

# Identification of transgene flanking sequences in a pre-market safety assessed sugarcane in Argentina

Ramón Enrique<sup>1\*</sup>, Daniel Kurth<sup>2</sup>, Enrique Ibarra-Laclette<sup>3</sup>, Aldo Noguera<sup>1</sup>, Björn Welin<sup>1</sup> and Atilio Pedro Castagnaro<sup>1</sup>

Crop Breeding and Applied Biotechnology  
21(3): e36862133, 2021  
Brazilian Society of Plant Breeding.  
Printed in Brazil  
<http://dx.doi.org/10.1590/1984-70332021v21n3a42>

**Abstract:** A prerequisite for commercial release of a transgenic crop variety in Argentina and other countries in the world is the identification and sequence analysis of the transgene(s) insertion site in the plant genome. Southern blot analysis showed a high number of transgene inserts in the glyphosate-resistant transgenic sugarcane line TUC 87-3RG. By using a targeted sequence enrichment approach, coupled with second-generation sequencing (SGS) and bioinformatic analyses, we were able to identify and subsequently analyze all flanking sequences for these transgene insertions in the sugarcane genome. PCR reactions confirmed 18 out of 29 candidate contigs as junction sequences that did not match essential genes in published sugarcane genomes. The methodology outlined here represents an efficient strategy for characterization of flanking sequences corresponding to multiple inserts of a transgene in a large, complex, and not fully sequenced plant genome such as that of sugarcane.

**Keywords:** Sugarcane, targeted sequence capture, SGS, flanking sequences, multiple inserts


## INTRODUCTION

Sugarcane (*Saccharum* hybrid) is an important crop for sugar and bioenergy production. Breeding programs around the world search for higher yielding cultivars that satisfy the demands of agricultural and industrial production (Carneiro et al. 2020).

Due to the extremely complex genome and genetic inheritance of sugarcane (Garsmeur et al. 2018), conventional breeding is slow, laborious, and costly. Another consequence of the highly complex genetics of this crop is difficulty in introducing a specific allele by introgression through successive backcrosses while maintaining the desired phenotype of the parental variety. For that reason, genetic transformation is a very valuable tool in sugarcane breeding, as this technology permits direct introduction of an agronomic or industrial trait to increase yields, lower production costs, or help generate more sustainable production.

We generated a herbicide resistant (HR) sugarcane variety TUC 87-3RG (Noguera et al. 2015), by genetic transformation within the sugarcane breeding program of the *Estación Experimental Agroindustrial Obispo Colombres* (EEAOC). This transgenic event was approved in 2015 for commercial release by the three regulatory entities evaluating transgenic crops in Argentina, the *Comisión*

**\*Corresponding author:**

E-mail: [renrique@eeaoc.org.ar](mailto:renrique@eeaoc.org.ar)  
 ORCID: 0000-0002-8292-3639

**Received:** 10 February 2021

**Accepted:** 05 July 2021

**Published:** 3 September 2021

<sup>1</sup> Instituto de Tecnología Agroindustrial del Noroeste Argentino (ITANOA), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Estación Experimental Agroindustrial Obispo Colombres (EEAOC), Av. William Cross 3150, Las Talitas (T4101XAC), Tucumán, Argentina

<sup>2</sup> Planta Piloto de Procesos Industriales y Microbiológicos (PROIMI), CONICET, San Miguel de Tucumán, Argentina

<sup>3</sup> Institute of Ecology, Veracruz, México

Nacional Asesora de Biotecnología Agropecuaria (CONABIA), the Servicio Nacional de Sanidad y Calidad Agroalimentaria (SENASA), and the Dirección Nacional de Mercados Agroalimentarios (DNMA). However, due to concerns regarding export markets, commercialization of this transgenic variety was put on hold by the Argentinean sugar industry.

During the regulatory process, extensive field testing was carried out to properly evaluate potential impact on fauna and flora, agronomic characteristics, chemical composition, and food safety (<https://www.argentina.gob.ar/agricultura/bioeconomia/biotecnologia/conabia>). In addition, molecular, genetic, and biochemical studies were conducted in order to ensure stable expression of the introduced gene(s), new protein concentrations, and proof of concept.

Another important characterization of genetically modified (GM) crops is the identification and analysis of the sequence of the insertion site(s) of the transgene(s) in the plant genome. Molecular characterization of inserted DNA and associated native flanking sequences consists of determining the number of insertion sites, the insert copy number at each insertion site, the DNA sequence of each inserted DNA, and the sequence of the native locus at each site (Kovalic et al. 2012). Currently, insert and copy numbers of the event, presence or absence of the genetic backbone, and genetic stability characterization is achieved by Southern blot analysis. The sequence of the insert and the DNA sequence of genomic flanking regions are commonly determined by means of sequencing overlapping polymerase chain reaction (PCR) fragments spanning the inserted DNA (Li et al. 2019). However, more recently, second-generation sequencing (SGS) and junction sequence analysis (JSA) have been applied for a complete molecular characterization of transgene insertions in GM events as they offers several advantages over the classical methods, such as simplicity, efficiency, and consistency. The latter method was successfully used for molecular characterization of two commercial soybean events, MON17903 and MON87704 (Kovalic et al. 2012), by using available reference genomes of soybean (Schmutz et al. 2010) and the LLRICE62 and TT51-1 rice events by using reference genomes of *Oryza sativa ssp. japonica* (Yang et al. 2013). The success of this strategy, however, depends on the availability of a good reference genome for the transgenic plant species. Given the absence of a specific reference genome in the case of commercial sugarcane varieties, a strategy of *de novo* sequence assembly of the transgenic variety comparing all generated reads has to be applied to find overlaps. Currently, there are three assemblies of sugarcane genomes that could be useful in this kind of analysis (Garsmeur et al. 2018, Zhang et al. 2018, Souza et al. 2019).

The target-enrichment strategy involves the selection of sequences of interest from a whole genome DNA library. To capture them, appropriate hybridization methods based on magnetic beads or microarrays associated with specific probes are used. Theoretically, only DNA fragments containing entirely or partially known regions will be captured (enriched) and subsequently sequenced. Nevertheless, although this strategy has been applied to different plant genome projects (Zhou and Holliday 2012, Clarke et al. 2013, Dubose et al. 2013, Dasgupta et al. 2015), to the best of our knowledge, no study has reported detection of transgene insertions in a plant genome.

In this study, we describe an approach of targeted sequence enrichment coupled with SGS technology and bioinformatics to identify and subsequently analyze flanking sequences for multiple transgene insertions in the genome of the transgenic sugarcane event TUC 87-3RG as required by GM crop regulatory measures in Argentina.

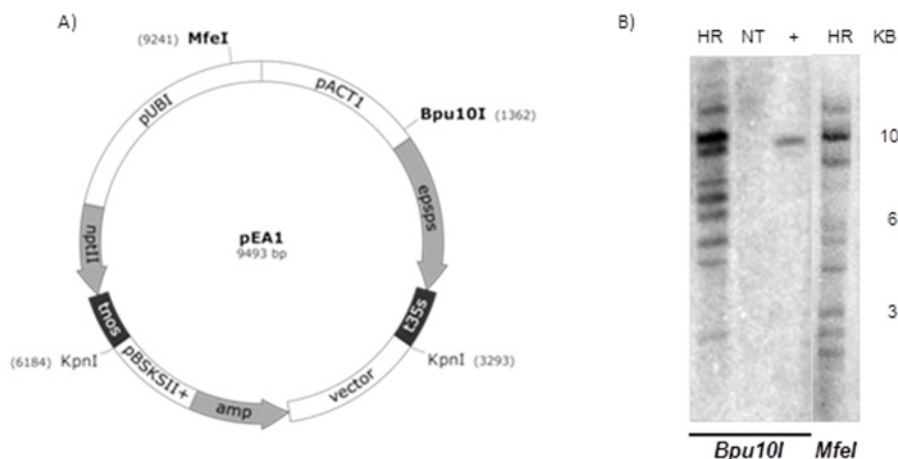
## MATERIAL AND METHODS

### Plant and reference materials

The HR sugarcane variety TUC 87-3RG was generated by transformation of the elite variety RA 87-3 with a gene construct (pEA1) containing the genes CP4 *epsps* for glyphosate resistance and *nptII* for kanamycin selection, as previously described by Noguera et al. (2015) (Figure 1A). The elite variety RA 87-3, the parental variety of TUC 87-3RG, was used as genetic control material. DNA and sequence of pEA1 plasmid were used as positive hybridization control in Southern blot analysis and as reference for the bioinformatic analyses.

### DNA extraction

Total genomic DNA for all sequencing libraries, PCR amplification, and Southern blot analyses were extracted according to the CTAB method described by Aljanabi et al. (1999) using ground young frozen leaves of the glyphosate-resistant line TUC 87-3RG and the parental variety RA 87-3 as control.



**Figure 1.** A) Circular map of the plasmid pEA1 containing the *KpnI* cassette used for biolistic transformation to create the transgenic line TUC 87-3RG. Expression elements for *epsps* and *nptII* genes are represented by arrows. B) Molecular characterization of TUC 87-3RG. Analysis of copy number in the HR transgenic line by means of Southern blot using *Bpu10I* and *MfeI* enzymes with *epsps* probe. (NT) The non-transformed RA 87-3 variety is used as a control. A dilution of pEA1 was used as a positive control (+).

### Southern blot analysis

A 428-bp *epsps* gene fragment was PCR amplified using the pEA1 plasmid as a template. This fragment was gel-purified and used as a DNA probe to hybridize with genomic DNA from both TUC 87-3RG and the non-transformed parental variety RA 87-3. DNA samples (approximately 20 µg per sample) and the pEA1 plasmid (10 pg), previously digested with restriction enzymes *MfeI* or *Bpu10I*, were separated by electrophoresis on a 0.8% agarose gel and subsequently transferred to a nylon membrane (Hybond<sup>+</sup>, Amersham). Baked membranes were hybridized with the *epsps* gene fragment, previously labeled with <sup>32</sup>P-dCTP using the Prime-a-Gene Labeling System kit (Promega). Unincorporated nucleotides were removed by passing the reaction solution through a G-25 micro-column (GE Healthcare). After overnight hybridization at 65 °C, membranes were washed three times with different concentrations of a saline sodium citrate (SSC) buffer (2X, 1X, and 0.5X) with 0.1% sodium dodecyl sulphate (SDS), exposed to a phosphor screen (GE Healthcare), and scanned after 24 h of exposure using a PhosphorImager Storm 845 (GE Healthcare).

### Targeted sequence capture of pEA1 sugarcane DNA fragments

In order to obtain a sequence-enriched genomic library from TUC 87-3RG, the solution-based capture method SeqCap EZ Library (Roche NimbleGen, Inc., Madison, WI, USA) was used. In collaboration with NimbleGen, a custom probe set for plasmid pEA1 was developed that allowed genome regions containing hybrid sequences of pEA1 and sugarcane genome DNA to be captured. One mg of each DNA sample was sheared by nebulization to generate 1.5-kb-size fragments using the Invitrogen Nebulizer kit (Invitrogen, Carlsbad, CA, USA). A sequence-enriched library was obtained by hybridization and capture of DNA samples with the custom probe set according to manufacturer’s guidelines. The size and mass of the enriched genome library were assessed using an Agilent DNA 7500 chip and a Bioanalyzer 2100 (Agilent) (Supplemental Figure 1). Quantitative PCR was used to validate the functionality of the library by using the KAPA Library Quantification Kit (Roche).

### Enriched library sequencing

The RAPID GS-FLX Titanium Library Preparation Kit was used to prepare 454 DNA libraries from sequence-enriched genome samples; libraries were individually amplified by emPCR (emulsionPCR) using the 454 Titanium kit, according to the manufacturer’s specifications and default parameters. Libraries were sequenced using the GS-FLX Titanium 454 Sequencing platform (*Servicios Genómicos*, Langebio-Cinvestav, Guanajuato, Mexico).

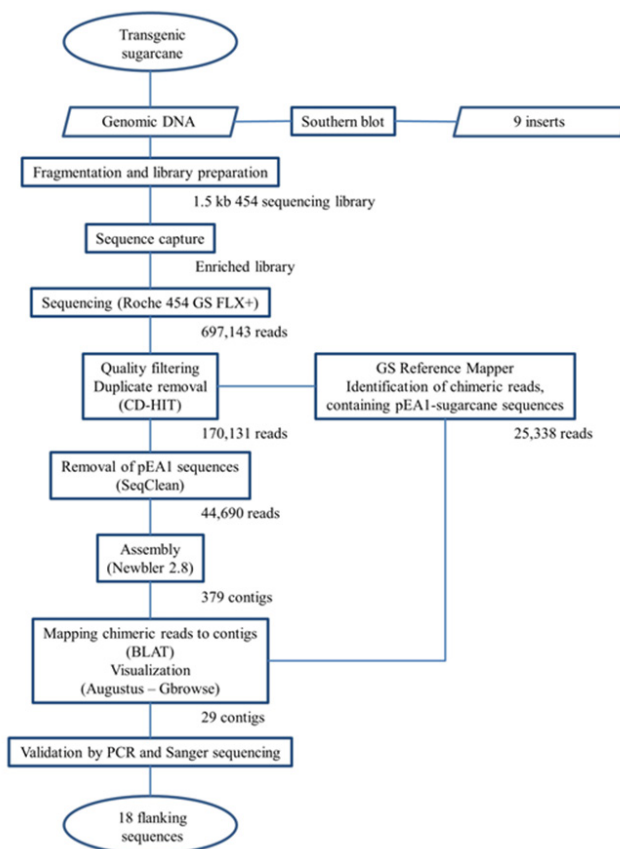
## Bioinformatics analysis

The bioinformatics analysis developed and described here is aimed at accurately detecting and characterizing chimeric sequences resulting from pEA1 insertions in the sugarcane genome, as shown in Figure 2.

First, using the CDHIT program (Niu et al. 2010), natural and artificial duplicate reads were removed from the data set generated by the GS-FLX Titanium sequencer. Additionally, reads with average quality of less than 20 (phred score) were discarded. With the aim of filtering out contaminating sequences, the remaining reads were compared against the human genome and a total of 5,153 distinct microbial genomes available in the RefSeq database (<ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/>). SeqClean (<http://www.tigr.org/tdb/tgi/software/>) was used to remove the sequences of the pEA1 vector from the data collection and to remove the vector from those pEA1 sugarcane chimeric reads. Filtered reads were mutually aligned and assembled using Newbler (version 2.8, Roche), and the resulting contigs were used as references to identify pEA1 sugarcane chimeric reads. First, filtered and contamination-free reads were mapped against the pEA1 vector using the gsMapper (module of Newbler), and then, using a custom perl program, the read identifiers that were partially aligned were pulled out from the 454ReadStatus.txt accessory file created by Newbler. All partially aligned reads were used to create “hints” by the BLAT mapping tool (Kent 2002). Finally, Gbrowse was used to get a visual representation of the alignments, and the putative flanking regions were manually characterized, as described in the Results and Discussion section. Alternatively, a visual representation was obtained by using other tools like WebGBrowse (Podcheti et al. 2009) or the AUGUSTUS web tool ([bioinf.uni-greifswald.de/augustus/submission.php](http://bioinf.uni-greifswald.de/augustus/submission.php)) (Figure 3).

## Locus-specific PCR and Sanger sequencing

Specific primers were designed that corresponded to the DNA insert and the flanking genomic DNA, based on the sequence determined from the transformation plasmid and from the genomic DNA of candidate contigs obtained from the bioinformatics analyses (Figure 4). Each fragment was amplified by PCR using a forward primer and corresponding reverse primer (data not shown) to detect chimeric DNA corresponding to junction sequences. PCR amplification of each fragment was performed using 30 cycles according to the following conditions: denaturation at 95 °C for 60 s, annealing at the specific temperature for the respective amplification reaction for 45 s, and extension at 72 °C for 60 s. Amplification products were analyzed in 1.5 % (w/v) agarose gels stained with Gel Red (Biotium, USA). Before the PCR products were sequencing, they were purified using the Pure Link Quick Purification Kit (Invitrogen, Germany) according to the manufacturer’s protocols. Purified reactions were sequenced by the Biotechnology Institute Sequencing Service, INTA Castelar, using an ABI 3130 Capillary DNA analyzer (Applied Biosystems, USA). All sequences were aligned against the NCBI database and published sugarcane genomes using the BLAST (Basic Local Alignment and Search Tool) algorithm.



**Figure 2.** Overall workflow for analyses of flanking sequences for inserts in the genome of transgenic sugarcane TUC 87-3RG.

## RESULTS AND DISCUSSION

### Southern blot analyses

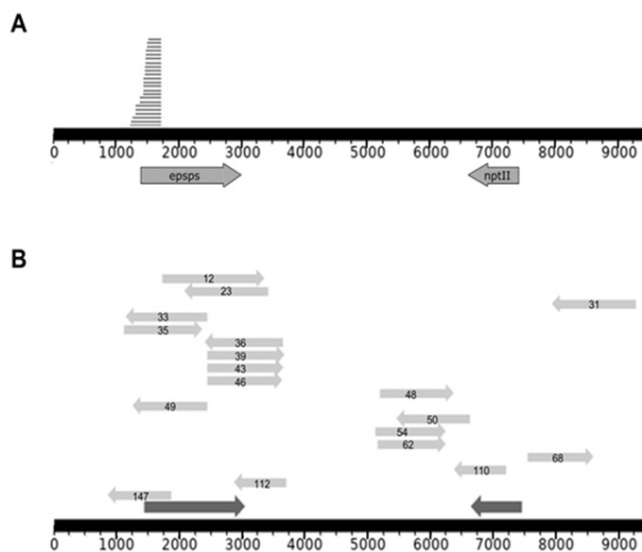
In order to study the number of CP4 *epsps* transgene insertions in the plant genome of the transgenic sugarcane variety TUC 87-3RG, Southern blot analyses were performed using the restriction enzymes *Bpu10I* and *MfeI*, which cut once in the pEA1 plasmid (Figure 1A). From the hybridization pattern of Southern blot analysis, it was evident that the TUC 87-3RG event contained multiple inserts (9) when using the *epsps* gene (Figure 1B) as a probe. Major technical concerns when evaluating TUC 87-3RG were the high number of transgene insertions and how to identify all insertion sites and their flanking sequences. Previous methods used for determining transgene flanking sequences were all based on PCR methods of genomes with at least one reference sequence and small numbers of inserts. As no reference genome was available for sugarcane at the time of analysis, a novel approach was needed, considering the vast complexity and size of the sugarcane genome.

### Sequence capture enrichment

To capture regions from the genome of TUC 87-3RG that contained sequences of the transgene, including hybrid pEA1 DNA sequences or only sequences from the transformation vector, 2.1 million riboprobes of 50-base size, designed by NimbleGen, were used. Genomic DNA isolated from TUC 87-3RG was sheared in 1.5-kb fragments and hybridized to the specific riboprobes to generate a sub-library of fragments containing the pEA1 sequence. All sub-libraries were quantified and characterized prior to sequencing (Supplemental Figure 1) in an Agilent 2100 Bioanalyzer. Typically, a 900-2000 bp broad library peak was observed, which peaked at around 1385 bp. The Sequence Capture method by NimbleGen allowed us to generate a plasmid sequence-enriched library from TUC 87-3RG of 1.5-kb DNA fragments containing pure transgene sequence and hybrid sequences of transgene-genomic DNA.

### 454-sequencing of captured libraries and junction sequence bioinformatics analysis

The 454-sequencing run of captured libraries resulted in a total of 697,143 reads (324,635,840 total bp) with an average length of 466 bp. These reads were filtered by quality, duplicates were removed, and reads matching human or microbial genomes were eliminated, generating a “clean” dataset with 170,131 reads. Next, 25,338 pEA1 sugarcane chimeric reads were obtained by mapping this dataset to the vector sequence with the Newbler software. The clean dataset was further filtered by removing reads containing pEA1 sequence with SeqClean (Chen et al. 2007), generating 44,690 reads. The remaining reads after filtering were then assembled into contigs with Newbler, which were used as a reference to identify pEA1 sugarcane chimeric reads. Even though a total of 379 contigs were generated, chimeric reads were mapped to only a few of them. Further curating of contigs for possible pEA1 flanking sequences produced 29 contigs containing a possible chimeric sequence. As expected, the pEA1 sugarcane chimeric sequence only mapped to one extreme of each contig, and coverage over the sequence decreased from this extreme to the opposite extreme (Figure 3A). Results observed from the 454-sequencing did not correspond to the expected parameters of a classic run using the FLX + platform (454 Sequencing System Software Manual, v 2.8; General Overview and Data File Format; 454 Life Sciences Corp. 2012), mainly due to the nature of the sample (enrichment and subsequent amplification of very small target regions, 1.5 kb), which was expected to have a highly repeated amplicon-like behavior known to generate a high light signal, as well as a high consumption of nucleotides. As a result, a lower average length of reads is observed



**Figure 3.** Analysis of contigs obtained from 454 sequencing. A. “junction reads” mapping for both contigs and pEA1 plasmid were mapped on pEA1 to identify the insertion site using Augustus software. The figure shows junction reads from contig 12. B. Map of the pEA1 plasmid showing the *epsps* and *nptII* genes (dark arrows). 18 selected contigs are shown above the plasmid, indicating the insertion site.



and the percentage of mixtures increased (Quince et al. 2011). Furthermore, the pearl load was reduced from 2,000,000 pearls per region to 1,650,000 to attempt to prevent cross-talking.

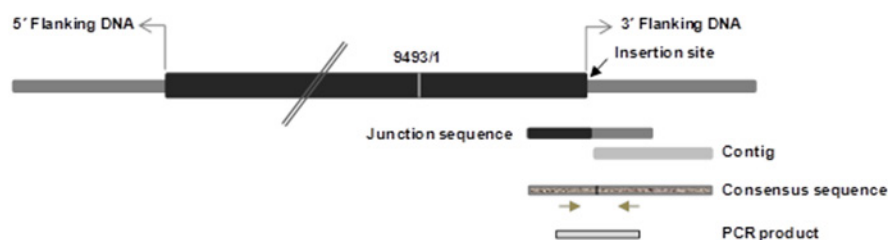
After bioinformatic analyses were conducted, 29 contigs were assembled as junction sequence candidates. The TUC 87-3RG event has 9 inserts as determined by Southern blot assays; therefore, at least 18 transgene flanking sequences were expected. Two contigs were discarded, as they contained partial pEA1 sequences, which might have arisen from defective trimming with SeqClean and might be misassemblies. From the remaining contigs, the 18 with best coverage in the vector-sugarcane junction were selected. Biolistic transformations have been shown to produce varied genome sequence disruptions, including insertions of smaller fragments or even genome rearrangements (Liu et al. 2019). Southern blot analyses would only recognize the disruptions where a full or even partial probe would hybridize, in this case the 428-bp *epsps* gene fragment, but might miss other insertions of a partial pEA1 sequence. The targeted sequence capture methodology would recover sequences from the whole 9.5-kb plasmid and identify plasmid-sugarcane junctions, arising even from partial or complex insertions not containing the probes used for Southern blot.

Thus, our sequence data could indicate the presence of additional junctions and partial transgene insertions not detected by the Southern blot analyses conducted.

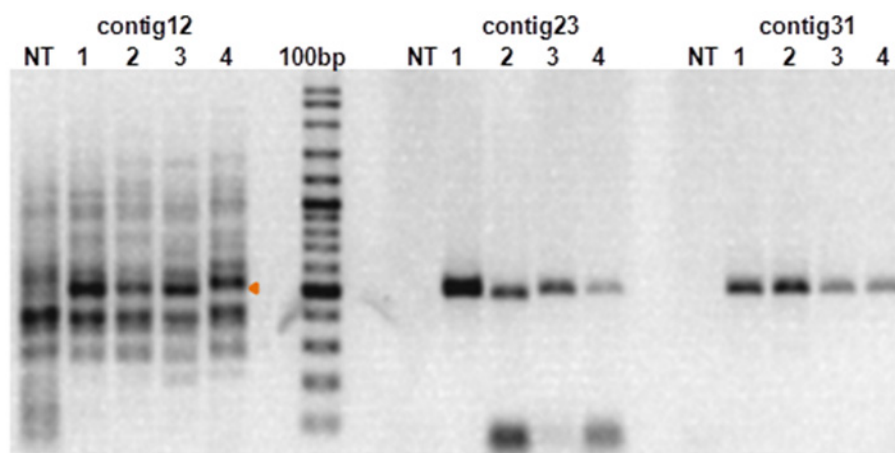
### Molecular characterization of flanking sequences in sugarcane genomes by PCR and sequencing

In order to verify that the aforementioned contigs indeed contained junction sequences of the transgene in the TUC 87-3RG genome, this region was amplified by conventional PCR. Reads associated with each contig were mapped against the sequence of the vector pEA1 and visualized using the AUGUSTUS software ([bioinf.uni-greifswald.de/augustus/submission.php](http://bioinf.uni-greifswald.de/augustus/submission.php)). Based on this analysis, 18 different contigs with coverage of the junction region were chosen for PCR amplification. The transgene cut sites were identified in each of the cases, as well as the 5' or 3' position relative to the insertion. For a summary of the results, see Figure 3B.

Identification of the potential transgene integration sites in the TUC 87-3RG genome allowed us to design primers for PCR analysis, which are anchored on both sides of the integration site, one within the transgene and the other at the sugarcane genome (Figure 4). According to the design, we would expect to observe a single band from the PCR products of transgenic sugarcane plants and no band from non-transformed parental line RA 87-3 plants. Consistent with our expectations, we were only able to observe distinctive PCR bands with the predicted size (Figure 5) for the majority of candidate contigs. However, in a few cases (contig12), we observed multiple bands in both DNA samples from TUC 87-3RG and control plants. In spite of these polymorphisms, the band of expected size was only present in DNA isolated from the transgenic variety. We attribute this phenomenon to the lack of specificity of some primers and the complexity of the sugarcane genome (Garsmeur et al. 2018). Moreover, sequences of amplified PCR products were compared to the proposed contigs and corroborated their identity. As we expected, a portion of the sequence matched the predicted inserted vector, and the remaining sequence came from the HR sugarcane genome. These flanking sequences for the inserts did not match essential genes when aligned against sugarcane genomes using BLASTn tools, and they did not generate new open reading frame (ORF) of potential allergens or toxic proteins in the genome of TUC 87-3RG (data not



**Figure 4.** Schematic diagram of the steps in designing a primer set for analysis of junction sequences. A representative hybrid read, pEA1 plant genome, specific for each candidate contig is mapped against pEA1 to determine the insertion site, by means of Augustus software. A consensus sequence that spanned the junction sequence and contig is obtained with Seqman Lasergene software and used to design the specific primers (grey arrows).



**Figure 5.** Molecular validation of flanking sequences in sugarcane genomes by PCR. A single band with the expected size was observed for PCR amplification of junction sequences corresponding to contigs 23 and 31. No band was obtained in the control plant (NT). (1-4) TUC 87-3RG event (under different PCR conditions). The yellow arrow shows the right band with expected size in samples with polymorphisms. The expected band is absent in the DNA sample from the control plant (NT).

shown). In fact, the majority of flanking sequences did not match any published sugarcane genome, and of the remaining sequences, matches were mostly in repetitive regions or intergenic regions, and a few in hypothetical proteins.

It is important to note that when this study was carried out, no better technology than the 454 FLX+ Roche sequencing platform was available to generate longer reads. In addition, as already pointed out, there was no complete reference genome for sugarcane where mapping of the junction reads could have been employed. The assembly of such complex genomes is still a very difficult task (Thirugnanasambandam et al. 2018), although Garsmeur et al. (2018) reported the first BAC-based monoploid genome sequence of sugarcane. After that, an allele-defined chromosome-level assembly of an ancestral (*Saccharum spontaneum*) (Zhang et al. 2018) and a gene-space assembly for a sugarcane variety (SP80-3280) (Souza et al. 2019) were published. In the future, these assemblies might provide essential genome templates for aligning sequencing data, such as genotyping-by-sequencing, WGS, and RNA-Seq data to explore hom(oe)ologous allelic variation and perform genetic (e.g., QTL and GWAS) and genomic studies in cultivars and sugarcane germplasm. As the field of genomics evolves, there is growing awareness in the scientific community of the importance in obtaining long-read data. In this sense, third-generation sequencing systems such as PacBio (Pacific Biosciences, USA) and MinION (Oxford Nanopore Technologies, UK) allow for DNA sequencing, and they achieve long reads with uniform coverage. Long sequence reads of ~ 15 Kb in size improve mappability for resequencing and simplify *de novo* assembly traits of great importance in these kinds of studies. Contigs generated by the assembly of reads from these platforms could span longer regions of the genome, including genome flanking sequences from either side of a transgene, thereby facilitating the study of foreign DNA insertions in a transgenic genome.

In conclusion, we were able to identify the junction sequences of the nine transgene insertions found in the TUC 87-3RG genome by using targeted sequence enrichment coupled with SGS and bioinformatic analysis, responding to the requirements of GM crop regulation in Argentina. These sequences are the hallmark to identify TUC 87-3RG from other transgenic events and sugarcane cultivars.

## REFERENCES

- Aljanabi S, Forget L and Dookun A (1999) An improved and rapid protocol for the isolation of polysaccharide- and polyphenol-free sugarcane DNA. *Plant Molecular Biology Reporter* **17**: 1-8.
- Carneiro MS, Chapola RG, Fernandes Junior AR, Cursi DE, Balsalobre TWA and Hoffmann PH (2020) RB985476-a sugarcane cultivar with high agro-industrial yield and disease resistance RB985476-a sugarcane cultivar with high agro-industrial yield and disease resistance. *Crop Breeding and Applied Biotechnology* **20**: e304020210.
- Chen YA, Lin CC, Wang CD, Wu HB and Hwang PI (2007) An optimized procedure greatly improves EST vector contamination removal. *BMC Genomics* **8**: 416.
- Clarke WE, Parkin IA, Gajardo HA, Gerhardt DJ, Higgins E, Sidebottom

- C, Sharpe AG, Snowdon RJ, Federico ML and Iniguez-Luy FL (2013) Genomic DNA enrichment using sequence capture microarrays: a novel approach to discover sequence nucleotide polymorphisms (SNP) in *Brassica napus* L. **PLoS One** **8**: e81992.
- Dasgupta MG, Dharanishanthi V, Agarwal I and Krutovsky K V (2015) Development of genetic markers in *Eucalyptus* species by target enrichment and exome sequencing. **PLoS One** **10**: e0116528.
- Dubose AJ, Lichtenstein ST, Narisu N, Bonnycastle LL, Swift AJ, Chines PS and Collins FS (2013) Use of microarray hybrid capture and next-generation sequencing to identify the anatomy of a transgene. **Nucleic Acids Research** **41**: e70.
- Garsmeur O, Droc G, Antonise R, Grimwood J, Potier B, Aitken K, Jenkins J, Martin G, Charron C, Hervouet C, Costet L, Yahiaoui N, Healey A, Sims D, Cherukuri Y, Sreedasyam A, Kilian A, Chan A, Van Sluys M-A, Swaminathan K, Town C, Bergès H, Simmons B, Glaszmann JC, van der Vossen E, Henry R, Schmutz J and D'Hont A (2018) A mosaic monoploid reference sequence for the highly complex genome of sugarcane. **Nature Communications** **9**: 2638.
- Kent WJ (2002) BLAT – The blast-like alignment tool. **Genome Research** **12**: 656-664.
- Kovalic D, Garnaat C, Guo L, Yan Y, Groat J, Silvanovich A, Ralston L, Huang M, Tian Q, Christian A, Cheikh N, Hjelle J, Padgett S and Bannon G (2012) The use of next generation sequencing and junction sequence analysis bioinformatics to achieve molecular characterization of crops improved through modern biotechnology. **Plant Genome** **5**: 149-163.
- Li F, Fu C and Li Q (2019) A simple genome walking strategy to isolate unknown genomic regions using long primer and RAPD primer. **Iranian Journal of Biotechnology** **17**: 89-93.
- Liu J, Nannas NJ, Fu F-F, Shi J, Aspinwall B, Parrott WA and Dawe RK (2019) Genome-scale sequence disruption following biolistic transformation in rice and maize. **The Plant cell** **31**: 368-383.
- Niu B, Fu L, Sun S and Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. **BMC Bioinformatics** **11**: 187.
- Noguera A, Enrique R, Perera MF, Ostengo S, Racedo J, Costilla D, Zossi S, Cuenya MI, Filippone MP, Welin B and Castagnaro AP (2015) Genetic characterization and field evaluation to recover parental phenotype in transgenic sugarcane: a step toward commercial release. **Molecular Breeding** **35**: 115.
- Podecheti R, Gollapudi R and Dong Q (2009) WebGBrowse--a web server for GBrowse. **Bioinformatics** **25**: 1550-1551.
- Quince C, Lanzen A, Davenport RJ and Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. **BMC Bioinformatics** **12**: 38.
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X-C, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC and Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. **Nature** **463**: 178-183.
- Souza GM, Van Sluys MA, Lembke CG, Lee H, Margarido GRA, Hotta CT, Gaiarsa JW, Diniz AL, Oliveira MDM, Ferreira SDS, Nishiyama MY, Ten-Caten F, Ragagnin GT, Andrade PDM, De Souza RF, Nicastro GG, Pandya R, Kim C, Guo H, Durham AM, Carneiro MS, Zhang J, Zhang X, Zhang Q, Ming R, Schatz MC, Davidson B, Paterson AH and Heckerman D (2019) Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. **GigaScience** **8**: 1-18.
- Thirugnanasambandam PP, Hoang NV and Henry RJ (2018) The challenge of analyzing the sugarcane genome. **Frontiers in Plant Science** **9**: 616.
- Yang L, Wang C, Holst-Jensen A, Morisset D, Lin Y and Zhang D (2013) Characterization of GM events by insert knowledge adapted re-sequencing approaches. **Scientific Reports** **3**: 2839.
- Zhang J, Zhang Xingtang, Tang H, Zhang Q, Hua X, Ma X, Zhu F, Jones T, Zhu X, Bowers J, Wai CM, Zheng C, Shi Y, Chen S, Xu X, Yue J, Nelson DR, Huang L, Li Z, Xu H, Zhou D, Wang Y, Hu W, Lin Jishan, Deng Y, Pandey N, Mancini M, Zerpa D, Nguyen JK, Wang L, Yu L, Xin Y, Ge L, Arro J, Han JO, Chakrabarty S, Pushko M, Zhang W, Ma Yanhong, Ma P, Lv M, Chen F, Zheng G, Xu J, Yang Z, Deng F, Chen X, Liao Z, Zhang Xunxiao, Lin Z, Lin H, Yan H, Kuang Z, Zhong W, Liang P, Wang Guofeng, Yuan Y, Shi J, Hou J, Lin Jingxian, Jin J, Cao P, Shen Q, Jiang Q, Zhou P, Ma Yaying, Zhang Xiaodan, Xu R, Liu J, Zhou Y, Jia H, Ma Q, Qi R, Zhang Z, Fang J, Fang H, Song J, Wang M, Dong G, Wang Gang, Chen Z, Ma T, Liu H, Dhungana SR, Huss SE, Yang X, Sharma A, Trujillo JH, Martinez MC, Hudson M, Riascos JJ, Schuler M, Chen LQ, Braun DM, Li L, Yu Q, Wang J, Wang K, Schatz MC, Heckerman D, Van Sluys MA, Souza GM, Moore PH, Sankoff D, VanBuren R, Paterson AH, Nagai C and Ming R (2018) Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. **Nature Genetics** **50**: 1565-1573.
- Zhou L and Holliday JA (2012) Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. **BMC Genomics** **13**: 703.