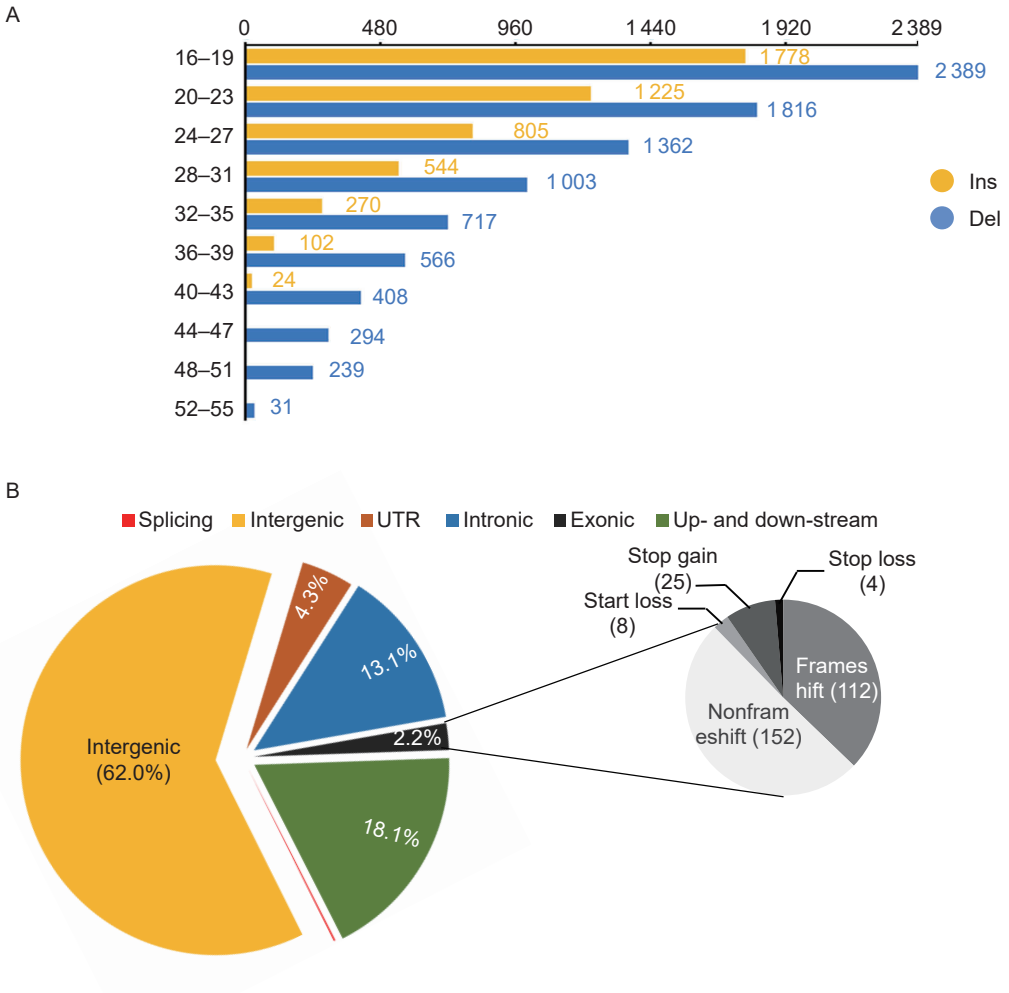

Fig. 1 Identification of long-insertions and deletions (InDels) variants in the re-sequenced Zhongpin 03-5373 (ZP) and Zhonghuang 13 (ZH) genomes. A, the length distribution of InDels (>15 bp) detected between ZP and ZH. Ins and Del represent insertions and deletions, respectively. B, location, distribution and functional annotation of InDel variants based on reference genome Williams 82.a2.v1.

density genetic map containing 508 molecular markers, and mapped QTL *qPH16* for plant height in Sanya City, Hainan Province, China (18.2°N, 109.5°E) in a 960-kb genome region flanked by two SNP markers, Map-3031 and Map-3141. A total of 35 long-InDel variants in the *qPH16* genomic region were all used to develop InDel markers, which were distributed in the intergenic region, up-down stream, UTR, introns, and exons (Appendix E). Among them, 26 (74.3%) of 35 InDel markers showed clear polymorphisms between parents ZP and ZH, while the remaining nine markers failed to amplify PCR products or had unpredicted polymorphisms (Appendix E). The results indicated that the authenticity of the InDels was higher than that (68%) in a previous study (Song *et al.* 2015). Given the limited population size and recombination events, three long-InDel markers covering the entire region were used to genotype 242 RILs derived from ZP×ZH (Li *et al.* 2016). Then, regional QTL mapping was performed using seven markers, including four SNP markers identified in a previous study (Liu *et al.* 2013), and plant height data were phenotyped in Sanya City in 2009 and 2010. In the current study, *qPH16* was mapped in a 13.22-cM genetic region flanked by Map-3141 and InDel16-29, explaining 4.1 and 7.1% of the phenotypic variance, respectively (Fig. 4). The genomic interval of *qPH16* was narrowed down from 960 to 477.55 kb by adding three InDel markers. Collinearity analysis indicated that the order of markers surrounding *qPH16* was consistent with that on reference genome Williams 82.a2.v1, and the genomic region of *qPH16* contained 65 annotated genes.

A total of 502 variations within the *qPH16* region between the ZP and ZH genomes were obtained to further analyze candidate genes, including 415 SNPs, 79 short-InDels and 8 long-InDels. Among the 502 variations, 34 with large-effects and predicted to affect

Table 1 Summary of long-insertions and deletions (InDels) variations identified between Zhongpin 03-5373 (ZP) and Zhonghuang 13 (ZH) among 20 soybean chromosomes[1)]

Chromsome ID	Genetic region						Up- and down-stream			Intergenic region	Total	Density (InDels/Mb)
	Exon	5′ UTR	3′ UTR	5′ UTR and 3′ UTR	Intron	Splicing	Subtotal	Up-stream	Down-stream	Up-stream and down-stream	Subtotal			
Chr01	5	7	8	0	52	0	72	49	28	4	81	366	519	9.13
Chr02	18	10	14	0	68	1	111	54	50	6	110	246	467	9.61
Chr03	18	23	15	0	139	1	196	75	69	9	153	668	1017	22.22
Chr04	7	10	13	0	58	0	88	48	48	4	100	218	406	7.75
Chr05	10	15	15	0	81	2	123	49	58	10	117	259	499	11.81
Chr06	25	14	23	0	115	0	177	84	67	8	159	519	855	16.63
Chr07	22	11	17	1	113	1	165	74	51	6	131	359	655	14.65
Chr08	20	17	19	0	80	1	137	69	44	8	121	239	497	10.39
Chr09	10	19	21	0	86	2	138	68	66	1	135	406	679	13.53
Chr10	12	17	18	0	71	0	118	45	50	11	106	348	572	11.09
Chr11	5	3	6	0	48	1	63	34	22	4	60	198	321	9.23
Chr12	16	12	15	0	88	1	132	45	52	8	105	261	498	12.42
Chr13	16	19	21	0	85	2	143	58	78	4	140	473	756	16.48
Chr14	9	6	13	0	75	2	105	43	53	8	104	191	400	8.16
Chr15	17	11	13	0	105	1	147	85	81	4	170	839	1156	22.34
Chr16	22	12	14	0	100	0	148	61	70	7	138	403	689	18.19
Chr17	7	8	12	0	77	1	105	63	44	3	110	502	717	17.22
Chr18	27	21	32	0	165	0	245	94	88	13	195	806	1246	21.48
Chr19	11	11	20	0	86	2	130	51	42	3	96	635	861	16.97
Chr20	24	11	24	0	102	0	161	51	62	7	120	482	763	15.93

[1)] UTR, untranslated regions; splicing, change of splice donor or acceptor sites; up-stream, overlaps 1-kb region upstream of transcription start site; down-tream, 1-kb region.

encoded peptide sequences were located in the exon regions of 14 annotated genes. Further analysis of the expression patterns of the 14 annotated genes of Williams 82 indicated that *Glyma.16G163800* showed a higher expression level in leaf, shoot apical meristem and stem tissue (https://phytozome) than other genes. Its paralogous gene in *Arabidopsis*, *AT1G34000* encoding a one-helix protein (OHP2), was related to responses to different light levels (Hey *et al.* 2020, 2018). Therefore, it was suggested that *Glyma.16G163800* may be considered as the candidate gene located within environment-specific QTL *qPH16*.

4. Discussion

Molecular markers play a key role in exploring the genetic basis of important agronomic traits in soybeans. In early studies, RFLP, AFLP and SSR markers were widely used for genetic analysis (Mansur *et al.* 1993, 1996; Lark *et al.* 1995; Concibido *et al.* 1997; Yan *et al.* 2015), however, this largely limited the progress of genetic research due to low-resolution genetic maps with large gaps. In recent years, whole genome sequencing technology has significantly improved traditional genetic studies by supplying cost-effective sequence information for the development of new genetic markers (Song *et al.* 2015). Millions of short-InDels were identified against the reference genome Williams 82 (Li *et al.* 2013; Zhou *et al.* 2015), and some of them have been developed as InDel markers for genetic studies (Li *et al.* 2014; Song *et al.* 2015). Compared with RFLP, AFLP, SSR and SNP markers, insertion and deletion markers can be more easily amplified and detected by regular electrophoresis gel systems (Li *et al.* 2014). In order to thoroughly capture InDel variations, the long-InDels of >15 bp in length were analyzed on the re-sequenced ZP and ZH genomes. A total of 13 573 long-InDels were obtained, and its number was much smaller than that (102 491) of small InDels (Yang *et al.* 2020). Short- and long-InDels also had different density distributions among the chromosomes. The density of

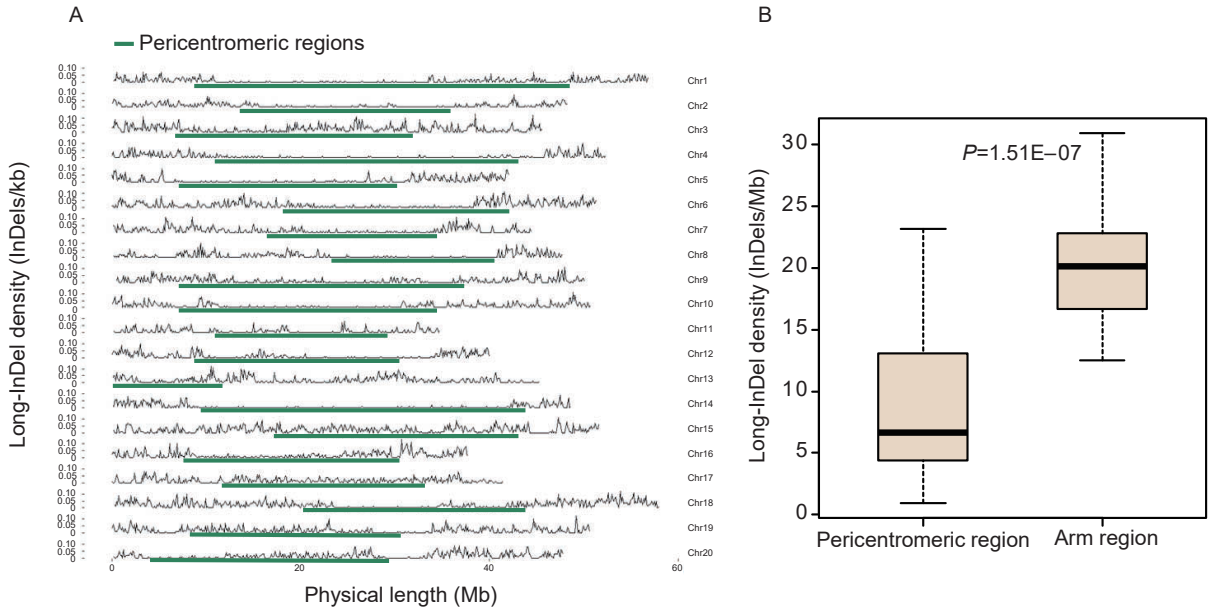

Fig. 2 Physical distribution of long-insertions and deletions (InDels) on 20 chromosomes of soybean. A, density distribution of long-InDels based on continuous 100-kb windows, and the gray underlining represents pericentromeric regions (Song *et al*. 2016). B, the average density of long-InDels on pericentromeric and arm regions. Significance was assessed by the *T*-test using SAS 9.4.

Table 2 Physical distribution of long-insertions and deletions (InDels) within arm and pericentromeric regions

Chromosome ID	Number of long-InDels				Density (long-InDels/Mb)			
	Arm region			Pericentro-meric region	Arm region			Pericentro-meric region
	Arm1	Arm2	Subtotal		Arm1	Arm2	Subtotal	
Chr01	152	147	299	220	18.77	15.64	17.09	5.60
Chr02	271	152	423	44	16.94	14.62	16.02	1.98
Chr03	238	263	501	516	34.49	21.21	25.96	19.47
Chr04	185	159	344	62	17.79	17.87	17.82	1.87
Chr05	120	303	423	76	18.75	25.25	22.99	3.19
Chr06	327	214	541	314	17.97	30.57	21.47	11.98
Chr07	337	217	554	101	19.04	21.70	20.00	5.98
Chr08	290	190	480	17	12.66	25.68	15.84	0.97
Chr09	138	222	360	319	21.56	19.48	20.22	9.85
Chr10	152	247	399	173	22.03	16.92	18.56	5.77
Chr11	140	62	202	119	12.28	13.19	12.55	6.40
Chr12	171	186	357	141	20.85	24.47	22.59	5.83
Chr13	0	513	513	243	0	15.78	15.78	18.27
Chr14	231	105	336	64	23.81	19.81	22.40	1.88
Chr15	398	185	583	573	21.75	21.26	21.59	23.20
Chr16	172	336	508	181	20.72	30.55	26.32	9.78
Chr17	243	85	328	389	16.99	14.66	16.32	18.09
Chr18	525	563	1 088	158	25.61	38.30	30.91	6.93
Chr19	173	328	501	360	19.44	20.00	19.80	14.17
Chr20	58	392	450	313	18.13	27.61	25.86	10.26

short-InDels on Chr18 was the largest (Yang *et al.* 2020), however, the density of long-InDels on Chr15 was the largest. The long-InDels detected in this study, combined with previously reported short-InDels, will be helpful in the development of InDel markers for genetic analysis of important agronomic traits in soybeans.

In a previous study, an environment-specific (such as Sanya) QTL *qPH16* was primarily mapped in a 960-kb genomic region in Chr16 using the RILs derived from ZP×ZH (Liu *et al.* 2013), which were partially overlapped with the other two QTLs of plant height, *Plant height 5-13* and *25-2* (https://www.soybase.org/). *Plant height 5-13* was obtained in an F_4 population derived from Young and PI416937 and flanked by BARC-039399-07318 and

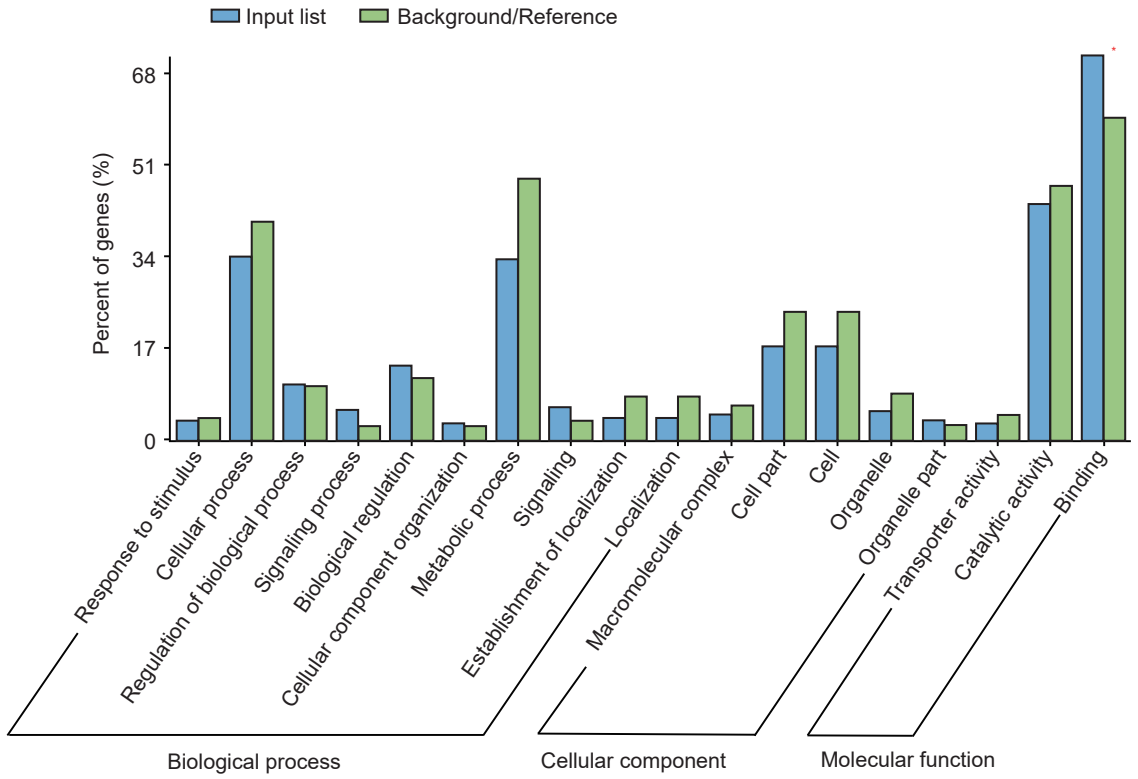

Fig. 3 Gene Ontology (GO) annotation and enrichment of 305 genes carrying large-effect insertions and deletions (InDels). GO annotations were grouped into three term types, including cellular component, molecular function and biological process. Significance (*P*-value) of GO enrichment was corrected by the false discovery rate (FDR) method, * represents $P<0.05$.

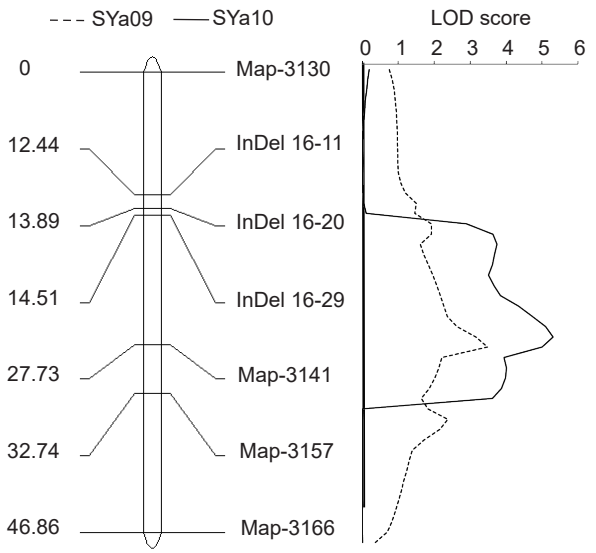

Fig. 4 Fine-mapping of QTL *qPH16* in SYa09 and SYa10. LOD, likelihood-of-odds.

Sat_396 in Chr16 (Lee *et al.* 1996). *Plant height 25-2* was identified in backcross populations derived from Beeson 80 and LG92-1143 (Guzman *et al.* 2007). In order to validate and use those long-InDels, a previously reported QTL for plant height, namely *qPH16*, was fine-mapped by developed long-InDel markers and *qPH16* was narrowed down to a 477.55-kb genomic region, including 65 annotated genes. Based on the function of the paralogous gene in *Arabidopsis*' response to different light levels, the candidate gene *Glyma.16G163800* for environment-specific QTL *qPH16* may be associated with the short-day condition in Sanya's environment. Therefore, this indicates that the long-InDel loci could provide useful information for developing InDel markers and increasing the marker density to further narrow down the targeted-QTL region.

Zhonghuang 13 has been the most widely planted soybean accession in China in the past 20 years due to its high yield, wide adaptability, and high stress tolerance (Li *et al.* 2014; Shen *et al.* 2018). The line ZP exhibits resistance to multiple SCN and tolerance to drought (Li *et al.* 2014, 2016). In this study, by comparing the ZH and ZP genomes, a total of 319 large-effect long-InDels were obtained located in 305 annotated genes that were most significantly enriched in ADP binding (GO:0043531) containing 14 R genes by GO enrichment analysis. The R genes with LRRs and an NB-ARC domain are an essential part of the reaction to different external stimulants (Dangl

and Jones 2001; Lam *et al.* 2010). The result of co-location between enriched R genes and the reported QTL indicated that those genes may be involved in resistance to different environmental stresses. For example, six of these genes were located in a QTL region for resistance to SCN, five were located in a QTL region for resistance to *Phytophthora sojae* and three were located in a QTL region for resistance to *Sclerotinia sclerotiorum* (Appendix D). At the same time, some genes were co-located in multiple-QTLs related to different diseases, which may serve as broad spectrum disease resistance.

5. Conclusion

This study shows that long-InDel markers along with short-InDels, SNPs and SSRs can be used for increasing the marker density of current genetic linkage maps with low marker density, and will be valuable for fine mapping QTLs. The developed long-InDel markers linked to plant height in this study can be utilized in marker-assisted selection and cloning of *qPH16*.

Acknowledgements

This research was supported by the National Key R&D Program of China (2016YFD0100201 and 2020YFE0202300) and the Agricultural Science and Technology Innovation Program (ASTIP) of the Chinese Academy of Agricultural Sciences.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Appendices associated with this paper are available on http://www.ChinaAgriSci.com/V2/En/appendix.htm

References

Arahana V S, Graef G L, Specht J E, Steadman J R, Eskridge K M. 2001. Identification of QTLs for resistance to *Sclerotinia sclerotiorum* in soybean. *Crop Science*, **41**, 180–188.

Britten R J, Rowen L, Williams J, Cameron R A. 2003. Majority of divergence between closely related DNA samples is due to InDels. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4661–4665.

Concibido V C, Lange D A, Denny R L, Orf J H, Young N D. 1997. Genome mapping of soybean cyst nematode resistance genes in 'Peking', PI 90763, and PI 88788 using DNA markers. *Crop Science*, **37**, 258–264.

Dangl J L, Jones J D G. 2001. Plant pathogens and integrated defence responses to infection. *Nature*, **411**, 826–833.

Doyle J J, Doyle J L. 1990. Isolation of plant DNA from fresh tissue. *Focus*, **12**, 13–15.

Fang C, Ma Y M, Wu S W, Liu Z, Wang Z, Yang R, Hu G H, Zhou Z K, Yu H, Zhang M, Pan Y, Zhou G A, Ren H X, Du W G, Yan H R, Wang Y P, Han D Z, Shen Y T, Liu S L, Liu T F, *et al.* 2017. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biology*, **18**, 161.

Ferreira M F D S, Cervigni G D L, Ferreira A, Schuster I, Moreira M A, Santana F A, Pereira W D, Barros E G D, Moreira M A. 2011. QTLs for resistance to soybean cyst nematode, races 3, 9, and 14 in cultivar Hartwig. *Pesquisa Agropecuária Brasilra*, **46**, 420–428.

Guo X M, Wang D C, Gordon S G, Helliwell E, Smith T, Berry S A, St Martin S K, Dorrance A E. 2008. Genetic mapping of QTLs underlying partial resistance to *Sclerotinia sclerotiorum* in soybean PI 391589A and PI 391589B. *Crop Science*, **48**, 1129–1139.

Guzman P S, Diers B W, Neece D J, St Martin S K, LeRoy A R, Grau C R, Hughes T J, Nelson R L. 2007. QTL associated with yield in three backcross-derived populations of soybean. *Crop Science*, **47**, 111–122.

Hey D, Grimm B. 2018. ONE-HELIX PROTEIN2 (OHP2) is required for the stability of OHP1 and assembly factor HCF244 and is functionally linked to PSII biogenesis. *Plant Physiology*, **177**, 1453–1472.

Hey D, Grimm B. 2020. ONE-HELIX PROTEIN1 and 2 form heterodimers to bind chlorophyll in photosystem II biogenesis. *Plant Physiology*, **183**, 179–193.

Lam H M, Xu X, Liu X, Chen W B, Yang G H, Wong F L, Li M W, He W M, Qin N, Wang B, Li J, Jian M, Wang J, Shao G H, Wang J, Sun S S, Zhang G Y. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics*, **42**, 1053–1059.

Ji F Y, Wei W, Liu Y, Wang G P, Zhang Q, Xing Y, Zhang S H, Liu Z H, Cao Q Q, Qin L. 2018. Construction of a SNP-based high-density genetic map using Genotyping by Sequencing (GBS) and QTL analysis of nut traits in Chinese chestnut (*Castanea mollissima* Blume). *Frontiers in Plant Science*, **9**, 816.

Jiao Y Q, Vuong T D, Liu Y, Meinhardt C, Liu Y, Joshi T, Cregan P B, Xu D, Shannon J G, Nguyen H T. 2015. Identification and evaluation of quantitative trait loci underlying resistance to multiple HG types of soybean cyst nematode in soybean PI 437655. *Theoretical and Applied Genetics*, **128**, 15–23.

Lark K G, Chase K, Adler F, Mansur L M, Orf J H. 1995. Interactions between quantitative trait loci in soybean in which trait variation at one locus is conditional upon a specific allele at another. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 4656–4660.

Lee S, Jun T H, Michel A P, Mian M A R. 2015. SNP markers linked to QTL conditioning plant height, lodging, and maturity in soybean. *Euphytica*, **203**, 521–532.

Lee S, Mian M A R, McHale L K, Wang H, Wijeratne A J, Sneller

C H, Dorrance A E. 2013. Novel quantitative trait loci for partial resistance to *Phytophthora sojae* in soybean PI 398841. *Theoretical and Applied Genetics*, **126**, 1121–1132.

Lee S H, Bailey M A, Mian M A R, Carter T E, Ashley D A, Hussey R S, Parrott W A, Boerma H R. 1996. Molecular markers associated with soybean plant height, lodging, and maturity across locations. *Crop Science*, **36**, 728–735.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H H, Ribaut J M, Li Z L, Wang J K. 2008. Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. *Theoretical and Applied Genetics*, **116**, 243–260.

Li X P, Han Y P, Teng W L, Zhang S Z, Yu K F, Poysa V, Anderson T, Ding J J, Li W B. 2010. Pyramided QTL underlying tolerance to *Phytophthora* root rot in mega-environments from soybean cultivars 'Conrad' and 'Hefeng 25'. *Theoretical and Applied Genetics*, **121**, 651–658.

Li Y H, Liu B, Reif J C, Liu Y L, Li H H, Chang R Z, Qiu L J. 2014. Development of insertion and deletion markers based on biparental resequencing for fine mapping seed weight in soybean. *Plant Genome*, **7**, 1–8.

Li Y H, Shi X H, Li H H, Reif J C, Wang J J, Liu Z X, He S, Yu B S, Qiu L J. 2016. Dissecting the genetic basis of resistance to soybean cyst nematode combining linkage and association mapping. *Plant Genome*, **9**, 1–11.

Li Y H, Zhao S C, Ma J X, Li D, Yan L, Li J, Qi X T, Guo X S, Zhang L, He W M, Chang R Z, Liang Q S, Guo Y, Ye C, Wang X B, Tao Y, Guan R X, Wang J Y, Liu Y L, Jin L G, *et al.* 2013. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics*, **14**, 579.

Liu N X, Li M, Hu X B, Ma Q B, Mu Y H, Tan Z Y, Xia Q J, Zhang G Y, Nian H. 2017. Construction of high-density genetic map and QTL mapping of yield-related and two quality traits in soybean RILs population by RAD-sequencing. *BMC Genomics*, **18**, 466.

Liu Y C, Du H L, Li P C, Shen Y T, Peng H, Liu S L, Zhou G A, Zhang H K, Liu Z, Shi M, Huang X H, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C Z, Tian Z X. 2020. Pan-genome of wild and cultivated soybeans. *Cell*, **182**, 162–176.

Liu Y L, Li Y H, Reif J C, Mette M F, Liu Z X, Liu B, Zhang S S, Yan L, Chang R Z, Qiu L J. 2013. Identification of quantitative trait loci underlying plant height and seed weight in soybean. *Plant Genome*, **6**, 1–11.

Mansur L M, Orf J, Lark K G. 1993. Determining the linkage of quantitative trait loci to RFLP markers using extreme phenotypes of recombinant inbreds of soybean (*Glycine max* L. Merr.). *Theoretical and Applied Genetics*, **86**, 914–918.

Mansur L M, Orf J H, Chase K, Jarvik T, Cregan P B, Lark K G. 1996. Genetic mapping of agronomic traits using recombinant inbred lines of soybean. *Crop Science*, **36**, 1327–1336.

Patil G, Chaudhary J, Vuong T D, Jenkins B, Qiu D, Kadam S, Shannon G J, Nguyen H T. 2017. Development of SNP genotyping assays for seed composition traits in soybean. *International Journal of Plant Genomics*, **2017**, 1–12.

Qiu B X, Arelli P R, Sleper D A. 1999. RFLP markers associated with soybean cyst nematode resistance and seed composition in a 'Peking'×'Essex' population. *Theoretical and Applied Genetics*, **98**, 356–364.

Shen Y T, Liu J, Geng H Y, Zhang J X, Liu Y C, Zhang H K, Xing S L, Du J C, Ma S S, Tian Z X. 2018. *De novo* assembly of a Chinese soybean genome. *Science China* (Life Sciences), **61**, 871–884.

Shi Z, Liu S M, Noe J, Arelli P, Meksem K, Li Z L. 2015. SNP identification and marker assay development for high-throughput selection of soybean cyst nematode resistance. *BMC Genomics*, **16**, 314.

Song Q J, Jenkins J, Jia G F, Hyten D L, Pantalone V, Jackson S A, Schmutz J, Cregan P B. 2016. Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly *Glyma1.01*. *BMC Genomics*, **17**, 33.

Song X F, Wei H C, Cheng W, Yang S X, Zhao Y X, Li X, Luo D, Zhang H, Feng X Z. 2015. Development of InDel markers for genetic mapping based on whole genome resequencing in soybean. *G3 — Genes Genomes Genetics*, **5**, 2793–2799.

Springer N M, Ying K, Fu Y, Ji T M, Yeh C T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez A L, Barbazuk W B, Jeddeloh J A, Nettleton D, Schnable P S. 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics*, **5**, e1000734.

Tucker D M, Maroof M A S, Mideros S, Skoneczka J A, Nabati D A, Buss G R, Hoeschele I, Tyler B M, St Martin S K, Dorrance A E. 2010. Mapping quantitative trait loci for partial resistance to *Phytophthora sojae* in a soybean interspecific cross. *Crop Science*, **50**, 628–635.

Wang D, Graef G L, Procopiuk A M, Diers B W. 2004. Identification of putative QTL that underlie yield in interspecific soybean backcross populations. *Theoretical and Applied Genetics*, **108**, 458–467.

Wang K, Li M Y, Hakonarson H. 2010. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, **38**, e164.

Xia N, Yan W B, Wang X Q. Shao Y P, Yang M M, Wang Z K, Zhan Y H, Teng W L, Han Y P, Shi Y G. 2019. Genetic dissection of hexanol content in soybean seed through genome-wide association analysis. *Journal of Integrative Agriculture*, **18**, 1222–1229.

Yan L, Li Y H, Yang C Y, Ren S X, Chang R Z, Zhang M C, Qiu L J. 2015. Identification and validation of an over-dominant QTL controlling soybean seed weight using populations derived from *Glycine max*×*Glycine soja*. *Plant Breeding*, **133**, 632–637.

Yang L, Tian Y, Liu Y L, Reif J C, Li Y H, Qiu L J. 2021. QTL mapping of *qSCN3-1* for resistance to soybean cyst nematode in soybean line Zhongpin 03-5373. *The Crop*

Journal, **9**, 351–359.

Yu C, Wan H H, Peter M B, Cheng B X, Luo L, Pan H T, Zhang Q X. 2021. High density genetic map and quantitative trait loci (QTLs) associated with petal number and flower diameter identified in tetraploid rose. *Journal of Integrative Agriculture*, **20**, 1287–1301.

Yue P, Sleper D A, Arelli P R. 2001. Mapping resistance to multiple races of *Heterodera glycines* in soybean PI 89772. *Crop Science*, **41**, 1589–1595.

Zhang G R, Gu C H, Wang D C. 2009. Molecular mapping of soybean aphid resistance genes in PI 567541B. *Theoretical and Applied Genetics*, **118**, 473–482.

Zhang J P, Song Q J, Cregan P B, Jiang G L. 2016. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theoretical and Applied Genetics*, **129**, 117–130.

Zhou Z K, Jiang Y, Wang Z, Gou Z H, Lyu J, Li W Y, Yu Y J, Shu L P, Zhao Y J, Ma Y M, Fang C, Shen Y T, Liu T F, Li C C, Li Q, Wu M, Wang M, Wu Y S, Dong Y, Wan W T, *et al.* 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, **33**, 408–414.

Executive Editor-in-Chief ZHANG Xue-yong
Managing Editor WANG Ning