JARQ 57 (3), 205-215 (2023) https://www.jircas.go.jp

Whole-genome Sequence Analysis to Confirm the Absence of Transgene in a Rice Line Made by Gene Targeting

Junichi MANO[1]*, Keita TSUKAHARA[1,2], Kazuto TAKASAKI[2], Satoshi FUTO[2], Ayako NISHIZAWA-YOKOI[3], Seiichi TOKI[3,4,5,6], Reona TAKABATAKE[1] and Kazumi KITTA[1]

[1] Institute of Food Research, National Agriculture and Food Research Organization, Tsukuba, Japan
[2] Fasmac Co. Ltd., Atsugi, Japan
[3] Institute of Agrobiological Sciences, National Agriculture and Food Research Organization, Tsukuba, Japan
[4] Graduate School of Nanobioscience, Yokohama City University, Yokohama, Japan
[5] Kihara Institute for Biological Research, Yokohama City University, Yokohama, Japan
[6] Department of Plant Life Science, Faculty of Agriculture, Ryukoku University, Otsu, Japan

Abstract

The recent development of homologous recombination-mediated gene targeting (GT) techniques has made it easy to modify nucleotide sequences in plant genomes. Backcross breeding following GT provides transgene-free plant lines with high probability. However, owing to the possibility of unintentional transgene introduction during genetic transformation, analytical methods may be necessary to reliably and comprehensively detect transgenes remaining on a plant genome. Herein, we conducted an analysis based on whole-genome sequencing. We analyzed the GT rice DNA using a next-generation sequencer, and we performed data analysis to find the transgene integration on the genome. All results supported the absence of transgenes in GT rice. Our genome integrity evaluation method could also be applicable to plant lines produced using other genome-editing technologies such as CRISPR/Cas9.

Discipline: Food
Additional key words: genome editing, rice, whole-genome sequence analysis

Introduction

Gene targeting (GT) is a genetic technique that uses homologous recombination (HR) to alter a specific DNA sequence in an endogenous gene at its original locus in the genome (Sanagala et al. 2017). For example, GT with positive–negative selection and the subsequent excision of the positive selectable marker gene using piggyBac transposon afforded precise and efficient genome modifications in rice (Nishizawa-Yokoi et al. 2015). As the final step of this method, the piggyBac transposase gene can be segregated out based on Mendel's laws of heritage. Subsequent confirmation by polymerase chain reaction (PCR) and/or southern blotting provides rice lines without the transgenes (Nishizawa-Yokoi et al. 2015).

Further, other new research tools for plant genome editing also include targeted mutagenesis using sequence-specific nucleases, such as transcription activator-like effector nuclease (TALEN) and clustered regularly interspaced short palindromic repeats/Crispr-associated protein 9 (CRISPR/Cas9). For the targeted mutagenesis of plant genomes, it is typically necessary to integrate these sequence-specific nuclease genes into plant genomes (Wolt et al. 2016). As in the case for GT, the sequence-specific nuclease genes can be removed using the conventional breeding technique.

Over the past two decades, many traditional genetically modified (GM) crops that hold transgenes on their genomes have been commercialized. Along with international guidelines, regulatory authorities in large parts of the world have imposed mandatory safety

*Corresponding author: jmano@affrc.go.jp
Received 30 August 2022; accepted 6 December 2022.

assessments as consequence developers are required to obtain and provide exhaustive data about their GM events. Lusser et al. (2012) reported that it takes ~5.5 years to prepare a regulatory dossier for a GM event, at the cost of $35 million. This tremendous cost is a major issue for the commercialization of GM crops. In contrast, crops produced via HR-mediated GT or targeted mutagenesis do not have transgenes (new improved crops), and the modifications are limited to the target regions of the genome. Unintentional impacts caused by the modification to the overall metabolisms and cell functions are likely to be smaller in these new improved crops compared with traditional GM crops. There is a possibility that they are commercialized with a limited degree of analysis for the required safety assessment.

Many improved crop lines have been developed via GT or targeted mutagenesis and some of them await commercial use. In some countries, regulatory authorities authorized a part of genome-edited plants (Sprink et al. 2016). However, active debates on how to deal with genome-edited crops are ongoing. Despite intense debates about the appropriate regulation of GT and targeted mutagenesis in many parts of the world, a consensus among the many stakeholders has not been reached. Unintended side effects of gene introduction are an important topic in discussions on the regulation of these new breeding techniques.

Several research groups have been examining the off-target integration of exogenous DNA along with genetic transformation. *Agrobacterium*-mediated transformation has sometimes resulted in the introduction of a vector sequence (Latham et al. 2006, Schouten et al. 2017). In addition, the particle bombardment method enables the deletion and extensive scrambling of transgene and chromosomal DNA (Latham et al. 2006). The commonly used PCR and/or southern blotting techniques cannot detect short DNA fragments of a transgene if any of the fragments are randomly introduced into the host plant genome. These previous reports highlighted the need for new analytical methods to detect short transgenes that may remain present in the improved plants. If the nonexistence of such transgene integration can be proved, this particular barrier to the commercialization of genome-edited plants could be surmounted.

Many research groups have described the characterization of recombinant DNA in GM crops with the use of next-generation sequencers (NGSs) (Arulandhu et al. 2016, Guttikonda et al. 2016, Kovalic et al. 2012, Yang et al. 2013, Zastrow-Hayes et al. 2015). These reports showed that NGSs worked well for the analyses of transgenes in plant genomes. Advances in sequencing technologies will improve the analytical performance of NGSs and reduce the analytical costs. Therefore, we selected a whole-genome sequencing-based approach for the evaluation of genome-edited crops in the present study, and we used an herbicide-tolerant rice line developed by HR-mediated GT (Nishizawa-Yokoi et al. 2015) as a model new breeding technique-improved plant. We demonstrated the NGS analysis of the rice sample and subsequent data analysis to prove the nonexistence of transgenes that could be introduced during the breeding processes.

Materials and methods

1. Materials

It is well known that the rice plant acquires tolerance to the herbicide bispyribac sodium by introducing two-point mutations (W548L and S627I) in the acetolactate synthase (ALS) gene (Nishizawa-Yokoi et al. 2015). To establish a universal strategy for producing mutant plants harboring only the desired mutation in the target locus, the W548L and S627I mutations were introduced into the ALS gene locus via GT with positive–negative selection and subsequent excision of the positive selectable marker gene from the ALS gene locus using piggyBac transposon (Nishizawa-Yokoi et al. 2015). An overview of GT is given in Figure 1. Briefly, callus was induced from a rice cultivar, Nipponbare. The GT vector—which has two expression cassettes of diphtheria toxin A subunit (DT-A) gene as a negative selection marker and a 6.4-kb fragment containing acetolactate synthase (ALS) gene with two-point mutations and piggyBac transposon harboring the expression cassette of hygromycin tolerance gene as a positive selection marker—was introduced into *Agrobacterium*. The cells of *Agrobacterium* were inoculated into the rice calli. A transgenic line was selected on the medium containing hygromycin B and the antibiotic meropenem, which kills *Agrobacterium*. Because most of the T-DNA introduced cells were killed by the expression of the DT-A gene, only the cells carrying HR events at the ALS locus were selected on the medium. Genomic DNA extracted from hygromycin-resistant calli was subjected to PCR analysis for the identification of GT calli. The GT callus lines were cultured for 4 weeks. The GT calli were again infected with *Agrobacterium* harboring the piggyBac transposase (PBase) gene and neomycin tolerance gene in T-DNA. The PBase gene was used for the removal of the hygromycin tolerance gene integrated between the mutated ALS genes in the first genetic recombination. The PBase-expressing calli were selected on medium containing geneticin and meropenem and were

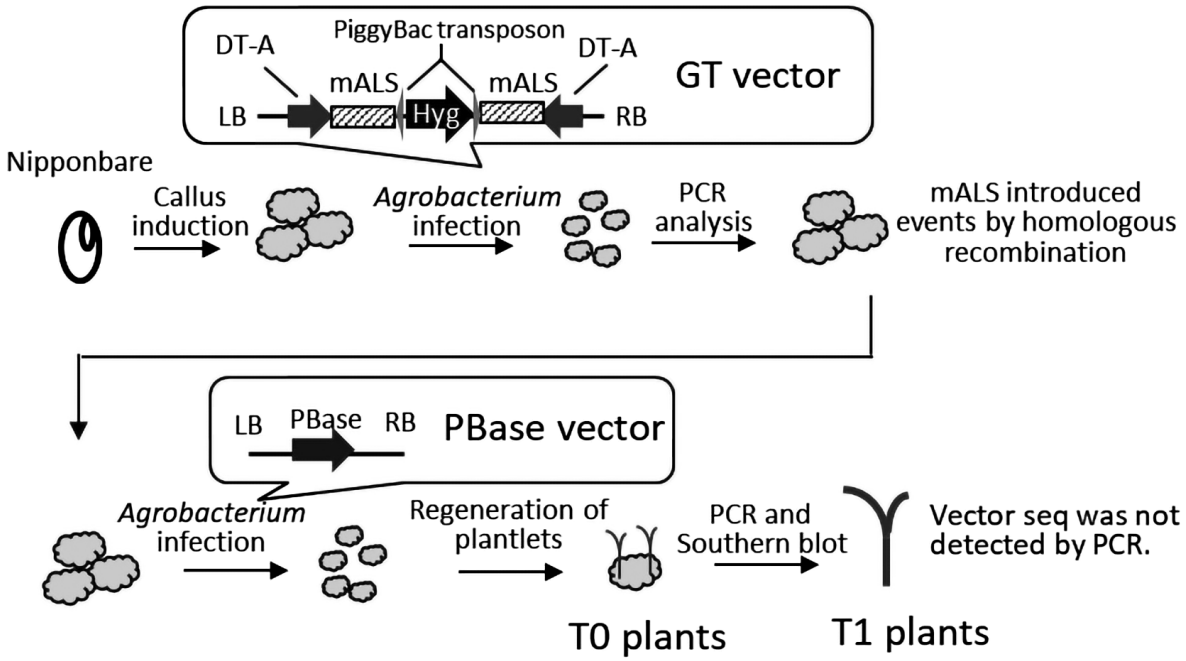

Fig. 1. Overview of gene targeting and the preparation of T1 generation plants for whole-genome sequencing in this study

DT-A, expression cassette of diphtheria toxin A subunit; LB, left border; mALS, mutated acetolactate synthase gene; RB, right border

regenerated. The regenerated plants were subjected to a marker excision analysis using cleaved amplified polymorphic sequences. We obtained T1 plants from self-pollinating marker-free T0 plants containing the mutated ALS gene and subjected them to a segregation analysis of the mutated ALS gene and PBase cassette, as well as an analysis on the transcript level of the ALS gene and an herbicide susceptibility test. Finally, a T1 plant in which no transgenes were detected by PCR (other than point mutations in the ALS gene) was selected. We used its leaves for the whole-genome sequencing in this study.

For the preparation of DNA samples spiked into rice DNA for whole-genome sequence analysis, we used reference materials of commercially available GM maize and soybean. A maize flour sample, including a GM event (i.e., MON810), and a flour sample of a GM soybean event (i.e., MON89788) were purchased from Sigma-Aldrich (St. Louis, MO, USA) and American Oil Chemists' Society (Urbana, IL), respectively.

2. Sample preparation for the whole-genome sequence analysis

For the DNA extraction from the rice sample, 2 g of rice leaves was washed with ultrapure water and ground using a mortar and pestle with liquid nitrogen. The ground sample was mixed with 7 mL of cetyltrimethyl ammonium bromide (CTAB) buffer at 60°C and incubated at 56°C for 20 min (Murray & Thompson 1980). The sample mixed with 7 mL of chloroform and isoamylalcohol (24:1) was then incubated for 20 min. After centrifugation at 3,000 rpm for 10 min, the supernatant was transferred to another plastic tube. The sample was remixed with 7 mL of chloroform and isoamylalcohol (24:1), and the supernatant was obtained. Ethanol precipitation was performed, and the pellet of DNA was dried. The DNA pellet was dissolved with 400 µL of sterile ultrapure water and 1 µL of RNase A (100 mg/mL, Nippon Gene, Tokyo) and then incubated at 37°C for 60 min. The solution was mixed with phenol and chloroform (1:1) and vortexed. After centrifugation, the supernatant was subjected to ethanol precipitation, and the obtained DNA pellet was solved into 150 µL of sterile ultrapure water.

For positive control in the data analysis, we spiked simulated transgene into the prepared rice genomic DNA sample. Genomic DNA was extracted from MON810 maize flour and MON89788 soybean flour using a DNeasy Plant Maxi kit (Qiagen, Hilden, Germany) in accordance with the manufacturer's protocol. Next, the PCR products of transgenes in MON810 and MON89788 were prepared. The reaction mixture (20 µL) was composed of 2 µL of 10× buffer for KOD-Plus, 0.2-mM dNTPs, 0.25-mM MgSO4, 300 nM of each primer, 40 ng of DNA template, and 0.1 unit of KOD-Plus DNA

polymerase. The thermal cycling conditions were as follows: 2 min at 94°C, 35 cycles of 15 s at 94°C, 0.5 min at 60°C, 6 min at 68°C, and finally 7 min at 68°C.

For the amplification of MON810 transgene (3,589 bases), MON810 Fw 5′-ttttttggccggccGCTA TCTGTCACTTTATTGTGAA-3′ and MON810 Rv 5′-ttttttggccggccGGTCGGTGCAGCCCACA-3′ were used as primers. For the amplification of MON89788 transgene (3,679 bases), MON89788 Fw 5′-ttttttcctagg GTATGACGAACGCAGTGAC-3′ and MON89788 Rv 5′-ttttttcctaggGATGCGGCCGCTTCGAG-3′ were used as primers. We designed small character regions in the primer sequences for the restriction enzyme treatment, but we did not use them in the present study.

After the agarose gel electrophoresis of the PCR amplicons, we extracted DNA from the gel using a MinElute DNA gel extraction kit (Qiagen), and the DNA was then purified by ethanol precipitation. UV absorbance was measured at 260 nm, and the DNA concentration was calculated. The copy numbers of each amplicon were calculated from the theoretical molecular weight of the amplicon and the measured UV absorbance.

The copy number of the rice haploid genome can be measured by an established real-time PCR assay for rice sucrose phosphate synthase gene (Genbank Accession No. D45890.1) (Takabatake et al. 2015). Real-time PCR assay was used to calculate the copy number in the DNA sample from the GT rice. Based on the copy number of the rice genome, the MON810 and MON89788 amplicons were mixed into the rice DNA sample. This resulted in a sample containing 606,000 copies/μL of rice genome, 60,600 copies/μL of MON89788 amplicon, and 6,060 copies/μL of MON810 amplicon. This mixed DNA sample was subjected to the following whole-genome sequencing.

3. Whole-genome sequencing

The rice DNA sample containing MON89788 and MON810 amplicons were fragmented, and then a genome shotgun library for pair-end sequencing using a TruSeq DNA PCR-Free Library Preparation Kit (Illumina, San Diego, CA) was prepared. The library sample was then subjected to a single-lane analysis of a next-generation sequencer (HiSeq 2000, Illumina). The read length was set as 100 bases.

4. Data analyses

As a reference sequence of rice genome, we used the file "Os-Nipponbare-Refference-IRGSP-1.0" in FASTA format from the website of The Rice Annotation Project Database: http://rapdb.dna.affrc.go.jp/download/irgsp1.html. The vector sequences used for GT were named "GT vector" and "PBase vector" (Fig. 1).

Data analysis for detecting homologous sequence was performed between recombinant vectors and the rice line made by GT. The workflow of the data analysis is summarized in Figure 2. First, the raw data of the reads from the HiSeq 2000 were analyzed using the read-trimming tool Trimmomatic ver. 0.32 (Bolger et al. 2014), and the adapter sequences and low-quality reads were removed. The cleaned reads were then mapped to the reference sequence using the software package BWA ver. 0.7.10 (Li 2013). The mapped data were merged into one data file using SAMtools ver. 0.1.19 (Li 2009).

The nucleotide sequences that differed from the reference sequence were extracted, and the variants were filtered under the following parameters: minimum root mean square mapping quality, 15; minimum read depth, 8; maximum read depth, 10,000,000; minimum (min.) number of alternate bases, two; single-nucleotide polymorphisms (SNPs) within INT base pairs (bp) around a gap to be filtered, 3; window size for filtering adjacent gaps, 10; min. *P*-value for strand bias (given PV4), 0.0001; min. *P*-value for baseQ bias, 1.00E-100; min. *P*-value for mapQ bias, 0; min. *P*-value for end distance bias, 0.001. The inserted sequences of >10 bases were listed, and we performed a homology search with the vector sequences by a nucleotide BLAST (Basic Local Alignment Search Tool) of BLAST+ (National Center for Biotechnology Information, https://blast.ncbi.nlm.nih.gov/Blast.cgi).

The reads unmapped to the reference sequence were extracted using SAMtools, and we submitted them to a *de novo* assembly by Velvet. The vector sequences were split into 24 bases, and their lists were used as a database. The contigs were then subjected to a homology search with the database of the split vector sequences by a nucleotide BLAST in BLAST+. Subsequently, a contig that completely matched the vector sequence was searched in the online nucleotide BLAST with the National Center for Biotechnology Information (NCBI's nonredundant database).

Finally, to evaluate the analytical performance of whole-genome sequencing, the MON89788 and MON810 amplicon sequences were searched from the contigs using BLAST+. In addition, the reads unmapped to the reference rice genome were mapped to the MON89788 and MON810 amplicon sequences with BWA, and the coverages were calculated using SAMtools ver. 0.1.19.

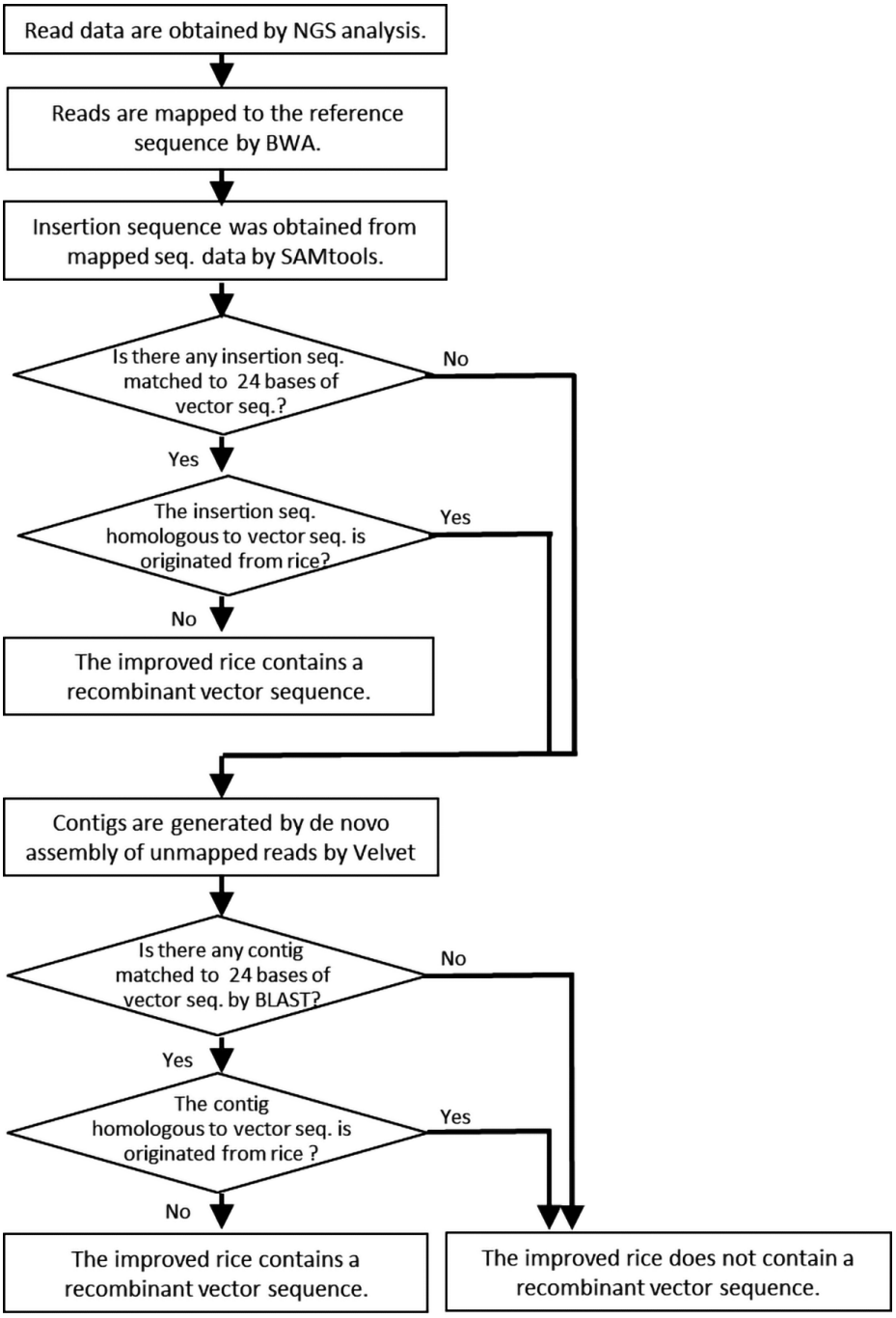

Fig. 2. Schematic of data analyses for detecting homologous sequence between recombinant vectors and the rice line made by gene targeting

Results and discussion

1. Threshold sequence length for the recognition of transgenes

To detect short transgene integration in the GT rice genome, it is necessary to find sequences that are homologous to the recombinant vectors, i.e., the GT and PBase vectors, from the NGS read data. Because short DNA sequences tend to be fortuitously homologous between the recombinant vector and the rice genome, we needed to predefine the minimum sequence length to be recognized as a transgene. We calculated the possibility that one or more parts of a vector sequence of a certain length completely matched with the GT rice genome. We set x as the length of the nucleotide sequence that completely matched between the genome and the vector.

The GT rice sample was cultivated using two recombinant vectors as described in the Materials and methods section. The total lengths of the vector sequences were approx. 4×10^4 bases. We set the rice haploid genome size as approx. 4×10^8 bases. If genome sequences have no bias in sequence, the probability y that one or more part of a vector sequence at the length of x bases is found in the rice genome was calculated using the following formula:

$$y = 1 - \left(1 - \frac{4 \times 10^4}{4^x}\right)^{2 \times 4 \times 10^8}$$

The relationship between x and y is illustrated in Figure 3. The probability decreases as the matching sequence length becomes longer. A perfect match of 24 bases occurs at the probability of 10.7%. The international guidelines note that 6-8 individual residues in an amino acid sequence of a recombinant protein should be evaluated for safety evaluations of new GM events (FAO/WHO 2001, CODEX 2003, Herman et al. 2009). We speculated that it may be necessary to recognize at least 24 bases corresponding to eight amino acids as transgene introduction. Therefore, the 24 bases were determined as the threshold length of a transgene in this study. The perfect matches of the vector sequence and the GT rice genome at <24 bases were ignored.

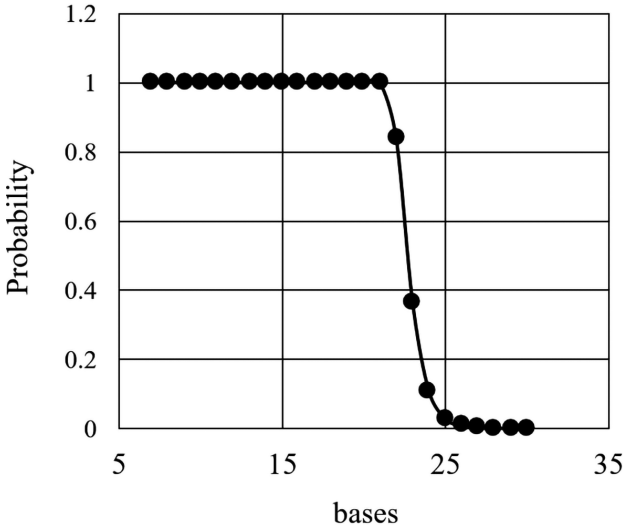

Fig. 3. Theoretical probability of at least one partial sequence of recombinant vector perfectly matching the rice genome
The probability is strongly dependent on the length of partial sequence.

2. Whole-genome sequencing

To prevent genomes of different variants from being included in the samples, we used leaves from one plant body of the GT rice. We selected a PCR-free kit for the NGS library preparation to avoid PCR bias leading to low coverage of the genome sequence. The rice genome size was estimated as 0.39 G bases (International Rice Genome Sequencing Project 2005), and we used the NGS HiSeq 2000, which provides sufficient read data. To generate high-quality and alignable sequence data, we performed pair-end sequencing. The results of the whole-genome sequencing are summarized in Table 1 (A). The read data were subjected to analysis using the Trimmomatic to remove adapter sequences and low-quality reads. The cleaned reads were then mapped to the reference sequence using BWA. The results are summarized in Table 1B. The total length of the mapped read data was >100-fold that of the reference genome (376.3 M of the mapped reads equals the 37.6 G bases; the reference sequence is 0.37 G bases). Sims et al. (2014) suggested that an average mapped read depth of 50× would be required for the reliable calling of single-nucleotide variants and small indels across 95% of the genome. We concluded that a sufficient amount of DNA sequence was provided for detecting small insertion sequences in the GT rice genome.

In addition, it is well known that no less than 10× sequence depth is required for the reliable detection of SNPs (Illumina website https://jp.illumina.com/science/education/sequencing-coverage.html). In the present study, 99.7% of the reference sequence was covered by the read data at the 10× sequence depth Table 1 (B). This value was also a good indicator for the coverage of a sample genome via NGS analysis.

3. Confirmation of the absence of transgene in the region corresponding to the rice reference genome

The SNPs and indels were picked up from the mapping results. It was confirmed that the two-point mutations on the ALS gene were present in the GT rice (data not shown). The total numbers of SNPs and indels were counted as shown in Figure 4 (A). The relationship between the length and the number of inserted sequences is illustrated in Figure 4 (B). The only three inserted sequences were of ≥24 bases (Fig. 4 (B)). Because complete 24 base match between the inserted sequences and vector sequences was determined as a criterion to identify unintentional transgene introduction, we subjected these three sequences to a homology search with the GT and PBase vector sequences. Because no match with the nucleotide sequences of GT and PBase vectors was found, we concluded that transgenes derived from vectors were not present in the region corresponding to the rice reference genome sequence. For reference, all

Table 1. Results of NGS analysis (A) and mapping to the rice reference genome sequence (B)

(A) Summarization of NGS analysis

	Yield (Mbp)	The number of cluster	% Q30	Mean Q
Index 1	23,101	115,505,978	79.03	32.00
Index 2	27,770	138,851,444	79.07	32.02
Total	50,871	254,357,422	Not calculated	Not calculated

(B) Summarization of mapping to the reference genome

Total reads (M)	Cleaned reads (M)	Reads mapped (M)	Reads unmapped (M)	Mapping rate (%)	Reference covered		
					1×depth	10×depth	30×depth
508.7	385.6	376.3	9.3	97.19	99.9%	99.7%	96.7%

(A)

Type of mutation	Frequency
SNPs	41,724
Indels	4,890

(B)

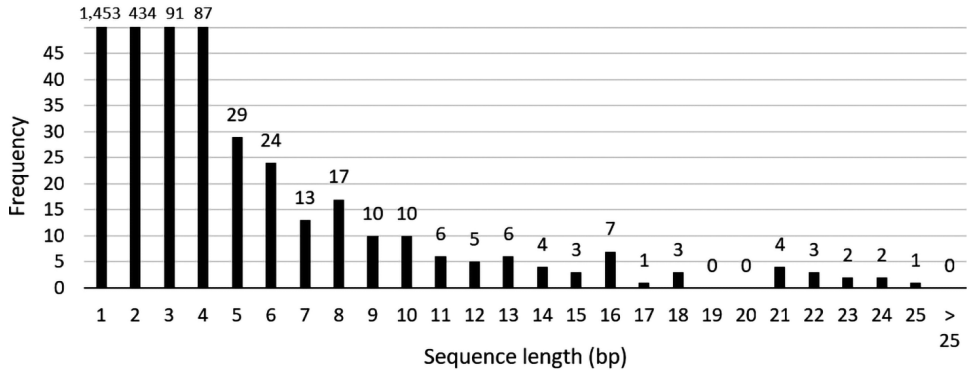

Fig. 4. Summarized data of SNPs and indels in the rice line made by gene targeting
A, the number of total SNPs and indels; B, sequence length composition of insertion

the inserted sequences of >10 bases were additionally subjected to a homology search with the vector sequences. The longest inserted sequence completely matched with the vector sequences was just 12 bases in length (Table 2). Based on the 24-base criterion, these matched sequences were too short to identify as unintentional transgene introduction.

4. Confirmation of the absence of transgenes in regions other than the rice reference genome

The reference sequence of the rice genome is 0.37 G bases, and the rice genome size is estimated as 0.39 G bases (International Rice Genome Sequencing Project 2005). This difference comes from the difficulty to read the nucleotide sequences that have secondary structure and/or sequence repeats such as a centromere

Table 2. Homology search of inserted sequences of >10 bases against the nucleotide sequences of GT and PBase vectors

Insert name	Reference vector	Alignment length (bp)	Insert start position	Insert end position	Reference start position	Reference end position	E-value	Score	Alignment sequence
chr05_9087326	GT	11	3	13	13,181	13,191	0.18	21.4	AAAATTCAAAA
chr11_3216779	GT	7	1	7	743	749	7.1	14	AATTAAG
chr08_17446362	GT	8	2	9	2,720	2,727	2	15.9	AATTATGA
chr09_1505636	GT	9	4	12	2,145	2,153	0.55	17.7	ACAAGTATG
chr09_1505637	GT	9	4	12	2,145	2,153	0.55	17.7	ACAAGTATG
chr06_510023	GT	9	6	14	3,447	3,455	0.92	17.7	ATATACTGT
chr06_510025	GT	9	1	9	3,447	3,455	0.92	17.7	ATATACTGT
chr06_8529478	GT	10	1	10	11,476	11,485	0.15	19.6	ATATATATAT
chr04_9839049	GT	8	2	9	2,094	2,087	2	15.9	ATATGCCA
chr08_16257741	GT	9	1	9	1,775	1,767	0.55	17.7	ATGAATATG
chr02_19391315	GT	9	4	12	12,413	12,405	0.55	17.7	ATTCAGGTC
chr07_13315706	GT	9	1	9	4,973	4,965	0.92	17.7	CAATTCCCT
chr08_3227892	GT	8	7	14	3,278	3,271	2.6	15.9	CAGCAGAC
chr03_31805197	GT	9	2	10	21,930	21,922	0.55	17.7	CAGGAAAGA
chr08_3235409	GT	8	3	10	13,764	13,757	2	15.9	CGCTGTGT
chr05_5791116	GT	12	7	18	5,075	5,086	0.028	23.3	GCCGCCGCCGCC
chr01_30647503	GT	11	3	13	20,129	20,119	0.13	21.4	GCCTTCCATCC
chr01_30647507	GT	11	3	13	20,129	20,119	0.13	21.4	GCCTTCCATCC
chr06_25435809	GT	9	3	11	1,332	1,324	0.55	17.7	GTAGAGAGG
chr12_11862516	GT	10	1	10	13,752	13,761	0.2	19.6	GTGAAACACA
chr04_7922290	GT	8	2	9	3,600	3,593	3.3	15.9	TAAATCGT
chr12_22666145	GT	8	10	17	11,599	11,606	4	15.9	TAGTTCAA
chr01_21833937	GT	11	4	14	439	449	0.099	21.4	TATACATATAT
chr04_928347	GT	8	1	8	473	466	2	15.9	TATTATTC
chr10_7257912	GT	7	4	10	1,914	1,908	7.1	14	TGAACGA
chr08_16257742	GT	8	2	9	1,774	1,767	2	15.9	TGAATATG
chr03_24407222	GT	9	6	14	6,275	6,283	0.92	17.7	TGCCCAAGC
chr10_8063252	GT	8	4	11	10,620	10,613	2	15.9	TGCGTCTC
chr04_927614	GT	9	2	10	3,856	3,848	0.55	17.7	TGTGTTTAG
chr04_927616	GT	9	2	10	3,856	3,848	0.55	17.7	TGTGTTTAG
chr11_3216773	GT	8	3	10	6,680	6,687	2	15.9	TTAAGCTT
chr08_6751352	GT	9	15	23	2,895	2,887	2.2	17.7	TTAATAGTT
chr10_8096582	GT	8	1	8	4,152	4,145	2	15.9	TTAATTGT
chr08_19728774	GT	9	5	13	8,434	8,442	0.55	17.7	TTAGCCTTT
chr10_2469708	GT	10	5	14	9,011	9,002	0.2	19.6	TTCCTTTTCC
chr01_15530723	GT	10	1	10	7,089	7,080	0.15	19.6	TTGTTACTTT
chr12_15909059	Pbase	10	6	15	5,182	5,191	0.51	19.6	ATTTGAGTTG
chr01_41520031	PBase	11	1	11	1,865	1,875	0.043	21.4	CACATCATCAG
chr04_4507544	PBase	11	2	12	1,102	1,092	0.043	21.4	GAAGAAGAGGA
chr10_7943162	PBase	9	1	9	6,900	6,892	0.55	17.7	GTATGTATA
chr02_22795020	PBase	9	1	9	5,344	5,352	0.55	17.7	TCAGCCGCC
chr04_33800682	PBase	9	3	11	1,922	1,914	0.73	17.7	TCAGGGTCT
chr10_2469711	PBase	10	4	13	5,423	5,432	0.2	19.6	TTTTCCTCTT

by sequencing, and it is also difficult to assemble them in the subsequent data analysis (Claros et al. 2012). The nucleotide sequences corresponding to approx. 5% of the whole rice genome are thus not included even in the latest reference sequence.

To evaluate the absence of transgenes in the unexplored regions in the rice genome, we extracted the reads unmapped to the reference sequence data and submitted them to the *de novo* assembly. We obtained 62,950 contigs. We then performed a homology search of contigs with the database of 24-base sequences from the vector sequences by nucleotide blast. Only one contig was matched with the vector sequences. To determine whether the matched contig sequence is derived from the rice genome, we performed a homology search using a nucleotide BLAST with the nr database. Because that contig sequence matched the rice actin2 gene (Genbank Accession No. EU155408.1), except for one indel, we concluded that the contig was not a transgene derived from the recombinant vector (Fig. 5).

For the performance evaluation of this analysis, we used the sequence data of spiked DNAs as an indicator. We added MON89788 and MON810 amplicons into the rice genome sample at the copy numbers of 1/10 and 1/100, respectively. The sequence data corresponding to these input DNAs should be classified into the reads unmapped to the reference rice genome, as expected. We performed the mapping of the read data to the MON89788 and MON810 amplicon sequences, and the coverage calculated using SAMtools for MON89788 and MON810 was 3.13 and 0.16, respectively. Because the MON89788 and MON810 amplicons were spiked at the copy numbers of 10% and 1% of the genome, the sequence depth for each amplicon was expected to be 10 and 1, respectively. The results were dependent on the amount of spiked DNA although the actual data were smaller than expected.

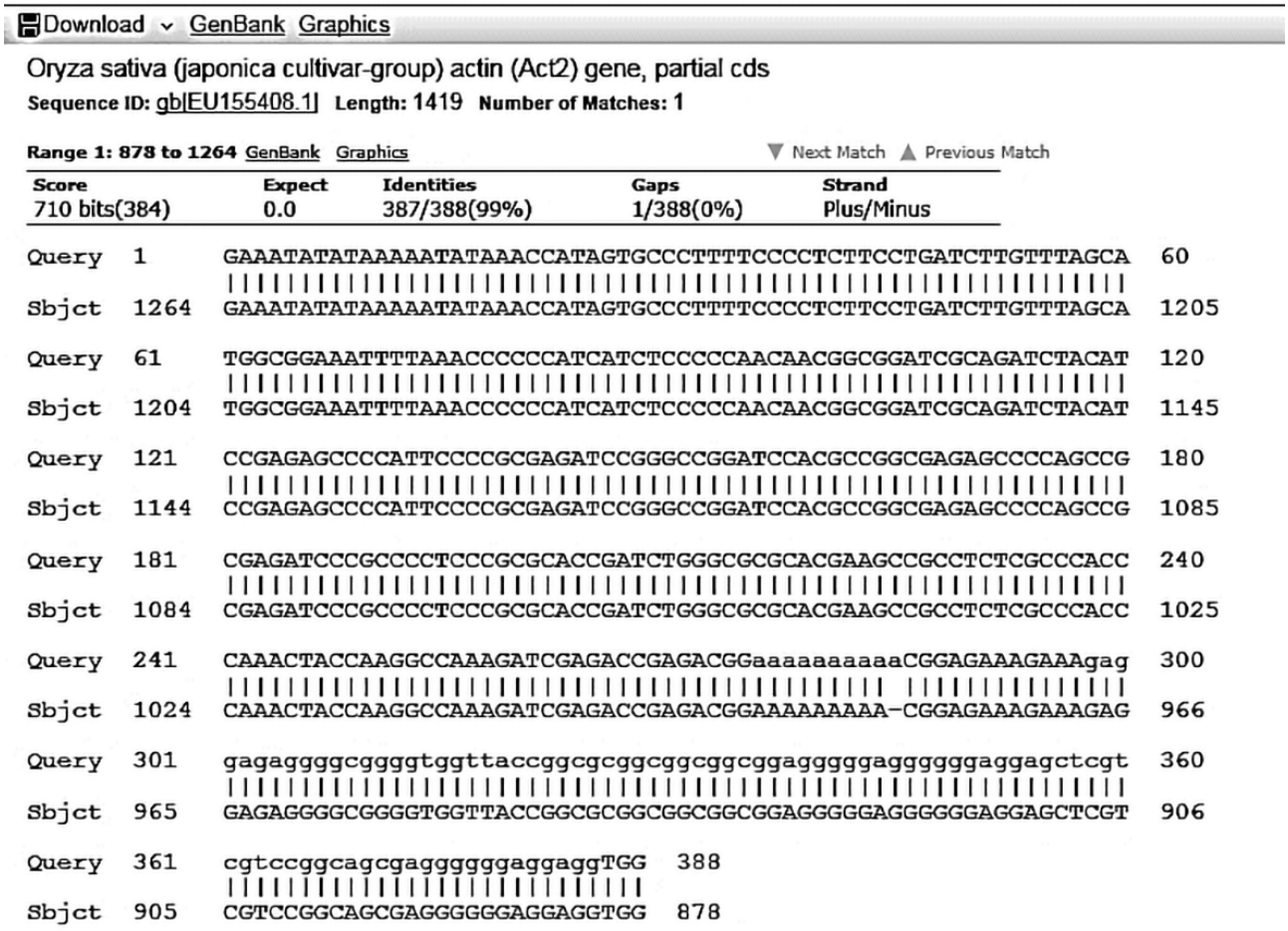

Fig. 5. Contig sequence derived from the reads unmapped to the rice reference genome showed the homology with the *Oryza sativa* actin2 gene

The contig sequence was shown as "query" in the result of nucleotide BLAST. The lower-case letters in the query sequence indicates low complexity regions that were automatically filtered in the BLAST search.

In addition, the MON89788 and MON810 amplicon sequences spiked into rice genome were searched from total contigs obtained by the de novo assembly (62,950 contigs). We detected contigs that matched with the MON89788 amplicon. Contigs corresponding to the MON810 amplicon were not obtained because the amount of read data was too small. These results indicated that the contig data prepared by the *de novo* assembly included genome sequence although its copy number was 1/10th of the genomic DNAs. The copy number of a part of a genome tends to be small in a sequence library, which is attributed to their high GC content and/or palindromic sequences (Star et al. 2014). For example, the sequences with high or low GC content showed coverage that is reduced to 1/5th of the average value by the HiSeq system (Ross et al. 2013). These previously reported data indicated that the sequencing run in the present study produced sufficient amounts of reads for the analysis of difficult sequences. We concluded that our NGS data could cover unexplored regions other than the reference sequence on the rice genome, and it would include the sequences that tend to be difficult to read owing to the sufficient sequencing depth.

Conclusion

We demonstrated an analysis that confirms the absence of transgenes that are of ≥24 bases in GT rice using whole-genome sequencing. Some research groups have recently revealed that the *Agrobacterium*-mediated genetic recombination sometimes causes short transgene integration into plant genomes (Schouten et al. 2017). Therefore, the evaluation of genome integrity described herein may be required in the future for the commercialization of GT plants. For HR-mediated GT techniques and targeted mutagenesis methods using TALENs and CRISPR/Cas9, a transgene is integrated into the rice genome, and the transgene is commonly removed by conventional breeding. Our genome integrity evaluation method summarized in Figure 2 can also be applied to plants made using genome-editing techniques.

The whole-genome sequencing approach can detect a short transgene sequence, and this approach provides more precise evaluations compared with conventional methods such as PCR or southern blotting. In general, there is no way to prove the nonexistence of something, i.e., "devil's proof." Certainly, whole-genome sequencing does not completely cover the "whole genome," and we cannot prove the absolute absence of transgenes. However, the NGS analysis provides a large amount of sequence data that does not include transgenes, and this enables us to evaluate the degree of the reliability of the conclusion that there is no transgene in the genome. This characteristic of the whole-genome sequencing approach is completely different from the conventional methodology used to determine the absence of transgenes, such as PCR and southern blotting.

The performance of sequencing technologies is rapidly improving. Long-read sequencers such as PacBio and Oxford Nanopore are also commonly used. The completeness of the coverage of a genome will continue to increase, and a decline in the cost is expected. Advances in sequencing technologies would enhance the usefulness of our approach to the evaluation of genome integrity.

Acknowledgements

We thank the late Professor Hiroshi Kamada at the University of Tsukuba for his kind advice to start this research. This research was partly funded by grants from the Ministry of Agriculture, Forestry and Fisheries of Japan, "Research Project for Genomics-based Technology for Agricultural Improvement GRA-201-1-1."

References

Arulandhu, A. J. et al. (2016) DNA enrichment approaches to identify unauthorized genetically modified organisms (GMOs). *Anal. Bioanal. Chem.*, **408**, 4575-4593.

Bolger, A. M. et al. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.

Claros, M. G. et al. (2012) Why assembling plant genome sequences is so challenging. *Biology (Basel)*, **1**, 439-459. https://doi.org/10.3390/biology1020439.

CODEX Alimentarius. Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants CAC/GL 45-2003, 2003. http://www.fao.org/fileadmin/user_upload/gmfp/docs/CAC.GL_45_2003.pdf. Accessed on 15 February 2018.

FAO/WHO. Evaluation of allergenicity of genetically modified foods, 2001. http://www.fao.org/fileadmin/templates/agns/pdf/topics/ec_jan2001.pdf. Accessed on 15 February 2018.

Guttikonda, S. K. et al. (2016) Molecular characterization of transgenic events using next generation sequencing approach. *PLoS ONE*, **11**, e0149515.

Herman, R. A. et al. (2009) Value of eight-amino-acid matches in predicting the allergenicity status of proteins: an empirical bioinformatic investigation. *Clin. Mol. Allergy*, **7**, 9.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793-800.

Kovalic, D. et al. (2012) The use of next generation sequencing and junction sequence analysis bioinformatics to achieve molecular characterization of crops improved through modern biotechnology. *Plant Genome*, **5**, 149-163.

Latham, J. R. et al. (2006) The mutational consequences of plant transformation. *J. Biomed. Biotechnol.*, **25376**, 1-7.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]. https://arxiv.org/abs/1303.3997. Accessed on 13 June 2022.

Li, H. et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

Lusser, M. et al. (2012) Deployment of new biotechnologies in plant breeding. *Nat. Biotechnol*., **30**, 231-239.

Murray, M. G. & Thompson W. F. (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.*, **8**, 4321-4325.

Nishizawa-Yokoi, A. et al. (2015) Precision genome editing in plants via gene targeting and piggyBac-mediated marker excision. *Plant J*., **81**, 160-168.

Ross, M. G. et al. (2013) Characterization and measuring bias in sequence data. *Genome Biol*., **14**, R51.

Sanagala, R. et al. (2017) A review on advanced methods in plant gene targeting. *J. Genet. Eng. Biotechnol*., **15**, 317-321.

Schouten, H. J. et al. (2017) Re-sequencing transgenic plants revealed rearrangements at T-DNA inserts, and integration of a short T-DNA fragment, but no increase of small mutations elsewhere. *Plant Cell Rep.*, **36**, 493-504.

Sims, D. et al. (2014) Sequencing depth and coverage: Key considerations in genomic analyses. *Nat. Rev. Genet*., **15**, 121-132.

Sprink, T. et al. (2016) Regulatory hurdles for genome editing: Process- vs. product-based approaches in different regulatory contexts. *Plant Cell Rep.*, **35**, 1493-1506.

Star, B. et al. (2014) Palindromic sequence artifacts generated during next generation sequencing library preparation from historic and ancient DNA. *PLoS ONE*, **9**, e89676.

Takabatake, R. et al. (2015) Comparison of the specificity, stability, and PCR efficiency of six rice endogenous sequences for detection analyses of genetically modified rice. *Food Cont*., **50**, 949-955.

Wolt, J. D. et al. (2016) The regulatory status of genome-edited crops. *Plant Biotechnol. J*., **14**, 510-518.

Yang, L. et al. (2013) Characterization of GM events by insert knowledge adapted re-sequencing approaches. *Sci. Rep.*, **3**, 2839.

Zastrow-Hyes, G. et al. (2015) Southern-by-sequencing: a robust screening approach for molecular characterization of genetically modified crops. *Plant Genome*, **8**, 1-15.