bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Identification of Genomic Insertion and Flanking Sequences of the Transgenic Drought-tolerant Maize Line “SbSNAC1-382” using the Single Molecular Real-Time (SMRT) Sequencing Method

Tingru Zeng, Dengfeng Zhang, Yongxiang Li, Chunhui Li, Xuyang Liu, Yunsu Shi, Yanchun Song, Yu Li, Tianyu Wang

Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China

Abstract

Safety assessment of genetically modified (GM) crops is crucial in the phase of product development before the GM crops are put on the market. Characteristics of flanking sequences of exogenous insertion sequences are essential for the safety assessment and marking of transgenic crops. In this study, we used the methods of genome walking and whole genome sequencing (WGS) to identify the flanking sequence characteristics of a *SbSNAC1* transgenic drought-tolerant maize line “SbSNAC1-382”, but both of the methods failed. Then, we constructed a genomic fosmid library of the transgenic maize line, which contained 4.18×10^5 clones with an average insertion fragment of 35 kb, covering 5.85 times of the maize genome. Subsequently, three positive clones were screened by pairs of specific primers and one of the three positive clones was sequenced by using the Single Molecule Real-Time (SMRT) sequencing technology. More than 1.95 Gb sequence data (~10^5 × coverage) for the sequenced clone was generated. The junction reads mapped to the boundaries of T-DNA and the flanking sequences in the transgenic line were

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

identified by comparing all sequencing reads with the maize reference genome and the sequence of transgenic vector. Furthermore, the putative insertion loci and flanking sequences were confirmed by PCR amplification and Sanger sequencing. The results indicated that two copies of the exogenous T-DNA fragments were inserted in the same genomic site. And the exogenous T-DNA fragments were integrated at the position of Chromosome 5: 177155650 to 177155696 in the transgenic line 382. Herein, we have demonstrated the successful application of the SMRT technology for the characterization of genomic insertion and flanking sequences.

Keywords: transgenic maize, flanking sequence, fosmid library, SMRT sequencing

Introduction

Since genetically modified (GM) crops were first introduced in the U. S. in the mid-1990s, they have become widely adopted by growers of many countries in the world [1]. In 2017 alone, 189.8 million hectares of GM crops were planted worldwide [2]. It is an international consensus that GM crops could be commercialized after they are proven to be safe. As a result, extensive testing and comprehensive analyses of transgenic lines with excellent objective traits are necessary for biosafety assessment before being approved and entering into market. Among these, molecular characterization of GM crops at the chromosomal level including insertion sequences, sites, copy numbers and flanking sequences is essential for the safety assessment and specific detection of GM crops [3, 4]. Furthermore, identification of T-DNA flanking

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sequences of GM crops and the development of specific detection methods are useful for breeding program, and important for bio-risk management to ensure food, feed and environmental safety [5, 6].

Traditionally, exogenous fragments flanking sequences of transgenic plants are obtained by various PCR-based methods according to the T-DNA sequence information [7-10]. Among them, thermal asymmetric interlaced PCR (TAIL-PCR) and genome walking are often used to isolate and clone T-DNA flanking sequences [9, 10]. Using the TAIL-PCR method by sequencing, several T-DNA flanking sequences were identified and characterized in transgenic maize [11], soybean [12], cotton [13], and alfalfa [14]. However, these PCR-based methods are laborious and expensive.

With the emergence and rapid development of next-generation sequencing (NGS) technology over the past few years, molecular characterization of insert locations, copy numbers, integrity, and stability of transgenic crops can be implemented in a relatively short time and at acceptable cost [15]. Up to now, a number of the flanking sequences of exogenous genes in GM plants such as *Arabidopsis* [16], rice [17], soybean [6], and maize [18] have been identified by the NGS method.

Taken together, both the PCR-based method and the NGS technology enabled us to successfully characterize both single and stacked transgenic events [15]. However, these approaches are difficult to identify all insertion loci and their flanking sequences of transgenic events with complex genome sequences or intricate modifications or

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

rearrangements of exogenous fragments [6, 19].

Maize is one of the most important crops in the world and 31% of the GM crops' growing area planted annually in the world are GM maize [1]. It is important to evaluate the safety of the GM maize, especially to identify the flanking sequence of exogenous genes in the GM maize. However, maize has larger genome and more repetitive sequences compared with soybean, cotton and rice, and it is difficult to identify flanking sequences of inserted genes [20]. In addition, transgenic maize events may often contain a part of or the entire vector backbone. In other cases, a partial copy of T-DNA inserts and the connection takes place outside the expected boundary [21, 22]. Therefore, for the acquisition of flanking sequences integrated into larger genomes and complex insertion fragments, the accurate flanking sequences can often be found by constructing DNA libraries. Turning genomes into countless fragments by physical or biological means cloned in fosmid or BAC vectors were a mainstay of genome projects during the Sanger-based sequencing era [23, 24]. Compared with other libraries, fosmid libraries have the advantages of short cloned fragments (about 40 kb), single copy insertion and easier to generate [25]. Because inserts in the fosmid libraries are generated randomly by ultra-sound rather than by enzymatic digestion, inserts in the fosmid libraries can avoid potential clone biases. It is suitable for physical mapping, gene cloning and chromosome mapping of gene fragments [6, 26]. Recently emerged single-molecule based NGS technology generate longer reads (20 kb) at increased coverage depth and is particularly important in resolving the challenges in characterization of transgenic events with insert locations

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in repetitive and low complexity regions of a genome [27]. As a result, using the fosmid libraries and the single-molecule based NGS technology might be suitable for identifying T-DNA flanking sequences of transgenic lines with complex genome sequences or intricate modifications or rearrangements of exogenous fragment.

Recently, we developed one transgenic line "SbSNAC1-382" by over-expression of *SbSNAC1* from sorghum, which conferred drought tolerance without a cost of crop productivity under well-watered conditions. Southern blots confirmed that the transgenic line SbSNAC1-382 was a two-copy insertion event, and the two copies might be inserted at the same genome location. In order to obtain the flanking sequence of the target gene of the transgenic maize event, after the failure of the genome walking method and the whole genome sequencing method, the single molecule real-time sequencing was used to identify the accurate flanking sequences of the inserted gene. Molecular characterization of the drought-tolerant transgenic maize at nucleic acid level will provide precise information for regulatory submissions and facilitate utilization of the line in future breeding program.

Materials and methods

Plant materials

SbSNAC1 with *Bst*EⅡ and *Nco*I enzymatic restriction sites were recombined into the pCAMBIA3301 vector under control of the cauliflower mosaic virus (CaMV) 35 S promoter, resulting in 35S::*SbSNAC1* constructs (S1 Fig). The constructed vector was transferred into maize hybrid HiII by *Agrobacterium*-mediated method. Positive transgenic events backcrossed with the inbred line "Zheng58" for six generations, and

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the resulting "SbSNAC1-382" was used in subsequent flanking sequence identification.

DNA isolation and Southern blot analysis

Genomic DNA for leaf samples of the transgenic event and the non-transgenic control was isolated using the CTAB method [28].

Thirty micrograms of the genomic DNA from the transgenic event and the non-transgenic control were digested with the restriction enzymes of *BstE*II and *Nco*I overnight. The resolved genomic DNA was then transferred to the positively charged nylon membranes (Hybond-N$^+$, Amersham Pharmacia Biotech) using a model 785 vacuum blotter system (Bio-Rad). The *Bar* amplified fragment (Table 1) labeled by DIG high primer DNA labeling (Roche, Cat. No. 11585614910) and purified using a high pure PCR product purification kit (Roche, No. 11732668001). The DNA blots were prehybridized at 42°C for 1 h in DIG easy hyb granule and then hybridized to denatured DIG-labeled probes for 20 h. The blots were then washed twice with 2×SSC and 0. 1% (w/v) SDS for 15 min each and washed twice with 1×SSC and 0. 1% (w/v) SDS for 15 min each. Immunological detection of the probes was carried out in accordance with the manufacturer's instructions for the DIG high primer DNA labeling and detection starter kit II.

Table 1. Primers used in this study.

Primer	Sequence (5'-3')
Bar F	GAAGTCCAGCTGCCAGAAAC
Bar R	GTCTGCACCATCGTCAACC
SbNACS3	GACCGCAAGTACCCAAACGG
SbNACA3	CACCCAGTCATCCAGCCTGAG

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SbNACS4	GGGACCGCAAGTACCCAAACG
SbNACA4	GCTGCGCTTCTCGCTCCTCT
NosF1	GAATCCTGTTGCCGGTCTTG
NosR1	TTATCCTAGTTTGCGCGCTA
35F1	GCTCCTACAAATGCCATCATTGC
35R1	GATAGTGGGATTGTGCGTCATCCC
zsp1	TATCCCTGGCTCGTCGCCGA
zsp2	AGGGCTTCAAGAGCGTGGTCGCT
zsp3	CCGTCACCGAGATTTGACTCGAGTTTC
YZP1	AGAATCATACACCAGTAACAAGCC
YZP2	CAGTACATTAAAAACGTCCGCA
YZP3	ACTAAAATCCAGATCCCCCGAA
YZP4	TTCACACAAGGAAACAGCTATGA
YZP5	CGATTAAGTTGGGTAACGCCA
YZP6	CTTCGCAAGACCCTTCCTCT
YZP7	TCCCTCTCCCTCCTCATCAC
YZP8	AGATTTTCTTCTTGTCATTGGG
YZP9	CTAGAGCAGCTTGAGCTTGG
V1	GGTTTCGCTCATGTGTTGAGC
G1	AGTGCACATTGCAATCCTACAAG
G2	CCTAAGTTCATGCAACTAGAGGTTTCA

Genome walking method

The 5' flanking sequence of the insertion sequence was obtained by the Genome Walking Kit according to the manufacturer instructions (TaKaRa, Dalian, China). The random primers were provided by the Genome Walking Kit and the specific primers designed based on theoretical insertion sequences (first round zsp1; second round zsp2; third round zsp3, Table 1). The specific PCR products were gel purified by using the DNA Gel Extraction Kit (Axygen, USA) and cloned to the pMD-18 vector system (Takara), and then sequenced by the Shanghai Sangon Company.

Whole genome sequencing

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A total of 1.5 μg DNA per sample was used as input material for the DNA sample preparations. Sequencing libraries were generated by using the Truseq Nano DNA HT Sample preparation Kit (Illumina USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, the DNA sample was fragmented by sonication to a size of 350 bp, then DNA fragments were end polished, A-tailed, and ligated with the full-length adapter for Illumina sequencing with further PCR amplification. At last, PCR products were purified (AMPure XP system) and libraries were analyzed for size distribution by Agilent2100 Bioanalyzer and quantified using real-time PCR. These libraries constructed above were sequenced by Illumina HiSeq4000 platform and 150 bp paired-end reads were generated with insert size around 350 bp.

Construction and screening of the fosmid library

DNA was interrupted by ultra-sound and separated by the method of Pulsed Field Gel Electrophoresis (PFGE). DNA fragments from 38-48 kb were recovered and end-repaired the sheared DNA to blunt and 5'-phosphorylated ends. The fosmid library was constructed with the Copy Control™ HTP Fosmid Library Production Kit (Epicenter, USA) using the pCC2FOS™ Vector and EPI300-T1^R plating cells.

Three pairs of vector-specific primers were designed to screen positive clones from the fosmid library (SbNACS3/SbNACA3; SbNACS4/SbNACA4; Bar F/Bar R, Table 1). In the initial screening of the library, three pairs of primers were used to detect the library, and a positive colony was obtained. Colony PCR reaction contained 10 μl 2×La Taq Mix (Takara), 1 μl colony, 0.5 μl forward and reverse primer, 8 μl

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ddH_2O. The procedure of PCR was as follows: 95℃ for 5 min; 95℃ for 30 sec; 60℃ for 30 sec; 72℃ for 30 sec; and a final extension at 72℃ for 5 min; 32 cycles. When a positive clone was identified, the positive colony diluted 2×10^6 times with LB liquid media was plated on LB solid medium, and monoclones were picked and subjected to colony PCR.

Single Molecule Real-Time Sequencing

10 μg of the monoclonal plasmid was extracted and purified. The PacBio libraries were constructed using plasmid that was mechanically sheared to a size of ~22 kb, using Covaris g-TUBE (Covaris, Inc. , Woburn, MA) according the manufacturer's instructions. PacBio SMRTbell libraries were prepared by ligation of hairpin adaptors at both ends of the DNA fragment using the PacBio DNA template preparation kit 2.0 for SMRT sequencing on the PacBio RS II machine (Pacific Biosciences of California, Inc., Menlo Park, CA). Bluepippin preparation system (SAGE science, Beverly, MA) was used to enrich more than 7 kb fragments in the library. Then, the quality of the library was tested by the Agilent Bioanalyzer 2100 kit (Agilent Technology, Inc. , Santa Clara, CA). Sequencing was performed on the PacBio RS II instrument as per the manufacturer's recommendations.

Results

Southern blot analysis of the transgenic line

In order to determine the transgene copy number, a Southern blot analysis were performed by using probes designed to hybridize the *Bar* gene in the T-DNA sequences. The results showed the transgenic line had two copies of insertion of the

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

exogenous sequences when *Hind*III and *Eco*RI were used to digest the DNA of the transgenic line (Fig 1). On the other hand, there was only one band when the DNA of the transgenic line was digested with the restriction endonucleases of *Bgl*II and *Dra*I for which there are no restriction sites in the insertion sequences (Fig 1). As a result, it might be two copies of insertion sequences at the same genomic location of the transgenic maize line.

Figure 1. Southern blot analysis of the SbSNAC1-382 line.1 to 4 digested DNA of the transgenic line by *Hind*Ⅲ, *Eco*RI, *Bgl*Ⅱ and *Dra*I, respectively; 5, digested plasmids as positive controls, M, marker.

Genome walking for detecting flanking sequences

Three nested specific primers (zsp1, zsp2, zsp3) were designed according to the sequences adjacent to the T-DNA left border. According to the instructions of the Genome Walking Kit (Takara-Bio, Dalian, China), with nested specific primers and four degenerate primers, three rounds of nested PCR reaction were completed and specific band was obtained (lane 10 of Fig 2). The sequencing results demonstrated that the specific PCR fragment contained 1227 bp in length. By aligning with the maize genome sequence on Maize GDB (www. maizegdb. org) and the T-DNA sequence, it showed that the fragment was made up of 932 bp of non-insert DNA and 295 bp of insert DNA. As expected, the 295 bp inserted DNA was identical to the sequence which was adjacent to the T-DNA left border. The 932 bp of non-insert DNA was identical to the maize genome sequence which is located between 177155650 - 177156582 bp on Chromosome 5. However, the flanking sequence

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

adjacent to the T-DNA right border could not be identified with multiple nested specific primers and degenerate primers using the same method.

Figure 2. Genome walking results for 5' flanking sequence. 1-4 lanes are the first amplification results of specific primer zsp1 and degenerate primer AP1-AP4, respectively; 5-8 lanes are the second amplification results of specific primer zsp2 and degenerate primer AP1-AP4, respectively; 9-12 lanes are the third amplification results of specific primer zsp3 and degenerate primer AP1-AP4, respectively; M: marker.

WGS for detecting flanking sequences

We attempted to use WGS technique to identify flanking sequences on both sides of the insertion sequence. Sequencing libraries were sequenced by Illumina HiSeq4000 platform and 150 bp paired-end reads were generated with insert size around 350 bp. After quality control processing, a total of 144.6 billion clean reads for the transgenic line were obtained (Table 2). Among them, 97. 66% of the reads could be mapped to the reference genome, accounting for ~64.57 × coverage of the maize genome. Furthermore, about 93.66% of the genome had at least one-fold coverage and 88.57% had at least four-fold coverage. Therefore, the above data indicate that the quality of sequencing was qualified and met the requirements of analysis.

In order to identify flanking sequences of putative insertion sites of exogenous fragments, all clean reads were mapped to the sequence of *pCAMBIA3301-SbSNAC1* vector and the maize reference genome. The putative flanking sequence of SbSNAC1-382 line was characterized based on junction reads in which one end of which maps to the vector sequence and the other end to the maize genome. After

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

detailed analysis, five putative flanking sequences were found. One of the five possible flanking sequences was consistent with the Genome Walking's results. The total length of the fragment was 263 bp. The 150 bp DNA sequence was identical to the sequence adjacent to the T-DNA left border, and the 113 bp DNA sequence was identical to the maize genome. Unfortunately, that the other four putative flanking sequences were not true according to the PCR results. As a result, the flanking sequence adjacent to the T-DNA right border of the SbSNAC1-382 line was still not identified by using the WGS technology.

Table 2. The summary of sequence data of WGS.

Index	Value
Clean reads (bp)	144,595,902,900
Q20 (%)	> 90
Q30 (%)	> 85
Mapped reads (bp)	932, 861, 603
Total reads (bp)	963, 972, 686
Mapping rate (%)	96. 77
Average depth (X)	64.57
Coverage at least 1X (%)	93.66
Coverage at least 4X (%)	88.57

Fosmid library construction and positive clone screening

To identify the flanking sequence adjacent to the T-DNA right border of the SbSNAC1-382 line, we constructed a fosmid library of the SbSNAC1-382 line (Takara, Dalian, China), with the recombination rate of 100% (S2 Fig). The original library was diluted and the number of colonies was counted. The library contained about 4.18 × 10^5 clones, and the average length of the inserted fragments was about 35 kb, which could achieve 5.85 times of the maize genome coverage. According to

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the Clarke-Carbon formula [29], the probability of screening any gene or sequence from the constructed library was 99. 71%.

In order to screen the target clone from the fosmid library, five pairs of primers were designed (SbNACS3/SbNACA3; SbNACS4/SbNACA4; Bar F/Bar R, NosF1/NosR1, 35F1/35R1, Table 1) according to the T-DNA sequences. According to the results of PCR methods, three positive clones were identified and stored for single-molecule real-time sequencing.

Single-molecule real-time sequencing and Sanger sequencing

One of the three positive clones screening from fosmid library was selected for sequencing. After the processing of quality control, yielding a total of 1.95 Gb in 100,544 clean reads with a mean length of 9.5 Kb, an N50 length of 12.5 Kb (Table 3).

To determine the hypothetical insertion sites of exogenous fragments, we constructed a local BLAST data library of single-molecule real-time sequencing data. According to the BLAST results of T-DNA sequences with the local data library, it confirmed the results of Southern blot that the exogenous sequences were composed of two copies of the *SbSNAC1* gene and the *bar* gene at the same maize genomic location. And the flanking sequences of both right border and left border were identified. In order to confirm the flanking sequences of the SbSNAC1-382 transgenic line, specific PCR primers were designed based on the putative genomic sequences and the insertion sequences. When using primer pairs with one primer annealing within putative flanking sequences (YZP2, YZP3, G1, G2, Table 1) and the other

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

annealing to the insertion sequence (YZP1, V1, Table 1), the gel electrophoresis revealed that PCR reactions of primer pairs (YZP1/YZP2; YZP1/YZP3; V1/G1; V1/G2, Table 1) had generated products with single band in the transgenic event 382 while no correct product could be detected from the non-transgenic control of Zheng58 or negative control of water (Fig 3). In addition, YZP3/G2 (Table 1) primers were used to amplify the whole length of the inserted sequence and sanger sequencing of the PCR products showed that the sequence was basically the same as that of the single-molecule sequencing, except for a few bases. Therefore, the exogenous sequence of the SbSNAC1-382 line was integrated at the physical position of Chr. 5: 177,155,650 to 177,155,696 with a 46 bp deletion (Fig 4). Furthermore, the exogenous fragment was inserted into the intergenic region of the maize genome, and no functional genes were interrupted by the inserted sequence.

In order to verify the results of single-molecule sequencing, we designed a series of primers on the insertion sequence (YZP2/YZP5; YZP4/YZP7; YZP6/YZP9; YZP8/G1, Table 1). The PCR products obtained by these primers were sequenced and compared with the results of single-molecule sequencing. It was found that the two sequencing results were basically the same, showing that single-molecule sequencing had a high accuracy. Further analysis of the structure of the insertion sequence revealed that the exogenous sequence contained two insertion sequences with tandem repetition and opposite direction. Because the restriction endonucleases *Hind*Ⅲ and *Eco*RI are between *bar* and *SbSNAC1*, there would be two bands after digestion with these two endonucleases. Meanwhile, for the genome digested with *Dra* Ⅰ and *Bgl* Ⅱ

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with no sites in the insertion sequence, there was only a single band in the Southern blot results. The special structure of the insertion sequence explained the results of Southern blot. Because of the special structure of the insertion sequence, neither the Sanger sequencing method nor the second generation sequencing method could obtain the cloned sequence.

Table 3. Statistics of single-molecule real-time sequencing for plasmid DNA

Index	Value
Polymerase read bases (bp)	1,949,620,057
Number of polymerase reads	100,544
Post-filter mean read length (bp)	19,390
Polymerase read N50 (bp)	28,065
Polymerase read quality	0.83
Mean subread length (bp)	9,509
Subreads N50 (bp)	12,507
Number of subreads	204,500

Fig 3. PCR validation of transgenic insertion sites. (A) PCR verification of 5' end of inserted sequence. 1, 2, 3 and 4, 5, 6 primer YZP1/YZP2 and YZP1/YZP3 amplified in the transgenic line, negative control Zheng58, negative control of water, respectively. M: marker. (B) PCR verification of 3' end of inserted sequence. 1, 2, 3 and 4, 5, 6 primer V1/G1 and V1/G2 amplified in the transgenic line, negative control Zheng58, negative control of water, respectively.

Fig 4. Schematic diagram of insertion loci and flanking sequences of SbSNAC1-382. The numbers under the line of Chr. 5 indicates physical positions on the chromosome. The arrows indicate the position of the validation primer.

Discussion

Detailed molecular characteristics of flanking sequences of insertions play an important role in safety assessment of genetically modified crops [30]. Traditionally, the PCR-based methods such as Tail-PCR and genome walking were used to determine locations of integration sites and junction sequences between exogenous sequences and host genome [9, 31]. With the continuous improvement of technology, the flanking sequence of single T-DNA copy insertion transgenic lines can be

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

obtained quickly and cheaply by these PCR-based methods. Charles et al. amplified the 5' flanking sequence of insertion sequence of 75 *Mu* maize mutant lines based on the PCR method, but the flanking sequences of 20% of the lines could not be obtained by this method in their study [32]. These PCR-based methods may not work well if the deletion, modification or rearrangement occurred in exogenous insertion sequences. On the other hand, high level of duplication or repetitive genome sequences adjacent to the exogenous fragment insertion location might increase the difficulty of identifying the flanking sequences. The maize genome size is about 2.3-2.5 G and nearly 85% of the maize genome is composed of hundreds of families of transposable elements, dispersed unevenly across the whole genome [33, 34]. In our research, the genome walking method was also used to amplify the flanking sequence of the insertion sequences, but only one end of the flanking sequence was identified due to the complex structure of the insertion sequences. As a result, using the PCR-based methods to identify the flanking sequences of complex exogenous fragments of transgenic lines might be a challenge in the maize genome.

With the emergence and development of high throughput next generation sequencing (NGS) technology, the cost of whole genome sequencing has been greatly reduced (Table 4). The NGS technology has been widely used in different species to discover genome structural variation, rearrangement, and so on [35-37], with some advantages including high throughput, no need for large amounts of DNA, time and labor saving [38]. Compared with other methods, the WGS combined with targeted bioinformatics analysis has become a sensitive and efficient method for identifying

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

molecular characteristics of GM crops. Guo et al. used the WGS technology to sequence and analyze the sequence information of two GM soybean events, and successfully identified from one single read analysis [6]. In the work of Kiran et al., by using the NGS method together with the PCR amplification to identify the T-DNA insertion site and flanking sequence of the GM maize IE09S034 at the 3’ end [18]. Although several NGS-based methods have been developed to identify the molecular characteristics of genetically modified crops, some examples often fail to identify insertion sites and flanking sequenced in GM crops. Park et al. used the WGS technology to identify the flanking sequences of three GM rice materials, but one failed to identify the flanking sequences of GM rice [39]. The authors of this article points out that if they can get a longer reads, this problem may not arise. The same problem has arisen in the course of our research. We used the WGS method to sequence the transgenic maize lines. After detailed analysis, only the one end flanking sequence of the insertion fragments was identified. Generally, the NGS technology using to identify the flanking sequences might be efficient if the transgenic line has one or two copies of insertion or stacked transgenic events. On the other hand, the clean reads of the WGS technology are usually only about 150 bp, and assembling the flanking sequences requires a large number of reads in the insertion region to be spliced together, which is a huge challenge for the genome with a large number of repetitive sequences. In our study, ~64.57 × coverage of the maize genome were sequenced, and only one end flanking sequence was identified. The insertion sequence consisted of two copies of T-DNA sequences, and it had no further clear

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sequence information, which increased the difficulty of identifying the flanking sequence using the WGS technology. Increasing the sequence coverage by deep sequencing might be helpful to identify the other end flanking sequence. But it is still difficulty to characterize the structure of the exogenous sequences using the WGS technology.

The fosmid technology has been applied in genomics of many species, such as rice [40], maize [41], and human [42]. Compared with the BAC library, the construction of fosmid library is simpler and faster. Furthermore, average length of the insertion sequences of the fosmid library is 38-48 kb which might be suitable to identify the flanking sequences and characterize the structure of the insertion sequence of transgenic lines. On the other hand, read length of single-molecule real-time DNA sequencing might be 10-20 kb, which may also contribute to characterize the insertion sequence of transgenic events. In our study, three positive clones were accurately identified from the fosmid library using the PCR method with three pairs of specific primers. Furthermore, with the SMRT sequencing technology, the flanking sequences were identified and the structure of exogenous insertion sequences was characterized. Although the use of the method of building fosmid libraries and the third generation sequencing to obtain flanking sequences of GM crops is more time-consuming and costly than the method based on PCR and WGS, it is more reliable for some GM crops with complex genomic or insertion sequence structure. In identifying the flanking sequences of GM crops, the method of constructing fosmid libraries combined with the third-generation sequencing

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

technology is not a high-throughput method, it is more time-consuming and costly, but it is more reliable to effectively identify the flanking sequences and characterize the insertion sequences with deletion, modification or rearrangement. As a result, when identifying the flanking sequences of genetically modified crops, different methods should be flexibly selected according to their genomic characteristics and the internal structure of insertion sequences (Table 4).

Table 4. Characteristic comparison of three methods for obtaining flanking sequences.

Method	Time	Cost	Insertion sequence structure	Genome complexity	Flux level
PCR-based	Short	Cheap	Simple	Simple	Low
NGS	Short	Cheap	Simple	Simple	High
Fosmid+sequencing	Long	Expensive	Complex	Complex	Low

Supporting Information

S1 Figure. Vector for transgenic line.

S2 Figure. Electrophoretogram of fosmid clones digested with *Not Ⅰ*. 1-16: Insert fragments; M: marker.

Acknowledgments

This work was carried out with the support of Innovation Program of Chinese Academy of Agricultural Sciences and the Major Projects of Genetically Modified Organisms, China (2016ZX08003004).

References

1. Graham B and Peter B. Environmental impacts of genetically modified (GM) crop use 1996–2015: Impacts on pesticide use and carbon emissions. GM Crops Food. 2017; 8:2, 117-147. doi: org/10.1080/21645698.2017.1309490 PMID: 28414252

2. ISAAA. Global Status of Commercialized Biotech/GM Crops in 2017: Biotech Crop Adoption Surges as Economic Benefits Accumulate in 22 Years. 2017; Vol. 53.

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3. Codex Alimentarius Commission. Guideline for the conduct of food safety assessment of foods derived from recombinant-DNAplants. CAC/GL. 2003; 45, 1–18.

4. European Food Safety Authority. Guidance on the environmental risk assessment of genetically modified plants. European Food Safety Authority. 2010; 8:1879.

5. Fraiture M, Herman P, Lefevre LT, Loose MD, Deforce D, Roosens NH. Integrated DNA walking system to characterize abroad spectrum of GMOs in food/feed matrices. BMC Biotechnol. 2015; 15:76, 1-11. doi: org/10.1186/s12896-015-0191-3 PMID: 26272331

6. Guo B, Guo Y, Hong H, Qiu LJ. Identification of genomic insertion and flanking sequence of G2-EPSPS and GAT transgenes in soybean using whole genome sequencing method. Front Plant Sci. 2016; Jul 12;7:1009. doi: org/10.3389/fpls.2016.01009 PMID: 27462336

7. Triglia T, Peterson MG, Kemp DJ. A procedure for *in vitro* amplification of DNA segments that lie outside the boundaries of known sequences. Nucleic Acids Res. 1988; Aug 25; 16(16):8186. doi: org/10.1093/nar/16.16.8186 PMID: 3047679

8. Jones DH, Winistorfer S C. Sequence specific generation of a DNA panhandle permits PCR amplification of unknown flanking DNA. Nucleic Acids Res. 1992; Feb11; 20(3): 595-600. doi: org/10.1093/nar/20.3.595 PMID: 1371352

9. Liu YG, Mitsukawa N, Oosumi T, Whittier RF. Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. Plant J. 1995; 8: 457-463.

10. Ji J, Braam J. Restriction site extension PCR: A novel method for high-throughput characterization of tagged DNA fragments and genome walking. PLoS ONE. 2010;May 11; 5(5): e10577. doi: org/10.1371/journal.pone.0010577 PMID: 20485508

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

11. Yang L, Xu S, Pan A, Yin C, Zhang K, Wang Z, et al. Event specific qualitative and quantitative polymerase chain reaction detection of genetically modified MON863 maize based on the 5'-transgene integration sequence. J Agric Food Chem. 2005 Nov 30; 53(24): 9312-8. doi: org/10.1021/jf051782o PMID: 16302741

12. Wang XB, Jiang LX, Wei L, Liu L, Lu W, Li WX. Integration and insertion site of EPSPs gene on the soybean genome in genetically modified glyphosate-resistant soybean. Acta Agron Sin. 2010; 36, 365–375. doi: org/10.3724/SP.J.1006.2010.00365

13. Akritidis P, Pasentsis K, Tsaftaris AS, Mylona PV, Polidoros AN. Identification of unknown genetically modified material admixed in conventional cotton seed and development of an event-specific detection method. Electron J Biotechn. 2008; 11, 7683. doi: org/10.2225/vol11-issue2-fulltext-11

14. Sun L, Gill US, Nandety RS, Kwon S, Mehta P, Dickstein R, et al. Genome-wide analysis of flanking sequences reveals that *Tnt1* insertion is positively correlated with gene methylation in *Medicago truncatula*. Plant J. 2019; Jun; 98(6): 1106-1119. doi: org/10.1111/tpj.14291 PMID: 30776165

15. Guttikonda SK, Marri P, Mammadov J, Ye L, Soe K, Richey K, et al. Molecular characterization of transgenic events using next generation sequencing approach. PLoS ONE. 2016; Feb 23; 11(2): e0149515. doi: org/10.1371/journal.pone.0149515 PMID: 26908260

16. Inagaki S, Henry IM, Lieberman MC, Comai L. High-throughput analysis of T-DNA location and structure using sequence capture. PLoS ONE. 2015; Oct 7; 10(10): e0139672 doi: org/10.1371/journal.pone.0139672 PMID: 26445462

17. Park D, Kim DG, Jang G, Lim JS, Shin YJ, Kin J, et al. Efficiency to discovery transgenic

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

loci in GM rice using next generation sequencing whole genome re-sequencing. Genomics Inform. 2015; Sep; 13(3): 81-5. doi: org/0.5808/GI.2015.13.3.81 PMID: 26523132

18. Siddique K, Wei J, Li R, Zhang D, Shi J. Identification of T-DNA insertion site and flanking sequence of a genetically modified maize event IE09S034 using Next-generation sequencing technology. Mol Biotechnol. 2019; Sep; 61(9): 694-702. doi: org/10.1007/s12033-019-00196-0 PMID: 31256331

19. Daniela W, Leif S, Joachim B, Lutz G. Next-generation sequencing as a tool for detailed molecular characterization of genomic insertions and flanking regions in genetically modified plants: a pilot study using a rice event unauthorized in the EU. Food Anal Method. 2013; 6, 1718. doi: org/10.1007/s12161-013-9673-x

20. Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, et al. Physical and genetic structure of the maize genome reflects its complex evolutionary history. PLoS Genet. 2007; Jul; 3(7): e123. doi: org/10.1371/journal.pgen.0030123 PMID: 17658954

21. Oltmanns H, Frame B, Lee LY, Johnson S, Li B, Wang K, et al. Generation of backbone-free, low transgene copy plants by launching T-DNA from the *Agrobacterium* chromosome. Plant Physiol. 2010; Mar; 152(3): 1158-66. doi: org/10.1104/pp.109.148585 PMID: 20023148

22. Sylvie DB, Chris DW, Marc VM, Ann D. T-DNA vector backbone sequences are frequently integrated into the genome of transgenic plants obtained by *Agrobacterium*-mediated transformation. Mol Breeding. 2000; 6: 459–468. doi: org/10.1023/A:102657552

23. Kim UJ, Shizuya H, de Jong PJ, Birren B, Simon MI. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. Nucleic Acids Res. 1992; Mar 11; 20(5): 1083-5.

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

doi: org/10.1093/nar/20.5.1083 PMID: 1549470

24. Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, et al. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl AcadSci U S A. 1992; Sep 15; 89(18): 8794-7. doi: org/10.1073/pnas.89.18.8794 PMID: 1528894

25. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008; May 1; 453(7191): 56-64. doi: org/10.1038/nature06862 PMID: 18451855

26. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structuralvariation of the human genome. Nat Genet. 2005; Jul;37(7):727-32. doi: org/10.1038/ng1562 PMID: 15895083

27. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009; Jan 2; 323(5910): 133-8. doi: org/10.1126/science.1162986 PMID: 19023044

28. Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW. Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proc Natl AcadSci U S A. 1984; Dec; 81(24): 8014-8. doi: org/10.1073/pnas.81.24.8014 PMID: 6096873

29. Clarke L, Carbon J. A colony bank containing synthetic CoI EI hybrid plasmids representative of the entire *E. coli* genome. Cell. 1976; 9(1): 91-99.

30. Yang L, Wang C, Holst-Jensen A, Morisset D, Lin Y, Zhang D. Characterization of GM events by insert knowledge adapted re-sequencing approaches. Sci Rep. 2013; Oct 3; 3: 2839.

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

doi: org/10.1038/srep02839 PMID: 24088728

31. Nakayama T, Soma M, Rahmutula D, Ozawa Y, Kanmatsuse K. Isolation of the 5'-flanking region of genes by thermal asymmetric interlaced polymerase chain reaction. Med SciMonit. 2001; 7(3): 345~349. PMID: 11386007

32. Dietrich CR, Cui F, Packila ML, Li J, Ashlock DA, Nikolau BJ, et al. Maize *Mu* transposons are targeted to the 5’-untranslated region of the *gl8* gene and sequences flanking *Mu* target-site duplications exhibit nonrandom nucleotide composition throughout the genome. Genetics. 2002; Feb; 160(2): 697-716. PMID: 11861572

33. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: Complexity, diversity, and dynamics. Science. 2009; Nov 20; 326(5956): 1112-5. doi: org/10.1126/science.1178534 PMID: 19965430

34. Hirsch CN, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, et al. Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell. 2016; Nov; 28(11): 2700-2714. doi: org/10.1105/tpc.16.00353 PMID: 27803309

35. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008; Jun; 40(6): 722-9. doi: org/10.1038/ng.128 PMID: 18438408

36. Hormozdiari F, Hajirasouliha I, McPherson A, Eichler EE, Sahinalp SC. Simultaneous structural variation discovery among multiple paired-end sequenced genomes. Genome Res. 2011; Dec; 21(12): 2203-12. doi: org/10.1101/gr.120501.111 PMID: 22048523

bioRxiv preprint doi: https://doi.org/10.1101/858951; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

37. Dubose AJ, Lichtenstein ST, Narisu N, Bonnycastle LL, Swift AJ, Chines PS, et al. Use of microarray hybrid capture and next-generation sequencing to identify the anatomy of a transgene. Nucleic Acids Res. 2013; Apr 1; 41(6): e70. doi: org/10.1093/nar/gks1463 PMID: 23314155

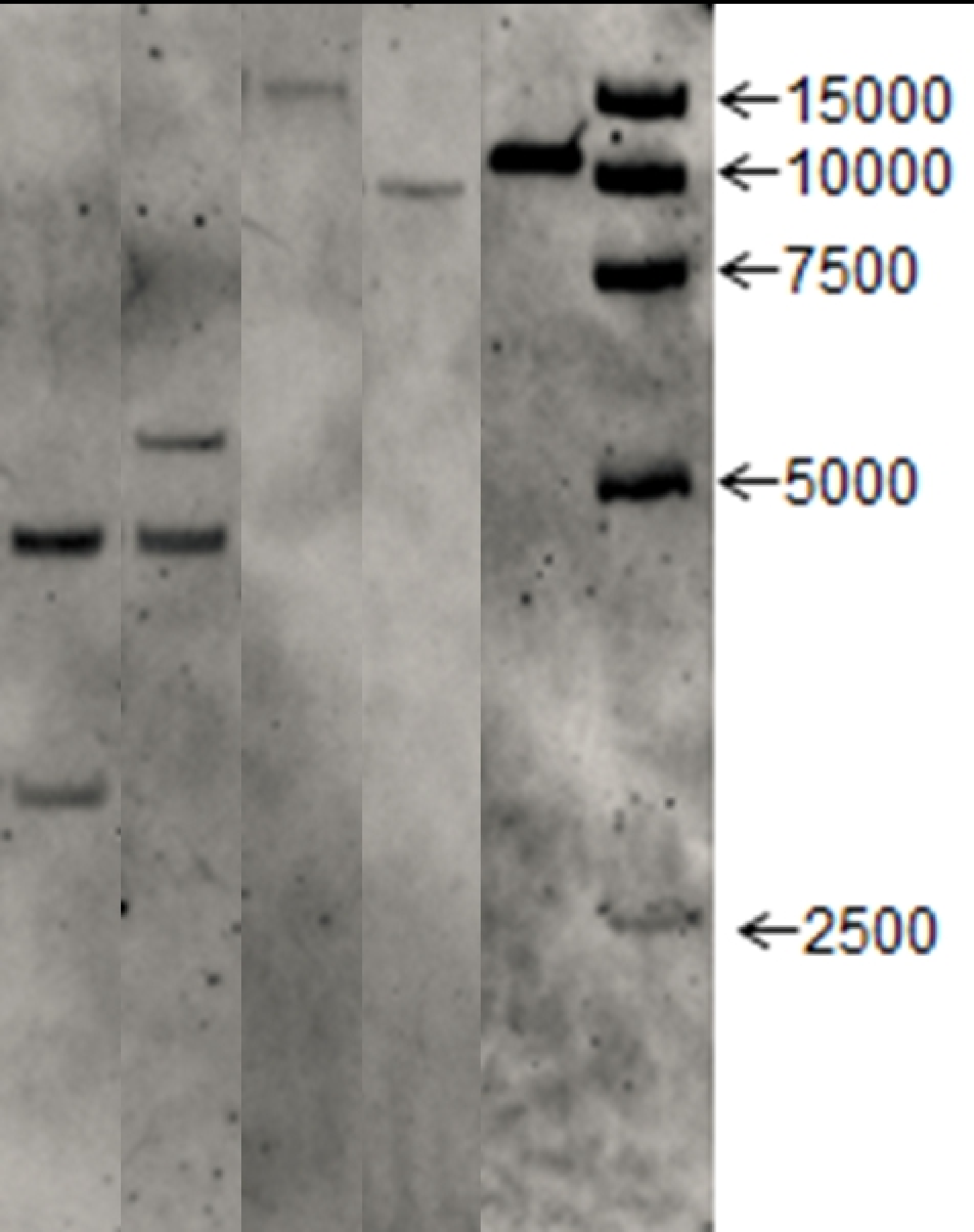
38. Pauwels K, De Keersmaecker SC, De Schrijver A, Jardin PD, Roosens NH, Herman P . Next-generation sequencing as a tool for the molecular characterization and risk assessment of genetically modified plants: add value or not. Trends Food Sci Tech. 2015; 45, 319–326. doi: org/10.1016/j.tifs.2015.07.009

39. Park D, Park SH, Ban YW, Kim YS, Park KC, Kim NS, et al. A bioinformatics approach for identifying transgene insertion sites using whole genome sequencing data. BMC Biotechnol. 2017; Aug 15; 17(1): 67. doi: org/10.1186/s12896-017-0386-x PMID: 28810845

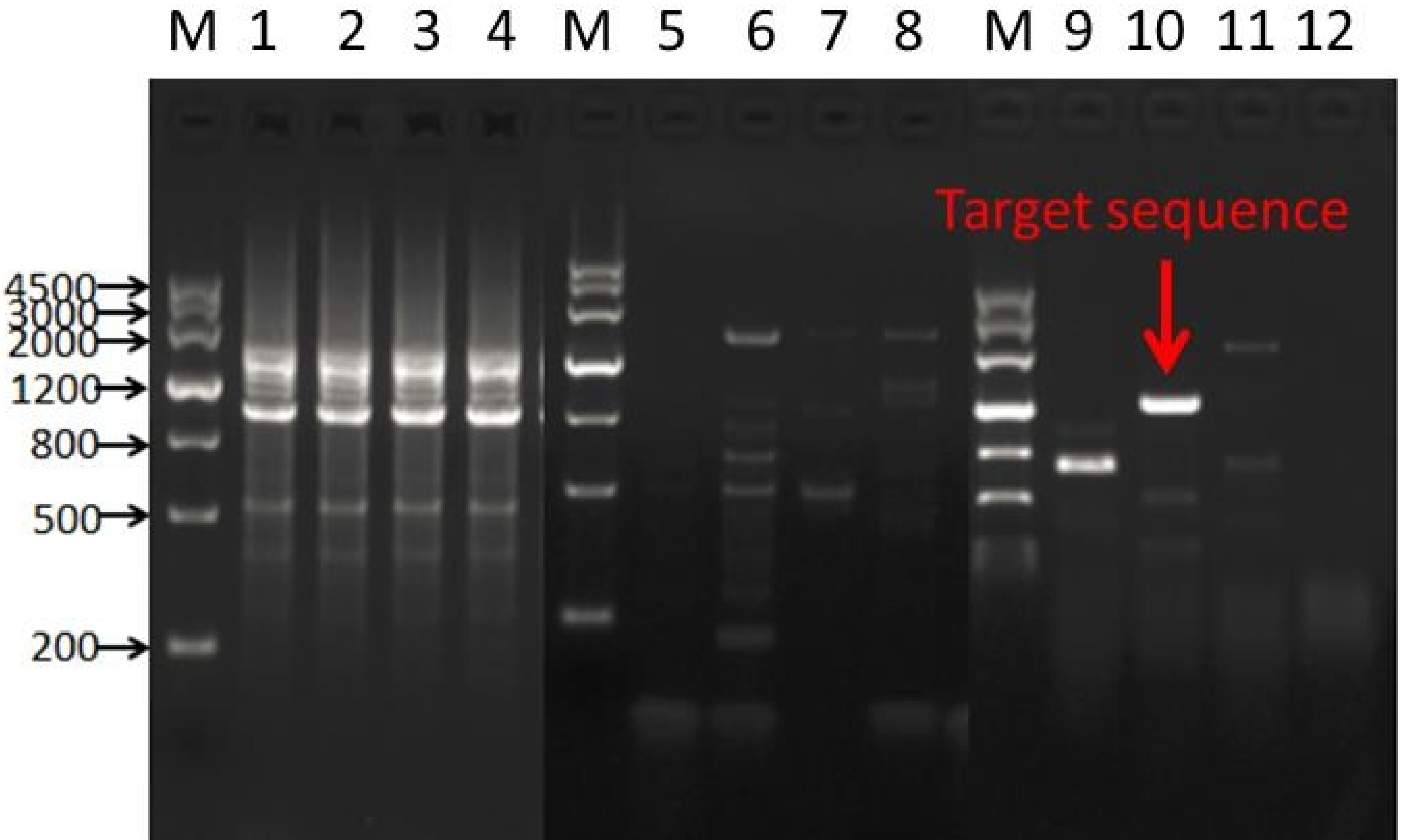
40. Ammiraju JS, Yu Y, Luo M, Kudrna D, Kim H, Goicoechea JL, et al. Random sheared fosmid library as a new genomic tool to accelerate complete finishing of rice (*Oryza sativa* spp. Nipponbare) genome sequence: sequencing of gap-specific fosmid clones uncovers new euchromatic portions of the genome. Theor Appl Genet. 2005; Nov; 111(8): 1596-607. doi: org/10.1007/s00122-005-0091-3 PMID: 16200416

41. Liu CX, Liu XL, Lei L, Guan HY, Cai YL. Fosmid library construction and screening for the maize mutant gene *Vestigial glume 1*. The Crop Journal. 2016; 4(1): 55-60. doi: org/CNKI:SUN:CROP.0.2016-01-007

42. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004; Oct 21; 431(7011): 931-45. doi: org/10.1038/nature03001 PMID: 15496913



15000
10000
7500
5000
2500


M
1
2
3
4
M
5
6
7
8
M
9
10
11
12
4500
3000
2000
1200
800
500
200
Target sequence