The Use of Next Generation Sequencing and Junction Sequence Analysis Bioinformatics to Achieve Molecular Characterization of Crops Improved Through Modern Biotechnology

David Kovalic,* Carl Garnaat, Liang Guo, Yongpan Yan, Jeanna Groat, Andre Silvanovich, Lyle Ralston, Mingya Huang, Qing Tian, Allen Christian, Nordine Cheikh, Jerry Hjelle, Stephen Padgette, and Gary Bannon

Abstract

The assessment of genetically modified (GM) crops for regulatory approval currently requires a detailed molecular characterization of the DNA sequence and integrity of the transgene locus. In addition, molecular characterization is a critical component of event selection and advancement during product development. Typically, molecular characterization has relied on Southern blot analysis to establish locus and copy number along with targeted sequencing of polymerase chain reaction products spanning any inserted DNA to complete the characterization process. Here we describe the use of next generation (NexGen) sequencing and junction sequence analysis bioinformatics in a new method for achieving full molecular characterization of a GM event without the need for Southern blot analysis. In this study, we examine a typical GM soybean \([Glycine \text{ max} (L.) \text{ Merr.}]\) line and demonstrate that this new method provides molecular characterization equivalent to the current Southern blot-based method. We also examine an event containing in vivo DNA rearrangement of multiple transfer DNA inserts to demonstrate that the new method is effective at identifying complex cases. Next generation sequencing and bioinformatics offers certain advantages over current approaches, most notably the simplicity, efficiency, and consistency of the method, and provides a viable alternative for efficiently and robustly achieving molecular characterization of GM crops.

Molecular Characterization is a key step in the assessment of genetically modified (GM) crops for regulatory approval. Current characterization methods are designed to accurately establish a number of important properties of the inserted DNA in the GM crop’s genome and this characterization is a key step in the production of improved GM crops. These characterizations are required for selection of the events with the most favorable molecular profile throughout research activities (Heck et al., 2005; Vaughn et al., 2005; Cerny et al., 2010) and are also performed as a last step before commercialization, where it is a crucial part of the comparative assessment process endorsed by Codex and used by most global regulatory organizations (Codex Alimentarius, 2003).

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. Current methods also establish a description of any genetic rearrangements that may have occurred at the insertion site as a consequence of transformation and demonstrate the absence of unintended plasmid DNA (i.e., non-transfer DNA [T-DNA] backbone sequence) within the GM event. Generational stability analysis, which demonstrates the stable heritability of inserted DNA


Abbreviations: ABI, Applied Biosystems, Inc.; CTAB, hexadecyltrimethylammonium bromide; EDTA, ethylenediaminetetraacetic acid; GM, genetically modified; JSA, junction sequence analysis; MW, molecular weight; NexGen, next generation; PCI, phenol:chloroform:isoamyl alcohol; PCR, polymerase chain reaction; T-DNA, transfer DNA; TE, Tris-EDTA; WGS, whole-genome shotgun.
sequences over a number of breeding generations, is also routinely conducted.

Currently, molecular characterization is achieved by a combination of Southern blot analysis (which determines insert and copy numbers of the event, presence or absence of the backbone, and generational stability [Southern, 1975]) and sequencing overlapping polymerase chain reaction (PCR) fragments spanning the inserted DNA (which determines the exact sequence of the insert and the DNA sequence of native genomic flanking regions) and the insertion site within the native genome (alternately referred to as the "host genome" in some sources [Codex Alimentarius, 2009]). Using next generation (NexGen) DNA sequencing (e.g., 454 [454 Life Sciences]), Illumina (Illumina Inc.), and SOLiD (Applied Biosystems, Inc. [ABI]), which provide dramatically increased sequencing throughput versus capillary electrophoresis sequencing technologies (e.g., ABI 3700 series), it previously has been established that it is possible to achieve comprehensive coverage of complex genomes given that sufficiently deep sequencing is performed (Wang et al., 2008) and it has also been shown that such deep sequencing can serve as the basis for accurate whole-genome studies (Ajay et al., 2011). Here we use deep Illumina sequencing and a novel data analytical workflow to demonstrate that it is possible to remove the need for Southern blot analysis and to conduct accurate molecular characterization using solely DNA sequencing and bioinformatics.

The overall strategy for this new molecular characterization method is to produce DNA sequence fragments that comprehensively cover the entire genome of test and control plants (i.e., the GM event under investigation and the parent line from which it was derived) and use bioinformatic tools to analyze these DNA fragments. These bioinformatic analyses establish the insert and copy number and the presence or absence of backbone sequences. It is worth noting that precisely the same endpoints of molecular characterization are reached as with the current Southern blot-based method (Fig. 1). There are multiple advantages to using NexGen sequencing and bioinformatics, most notably the simplicity and the consistency of the method as opposed to Southern blot studies, which require customized experimental design for every event; the method described here is essentially identical for all events. The new sequencing-based method also overcomes many technical challenges inherent in Southern blot analyses (e.g., false positive hybridization bands resulting from incomplete digestion or star activity [Wei et al., 2008]) and the need for radioactive ³²P-labeled probes. This new method will also provide higher reproducibility, because it is less dependent on complex lab-based procedures and it decreases the time and resources required to complete molecular characterization studies.

The bioinformatic analysis developed and described here is based on accurately detecting and characterizing novel chimeric sequences resulting from insertions into the native genome as shown in Fig. 2. Throughout this paper such novel chimeric sequences are referred to as “junction sequences” since the insertion of novel DNA into parental plant genomic DNA creates new junctions between the introduced DNA and the native genomic DNA sequence; the sequencing and bioinformatics analytics used to achieve this characterization is termed “junction sequence analysis” (JSA).

In this paper we describe the NexGen sequencing and bioinformatics method and demonstrate its functional equivalence to current Southern blot-based molecular characterization by using both of these methods to establish molecular characterization of a typical GM event (MON17903). We also examine an event containing in vivo DNA rearrangement of multiple T-DNA inserts (MON87704) to demonstrate that the new method is effective at identifying complex cases. We discuss some potential applications, practical advantages, and considerations of the new method.

Materials and Methods

Plant and Reference Materials

Test events MON17903 and MON87704 were generated during crop research and testing. The control materials, designated A3244 or A3525, are the conventional soybean varieties that are the respective parents of the test events. The test and control materials have similar genetic backgrounds with the exception of Agrobacterium-inserted T-DNA cassettes. These lines were characterized using genomic DNA whose extraction is described below.

The transformation plasmids (Fig. 3) were used in conjunction with the control material to develop the GM events studied here. These plasmid DNAs serve as positive controls for Southern blots and are used as spike-in control for the DNA sequencing experiments. The known DNA sequences of the transformation plasmids also serves as reference for bioinformatic analyses as described below.

Genomic DNA Isolation and Quantification

Genomic DNA for NexGen sequencing library construction was isolated from seeds of the test materials and the conventional soy controls. Seeds were treated to remove surface contaminants by agitation in a 0.05% (v/v) Tween-20 solution for 30 sec and rinsing with tap water followed by submerging in 0.5% NaOCl for 1 min and rinsing with tap water followed by two 1-min washes with 1% HCl. The seeds were dried and then processed to a fine powder using a Harbil paint shaker (Fluid Management) for 3 min. For Southern blot analysis, tissue from newly expanding leaves was used for DNA isolation. Total genomic DNA for all sequencing library preparations and Southern blot analyses was isolated using a modified hexadecyltrimethylammonium bromide (CTAB) procedure (Rogers and Bendich, 1985): Approximately 2.5 g powder was transferred to 50 mL conical tubes with 25 mL of CTAB extraction buffer (1.5% [w/v] CTAB, 75 mM Tris-HCl pH 8.0, 100 mM ethylenediaminetetraacetic acid [EDTA] pH 8.0, 1.05 M NaCl, and 0.75% [w/v] polyvinylpyrrolidone [molecular weight
KOVALIC ET AL.: JUNCTION SEQUENCE ANALYSIS

Figure 1. Molecular methods and their resultant characterization endpoints. Panel A details the current methodology (reliant on Southern blot analysis): the junction sequence analysis (JSA) method described in this paper and the resultant characterization endpoints. Panel B shows locus-specific polymerase chain reaction (PCR) and sequencing methods and endpoints, which are common to both the Southern-based and the JSA characterization method.

Figure 2. Illustration of a simple transformation event leading to detectable novel junction sequences. Schematic detail of a simple example transformation event, wherein single copy of a transfer DNA (T-DNA) cassette has been inserted at a single locus within the genome. The relationship of the control and test loci are shown along with the inserted T-DNA cassette of the test material and two novel chimeric sequences detectable around the two T-DNA to genomic flanking junctions. These novel sequences are characteristic of the transformed locus and are detectable by next generation (NexGen) sequencing and bioinformatics. Throughout this paper we term these sequences “junction sequences.”
[MW] 40,000]. The samples were incubated at approximately 65°C for 60 min with intermittent mixing. Equal volumes of phenol:chloroform:isoamyl alcohol (25:24:1) (PCI) were added to the samples and mixed by inversion for 15 min. The samples were centrifuged at 12,000 × g for 10 min and the aqueous phase transferred to a clean tube. The PCI extraction and centrifugation steps were repeated two more times. The aqueous phase was extracted with equal volume of chloroform and centrifuged as before. The aqueous phase was transferred to a clean tube and DNA was precipitated by mixing in two volumes of 100% ethanol. The DNA was centrifuged at 5000 × g for 5 min and the DNA pellets were rinsed with 70% ethanol and air dried. The DNA was dissolved in Tris-EDTA (TE) buffer and then precipitated by the addition of equal volumes of precipitation buffer (20% w/v polyethylene glycol [MW 5000] and 0.25 M NaCl). After incubation at 37°C for 15 min, the DNA was centrifuged at 5000 × g for 15 min and rinsed with 70% ethanol. The DNA was air dried and redissolved in TE buffer (pH 8.0). All extracted DNA was quantified using a Qubit Fluorometer (Invitrogen) and was stored in a 4°C refrigerator or a −20°C freezer.

**Sequencing Library Preparation**

Genomic and plasmid DNA libraries were prepared using TruSeq library kits (Illumina) in accordance with the manufacturer’s protocols. The DNA was sheared using a Covaris S2 ultrasonicator (Covaris, Inc.), and the resultant double-stranded DNA fragments were end-repaired, A-tailed, and ligated to indexed adapters. Ligation products were purified using a Pippin Prep DNA size selection system (Sage Science Inc.) to select DNA for fragments of an average insert size of 280 bp. The DNA fragments with adaptor molecules were selectively enriched by 10 cycles of PCR, and the libraries were quantified using a Bioanalyzer 2100 using a High-Sensitivity DNA kit (Agilent Technologies). Indexed adapters were used to allow for the multiplexing of samples for sequencing and to ensure sample and data integrity. Eight separate indexed libraries were produced for both the test and control soy samples, and these eight libraries were used to create test and control library pools (in equimolar ratios) before sequencing.

To produce “spike-in” positive control samples for sequencing, two plasmid DNA libraries were created as described above and then diluted to either one or one-tenth soy genome equivalents (concentrations equivalent to single copy or one-tenth copy per genome representation of the plasmid DNA) before pooling with samples produced from the control material (multiplexed as described above).

**Whole-Genome Sequencing**

The pooled sequencing library samples (test, control, and plasmid spike-in samples prepared as described above) were sequenced using Illumina HiSeq technology (Illumina, Inc.) following the manufacturer’s procedure that produced paired-end short sequence reads (approximately 100 bp long). Each pooled sample library was run on separate HiSeq lanes. Sufficient numbers of sequence fragments (reads) are obtained (>75x effective genome coverage, determined as described below) to comprehensively cover the genomes of the sequenced samples (Clarke and Carbon, 1976; Wang et al., 2008; Ajay et al., 2011).

**Sequencing Read Selection**

Reads for further analysis were selected based on their similarity to various “selection query sequences”
Whole-genome high-throughput sequence reads were selected on the basis of sequence similarity to the appropriate query sequence using the local alignment software BlastAll (version 2.2.21; Altschul et al., 1990, 1997). The DNA sequencing reads with a match to the query sequence having an e-value of 1 × 10^{-5} or less and having a match length of at least 30 bases with at least 96.7% sequence identity were collected. The selection criteria had previously been established as providing the best possible sensitivity and specificity based on a parameter optimization study where many different potential parameters sets were systematically evaluated for performance (not shown).

To identify and remove selected reads originating from native soybean genome sequence (specifically from any regions that have similarity to the selection query sequence), the collected reads were compared against the reference genome DNA sequence of the native soybean genome (Schmutz et al., 2010) using BlastAll (version 2.2.21; Altschul et al., 1990, 1997) and reads were removed from the selected collection if they had better sequence alignment quality (bit score > 5 larger) or a longer alignment length (>2 bases longer) when compared to the reference genome DNA sequence versus comparison to the selection query sequence in the same region of the read (match region overlap of at least 10 bp).

For the final selected dataset both reads of the paired-end sequences were collected in all cases and used as input to the sequencing read quality refinement step as described below.

**Sequencing Read Quality Refinement**

A superior quality segment (bases 3 through 42) of each selected DNA sequence was compared to all other selected sequencing reads with the alignment software Bowtie version 0.12.3 (Langmead et al., 2009) allowing up to one mismatch. If multiple read pairs were matched at both paired reads, such read pairs were deemed redundant and only the best sequence quality pair of reads (defined by phred sequence quality score [Ewing and Green, 1998; Ewing et al., 1998]) was kept for further analysis.

Computer software Novoalign version 2.06.09 (Novocraft Technologies, 2008) was used to remove the unaligned adaptor sequences at both ends of the sequencing reads. Low quality read ends (with phred quality scores of 12 or lower) were trimmed from the reads. Only reads of 30 bases or longer after adaptor and quality trimming were collected for further analysis.

**Effective Sequencing Depth Determination**

A well-known single copy locus, lectin (Le1) (GenBank accession version K00821.1), was selected from the soybean (Glycine max) genome and used to estimate the effective sequence coverage depth for each sample sequenced. Whole-genome high-throughput sequence reads were selected as described above using the Le1 sequence as a selection query, quality refined as described above, and aligned to the known Le1 sequence, and the sequencing depth distribution at this locus was calculated based on the resultant alignment. The average sequencing depth is defined as the average number of sequencing reads across the alignment. For example, at 100x average depth of coverage we expect any base in the genome was independently sequenced 100 times in our experiment.

**Junction Sequence Read Selection**

The sequence of the transformation plasmid PV-GMGOX20 or PV-GMPQ/HT4404 were used as a selection query (as described above) to find all reads that were either fully matched or partially matched to the plasmid sequence, with the latter being characteristic of chimeric junction sequences. A junction sequence was characterized by the presence of both a transformation plasmid sequence and either nonplasmid sequence or noncontiguous plasmid sequence; these sequences are likely to be derived from either the native soybean genome flanking an insertion or from rearranged plasmid. The selected reads were quality refined as described above.

**Junction Detection**

Selected and quality-refined reads were then aligned against the whole transformation plasmid to detect junction sequences. Reads with partial match to the transformation plasmid (at least 30 bases and 96.7% identity) were collected as potential junction sequences and their match cutoff position (junction point) on the plasmid were noted. A custom developed Perl script was used to identify the junction points on the transformation plasmid and their supporting junction reads. For each junction position, all supporting junction reads were aligned at the 30 bases proximal to the junction position. The remaining bases of these reads are sorted to show the alignment and the consensus of the flanking junction sequences past the junction point.

**Locus-Specific Polymerase Chain Reaction and Sequencing Analysis**

Primers were designed specific to the DNA insert and the flanking genomic DNA based on the sequence determined from the transformation plasmid and from the genomic DNA of the GM soybean MON17903 or MON87704. The primers used to amplify three overlapping regions of DNA that span the entire length of the insert in MON17903 and a portion of the flanking sequences are shown in Fig. 4. The PCR analyses were conducted using 100 ng of genomic DNA template in a 50 μL reaction volume containing a final concentration of 1.5 mM MgCl₂, 0.2 μM of each primer, 0.2 mM of each deoxyribonucleotide triphosphate, and 0.02 units/μL of Phusion Hot Start II High Fidelity DNA Polymerase (New England Biolabs). The amplification was performed under the following cycling conditions: 98°C for 30 sec and then 35 cycles of 98°C for 10 sec, 65°C for 15 sec, and
72°C for 1 min. Aliquots of each product were separated on 1.0% (w/v) agarose gels and visualized by ethidium bromide staining to verify that the products were of the expected size. Before sequencing of the PCR products, they were purified using the QIAquick PCR Purification Kit (Qiagen, Inc.). The PCR products were sequenced on an ABI 3730 capillary machine (Applied Biosystems) using BigDye Terminator chemistry kits. The DNA sequences were assembled to generate a consensus sequence spanning the entire insert using standard assembly software (CLC bio, 2012). Similar strategy and methods were used to analyze MON87704.

Southern Blot Analysis
Ten micrograms of total genomic DNA were digested in parallel with positive control samples (plasmid PV-GMGOX20 DNA or probe template spiked into 10 μg of control soybean DNA) and subjected to Southern blot analyses. Digested genomic DNA was separated using gel electrophoresis using a 0.8% (w/v) agarose gel. All Southern blot analyses were based on the method of Southern (Southern, 1975). Probe template DNA containing sequences of the transformation plasmid PV-GMGOX20 was prepared by PCR amplification. Approximately 25 ng of each probe template was labeled with 32P-deoxycytidine triphosphate (~6000 Ci mmol⁻¹; GE Healthcare Biosciences) by the random priming method (RadPrime DNA Labeling System, Life Technologies). Multiple exposures of each blot were then generated using Kodak Biomax MS film (Eastman Kodak) in conjunction with one Kodak Biomax MS intensifying screen in a −80°C freezer.

Results and Discussion
In this paper we describe a new method for molecular characterization of GM crops, JSA; the various steps of the new JSA method are depicted in Fig. 5. This method uses deep NexGen sequencing, locus-specific sequencing, and bioinformatics to achieve the goals of molecular characterization in place of Southern blot analysis as shown in panel A in Fig. 1.

Sequencing and bioinformatics analysis, which is based on the production of an assembled 3x draft genomic sequence (i.e., containing an average depth of sequence coverage of 3x), has previously been used for molecular characterization of a papaya (Carica papaya L.) GM event (Ming et al., 2008). In this sequencing of the papaya genome the authors used 1.6 million whole-genome shotgun (WGS) reads derived from a variety of plasmid and bacterial artificial chromosome libraries, which were constructed with a range of insert sizes. These WGS reads totaled approximately 3x genome coverage and were used to assemble a draft genome sequence of the GM papaya line SunUp. From the perspective of molecular characterization the most notable difference between the papaya study and the method presented here is the completeness of the genomic DNA dataset being used to achieve the characterization, hence the completeness of the characterization itself. Using a previously established and accepted method (Clarke and Carbon, 1976), the 3x papaya dataset is predicted to represent approximately 95% of the genome sequenced (the authors estimated 75% genome coverage with the assembled contigs) whereas the 75x coverage used in our method is predicted by the
same theory to provide genome coverage that would not miss a single basepair in the soybean genome. It previously has been experimentally established that it is possible to achieve comprehensive coverage of complex genomes that form the foundation for accurate whole-genome studies given deep NexGen sequencing (Wang et al., 2008; Ajay et al., 2011); this is notwithstanding known biases in NexGen sequencing techniques, including the Illumina sequencing by synthesis method used here (Minoche et al., 2011).

In this study effective experimental coverage was estimated by examining a well-known single copy locus, and sufficient depth of coverage to allow for a comprehensive dataset was observed (>75x) for every sample. We have compared values of estimated coverage obtained using Le1 as a representative locus vs. whole-genome mapping (i.e., in this case mapping all reads vs. the complete transcriptome of soybean) and found that by using a carefully chosen, conserved, and representative single-copy locus we achieve good estimations of average genome coverage (not shown). The use of such regions is considerably more efficient than whole-genome mapping, with good depth of coverage estimates being generated using a fraction of the compute resources in considerably less time. Genomic regions with properties similar to Le1 may be selected for use in rapidly estimating experimental coverage in other species.

It is important to note that even though previous studies have established that sufficiently deep Illumina sequencing can provide comprehensive datasets that allow for accurate whole-genome studies, the method presented here also includes experimental controls for detection sensitivity (i.e., positive spike-in controls). These experimental controls demonstrate attainment of the required sensitivity for the precise sequences required to achieve accurate characterization (i.e., the entire transformation plasmid). These controls are run with every experiment and it should be noted that these controls are exactly the same as are those currently run in conventional Southern blot analyses.

Also of note is that although the method presented here provides >75x coverage of the genomes under study, accurate assembly of these complete genomes is not technically possible using currently available sequence assembly tools. This is due to the nature of the sequences generated in this study (i.e., short reads of a single short insert length [Miller et al., 2010]); notwithstanding this limitation on sequence assembly, these sequences represent datasets sufficient for achieving precise molecular characterization of transformed DNA in GM crops.
Southern Blot Analysis

The DNA insert of MON17903 was characterized by Southern blot analyses with a set of probes that spanned the entire transformation plasmid (Fig. 3). The selection and design of the probes allowed for the determination of the copy number and the number of insertion sites as well as the presence or absence of all sequences from the transformation plasmid. Figure 4 shows a linear map depicting restriction sites within the DNA insert as well as the soybean genomic DNA flanking the insert in the GM soybean MON17903. The restriction enzymes AseI and StyI both cleave once within the inserted DNA and once within the known 5’ and 3’ flanking sequences. Restriction digestion of genomic DNA with AseI or StyI generates a specific set of DNA fragments (fingerprint), each of which contains a part of the DNA insert and a genomic flanking sequence. The size and number of the DNA fragments that make up the fingerprint are dictated by the genomic flanks, which are unique to each insert. Therefore, the insert and copy number of DNA inserts in a GM event can be determined by the DNA fingerprints in Southern blot analysis. Expected bands were produced for each probe set from the T-DNA region and restriction enzyme combination and no unexpected bands were detected showing that the T-DNA insert is intact and contained within a single locus (panels A and B in Fig. 6).

In addition, three overlapping probes that span the entire backbone of PV-GMGOX20 were used for Southern blot analysis. MON17903 DNA digested with AseI or StyI showed no detectable hybridization signal (panels A, B, and C in Fig. 7). Thus, no additional elements from the transformation
vector linked or unlinked to intact gene cassettes are detectable in the genome of the MON17903 event.

Locus-Specific Polymerase Chain Reaction and Sequencing

The organization of the inserts in MON17903 and MON87704 were confirmed using PCR analysis by amplifying overlapping regions of DNA that span the entire length of the insert and the approximately 1 kb genomic DNA flanking each end of the insert. The positions of the PCR products relative to the MON17903 insert are shown in Fig. 4.

For either event amplicon production and sequencing were followed by assembly and generation of a consensus sequence spanning the entire locus. The resulting consensus sequence was used in pairwise alignments with the previously determined sequence of the transformation plasmid (either PV-GMGOX20 or PV-GMPQ/HT4404, respectively). For either event the resultant alignment of the insert sequence in the GM soybean to the T-DNA sequence of the transformation plasmid demonstrated that the expected DNA sequence from the plasmid vector was integrated into the soybean genome and was not unexpectedly altered during transformation.

Taken together, the results from Southern blot analysis and locus-specific PCR and sequencing demonstrate that the GM soybean event MON17903 contains a single copy of the inserted T-DNA at a single genomic locus. These results also demonstrate that no unintended backbone DNA fragments have been inserted into the genome of MON17903. Similar data for MON87704 show the co-integration of the T-DNA I and T-DNA II sequences at a single locus along with in vivo rearrangement of the relative orientation of these two T-DNAs as shown in Fig. 8, panel A.

Whole-Genome Sequencing, Effective Coverage, and Experimental Limit of Detection

The whole-genome libraries for test and control materials were sequenced to a depth of >75x effective coverage for each sample (at >75x average depth of coverage we expect, on average, each basepair in the genome was independently sequenced >75 times). Effective genome coverage was experimentally determined using a well-known single copy locus in the soybean genome, the Glycine max lectin gene (Le1). Selecting sequences similar to the known endogenous single copy locus Le1 (endogenous selection query sequence in Fig. 9) and mapping these to the known gene sequence enables us to estimate coverage and demonstrates that both samples are adequately sequenced to a depth of coverage of >75x, which is required to ensure comprehensive datasets for analysis. This level of coverage is predicted to provide full genome coverage for soybean (Clarke and Carbon, 1976); for both the test and control samples we observed suitably deep coverage of both the Le1 reference gene and the positive spike-in control (see below), demonstrating that each sample is adequately sequenced in the experiment (Supplemental Table S1).

By conducting an Effective Sequencing Depth Determination (as described above, using the plasmid spike-in sequencing reads and the known sequence of the transformation plasmid PV-GMGOX20 or PV-GMPQ/HT4404, respectively) we observed complete coverage of
Figure 8. Schematic representation of the insert DNA in soybean MON87704. Panel A represents a linear map of the transfer DNA (T-DNA) I and T-DNA II arrangement in the transformation plasmid, PV-GMPQ/HT4404. The numbers at the bottom are the corresponding coordinates on PV-GMPQ/HT4404. Panel A also represents a linear map of the arrangement of the two T-DNAs as they exist in MON87704. The numbers at the bottom are the corresponding coordinates on MON87704. In MON87704, the T-DNA II region between two vertical dotted lines is inverted and linked in tandem to the T-DNA I region at its partial right border. Panel B: The numbers correspond to the basepair locations of the T-DNA I and T-DNA II regions in the transformation plasmid PV-GMPQ/HT4404 and those of in MON87704 insert sequence. Also shown is the correspondence of the junction sequences determined from high throughput sequencing and junction sequence analysis (JSA) bioinformatics with the full insert sequence characterized by in planta locus-specific sequence. Junction points are indicated by the “^” character in the alignment sequence text.
our reference material at both the one genome equivalent and one-tenth genome equivalent (100% nucleotide identity over 100% of the transformation plasmid; not shown). This result demonstrates that all basepairs of the transformation plasmid are observed by the sequencing and bioinformatics performed in this study down to a level of at least one-tenth genomic equivalent and hence the achievement of a detection limit of at least one-tenth copy number for the plasmid DNA sequences in this study.

### Sequencing Read Selection

To conduct a detailed study of all sequencing reads that were either wholly or partly sourced from the transformation plasmid we selected from the full dataset only those reads derived from the transformation plasmid based on sequence similarity. To perform this selection, we used the well-established sequence database search tool BlastAll (using optimized parameters as described in the Materials and Methods section above) (Altschul et al., 1990, 1997). The results of sequence read selection of either the MON17903 test or A3244 control samples using the DNA sequence from the PV-GMGOX20 transformation plasmid T-DNA, backbone or the endogenous control sequence (Le1) are shown in Supplemental Table S2 and Fig. 9.

Examination of Fig. 9 (T-DNA selection query sequence) shows that T-DNA sequences are observed only in the test sample and not in the control. This is expected, since only the test sample had been experimentally transformed with the PV-GMGOX20 plasmid. No sequences similar to the backbone portion of the transformation plasmid were detected in either the test or control samples (backbone selection query sequence in Fig. 9) demonstrating that backbone-like sequences are not naturally present in the *Glycine max* variety used for transformation and that no backbone sequences were introduced during the transformation event that produced the GM soybean MON17903. The selected sequences were used for further experimental characterizations (the Le1 gene sequences were used to establish effective sequencing depth as discussed above, and any plasmid-like sequences were used to establish locus and copy number as discussed below).

### Junction Detection

Using the selected dataset of reads derived from the test samples we identified and classified novel junction sequences characteristic of genomic insertion events. All reads with partial matches to the transformation plasmid were classified and characterized. In the case of the test material, MON17903, only two classes of junction sequences were observed (Fig. 10). Consistent with the Southern blot analysis, the observation of a single pair of junction sequence classes is characteristic of a
single insertion site with no internal rearrangements or duplications. The junction sequences detected showed the expected left and right border sequences from the transformation plasmid (panel B in Fig. 10) and flanking regions from the same native soybean genomic locus.

The organization of the single inserted locus and its presence as a single copy is demonstrated by locus-specific sequencing of PCR products (identical to the Southern blot-based methodology, described above). The alignment of this sequence data with the detected junction sequences (100% nucleotide identity and 100% match length of the consensus junction sequences JSC-A and JSC-B with the expected junction points [panel C in Fig. 10]) confirms that the two observed junction sequence classes are linked and spanned by known, expected, and contiguous single copy DNA sequence at a single locus within the soybean genome. The nature of any modifications to the native locus at which the insertion occurred is studied by comparison to wild-type locus determined by sequencing of specific PCR products (identical to the current methodology; not shown). These data uniquely identify the insertion locus and the detail of the precise insertion as well as establishing the overall number of insertion loci.

To demonstrate the method is capable of detecting DNA inserts with more complicated features such as rearrangement and multiple inserts, another Agrobacterium-mediated transformation event MON87704 was analyzed. MON87704 is a soybean event transformed with the binary vector PV-GMPQ/HT4404 that contains two T-DNAs (Fig. 3). Previous molecular characterization showed that the two T-DNAs (T-DNA I and T-DNA II, respectively) co-integrated into one locus of the MON87704 genome as shown in panel A in Fig. 8. Genomic DNA from MON87704 was subjected to sequencing and JSA using the same method as described for MON17903. From this analysis, four junction sequence classes are detected that characterize the event, two representing the T-DNA to genome junctions and two at the point where T-DNA I and T-DNA II sequences from the binary vector were joined during transformation to create novel junction sequence characteristic of the rearrangement (panel B in Fig. 8). In this complex case the analysis described above is able to detect all of the expected junction sequences created on the insertion of multiple T-DNAs and their in vivo rearrangement.

The organization of the complex inserted MON87704 locus and the presence of two complete rearranged T-DNAs is demonstrated by sequencing of locus-specific PCR products. The alignment of this sequence data with the detected junction sequences (100% nucleotide identity and 100% match length of the consensus junction sequences JSC-C, JSC-D, JSC-E, and JSC-F with the expected junction points [panel B in Fig. 8]) confirms that the four observed junction sequence classes are linked and spanned by the known, expected, and contiguous DNA sequence.

**Junction Sequence Analysis-Based Molecular Characterization – Conclusions**

Taken together the results from JSA and locus-specific PCR and sequencing (i.e., detection of a single linked pair of junction sequences; Fig. 10) demonstrate that the GM soybean MON17903 contains a single copy of the T-DNA inserted at a single genomic locus. No detectable backbone sequences in
Junction Sequence Class A:

![DNA sequence diagram]

Junction Sequence Class B:

![DNA sequence diagram]

Panel B. Detected Junction Sequences
the test dataset (Fig. 9, Backbone) along with the appropriate controls for experimental comprehensiveness (Le1 gene) and detection sensitivity (plasmid spike-in controls) demonstrates that no unintended backbone DNA fragments have been inserted during the transformation process.

The molecular characterization and conclusions are identical for MON17903 by both the Southern blot-based method and the JSA method. Therefore, these two methods are functionally equivalent methods for such characterizations. Additionally, we show that the method is capable of detecting complex cases of multiple T-DNAs and rearrangements of inserted DNA as in the case of MON87704.

**Other Considerations**

Since the sequencing coverage performed for this method was deep (>75x) and parameters chosen for read enhancement are optimized to be sensitive, extremely small amounts of sample contamination can be detected. Even though any contaminating sequences are easily distinguished from true transformation plasmid derived reads through sequence analysis, we have found it useful to take a variety of steps that help avoid such low-level contamination. Specifically, potential surface contaminants (e.g., adventitiously present bacteria, laboratory dust, etc.) are removed by washing our starting materials as described in the Material and Methods section. All lab procedures are performed in a dedicated clean facility equipped with the appropriate contamination management equipment, work practices, and policies (e.g., positive flow hoods, dedicated equipment, restricted access). These precautions have proven to be effective at eliminating contaminants from our current studies, including the data presented here.

Occasionally, T-DNA constructs will be designed to contain elements similar or identical to endogenous elements found in the transformation line (e.g., promoters, introns, and coding sequences). To accurately distinguish between transformed and endogenous sequences a screen is conducted against the endogenous genome sequence (as described in the Sequencing Read Selection section of the Materials and Methods above). In practice this screen has proved successful in identifying and removing reads originating from the native genome sequence, that is, endogenous regions that have similarity to the selection query sequence. This screen is conducted using either a reference genome sequence or the control NexGen dataset (the later being generated with each experiment is an asset in cases where no reference genome sequence is available). This screen has proved effective at preventing the detection of false junction sequences arising from endogenous sequences.

The JSA method described here provides a number of advantages compared to current Southern blot-based methods. Notably JSA is experimentally simpler, and its experimental design remains consistent across different constructs, transformation events, and species. Unlike Southern blot analysis there is no experimental variation based on T-DNA composition or insertion locus, both of which necessitate the need for unique experimental designs in every case when using Southern blots. The greater simplicity brings with it increases in experimental efficiency, greater ease of data interpretation, and greater predicted reproducibility due to the decreased dependence on custom and complex lab procedures. In contrast to Southern blot analysis, the JSA method also has the potential for scale-up and automation as required. Planning and capacity management of labs using the new method will also be improved. Currently the cost...
associated with this method is significantly less than alternatives within a USEPA Good Laboratory Practice environment as specified in 40 CFR part 160 (requiring less than 50% of the funding and manpower vs. Southern blot-based studies), and these costs are projected to decrease further as sequencing technologies advance.

It is interesting to note that the method described here provides a digital sequencing-based method analogous to the previously described “Detection of Specific Sequences Among DNA Fragments Separated by Gel Electrophoresis” (Southern, 1975). Similar to the previous study, here we describe the method and demonstrate its use on a limited set of test cases; however, practically in our hands this new method has proved successful in all cases we have tested so far and we expect it to be widely applicable regardless of specific sequence under detection or its genome location.

Finally, we note that generational stability studies (the stable maintenance of any particular loci across a number of generations of progeny) can be achieved by using JSA described here on the various generations under study. Stable integrations will produce consistent patterns of detected junction sequence among all generations studied.

Conclusions

Using a typical GM soybean event we have demonstrated that NexGen sequencing and JSA bioinformatics provides molecular characterization equivalent to the current Southern blot-based method. Additionally we have demonstrated that the method presented is capable of detecting complex events including those with multiple T-DNAs and sequence rearrangements. Next generation sequencing and JSA bioinformatics offers multiple advantages over the current Southern blot-based approach to molecular characterization, most notably the simplicity, efficiency and consistency of the method, and it provides a viable alternative for use in efficiently and robustly establishing the molecular characteristics of crops improved by modern biotechnology.

Supplemental Information Available

The following supplemental information is included with this article:

Supplemental Table S1. Summary data for the next generation (NexGen) sequencing.

Supplemental Table S2. Selection of sequences and data normalization for test and control data sets using backbone, endogenous (internal standard), or transfer DNA (T-DNA) sequence.

Acknowledgments

We thank Jason Ward, Tracey Reynolds, Jan Verhaert, Shuichi Nakai, Chrisry Lawrence, Austin Burns, and Jim Masacci for providing critical suggestion to improve this paper. For freely providing their technical advice and assistance we thank Dan Ader, Bo-Xing Qiu, Xiexfeng Zhou, Randy Kerstetter, Kim Lawry, Dan Stoffey, Holly Gossman, Kelly Klug, Amanda Albers, Phil Ltrapelle, and Elise Waldman. We also thank Tracey Cavato, Erik Jacobs, and Ron Lirette for their overall guidance and support. Due to the fact that the Editor Dave Somers and the authors of this paper are from the same institution, Charlie Brummer assumed the role of editor and handled all aspects of the review of this paper. Special thanks are extended to Charlie.

References


CLC bio. 2012. CLC main workbench v. 6.6.6, assemble sequences. CLC bio, Cambridge, MA.


