Comprehensive Characterization of Transgene Integration in a Similar Intragenesis Rice Event Using Paired-end Whole-genome Sequencing Approach

Wenting Xu

Shanghai Jiao Tong University

Hanwen Zhang

Shanghai Jiao Tong University

Yuchen Zhang

Shanghai Jiao Tong University

Rong Li

Shanghai Jiao Tong University

Litao Yang (✉ yyltt@sjtu.edu.cn)

Shanghai Jiao Tong University <https://orcid.org/0000-0002-4515-5722>

Article

Keywords: molecular characterization, rice event G281, intragenesis, paired-end whole-genome sequencing, T-DNA insertion

Posted Date: October 12th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-951165/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Version of Record: A version of this preprint was published at Communications Biology on July 5th, 2022. See the published version at <https://doi.org/10.1038/s42003-022-03608-1>.

Abstract

Efficient, accurate molecular characterization of genetically modified (GM) organisms is challenging, especially for novel transgenic products of *cisgenesis/intragenesis* transferred with genes/elements of recipient species. Herein, GM rice event G281, involving transfer with native promoters and an RNA interference (RNAi) expression cassette in a process similar to *intragenesis*, was subjected to molecular characterization using paired-end whole genome sequencing (PE-WGS). The results showed that transgenes integrated at rice chromosome 3 locus 16,439,674 included a 36 bp deletion of rice genomic DNA, and the whole integration contained two copies of the complete transfer DNA (T-DNA) in a head-to-head arrangement. No unintended insertion or backbone sequence of the transformed plasmid were observed at the whole genome level. Molecular characterization of the G281 event will assist risk assessment and application for a commercial license. Additionally, the findings demonstrate the applicability of PE-WGS for molecular characterization of *cisgenesis/intragenesis* crops.

1. Introduction

The commercialization of genetically modified organisms (GMO) has brought great benefits to farmers, consumers, and the environment (ISAAA, 2019). At the request of all countries and regions, the biosafety of each GMO event should be thoroughly evaluated in terms of molecular characterization, food safety, and environmental safety. Only events that satisfy risk assessment should be approved for production and commercialization (Sparrow PA, 2010).

Molecular characterization of GMOs at the genome level often generates information on exogenous DNA integration and its inheritable stability, and a full understanding of exogenous DNA integration is a fundamental step in molecular characterization (Miraglia, et al., 2004). Exogenous DNA integration in GMOs includes integrated exogenous genes, integration sites, flanking DNA sequences of all inserts, the copy number of all inserts, and the presence or absence of unexpected vector backbone sequences (Codex, 2003). This analysis plays a pivotal role in GMO biosafety supervision, quantitative detection, establishment of detection methods, and analysis of the content and traceability of GMO products (Yang et al., 2013). This information is crucial for labeling for assessing the phenotype structural stability, food safety, and commercial value of new GM materials (Kuiper, et al., 2010; Guttikonda, Marri et al. 2016). In the guidelines for GMO risk assessment, several techniques have been specified for probing exogenous DNA integration including PCR-based chromosome walking, Real-time PCR, and Southern blotting, etc. (Li et al., 2017) PCR-based chromosome walking approaches coupled with Sanger sequencing are used for identifying exogenous DNA integration and flanking recipient genomic regions, but these techniques also have shortcomings, especially sequence bias and failure to generate comprehensive data, which can lead to missing insertion sites, especially for GM events with multiple exogenous DNA insertions (Li et al., 2017). Real-time PCR and Southern blotting are recommended for determining the copy number of exogenous genes and unintended insertion of backbone sequences of transformed vectors (Li et al., 2017). However, real-time PCR is easy to acquire ambiguous results when studying genes with multiple copies and/or tandem repeats of exogenous genes (Li et al., 2017). Southern blotting requires a large number of samples, takes a long time, requires technically accomplished researchers, and the selection of different enzyme digestion sites can lead to differences in copy number (Li et al., 2017). Molecular characterization of most commercialized GM crop events in the past few decades has employed the above methodologies, but publically available data is not always complete and accurate due to the inherent limitations of the methodologies (Yang et al., 2013).

High-throughput next-generation resequencing (NGS) is increasingly used in many research areas as the cost decreases dramatically. NGS has been successfully coupled with bioinformatics pipelines for full or partial molecular characterization of GMO events based on the resequencing technique. Zhang et al., identified integration sites and

their flanking DNA sequences in transgenic cattle (Zhang et al., 2012). Two transgenic glyphosate-tolerant soybean insertion sites and flanking sequences were identified using whole-genome sequencing (WGS) (Guo et al., 2016). The precise insertion loci and copy number of transfer DNAs (T-DNAs) in transgenic rice plant lines SNU-Bt9–5, SNU-Bt9–30, and SNU-Bt9–109 were determined using NGS-based molecular characterization methods, and the data were comparable with those obtained from Southern blotting analysis (Doori et al., 2017). Furthermore, corresponding bioinformatics pipelines for analysis of NGS data have been developed (Yang et al., 2013; Zhang et al., 2020). NGS technology combined with target enrichment was used to identify genetic integrations from complex food/feed samples (Zastrow-Hayes et al., 2015; Košir et al., 2018). Previous studies have successfully employed WGS coupled with paired-end sequencing (PE-WGS) to identify transgene insertion sites and their flanking sequences. However, molecular characterization using current PE-WGS still remains challenging, especially for transgene insertions with endogenous recipient DNA or DNA sequences that are highly similar to host species (Vanblaere et al., 2014; Li et al., 2017).

As recombinant techniques develop and more gene functions are clearly annotated, novel GM crops are being produced via *cisgenesis* and *intragenesis*, along with traditional *transgenesis* introduction of exogenous genes (Holme et al., 2013). *Cisgenesis* involves using donor DNA fragment from the species itself or from a cross-compatible species without vector backbone DNA except T-DNA border sequences (Schouten et al., 2008). *Intragenesis* is more similar to *cisgenesis*, but *intragenesis* allows the creation of novel combinations of DNA fragments, such as combining functional genetic elements like promoters and terminators, and antisense or RNA interference (RNAi) cassettes (Schouten et al., 2008). Although *cisgenesis* and *intragenesis* mostly involve their own DNA fragments, risk assessment should still be performed, including molecular characterization (Eckerstorfer et al., 2019; Paraskevopoulos et al., 2021). However, reports on molecular characterization of *cisgenesis* and *intragenesis* in crops are limited, and there is only one report on deciphering the TDNA insertion of partial cisgenic lines of apple ‘Gala’ using inverse PCR (Vanblaere et al., 2014).

In order to explore the potential of NGS in molecular characterization of *cisgenesis/intragenesis*, transgenic rice event G281, generated by transforming the binary vector p1300-S450RNAi-hFL-G6 into rice line Xiushui-110 via *Agrobacterium*-mediated transformation, was used as an example similar to *intragenesis* (Lin et al., 2010). We performed molecular characterization of the G281 event using PE-WGS and a dedicated pipeline, and discussed the characterization of cisgene integrations in *cisgenesis/intragenesis*.

2. Materials And Methods

2.1 Plant materials

Transgenic rice event G281 is associated with high sensitivity to bentazon, high tolerance to glyphosate, and high expression of human lactoferrin (hLF) protein in seed, according to Zhejiang University, China. T-DNAs including tandem arrays of three expression cassettes of *hLF*, *G6 5-enolpyruvylshikimate-3-phosphate synthase (G6 EPSPS)*, and the RNAi of rice *CYP81A6 gene* were introduced into the non-GM line Xiushui-110 via *Agrobacterium*-mediated transformation (Lin et al., 2010). Seeds of transgenic rice event G281 and its recipient line (Xiushui 110) were kindly supplied by the developer. Rice plants of G281 and its recipient line were grown in a greenhouse in Shanghai Jiao Tong University, China, and fresh leaves were sampled for molecular characterization. Fresh sampled leaves were washed, dried, and ground into powder in liquid nitrogen for DNA extraction. Rice genomic DNA was extracted and purified using a Plant Genomic DNA Kit (Cat. no. DP305-3; TIANGEN). The quality and quantity of extracted DNA was evaluated with a Nanodrop ND-8000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and by 1%

agarose gel electrophoresis, respectively. Purified DNA samples were stored and used for paired-end whole-genome sequencing (PE-WGS), PCR verification, and droplet digital PCR (ddPCR) analyses. In PCR and ddPCR analyses, three individual biological replicates were performed.

2.2 PE-WGS

Approximately 5 µg of genomic DNA from GM or wild-type (WT) rice lines was used for PE library construction with an Illumina Paired-end Preparation Kit (Illumina Inc, USA) with an average insert size of 500 bp in length. Library quality was evaluated by Pico-Green (Quant-iT; Invitrogen, USA) and an Agilent Bioanalyzer DNA 1000 kit (Agilent Technologies, USA). Constructed libraries were subjected to sequencing on an Illumina HiSeq 2000 DNA sequencer (Illumina Inc, USA) by BGI-Shenzhen (Shenzhen, China), and 100 bp PE reads were generated. Raw sequencing data were preprocessed to remove index sequences, short and low-quality reads, and read lengths were unified. The resulting 100 bp trimmed sequencing read data were used for further analysis.

2.3 Bioinformatics pipeline for PE-WGS data analysis

PE-WGS data were analyzed using our previously developed bioinformatics pipeline with slight modifications to add functionality and avoid false-positive reads from host native genome DNAs (Figure 1). To reduce interference from native genome DNAs, false-positive reads were filtered by searching for the positions of similar sequences between plasmids and references using BLASTN (Chen et al., 2015). Candidate results were generated after blocking the positions $\pm$ the insert sizes of homologous sequences. Burrows-Wheeler Aligner (BWA version 0.6.2) (Li and Durbin, 2009), SPAdes Genome Assembler (SPAdes version 3.11) (Bankevich et al., 2012), Sequence Alignment/Map tools (SAM tools version 1.6) (Li et al., 2009), and Integrative Genomics Viewer (IGV, version 2.4) (Robinson et al., 2011) formed the pipeline for read mapping, assembly, alignment storage, and visualizing mapping results, respectively. Transformation plasmid sequences of G281 and rice reference gene sequences (RefSeq assembly accession: GCF_001433935.1) were used as reference sequences in the analysis.

2.4 Conventional PCR and Sanger sequencing

Primer pairs used for verifying transgene insertion sites and their flanking sequences were designed based on the candidate read pairs identified in PE-WGS analysis (**Supplementary Table S1**). Conventional PCR was performed with a Veriti thermocycler (ThermoFisher Scientific) in a total volume of 50 µl, including 5 µl of 10× Ex Taq Buffer (Takara, Japan), 5 µl of dNTPs (2 mM), 0.25 µl of Ex Taq polymerase (Takara), 2 µl of 10 µM forward primer, 2 µl of 10 µM reverse primer, 2 µl of DNA template, and 33.75 µl of RNase-free and DNase-free water. Each reaction was performed in triplicate, and thermal cycling included 5 min at 98°C, followed by 35 cycles of 30 s denaturation at 94°C, 30 s annealing at 55°C, and 1 min extension at 72°C, then an additional extension step at 72°C for 7 min. Amplified products were purified, sent to Invitrogen (Shanghai, China) for Sanger sequencing, and the sequencing results were used as query sequences for BLASTN searching (Camacho et al., 2008).

2.5 Droplet Digital PCR

The ddPCR primers and probes for *G6 EPSPS* and hFL genes were designed using beacon designer software version 8.0, and our previously reported primers and probes for rice endogenous reference gene *SPS* were also used (Yang, 2018). All primers and probes were purchased from Invitrogen (Shanghai, China) and are listed in **Supplementary Table S2**. The ddPCR assays were performed on a QX200 Droplet PCR platform (Bio-Rad, Pleasanton, CA, USA) with a final volume of 20 µl comprising 10 µl of 2× ddPCR Supermix (Bio-Rad, Pleasanton, CA, USA), 1 µl of 10 µM forward primer, 1 µl of 10 µM reverse primer, 0.5 µl of 10 µM probe, 1 µl of DNA template, and 6.5 µl of RNase-free and DNase-free water. After preparation of the 20 µl reaction mix for ddPCR, droplets were generated in 8-well

cartridges using a QX200 droplet generator (Bio-Rad, Pleasanton, CA, USA). Water-in-oil emulsions were transferred to a 96-well plate and amplified in a T100 PCR cycler (Bio-Rad, Pleasanton, CA, USA). The procedure for ddPCR assays included 5 min at 95°C, followed by 40 cycles of a two-step thermal profile comprising 30 s at 95°C and 60 s at 58°C at a ramp rate of 2.0°C/s. After cycling, each sample was incubated at 98°C for 10 min then cooled to 4°C. After PCR amplification, PCR plates were transferred to a QX200 droplet reader (Bio-Rad, Pleasanton, CA, USA), and data acquisition and analysis were performed using QuantaSoft (Bio-Rad, Pleasanton, CA, USA). All reactions were repeated in triplicate with three parallels.

3. Results

3.1 Sequencing data from PE-WGS

Transgenic rice G281 (GM) and non-GM rice Xiushui 110 (WT) were sequenced, and 11.7 GB and 12.2 GB data comprising 54,130,964 and 54,819,968 pairs of 100 bp reads were acquired, respectively. General sequencing coverage was up to 28.91× for GM rice and 29.28× for WT rice. Additionally, both ends of the 53,545,140 and 54,233,374 read pairs were mapped to the rice reference genome (**Table 1**). The calibration weight R with values of 0.986 for both GM and WT was used to modify bias among the mapped sequencing data, which was calculated from the percentage of read pairs that mapped to the reference genome among the total number of trimmed paired reads.

3.2 Exogenous DNA integration sites and flanking sequences

All trimmed read pairs were analyzed according to the pipelines depicted in **Figure 1**. For the G281 line, 53,545,140 read pairs and 2604 read pairs mapped to the rice reference genome and the transformed plasmid, respectively. A total of 322 read pairs were extracted for which one end mapped to the rice reference genome and the other mapped to the transformed plasmid using the original TranSeq pipeline without false-positive reads removed. Using the modified TranSeq pipeline, 2284 false-positive read pairs were discarded and only 69 read pairs were retained as candidate reads for determining the T-DNA insertion site and its flanking sequences (**Table 1**). Among the 69 potential candidate reads, 62 read pairs mapped to Chr 03, two pairs mapped to Chr 02, and one pair mapped to Chr 04, 06, 08, 11, and 12. After considering the sequencing depth (28.91×) and BLASTN analysis, the 62 read pairs mapped to Chr 03 were confirmed as true candidates (**Table 2**). The sporadic read pairs mapped to other chromosomes might be due to data jitter, sequencing errors, and host genome noise. IGV visualization of these candidate read pairs showed that only one transgene integration was introduced at locus 16,439,674 of Chr 03 based on clustering pattern analysis with reference rice (GCA_001433935.1; **Figure 2A**). The assembled eight contigs (**Supplementary Sequence File 1**) generated from the 62 read pairs in *de novo* analysis clearly showed the T-DNA integration site at Chr 03 locus 16,439,674 with a 36 bp rice host genome DNA deletion, and two contigs with lengths of 478 bp and 434 bp comprising the 5' and 3' end flanking sequences adjacent to transgene insertion (**Figure 2B**). Further conventional PCR results showed that the expected DNA fragments covering the adjacent regions of the inserted site were only observed in G281 rice, while no amplicons were observed in WT rice. Sanger sequencing of the expected amplicons confirmed that the obtained integration site and flanking sequence information were identical with those from PE resequencing analysis (**Figure 2C**).

The results of WT resequencing data yielded 54,819,968 and 322 read pairs that mapped to the rice reference genome and the transformed plasmid, respectively (**Table 1**). Only five read pairs were extracted as candidates in which one end mapped to the rice reference genome and the other mapped to the transformed plasmid, including

four reads pairs and one read pair that appeared to match Chr 02 and Chr 07, respectively (**Supplementary Table S3**). Further BLASTN analysis of these read pairs showed that all five were false-positives due to homologous sequences in transformation vectors or sequencing background noise. The WT resequencing results also confirmed its non-GMO authenticity, and the data were helpful for further identifying transgene integration in *intragenesis* rice line G281.

3.2 Copy number of exogenous DNA

To evaluate the copy number of the inserted transgenes based on the resequencing data, the formula

$$\text{CopyNumber}_{\text{Seq.Data}} = \frac{\text{ADT}}{\text{D} \times \text{R}}$$
 was used to calculate the copy number. In the formula, ADT represents the Average sequencing Depth of the Target gene, D represents the general sequencing depth, and R is the relative relationship between mapped reads and the host genome. To improve the accuracy of copy number evaluation and decrease the sequence bias in resequencing, the single copy rice endogenous reference gene, *SPS*, was used as a calibrator. The calculated copy number for the *G6 EPSPS* gene, the *hLF* gene, and the rice *Gt1* promoter was 2.11, 1.95, and 3.1, respectively (**Table 2**), indicating that there were two copies of each transferred into the rice genome.

Furthermore, the *G6 EPSPS* and *hLF* gene copy number was also analyzed using the ddPCR method with gene-specific primers and TaqMan probes to compare with the copy number from NGS analysis. The copy number was calculated as the ratio of transgene and rice endogenous reference gene *SPS* (copy/copy). The ddPCR results showed that the copy number of *G6 EPSPS* and *hLF* genes was 1.87 and 1.96, respectively (**Table 2**). The results from both resequencing data and ddPCR suggested that the copy number of *G6 EPSPS* and *hLF* genes was two, and the copy number from NGS analysis was slightly higher than that from ddPCR. Thus, based on the NGS and ddPCR results, we concluded that two copies of plasmid T-DNAs were inserted into the rice genome.

3.3 Arrangement and full sequencing of whole transgene integration

A total of 2604 read pairs in which both ends were well mapped to the transformed plasmid were extracted for rice G281. These extracted reads were used to assemble transgene integration using SPAdes. A total of four contigs were obtained, including one with a length of 9145 bp, while the other three were shorter (279 bp, 266 bp, and 213 bp; **Supplementary Sequence File 2**). BLASTN results showed that only the 9145 bp contig corresponded to transgene integration, which contained three RNAi expression cassettes; one for *CYP81A6* including a 3' untranslated region (UTR), the RNAi sequence, and the *CaMV35s* promoter; one for *G6 EPSPS* (*Pepc* terminator, *G6 EPSPS* gene, and *ubiquitin* promoter); and one for *hLF* (*Pepc* terminator, *hLF* gene, and rice *glutelin* promoter, *Gt1*) in a tandem arrangement (**Figure 3**). Based on the copy number evaluation and the single transgene integration site in Chr 03, we concluded that the transgenic T-DNA inserted into rice Chr 03 locus 16,439,674 included two copies of complete T-DNAs. Furthermore, the flanking sequences adjacent to the integration site showed that both the 5' and 3' ends of rice genome DNAs were connected to the 3' UTR of the RNAi expression cassette, indicating that the two T-DNAs were arranged head-to-head. The conventional PCR results also confirmed that these two copies of entire T-DNA sequences were inserted in a head-to-head rearrangement. A schematic diagram of the whole DNA insertion is shown in **Figure 3**.

3.4. Unintended integration and plasmid backbone residual insertion

Unintended DNA integration and plasmid backbone residual insertions often occur during transgene transformations, due to the random nature of exogenous DNA integration, especially when using particle bombardment transformation. By mapping the cleaned raw NGS data for both GM and WT rice lines to the transformation plasmid sequence, we found no unintended integration or insertion of plasmid backbone residual sequences in G281, other than the T-DNA region (**Figure 4**). We also performed conventional PCR analysis targeting the plasmid backbone sequences and did not obtain any positive amplicons from G281 DNA (data not shown), further supporting the conclusion that there were no residual transformation plasmid backbone sequences in G281.

4. Discussion

Clear molecular characterization of GM events is required to approve GMOs and GM products for commercialization, including new transgenic crops produced using DNA recombinant techniques such as *cisgenesis* and *intrasgenesis* (Turnbull et al., 2021). For a comprehensive understanding of transgene integration, NGS has been often accepted and employed to decipher the transgene integration of GMOs. Advantages of NGS include higher efficiency, accuracy, and reliability than with traditional sequencing methods (Arulandhu et al., 2018). However, most published research has focused on transgene integration in conventional transgenic crops/animals, and few studies have explored the possibility of using NGS approaches for molecular characterization of *cisgenesis* and *intrasgenesis*. In the present work, the G281 transgenic rice event was generated by transforming a plasmid containing three gene expression cassettes. In the plasmid, the 5' end of T-DNA is connected to the RNAi sequence of the rice *CYP81A6* gene, and its 3' end is connected to the rice Gt1 promoter. According to the definition of *intrasgenesis*, the G281 line can be regarded as having a structure similar to that acquired by *intrasgenesis*, although the exogenous G6 EPSPS and hLF genes are derived from other species (Lin et al., 2010). In the present work, we tried to develop a suitable pipeline to characterize transgene integration by *intrasgenesis*, employing G281 as an example. We performed successful molecular characterization of G281, including the transgene insertion site and its flanking sequences, copy number of transgenes, unintended insertions, and transformed plasmid backbone residues.

The DNAs of both ends of the T-DNA region are from the rice host genome in G281, making it difficult to identify the integration site. Using previously developed pipelines, numerous false-positive read pairs from the rice host genome are extracted, decreasing the efficiency and accuracy of screening and the identification of candidate reads. Herein, 322 read pairs were extracted as candidates from G281 data using the TranSeq pipeline. However, only 69 read pairs remained when we used the modified pipeline with an additional step restricting the positions $\pm$ insert sizes of homologous sequences. According to the filtering of read pairs, the transgene integration site was easily identified with high accuracy, whereas more than 75% of extracted reads were false-positives using the original TranSeq pipeline (Yang et al., 2013). These results indicate that the modified pipeline avoided the influence of host genome DNA, hence it could be employed for analysis of new *cisgenesis/intrasgenesis* lines.

To determine the sequence and structure of transgene integration, we combined the results of copy number evaluation and *de novo* analysis, revealing one entire transgene integration at rice Chr 03 involving two repeated T-DNA regions of transformed plasmid, and the two repeats are arranged in a head-to-head manner. In general, the whole transgene integration could be assembled based on reads mapped to the transformed plasmid. However, this is difficult when multiple copies of T-DNAs are inserted. Since the reads from NGS data are very short, identifying DNAs with multiple repeats is challenging (Park et al., 2017). In G281, the whole transgene integration was not

obtained directly from *de novo* analysis. We only assembled one repeat of the T-DNA region with a length of 9145 bp. Based on the sequences of flanking sequences and data from PCR amplification of the links between two repeats, we confirmed the whole structure and sequence of the transgene integration. The result of copy number evaluation corrected the previous reported result of only one single copy of transgene insertion in G281 line from Southern blotting analysis (Lin et al., 2010). Some recent reports have used third-generation sequencing (TGS) approaches to reveal whole transgene integration (Nicholls et al., 2019; Peng et al., 2021; Chen et al., 2021; Giraldo et al., 2021). However, analyzing whole transgene integration involving complex structures and/or rearrangements remains difficult using TGS approaches because of the absence of effective data analysis pipelines and quite high costs.

Molecular characterization of the G281 line using the developed pipeline showed that the approach is suitable for investigating *intragenesis* lines, even though *intragenesis* is more complex than the G281 event. When exploring *intragenesis* lines, control lines are more important and should be sequenced at the same time. By comparing NGS data between *intragenesis* and control lines, false-positive reads can be filtered out using dedicated pipelines. For example, 262 read pairs extracted for the WT line were identified as potential candidates for determining the transgene integration site. Obviously, there were no transgenes transferred in the WT line, hence these 266 read pairs must be derived from homogenous sequences between the host genome and the transformed plasmid, sequencing bias, or sequencing data contamination. According to the reads extracted from the WT line, we also found reads that were similar to those in G281, hence the similar reads were false-positives. Therefore, control lines are essential for analysis of *intragenesis* lines.

In summary, we modified our previously established TranSeq bioinformatics pipeline, and performed full molecular characterization of GM rice G281 using a PE-WGS approach with the modified pipeline. We identified only a single integration site in Chr 03 with two head-to-head repeats of complete T-DNAs inserted in G281. No residual transformation plasmid backbone sequences or unintended integration of exogenous DNA was detected in G281. The results confirm that the PE-WGS and modified TranSeq pipeline approach can be successfully applied to characterize transgene integration in *intragenesis* lines.

Declarations

Acknowledgements

This work was supported by the National Transgenic Plant Special Fund (Grant No. 2016ZX08012-002), and.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- ISAAA. (2019). Global Status of Commercialized Biotech/GM Crops in 2019: Biotech Crops Drive Socio-Economic Development and Sustainable Environment in the New Frontier. /SAAA Brief No. 55. ISAAA: Ithaca, NY.
- Sparrow, P. A. C. (2010). GM risk assessment. *Mol. Biotechnol.* 44, 267-75. doi: 10.1007/s12033-009-9237-x
- Miraglia, M., Berdal, K.G., Brera, C., Corbisier, P., Holst-Jensen, A., Kok, E.J., Marvin, H.J., Schimmel, H., Rentsch, J., Van Rie, J.P., Zagon, J.(2004) Detection and traceability of genetically modified organisms in the food production chain.

Food Chem. Toxicol. 42, 1157-80. doi: 10.1016/B978-0-85709-068-3.50012-3

Codex Alimentarius. (2003). Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants. CAC/GL. 45, 1–18.

Yang, L.T., Wang, C.M., Holst-Jensen, A., Morisset, D., Lin, Y.J., Zhang, D.B. (2013). Characterization of GM events by insert knowledge adapted re-sequencing approaches. Sci. Rep. 3–2839. doi: 10.1038/srep02839

Kuiper, H. A., Kleter, G. A., Noteborn, H. P. J. M., Kok, E. J. (2001). Assessment of the food safety issues related to genetically modified foods. Plant J. 27, 503-528. doi: 10.1046/j.1365-313X.2001.01119.x

Guttikonda, S. K., Marri, P., Mammadov, J., Ye, L., Soe, K., Richey, K., Cruse, J., Zhuang, M. Z., Gao, Z., Evans, C., Rounsley, S., Kumpatla, P. (2016). Molecular Characterization of Transgenic Events Using Next Generation Sequencing Approach. PLoS One. 11. doi: 10.1371/journal.pone.0149515. eCollection 2016.

Kamle, S., Ali, S. (2013). Genetically modified crops: detection strategies and biosafety issues. Gene. 522, 123-32. doi: 10.1016/j.gene.2013.03.107. Epub 2013 Apr 6

Li, R., Quan, S., Yan, X. F., Biswas, S., zhang, D.B., Shi, J. X. (2017). Molecular characterization of genetically-modified crops: challenges and strategies. Biotechnol Adv. 35, 302-309. doi: 10.1016/j.biotechadv.2017.01.005. Epub 2017 Jan 26.

Zhang, R., Yin, Y. L., Zhang, Y. J., Li, K., Zhu, H. X., Gong, Q., Wang, J. W., et.al. (2012). Molecular characterization of transgene integration by next-generation sequencing in transgenic cattle. PLoS One. 7. doi: 10.1371/journal.pone.0050348. Epub 2012 Nov 21.

Guo, B.F., Guo, Y., Hong, H. L., Qiu, L. J. (2016). Identification of Genomic Insertion and Flanking Sequence of G2-EPSPS and GAT Transgenes in Soybean Using Whole Genome Sequencing Method. Front Plant Sci. 7, 1009. doi: 10.3389/fpls.2016.01009. eCollection 2016.

Doori, P., Park, S., Ban, Y. K., Kim, Y.S., Park, K., Kim, N., et al. (2017). A bioinformatics approach for identifying transgene insertion sites using whole genome sequencing data. BMC Biotechnol. 15,17-67. doi: 10.1186/s12896-017-0386-x.

Zhang, Y. C., Zhang, H. W., Qu, Z., Zhang, X. J., Cui, J. J., Wang, C.H., Yang, L.T. (2020). Comprehensive analysis of the molecular characterization of gm rice g6h1 using a paired-end sequencing approach. Food Chem. 309, 125760. doi: 10.1016/j.foodchem.2019.125760. Epub 2019 Oct 23.

Zastrow-Hayes, G. M., Lin, H., Sigmund, A. L., Hoffman, J.L., Alarcon, C.M., Hayes, K.R., et al. (2015). Southern-by-Sequencing: A Robust Screening Approach for Molecular Characterization of Genetically Modified Crops. Plant Genome. 8. doi: 10.3835/plantgenome2014.08.0037.

Vanblaere, T., Flachowsky, H., Gessler, C., Broggini, G. A. L. (2014). Molecular characterization of cisgenic lines of apple 'Gala' carrying the Rvi6 scab resistance gene. Plant Biotechnol J. 12, 2-9. doi: 10.1111/pbi.12110. Epub 2013 Sep 3.

Holme, I.B., Wendt, T., Holm, P. B. (2013). Intragenesis and cisgenesis as alternatives to transgenic crop development. Plant Biotechnol J. 11, 395-407. doi: 10.1111/pbi.12055. Epub 2013 Feb 20.

Schouten, H. J., Jacobsen, E. (2008). Cisgenesis and intragenesis, sisters in innovative plant breeding. *Trends Plant Sci.* 13, 260-1. doi: 10.1016/j.tplants.2008.04.005. Epub 2008 May 15

European Food Safety Authority (EFSA), Paraskevopoulos, K., Federici, S. (2021). Overview of EFSA and European national authorities' scientific opinions on the risk assessment of plants developed through New Genomic Techniques. *EFSA J.* 19. doi: 10.2903/j.efsa.2021.6314

Eckerstorfer, M. F., Dolezel, M., Heissenberger, A., Miklau, M., Reichenbecher, W., Steinbrecher, R.A., Waßmann, F. (2019). An EU Perspective on Biosafety Considerations for Plants Developed by Genome Editing and Other New Genetic Modification Techniques (nGMs). *Front Bioeng Biotechnol.* 7, 31. doi: 10.3389/fbioe.2019.00031. eCollection 2019.

Lin, C. Y., Nie, P., Lu, W., Zhang, Q., Li, J., Shen, Z.C. (2010). A selectively terminable transgenic rice line expressing human lactoferrin. *Protein Expr Purif.* 74, 60-4. doi: 10.1016/j.pep.2010.04.015. Epub 2010 Apr 28.

Chen, Y., Ye, W., Zhang, Y., Xu, Y. (2015). High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Res.* 43, 7762-8. doi: 10.1093/nar/gkv784. Epub 2015 Aug 6.

Li, H., Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 25, 1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 19, 455-477. doi: 10.1089/cmb.2012.0021. Epub 2012 Apr 16.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 25, 2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol.* 29, 24–6. doi: 10.1038/nbt.1754.

Yang, Y., Li, L., Yang, H., Li, X., Zhang, X. J., Xu, J. F., Zhang D. B., et al. (2018). Development of Certified Matrix-Based Reference Material as a Calibrator for Genetically Modified Rice G6H1 Analysis. *J Agric Food Chem.* 66, 3708-3715. doi: 10.1021/acs.jafc.8b00468. Epub 2018 Apr 3.

Turnbull, C., Lillmo, M., Hvoslef-Eide, T. A. K. (2021). Global Regulation of Genetically Modified Crops Amid the Gene Edited Crop Boom - A Review. *Front Plant Sci.* 12, 630396. doi: 10.3389/fpls.2021.630396.

Arulandhu, A. J., Jeroen, V. D., Martijn, S., Rico, H., Marleen, V., Bonnie, M., Li, R., Yang, L., Shi, J., (2018). NGS-based amplicon sequencing approach; towards a new era in GMO screening and detection. *Food Control.* 93, 201-210. doi: 10.1016/j.foodcont.2018.06.014

Nicholls, P. K., Bellott, D. W., Cho, T. J., Pyntikova, T., Page, D.C. (2019). Locating and Characterizing a Transgene Integration Site by Nanopore Sequencing.

G3 (Bethesda). 9, 1481-1486. doi: 10.1534/g3.119.300582.

Peng, C., Mei, Y. T., Ding, L., Wang, X. F., Chen, X. Y., Wang, J. M., et al. (2021). Using Combined Methods of Genetic Mapping and Nanopore-Based Sequencing Technology to Analyze the Insertion Positions of G10evo-EPSPS and Cry1Ab/Cry2Aj Transgenes in Maize. *Front Plant Sci.* 12, 690951. doi: 10.3389/fpls.2021.690951. eCollection 2021

Chen, X. H., Dong, Y., Huang, Y. L., Fan, J., Yang, M. S., Zhang, J. (2021). Whole-genome resequencing using next-generation and Nanopore sequencing for molecular characterization of T-DNA integration in transgenic poplar 741. *BMC Genomics.* 22, 329. doi: 10.1186/s12864-021-07625-y

Giraldo, P. A., Shinozuka, H., Spangenberg, G. C., Smith, K. F., Cogan, N. O. I. (2021). Rapid and Detailed Characterization of Transgene Insertion Sites in Genetically Modified Plants via Nanopore Sequencing. *Front Plant Sci.* 11, 602313. doi: 10.3389/fpls.2020.602313. eCollection 2020.

Tables

Table 1. Sequencing data estimates and mapping results analysis

Sample	Total trimmed read pairs	Sequencing depth (D)	Read pairs mapped to host genome	Coverage calibrator (R)	Read pairs mapped to plasmid	Reads clustered around insertion regions	
						Not filtered	Filtered
Xiushui	54819968	29.28	54233374	0.989	332	262	5
G281(GM)	54130964	28.91	53545140	0.989	2604	322	69

Table 2. Estimation of the copy number of transgenes in rice line G281 using resequencing data and ddPCR

Gene	PE-WGS				ddPCR assay					
	ADT	D	R	Copy number	Quantified copies				RSD (%)	Copy number
					1	2	3	Mean		
HLF	48.93	28.91	0.989	1.95	37160	37760	37150	37356.7	0.94	1.96
EPSPS	53.76	28.91	0.989	2.11	35680	35800	35320	35600	0.70	1.87
GT1	77.75	28.91	0.989	3.10	/	/	/	/	/	/
sps	25.11	28.91	0.989	/	18980	19000	19100	19026.7	0.34	/

“/” means no detected.

Figures

Figure 1

Modified bioinformatics pipeline for molecular characterization of transgenic plants using the PE-WGS approach.

Figure 2

The transgene integration site and its flanking sequence from PE-WGS and PCR analyses. (A) IGV view of 69 candidate read pairs crossing the insertion site of chromosome 3. The red line indicates the transgene insertion site in the rice genome. (B) Location of the transgene integration site according to the assembled contigs from candidate read pairs. (C) Conventional PCR amplification results for the flanking sequence. Lane M, DL2000 markers; Lane GM, G281; Lane WT, Xiushui 110; Lane NTC, no template control.

Figure 3

Schematic diagram of whole transgene integration in G281 rice. RB, right border of T-DNA; LB, left border of T-DNA; Gt1, rice glutelin promoter; hLF, human lactoferrin; G6 EPSPS, the 5-enolpyruvylshikimate-3-phosphate synthase isolated from *Pseudomonas putida* fused with a corn chloroplast transit peptide at the N-terminus (gb: EU169459); RNAi, reverse repeat sequence for RNA interference against CYP81A6; CaMV35S, cauliflower mosaic virus 35S promoter; ZmUbi, polyubiquitin-1 promoter of *Zea mays*; PEPC, phosphoenolpyruvate carboxylase terminator of *Zea mays*.

Figure 4

Read pairs mapped to the entire transformed plasmid from WT and GM paired-end sequencing data.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementarySequenceFile2.docx](#)
- [SupplementarySequenceFile1.docx](#)
- [nrreportingsummary01D.pdf](#)