TDL.ORG TEXAS DIGITAL LIBRARY

Journal of Regulatory Science

http:\\journalofregulatoryscience.org

Regulatory Science

Journal of Regulatory Science 6(1) (2018) 1–14

Evaluation of Whole Genome Sequencing and an Insertion Site Characterization Method for Molecular Characterization of GM Maize

Rebecca Cade[a,*], Kristina Burgin[a], Kelly Schilling[b], Tae-Jin Lee[a], Peter Ngam[b], Nico Devitt[b], Diego Fajardo[b]

[a]*Syngenta Crop Protection, LLC, 9 Davis Drive, Research Triangle Park, NC 27709*
[b]*National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, NM 87505*

Abstract

Molecular characterization of genetically modified (GM) crops has traditionally been accomplished through a combination of Southern blot, polymerase chain reaction (PCR), and Sanger sequencing analyses. This characterization generates critical information that is used in downstream safety assessment of GM crops and development of GM detection methods. Next generation sequencing (NGS) technologies, such as whole genome sequencing (WGS), have shown the potential to replace some or all of these techniques for molecular characterization of GM crops. This paper describes the characterization of two GM maize events using NGS for WGS in combination with an insertion site characterization (ISC) method. The sensitivity of the method is also compared to that of Southern blot analysis through detection of small insert fragments. Our results demonstrate that WGS is at least as sensitive as Southern blot analysis for determining the insert copy number, presence or absence unintended insertions, and for characterization of small fragment insertions. These results support the conclusion that WGS along with an appropriate insertion site characterization method are a suitable alternative to Southern blot analyses for molecular characterization of GM maize.

Keywords: genetically modified crop, insertion site characterization, molecular characterization, molecular characterization, next generation sequencing, whole genome sequencing, Southern blot

Abbreviations: GM, genetically modified; PCR, polymerase chain reaction; NGS, Next generation sequencing; WGS, whole genome sequencing; ISC, insertion site characterization; KOGs, euKaryotic clusters of Orthologous Groups; PMI, phosphomannose isomerase; CEGMA, Core Eukaryotic Genes Mapping Approach; T-DNA, transferred DNA

1. Introduction

Detailed molecular characterization of GM crops is required by various regulatory agencies prior to product registration. Several groups have provided data requirements for molecular characterization of GM crops including: the European Food Safety Authority [7], the Food and Agricultural Organization [8], and the Organization for Economic Cooperation and Development [19]. Molecular characterization provides critical DNA sequence information that is required for downstream safety assessment of GM crops and has typically been performed using a combination of Southern blot, PCR, and Sanger sequencing technologies. These analyses are used to determine and describe the following for a transgenic event: the insert copy number, presence or absence of transformation plasmid backbone (DNA sequence outside of the transferred-DNA [T-DNA]), location and integrity of the insertion site in the plant genome, the exact sequence of the inserted DNA and flanking host DNA, and the genetic stability of the insert through multiple generations and across the breeding pedigree.

The use of NGS technologies has been shown to be a suitable alternative for addressing molecular characterization endpoints that have been traditionally addressed by Southern blot analysis (reviewed in [10]). NGS technologies, including WGS, can be effectively used to determine insert copy number through detection of unique junctions between a transgenic insert and the adjacent genomic sequences, and can also be used to detect unintended integrations such as transformation plasmid backbone. While multiple studies [9, 11, 31] have shown the potential applications of WGS for molecular characterization of GM crops, only one study has been performed with GM maize [31] and it employed target capture instead of WGS. Maize is more challenging compared to an inbred crop such as soybean, due to its large complex genome, variability between genotypes, and abundance of repetitive sequences [13, 17, 23, 25]. While

*Corresponding author: Rebecca Cade, Phone: 919-281-7258, Email: rebecca.cade@syngenta.com

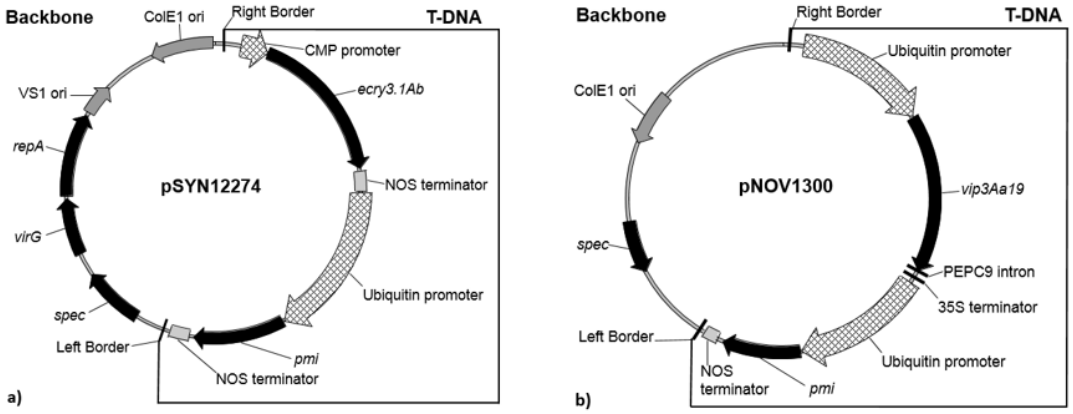


Figure 1: Transformation plasmid maps: a) map of pSYN12274, transformation plasmid used to generate event 5323, b) map of pNOV1300, transformation plasmid used to generate event MIR152.

Zastrow-Hayes et al. [31] demonstrated the ability to detect inserted transgene fragments as small as 35 bp with target capture techniques, the level of sensitivity has yet to be determined for WGS.

In this paper we characterize two GM maize events with WGS and a newly developed ISC method in order to address two questions: (1) Can WGS and the ISC method be used for molecular characterization of GM maize, and (2) is the sensitivity of WGS and the ISC method comparable to the sensitivity of Southern blot analysis? The ISC method developed for these analyses, while unique in some aspects (e.g. alignment tools), follows a design principle that has previously been demonstrated to be effective in analyzing T-DNA inserts in GM crops [11, 30], sequence alignment to the appropriate reference, assembly, and realignment for analysis.

The two maize events were analyzed by WGS and the ISC method to determine insert copy number, presence or absence of transformation plasmid backbone DNA, location and integrity of the insertion site, and insert sequence integrity. The results of these analyses were then compared to previous characterization results generated by Southern blot or Sanger sequencing analysis. To compare sensitivity of the WGS and the ISC method with the Southern blot analysis method, five small DNA fragments containing sequences from the transformation plasmid were synthesized and spiked into the GM maize DNA and attempted to be detected by both methods.

Here we demonstrate the utility of WGS and the ISC method, the ability to characterize T-DNA inserts in a complex genome, sensitivity of the technology as compared to the traditional technologies, and ultimately its suitability as a method for molecular characterization of GM crops.

2. Material and Methods

2.1. Plant Material

Two *Agrobacterium*-mediated transformation events, 5323 maize and MIR152 maize, previously characterized by Sanger sequencing and Southern blot analyses, were used in this study. Transformations were performed using the inbred JHAX maize (5323) and a B73 maize hybrid (MIR152) and transformation plasmids pSYN12274 (5323) and pNOV1300 (MIR152) (Figure 1). DNA from the transformation plasmids and JHAX maize were used as controls in Southern blot analyses, and the transformation plasmid sequences served as references for sequence analysis.

2.2. Genomic DNA Preparation

Seed of events 5323 and MIR152, and nontransgenic JHAX maize were germinated and grown in a greenhouse for four weeks. Leaves from ten plants from each line, confirmed to contain the transgenic insert by real-time PCR, were harvested on dry ice and stored at -80 °C prior to grinding and DNA isolation. High molecular weight genomic DNA for library preparation and subsequent Illumina sequencing was prepared using a modification of a method by Zhang et al. [32]. DNA for Southern blot analysis was prepared using a modification of a method by Murray and Thompson [18]. The DNA concentrations were measured using a Quant-iT^{TM} PicoGreen$^{®}$ dsDNA kit on a Turner Biosystems TBS-380 Mini-Fluorometer.

2.3. Small Fragment Preparation

A common sequence between the pSYN12274 and pNOV1300 transformation plasmids was selected to represent small unintended insert fragments in the 5323 and MIR152 maize genomes. Five DNA fragments containing partial sequences of varying sizes (25, 50, 100, 150, and 200 bp) from this common sequence, the phosphomannose isomerase (PMI) gene, were synthesized at GENEWIZ, LLC. (South Plainfield, NJ). Each PMI sequence was flanked by 500 bp from maize chromosome 5 to simulate an unintended partial insertion from the transformation plasmid, and cloned into pUC57 plasmids.

2.4. Library Preparation

Genomic and plasmid DNA libraries were prepared using Kapa HTP DNA Library Preparation kit (Kapa Biosystems, Wilmington, MA). The DNA was fragmented using a Covaris

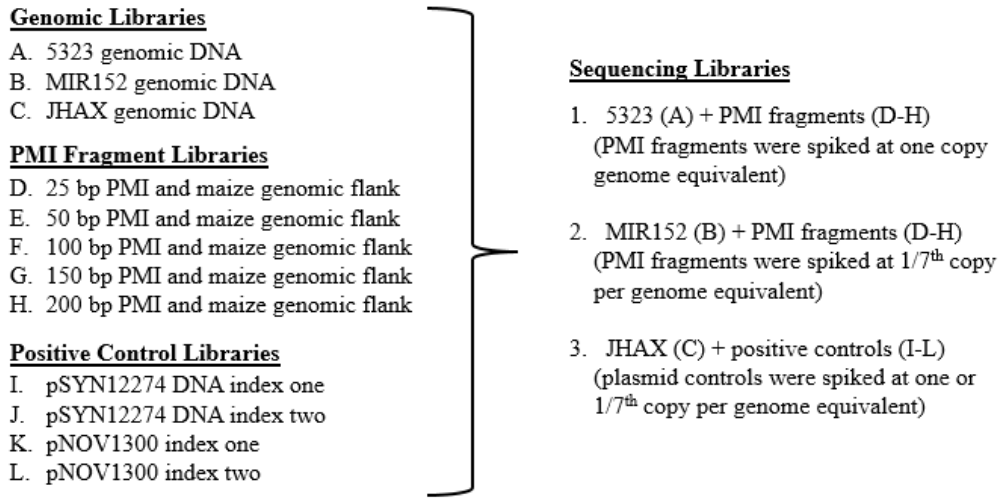


Figure 2: Outline of the twelve libraries generated and the subsequent combination strategy for sequencing. Three genomic DNA libraries were generated (A-C), to which PMI fragment (D-H) or positive control libraries (I-L) were spiked. Two indexed libraries were generated for each of the plasmid vector positive controls so that both the one copy and one-seventh copy representations could be spiked into a single JHAX library and subsequently differentiated (3).

S2 300 watt Focused-ultrasonicator (Covaris, Inc., Woburn, MA) with 5% Duty Factor, 200 cycles per burst, and 60 seconds treatment. The sheared, double-stranded DNA fragments were first end-repaired, A-tailed, and ligated to TruSeq index adapters (Illumina, San Diego, CA). Then, DNA library fragments were enriched by PCR according to the manufacturers protocol (KAPA HTP Library Preparation Kit). Size selection was performed to 600 bp using the Blue Pippin Instrument (Sage Science, Beverly, MA). The final library was validated using a Bioanalyzer 2100 7500 DNA chip (Agilent Technologies, Santa Clara, CA) and Nanodrop™ spectrophotometer (Life Technologies, Austin, TX).

Twelve indexed, paired-end libraries were generated; three maize genomic libraries (5323, MIR152, and JHAX), five plasmid libraries from the fragments containing the PMI sequences (25, 50, 100, 150, and 200 bp), and four positive control plasmid libraries (Figure 2).

2.5. Sequencing Library Preparation

To simulate insertion of unintended small fragments, the five PMI-containing fragment libraries were spiked into the 5323 genomic DNA library at one copy per genome equivalent and into the MIR152 genomic DNA library at one-seventh copy per genome equivalent. This was determined by weight and calculated using the following formula [1] and based on a maize genome size of 2.67 x 10^9 bp.

$$\left\{\left(\frac{\text{positive assay control size (bp)}}{\text{genome size (bp)} \times \text{ploidy}}\right) \times \mu\text{g loaded}\right\} \times 1 \times 10^{6} = \text{pg for 1 copy}$$

The plasmid pSYN12274 and pNOV1300 libraries were spiked into the JHAX maize genomic DNA library at either one copy or one-seventh copy per genome equivalents as described above, and served as a positive control. A separate negative control library was not generated for JHAX since its genome had previously been sequenced to over 200-fold coverage (JHAX v4; unpublished). Following combination, the libraries were diluted to 10 nM and evaluated by quantitative polymerase chain reaction (qPCR) to ensure equal flow cell loading.

2.6. *Illumina*® Sequencing

Sequencing was performed using Illuminas HiSeq2000 (Illumina, San Diego, CA) instrument following the manufacturers protocol to produce paired-end sequence reads (2 x 100 bp). Initial quality control of the sequenced reads was performed using the CASAVA software (Illumina, Inc., San Diego, CA). Reads containing ≥20 bp of adapter sequence were discarded.

2.7. Genome Coverage Determination

Single and/or low copy eukaryotic orthologous genes (KOGs) in the JHAX draft maize reference genome were identified using CEGMA (Core Eukaryotic Genes Mapping Approach, v2.4; [20]), and a subset was retained as single copy based on relative coverage (BEDTools v2.25.0; [21]) (previous, unpublished analyses). Coordinates of the subset KOGs on the JHAX v4 and B73 v5 (unpublished) draft maize reference genomes were determined (Genomic Mapping and Alignment Program, v2016-06-09; [29]). Conservatively, only KOGs aligning to the references with >99% length and identity were retained for coverage calculations.

Alignments of the three genomic sequencing libraries to either the JHAX v4 (5323 and JHAX) or B73 v5 (MIR152)

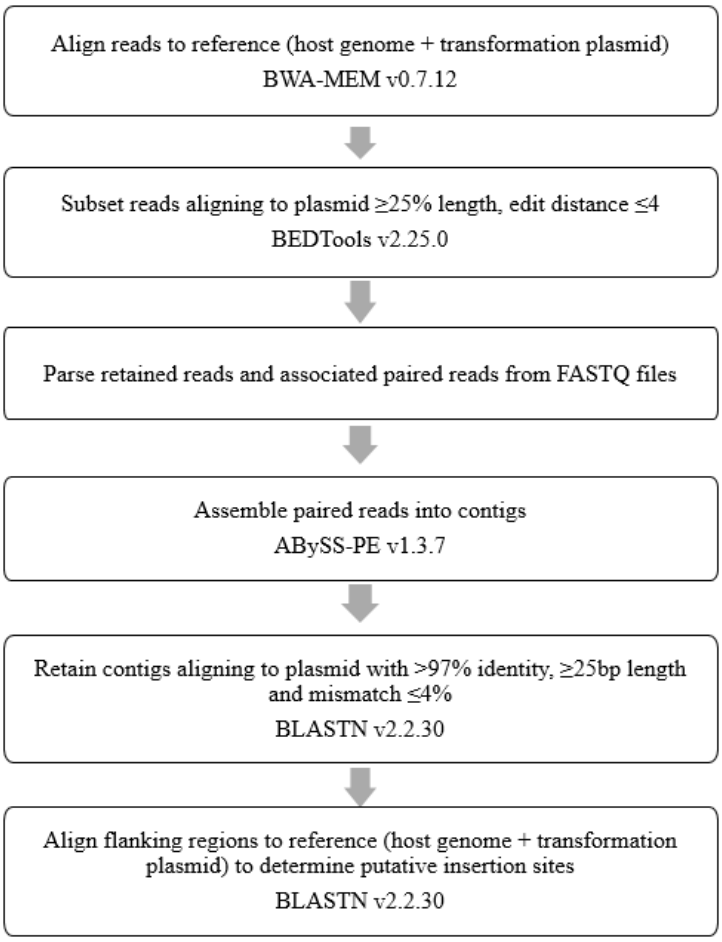


Figure 3: Flowchart summarizing the ISC method.

references were used to determine single-copy gene coverage. Reads aligning to coordinate intervals of the retained KOGs at ≥25% length of the read alignments (BEDTools) with edit distance ≤4 were retained. The BEDTools were used to calculate median coverage of each of the defined KOGs as well as the overall median coverage, considered the single copy coverage for the given reference.

2.8. Read Alignment and Assembly for Insert Site Characterization

Inputs required for the ISC method are: paired-end sequencing data from the transgenic line intended for characterization; a high-quality reference sequence of the host genome, or closely-related genome; and the reference sequence of the transformation plasmid used to generate the event.

Paired-end reads for a given sequencing library were pooled and aligned using BWA-MEM v0.7.12 [15] to a concatenated reference consisting of: the most closely related, well-characterized maize reference genome, the transformation plasmid, and the PMI fragments' pUC57 plasmid backbone. The 5323 sequencing library was aligned to the JHAX v4 maize genome, pSYN12274 transformation plasmid, and pUC57 plasmid backbone; the MIR152 sequencing library was aligned to the B73 v5 maize genome, pNOV1300 transformation plasmid, and pUC57 plasmid backbone. Pooled reads were aligned to a reference sequence containing all anticipated sources of read data in order to minimize alignment bias and therefore reduce bias during assembly, as well as coverage calculations. Reads were retained from the 5323 and MIR152 sequencing libraries aligning to the T-DNA plasmid reference at ≥25% length of the read alignment, using BEDTools with an edit distance of ≤4. FASTQ sequences of the retained reads and their associated paired reads were parsed from the original FASTQ files using a custom Perl script. Paired reads were assembled using ABySS-PE v1.3.7 [24]. Alignment rates were calculated for the initial BWA-MEM read alignments as well as for alignments of the filtered subsets of reads to the resultant ABySS contigs using SAMTools flagstat v1.3 [14].

Resultant ABySS contigs were aligned to their respective transformation plasmid using BLASTN v2.2.30 [3]. If a portion of a contig aligned to the transformation plasmid with >97% identity, ≥25 bp length, and ≤4% mismatches, the entire contig was retained for further analysis. Flanking regions of retained contigs were aligned to the JHAX v5 or B73 v5 reference using BLASTN to determine the putative insertion site(s) in the host genome (Figure 3).

Contigs containing junction sequences were then used to determine insert copy number, presence/absence of transformation plasmid backbone DNA, and further characterize the putative insertion sites. Contig subsequences aligning to the transformation plasmid T-DNA were also analyzed to determine integrity of the insert(s) sequence.

Portions of the contigs aligning to the transformation plasmid were also aligned to the host genome as a quality control step; if the contig T-DNA subsequence aligned to the host genome equally as well or better than to the transformation plasmid (based on identity, length and mismatch), it was considered a potential false positive.

2.9. Coverage of Putative Insertion(s)

In order to estimate coverage of the T-DNA inserts and PMI fragments using the ISC method, the filtered reads used in the ABySS assembly were realigned to the entire set of resultant ABySS contigs using BWA-MEM. Primary read alignments were used for coverage calculations. The BEDTools were used to calculate median coverage of each of the putative T-DNA insertions and PMI fragments. Genome equivalents were calculated by dividing the estimated coverage of the putative genomic insertion by the estimated single copy coverage of the host genome.

Coverage calculations were validated using indexed reads. A read overlap of 25% in the T-DNA/PMI region was required to mimic the ISC method. Coverage values were also calculated without the 25% overlap requirement; these values were considered the actual coverage of the T-DNA/PMI region. Reads associated with a given maize line or PMI fragment were parsed by index and aligned to the most closely-related reference using BWA-MEM. Reads aligning to the T-DNA or PMI fragment using BEDTools with an edit distance of ≤4, were retained. The BEDTools "coverage" and "group by" modules were used to calculate median coverage from primary read alignments of each of the putative T-DNA insertions and PMI fragments.

2.10. Presence of Vector Backbone

The JHAX sequencing library was spiked with libraries from plasmid pSYN12274 and plasmid pNOV1300 as positive controls to demonstrate that vector backbone could be detected at the various levels of incorporation. Sequencing library reads aligning to the pSYN12274 or pNOV1300 vector backbone regions were analyzed as potential vector backbone sequence. Plasmids pSYN12274 and pNOV1300 were aligned to JHAX v5, B73 v5, and/or pUC57 plasmid backbone using BLASTN to identify regions of common sequence.

2.11. Southern Blot Analysis

Southern blot analyses were performed using standard molecular biology techniques [5]. Each lane contained 7.5 μg of maize genomic DNA that was digested with the appropriate restriction enzyme(s). The pUC57 constructs containing the five PMI fragments flanked by maize genomic sequence were linearized by restriction digest prior to being analyzed by Southern blot analysis. Southern blot analysis samples included genomic DNA from the event of interest, nontransgenic maize genomic DNA, nontransgenic maize genomic DNA individually spiked with the five PMI-containing fragments, and positive assay controls. The samples were loaded onto 1% agarose gels and the DNA fragments were separated by electrophoresis in 1X tris-acetate-EDTA buffer. A radioactive probe containing PMI sequence was generated by PCR from the transformation plasmid, and labeled with alpha-phosphorus-32-deoxycytidine triphosphate ([α-32P]-dCTP) by random priming using the GE Healthcare Megaprime™ DNA labeling system. Membranes were washed and subjected to imaging with a Molecular Dynamics Storm™ 860 PhosphorImager™.

3. Results and Discussion

3.1. Insert Copy Number Determination

To determine the copy number and intactness of the inserted T-DNA using the ISC method, the number of unique junction sequences were determined and analyzed for each event. An insertion site in the host genome has a pair of genome-to-insert junction sequences; one on the 5′ end and one on the 3′ end of the inserted sequence. The junctions are chimeric sequences consisting of a maize genomic DNA and transformation plasmid DNA. Therefore, the BWA-MEM v0.7.12 [15] aligner was chosen for determination of insert copy number because of its ability to process chimeras by generating alignments for subsequences of each chimeric read to each applicable reference. This characteristic enabled parsing of reads that aligned partially (≥25bp) to the T-DNA plasmid that might otherwise be missed using a global aligner. The 25-bp overlap was required to increase the likelihood of a statistically significant alignment in the T-DNA/PMI region and reduce false positives. Soft clipping the read alignments allowed for entire read sequence to be retained for assembly across the putative junctions.

Assembly of reads from the 5323 sequencing library resulted in six putative junction-containing contigs that aligned to the pSYN12274 transformation plasmid. An additional contig was generated that aligned completely to the backbone region of pSYN12274 and was retained for further analysis of potential backbone presence. The longest of the seven contigs, with a putative length of 6376 bp, and originating from positions 74 bp to 6449 bp on the pSYN12274 plasmid reference, represented a full T-DNA insertion in the maize genome (Figure 4). The portion of this contig aligning to pSYN12274 realigned at 99.86% identity and the two junction sequences identified were identical to junction sequences identified previously by PCR and Sanger sequence analyses (data not shown).

The next five contigs represented each of the five PMI-containing fragments (25 bp to 200 bp) spiked into the sequencing library in order to determine sensitivity of detecting small fragments with this method. Alignments of these contigs with the PMI reference sequences can be found in Supplementary Figure 1.

The final contig was generated from 5323 sequencing library reads aligning to the pSYN12274 transformation plasmid backbone. The majority of these read alignments were due to

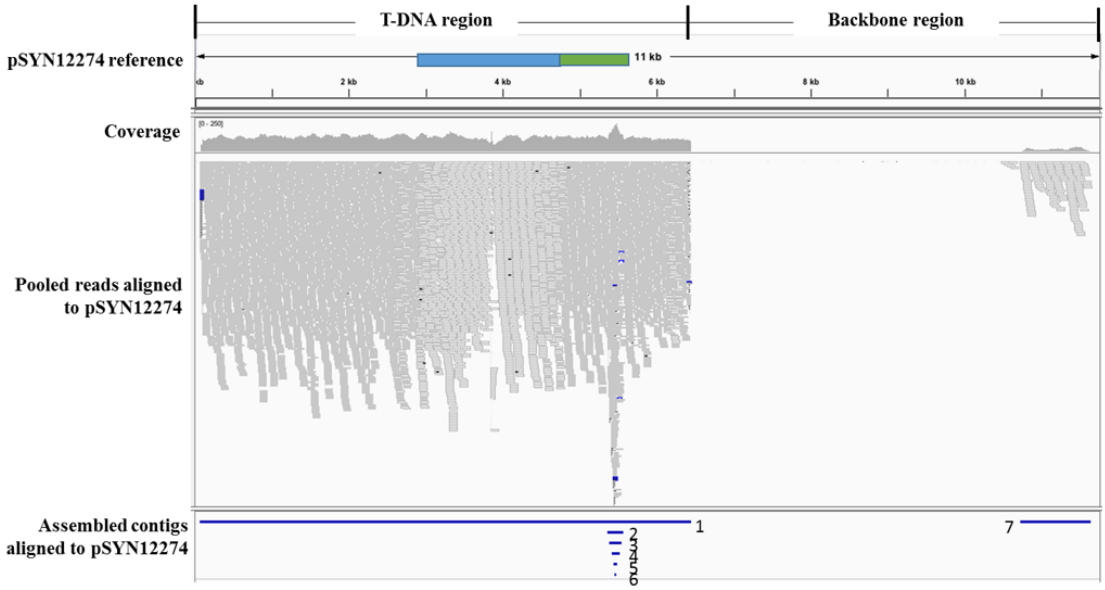


Figure 4: Generated using IGV [26]: Coverage depth of aligned reads (data range 0-250x). *Middle track:* Pooled reads from the 5323 sequencing library (including spiked-in PMI fragment plasmids) aligned to the pSYN12274 plasmid reference. (*Read shading: light = multimapped; medium = uniquely aligned; dark = chimeric*). Blue region of the pSYN12274 reference indicates Ubiquitin promoter and the green shaded region represents the PMI gene. Reads aligning to the backbone region have sequence in common between pSYN12274, JHAX v5 chromosome 7, and pUC57 (PMI fragment plasmid backbone). *Bottom track:* Assembled contigs including the 5323 insertion (1), PMI fragments (2-6), and backbone sequence aligned to the pSYN12274 plasmid reference (7).

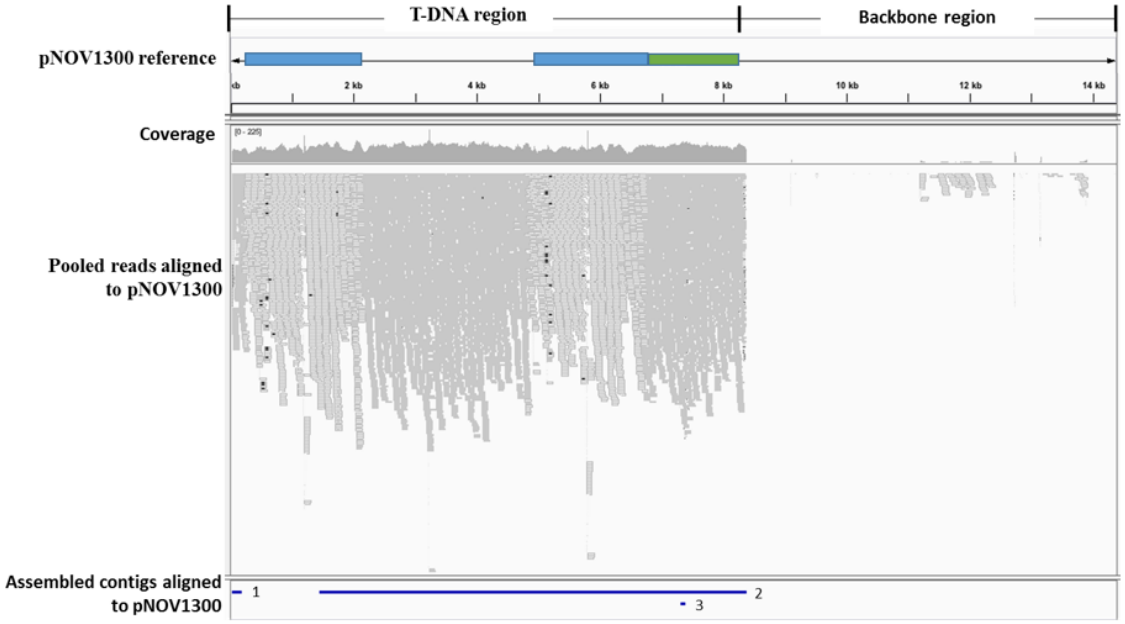


Figure 5: Generated using IGV [26]: *Top track:* Coverage depth of aligned reads (data range 0-225X). *Middle track:* Pooled reads from the MIR152 sequencing library (including spiked-in PMI fragment plasmids) aligned to the pNOV1300 plasmid reference. (*Read shading: light = multimapped; medium = uniquely aligned; dark = chimeric*). Blue regions in the pNOV1300 reference indicate Ubiquitin promoters and the green shaded region represents the PMI gene. Reads aligning to the backbone region have sequence in common between pNOV1300 and pUC57 (PMI fragment plasmid backbone). *Bottom track:* Assembled contigs including the MIR152 insertion (1) and (2) and 200-bp PMI fragment (3).

the sequence in common between pSYN12274 and the PMI fragments' pUC57 plasmid backbone.

Assembly of reads from the MIR152 sequencing library resulted in three putative junction contigs which aligned to the pNOV1300 transformation plasmid. The first two contigs (contigs 1 and 2), each containing one junction, represented the 5′ and 3′ junctions of a full T-DNA insertion into the maize genome (Figure 5). The putative length of the full MIR152 T-DNA insert is 8350 bp, with junction alignments indicating the insert originated from positions 33 bp to 8382 bp on the pNOV1300 transformation plasmid reference and inserted in the reverse (3′ to 5′) orientation. The portions of the MIR152 assembled contigs containing junction sequences realigned to pNOV1300 with ≥99.99% identity and the junction sequence identified was identical to the 5′ and 3′ junction sequences discovered through PCR and Sanger sequence analyses (data not shown).

The third contig contained one junction of the 200-bp PMI fragment spiked into the MIR152 sequencing library. The PMI fragments spiked into MIR152 were at much lower genome equivalents than had been intended, therefore this was the only fragment spiked into the MIR152 sequencing library that was detected.

The number of unique junctions obtained for both 5323 and MIR152, excluding the spiked PMI fragments, is consistent with data previously obtained through Southern blot analysis (data not shown). A single pair of insert-to-genome junction sequences was obtained for both events.

3.2. Determination of Backbone Presence/Absence

The presence or absence of transformation plasmid backbone DNA in a transgenic event genome can be evaluated by aligning WGS reads against the entire transformation plasmid reference, including the backbone region. Reads aligning to the plasmid backbone can originate from three sources: (1) backbone sequence that was incorporated during transformation, (2) sequence homologous to both the plasmid backbone and the host genome, and/or (3) bacterial contamination with sequence homologous to the plasmid backbone sequence.

The transformation plasmids pSYN12274 and pNOV1300 were used as positive controls to ensure method sensitivity for detecting backbone sequence. Using the ISC method it was determined that both the 5323 and MIR152 genomes were free of backbone sequence insertions. The filtered subset of 5323 pooled sequencing library reads contained 228 reads (3.1%) that aligned to pSYN12274 in the backbone region (Figure 4). Likewise, 251 reads (2.8%) of the filtered subset of MIR152 sequencing library reads aligned to pNOV1300 in the backbone region (Figure 5). This was expected due to the contribution from reads originating from the PMI fragments' pUC57 plasmid backbone spiked in both sequencing libraries.

Bacterial plasmids and other sources of contamination introduced in the laboratory can be a challenge due to the sensitivity of NGS technologies [31, 16]. Contamination can be minimized by using good laboratory practices and preparing DNA and sequencing libraries in dedicated clean room environments [11] but it is difficult to eliminate. To determine if sequence reads are false positives caused by contamination or endogenous sequence similarity they can be assessed for level of coverage relative to single copy genes, compared to the non-GM reference sequence, and screened against known contamination sources.

Less than 0.2% of the indexed 5323 genomic sequencing library reads aligned to a ∼1-kb region of the pSYN12274 backbone sequence. The 1X median coverage level of these reads was much lower than that of the expected one copy per genome equivalent coverage level for the 5323 sequencing library (105X), which would have been expected for a true backbone sequence insertion. A BLASTN alignment demonstrated that there is a simliar sequence in the JHAX v5 reference genome on chromosome 7.

The conclusions drawn from WGS and the ISC method (no vector backbone presence) were identical to the conclusion drawn from previous Southern blot studies (data not shown). Therefore, these two methods are functionally equivalent methods for determination of presence or absence of backbone sequence.

3.3. Chromosomal Location

Determination of the genomic DNA sequence flanking the insertion site is necessary for identifying the chromosomal location of the insert and for the development of event-specific PCR detection assays necessary for commercialization of a GM event. This analysis is typically achieved through genome walking PCR methods [2]. Genome walking can be labor intensive and often fails to characterize flanking DNA if the restriction enzyme sites necessary to perform the method are not present in close proximity to the insert or if the primer target sequence is missing in the transformation event because of truncation or rearrangement.

The genomic flanking sequence obtained during ISC allows for the insert's chromosomal location to be determined when enough genomic sequence is obtained. Paired-end read data were used for ABySS assembly in order to assemble long contiguous sequences that would not have been obtained by single-end reads. The additional flanking sequence information that can be obtained by this method is beneficial when other methods for obtaining flanking sequence have been unsuccessful. The flanking sequence that was identified using WGS and the ISC method was compared to the host reference genome to identify putative insertion sites. For 5323, genomic flanking sequences of 423 bp at 5′ end and 508 bp at 3′ end were obtained; these sequences aligned to chromosome 5 in the JHAX v5 draft reference genome at 100% identity. The putative insertion site for the 5323 full T-DNA insert was determined to be located on JHAX v5 chromosome 5 (Figure 6a). In addition, it was determined that 56 bp of genomic sequence was deleted at the insertion site.

For MIR152, genomic flanking sequences of 89 bp at 5′ end and 279 bp at 3′ end were obtained; these sequences aligned to chromosome 1 in the B73 v5 reference genome at 100% identity. The 279 bp of 3′ sequence obtained was previously

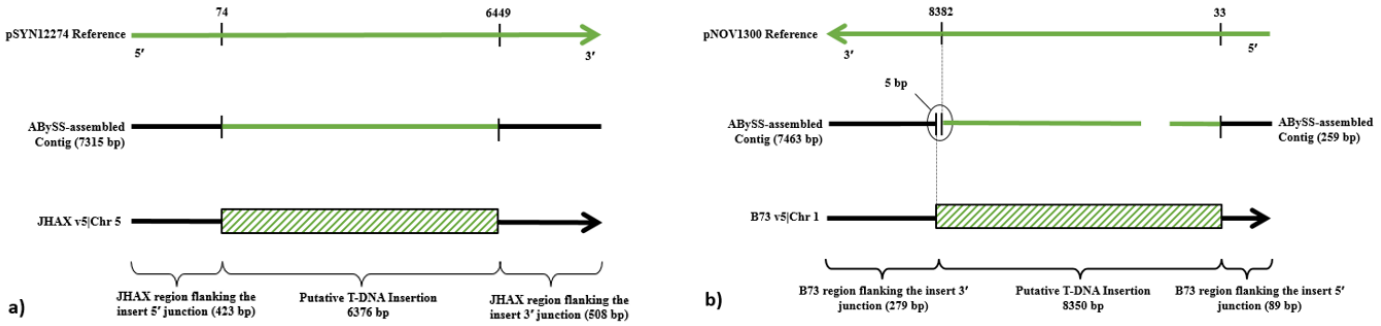


Figure 6: a) figure illustrating the putative ~6-kb 5323 T-DNA insertion region in JHAX v5 (hashed area), flanking regions, and orientation relative to the pSYN12274 plasmid reference and ABySS-assembled 7315-bp contig, b) figure illustrating the putative ~8-kb MIR152 T-DNA insertion region in B73 v5 (hashed area), flanking regions, and orientation relative to the pNOV1300 plasmid reference and ABySS-assembled contigs. A 5-bp region in the 7463 bp contig could not be assigned to either the pNOV1300 or B73 v5 reference. (Note: Figures are not to scale.)

unattainable using genome-walking PCR methods. The putative insertion site for the MIR152 T-DNA insert was determined to be on B73 v5 chromosome 1 (Figure 6b). A 5-bp region of the contig (approximately 86X coverage) at the putative 3′ insert-to-genome junction did not align to either pNOV1300 or B73 chromosome 1. Insertions of such "filler" DNA at the insert-to-genome junctions are common with *Agrobacterium*-mediated transformation [4, 27, 28]. In addition, it was determined that 16 bp of genomic sequence was deleted at the insertion site.

The coordinates for each of the PMI putative insertion sites were also determined, and corresponded to the region of maize chromosome 5 from which they were designed. The amount of flanking sequence obtained using WGS for both the 5323 and MIR152 events and the simulated PMI insertion events is comparable to what can be obtained using genome walking methods.

3.4. Insert Sequence Integrity

Full event characterization, including analyzing the integrity of the insert sequence, was attempted; however, WGS could not provide complete unambiguous sequence of the full T-DNA inserts. One explanation for this was that both 5323 and MIR152 inserts contain maize polyubiquitin promoters [6] and therefore sequence reads in these regions could not be mapped uniquely to the inserts. MIR152 contains two polyubiquitin promoters, further adding complexity to the analysis. Because the genome-to-insert junctions identified with ISC are unique sequences, ISC is not negatively impacted by these repetitive regions. It is important to note that the ABySS assemblies aligning to the T-DNA regions are considered draft level assemblies, and further analysis (e.g. single-molecule long-read sequencing, PCR and/or Sanger sequencing) would be required to conclusively define variants such as single nucleotide polymorphisms and insertions/deletions.

3.5. Genomic Reference Choice

The absence of a host reference genome and presence of endogenous genes in transgenic inserts can generate unintended false-positive junctions [31]. However, in the absence of a host reference genome, a closely-related genome can provide an adequate alternative. The B73 maize reference genome used for MIR152 alignments, though not isogenic, was also sufficient because the transformation line analyzed was a hybrid derived from B73 maize.

In this study the draft JHAX assembly was sufficient to identify false-positive junctions resulting from the endogenous maize polyubiquitin promoter (Table 1). The maize JHAX v4 reference was used for initial 5323 read alignments because of its well-characterized polyubiquitin region in common with pSYN12274. In order to determine putative insertion site coordinates on the more recent reference genome, flanking regions were aligned to JHAX v5 to obtain chromosomal location.

3.6. Genome Coverage Determination

To ensure that the WGS data generated for this study were high quality, the median coverage for each of the genomic sequencing libraries (5323, MIR152, and JHAX) was determined. Maize gene copy numbers can vary greatly depending on genetic background [20], therefore, CEGMA was used to obtain a set of single-copy maize genes that could serve as representatives for assessing coverage generated against single-copy regions. Multiple putative single-copy genes were used to assess coverage to avoid potential bias that could be introduced from using a gene thought to be single-copy but was in fact multiple-copy. The CEGMA analysis identified 229 complete and partial KOGs in the JHAX genome. KOGs aligning to JHAX at ≥99% length and >70% identity (MUMmer v3; [12]) were retained as putative single-copy genes based on relative coverage (previous analyses; unpublished). Of the 37 KOGs retained, 36 aligned to JHAX v4 and B73 v5 with >99% length and identity, and were used to determine genome coverage. The median genome coverage calculated using the single copy KOGs was 105X for event 5323, 93.5X for event MIR152, and 109X for the JHAX library.

The minimum coverage required for characterization of the insert-to-genome junctions of the ~6-kb putative 5323 insert,

Transformation Plasmid	Start	Stop	Reference	Start	Stop	Identity (%)	Length (bp)	Description
pSYN12274	2854	3850	JHAX v5|chromosome 5	99409429	99408433	100.00	997	Ubiquitin region[a]
pSYN12274	3870	4846	JHAX v5|chromosome 5	99404109	99403133	100.00	977	Ubiquitin region[a]
pNOV1300	200	2192	B73 v5|chromosome 5	84635058	84633067	99.95	1993	Ubiquitin region
pNOV1300	4798	6790	B73 v5|chromosome 5	84635058	84633067	99.95	1993	Ubiquitin region

Table 1: Locations of maize endogenous ubiquiton regions relative to the pSYN12274 plasmid, pNOV1300 plasmid, and JHAX v5, and B73 v5 host genomes.

	Coverage Calculations using ISC Method (pooled reads)		Coverage Calculations for Validation (indexed reads, with 25% overlap requirement)		Coverage Calculations for Validation (indexed reads, without overlap requirement)	
	Median Coverage	Genome Equivalent	Median Coverage	Genome Equivalent	Median Coverage	Genome Equivalent
5323 Sequencing Library (105X single copy coverage)						
5323 T-DNA insert	107X	1	105X	1	105X	1
25 bp PMI fragment	4X	1/26	5X	1/21	11X	1/9
50 bp PMI fragment	17X	1/6	17X	1/6	18.5X	1/5
100 bp PMI fragment	22X	1/4	22X	1/4	23X	1/4
150 bp PMI fragment	9X	1/11	10X	1/10	10X	1/10
200 bp PMI fragment	25X	1/4	27X	1/3	28X	1/3
MIR152 Sequencing Library (93.5X single copy coverage)						
MIR152 T-DNA insert	72X (5' junction)/ 103X (3' junction)	3/4 (5' junction)/ ~1 (3' junction)	97X	1	97X	1
25 bp PMI fragment	Not Detected	Not Detected	1X	1/93.5	2X	1/46
50 bp PMI fragment	Not Detected	Not Detected	3X	1/31	4X	1/23
100 bp PMI fragment	Not Detected	Not Detected	6X	1/15	6X	1/15
150 bp PMI fragment	Not Detected	Not Detected	3X	1/32	3X	1/32
200 bp PMI fragment	6X (3' junction)	1/16 (3' junction)	10X	1/9	12X	1/7

Table 2: Coverage of putative insertions. Genome equivalents were calculated by dividing the estimated coverage of the putative genomic insertion by the estimated single copy coverage of the host genome.

as well as the minimum coverage needed to generate a draft assembly of the entire ~6-kb region and flanking regions were determined by assembling subsets of reads. The insert-to-genome junctions of the putative 5323 ~6-kb insert were characterized with as little as one lane of data, with an approximate coverage depth of 11X. Three lanes of data, generating approximately 35X coverage, were required for a draft assembly of one contig containing the entire ~6-kb T-DNA region and flanking regions. The establishment of a minimum depth of coverage is an important validation step for this method and will ensure adequate coverage for molecular characterization of other GM events.

3.7. Coverage of Putative Insertion Sites

In order to estimate the depth of sequence coverage of each insert, the filtered reads used in the ABySS assembly were realigned to the entire set of resultant ABySS contigs. This coverage determination was then compared to the respective genomic sequence library coverage in order to approximate genome equivalents for each insertion. The calculated genome equivalents were used to determine the sensitivity of this method. The median coverage for the ~6-kb 5323 insert was 107X (Table 2) which is approximately one-genome equivalent of the calculated single-copy gene coverage for 5323 (105X). The sequence coverages of the PMI fragments calculated ranged from 1/4 to 1/26 per genome equivalent. These values were validated by calculating coverage using the reads parsed by index and requiring a 25% overlap in the PMI region as is required in the ISC method. The median coverage calculated for the ~6-kb 5323 insert was 105X. The median coverage calculations for the indexed PMI fragments spiked into the 5323 library ranged from 1/3 to 1/21 per genome equivalent.

Median coverage was also calculated for the indexed samples in the 5323 sequencing library without the 25% read overlap to determine the impact of this requirement on the coverage calculations (Table 2). The 25% read overlap requirement has the greatest impact on the 25 bp PMI fragment since it is the shortest fragment, decreasing median coverage from 11X to 5X.

The median coverage calculated for the ~7.5-kb MIR152 fragment containing the 3′ junction was 103X (Table 2) and 72X for the 259-bp fragment containing the 5′ junction, which vary from the one-genome equivalent of 93.5X calculated for MIR152 single-copy coverage. For MIR152, two contigs were used in the ISC method, each containing a junction sequence, as opposed to one contig as seen with 5323. Lower read realignment rates to the assembled ABySS contigs (84.85%) when compared to 5323 (92.84%) might be due to this fragmented assembly. One possible explanation for the fragmented assembly is the repetitive nature of the MIR152 insert, which contains two polyubiquitin promoter regions (Figure 1b). The collapse of these two polyubiquitin regions into one assembled contig could also account for the increased median coverage (103X) of the contig containing the 3′ junction, as opposed to the single-copy gene coverage of 93.5X for MIR152. Another

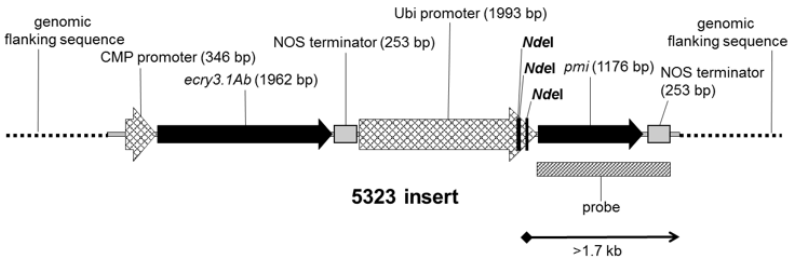
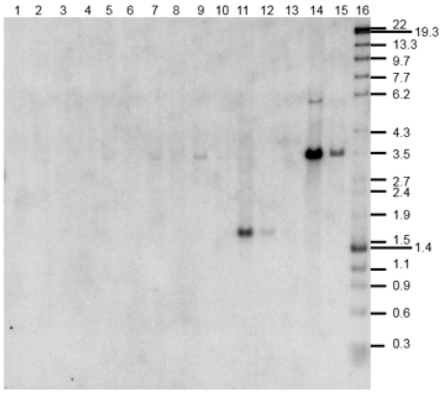


Figure 7: Map of the 5323 insert, including the location of the Southern blot analysis restriction sites and PMI-specific probe, and the size of the expected hybridization band.



Lane 1: Nontransgenic maize (*NcoI*) spiked with the 25 bp PMI fragment (1 copy).
Lane 2: Nontransgenic maize (*NcoI*) spiked with the 25 bp PMI fragment (1/7th copy).
Lane 3: Nontransgenic maize (*NcoI*) spiked with the 50 bp PMI fragment (1 copy).
Lane 4: Nontransgenic maize (*NcoI*) spiked with the 50 bp PMI fragment (1/7th copy).
Lane 5: Nontransgenic maize (*NcoI*) spiked with the 100 bp PMI fragment (1 copy).
Lane 6: Nontransgenic maize (*NcoI*) spiked with the 100 bp PMI fragment (1/7th copy).
Lane 7: Nontransgenic maize (*NcoI*) spiked with the 150 bp PMI fragment (1 copy).
Lane 8: Nontransgenic maize (*NcoI*) spiked with the 150 bp PMI fragment (1/7th copy).
Lane 9: Nontransgenic maize (*NcoI*) spiked with the 200 bp PMI fragment (1 copy).
Lane 10: Nontransgenic maize (*NcoI*) spiked with the 200 bp PMI fragment (1/7th copy).
Lane 11: Nontransgenic maize (*NcoI*) spiked with the PMI-specific probe (1 copy).
Lane 12: Nontransgenic maize (*NcoI*) spiked with the PMI-specific probe (1/7th copy).
Lane 13: Nontransgenic maize DNA (*NcoI*).
Lane 14: 5323 genomic DNA (*NdeI*) (1 copy).
Lane 15: 5323 genomic DNA (1/7th) and nontransgenic maize DNA (*NdeI*).

Figure 8: Southern blot analysis performed with the PMI-specific probe to compare sensitivity of this analysis with WGS and the ISC method.

possible explanation for the coverage discrepancy is that the host genome for the MIR152 line is a hybrid of B73 and another maize line, potentially affecting single-copy gene coverage calculations. Only one junction of one of the five PMI fragments (200 bp) spiked into the MIR152 sequencing library was detected using the ISC method. The coverage for this fragment was calculated as 1/16 per genome equivalent.

Median coverage calculated from indexed reads with and without the 25% read overlap requirement for the entire ~8-kb MIR152 insert was 97X, or approximately one-genome equivalent of the 93.5X single copy coverage. Median coverage calculation for the indexed PMI fragments show the fragments were spiked-in at very low levels (less than the attempted one-seventh copy per genome equivalent), providing a possible explanation as to why most were not detected using the ISC method. The fragment coverages ranged from as low as 1X (1/93.5 per genome equivalent) to 10X (1/9 per genome equivalent) (Table 2) with the required 25% read overlap, and from 2X (1/46 per genome equivalent) to 12X (1/7 per genome equivalent) without the requirement.

As shown from the coverage calculations using indexed reads, the ISC method generates coverage values that closely approximate the actual coverage of the T-DNA/PMI inserts when the assembled contig contains both the 5′ and 3′ genome-to-insert junctions. Because the 25% read overlap requirement decreases coverage calculations for the smallest fragments (25 to 50 bp), the indexed read coverages without the 25% overlap calculated in this study should be used when considering sensitivity and minimum coverage requirements in future analyses.

3.8. Sensitivity Analysis

Each of the PMI-containing pUC57 plasmids used during sequence analysis were also analyzed by Southern blot in order to compare the sensitivity for detecting fragments of these sizes by Southern blot versus WGS. Southern blot analysis was performed using a PCR-generated probe containing sequence from the PMI gene which is present in the inserts of both MIR152 and 5323. A map of the 5323 insert, indicating the locations of the PMI-specific probe and restriction sites, is shown in Figure 7. The Southern blot also included lanes with several positive and negative controls. Positive controls included genomic DNA from event 5323 which was meant to demonstrate detection of the PMI sequence in the genome of the event of interest. The negative control was nontransgenic maize genomic DNA which was meant to demonstrate absence of these sequences in the nontransgenic maize genome. Finally, the positive assay control was the PMI-specific probe to demonstrate sensitivity of the assay.

Each PMI-containing DNA fragment was expected to produce a 3.7- to 3.9-kb hybridization band (Figure 8). A 3.7- to 3.9-kb band was detected in lanes containing the 100-, 150-, and 200-bp PMI-containing fragments representing one copy; the 200-bp PMI-containing fragment was the only fragment detected at one-seventh copy (Figure 8). The expected bands were

produced from the controls included in these analyses. Digestion of genomic DNA from event 5323 with *Nde*I resulted in an extra hybridization band of approximately 5.7 kb, which can be attributed to incomplete digestion, a technical difficulty often seen with the use of restriction enzymes. No bands were produced in the lane containing the nontransgenic maize DNA and an approximately 1.5 kb hybridization band was produced in the lane containing the 1.5-kb PMI-specific probe, included as the positive assay control, as expected.

The sensitivity for these analyses was determined to be at 100 bp for a one copy per genome equivalent fragment; however, it is important to note that the sensitivity of Southern blot analyses may be influenced by GC content, length and specific activity of the probe, length of DNA fragment being probed, and the stringency of the washing conditions [22]. Thus, the sensitivity of Southern blot analysis may be greater.

The PMI-containing fragments (25 bp to 200 bp) spiked into the 5323 sequencing library were used to determine the sensitivity of WGS and the ISC method. All five PMI-containing fragments spiked into the 5323 sequencing library were correctly identified as confirmed by comparing results to the known sequences of the corresponding plasmid (Supplementary Figure 1). The 5323 assembled contigs containing PMI fragment sequences realigned to pSYN12274 with percent identities ranging from 99.35-100%. The smallest fragment detected was 25 bp, spiked in at a coverage level of 11X (1/9 copy per genome equivalent) based on indexed reads (Table 2). The lowest coverage level detected was the 150-bp PMI fragment at 10X (1/10 copy per genome equivalent).

The five PMI-containing fragments (25 bp to 200 bp) were also spiked into the MIR152 sequencing library, at low levels ranging from 2X to 12X coverage (Table 2). Only the 200-bp PMI-containing fragment was detected at 12X coverage, or 1/7 per genome equivalent, based on indexed reads.

Based on these results, a minimum of 11X coverage (equal to 1/9 per genome equivalent in the 5323 sequencing library) should be considered when trying to detect unintended insertions as small as 25 bp. The WGS and the ISC method was determined to be more sensitive than Southern blots for detection of small fragments, where the smallest fragment detected was 100 bp at one copy and 200 bp at one-seventh copy per genome equivalent.

4. Conclusion

In this paper we have demonstrated that WGS and the ISC method developed here were able to characterize transgenic events in GM maize by achieving the same molecular characterization endpoints as were traditionally achieved by Southern blot analysis. Although preliminary insert sequence was obtained, this method requires additional analysis (traditionally achieved through PCR and Sanger sequencing) to fully evaluate insert sequence integrity. While the events analyzed were relatively simple (one copy, no re-arrangements) the method should be capable of detecting multiple copies and re-arrangements which would be present as unique junction sequences. The ability to characterize complex insertions with WGS has also previously been demonstrated with other similar methods [9, 11, 30].

Comparison of sensitivity demonstrated that WGS was more sensitive than Southern blot analysis for detecting small insert fragments. With use of WGS and the ISC method, fragments as small as 25 bp were detectable when a minimum coverage of 11X is obtained for the fragment. Because of the sensitivity of WGS, low-level library contamination can present as false positive read alignments, particularly when the T-DNA plasmid backbone contains microbial sequence.

NGS technologies can provide a higher-throughput alternative to Southern blot analysis and offers several advantages, including flexibility and sequence-level detail that is less subjective than Southern blot analyses. The results of these analyses support a conclusion that WGS and an appropriate insertion site characterization method are a suitable alternative for the Southern blot analysis method traditionally used for molecular characterization of GM maize

5. Declaration of Conflicting Interest

The authors declare that there is no conflict of interest.

6. Acknowledgement

The authors would like to acknowledge Mark Rose Andrew Farmer, Laura Kavanaugh, and Bob Dietrich for valuable technical discussion and help with the experimental design.

7. Article Information

This article was received October 24, 2017, in revised form December 21, 2017, and made available online March 30, 2018.

8. References

[1] Arumuganathan, K., & Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter, 9*(3), 208-218.

[2] Arnold, C., & Hodgson, I. J. (1991). Vectorette PCR: a novel approach to genomic walking. *Genome Research, 1*(1), 39-42.

[3] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*(1), 421.

[4] Chilton, M. D. M., & Que, Q. (2003). Targeted integration of T-DNA into the tobacco genome at double-stranded breaks: new insights on the mechanism of T-DNA integration. *Plant Physiology, 133*(3), 956-965.

[5] Chomczynski, P. (1992). One-hour downward alkaline capillary transfer for blotting of DNA and RNA. *Analytical Biochemistry, 201*(1), 134-139.

[6] Christensen, A. H., Sharrock, R. A., & Quail P. H. (1992). Maize polyubiquitin genes: structure, thermal perturbation of expression and transcript splicing, and promoter activity following transfer to protoplasts by electroporation. *Plant Molecular Biology*, 18:675689.

[7] EFSA Panel on Genetically Modified Organisms (GMO). (2011). Scientific Opinion: Guidance for risk assessment of food and feed from genetically modified plants (Rep. No. 9). *EFSA Journal*, 9, 2150.

[8] Food and Agricultural Organization of the United Nations (FAO). (2003). *Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants* (Rep. No. CAC/GL 45-2003). Retrieved October 20, 2017 from www.fao.org/input/download/standards/10021/CXG_045e.pdf

[9] Guttikonda, S. K., Marri, P., Mammadov, J., Ye, L., Soe, K., Richey, K., . . . & Kumpatla, S. P. (2016). Molecular characterization of transgenic events using next generation sequencing approach. *PLoS One, 11*(2), e0149515.

[10] Holst-Jensen, A., Spilsberg, B., Arulandhu, A. J., Kok, E., Shi, J., & Zel, J. (2016). Application of whole genome shotgun sequencing for detection and characterization of genetically modified organisms and derived products. *Analytical and Bioanalytical Chemistry, 408*(17), 4595-4614.

[11] Kovalic, D., Garnaat, C., Guo, L., Yan, Y., Groat, J., Silvanovich, A., . . . & Bannon, G. (2012). The Use of Next Generation Sequencing and Junction Sequence Analysis Bioinformatics to Achieve Molecular Characterization of Crops Improved Through Modern Biotechnology. *The Plant Genome, 5*(3), 149-163.

[12] Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5:R12.

[13] Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., . . . & Wang, J. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature genetics, 42*(11), 1027-1030.

[14] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078-9.

[15] Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760.

[16] Lusk R. W. (2014). Diverse and Widespread Contamination Evident in the Unmapped Depths of High Throughput Sequencing Data. *PLoS One, 9*, e110808.

[17] Messing, J., & Dooner, H. K. (2006). Organization and variability of the maize genome. *Current Opinion in Plant Biology, 9*(2), 157-163.

[18] Murray, M. G., & Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research, 8*(19), 4321-4326.

[19] Organisation for Economic Cooperation and Development (OCED). (2010). *Consensus Document on Molecular Characterization of Plants Derived from Modern Biotechnology*. Retrieved October 20, 2017 from https://www.oecd.org/science/biotrack/46815346.pdf

[20] Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics, 23*(9), 1061-1067.

[21] Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842.

[22] Schatz, D. (1989). Southern Blotting and Hybridization. 1987. In Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., & Smith, J. A. and K. Struhl. (Ed.), *Current protocols in molecular biology*. 6.2.91-2.9.13: 2.9.1-2.9.13. New York: John Wiley and Sons Inc.

[23] Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., . . . & Wilson, R. K. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science, 326*(5956), 1112-1115.

[24] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research, 19*(6), 1117-1123.

[25] Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C. T., Jia, Y., . . . & Schnable, P. S. (2009). Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genetics, 5*(11), e1000734.

[26] Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics 14*, 178-192 (2013).

[27] Tzfira, T., Li, J., Lacroix, B., & Citovsky, V. (2004). *Agrobacterium* T-DNA integration: molecules and models. *Trends in Genetics, 20*: 375-383.

[28] Windels, P., De Buck, S., Van Bockstaele, E., De Loose, M., & Depicker, A. (2003). T-DNA Integration in Arabidopsis Chromosomes. Presence and Origin of Filler DNA Sequences. *Plant Physiology, 133*: 2061-2068.

[29] Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics, 21*(9), 1859-1875.

[30] Yang, L., Wang, C., Holst-Jensen, A., Morisset, D., Lin, Y., & Zhang, D. (2013). Characterization of GM events by insert knowledge adapted re-sequencing approaches. *Scientific Reports, 3,* 2839.

[31] Zastrow-Hayes, G. M., Lin, H., Sigmund, A. L., Hoffman, J. L., Alarcon, C. M., Hayes, K. R., . . . & Beatty, M. K. (2015). Southern-by-Sequencing: A Robust Screening Approach for Molecular Characterization of Genetically Modified Crops. *The Plant Genome, 8*(1).

[32] Zhang, H. B., Zhao, X., Ding, X., Paterson, A. H., & Wing, R. A. (1995). Preparation of megabasesize DNA from plant nuclei. *The Plant Journal, 7*(1), 175-184.

Supplementary Figure 1: Alignment of PMI Junction Contigs to Reference Sequence

Junction contig 2

```
Junction contig 2       AATCTCACAT TCGATACTTG AGAACGGGTT ATAGTAGATA ACAAGATAAT
200bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 2       AGGGCAGAAT CATGAAAGAT CAGAGATTCG GATGATAAGG TCACAACATG
200bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 2       ATTTTTACAA CAGCCTGATG CCGAACGTTT AAGCGAACTG TTCGCCAGCC
200bp PMI reference     .TTTTTACAA CAGCCTGATG CCGAACGTTT AAGCGAACTG TTCGCCAGCC

Junction contig 2       TGTTGAATAT GCAGGGTGAA GAAAAATCCC GCGCGCTGGC GATTTTAAAA
200bp PMI reference     TGTTGAATAT GCAGGGTGAA GAAAAATCCC GCGCGCTGGC GATTTTAAAA

Junction contig 2       TCGGCCCTCG ATAGCCAGCA GGGTGAACCG TGGCAAACGA TTCGTTTAAT
200bp PMI reference     TCGGCCCTCG ATAGCCAGCA GGGTGAACCG TGGCAAACGA TTCGTTTAAT

Junction contig 2       TTCTGAATTT TACCCGGAAG ACAGCGGTCT GTTCTCCCCG CTATTGCTGA
200bp PMI reference     TTCTGAATTT TACCCGGAAG ACAGCGGTCT GTTCTCCCCG CTATTGCTGA

Junction contig 2       ATTCACAAGG AAAAAGATCA CTAGATCCAT GCGAAAGGAG AGGTAGGCAA
200bp PMI reference     A......... .......... .......... .......... ..........

Junction contig 2       CAAGATCAGC TGGATGATCA ACAGGAATGC TATGAAGTTT TAGGGGCAAG
200bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 2       GAATTTATGG AAAGAAACAT GGCCTTGATA GGGTTTGCGC A
200bp PMI reference     .......... .......... .......... .......... .
```

Junction contig 3

```
Junction contig 3       TGTTTTGGTC GATGATGTGC ATGTGGTTAT ATGTGTGTAA CTGTATAATT
150bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 3       TTATAAATGG ACGCGTGTAG GGAAGAAATG TTTAAGCGAA CTGTTCGCCA
150bp PMI reference     .......... .......... .........G TTTAAGCGAA CTGTTCGCCA

Junction contig 3       GCCTGTTGAA TATGCAGGGT GAAGAAAAAT CCCGCGCGCT GGCGATTTTA
150bp PMI reference     GCCTGTTGAA TATGCAGGGT GAAGAAAAAT CCCGCGCGCT GGCGATTTTA

Junction contig 3       AAATCGGCCC TCGATAGCCA GCAGGGTGAA CCGTGGCAAA CGATTCGTTT
150bp PMI reference     AAATCGGCCC TCGATAGCCA GCAGGGTGAA CCGTGGCAAA CGATTCGTTT

Junction contig 3       AATTTCTGAA TTTTACCCGG AAGACAGCGT GAAATAGAAA AGAACTCGAG
150bp PMI reference     AATTTCTGAA TTTTACCCGG AAGACAGCG. .......... ..........

Junction contig 3       TATTTTTATT TTGATAGGAA AATATGCGAC GAGA
150bp PMI reference     .......... .......... .......... ....
```

Junction contig 4

```
Junction contig 4       TCTCGTCAAA AACAAAGACA AGAGACATAA ATATCCAATA CAAAAGGAAA
100bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 4       CCAGAGAGGT AGTGGTATTT TTTTCTTTCT TGGTGGCTAA GCATCGCTCA
100bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 4       CCCTGTGATG CAAAAATCTA CCAGAGACAA GTATAGCCAA GACCATCAAA
100bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 4       TAAAGAGACA ATTTAGCAAA CAATCCAAAT CAAGATCNAT CTTCAAAGTC
100bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 4       ACAATCTTGG AAGAGTTTCT TTGCCCTTTT TTGGCAGGGG AAGGGTGGGT
100bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 4       AAGTCCTATC AGTAGAGTTG AATATGCAGG GTGAAGAAAA ATCCCGCGCG
100bp PMI reference     .......... ......GTTG AATATGCAGG GTGAAGAAAA ATCCCGCGCG

Junction contig 4       CTGGCGATTT TAAAATCGGC CCTCGATAGC CAGCAGGGTG AACCGTGGCA
100bp PMI reference     CTGGCGATTT TAAAATCGGC CCTCGATAGC CAGCAGGGTG AACCGTGGCA

Junction contig 4       AACGATTCGT TTAATTTTGC AATCAACAAT AGGATAAGAT CTCATATGTA
100bp PMI reference     AACGATTCGT TTAATT.... .......... .......... ..........

Junction contig 4       TTATGGAAAC AAATAAGTAG ATTTTTGCGT TACAAAGGTT ACCTTTTT
100bp PMI reference     .......... .......... .......... .......... ........
```

Junction contig 5

```
Junction contig 5       GATCAGTCAT TGTACTTCTT CTATTAGGGT CTACATTTTA TCAAGGTCAG
50 bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 5       TTATTGTAGT TCACCCTGCT GGCTATCGAG GGCCGATTTT AAAATCGCCA
50 bp PMI reference     .......... TCACCCTGCT GGCTATCGAG GGCCGATTTT AAAATCGCCA

Junction contig 5       GCGCGCGGGA TATCAGGATC AATCATTATA TTTTAATCAG TGTCAGTCAA
50 bp PMI reference     GCGCGCGGGA .......... .......... .......... ..........

Junction contig 5       TGTATTTATT AAGGTCAATC ATTGTATT
50 bp PMI reference     .......... .......... ........
```

Junction contig 6

```
Junction contig 6       AAGTCATTCT GTTACAATTC TAGTCATCAC ATGTCATTTA GTCATTTTAT
25 bp PMI reference     .......... .......... .......... .......... ..........

Junction contig 6       GACTTATTTA AAATATTTCA TATTGTCAGC GATTTTAAAA TCGGCCCTCG
25 bp PMI reference     .......... .......... ........GC GATTTTAAAA TCGGCCCTCG

Junction contig 6       ATAACAGTTG TTACAAGACT
25 bp PMI reference     ATA....... ..........
```