

BSA-SEQ-BASED METHOD FOR LOCATING KEY GENETIC SEGMENTS OF PEDUNCLE LENGTH IN BREWING DWARF SORGHUM [*SORGHUM BICOLOR* (L.) MOENCH]

YANG, G. D.¹ – LI, J. H.¹ – HU, Z. Y.^{1*} – HAO, Z. Y.¹ – SUN, B. S.¹ – LIU, C. S.¹ – WANG, Q.² – MENG, X. X.² – GUO, W.³

¹*Keshan Branch of Heilongjiang Academy of Agricultural Sciences, 161005 Heilongjiang, China*

²*Crop Resources Institute of Heilongjiang Academy of Agricultural Sciences, 150086 Heilongjiang, China*

³*Quality & Safety Institute of Agricultural Products, Heilongjiang Academy of Agricultural Science, Harbin, China*

The first two authors contributed equally to this paper.

**Corresponding author
e-mail: Liping04230319@126.com*

(Received 11th May 2023; accepted 19th Jul 2023)

Abstract. Sorghum peduncle length is one of the key agronomic traits in sorghum production and has a critical impact on sorghum mechanization for harvesting. In this study, an F2 segregating population was constructed with significantly different long-peduncle KY133B and short-peduncle KY123B as parents, and the sorghum peduncle length gene was preliminarily located using BSA-seq technology. The association analysis was performed by ED algorithm and SNP-index (or InDel-index) algorithm, and finally the sorghum peduncle length association region was targeted on chromosome 7 and 10. In this study, the BSA-seq technique was used to rapidly and efficiently locate key genetic segments of sorghum peduncle length, which laid the foundation for subsequent functional validation and molecular studies of peduncle length genes.

Keywords: *sorghum, peduncle length, BSA-seq, SNP; InDel*

Introduction

Sorghum [*Sorghum bicolor* (L.) Moench, 2n=20] is the fifth most important cereal crop in the world after corn, rice, wheat, and barley, mainly used for food and forage (Dabija et al., 2021). In the 1980s, China's sorghum breeding goals began to change, from the original goal of high yield breeding for food to brewing, forage and other goals of special sorghum breeding. Brewing sorghum gradually became the main body of sorghum breeding and production, planted area accounted for about 85% of the national sorghum production. The selection of dwarf stalks, dense planting, resistance, suitable for mechanization of brewing sorghum varieties become particularly urgent. Sorghum peduncle length is one of the key agronomic traits in sorghum production and plays an important role in mechanized harvesting. QTL localization studies for agronomic traits of sorghum such as plant height (Lin et al., 1995), panicle length (Rami et al., 1998), post-flowering greenness (Kassahun et al., 2010), and root growth angle (Mace et al., 2012) have been widely reported, but less on peduncle length, and the method of localizing peduncle length is based on constructing genetic maps for QTL localization. Su (2012) identified one QTL associated with peduncle length in sorghum, located on

chromosome 1; Zhai (2010) identified one QTL associated with peduncle length, located on chromosome 6; Ding et al. (2023) identified ten QTL associated with peduncle length, localized on chromosomes 1, 3, 6, 7, and 8 respectively. In this study, BSA-seq technique was compared with the sorghum reference genome to rapidly and efficiently locate the sorghum peduncle length association region, which laid the foundation for subsequent functional validation and molecular studies of peduncle length genes.

Materials and Methods

Test materials

The F₂ segregating population was constructed by using long-peduncle KY133B and short-peduncle KY123B as parents (*Figure 1*) and harvested with a total of 339 plants. Phenotypic analysis of peduncle length traits was performed using SPSS 16.0, and the phenotypic data of peduncle length (the distance from the last node of the stalk to the base of the panicle) were judged to be in accordance with normal distribution.

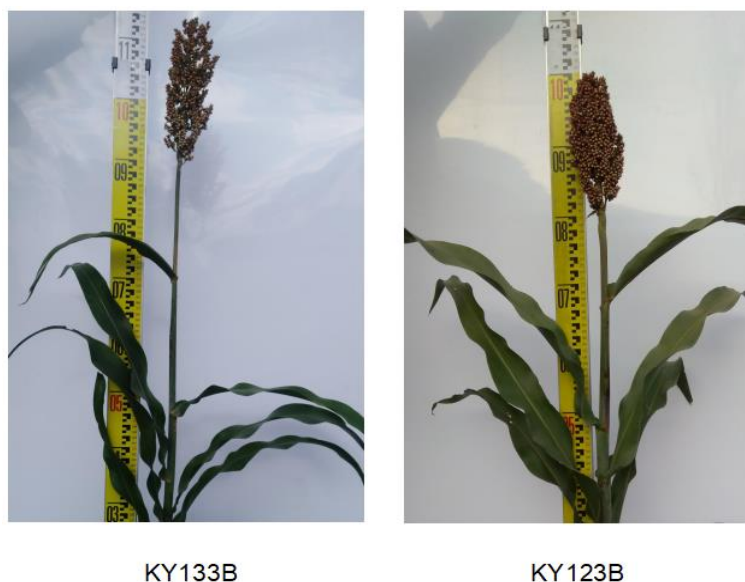


Figure 1. Parents

Constructing extreme mixing pools

All seeds of KY133B, KY123B, and F₂ populations were sown at the base of Keshan Branch of Heilongjiang Academy of Agricultural Sciences (48°01'N, 125°83'E), and 30 long-peduncle plants and 30 short-peduncle plants were identified to construct DNA mixing pools for each of the two extreme traits.

Mixing pool DNA extraction and sequencing

DNA extraction, quality control and sequencing were finished by Shanghai Majorbio Bio-Pharm Technology Co., Ltd. Sequencing was performed with Illumina NovaSeqTM platform, and the sequencing strategy was Illumina PE150 with a total sequencing read length of 300 bp. The reference genome was *Sorghum bicolor* NCBIv3 version (https://www.ncbi.nlm.nih.gov/genome/108?genome_assembly_id=321335).

Bioinformatics analysis

Comparison with the sorghum genome was performed by the BWA software (Li et al., 2009). Variant detection of SNP was implemented using GATK software (McKenna et al., 2010), and InDel sites were annotated using SNPEff software (Cingolani et al., 2012). High-quality SNP and InDel credible sites were obtained. Candidate regions associated with traits were calculated using the Euclidean distance (ED) algorithm (Hill et al., 2013) and the SNP-index(or InDel-index) algorithm (Fekih et al., 2013). The Δ SNP-index (or Δ InDel-index) values were counted. BLAST software (Altschul et al., 1997) was applied to compare with NR (Deng et al., 2006), GO (Ashburner et al., 2000), KEGG (Kanehisa et al., 2004), EggNOG (Huerta-Cepas et al., 2019) and Uniprot (2021) databases for in-depth annotation.

Results and Analysis

Data quality assessment

Using Illumina's sequencing platform, whole genome resequencing was performed on two offspring extreme mixed pools and two parents, and 27.75G of high-quality clean reads data were obtained after filtering. Q30 \geq 92.32%, GC content between 42.99% and 43.48%, the insert length was normally distributed, average sample to reference genome matching efficiency of 99.13%, the average genome coverage depth was about 35.48X, and the genome coverage was about 91.82% (at least one base coverage) (Table 1). The results showed that the data met the analysis criteria.

Table 1. Sample data and reference genome comparison results

Sample	Clean Reads	Clean Q30 (%)	Clean GC (%)	Mapped Ratio (%)	Coverage 1X (%)	Coverage 5X (%)	Average Depth of Coverage
Paternal pool	170994818	92.32	43.48	99.09	91.14	88.94	35.94
Maternal pool	170557062	92.58	42.99	99.25	91.44	89.38	35.9
Mixed pool of long peduncle	165951118	92.46	43.27	99.14	92.33	90.42	34.89
Mixed pool of short peduncle	167581834	92.77	43.42	99.05	92.38	90.49	35.18

F2 population phenotypic analysis

Phenotyping of 339 F2 plants was analyzed based on the mean, standard deviation, coefficient of variation, variation range, Skewness, Kurtosis and S-W test (Table 2). The absolute value of Skewness coefficient and Kurtosis coefficient is less than 1.96, and the distribution is normal by S-W test.

Table 2. Phenotypic analysis of peduncle length in F2 population

Traits	Mean (cm)	SD (cm)	CV (%)	Range (cm)	Skewness	Kurtosis	Shapiro-Wilk test
Peduncle Length	30.95	5.74	18.54	16.5~45.5	0.083	-1.053	0.994**

** : significant difference at the 0.01 probability levels

SNP and InDel association analysis

The ED and SNP-index (or InDel-index) algorithm was used for association analysis (Table 3).

Table 3. Gene region information statistics for association analysis

Method of association	Chromosome	Associated area Start (bp)	Associated area End (bp)	Area Size (Mb)	Number of gene	Number of effective mutation genes
ED	Chr.10	51040000	51520000	0.48	29	13
	Total			0.48	29	13
SNP-index(or InDel-index)	Chr.7	56061761	56233045	0.171	3	0
	Chr.7	56488110	57489673	1.002	86	21
	Chr.7	58085816	59418341	1.333	105	40
	Chr.10	6704981	6919926	0.215	16	4
	Chr.10	7390168	7409582	0.019	3	1
	Chr.10	7437745	7445806	0.008	1	1
	Chr.10	7460354	9779782	2.319	173	85
	Chr.10	10040745	10169611	0.129	7	4
	Chr.10	10224722	10279790	0.055	1	1
	Chr.10	10351177	11449674	1.098	42	25
	Chr.10	11530311	12215348	0.685	32	13
	Chr.10	13250087	13508590	0.259	4	1
	Chr.10	14530751	14683328	0.153	3	1
	Chr.10	15700653	17665682	1.965	46	3
	Chr.10	19804761	20095882	0.291	5	1
	Chr.10	20120578	20509675	0.389	10	2
	Chr.10	20540018	22679012	2.139	18	8
	Chr.10	22696405	23409374	0.713	7	2
	Chr.10	25410396	25748464	0.338	1	0
	Chr.10	25813814	27048540	1.235	4	3
	Chr.10	27878343	27999746	0.121	4	2
	Chr.10	28580778	28754411	0.174	1	1
	Chr.10	33470353	33699332	0.229	4	0
	Chr.10	33740229	34856631	1.116	16	2
	Chr.10	34871941	36979556	2.108	13	5
	Chr.10	38441713	39534226	1.093	5	0
	Chr.10	40558082	40758538	0.2	1	0
	Chr.10	41162274	41952700	0.79	16	0
	Chr.10	42930213	44219974	1.29	30	1
	Chr.10	44270048	45609222	1.339	25	11
	Chr.10	46755512	48119795	1.364	40	24
	Chr.10	48230242	48244051	0.014	1	0
Chr.10	48260257	48839953	0.58	13	7	
Chr.10	49085145	55267222	6.182	425	152	
Total				31.12	1161	421
Intersection area	Chr.10	51040000	51520000	0.48	29	13
	Total			0.48	29	13

Effective mutation gene: mutation gene of moderately and highly putative impact

According to the association threshold of the ED algorithm (Figure 2A), one chromosomal region associated with sorghum peduncle length was obtained, and the total length of genes annotated in the region was 0.48 Mb, containing a total of 29 genes, including 13 effective mutation genes; according to the association threshold of the SNP-index (or InDel-index) algorithm (Figure 2B), a total of 34 chromosomal regions associated with sorghum peduncle length were obtained, and the total length of the annotated chromosomal regions was 31.12 Mb, containing 1161 genes, including 421 effective mutation genes.

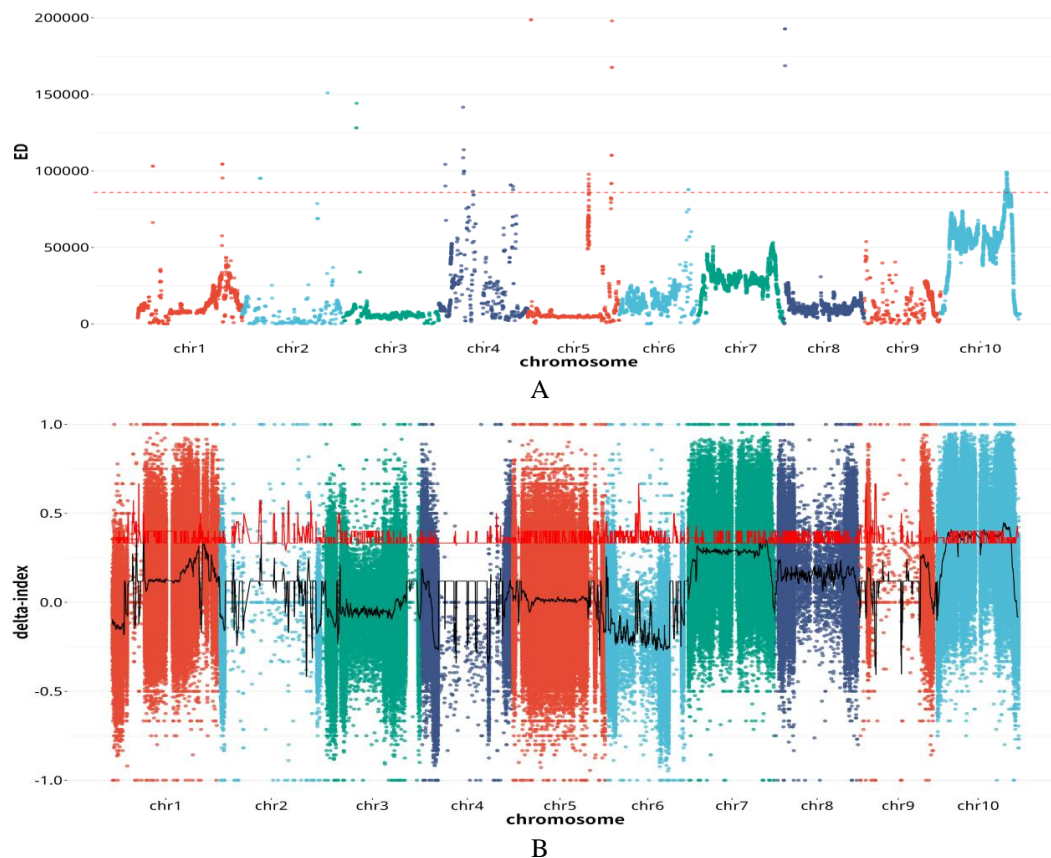


Figure 2. Distribution of association values on chromosomes. A: ED correlation analysis. The horizontal coordinate is the distribution of chromosomes, and each point represents the ED value of SNP (or InDel). The red dashed line represents the confidence level of 0.995 the threshold line. B: SNP-index (or InDel-index) correlation analysis. The horizontal coordinate is the distribution of chromosomes, and each dot represents the delta-index value of SNP (or InDel). The red line represents the confidence level of 0.95 the threshold line, the black line represents the delta-index value of SNP (or InDel) after fitting

Candidate segment screening and gene annotation

In this study, the results of the association analysis of the ED algorithm were intersected with those of the SNP-index (or InDel-index) algorithm, and the sorghum peduncle length association region was finally targeted on chromosome 10, with a total gene length of 0.48 Mb and 29 genes in the region. The BLAST software was applied to

annotate the genes in the candidate area with multiple databases in depth. The functionally annotated genes were gene-LOC8065033, gene-LOC8065034, gene-LOC8065035, gene-LOC8065037, gene-LOC8065039, gene-LOC8065040 gene-LOC8079323, gene-LOC8079325, and the functions of the remaining 21 genes are still unclear and to be verified (Table 4). One of the genes, gene-LOC8065033, controls the synthesis of the transcriptional elongation regulator MINIYO (IYO) and is speculated to be a candidate gene associated with sorghum peduncle length trait.

Table 4. Gene and function annotation in location range

Gene ID	Start (bp)	End (bp)	Gene function annotation
LOC8065033	51041203	51051697	transcriptional elongation regulator MINIYO
LOC8065034	51067297	51074272	serine/threonine-protein phosphatase PP2A-1 catalytic subunit
LOC8065035	51088997	51091343	protein far1-related sequence 5
LOC8065037	51336036	51345249	protein NRT1/ PTR family 4.6 isoform X2
LOC8065038	51422040	51430431	probable histone deacetylase 19
LOC8065039	51434295	51436378	pectate lyase
LOC8065040	51459580	51463289	receptor kinase-like protein Xa21
LOC8065041	51480362	51483286	uncharacterized protein LOC8065041 isoform X2
LOC8079310	51076106	51080033	putative receptor protein kinase ZmPK1
LOC8079311	51082103	51084859	putative receptor protein kinase ZmPK1
LOC8079312	51081797	51088681	putative receptor protein kinase ZmPK1
LOC8079313	51110482	51111051	uncharacterized protein LOC8079313
LOC8079314	51128255	51133018	uncharacterized protein LOC8079314 isoform X1
LOC8079315	51186408	51189882	hypothetical protein BDA96_10G231100
LOC8079319	51295599	51297503	uncharacterized protein LOC8072134
LOC8079320	51300732	51301650	probable glutathione S-transferase GSTU1
LOC8079321	51439955	51442414	uncharacterized protein LOC8079321
LOC8079323	51466537	51476092	ABC transporter A family member 7
LOC8079324	51493325	51498235	probable LRR receptor-like serine-/threonine-protein kinase At1g14390
LOC8079325	51498383	51501881	polyadenylate-binding protein 3
LOC8079326	51512873	51517274	probable LRR receptor-like serine-/threonine-protein kinase At1g34110
LOC110430693	51450324	51452230	hypothetical protein BDA96_10G232500
LOC110430928	51488346	51489613	hypothetical protein BDA96_10G233000
LOC110431047	51410519	51418494	uncharacterized protein LOC110431047 isoform X1
LOC110431173	51329212	51332919	uncharacterized protein LOC110431173
LOC110431241	51339606	51341007	hypothetical protein BDA96_10G231900
LOC110431255	51454603	51455502	probable pathogenesis-related protein ARB_02861 isoform X1
LOC110431325	51378066	51381259	hypothetical protein BDA96_10G232000
LOC110431432	51059075	51062201	uncharacterized protein LOC110431432

Discussion

QTL associated with sorghum peduncle length trait have been reported to be localized on chromosomes 1, 3, 6, 7, and 8 (Zhai, 2010; Su, 2012; Ding et al., 2023). In this study, the ED and SNP-index (or InDel-index) algorithm were used to locate the association region for the peduncle length trait on chromosomes 7 and 10. The difference from the results of previous studies may be related to the different materials used to construct the genetic population and the different ecological environment. Ding et al. (2023) localized the QTL associated with the peduncle length trait in the region of 58.85~62.98Mb on chromosome 7; Klein et al. (2001) localized the QTL associated with the peduncle length (panicle exertion) trait in the region of 58684345~59149792bp on chromosome 7; in this study, the region of 58085816~59418341bp on chromosome 7 associated with the peduncle length trait was also obtained by SNP-index (or InDel-index) algorithm. We believe that the SNP-index (or InDel-index) algorithm combines the sequencing information of the parents in the calculation, which has the characteristics of effective removal of background noise and good credibility. The ED algorithm does not combine the information of the parents in the calculation and needs to obtain relative confidence region by raising the association threshold, which is the reason why the ED algorithm locates fewer regions in this experiment. Studies found that detection of some QTLs associated with the peduncle length trait were affected by environmental effects (Klein et al., 2001; Ding et al., 2023). Klein et al. (2001) found that QTL associated with peduncle length (panicle exertion) in LG D gene region was detected only under supplemental irrigation, while QTL associated with peduncle length (panicle exertion) in LG E gene region was detected in all test environments (sprinkler irrigated and non-irrigated). In this study, among the associated regions obtained from the F2 population constructed by brewing dwarf sorghum material, the most obvious signal locus was on chromosome 10 from 51040000 to 51520000 bp, with a total gene length of 0.48 Mb and a total of 29 genes. Gene function annotation revealed that gene-LOC8065033 controls the synthesis of the transcriptional elongation regulator MINIYO (IYO). IYO is thought to be associated with cell proliferation and differentiation in Arabidopsis research reports and is the switch that initiates Arabidopsis differentiation (Sanmartín et al., 2001; Muñoz et al., 2017); IYO, as an RNA-binding protein that positively regulates cell differentiation, is the first elongation factor identified in plants interacting protein that is thought to have an important role in transcriptional elongation (Lijsebettens et al., 2014). There is evidence that IYO/RPAP1 subcellular distribution is critical in determining the fate of plant and animal stem cells, and the IYO/RPAP1 interactome is conserved from plants to animals (Muñoz et al., 2017; Lynch, 2018; Contreras et al., 2019). However, the regulatory mechanism regarding IYO is not yet clear. We speculate that the gene-LOC8065033 may be a candidate gene associated with the sorghum peduncle length trait.

Conclusion

In this experiment, using BSA-seq method, the association region of brewing dwarf sorghum peduncle length trait was finally targeted on chromosome 7 and 10, which laid the foundation for the subsequent cloning and molecular study of peduncle length genes.

Acknowledgements. This research was supported by the Research on the Breeding of New Varieties of Extremely Early Maturing Sorghum and Green and Efficient Cultivation Techniques (GA21B009-10) and QTL Localization and Regulatory Analysis of Sorghum Peduncle Length Genes (2021YYFF010) , and Genetic Analysis and QTL Localization of Brewing Dwarf Sorghum Primary Branch Stalk Length (XDZDB2023-02).

REFERENCES

- [1] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. (1997): Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. – *Nucleic acids research* 25(17): 3389-3402.
- [2] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G. (2000): Gene ontology: tool for the unification of biology. – *Nat Gene* 25: 25-29.
- [3] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Susan, J. L., Lu, X., Ruden, D. (2012): A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff (SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3). – *Fly* 6(2): 80-92.
- [4] Contreras, R., Kallemi, P., González-García, M. P., Lazarova, A., Sánchez-Serrano, J. J., Sanmartín, M., Rojo, E. (2019): Identification of domains and factors involved in MINIYO nuclear import. – *Frontiers in Plant Science* 10: 1044.
- [5] Dabija, A., Ciocan, M. E., Chetrariu, A., Codină, G. G. (2021): Maize and Sorghum as Raw Materials for Brewing, a Review. – *Applied Sciences* 11(7): 3139.
- [6] Deng, Y., Li, J., Wu, S., Zhu, Y., Chen, Y., He, F. (2006): Integrated nr database in protein annotation system and its localization. – *Computer Engineering* 32: 71-72.
- [7] Ding, Y. Q., Wang, C., Xu, J. X., Gao, X., Cheng, B., Cao, N., Zhang, L. Y. (2023): QTL localization of sorghum panicle traits based on high-density genetic mapping. – *Journal of Plant Genetic Resources*, pp. 1-13.
- [8] Fekih, R., Takagi, H., Tamiru, M., Abe, A., Natsume, S., Yaegashi, H., Sharma, S., Sharma, S., Kanzaki, H., Matsumura, H., Saitoh, H., Mitsuoka, C., Utsushi, H., Uemura, A., Kanzaki, E., Kosugi, S., Yoshida, K., Cano, L., Kamoun, S., Terauchi, R. (2013): MutMap+: Genetic mapping and mutant identification without crossing in rice. – *PLoS One* 8: e68529.
- [9] Hill, J. T., Demarest, B. L., Bisgrove, B. W., Gorski, B., Su, Y. C., Yost, H. J. (2013): MMAPP: mutation mapping analysis pipeline for pooled RNA-seq. – *Genome Res* 23: 687-697.
- [10] Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., Mering, C., Bork, P. (2019): eggNOG 5.0: a hierarchical, functionally, and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. – *Nucleic acids research* 47(D1): D309-D314.
- [11] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M. (2004): The KEGG resource for deciphering the genome. – *Nucleic Acids Res* 32: D277-D280.
- [12] Kassahun, B., Bidinger, F. R., Bash, C. T., Kuruvashetti, M. S. (2010): Stay-green expression in early generation *Sorghum bicolor* (L.) Moench QTL introgression lines. – *Euphytica* 172(3): 351-362.
- [13] Klein, R. R., Rodriguez-Herrera, R., Schlueter, J. A. (2001): Identification of genomic regions that affect grain-mould incidence and other traits of agronomic importance in sorghum. – *Theoretical and Applied Genetics* 102(2-3): 307-319.
- [14] Li, H., Durbin, R. (2009): Fast and accurate short read alignment with Burrows-Wheeler transform. – *Bioinformatics* 25(14): 1754-1760.

- [15] Lijsebettens, M. V., Dürr, J., Woloszynska, M., Grasser, K. D. (2014): Elongator and SPT4/SPT5 complexes as proxy to study RNA polymerase II transcript elongation control of plant development. – *Proteomics* 14(19): 2109-2114.
- [16] Lin, Y. R., Schertz, K. F., Paterson, A. H. (1995): Comparative-analysis of QTLs affecting plant height and maturity across the poaceaein reference to an interspecific sorghum population. – *Genetics* 141(1): 391-411.
- [17] Lynch, C. J., Bernad, R., Calvo, I., Nóbrega-Pereira, S., Ruiz, S., Ibarz, N., Martinez-Val, A., Graña-Castro, O., Gómez-López, G., Andrés-León, E., Angarica, V. E., Sol, A., Ortega, S., Fernandez-Capetillo, O., Roj, E., Munoz, J., Serrano, M. (2018): The RNA polymerase II factor RPAP1 is critical for mediator-driven transcription and cell identity. – *Cell Reports* 22(2): 396-410.
- [18] Mace, E. S., Singh, V., Van Oosterom, E. J., Hammer, G. L., Hunt, C. H., Jordan, D. R. (2012): QTL for nodal root angle in sorghum (*Sorghum bicolor* L. Moench) co-locate with QTL for traits associated with drought adaptation. – *Theoretical and Applied Genetics* 124(1): 97-109.
- [19] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytzky, A., Garimella, A., Altshuler, D., Gabriel, S., Daly, M., DePristo, M. (2010): The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. – *Genome Research* 20(9): 1297-1303.
- [20] Muñoz, A., Mangano, S., González-García, M. P., Contreras, R., Sauer, M., Rybel, B. D., Weijers, D., Sánchez-Serrano, J. J., Sanmartín, M., Rojo, E. (2017): RIMA-dependent nuclear accumulation of IYO triggers auxin-irreversible cell differentiation in Arabidopsis. – *The Plant Cell* 29(3): 575-588.
- [21] Ramí, J. F., Dufour, P., Trouch, G., Fliedel, G., Mestres, C., Davrieux, F., Blanchard, P., Hamon, P. (1998): Quantitative trait loci for grain quality, productivity, morphological and agronomical traits in sorghum (*Sorghum bicolor* L. Moench). – *Theoretical and Applied Genetics* 97(4): 605-616.
- [22] Sanmartín, M., Sauer, M., Munoz, A., Zouhar, J., Ordóñez, A., Ven, W., Caro, E., Sánchez, M., Raikhel, N., Gutiérrez, C., Sánchez-Serrano, J., Rojo, E. (2001): A molecular switch for initiating cell differentiation in Arabidopsis. – *Current Biology* 21(12): 999-1008.
- [23] Su, S. (2012): QTL localization study of morphological agronomic traits in sorghum. – Nanjing University, Nanjing.
- [24] The UniProt Consortium (2021): UniProt: The universal protein knowledgebase in 2021. – *Nucleic acids research* 49(D1): D480-D489.
- [25] Zhai, G. W. (2010): QTL localization of sugar content of sweet sorghum stalk sap and its related traits. – Zhejiang Normal University, Zhejiang.