Whole-Genome Resequencing using Next-Generation and Nanopore Sequencing for Molecular Characterization of T-DNA Integration in Transgenic Poplar 741

Xinghao Chen
Hebei Agricultural University https://orcid.org/0000-0002-1591-0424
Yan Dong
Hebei Agricultural University
Yali Huang
Hebei Agricultural University
Jianmin Fan
Hebei Agricultural University
Minsheng Yang (yangms100@126.com)
Hebei Agricultural University https://orcid.org/0000-0003-2488-267X
Jun Zhang
Hebei Agricultural University

Research article

Keywords: Transgenic safety assessment, Poplar, T-DNA, Integration site, Copy number

Posted Date: November 11th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-103623/v1

Version of Record: A version of this preprint was published at BMC Genomics on May 6th, 2021. See the published version at https://doi.org/10.1186/s12864-021-07625-y.

Abstract

Background

With the rapid development of transgenic technology, transgenic plants have been planted all over the world, and transgenic forest trees have also been commercialized. At the same time, the potential threat of transgenic plants to human health and the natural environment has aroused widespread concern. Therefore, safety assessments before field release and commercial planting of transgenic plants are necessary. By determining the copy number and integration sites of T-DNA in transgenic plants, the safety of transgenic plants at the genomic level can be assessed.

Results

In this study, we performed genome resequencing of Pb29, a transgenic high-resistance poplar 741 line that has been commercialized, using next-generation and Nanopore sequencing. The results revealed that there are two T-DNA insertion sites, located at 9,283,905–9,283,937 bp on chromosome 3 (Chr03) and 10,868,777–10,868,803 bp on Chr10. The accuracy of the T-DNA insertion locations and directions was verified using polymerase chain reaction amplification. Through sequence alignment, different degrees of base deletions were detected on the T-DNA left and right border sequences, and in the flanking sequences of the insertion sites. An unknown fragment was inserted between the Chr03 insertion site and the right flanking sequence, but the Pb29 genome did not undergo chromosomal rearrangement. It is worth noting that we did not detect the *API* gene in the Pb29 genome, indicating that Pb29 is a transgenic line containing only the *BtCry1AC* gene. On Chr03, the insertion of T-DNA disrupted a gene encoding TAF12 protein, but the transcriptional abundance of this gene did not change significantly in the leaves of Pb29. Additionally, except for the gene located closest to the T-DNA integration site, the expression levels of four other neighboring genes did not change significantly in the leaves of Pb29.

Conclusions

This study provides important molecular information for safety assessments and management of transgenic poplar 741. Our findings also provide a theoretical basis for safety assessments of other transgenic poplar.

Background

Poplar is one of the most widely distributed tree species owing to its rapid growth and strong adaptability to environmental changes [1]. It is one of the important industrial timber species that is widely used in the paper-making industry and panel processing. However, with the continuous increase of poplar planting area, the ensuing insect attack has become more and more serious, which has brought huge losses to forestry production [2]. In order to reduce the economic losses caused by insect pests, decrease the need for chemical pesticides, and protect the ecological environment, the cultivation of insect-resistant transgenic varieties is particularly important [3]. Transgenic technology is used commercially for growing

trees in China, which was the first country to commercialize transgenic poplar. As early as 2011, the planted area of transgenic insect-resistant poplar in China reached 450 ha [4]; thus, transgenic trees substantially impact the natural environment. At the same time, the safety of transgenic plants has gradually become an issue of widespread concern. Therefore, transgenic plants should be strictly assessed before commercial planting, to protect the natural environment and human life and health.

The genomes of transgenic plants have been modified at the molecular level. Therefore, when assessing the safety of transgenic plants, we should first perform a molecular analysis of T-DNA integration, to elucidate the T-DNA copy number and integration sites. Several studies have shown that the expression level of a foreign gene is inversely related to its copy number at the T-DNA insertion site [5, 6]; the flanking sequences of the T-DNA insertion site can also affect the expression of the foreign gene [7]. Thus, T-DNA copy number and insertion site locations are important factors affecting the expression of the foreign gene. Therefore, clarifying the T-DNA copy number and insertion site locations can help in the assessment of the safety of transgenic plants at the genomic level, and is in fact one of the primary means of assessing the safety of these plants.

There are many methods for locating the insertion sites of foreign genes in transgenic plants, most of which are based on polymerase chain reaction (PCR) amplification; these include thermal asymmetric interlaced PCR [8], inverse PCR [9], and adapter-ligated PCR [10]. Although these methods have been successfully applied to transgenic plants of species such as *Arabidopsisthaliana* [11] and rape [12], they are prone to false-positives, and are also time-consuming, laborious, and poorly reproducible. In recent years, with the continuous development of sequencing technology, next-generation sequencing (NGS) has been widely used for genome sequencing because of its high throughput capability, low cost, and accurate results. NGS has been successfully used to locate T-DNA insertion sites in transgenic soybean [13], rice [14], and birch [15]. However, the NGS reads are too short to accurately locate all of the T-DNA insertion sites in transgenic plants with complex T-DNA integration patterns or genomes. By contrast, third-generation sequencing technology, developed by Oxford Nanopore Technologies and PacBio, can produce longer reads, which can overcome the limitations of NGS such as short reads and bias due to GC content, although the accuracy is relatively low. Therefore, by combining NGS with third-generation sequencing technology, we can accurately and efficiently analyze overall genomic changes due to T-DNA mutations.

Transgenic poplar 741, which was cultivated by Hebei Agricultural University and the Institute of Microbiology of the Chinese Academy of Sciences, has been certified safe according to national standards for transgenic animals and plants. Transgenic poplar 741 were planted commercially from 2002 to 2007 after environmental impact and production tests. However, no molecular analysis of T-DNA integration in transgenic poplar 741 has been performed. In this study, we performed whole-genome resequencing of transgenic poplar 741 using NGS and Nanopore sequencing, and analyzed the copy number and insertion sites of the T-DNA as well as the flanking sequences at the T-DNA integration site. Our results provide important molecular information for the safety assessment and management of transgenic poplar 741, and a theoretical basis for safety assessments of other transgenic poplar.

Results

Results of NGS analysis

After performing quality-control checks, a total of 52.3 million clean reads were obtained from the raw reads, corresponding to more than 30× coverage of the reference genome. More than 92% of the sequencing data had Phred-like quality scores ≥ 30, indicating that the data were high quality (Table 1). After sequence alignment, nine junction reads on chromosome 03 (Chr03), and four on Chr10, were identified in the Pb29 sequence, indicating that there are two T-DNA insertion sites in the Pb29 genome (Table S1). Based on the physical positions of the junction reads, one insertion site is located at 9,283,937 bp on Chr03, and the other at 10,868,777 bp on Chr10. T-DNA is inserted in the reverse direction on Chr03, and in the forward direction on Chr10. However, further analysis revealed that only unilateral junction reads could be detected at both T-DNA insertion sites; ideally, junction reads should be detected on both sides of each insertion site (Fig. 1).

Table 1 The summary of sequence data from NGS.

Clean reads	Clean bases (Gb)	GC (%)	Q20(%)	Q30(%)
52,313,447	15.7	37.42	97.36	92.77

Confirmation of insertion sites and directions using PCR amplification

To verify the accuracy of the T-DNA insertion sites and directions, we designed 6 primers based on the flanking sequences of the T-DNA insertion sites and the T-DNA sequence (Fig. 2), and amplified the genomic DNA of poplar 741 and Pb29 using different primer combinations (Table 2). The results of PCR amplification revealed that the PCR runs using primer combinations 3, 4, 6, and 7 generated products with a single band for Pb29 in Fig. 3A, whereas no products were amplified for poplar 741 in Fig. 3B. When primer combinations 1, 2, 8, and 9 were used in the PCR, amplified bands were not produced for Pb29 or poplar 741, indicating that T-DNA was indeed inserted into Chr03 in the reverse direction and into Chr10 in the forward direction, thus verifying the NGS results. Meanwhile, the target band was observed after PCR runs using primer combinations 5 and 10 for both Pb29 and poplar 741, indicating that Pb29 is a heterozygous mutant created via T-DNA insertion (Fig. 3).

Table 2 The primer combinations and product size for verifying the insertion sites and directions.

No.	Primer combination	Product size (bp)	No.	Primer combination	Product size (bp)
1	Chr3u-F1 & 131#S5F	552	6	Chr10u-F2 & 131#S5F	818
2	Chr3d-R2 & 131#S2F	767	7	Chr10d-R2 & 131#S2F	745
3	Chr3u-F1 & 131#S2F	671	8	Chr10u-F2 & 131#S2F	937
4	Chr3d-R2 & 131#S5F	648	9	Chr10d-R2 & 131#S5F	626
5	Chr3u-F1 & Chr3d-R2	684	10	Chr10u-F2 & Chr10d-R2	928

Results of Nanopore sequencing analysis

To further verify the NGS results and determine whether chromosomal rearrangement occurred in the Pb29 genome due to T-DNA insertion, we used the third-generation sequencing technology developed by Oxford Nanopore Technologies to resequence the whole genomes of poplar 741 and Pb29. More than 96% of the clean reads of both poplar 741 and Pb29 mapped to the reference genome, corresponding to 40× and 39× coverage of the reference genome, respectively (Table 3). The depth of coverage was evenly distributed across both poplar 741 and Pb29 chromosomes, indicating that the genomic DNA of poplar 741 and Pb29 was sequenced in a random manner (Fig. 4).

The BAM file generated by comparing all junction reads with the *P. trichocarpa* reference genome was imported into (Integrative Genomics Viewer) IGV software for visual analysis. All junction reads only mapped to Chr03 or Chr10, and there was a gap between reads on both chromosomes. The two gaps, each formed by a T-DNA insertion that disrupted part of the genome sequence, matched the two T-DNA insertion sites in the Pb29 genome exactly. The two T-DNA insertion sites in the Pb29 genome are located at 9,283,905–9,283,937 bp on Chr03 and 10,868,777–10,868,803 bp on Chr10, consistent with the detection results obtained using NGS (Fig. 5).

Compared with the *P. trichocarpa* reference genome, evidence of many (Structural variation) SV events was seen in the genomes of both poplar 741 and Pb29, most of which were deletions or insertions of chromosome segments (Fig. 6). After removing the regions representing SV events of the same type at the same positions in the poplar 741 and Pb29 genomes, SV events > 1 kb are regarded as chromosomal rearrangements in the Pb29 genome caused by T-DNA insertion. However, we did not detect this type of event, indicating that the insertion of T-DNA did not cause large chromosomal rearrangements in the Pb29 genome.

Table 3 The summary of sequence data from Nanopore sequencing.

Sequence Data	741	Pb29	Sequence Data	741	Pb29
Clean reads	2,351,233	2,194,474	Clean bases(Gb)	20.6	19.9
N50Len	10,146	10,474	N90Len	6,147	6,414
MeanLen	8,778	9,072	unmapped	88,268	83,507
mapped	2,262,965	2,110,967	Mapped ratio(%)	96.25	96.19
Average depth	40	39	Coverage_ratio_1X(%)	84.99	84.8
Coverage_ratio_5X(%)	75.23	74.92	Coverage_ratio_10X(%)	69.49	69.12

T-DNA and flanking sequence analysis

Because Nanopore sequencing can be used to obtain longer reads, some junction reads contained complete T-DNA sequences. The complete T-DNA sequences at the two insertion sites were extracted and compared with the vector sequence. The results showed that the left and right border sequences of the T-DNA inserted on Chr03 were missing 26 and 3 bp, respectively, whereas the left and right border sequences of the T-DNA inserted on Chr10 were missing 35 and 34 bp, respectively (Fig. 7). It is worth noting that the 35S-API-Nos expression component was not detected in the T-DNA sequences at either insertion site; furthermore, both T-DNA sequences are exactly the same, indicating that the expression component of the *API* gene was not lost during the transformation process. Rather, it was not present in the expression vector in Agrobacterium before transformation (Fig. 8).

We compared isolated flanking sequences with the *P. trichocarpa* reference genome and found that fragments had been deleted from the flanking sequences at both insertion sites, as T-DNA insertion damaged the genome sequence at those sites (box with red outline in Fig. 9). The genome sequence at the T-DNA insertion sites on Chr03 and Chr10 was missing 33 and 27 bp, respectively, consistent with the results of the alignment analysis (Fig. 5). A short fragment (24 bp in length) was found between the T-DNA insertion site and the right flanking sequence on Chr03 in the Pb29 genome; this fragment could not be mapped to the *P. trichocarpa* reference genome. We analyzed the clean reads from poplar 741 found that reads mapped to the same positions essentially had the same sequences as the corresponding sections of the *P. trichocarpa* genome (Fig. 10), indicating that the 24-bp fragment did not arise from the difference between genomes but was instead caused by the insertion of an unknown fragment during the T-DNA integration process.

Analysis of the expression levels of genes located near the insertion sites

The genes within 20 kb upstream and downstream of the two T-DNA insertion sites were detected based on the genome annotation file of *P. trichocarpa*. The results showed that T-DNA was inserted 9,466 bp downstream of the LOC112326972 gene and 8,137 bp upstream of the LOC7475699 gene on Chr03, and 15,621 bp downstream of the LOC7498060 gene and 1,543 and 11,914 bp upstream of the LOC7498061 and LOC7498062 genes, respectively, on Chr10 (Table 4). (Fragments Per Kilobase Million) FPKM values

associated with the transcriptome data were used to compare the expression levels of the five neighboring genes. The results showed that except for the LOC7498061 gene, the expression levels of the other four genes in Pb29 leaves did not change significantly, indicating that the insertion of T-DNA did not significantly affect the expression levels of these four genes. The LOC7498061 gene is located closest to the T-DNA insertion site; its expression level was significantly upregulated in Pb29 leaves, indicating that the insertion of T-DNA in Pb29 affects gene expression within a certain range (Fig. 11).

Table 4 The genes located near the insertion sites.

Insertion location	Neighboring gene（< 20 kb）		Genomic location
Chr03：9283905-9283937	Upstream	LOC112326972	Chr03:9261716:9274439
	Downstream	LOC7475699	Chr03:9292074:9294391
Chr10：10868777-10868803	Upstream	LOC7498060	Chr10:10848741:10853156
	Downstream	LOC7498061	Chr10:10870346:10873516
		LOC7498062	Chr10:10880717:10883716

Analysis of the *TAFs* gene family

According to the results of whole-genome resequencing analysis, the T-DNA insertion site on Chr03 (9,283,895–9,283,937 bp) is located within the first exon of the LOC7478355 gene (9,283,876–9,291,377 bp). Therefore, the insertion of T-DNA disrupted the structure of the LOC7478355 gene. According to the National Center for Biotechnology Information (NCBI) analysis, the LOC7478355 gene, which belongs to the *TAFs* gene family, encodes a TAF12 protein, which is one of the core subunits constituting the basic transcription factor TFIID. To understand the impact that this disruption of the gene structure has on the function of this gene, we first analyzed the *TAFs* gene family to clarify the number of genes encoding TAF12 protein in the genome.

We identified 33 *TAFs* genes in the genome of *P. trichocarpa* through bioinformatics analysis. The 33 *PtTAFs* genes were renamed according to their chromosomal positions and the phylogenetic tree constructed with PtTAFs and AtTAFs proteins (Table 5; Fig. 12A). Within the *TAFs* gene family, there are three genes encoding TAF12 protein *PtTAF12*, *PtTAF12b*, and *PtTAF12c*. Through synteny analysis of the *PtTAFs* gene family, we identified five segmental duplication events involving 10 *PtTAF* genes that encode TAF7, TAF8, and TAF15 proteins. No duplicated segments containing genes encoding TAF12 protein were identified, indicating that *PtTAF12*, *PtTAF12b*, and *PtTAF12c* were not formed from segmental duplication occurring among the three genes (Fig. 12B). The RNA-seq results showed that the expression levels of the three genes in Pb29 leaves were slightly higher than those in poplar 741, but none of the differences were significant, indicating that the transcriptional abundance of the genes encoding TAF12 protein did not change significantly (Fig. 13).

Table 5 The physical characteristics of *TAFs* gene family in *Populus trichocarpa*.

Name	Gene symbol	Chr	Genomic location	Name	Gene symbol	Chr	Genomic location
TAF1	LOC7486075	7	13532968:13553923	*TAF9*	LOC7478099	1	14186976:14189156
TAF1b	LOC7472543	17	3964494:3986991	*TAF10*	LOC7465449	18	1025682:1027995
TAF2	LOC18105528	15	5635684:5667456	*TAF11*	LOC7491456	1	5828629:5834085
TAF4	LOC7462851	2	8999445:9006794	*TAF11b*	LOC112328249	7	19133:20885
TAF4b	LOC7497338	14	1709740:1716465	*TAF12*	LOC7478108	1	14349460:14356956
TAF5	LOC7489149	6	26638264:26645300	*TAF12b*	LOC7478355	3	9283876:9291377
TAF5b	LOC7465413	18	1568484:1574762	*TAF12c*	LOC7454916	6	6315941:6324206
TAF6	LOC7487521	1	36978924:36984663	*TAF13*	LOC7495223	14	9488112:9491114
TAF6b	LOC7468541	14	12685250:12694053	*TAF14*	LOC7453610	4	18542675:18546560
TAF7	LOC7457707	1	423095:426480	*TAF14b*	LOC7480178	7	991034:993419
TAF7b	LOC7463581	3	21459132:21462376	*TAF14c*	LOC7494288	9	10457158:10460675
TAF8	LOC7480583	4	10747129:10749249	*TAF15*	LOC18095136	1	28789532:28794606
TAF8b	LOC18100163	6	10071150:10074496	*TAF15b*	LOC18100420	6	15177976:15183124
TAF8c	LOC18103895	12	13966349:13966864	*TAF15c*	LOC7498166	8	5359361:5362704
TAF8d	LOC7457637	15	13330434:13331563	*TAF15d*	LOC18102067	9	7426288:7431562
TAF8e	LOC7486490	16	7582005:7583848	*TAF15e*	LOC7481594	10	17229566:17233844
TAF8f	LOC7484577	17	11185460:11187491				

Discussion

Whole-genome resequencing using NGS and Nanopore sequencing improved the accuracy of T-DNA insertion site analysis

Molecular information, such as the locations of T-DNA insertion sites and copy numbers, is necessary for more comprehensive safety assessments of transgenic plants. PCR-based methods are often used to elucidate T-DNA insertion sites and copy numbers. However, these methods are time-consuming, labor-intensive, and produce inaccurate results. When T-DNA integration patterns or the genomes of T-DNA mutants are relatively complex, PCR-based methods cannot be used to accurately determine all T-DNA insertion sites and copy numbers. For example, Gang et al. performed 120 rounds of PCR using 12 border primers and 10 arbitrarily degenerated primers, and located only two T-DNA insertion sites in a birch T-DNA mutant; in contrast, six T-DNA insertion sites were located via genome resequencing using NGS [16] Whole-genome resequencing is a more effective method for analyzing T-DNA insertion sites and copy

numbers. With the emergence and development of high-throughput NGS technology, NGS is now widely used to elucidate T-DNA insertion sites and copy numbers because of its high throughput capability and low cost. However, NGS reads are too short to obtain complete information on the T-DNA insertion sites [17]. In this study, although both NGS and Nanopore sequencing located two T-DNA insertion sites in the Pb29 genome, NGS only detected junction reads on one side of each insertion site. In contrast, complete T-DNA sequences and flanking sequences of T-DNA insertion sites were elucidated using Nanopore sequencing, because it can produce longer reads. Nanopore sequencing can also be used to analyze the entire genome of a T-DNA mutant and identify any chromosomal rearrangements due to T-DNA integration [18]. Therefore, NGS and Nanopore sequencing should be used together to analyze T-DNA mutants, to improve the accuracy of T-DNA insertion site analysis.

T-DNA insertion sites and copy numbers constitute important molecular information for safety assessments of transgenic plants

There have been several controversial incidents regarding the safety of genetically modified products, such as those involving *Bertholletia excelsa* [19] and the monarch butterfly [20]. Accordingly, the potential threats of genetically modified organisms to the environment and human health are of widespread concern. As a result, many countries have formulated legislation and established agencies to conduct safety assessments and management of genetically modified organisms. The T-DNA insertion site, which serves as a label for genetically modified materials, provides important molecular information for the screening and identification processes that are conducted during safety assessments of genetically modified materials, before the materials are released into the environment [21]. Additionally, due to the random location of T-DNA insertion sites and the existence of position effects [7], the euchromatin or heterochromatin region into which the T-DNA is inserted, and the flanking sequences of the T-DNA insertion sites, affect the expression activity of the foreign gene [22, 23]. This activity may also be correlated with the copy number of the foreign gene. For example, the fatty acid content in transgenic rape is positively correlated with the copy number of the thioesterase gene, which encodes an acyl-acyl carrier protein [24]. Furthermore, Cervera et al. found a significant negative correlation between the expression level of the *GUS* gene and its copy number in transgenic citrus [25]. Therefore, T-DNA insertion sites and copy numbers are closely related to the transcription level of the foreign gene, which is important to consider during safety assessments of transgenic plants [26]. In this study, we located two T-DNA insertion sites in the genome of Pb29, a transgenic poplar 741 line, at 9,283,905–9,283,937 bp on Chr03 and 10,868,777–10,868,803 bp on Chr10. According to the sequence information associated with the junction reads, T-DNA was inserted in opposite directions at those two insertion sites. The T-DNA sequences at both insertion sites did not result from tandem duplication; instead, a single-copy integration pattern was observed at both sites. The T-DNA insertion sites and directions elucidated via resequencing were further confirmed using PCR amplification. Our results provide molecular information that could aid safety assessment and management of transgenic poplar 741.

T-DNA and flanking sequence analysis

T-DNA integration into a receptor genome often results in base deletions on the T-DNA left and right border sequences, or to duplications, deletions, and inversions of DNA sequences in the receptor genome [27]; it can even induce chromosomal rearrangement [28]. Kim et al. analyzed a large number of transgenic rice plants and found a difference in the number of base deletions on the T-DNA left and right border sequences, with more deletions occurring on the left side [29]. In a birch T-DNA mutant, the integration of T-DNA led to the deletion or translocation of several chromosomal fragments [15]. Although we did not observe chromosomal rearrangement in the Pb29 genome, base deletions were observed in the T-DNA left and right border sequences, and in the genome sequence at the T-DNA insertion sites; this phenomenon is common in many genetically modified materials. However, we also found a 24-bp fragment that was inserted between the T-DNA insertion site and its right flanking sequence on Chr03; however, further analysis is needed to elucidate the specific source of the fragment. No *API* gene was detected within the T-DNA sequences at either insertion site, indicating that the *API* gene was not integrated into the Pb29 genome. The two T-DNA sequences are exactly the same, indicating that the *API* gene was not present in the expression vector in *Agrobacterium* before transformation. Therefore, Pb29 is a transgenic line that contains only one insect resistance gene.

Analysis of the expression levels of genes near the T-DNA insertion sites

T-DNA is randomly inserted into the genome [30, 31]. The introduction of exogenous genes may affect the regulation and expression of endogenous genes in plants [32]. When T-DNA is inserted into a coding gene, the function of the gene is affected. Furthermore, insertion into an intergenic region may affect the expression activity of upstream and downstream genes, resulting in unexpected effects [33]. In the transgenic rice 04Z11EM13 line, T-DNA was inserted into the fourth exon of the *OsBC1L4* gene, resulting in a mutant phenotype that exhibited fewer tillers and dwarfism [34]. Liu et al. analyzed the flanking sequences of the T-DNA insertion site in a rice flag leaf mutant and found that T-DNA insertion led to a significant reduction in the expression of the neighboring AK100376 gene, thus causing phenotypic change [35]. In the genome of the transgenic poplar 741 line Pb29, one T-DNA copy was inserted into the first exon of a gene encoding TAF12 protein, which belongs to the *TAFs* gene family, on Chr03. Through gene family analysis, we identified three genes encoding TAF12 protein in the genome of *P. trichocarpa*. However, the expression of these three genes in Pb29 leaves did not change significantly, which may be due to T-DNA being integrated into only one homologous chromosome, whereas poplar 741 is triploid and has three alleles for each gene. Of the neighboring genes at the two T-DNA insertion sites, except for the LOC7498061 gene, which is located closest to an insertion site, the expression levels of the other four genes did not change significantly in Pb29 leaves, implying that the insertion of T-DNA had little effect on the expression of endogenous genes in the Pb29 genome. Any changes in growth or physiology that may result from the significantly upregulated expression of the LOC7498061 gene in Pb29 need to be studied further.

Conclusions

In this study, we resequenced the whole genomes of poplar 741 and the transgenic poplar 741 line Pb29 using NGS and Nanopore sequencing. In the Pb29 genome, we found that the T-DNA sequence was inserted inversely into the 9,283,905–9,283,937-bp region on Chr03, and in the forward direction into the 10,868,777–10,868,803-bp region on Chr10. Both insertion sites exhibited a single-copy integration pattern, and the locations and directions of T-DNA insertion were confirmed using PCR amplification. After the T-DNA copies had been inserted into the genome, different degrees of base deletions were detected on the T-DNA left and right border sequences, and in the flanking sequences of the insertion sites. A fragment was found to be inserted between the insertion site and right flanking sequence on Chr03, and no chromosomal rearrangement was detected in the Pb29 genome. Only the *BtCry1Ac* gene was detected in the T-DNA sequence at both insertion sites; no *API* gene was detected, indicating that Pb29 is a transgenic line containing only one insect resistance gene. The insertion of T-DNA destroyed the structure of a gene encoding TAF12 protein on Chr03, but the transcriptional abundance of this gene did not change significantly in Pb29 leaves. Except for the LOC7498061 gene, which is located closest to a T-DNA insertion site, the expression of four other neighboring genes did not change significantly in Pb29 leaves. This study provides molecular information that is important for the safety assessment and management of transgenic poplar 741, as well as a theoretical basis for safety assessments of other transgenic poplar.

Methods

Plant materials

The experimental materials used in this study were tissue culture seedlings of poplar 741 and the transgenic poplar 741 line Pb29 cultured in our tissue culture room. Pb29 is the main transgenic poplar 741 line planted commercially. It theoretically carries two insect resistance genes (*BtCry1AC* and *API*) that confer high levels of resistance to lepidopteran pests, such as *Hyphantria cunea* and *Clostera anachoreta* [2, 36, 37]. The leaves of poplar 741 and Pb29 were collected, immediately placed into liquid nitrogen, and preserved at −80° for subsequent DNA extraction.

Genome resequencing using NGS and data analysis

Genome resequencing of Pb29 via NGS was performed by Biomics Co., Ltd. (Beijing, China). Genomic DNA was extracted using a plant DNA extraction kit (SENO Biological Technology Co., Ltd, Zhangjiakou, China) in accordance with the manufacturer's protocol and quantified using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The DNA was broken into fragments with an average length of 300 bp to construct the library. The library was then sequenced using the HiSeq sequencing platform (Illumina, San Diego, CA, USA), and 150-bp paired-end reads were generated. After sequencing, the raw data were initially screened to remove adapter sequences and low-quality reads ($Q < 20$) and thus obtain clean, high-quality data. Using AIM-HII software [38], the clean reads were compared against the *Populus trichocarpa* genome sequence and the vector sequence to identify junction reads. First, the junction reads that aligned with both the reference genome sequence and the vector sequence

were identified, and the T-DNA insertion sites and directions were then determined based on the alignment information associated with the junction reads.

PCR verification of T-DNA insertion sites and directions

To verify the T-DNA insertion sites and directions, we designed primers based on the flanking sequences of the T-DNA insertion sites and the T-DNA sequence (Table 6), and amplified the genomic DNA of poplar 741 and Pb29. The PCR products corresponding to the sides of the junction reads undetected by NGS were purified and ligated into the pUCm-T vector. After 12 h at 16°C, the plasmid was transformed into *E. coli* DH5α competent cells. The cells were shaken for 1 h in a constant-temperature shaker at 37°C, and then plated onto a Luria-Bertani agar plate containing ampicillin and cultured for 8 h at 37°C. Single colonies were selected and sent to Beijing Zhongke Xilin Biotechnology Co., Ltd. (Beijing, China) for sequencing, to determine the integrity of the flanking sequences.

Table 6 Primer sequences for verifying T-DNA insertion sites.

Primer name	Sequences (5' to 3')
131#S5F	TAGTGACCTTAGGCGACTTTTGAACG
131#S2F	ATTTGGGTGATGGTTCACGTAGTGG
Chr3u-F1	AGAGTACGCCCTTTGATTATTTGCT
Chr3d-R2	GCCTGACATTGCGGTGACATTCTGC
Chr10u-F2	CGACGAGATGCCTCCACCATTCTGA
Chr10d-R2	TCTTCTATGGTTGCTCCTGCTTTGT

Genome resequencing using Nanopore sequencing and data analysis

The genomic DNA of poplar 741 and Pb29 was resequenced using the Nanopore sequencing platform (Biomarker Technologies, Beijing, China). After extracting the genomic DNA of poplar 741 and Pb29, the purity, concentration, and integrity of the extracted DNA were inspected using a NanoDrop spectrophotometer, Qubit fluorometer (Invitrogen, Carlsbad, CA, USA), and 0.35% agarose gel electrophoresis, respectively. After passing the quality checks, the DNA samples were used to construct libraries and sequenced with Ligation Sequencing Kit 1D (SQK-LSK109; Oxford Nanopore Technologies, Oxford, UK). Low-quality reads, reads with adapters, and short sequencing reads (length < 500 bp) were filtered from the raw reads. Then, Minimap2 software (https://github.com/lh3/minimap2) [39] was used to compare the clean reads with the reference genome and vector sequences (at the same time). The junction reads thus obtained were saved in BAM file format, and the data were visualized with IGV software (http://www.broadinstitute.org/software/igv/) to locate the T-DNA insertion sites. By comparing the clean reads and reference genome sequence, information such as alignment rate and sequencing depth and coverage could be calculated.

T-DNA and flanking sequence analysis

Part of the flanking sequences were isolated from the junction reads obtained by NGS and the Sanger sequencing reads obtained from the PCR products; the other part of the flanking sequences and complete T-DNA sequences were extracted from the junction reads obtained by genome resequencing using Nanopore sequencing. All flanking and T-DNA sequences were compared with the genome and vector sequences, respectively, to determine the integrity of the flanking and T-DNA sequences at the insertion sites.

RNA-sequencing (RNA-Seq) analysis of the expression levels of genes located near the insertion sites

To detect whether the insertion of T-DNA affects the expression of upstream and downstream genes near the insertion site, we analyzed poplar 741 and Pb29 leaves using RNA-seq. First, healthy and mature leaves along a long branch within the upper parts of mature trees of poplar 741 and Pb29 growing in the test forest were collected. Then, total RNA was extracted from the leaves using a plant RNA extraction kit (SENO Biological Technology Co., Ltd.) in accordance with the manufacturer's instructions. The concentration and quality of the RNA samples were determined using a NanoDrop 2000 spectrophotometer and an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). After the quality of the RNA samples had been verified, a cDNA library of each sample was constructed and Illumina sequencing was performed by LC Bio Technology Co., Ltd. (Hangzhou, China). FPKM values were used to examine changes in the expression of genes upstream and downstream of the T-DNA insertion sites. Three biological replicates for each poplar line were sampled.

Identification of *TAFs* gene family members and expression of genes encoding TAF12 protein

The whole-genome and protein sequences of *P. trichocarpa* were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/genome/98). Identified TAFs protein sequences from *A. thaliana* (downloaded from the Arabidopsis Information Resource; https://www.arabidopsis.org/) were used as queries in BLASTP searches against the *P. trichocarpa* genome with an e-value cutoff of 1e-10. Redundant sequences were manually removed, and all candidate proteins were analyzed and verified using InterProScan (http://www.ebi.ac.uk/interpro/search/sequence-search) and the Conserved Domains Database (https://www.ncbi.nlm.nih.gov/cdd). A multiple sequence alignment of TAFs proteins was generated using ClustalW in MEGA 7 (https://www.megasoftware.net) with default parameters. A neighbor-joining phylogenetic tree was constructed based on the alignment results with the following settings: Poisson model, pairwise deletion, and 1,000 bootstrap replications. *PtTAFs* gene duplication events were analyzed using the Multiple Collinearity Scan toolkit (MCScanX; http://chibba.pgml.uga.edu/mcscan2) [40]. The expression levels of the genes encoding TAF12 protein in leaves were analyzed using RNA-seq.

Abbreviations

PCR: Polymerase chain reaction; NGS: Next-generation sequencing; IGV: Integrative Genomics Viewer; SV: Structural variation; FPKM: Fragments Per Kilobase per Million; NCBI: National Center for Biotechnology Information

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

All relevant data are within this ariticle and its additional files.

Competing interests

The authors declare that they have no conflict of interest.

Funding

This study was supported by the National Key Program on Transgenic Research (2018ZX08020002) and the Basic Research Plan Project of Hebei Province (18966801D).

Authors' contributions

XC and YD conducted the experiments, analyzed the data and wrote the manuscript. YH and JF revisied the manuscript. MY and JZ designed the experiments and edited the manuscript. All authors have read and approved the manuscript.

Acknowledgments

We would like to thank Textcheck (www.textcheck.com) for English language editing of this manuscript.

References

1. Feng H, Guo J, Wang W, Song X, Yu S. Soil depth determines the composition and diversity of bacterial and archaeal communities in a poplar plantation. Forests. 2019;10(7):550.
2. Wang G, Dong Y, Liu X, Yao G, Yu X, Yang M. The current status and development of insect-resistant genetically engineered poplar in China. Front Plant Sci. 2018;9:1048.

3. Zhao C, Wang J, Zhao J, Pang D, Zhang D, Yang M. Expression characteristics of *Bt* gene in transgenic poplar transformed by different multi-gene vectors. Scientia silvae sinicae. 2019;55(09):61-70.
4. Lu MZ, Hu JJ. A brief overview of field testing and commercial application of transgenic trees in China. BMC Proc. 2011;5(Suppl 7):O63.
5. Tang J, Scarth R, Fristensky B. Effects of genomic position and copy number of Acyl-ACP thioesterase transgenes on the level of the target fatty acids in *Brassica napus* Mol Breed. 2003;12:71-81.
6. Chen JM, Carlson AR, Wan JM, Kasha KJ. Chromosomal Location and Expression of Green Fluorescent Protein (*gfp*) Gene in Microspore Derived Transgenic Barley (*Hordeum vulgare*). Acta Genet Sin. 2003;30(8):697-705.
7. Chandler VL, Vaucheret H. Gene Activation and Gene Silencing. Plant Physiol. 2001;125:145-148.
8. Liu YG, Whittier RF. Thermal Asymmetric Interlaced PCR: Automatable Amplification and Sequencing of Insert End Fragments from P1 and YAC Clones for Chromosome Walking. Genomics. 1995;25:674-681.
9. Ochman H, Gerber AS, Hartl DL. Genetic applications of an inverse polymerase chain reactio. Genetics. 1988;120(3):621-623.
10. O'Malley RC, Alonso JM, Kim CJ, Leisse TJ, Ecker JR. An adapter ligation-mediated PCR method for high-throughput mapping of T-DNA inserts in the Arabidopsis genome. Nat Protoc. 2007;2(11):2910-2917.
11. Papazova N, Ghedira R, Glabeke SV, Bartegi A, Windels P, Taverniers I, et al. Stability of the T-DNA flanking regions in transgenic *Arabidopsis thaliana* plants under influence of abiotic stress and cultivation practices. Plant Cell Rep. 2008;27:749-757.
12. Yang K, Wu XL, Lang CX, Chen JQ. Isolation of the flanking sequences adjacent to transgenic T-DNA in *Brassica napus* genome by an improved inverse PCR method. Agric Sci Technol. 2010;11(2):65-68,139.
13. Guo B, Guo Y, Hong H, Qiu L. Identification of genomic insertion and flanking sequence of G2-EPSPS and GAT transgenes in soybean using whole genome sequencing method. Front Plant Sci. 2016;7:1009.
14. Park D, Kim D, Jang G, Lim J, Shin Y, Kim J. Efficiency to discovery transgenic loci in GM rice using next generation sequencing whole genome re-sequencing. Genomics Inform. 2015;13(3):81-85.
15. Gang H, Liu G, Zhang M, Zhao Y, Jiang J, Chen S. Comprehensive characterization of T-DNA integration induced chromosomal rearrangement in a birch T-DNA mutant. BMC genomics. 2019;20:311.
16. Gang H, Li R, Zhao Y, Liu G, Chen S, Jiang J. The birch GLK1 transcription factor mutant reveals new insights in chlorophyll biosynthesis and chloroplast development. J Exp Bot. 2019;70(12):3125-3138.
17. Williams-Carrier R, Stiffler N, Belcher S, Kroeger T, Stern DB, Monde RA, et al. Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator

lines of maize. Plant J. 2010;63:167-177.

18. Jupe F, Rivkin AC, Michael TP, Zander M, Motley ST, Sandoval JP, et al. The complex architecture and epigenomic impact of plant T-DNA insertions. PLoS Genet. 2019;15(1):e1007819.
19. Nordlee JA, Taylor SL, Townsend JA, Thomas LA, Bush RK. 1996. Identification of a Brazil-nut allergen in transgenic soybeans. N Engl J Med. 1996;334(11):688-692.
20. Losey JE, Rayor LS, Carter ME. Transgenic pollen harms monarch larvae. Nature. 1999;399(6733):214.
21. Xu J, Hu H, Mao W, Mao C. Identifying T-DNA insertion site(s) of transgenic plants by whole-genome resequencing. Hereditas (Beijing). 2018;40(8):676-682.
22. Hilder VA, Barker RF, Samour RA, Gatehouse AMR, Gatehouse JA, Boulter D. Protein and cDNA sequences of Bowman-Birk protease inhibitors from the cowpea (Vigna unguiculata Walp.). Plant Mol Biol. 1989;13:701-710.
23. Tinland B, Schoumacher F, Gloeckler V, Bravo-Angel AM, Hohn B. The Agrobacterium tumefaciens virulence D2 protein is responsible for precise integration of T-DNA into the plant genome. EMBO J. 1995;14:3585-3595.
24. Tang J, Scarth R, Fristensky B. Effects of genomic position and copy number of Acyl-ACP thioesterase transgenes on the level of the target fatty acids in *Brassica napus* Mol Breed. 2003;12:71-81.
25. Cervera M, Pina JA, Juárez J, Navarro L, Peña L. A broad exploration of a transgenic population of citrus: stability of gene expression and phenotype. Theor Appl Genet. 2000;100:670-677.
26. Yang L, Wang C, Holst-Jensen A, Morisset D, Lin Y, Zhang D. Characterization of GM events by insert knowledge adapted re-sequencing approaches. Scientific Reports. 20123;3(10):2839.
27. Forsbach A, Schubert D, Lechtenberg B, Gils M, Schmidt R. A comprehensive characterization of single-copy T-DNA insertions in the *Arabidopsis thaliana* Plant Mol Biol. 2003;52:161-176.
28. Ruprechta C, Carrollb A, Persson S. T-DNA-induced chromosomal translocations in *feronia* and *anxur2* mutants reveal implications for the mechanism of collapsed pollen due to chromosomal rearrangements. Mol Plant. 2014;7:1591-1594.
29. Kim SR, Lee J, Jun SH, Park S, Kang HG, Kwon S. Transgene structures in T-DNA-inserted rice plants. Plant Mol Biol. 2003;52:761-773.
30. Kim SI, Gelvin SB. Genome-wide analysis of agrobacterium t-dna integration sites in the arabidopsis genome generated under non-selective conditions. Plant J. 2007;51(5):779-791.
31. Magori S, Citovsky V. Epigenetic control of Agrobacterium T-DNA integration. Biochim Biophys Acta. 2011;1809(8):388-394.
32. Deng L, Deng X, Wei S, Cao Z, Tang L, Xiao G. Development and identification of herbicide and insect resistant transgenic plant B1C893 in rice. Hybrid Rice. 2014;29(1):67-71.
33. Jiang X, Xiao G. Detection of unintended effects in genetically modified herbicide-tolerant (GMHT) rice in comparison with nontarget phenotypic characteristics. Afr J Agric Res. 2010;5(10):1082-1088.

34. Dai XX. Isolation of flanking sequences from a rice T-DNA insertional mutant library and function study of *OsBC1L* family genes.D. Thesis, Huazhong Agricultural university, Wuhan, China. 2009.
35. Liu H, Lu H, Luo L, Zhu ML. Phenotypic analysis of a rice flag leaf mutant and T-DNA flanking genes. Plant Science Journal. 2017;35(5):708-715.
36. Tian YC, Zheng JB, Yu HM, Liang HY, Li CQ, Wang JM. Studies of Transgenic Hybrid Poplar 741 Carrying Two Insect-resistant Genes. Acta Bot Sin. 2000;42(3);263-268.
37. Zhang Y, Zhang J, Lan J, Wang J, Liu J, Yang M. Temporal and spatial changes in Bt toxin expression in Bt-transgenic poplar and insect resistance in field tests. J For Res. 2016;27(6):1249-1256.
38. Esher SK, Granek JA, Alspaugh JA. Rapid mapping of insertional mutations to probe cell wall regulation in Cryptococcus neoformans. Fungal Genet Biol. 2015;82:9-21.
39. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-3100.
40. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49-e.

Figures

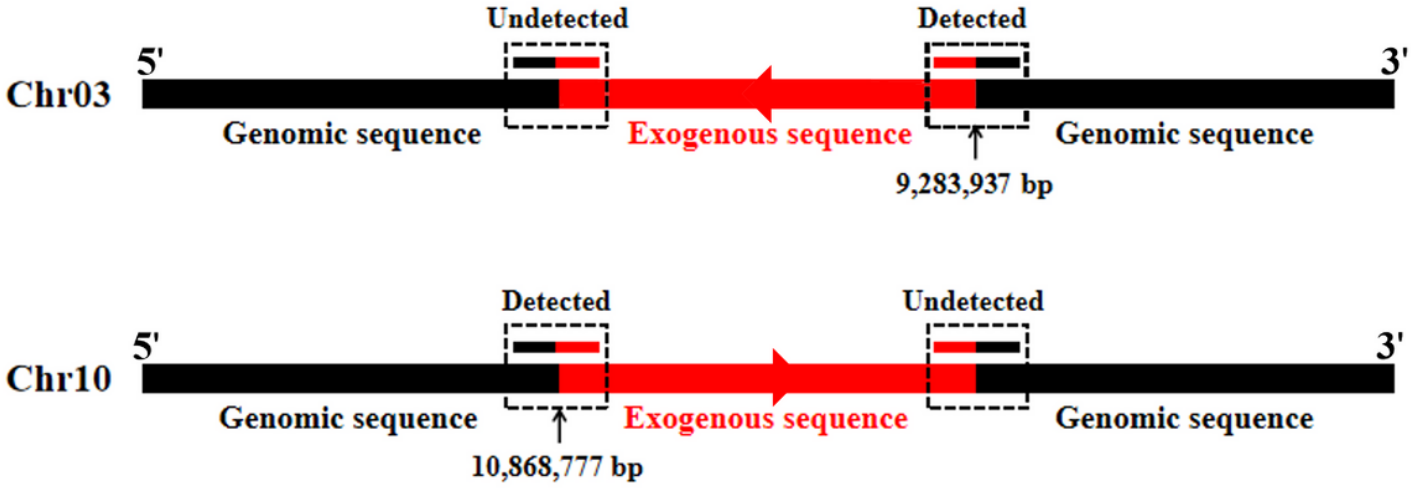

Figure 1

The detection results of T-DNA insertion sites obtained using NGS. The box with black dotted lines is the detected or undetected junction reads on both sides of the T-DNA insertion site.

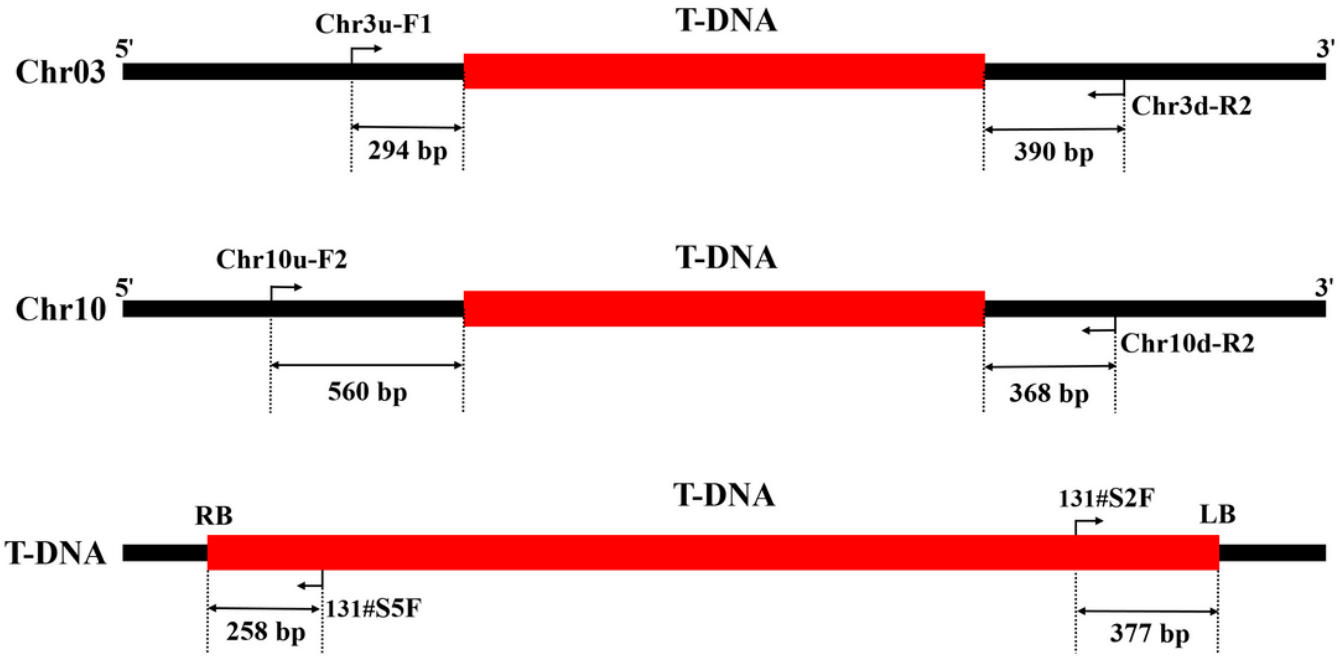

Figure 2

Schematic diagram of PCR primers for verifying the insertion sites and directions.

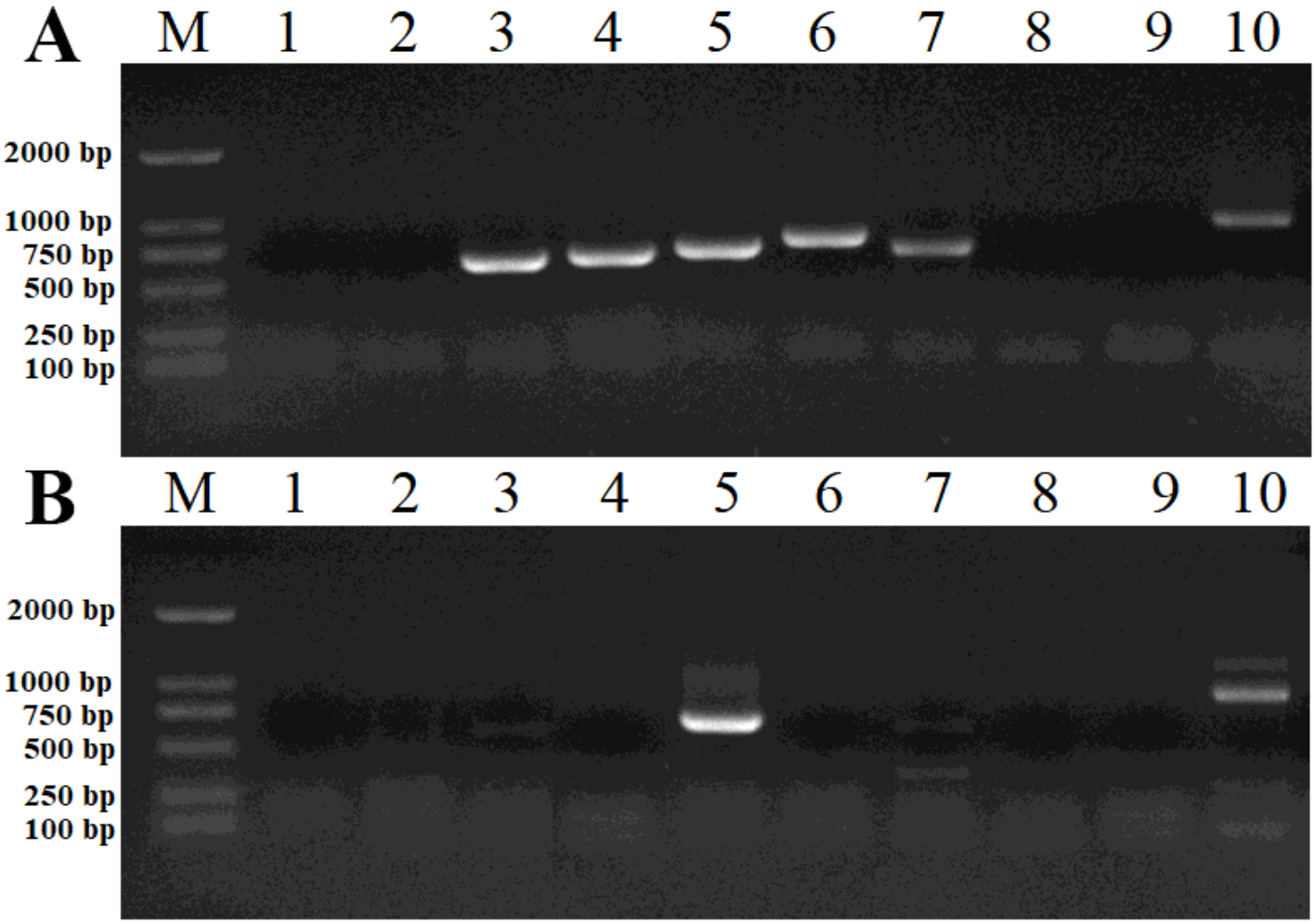

Figure 3

PCR verification result of T-DNA insertion sites and directions. (A) The results of PCR amplification of genomic DNA of poplar 741. (B) The results of PCR amplification of genomic DNA of Pb29.

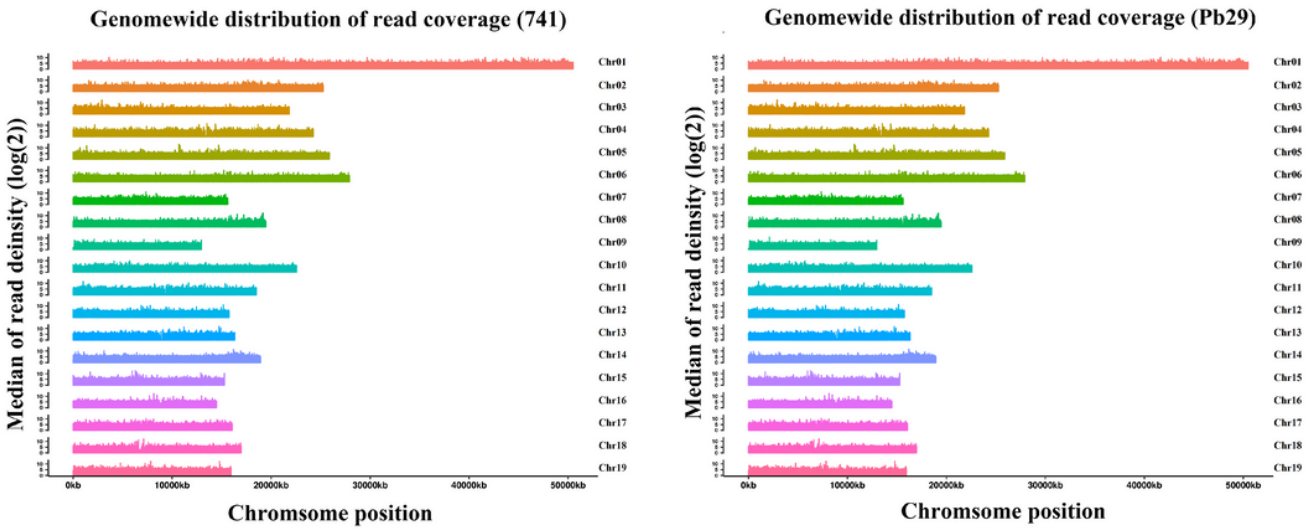

Figure 4

Genomewide distribution of read coverage of poplar 741 and Pb29.

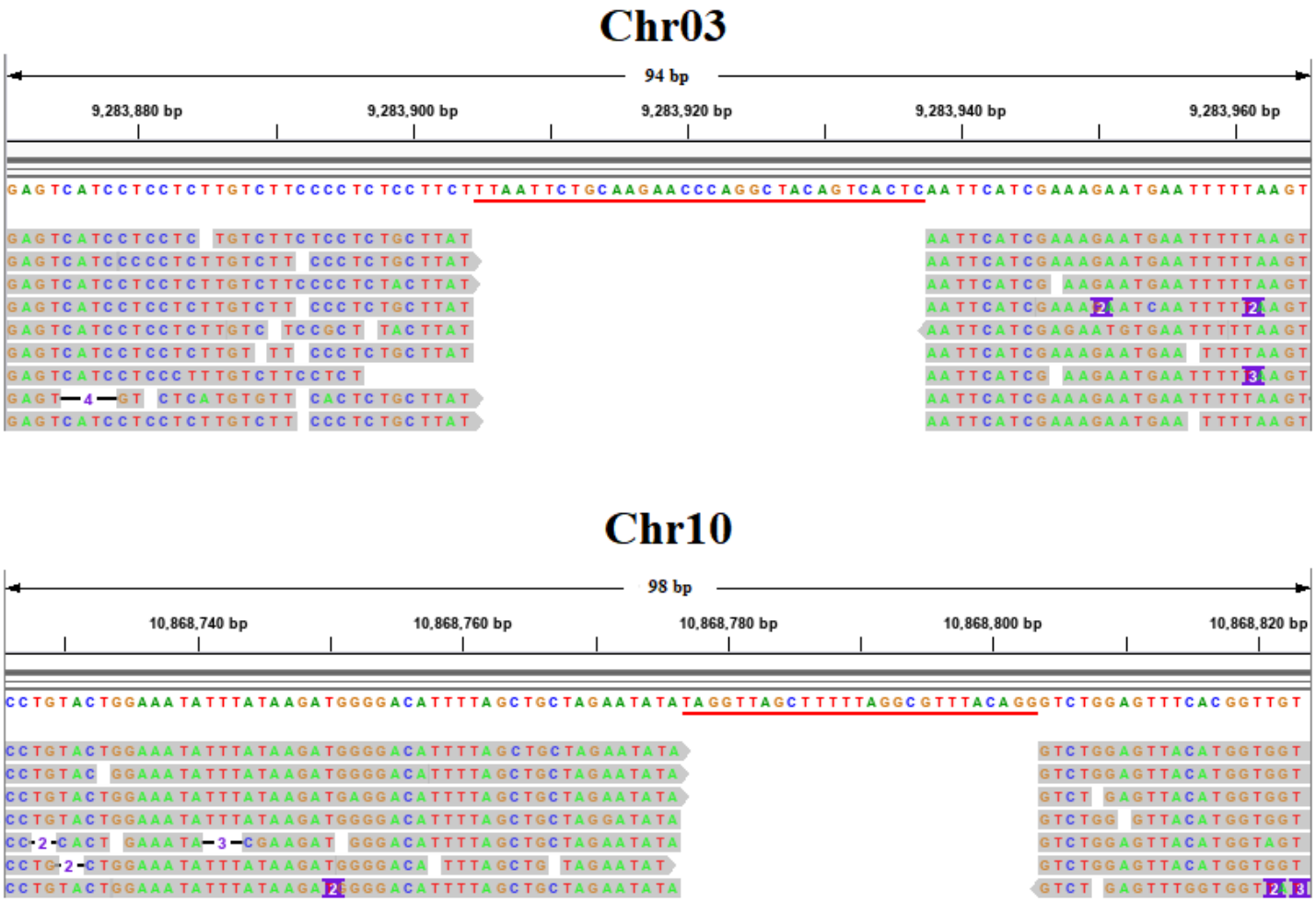

Figure 5

Visual analysis of junction reads using IGV software. The base sequences marked with the red line are the gaps that are not aligned to the reference genome.

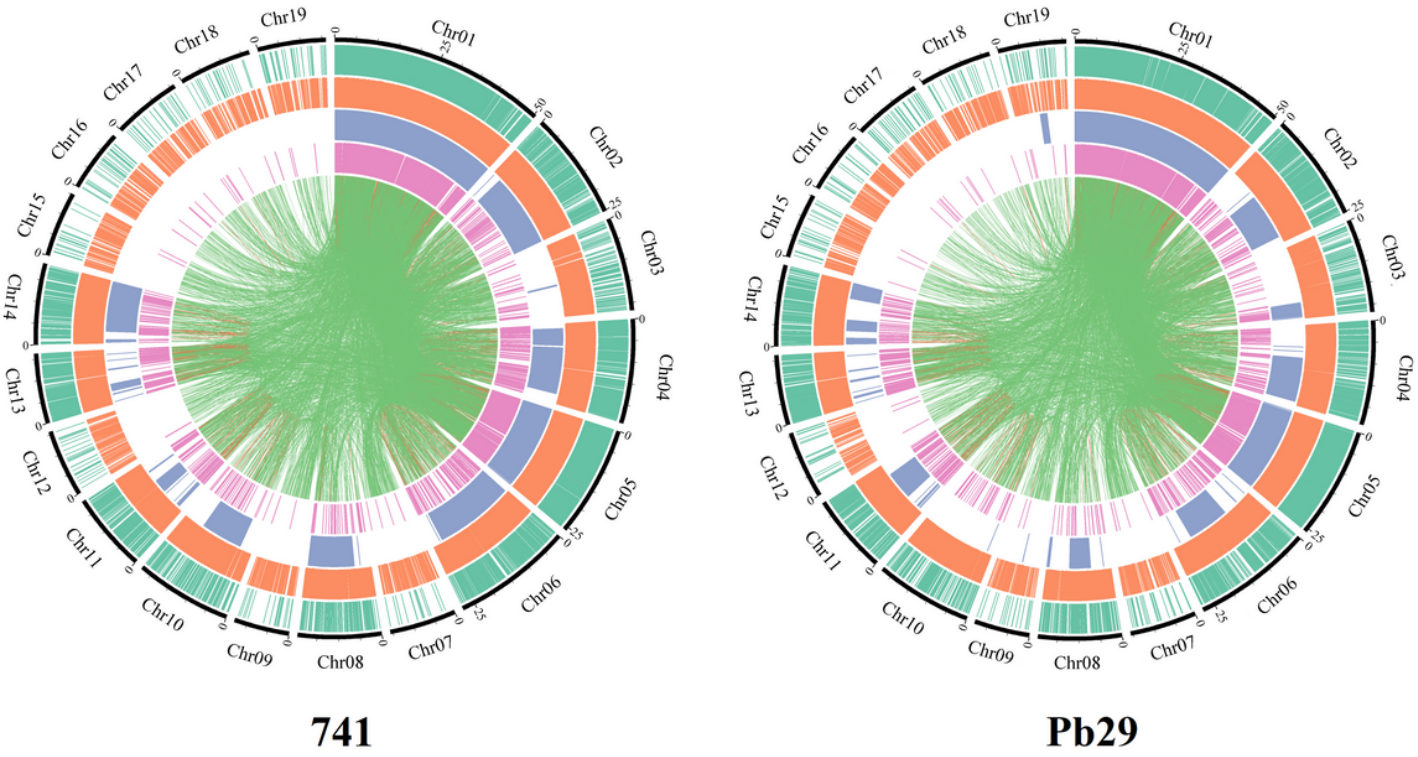

Figure 6

The distribution of SV variants on chromosomes in genomes of 741 poplar and Pb29 . From outside to inside: chromosome coordinates (Mb), insertion, deletion, inversion, duplication and translocation.

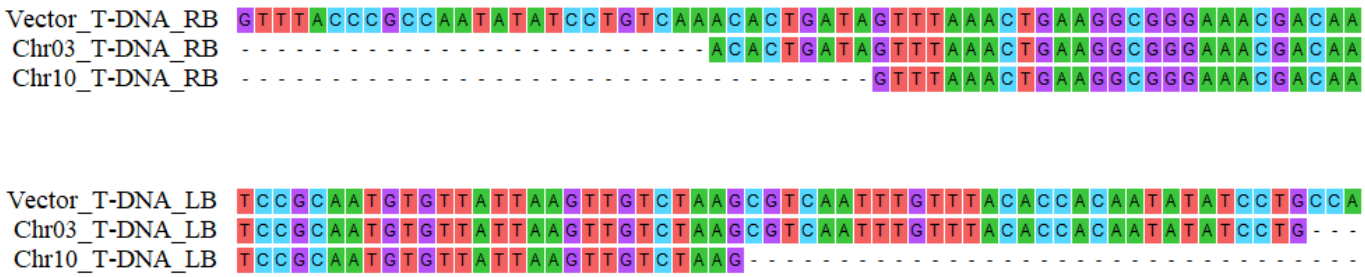

Figure 7

Analysis of the left and right T-DNA border sequences in both insertion sites.

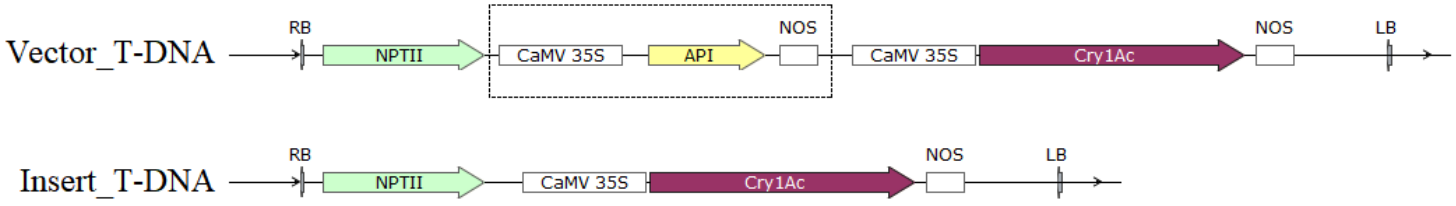

Figure 8

Analysis of inserted T-DNA sequences and vector T-DNA sequence.

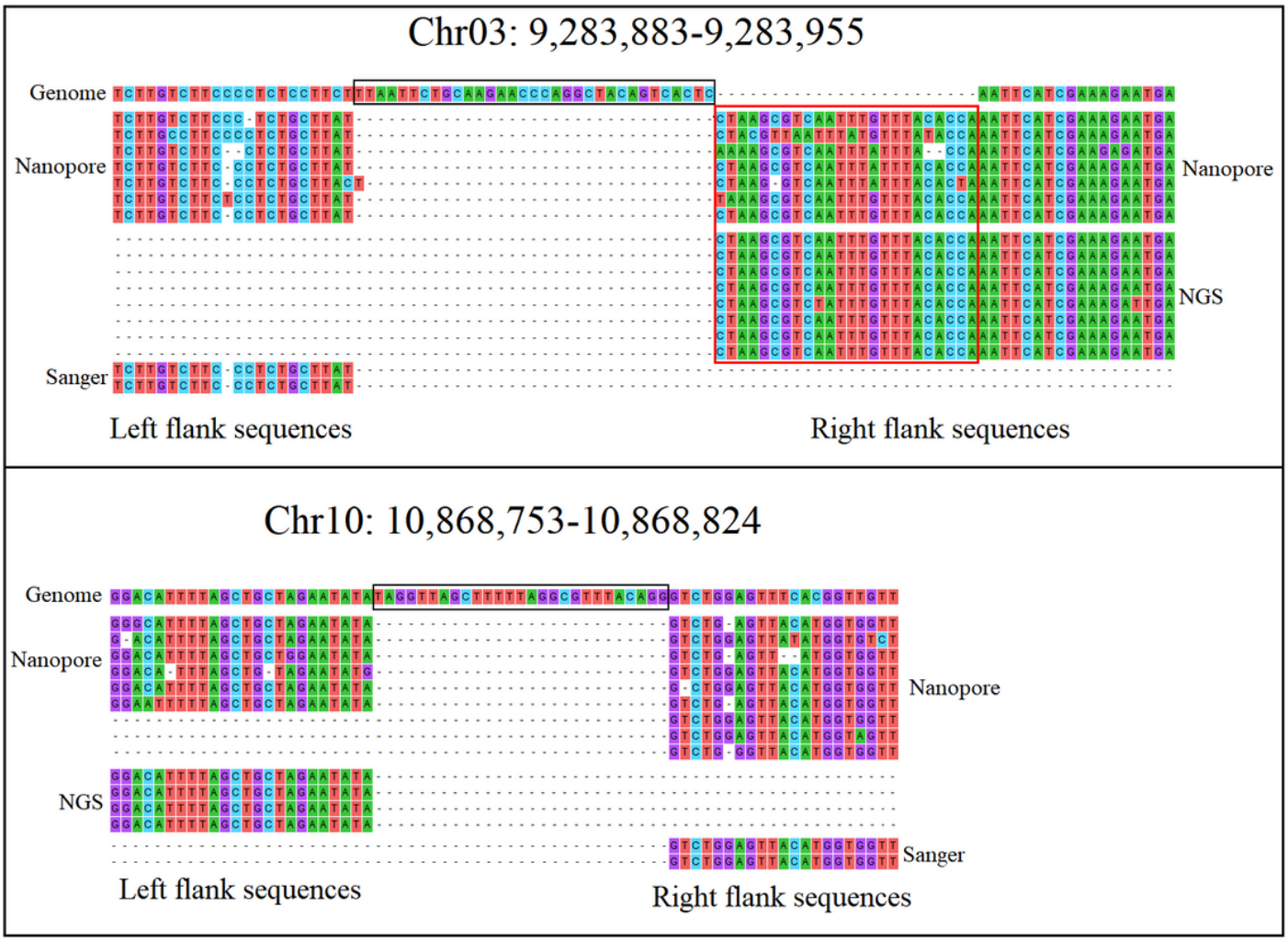

Figure 9

Analysis of flanking sequences of the both T-DNA insertion sites. The box with red outline is the base deletions occurred in the genome sequence and the box with black outline is the base insertions occurred in the genome sequence.

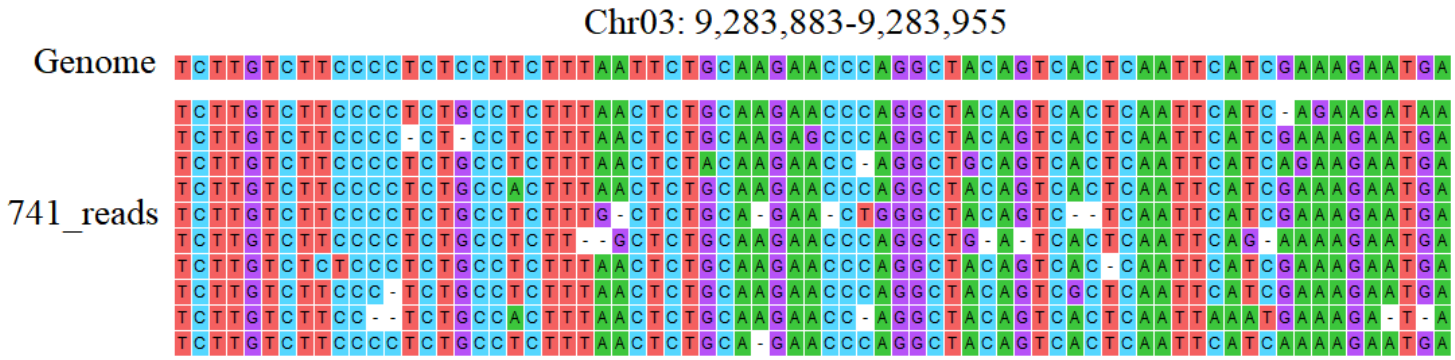

Figure 10

Partial alignment result of sequence data of poplar 741 with P. trichocarpa genome.

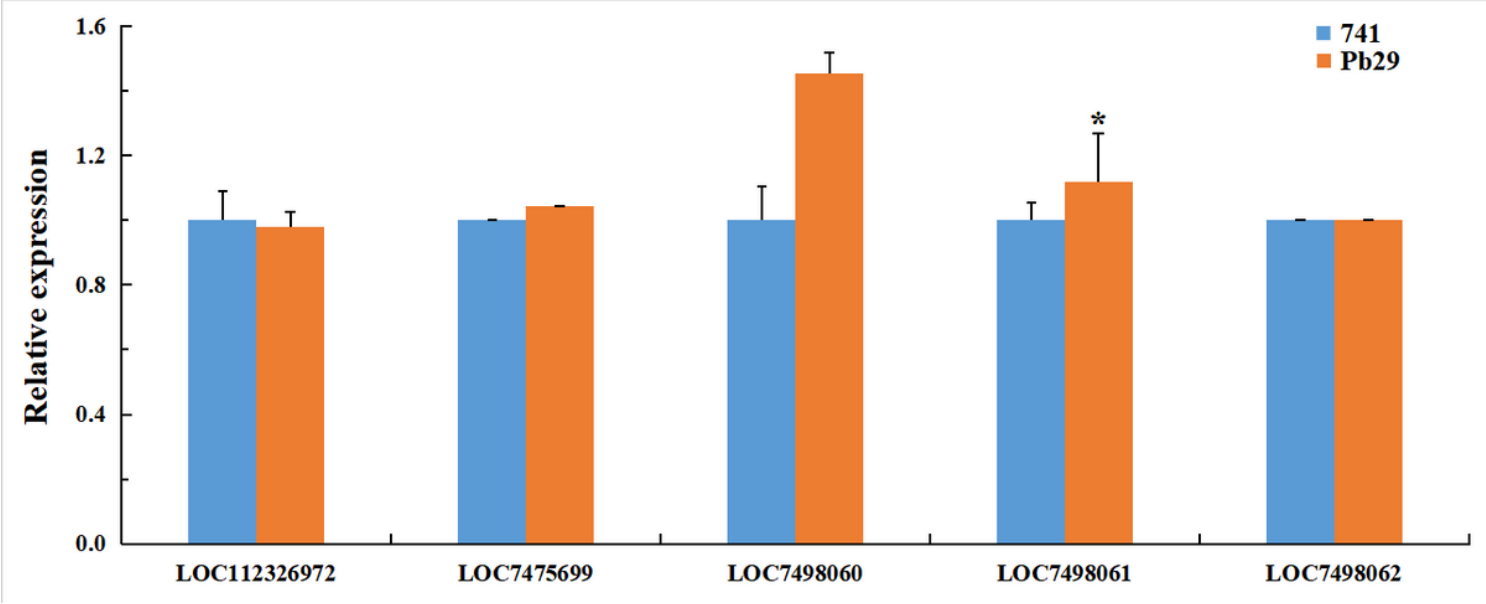

Figure 11

Analysis of the expression levels of genes located near the insertion sites in poplar 741 and Pb29. All data are presented as the mean ± SEM. (*, P < 0.05).

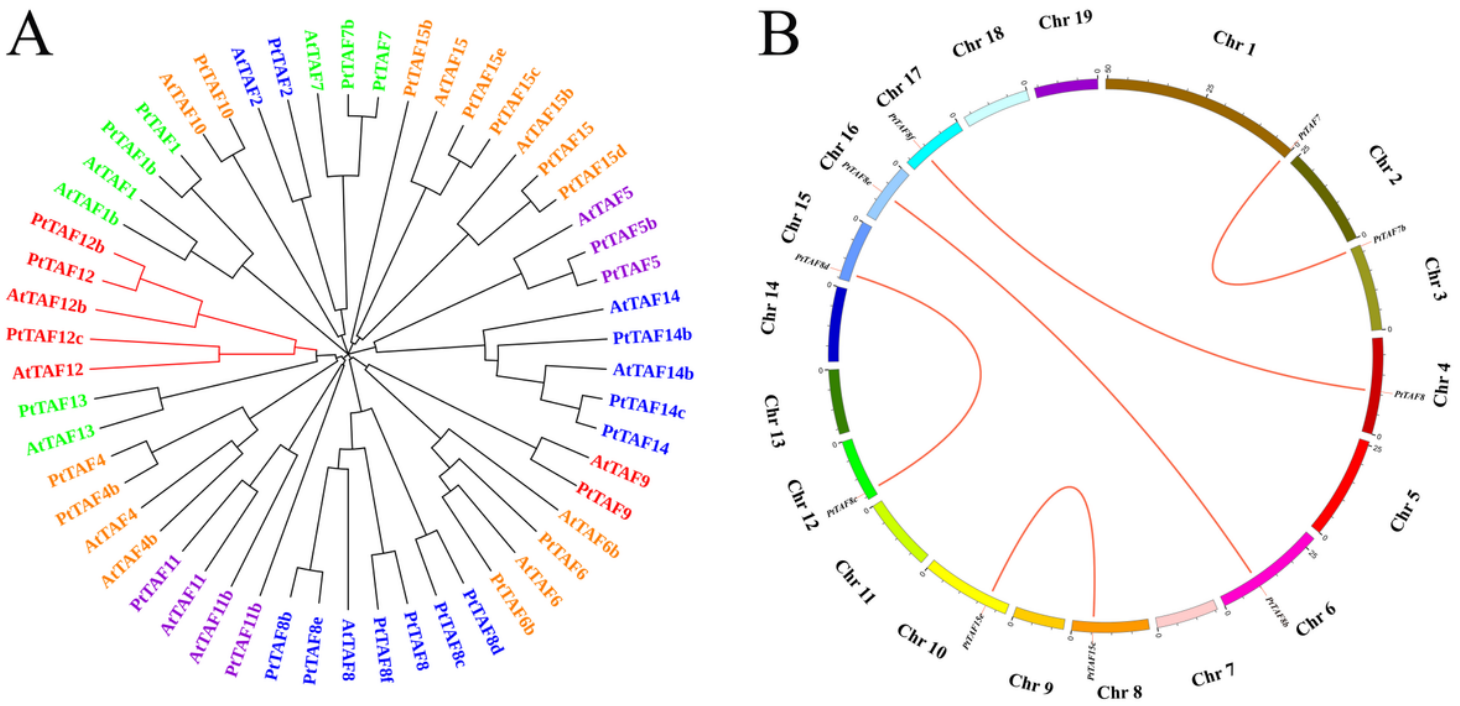

Figure 12

Phylogentic analysis (A) and synteny analysis (B) of PtTAFs family genes.

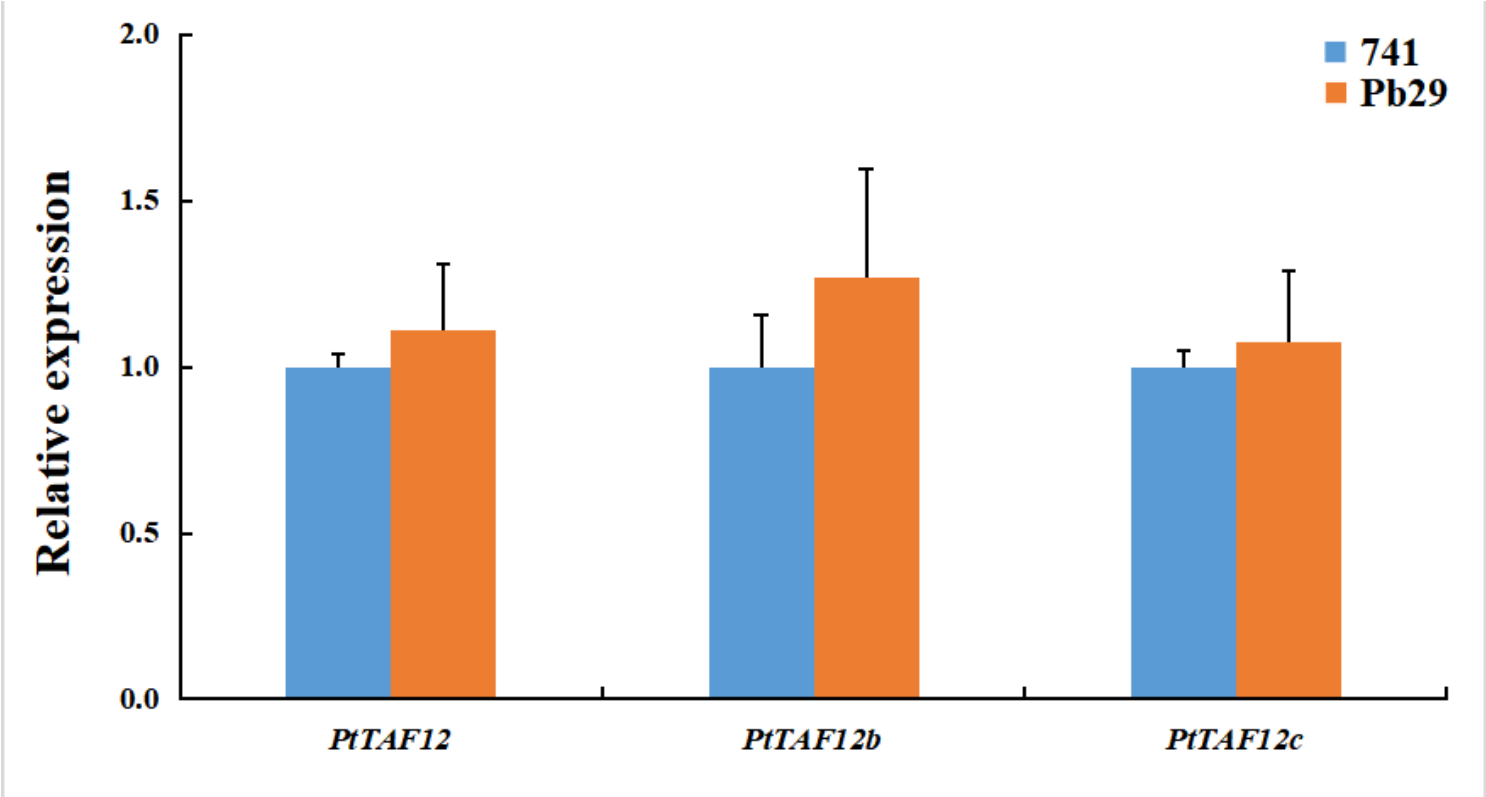

Figure 13

The relative expression of the genes encoding TAF12 protein in leaves of poplar 741 and Pb29. All data are presented as the mean ± SEM.

Supplementary Files

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

- TableS1.doc