Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Correspondence:

Genome-wide profiling of genetic variation in *Agrobacterium*-transformed rice plants*#

Wen-xu LI[§1,4], San-ling WU[§2], Yan-hua LIU[1], Gu-lei JIN[5], Hai-jun ZHAO[1], Long-jiang FAN[3], Qing-yao SHU[†‡1]

([1]*State Key Laboratory of Rice Biology, Institute of Crop Sciences, Zhejiang University, Hangzhou 310058, China*)
([2]*Analysis Center of Agrobiology and Environmental Sciences, Faculty of Agriculture, Life and Environment Sciences, Zhejiang University, Hangzhou 310058, China*)
([3]*IBM Biocomputational Laboratory, Institute of Crop Sciences, Zhejiang University, Hangzhou 310058, China*)
([4]*Institute for Wheat Research, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China*)
([5]*Hangzhou Guhe Information and Technology Co., Ltd., Hangzhou 310058, China*)
[†]E-mail: qyshu@zju.edu.cn

http://dx.doi.org/10.1631/jzus.B1600301

Agrobacterium-mediated transformation has been widely used in producing transgenic plants, and was recently used to generate "transgene-clean" targeted genomic modifications coupled with the clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas9) system. Although tremendous variation in morphological and agronomic traits, such as plant height, seed fertility, and grain size, was observed in transgenic plants, the underlying mechanisms are not yet well understood, and the types and frequency of genetic variation in transformed plants have not been fully disclosed. To reveal the genome-wide variation in transformed plants, we sequenced the genomes of five independent T_0 rice plants using next-generation sequencing (NGS) techniques. Bioinformatics analyses followed by experimental validation revealed the following: (1) in addition to transfer-DNA (T-DNA) insertions, three transformed plants carried heritable plasmid backbone DNA of variable sizes (855–5216 bp) and in different configurations with the T-DNA insertions (linked or apart); (2) each transgenic plant contained an estimated 338–1774 independent genetic variations (single nucleotide variations (SNVs) or small insertion/deletions); and (3) 2–6 new *Tos17* insertions were detected in each transformed plant, but no other transposable elements or bacterial genomic DNA.

During the past three decades, *Agrobacterium*-mediated transformations have been widely applied to a broad range of species, resulting in a number of transgenic (or genetically modified (GM)) plants being commercialized globally (http://www.isaaa.org). The importance of this technology may further increase with the widespread application of the CRISPR/Cas9 system to produce targeted genomic modifications of agricultural plants (Ledford, 2015). For assessment of both their safety and practical usefulness, a genome-wide analysis of genetic changes in transformed plants would provide the most comprehensive and fundamental information. NGS technologies, because of their high throughput and cost effectiveness, are appropriate choices for holistically examining genetic variation in transgenic plants. NGS-based methods have been used for characterizing complex cases of T-DNA insertions in soybean (Kovalic *et al*., 2012) and rice (Yang *et al*., 2013), and for examining other types of genetic variation in transgenic rice plants (Kawakatsu *et al*., 2013; Endo *et al*., 2014; Wei *et al*., 2016). Despite the vast amount of transgenic plants

‡ Corresponding author
§ The two authors contributed equally to this work
* Project supported by the Ministry of Science and Technology, China (No. SQ2015IM3600010) and the Fundamental Research Funds for the Central Universities (No. 2016XZZX001-09), China
# Electronic supplementary materials: The online version of this article (http://dx.doi.org/10.1631/jzus.B1600301) contains supplementary materials, which are available to authorized users
ORCID: Qing-yao SHU, http://orcid.org/0000-0002-9201-0593

produced and applied worldwide, the number of such studies and the representation of transgenic materials, are limited, our understanding of genome-wide variation in transformed plants is far from complete. Hence, more studies involving representative transgenic plants are needed.

In the present study, we produced transgenic rice plants via *Agrobacterium*-mediated transformation and used NGS to sequence five independent T_0 plants to assess the frequencies, spectra, and nature of variations introduced into the rice genome. A total of about 11.5 Gb of clean data were generated for each of the five transgenic lines (T027, T182, T204, T273, and T400), of which over 98% mapped to the reference genome (Table S1).

A bioinformatics analysis of NGS data showed that there was one T-DNA sequence in each transgenic line (Fig. 1). The left and right borders of the T-DNA insertions were complete and clear in T027, T204, and T400, but the right borders of T273 and T182 had short backbone DNA segments (Fig. 1). Furthermore, we detected separate backbone DNA insertions in three transgenic lines, i.e. T027, T182, and T204 (Fig. 1). In T027, the backbone (3.5 kb) and T-DNA fragments were inserted into two different chromosomes (chromosomes 1 and 7), while, in T182, both the backbone (5.2 kb) and T-DNA fragments were integrated into chromosome 1, but separated by a about 3-kb rice genomic fragment (Fig. 1). While only one end of the T-DNA insertion was precisely located on chromosome 1 in T182, neither end of the small backbone DNA fragment (855 bp) in T204 could be located to any chromosome (Fig. 1, Fig. S1, Table S2).

To assess whether the transgenic plant production process generated other genetic variations, SNVs and small indels were profiled genome-wide. A total of 8204 SNVs and small indels were identified in the five transgenic lines compared with the reference genome. Further examination indicated that 2593 of these were common across the five transgenic lines. The remaining 5611 genomic variations (4566 SNVs and 1045 indels) were considered to be derived from transformation and in vitro culture. Among the five transgenic lines, four lines with a short in vitro culture period had similar numbers of variations (502–766). Line T204, generated from a callus with a prolonged in vitro culture period, had many more variations than the other lines, particular for the number of SNVs (Table 1). A genomic position analysis indicated that

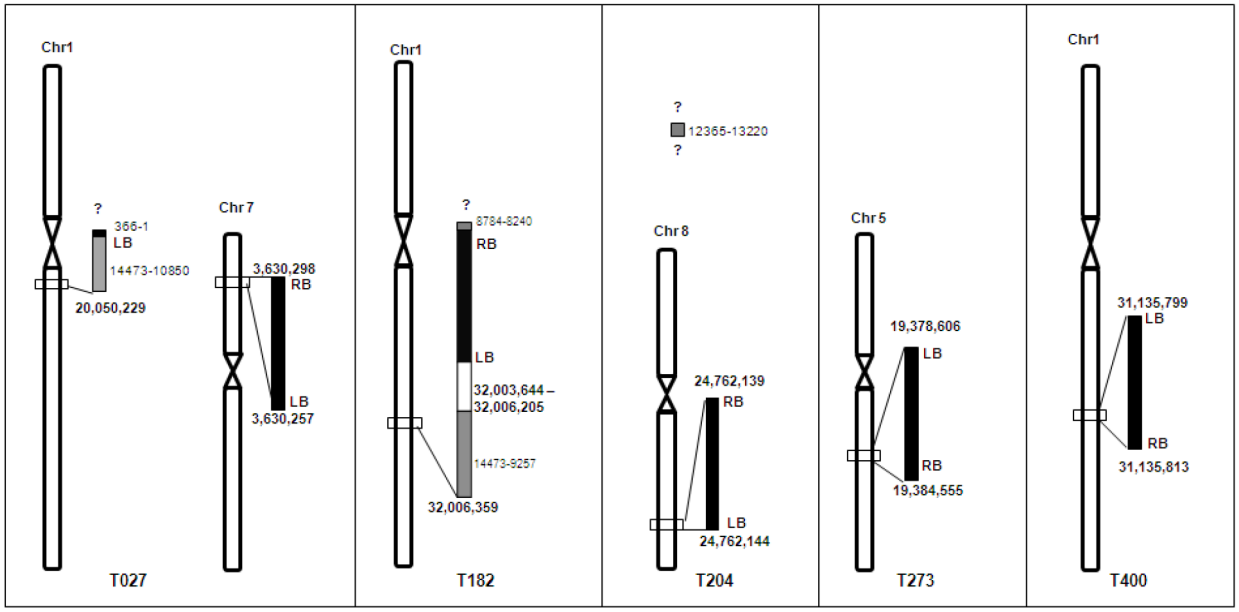

Fig. 1 Schematics of plasmid DNA insertions in different chromosomes detected by NGS followed by bioinformatics analysis of five transgenic rice plants

Rice genomic DNA is shown in empty boxes, and plasmid transfer and backbone DNA in black and gray filled boxes, respectively. The bold numbers around the insertions refer to the nucleotide number in respective rice chromosomes, and non-bold numbers refer to nucleotides in the expression vector. The insertions with defined positions in rice chromosomes are marked with lines and nucleotide numbers, while those that could not be located in rice chromosomes are marked with a question mark (?)

on average 26.36% of indels and 36.14% of SNVs were located in the coding regions (Table 1). Validation of randomly selected variations (n=200) by Sanger sequencing confirmed 57% of them as true mutations. Hence, the actual number of mutations in the transgenic plants was less than 5611.

The number of transposable elements (TEs) in the transgenic plants was also determined. Compared with Nipponbare, 18 new *Tos17* sequences, from 2 to 6 in each line, were identified across nine chromosomes (Fig. 2, Table S3). The new *Tos17* insertions were validated in all lines except T400, where the amplification of both ends was observed for only two of its five events, while one-ended amplification was observed for the other three insertions (Fig. S2). Interestingly, many of the *Tos17* insertions were located in the vicinity of chromosome ends (Fig. 2, Table S4).

Table 1 Number of unique genomic variations in five transgenic rice lines

Transgenic line	Type	Total number	Validation[1]	Coding region
T027	Indel	183	13/20	41
	SNV	319	14/20	119
T182	Indel	204	15/20	63
	SNV	309	13/20	124
T204	Indel	165	11/20	51
	SNV	3060	8/20	1085
T273	Indel	217	13/20	53
	SNV	388	12/20	139
T400	Indel	301	11/20	74
	SNV	465	14/20	147
Total		5611	114/200	1896

[1] The validation was performed by the Sanger sequencing of fragments encompassing SNV/indel sites (20 SNV and indel mutations were tested from each line)

Other TEs, including mPing, RN_363, and Tami2, were also examined. While a few differences were common to all the lines, none was unique to any single transgenic plant. For example, among over 40 mPings identified, three were present in all five transgenic lines, in chromosomes 1 (positions 40, 490, 011 bp), 3 (26, 791, 701 bp), and 12 (3, 328, 968 bp), respectively, but were not reported in the reference genome of Nipponbare. Conversely, one mPing (chromosome 1: 17, 514, 263 bp) reported in the reference genome of Nipponbare was not identified in any of the transgenic lines.

In addition to SNVs and small indels, 793 structural variations (SVs), such as deletions of chromosomal fragments, inversions, interchromosomal exchanges, and complex variations, were identified in the five transgenic lines through bioinformatic analysis. About one third (267/793) of the SVs were common to the five transgenic lines. Therefore, there were differences between the "Nipponbare" used for genetic transformation and the "Nipponbare" used as a reference. The numbers and profiles of SVs were quite similar among the five transgenic lines (Fig. S3).

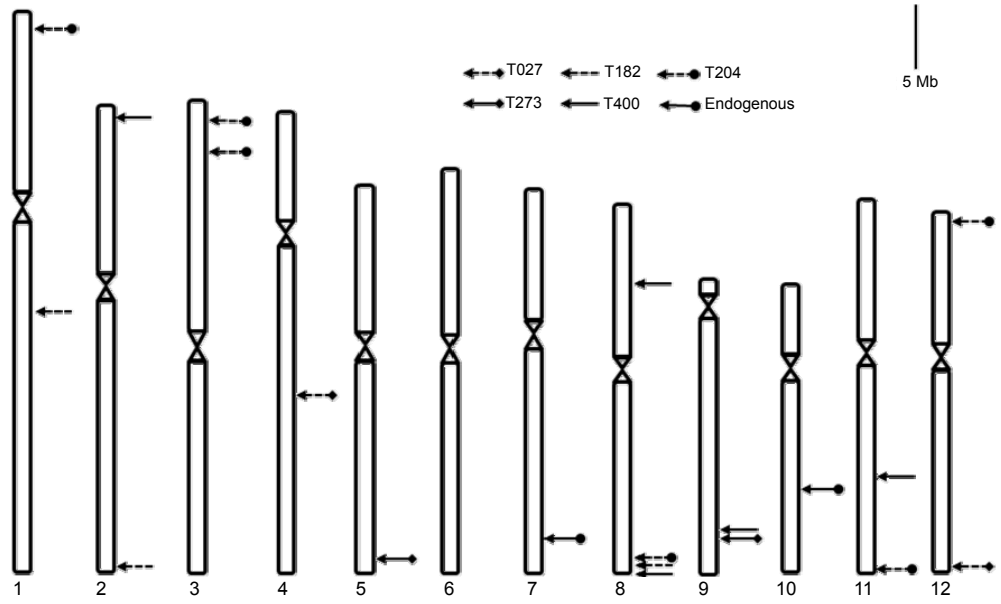

Fig. 2 Distribution of endogenous and newly inserted *Tos17* sequences on rice chromosomes

However, we were not able to validate any of the 10 randomly selected SVs in T182 by polymerase chain reaction (PCR) amplification (Table S5). No *Agrobacterium* genomic sequence was identified in the transgenic plants.

The present study revealed the complicated plasmid DNA integration scenarios and the types and frequency of genome-wide variations in transformed rice plants, greatly enriching our knowledge of the genetics and genomics of *Agrobacterium*-transformed plants.

Materials and methods

Production of transgenic rice

Transgenic rice plants were developed via *Agrobacterium*-mediated transformation of the *japonica* cultivar "Nipponbare" following the method of Hiei and Komari (2008). Rice calli induced from mature seeds were co-cultured with *A. tumefaciens* strain EHA105 containing plasmid p1301-amiMRP5-OleN (Fig. S4). Because the in vitro culture process is known to cause somaclonal variation (Sabot *et al.*, 2011), we analyzed one transgenic T_0 line (T204) having an extended in vitro culture period (about 5 months) and four T_0 lines (T027, T182, T273, and T400) with a short culture period (about 3 months) (Fig. S5).

Genome sequencing and bioinformatics analysis

Genomic DNA was extracted from the leaf tissues of five independent transgene positive T_0 lines using the method of Sambrook and Russell (2001). For NGS, DNA sequencing libraries were constructed according to the manufacturer's instructions. Short paired-end (PE) reads (90 bp) were generated using the Illumina Hiseq2000 sequencing platform of BGI (Shenzhen, China). Raw reads were pre-processed to remove adaptors and reads of low quality (Li *et al*., 2010). All sequences generated by this study have been deposited in GenBank (BioProject ID: PRJNA290293).

For read mapping and SNV calling and detection of small insertion/deletions (indels <10 bp), clean reads of each line were aligned to the reference genome using Burrows-Wheeler Aligner (BWA) v0.5.9 and SAMtools v0.1.18 (Li *et al*., 2009). The SNVs and indels were further annotated with the Michigan State University (MSU) Rice Genome Annotation Project Database using SnpEff v2.0.5d, and structural variations were analyzed by Break Dancer v1.1 (Chen *et al*., 2009).

The identification and prediction of inserted DNAs (plasmid DNA, TE, and *Agrobacterium* DNA) were enabled using PE read pairs. BWA v0.5.9 was used for read mapping (Li *et al*., 2009), and a custom in-house Perl script was written for locating insertion sites. To detect the presence of *Agrobacterium* DNA sequences in transgenic rice, the *Agrobacterium* C58 genome sequence was used as a reference (Li *et al.*, 2009).

PCR and DNA Sanger sequencing

The presence of plasmid DNAs, T-DNA, and *Tos17* insertion sequences revealed by NGS was further examined by PCR using site-specific primers (Tables S1 and S2). For indel or SNV validation, fragments encompassing the variation sites were amplified using event-specific primers. The variation was determined by clone sequencing of the amplified fragments (five clones for each SNV/indel).

Compliance with ethics guidelines

Wen-xu LI, San-ling WU, Yan-hua LIU, Gu-lei JIN, Hai-jun ZHAO, Long-jiang FAN, and Qing-yao SHU declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Chen, K., Wallis, J.W., McLellan, M.D., *et al*., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**(9): 677-681. http://dx.doi.org/10.1038/nmeth.1363

Endo, M., Kumagai, M., Motoyama, R., *et al*., 2014. Whole-genome analysis of herbicide-tolerant mutant rice generated by *Agrobacterium*-mediated gene targeting. *Plant Cell Physiol*., **56**(1):116-125. http://dx.doi.org/10.1093/pcp/pcu153

Hiei, Y., Komari, T., 2008. *Agrobacterium*-mediated transformation of rice using immature embryos or calli induced from mature seed. *Nat. Protoc.*, **3**(5):824-834. http://dx.doi.org/10.1038/nprot.2008.46

Kawakatsu, T., Kawahara, Y., Itoh, T., *et al*., 2013. A whole-genome analysis of a transgenic rice seed-based edible vaccine against cedar pollen allergy. *DNA Res*., **20**(6):623-631. http://dx.doi.org/10.1093/dnares/dst036

Kovalic, D., Garnaat, C., Guo, L., *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 Gen.*, **5**:149-163. http://dx.doi.org/10.3835/plantgenome2012.10.0026

Ledford, H., 2015. CRISPR, the disruptor. *Nature*, **522**(7554): 20-24. http://dx.doi.org/10.1038/522020a

Li, H., Handsaker, B., Wysoker, A., *et al.*, 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16):2078-2079. http://dx.doi.org/10.1093/bioinformatics/btp352

Li, R., Zhu, H., Ruan, J., *et al*., 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*., **20**(2):265-272. http://dx.doi.org/10.1101/gr.097261.109

Sabot, F., Picault, N., El-Baidouri, M., *et al*., 2011. Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. *Plant J*., **66**(2):241-246. http://dx.doi.org/10.1111/j.1365-313X.2011.04492.x

Sambrook, J., Russell, D.W., 2001. Molecular Cloning: A Laboratory Manual, 3rd Ed. Cold Spring Harbor Laboratory Press, NY. http://dx.doi.org/10.1086/394015

Wei, F.J., Kuang, L.Y., Oung, H.M., 2016. Somaclonal variation does not preclude the use of rice transformants for genetic screening. *Plant J.*, **85**(5):648-659. http://dx.doi.org/10.1111/tpj.13132

Yang, L., Wang, C., Holst-Jensen, A., 2013. Characterization of GM events by insert knowledge adapted re-sequencing approaches. *Sci. Rep.*, **3**:2839. http://dx.doi.org/10.1038/srep02839

List of electronic supplementary materials

Fig. S1 PCR validation of plasmid DNA in transgenic rice lines

Fig. S2 PCR validation of newly inserted *Tos17* with site specific primers

Fig. S3 Chromosome structural variations detected uniquely in transgenic lines

Fig. S4 Schematic diagram of artificial microRNA of *OsMRP5* expression plasmid p1301-amiMRP5-OleN

Fig. S5 Procedure used to produce transgenic plants from seed derived calli, and the particulars of 5 individual transgenic T_0 lines used for genome sequencing

Table S1 Summary of the genome sequencing of five transgenic rice lines

Table S2 Primers for validation of plasmid transfer and backbone DNA insertions

Table S3 Primers used for validation of newly inserted *Tos17* sequences

Table S4 Information about new *Tos17* insertions in transgenic rice

Table S5 Primers for PCR validation of structural variation in transgenic line T182

中文概要

题　目：农杆菌介导的水稻转化植株遗传变异特征的全基因组分析

目　的：揭示转基因水稻全基因组遗传变异的特征与频率。

创新点：通过单核苷酸分辨率揭示了农杆菌介导法转化水稻植株全基因组水平遗传变异的类型和频率以及外源 DNA 的整合模式。

方　法：应用 Illumina Hiseq2000 高通量测序技术测定了 5 个 T_0 代转基因水稻株系的基因组序列。结合生物信息学分析和聚合酶链反应（PCR）扩增，以及 Sanger 测序，我们检测和验证单核苷酸多态性（SNP）和 Indel 变化类型和数量，转移 DNA（T-DNA）及其载体骨架序列和转座子整合位点及特征，大片段的结构变异等遗传变异。

结　论：结果表明，农杆菌介导的水稻遗传转化，除 T-DNA 整合到水稻基因组外，还存在载体骨架整合现象；每个转基因水稻基因组的变异（SNP 和小片段的缺失插入）数目为 338–1774，与报道的组培过程中产生的变异类似；转基因水稻基因组仅存在 *Tos17* 转座子数量的变化，未检测到其他转座子数目和位置的变化。

关键词：水稻；遗传转化；全基因组；转移 DNA；转座子