INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
19–0003/201x/00–0–000–000
DOI: 10.17957/IJAB/15.1039
http://www.fspublishers.org

Full Length Article

Identification of Genome-wide Insertion and Deletions for Sugar Beet (*Beta vulgaris* L.) Using Next-generation Re-sequencing

Naixin Liu[1,2], Ling Wang[1*], Yumei Wu[2*], Zedong Wu[2] and Wang Xing[3]
[1]*College of Landscape Architecture, Northeast Forestry University, Harbin 150040, China*
[2]*Academy of Crop Sciences, Heilongjiang University, Harbin 150080, China*
[3]*National Beet Germplasm Medium Term Storeroom, Harbin 150080, China*
[*]For correspondence: wanglinghlj@126.com; zjzxwym@163.com

Abstract

Molecular markers, based on variation of DNA level, are valuable tools in both basic and applied research. Insertion and deletions are the second abundant genetic variations after single nucleotide polymorphism in the whole genomes. Sugar beet is one of the most important sugar crops in the world that provides one third of worldwide sucrose production. Only few insertion and deletion markers were developed in sugar beet, thus the genetic study of sugar beet is need of the time. In the present study, five sugar beet varieties were re-sequenced and genome-wide insertion and deletions were identified compared with reference genome. A total of 42.91 Gb of sequences were obtained, with average sequencing depth of 14.75 and 99.84% of sequences aligned to reference genome. About 1.57×10^7 genome-wide insertion and deletions were identified, with average of 314,040 per variety, most of which were distributed at intergenic regions. Among them, 2.33×10^5 insertion and deletions were distributed at coding sequence region, and the majority of them caused non-frame insertion/deletion. The lengths of insertion and deletions were diverse and numbers of insertion and deletions decreased with increase of their lengths. A total of 676,944 insertion and deletions with length more than 3 bp were suitable for developing Insertion/Deletion markers. Findings of this study added useful and valuable information and resources for developing Insertion/Deletion markers, which can be used in genetic study, molecular breeding and variety identification in sugar beet. © 2019 Friends Science Publishers

Keywords: Sugar beet; Next-generation re-sequencing; Insertion and deletions; Distribution and number; Length

Introduction

Molecular markers are based on variation of DNA level and provide more accurate estimate than phenotypic and pedigree information (Li *et al.*, 2010). Therefore molecular markers are valuable tool in both basic and applied research (Mahmood *et al.*, 2016). With the development of molecular technology, different types of molecular markers have been developed and used in genotyping (Devran *et al.*, 2016; Stevanato and Biscarini, 2016; Qiu *et al.*, 2017; Hu *et al.*, 2018), genetic diversity analysis (Bashir *et al.*, 2015), variety identification (Ali *et al.*, 2017), assisted breeding (Luo *et al.*, 2017) and phylogenetic analysis (Dai *et al.*, 2016), such as Simple Sequence Repeat (SSR) (Qiu *et al.*, 2012), Insertion/Deletion (InDel) (Liu *et al.*, 2013) and Single Nucleotide Polymorphism (SNP) (Qiu *et al.*, 2017).

Insertion-deletion (InDel) is insertion and/or deletion of nucleotides less than 1kb at DNA level, which is the second most abundant genetic variation in genomes after SNP(Liu *et al.*, 2015b) and has been successfully used in genetic studies in many crops, such as rice (*Oryza sativa* L.) (Liu *et al.*, 2015a), maize (*Zea mays* L.) (Liu *et al.*, 2015b), soybean (*Glycine* max L.) (Wang *et al.*, 2018), cucumber (*Cucumis sativus*) (Li *et al.*, 2013), pepper (*Capsicum annuum*) (Guo *et al.*, 2015) and apple (*Malus domestica* apple) (Liu *et al.*, 2017). InDel has large advantages such as co-dominant, multi-allelic, easy to use and inexpensive, thus it can be developed as desired molecular markers for genetic studies and crop breeding (Mahmood *et al.*, 2016). However, development of InDel markers is based on sequence information of two or more individuals, thus it is harder to develop these marker than other molecular markers.

Sugar beet (*Beta vulgaris* L.) is one of the most important sugar crops and provides one third of worldwide sucrose production (Joshi *et al.*, 2005). It's a diploid species with 2n=18 chromosomes and genome size of 758 Mb (Arumuganathan and Earle, 1991). A large number of beet materials and wild species were analyzed genetic diversity by morphology (Amirian *et al.*, 1981) and biochemistry (Srivastava *et al.*, 2007). This strategy has some disadvantages that it would cost much time, money and labor and have low accuracy. Till now, some types of molecular markers have been used in sugar beets, such as amplified fragment length polymorphism (AFLP) (Amirian *et al.*, 1981), SSR (Li *et al.*, 2010) and SNP (Stevanato and Biscarini, 2016). However, because of no sequence information obtained in sugar beet, there are few InDel markers developed in sugar beet.

To cite this paper: Liu, N., L. Wang, Y. Wu, Z. Wu and W. Xing, 201x. Identification of genome-wide insertion and deletions for sugar beet (*Beta vulgaris* L.) using next-generation re-sequencing. *Intl. J. Agric. Biol.* 00: 000–000

Next generation sequencing is powerful tool for discovering high density of molecular markers, because it could produce large quantity of sequences at a significantly lower cost (Liu *et al.*, 2015a). Moreover, this strategy enables more efficient re-sequencing of a large number of genomes, which could be compared and identified high density of InDels. However, the accuracy of next-generation re-sequencing is relative lower than first generation sequencing. The solution to solve this problem is sequencing more times. Along with increase of the sequencing depth, the accuracy increases. The first sugar beet variety KWS2320 has been sequenced using the Roche/454, Illumina and Sanger sequencing platforms (Dohm *et al.*, 2014) and was reported on NCBI (https://www.ncbi.nlm.nih.gov/) in 2014, with the size of reference genome about 567Mbp. The successful sequencing of KWS2320 brought us important hope to obtain whole genome sequences of other varieties and identify large number of InDels to develop markers.

In this study, five sugar beet varieties were re-sequenced using Illumina sequencing platform. Then, they were compared their genome sequences with the reference sugar beet KWS2320, respectively. Finally, both genome-wide InDels and InDels within coding sequence region (CDS) were identified. Findings of this study will add useful and valuable information and resources for developing Insertion/Deletion markers, which can be used in genetic studies, molecular breeding and variety identification in sugar beet.

Materials and Methods

Plant Materials and DNA Extraction

Five sugar beet varieties were used for whole-genome re-sequencing. Among them, ZT000549 and ZT000589 were elite varieties widely grown in north of China, and other three varieties *i.e.*, ZT000286, MA3001 and KWS1231 were introduced from Poland, Denmark and Germany, respectively. All varieties were grown in greenhouse of Heilongjiang University at Haerbin (45.6°N, 127.5°E) in China. Young leaves of 20 plants of each variety at 2-week seedling stage were bulked and genomic DNA was isolated using CTAB method with minor modification (Wen and Zhang, 2012).Then the extracted DNA was checked using 1% agarose gel. DNA with no degradation and pollution of RNA could be used for re-sequencing.

Whole-genome Re-sequencing by Illumina Sequencing Platform

The five sugar beets varieties were whole-genome re-sequenced using Illumina sequencing platform by Novogene Inc. (Beijing, China). First, DNA was checked its purity and concentration using Danodrop and Qubit (Thermo Fisher Inc., China). DNA with OD between 1.8-2.0 and concentration above 1.5 μg/μL could be used for re-sequencing. Secondly, it was randomly broken into about 350 bp fragments using Covaris ME220 (Covaris Inc., America). The fragments were used to constructed library using TruSeq Library Construction Kit (Takara Bio Inc., China). The library construction was as follows: The fragments were phosphorylate their ends, added A-talling, then ligate Index adapter and denature and amplify for final product. Finally, samples were sequenced by Illumina HiSeq platform.

Sequence Alignment and InDel Identification

Raw reads from Illumina HiSeq platform were first subjected to quality control procedure to obtain clean reads. Three types of reads were removed: The reads containing the Illumina library construction adapters; the reads containing more than 10% unknown bases; one end of the read containing more than 50% of low quality bases (Q <5). Then clean reads were aligned to the reference sequence by BWA software, a new read alignment package named Burrows-Wheeler Alignment tool (Li and Durbin, 2009). Duplications were removed using SAMtools software (Li *et al.*, 2009). InDels were detected using SAMtools software with the parameters as "mpileup -m 2 -F 0.002 -d 1000" and annotated using ANNOVAR software (Wang *et al.*, 2010).

Results

Sequence Statistics of Five Varieties

A total of 2.86×10^9 original reads and 42.91 Gb of sequences were obtained, with each of 4.90×10^8-6.32×10^8 reads and 7.35-9.49 Gb of sequences (Table 1). After quality controlling, 2.71×10^9 raw reads and 42.84 Gb (99.84%) of sequences could aligned to reference genome, with average of 5.42×10^8 raw reads and 8.57 Gb of sequences per variety. The sequencing depths of ZT000589, ZT000549, ZT000286, MA3001 and KWS1231 were 14.95, 15.09, 12.87, 16.41 and 14.41 respectively, with average of 14.75.

Genome-wide InDel Distributions

Genome-wide InDel distribution of all five varieties was listed in Table 2. There were 312464, 316499, 281356, 343025 and 316956 InDels between five varieties (ZT000589, ZT000549, ZT000286, MA3001 and KWS1231) and KWS2320 respectively. InDel numbers and distributions were similar between all five varieties and reference genome. For all five varieties, most InDels were distributed in intergenic regions (about 55% of all InDels), followed by genic region and ncRNA had the least InDels. Within genic regions, most InDels were distributed in introns, followed by upstream and downstream, and CDS had the least InDels.

Table 1: The statistics of sequence information of five sugar beets

Variety	Total reads	Raw bases (bp)	Mapping reads	Clean bases (bp)	Sequencing depth
ZT000589	58044762	8720521800	54576874	8706714300	14.95
ZT000549	60717192	9118802400	55359194	9107578800	15.09
ZT000286	48951380	7351715400	47019099	7342707000	12.87
MA3001	63160910	9487149300	60856505	9474136500	16.41
KWS1231	54788586	8229593700	52941659	8218287900	14.41
Average	57132566	8581556520	54150666	8569884900	14.75

Total reads: reads number of original sequence data; Raw bases: base number of original sequence data; Mapping reads: reads number which could be aligned to reference genome; Clean bases: base number which could be aligned to reference genome; Sequencing depth: clean bases divided by reference genome

Table 2: Distribution of genome-wide InDels between five sugar beets and reference genome, respectively

Location in the genome		ZT000589	ZT000549	ZT000286	MA3001	KWS1231
Genic region	Upstream	16112	16464	14227	17922	16520
	5'UTR	7810	7976	6949	8236	7756
	CDS	4828	4810	4473	4975	4824
	Intron	82214	81317	73649	85835	80073
	3'UTR	7680	7700	6842	7914	7459
	Downstream	15975	16001	14112	17294	15942
ncRNA	5'UTR	54	51	41	47	39
	CDS	1375	1397	1281	1446	1312
	Intron	3212	3165	2952	3316	3162
	3'UTR	49	48	45	61	45
Intergenic region		173055	177570	156785	195979	179824
Total		312364	316499	281356	343025	316956

Table 3: Distribution of InDels in CDS regions causing large variations between five sugar beets and reference genome, respectively

Result of InDels	ZT000589	ZT000549	ZT000286	MA3001	KWS1231
Stop gain	54	66	69	68	61
Stop loss	11	14	9	10	11
Frameshift deletion	790	817	760	872	826
Frameshift insertion	694	634	601	737	674
Non-frameshift deletion	1619	1623	1501	1642	1628
Non-frameshift insertion	1532	1535	1429	1534	1508
Total	4700	4689	4369	4863	4708

For InDels length, the trends were also the same for all five varieties (Fig. 1). InDels number decreased with increase of InDel length. All InDel could be divided into four types: more than 110,000 InDels (about 40% of all InDels) of each variety were 1 bp; 20,000-80,000 InDels were range from 2 to 4 bp; about 10,000 InDels were 5-8 bp; less than 7,000 InDels were more than 9 bp.

InDel Distribution in CDS Region

Like genome-wide InDels, numbers and distributions of InDels in CDS regions were also similar for all five varieties compared with KWS2320 (Table 3). There were 4828, 4810, 4473, 4975 and 4824 InDels within CDS between five varieties (ZT000589, ZT000549, ZT000286, MA3001 and KWS1231) and reference genome, and almost all of them caused significant change of encoding protein (Table 2). Most of them (about 1,500 InDels) caused nonframeshift deletion or insertion, which would insert or delete one or a few amino acids of encoding protein. About 600-800 InDels caused frameshift deletion or insertion, and below 100 InDels caused stop loss or stop gain. These two types of InDels would finally change sequence of encoding protein largely.

InDel numbers decreased with increase of length of InDels (Fig. 2). When InDel length was multiple of three, the numbers would be largely more than that of two adjacent lengths. For example, length of 1736 InDels was 3 bp while length of 212 and 171 InDels were 2 bp and 4 bp respectively for ZT000589.

Discussion

In the present study, a total of five sugar beet varieties were re-sequenced using Illumina sequencing platform and they were compared their genome-wide sequences with reference genome of KWS2320, respectively. Most InDels (about 55%) were distributed in intergenic regions (Table 2), it was consistent with some other species (Liu *et al.*, 2015b). Within genic regions, most InDels (more than 60%) were distributed in introns. These variation distributions may be caused by no change of gene functions or expressions. If InDels were distributed in CDS, the sequence of coding protein will be changed and gene function would be finally changed or even destroyed (Song *et al.*, 2007; Ishimaru *et al.*, 2013). If the InDels were distributed in promoter regions, the gene expressions would be increased or decreased (Bai *et al.*, 2017).

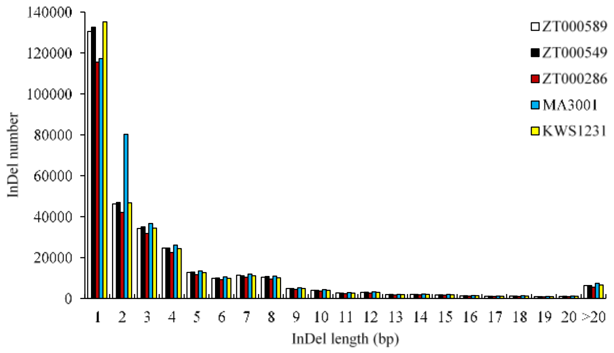

Fig. 1: Distribution of genome-wide InDel length between five sugar beets and reference genome, respectively

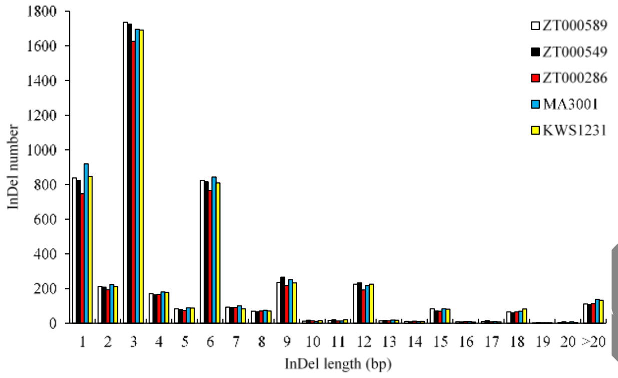

Fig. 2: Distribution of InDels length in CDS regions between five sugar beets and reference genome, respectively

These two types of variations would ultimately change plants phenotype or adaptability, thus most of them were eliminated. However, InDels distributed in intergenic or inton regions wouldn't cause so much change and had much probability to be retained (Leister, 2005).

The lengths of InDel were also diverse and InDel numbers decreased with increase of InDel lengths, either in whole genome-wide or in CDS region (Fig. 1 and 2). This phenomenon may be also resulted from higher survival probability of plants with shorter InDel length. Either insertion or deletion would damage genome, lose some genes and finally hurt plants. When the length of InDel was shorter, the damage was slighter, and the plant had more probability to survive. Thus, this variation had higher probability to be retained. Interestingly, numbers of InDel with length of multiple of three were much more than that of two adjacent lengths in CDS regions, and they were consistent with InDel location distributions in these regions (Table 3). We all know that three nucleotides encode one amino acid, thus InDels with length of multiple of three would cause non-frame shift, and this type of variation had relatively small change of coding protein sequences compared with frameshift or stop loss/gain. Individuals containing this variation would have higher survival probability, and it had more chances to be retained.

Till now, most genotyping was operated on polyacrylamise and agarose gel electrophoresis, and these two gels had different resolutions. PCR products with length of 60-100 bp and major differences equal or greater than 3 bp could be solved at polyacrylamise gel, while PCR products with length of 150-300 bp and major differences equal or greater than 8 bp could be solved at agarose gel (Liu *et al.*, 2015a; Liu *et al.*, 2015b). Thus, InDels with length equal or greater than 3 bp could be developed for InDel markers. In the present study, lengths of 135,639, 137,011, 123,776, 145,478 and 135,040 InDels were more than 3 bp between five varieties (ZT000589, ZT000549, ZT000286, MA3001) and KWS1231 respectively (Fig. 1), with average of 238.78 InDels per Mb. All of them were suitable for developing InDel markers with high throughput and would have large value for genotyping, genetic diversity analysis, variety identification, assisted breeding and phylogenetic analysis for sugar beet.

Conclusion

InDel numbers and distributions were similar between all five varieties and reference genome, either for genome-wide or CDS region. Among them, a total of 676,944 InDels with length more than 3 bp were suitable and had large values for developing InDels markers in the future.

Acknowledge

This work was supported by Fundamental Research Funds for the Central University (2572016EAJ6) and Sugar Crop Research System in China (CARS-170503).

References

Ali, A., J.D. Wang, Y.B. Pan, Z.H. Deng, Z.W. Chen, R.K. Chen and S.J. Gao, 2017. Molecular identification and genetic diversity analysis of Chinese Sugarcane (*Saccharum* spp. hybrids) varieties using SSR markers. *Trop. Plant Biol.*, 10**:** 1–10

Amirian, R., M.R. Naghavi, A.A.S. Busheri and M. Omidi, 1981. Evaluation of genetic diversity in accessions using morphological and AFLP markers. *Plant Physiol.*, 142: 135–147

Arumuganathan, K. and E.D. Earle, 1991. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.*, 9: 208

Bai, X., Y. Huang, Y. Hu, H. Liu, B. Zhang, C. Smaczniak, G. Hu, Z. Han and Y. Xing, 2017. Duplication of an upstream silencer of FZP increases grain yield in rice. *Nat. Plants*, 3: 885–893

Bashir, E.M.A., A.M. Ali, A.M. Ali, E.T.I. Mohamed, A.E. Melchinger, H.K. Parzies and B.I.G. Haussmann, 2015. Genetic diversity of Sudanese pearl millet (*Pennisetum glaucum* (L.) R. Br.) landraces as revealed by SSR markers, and relationship between genetic and agro-morphological diversity. *Genet. Resour. Crop Evol.*, 62: 579–591

Dai, S.F., J.Q. Jiang, Y.N. Jia, X.F. Xue, D.C. Liu, Y.M. Wei, Y.L. Zheng and Z.H. Yan, 2016. Molecular characterization and phylogenetic analysis of Wx genes from three Taeniatherum diploid species. *Biol. Plant.*, 60: 1–8

Devran, Z., A. Göknur and L. Mesci, 2016. Development of molecular markers for the Mi-1 gene in tomato using the KASP genotyping assay. *Hortic. Environ. Biotechnol.*, 57: 156–160

Dohm, J.C., A.E. Minoche, D. Holtgräwe, S. Capellagutiérrez, F. Zakrzewski, H. Tafer, O. Rupp, T.R. Sörensen, R. Stracke and R. Reinhardt, 2014. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature*, 505: 546

Guo, G.J., Q. Sun, J.B. Liu, B.G. Pan, W.P. Diao, G.E. Wei, C.Z. Gao and S.B. Wang, 2015. Development and application of pepper InDel markers based on genome re-sequencing. *Jiangsu J. Agric. Sci.,* 31: 1400–1406

Hu, H., W.K. Lv, Q.L. Li, X.X. Ou, J.Y. Xu, Z.X. Li, D.Y. Xing, L.W. Yang, J.L. Xu, X.J. Qiu, T.Q. Zheng, X.Y. Wang, J.F. Jiang and Z.Y. Liu, 2018. Characterization of main effects, epistatic effects and genetic background effects on QTL for yield related traits by two sets of reciprocal introgression lines in rice (*Oryza sativa*). *Intl. J. Agric. Biol.,* 20: 2125–2132

Ishimaru, K., N. Hirotsu, Y. Madoka, N. Murakami, N. Hara, H. Onodera, T. Kashiwagi, K. Ujiie, B. Shimizu, A. Onishi, H. Miyagawa and E. Katoh, 2013. Loss of function of the IAA-glucose hydrolase gene TGW6 enhances rice grain weight and increases yield. *Nat. Genet.,* 45: 707–11

Joshi, S.S., M.W. Pawar, S.S. Datir and D.B. More, 2005. Physiological studies and sucrose metabolism during root development in three sugar beet cultivars. *Sugar Tech.,* 7: 150–153

Leister, D., 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet.,* 21: 655–63

Li, H. and R. Durbin, 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25: 1754–60

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing, 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25: 2078–9

Li, J.Q., B. Schulz and B. Stich, 2010. Population structure and genetic diversity in elite sugar beet germplasm investigated with SSR markers. *Euphytica,* 175: 35–42

Li, S.G., D. Shen, B. Liu, Y. Qiu, X.H. Zhang, Z.H. Zhang, H.P. Wang and X.X. Li, 2013. Development and application of cucumber InDel markers based on genome re-sequencing. *J. Plant Genet. Res.,* 14: 278–283

Liu, B., Y. Wang, W. Zhai, J. Deng, H. Wang, Y. Cui, F. Cheng, X. Wang and J. Wu, 2013. Development of InDel markers for *Brassica rapa* based on whole-genome re-sequencing. *Theor. Appl. Genet.,* 126: 231–239

Liu, J., J. Li, J.T. Qu and S.Y. Yan, 2015a. Development of genome-Wide Insertion and Deletion polymorphism markers from next-generation sequencing data in rice. *Rice,* 8: 27

Liu, J., J.T. Qu, C. Yang, D.G. Tang, J.W. Li, H. Lan and T.Z. Rong, 2015b. Development of genome-wide insertion and deletion markers for maize, based on next-generation sequencing data. *BMC Genom.,* 16: 601

Liu, Y.X., J.H. Lan, S.H. Bai, X.H. Sun, C.X. Liu, Y.G. Zhang and H.Y. Dai, 2017. Screening of SNP and InDel markers to Glomerella leaf spot resistance gene locus in apple using HRM technology. *Acta Hortic. Sin.,* 44: 215–222

Luo, Y.C., T.C. Ma, A.F. Zhang, K.H. Ong, Z.X. Luo, Z.F. Li, J.B. Yang and Z.C. Yin, 2017. Marker-assisted breeding of Chinese elite rice cultivar 9311 for disease resistance to rice blast and bacterial blight and tolerance to submergence. *Mol. Breed.,* 37: 106

Mahmood, S., Z.H. Li, X.P. Yue, B. Wang, J. Chen and K.D. Liu, 2016. Development of INDELs markers in oilseed rape (*Brassica napus* L.) using re-sequencing data. *Mol. Breed.,* 36: 79

Qiu, X.J., K. Chen, W.K. Lv, X.X. Ou, Y.J. Zhu, D.Y. Xing, L.W. Yang, F.J. Fan, J.J. Yang, J.L. Xu, T.Q. Zheng and Z.K. Li, 2017. Examining two sets of introgression lines reveals background-independent and stably expressed QTL that improve grain appearance quality in rice (*Oryza sativa* L.). *Theor. Appl. Genet.,* 130: 951–967

Qiu, X.J., R. Gong, Y.B. Tan and S.B. Yu, 2012. Mapping and characterization of the major quantitative trait locus qSS7 associated with increased length and decreased width of rice seeds. *Theor. Appl. Genet.,* 125: 1717–1726

Song, X.J., W. Huang, M. Shi, M.Z. Zhu and H.X. Lin, 2007. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.,* 39: 623–30

Srivastava, S., P.S. Gupta, S.V. Kumar and S.H. Mohan, 2007. Genetic diversity analysis in Sugar beet (*Beta vulgaris* L.) using isozymes, RAPD and ISSR markers. *Cytol. Intl. J. Cytol.,* 72: 265–274

Stevanato, P. and F. Biscarini, 2016. Digital PCR as new approach to SNP genotyping in Sugar Beet. *Sugar Tech.,* 18: 429–432

Wang, J.L., L.P. Kong, K.C. Yu, F.G. Zhang, X.Y. Shi, Y.P. Wang, H.Y. Nan, X.H. Zhao, S.J. Lu and D. Cao, 2018. Development and validation of InDel markers for identification of QTL underlying flowering time in soybean. *Crop J.,* 6: 126–135

Wang, K., M. Li and H. Hakonarson, 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl. Acids Res.,* 38: e164

Wen, D. and C. Zhang, 2012. Universal Multiplex PCR: a novel method of simultaneous amplification of multiple DNA fragments. *Plant Methods,* 8: 32

(Received 02 January 2019; Accepted 12 February 2019)