Title:

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

David Kovalic1,*, Carl Garnaat1, Liang Guo2, Yongpan Yan1, Jeanna Groat1, Andre Silvanovich1, Lyle Ralston1, Mingya Huang1, Qing Tian1, Allen Christian1, Nordine Cheikh1, Jerry Hjelle1, Stephen Padgette1 and Gary Bannon1

Received ________________________.

* Corresponding Author email address: david.k.kovalic@monsanto.com

1 Monsanto Company, 800 North Lindbergh Blvd., St. Louis, MO  63167, USA

2 Bayer Crop Science, 2 T.W. Alexander Dr., RTP, NC  27709, USA
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 PCR products spanning any inserted DNA to complete the characterization process. Here we describe the use of 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 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 T-DNA inserts in order to demonstrate that the new method is effective at identifying complex cases. NexGen 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.
Introduction:

Molecular characterization is a key step in the assessment of 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 T-DNA backbone sequence) within the GM event. Generational stability analysis, which demonstrates the stable heritability of inserted DNA sequences over a number of breeding generations, is also routinely conducted.

Currently, molecular characterization is achieved by a combination of Southern blot analysis (that determines insert and copy numbers of the event, presence or absence of the backbone, and generational stability (Southern 1975)); and sequencing overlapping PCR fragments spanning the inserted DNA (that 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 NexGen DNA sequencing (e.g. 454...
(454 Life Sci. Branford CT), Illumina (Illumina Inc. San Diego, CA) and SOLiD (ABI, Foster City, CA)), which provide dramatically increased sequencing throughput versus capillary electrophoresis sequencing technologies (e.g. ABI 3700 series), it has previously been established that it is possible to achieve comprehensive coverage of complex genomes given 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 utilizing 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/copy number and the presence/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 (Figure 1, Panel A). 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 $^{32}$P-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 Figure 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 employed 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) in order 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 (Figure 3) were used in conjunction with the control material to develop the GM events studied here. These plasmid DNAs serves 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 samples were prepared for NexGen sequencing library construction, genomic DNA 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, rinsing with tap water, followed by submerging in 0.5% NaOCl for 1 min, 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, Wheeling, IL) for 3 mins. 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 utilizing 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 EDTA pH 8.0, 1.05 M NaCl, 0.75% (w/v) PVP (MW 40,000)]. The samples were incubated at ~65° for 60 minutes 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 minutes. The samples were centrifuged at 12000 x g for 10 minutes and the aqueous phase transferred to a clean tube. The PCI extraction and centrifugation steps were repeated 2 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 2 volumes of 100% ethanol. The DNA was centrifuged at 5000 x g for 5 minutes and the DNA pellets were rinsed with 70% ethanol and air dried. The DNA was dissolved in TE buffer and then precipitated by the addition of equal volumes of precipitation buffer (20% w/v PEG(MW 5,000), 0.25 M NaCl). After incubation at 37° for 15 minutes, the DNA was centrifuged at 5000 x g for 15 minutes and rinsed with 70% ethanol. The DNA was air dried and re-dissolved in TE buffer (pH 8.0).
All extracted DNA and quantified using a Qubit™ Fluorometer (Invitrogen, Carlsbad, CA) 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, San Diego, CA,) in accordance with the manufacturer’s protocols. DNA was sheared using a Covaris S2 ultrasonicator (Covaris, Inc., Woburn, MA), and the resultant dsDNA 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., Beverly, MA) to select DNA for fragments of an average
insert size of 280 bp. DNA fragments with adapter molecules were selectively enriched by 10 cycles of PCR, and the libraries were quantified using an Bioanalyzer 2100 using a High-Sensitivity DNA kit (Agilent Technologies, Santa Clara, CA). Indexed adapters were used to allow for the multiplexing of samples for sequencing and to ensure sample/data integrity. Eight separate indexed libraries were produced for both the test and control soy samples, and these 8 libraries were used to create test and control library pools (in equimolar ratios) prior to sequencing.

To produce “spike-in” positive control samples for sequencing, plasmid DNA libraries were created as described above and then diluted to 1 and 1/10 soy genome equivalents (concentrations equivalent to single copy or 1/10 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., San Diego, CA) 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).

Sequencing Read Selection

Reads for further analysis were selected based on their similarity to different “selection query sequences” depending on the intended analysis: either a well known single copy locus was used for
Effective Sequencing Depth Determination, or the transformation plasmid was used for JSA. These two analyses are described below.

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 (v. 2.2.21) (Altschul et al. 1990; Altschul et al. 1997). DNA sequencing reads with a match to the query sequence having an e-value $1e^{-5}$ 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).

In order to identify remove any selected reads originating from native soybean genome sequence (specifically from any regions which 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 (v. 2.2.21) 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 10bp).

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 3-42) of each selected DNA sequence was compared to all other selected sequencing reads with the alignment software Bowtie v. 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 v. 2.06.09 (http://www.novocraft.com) was used to remove the unaligned adapter 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 adapter 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, 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 100× 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 non-plasmid sequence or non-contiguous 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 in order 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 PCR 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. 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 as shown in Figure 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 dNTP, and 0.02 units/μl of Phusion Hot Start II High Fidelity DNA Polymerase (New England Biolabs (NEB), Beverly, MA). The amplification was performed under the following cycling conditions: 98°C for 30 sec; 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. Prior to sequencing of the PCR products, they were purified using the QIAquick PCR Purification Kit (Qiagen, Inc., Valencia, CA). The PCR products were sequenced on an ABI 3730 capillary machine (Applied BioSystems, Carlsbad, CA) 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, Cambridge, Massachusetts). 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 polymerase chain reaction (PCR) amplification. Approximately 25 ng of each probe template was labeled with 32P-dCTP (~6000 Ci mmol⁻¹; GE Healthcare Biosciences, Pittsburgh, PA) by the random priming method (RadPrime DNA Labeling System, Life Technologies,
Carlsbad, CA). Multiple exposures of each blot were then generated using Kodak Biomax MS film (Eastman Kodak, Rochester, NY) 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 Figure 5. This method utilizes deep NexGen sequencing, locus-specific sequencing and bioinformatics to achieve the goals of molecular characterization in place of Southern blot analysis as shown in Figure 1, panel A.

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 GM event (Ming et al. 2008). In this sequencing of the papaya genome the authors used 1.6 M whole-genome shotgun (WGS) reads derived from a variety of plasmid and BAC libraries which were constructed with a range of insert sizes. These WGS reads totaled ~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 methods (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 base in the soybean genome. It has previously been established experimentally 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 know biases in NexGen sequencing techniques, including the Illumina sequencing by synthesis method employed here (Minoche et al. 2011).
In this study effective experimental coverage was estimated by examining a well known single copy locus, and sufficient dept 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 corn) and found that by employing a carefully chosen, conserved and representative single-copy locus provides 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 which 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 theses 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 method 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 (Figure 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 Ase I and Sty I both cleave once within the inserted DNA and once within the known 5' and 3' flanking sequences. Restriction digestion of genomic DNA with Ase I or Sty I 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 (Figure 6, Panels A and B).

In addition, three overlapping probes that span the entire backbone of PV-GMGOX20 were used for Southern blot analysis. MON17903 DNA digested with Ase I or Sty I showed no detectable hybridization signal (Figure 7, Panels A, B, and C). 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 PCR 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 Figure 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 vector 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 Figure 10.
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 >75× average depth of coverage we expect, on average, each base 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). By selecting sequences similar to the known endogenous single copy locus Le1 (Figure 8, ‘Endogenous’ selection query sequence) 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); and 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 (Supplementary Table 1).

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 observe complete coverage of our reference material at both the 1 genome equivalent and 1/10 genome equivalent (100% nucleotide identity over 100% of the transformation plasmid, not shown). This result demonstrates that all bases of the transformation plasmid are observed by the sequencing and bioinformatics performed in this study down to a level of at least 1/10 genomic equivalent and hence the achievement of a detection limit of at least 1/10 copy number for the plasmid DNA sequences in this study.
Sequencing Read Selection

In order 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 alignment. 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). 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 endogenous control sequence (Le1) are shown in Supplementary Table 2 and Figure 8.

Examination of Figure 8 (‘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 (Figure 8, ‘Backbone’ selection query sequence) 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 Lec1 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 (Figure 9). 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 (Figure 9, Panel B) 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 (Figure 9, Panel C) confirms that the two observed junction sequence classes are linked and spanned by known, expected, 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.

In order 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 (Figure 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 Figure 10, Panel A. 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 which characterize the event, two representing the T-DNA/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 (Figure 10, panel B). In this complex case the analysis pipeline described above is able to detect all of the expected junction sequences created upon 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 (Figure 10, Panel B) 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; Figure 9) 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 the test dataset (Figure 8, lane B) along with the appropriate controls for experimental comprehensiveness (Lec1 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 >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). In order 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 any reads originating from the native genome sequence, i.e. any endogenous regions which 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 utilizing the new method will also be improved. Currently the cost associated with this method is significantly less than alternatives within a GLP environment (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 in practice 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 amongst all generations studied.
Conclusions:

Using a typical GM soybean event we demonstrated that NexGen sequencing and bioinformatics method produced equivalent data to that generated by the current Southern blot based methods in achieving molecular characterization of a GM event. Additionally we demonstrate that the method presented is capable of detecting complex events including those with multiple T-DNAs and sequence rearrangements. NexGen 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.
Acknowledgements:

We thank Jason Ward, Tracey Reynolds, Jan Verhaert, Shuichi Nakai, Chrissy Lawrence, Austin Burns and Jim Masucci for providing critical suggestion to improve this paper. For freely providing their technical advice and assistance we thank Dan Ader, Bo-Xing Qiu, Xuefeng Zhou, Randy Kerstetter, Kim Lawry, Dan Stoffey, Holly Gossman, Kelly Klug, Amanda Albers, Phil Latreille and Elise Waldman. We also thank Tracey Cavato, Erik Jacobs and Ron Lirette for their overall guidance and support.
References:


Figure Captions and Tables:

Figure 1. Molecular methods and their resultant characterization endpoints. Panel A details the current methodology (reliant on Southern blot analysis); the JSA method described in this paper and the resultant characterization endpoints. Panel B shows locus specific 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 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/genomic flanking junctions. These novel sequences are characteristic of the transformed locus and are detectable by NexGen sequencing and bioinformatics. Throughout this paper we term these sequences ‘Junction Sequences.

Figure 3. Panel A: Circular map of the transformation plasmid PV-GMGOX20 containing the T-DNA used for Agrobacterium mediated transformation to create MON17903. Panel B: Circular map of the transformation plasmid PV-GMPQ/HT4404 containing the two T-DNAs used for Agrobacterium mediated transformation to create MON87704.

Figure 4. A linear map of the insert and adjacent DNA flanking the insert in MON17903. Identified on the map are the locations of the T-DNA border regions as well as restriction sites with positions relative to the size of the linear map for enzymes used in the Southern analysis. The T-DNA insert is depicted by the thick black bar shown between the 5' and 3' genomic flanks. The position of T-DNA probes used for Southern blot analysis is shown (labeled Probe 1-3). The positions of the PCR amplicons produced for directed sequencing of the locus are also shown (PCR Amplicon 1-3). The resultant consensus sequence covers the entire T-DNA insert, 989 bp of 5' flanking sequence and 930 bp of 3' flanking sequence.

Figure 5. Genomic DNA from the test and control material were sequenced using Illumina HiSeq/TruSeq technology (Illumina, Inc.) that produces large numbers of short sequence reads approximately 100 bp in length. Sufficient numbers of these sequence fragments were obtained to...
comprehensively cover the genomes of each sample at >75× average coverage. Utilizing these genome sequence reads, bioinformatics search tools were used to select all sequence reads that are significantly similar (as defined in the text) to the transformation plasmid. Only the selected sequence reads were used in further bioinformatics analysis to determine the insert number by detecting and characterizing all junction sequences and the presence/absence of the plasmid backbone sequences by lack of detectable sequences, including the use of suitable controls for experimental comprehensiveness and sensitivity.
**Figure 6. Southern Blot Analysis of MON17903 Using T-DNA Probes.**

**Panel A (Probe 1 and Probe 3)**
The blot was hybridized with two $^{32}$P-labeled probes that spanned a portion of the T DNA sequence (Figure 4, Probe 1 and Probe 3). Each lane contains ~10 µg of digested genomic DNA isolated from leaf tissue. Lane designations are as follows:

<table>
<thead>
<tr>
<th>Lane</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Parental Conventional Control (Ase I)</td>
</tr>
<tr>
<td>2</td>
<td>MON17903 (Ase I)</td>
</tr>
<tr>
<td>3</td>
<td>Parental Conventional Control (Sty I)</td>
</tr>
<tr>
<td>4</td>
<td>MON17903 (Sty I)</td>
</tr>
<tr>
<td>5</td>
<td>Parental Conventional Control (Sty I) spiked with Probe 1 and Probe 3 (~1.0 genome equivalent)</td>
</tr>
<tr>
<td>6</td>
<td>Parental Conventional Control (Sty I) spiked with Probe 1 and Probe 3 (Bam HI / Sca I) (~0.1 genome equivalent)</td>
</tr>
<tr>
<td>7</td>
<td>Parental Conventional Control (Sty I) spiked with PV-GMGOX20 (Bam HI / Sca I) (~1.0 genome equivalent)</td>
</tr>
<tr>
<td>8</td>
<td>Parental Conventional Control (Ase I)</td>
</tr>
<tr>
<td>9</td>
<td>MON17903 (Ase I)</td>
</tr>
<tr>
<td>10</td>
<td>Parental Conventional Control (Sty I)</td>
</tr>
<tr>
<td>11</td>
<td>MON17903 (Sty I)</td>
</tr>
</tbody>
</table>

Arrows denote the size of the DNA, in kilobase pairs, obtained from the 1kb DNA Extension Ladder (Invitrogen) on the ethidium bromide stained gel. Long run (32 volts 14.5 hours, followed by 35 volts 8 hours) or Short run (35 volts ~8 hours) are indicated above the lanes.

**Panel B (Probe 2)**
The blot was hybridized with one $^{32}$P-labeled probe that spanned a portion of the T DNA sequence (Figure 4, Probe 3). Each lane contains ~10 µg of digested genomic DNA isolated from leaf tissue. Lane designations are as follows:

<table>
<thead>
<tr>
<th>Lane</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Parental Conventional Control (Ase I)</td>
</tr>
<tr>
<td>2</td>
<td>MON17903 (Ase I)</td>
</tr>
<tr>
<td>3</td>
<td>Parental Conventional Control (Sty I)</td>
</tr>
<tr>
<td>4</td>
<td>MON17903 (Sty I)</td>
</tr>
<tr>
<td>5</td>
<td>Parental Conventional Control (Sty I) spiked with PV-GMGOX20 (Bam HI / Sca I) (~1.0 genome equivalent)</td>
</tr>
<tr>
<td>6</td>
<td>Parental Conventional Control (Sty I) spiked with PV-GMGOX20 (Bam HI / Sca I) (~0.1 genome equivalent)</td>
</tr>
<tr>
<td>7</td>
<td>Parental Conventional Control (Ase I)</td>
</tr>
<tr>
<td>8</td>
<td>MON17903 (Ase I)</td>
</tr>
<tr>
<td>9</td>
<td>Parental Conventional Control (Sty I)</td>
</tr>
<tr>
<td>10</td>
<td>MON17903 (Sty I)</td>
</tr>
</tbody>
</table>
Arrows denote the size of the DNA, in kilobase pairs, obtained from the 1kb DNA Extension Ladder (Invitrogen) on the ethidium bromide stained gel. Long run (32 volts 14.5 hours, followed by 35 volts 8 hours) or Short run (35 volts ~8 hours) are indicated above the lanes.
Figure 7. Southern Blot Analysis of MON17903 genomic DNA Using Backbone Probes

Each blot was hybridized with one $^{32}$P-labeled probe that spanned a portion of the T DNA sequence, each lane contains ~10 µg of digested genomic DNA isolated from leaf tissue. Lane designations are as follows:

<table>
<thead>
<tr>
<th>Lane</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Parental Conventional Control (Ase I)</td>
</tr>
<tr>
<td>2</td>
<td>MON17903 (Ase I)</td>
</tr>
<tr>
<td>3</td>
<td>Parental Conventional Control (Sty I)</td>
</tr>
<tr>
<td>4</td>
<td>MON17903 (Sty I)</td>
</tr>
<tr>
<td>5</td>
<td>Parental Conventional Control (Sty I) spiked with PV-GMGOX20 (Bam HI / Sca I) [~1.0 genome equivalent]</td>
</tr>
<tr>
<td>6</td>
<td>Parental Conventional Control (Sty I) spiked with PV-GMGOX20 (Bam HI / Sca I) [~0.1 genome equivalent]</td>
</tr>
<tr>
<td>7</td>
<td>Parental Conventional Control (Ase I)</td>
</tr>
<tr>
<td>8</td>
<td>MON17903 (Ase I)</td>
</tr>
<tr>
<td>9</td>
<td>Parental Conventional Control (Sty I)</td>
</tr>
<tr>
<td>10</td>
<td>MON17903 (Sty I)</td>
</tr>
</tbody>
</table>

Panel A; Backbone Probe 1; Panel B: Backbone Probe 2; Panel C: Backbone Probe 3 (see Figure 4). Arrows denote the size of the DNA, in kilobase pairs, obtained from the 1kb DNA Extension Ladder (Invitrogen) on the ethidium bromide stained gel. Long run (30 volts 14.5 hours, followed by 40 volts 5.5 hours) or Short run (40 volts 5.5 hours) are indicated above the lanes.
Figure 8. Bar graph of sequence read selection results.
Three different query sequences (PV-GMGOX20 Backbone; the endogenous single copy Le1 gene; or PV-GMGOX20 T-DNA) were used to select
sequences from either the MON17903 test or A3244 control Illumina sequence datasets (labeled ‘Sequencing Sample’) as described in the text.
The number of selected bases (normalized per Kb of Selection Query Sequence length, and by sample effective coverage to a 1X depth) in each
dataset is shown in this graph.

Figure 9. Junction sequences detected by JSA.
Panel A: Linear map of the event illustrating the relationship of the detected junction sequences to the event locus (note the junction sequences,
schematically represented by red lines, are for illustration purposes and their lengths are not drawn to scale to aid viewing). Panel B: trimmed
nucleotide alignment of the detected junction sequences. The detected junction sequences are ordered by junction sequence class (defined by
break point as discussed in the text) and then by alignment start site. The junction point between the T-DNA border and genomic flanking
sequence is indicated by the “^” character. Both alignments are trimmed to include only the 30 plasmid bases proximal to the break point, as
well as all sequence of the genomic flank. Both alignments are shown 5’->3’ beginning with the detected plasmid sequence. Panel C: Full
consensus sequence for JSC-A and JSC-B showing perfect alignment to the independently determined in planta locus specific sequence. Insert-
genomic flank junction points are indicated by the “^” character.

Figure 10. Schematic Representation of the Insert DNA in Soybean MON87704
Panel A represents a linear map of the T-DNA I and T-DNA II arrangement in the transformation plasmid, PV-GMPO/HT4404. The horizontal
arrows indicate the sequence direction of the elements. The numbers at the bottom are the corresponding coordinates on PV-GMPO/HT4404.
Panel A also represents a linear map of the arrangement of the two T-DNAs as they exist in MON87704. The horizontal arrows indicate the
sequence direction of the elements. The numbers at the bottom are the corresponding coordinates on MON87704. In MON87704, the T-DNA II
region between 2 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 base locations of the T-DNA I and T-DNA II regions in the transformation plasmid PV-GMPO/HT4404 and those of in
MON87704 insert sequence. Also shown is the correspondence of the junction sequences determined from high throughput sequencing and 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.
**Supplementary Data:**

**Supplementary Table 1.**

Summary data for the NexGen sequencing conducted in this study. For each sample the raw data produced is presented in terms of total read number and total nucleotide number. Effective coverage is assayed by mapping and alignment of all raw data to a well known single copy locus within the soybean genome (Le1). The median and the range of effective coverage depth is shown for both samples.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sequencing Reads Produced</th>
<th>Raw Data (Gb)</th>
<th>Effective Coverage Depth of (x-fold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test (MON17903)</td>
<td>1,139,538,382</td>
<td>113.2</td>
<td>97x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>62x-139x</td>
</tr>
<tr>
<td>Control (A3244)</td>
<td>1,322,787,632</td>
<td>130.3</td>
<td>102x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60x-136x</td>
</tr>
</tbody>
</table>
Supplementary Table 2.
Selection of sequences and data normalization for test and control data sets using backbone, endogenous (internal standard) or T-DNA sequence.

<table>
<thead>
<tr>
<th>Query Sequence</th>
<th>Sample</th>
<th>Number of Selected Reads</th>
<th>Cumulative Length of Selected Reads (bp)</th>
<th>Mean Coverage (X-fold)</th>
<th>Normalized Selection Metric (kb of Selected Sequence / kb Query Sequence per 1x Coverage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backbone</td>
<td>Test (MON17903)</td>
<td>0</td>
<td>0</td>
<td>94.4</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Control (A3244)</td>
<td>0</td>
<td>0</td>
<td>100.0</td>
<td>0.00</td>
</tr>
<tr>
<td>Endogenous</td>
<td>Test (MON17903)</td>
<td>2017</td>
<td>195642</td>
<td>94.4</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>Control (A3244)</td>
<td>2135</td>
<td>207838</td>
<td>100.0</td>
<td>1.01</td>
</tr>
<tr>
<td>T-DNA</td>
<td>Test (MON17903)</td>
<td>4452</td>
<td>428599</td>
<td>94.4</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>Control (A3244)</td>
<td>0</td>
<td>0</td>
<td>100.0</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Panel A.

**Current Methodology**

**Southern Blot Analysis**

- Step 1: Restriction enzyme digestion of genomic DNA, gel electrophoresis and blotting
- Step 2: Probe production, hybridization and imaging
- Step 3: Image interpretation

**New Methodology**

**Junction Sequence Analysis (JSA)**

- Step 1: High throughput sequencing of genomic DNA samples and quality control. A collection of 100-mer sequences are generated which comprehensively cover the sample genome.
- Step 2: Selection of all 100-mers containing sequence statistically similar to that of the transformation plasmid
- Step 3: Bioinformatics to find and characterize all selected 100-mer sequences originating from transgenic insertions (Junction Sequence Detection)

**Molecular Characterization Achieved**

1) Insert number determined:
   - Band pattern (Southern) or Junction sequence pairs detected (JSA)
2) Presence or absence of unintended insertions:
   - No band with backbone probe (Southern) or No unintended junction sequences detected (JSA)

Panel B.

**Common Methods**

**Locus Specific PCR and Sequencing**

- Step 4: Directed sequencing across the transgenic locus from 5’ flank to 3’ flank
- Step 5: Directed sequencing across wild type locus

**Molecular Characterization Achieved**

1) Exact sequence of insert(s) determined
2) Organization, intactness and copy number of genetic elements demonstrated using insert sequence
3) Integrity and organization of the insertion site(s)
Panel B.

PV-GMPQ-HT4404
13088 bp
The Plant Genome: Published online 15 Oct. 2012; doi: 10.3835/plantgenome2012.10.0026
Sequence Fragment Generation and Selection.

- Billions of randomly distributed 100-mers (>75X Coverage)
- Thousands of selected 100-mers

Insert Number Determination.

- Illumina HiSeq/TruSeq Sequencing
- Plant Genome
- Query Sequence
- Construct Sequence
- T-DNA
- Backbone
- Flank
- Flank

Sequence Quality Refinement and Junction Sequence Analysis

The Plant Genome: Published online 15 Oct. 2012; doi: 10.3835/plantgenome2012.10.0026
Panel A.

Panel B.
Panel A.
Full JSC-A Consensus versus *In planta* sequence

Score = 325.6 bits (164),  Expect = 9E-93
Identities = 164/164 (100),  Gaps = 0/164 (0)
Strand = Plus/Plus

Query 1  | TTTTTTAATTTCGACGTGTCGTTAGTTAGTTGTAACCTGGTATCCCGACACGTCGTCGTTT | 60
         | ............................................................................... |
Sbjct 917 | TTTTTTAATTTCGACGTGTCGTTAGTTAGTTGTAACCTGGTATCCCGACACGTCGTCGTTT | 976

Query 61 | CCTAGTTCTAACTGGAAGGGGGAACGACAATCTGATCCCCATCAAGCTCTAGCTAGAGC | 120
         | ............................................................................... |
Sbjct 977 | CCTAGTTCTAACTGGAAGGGGGAACGACAATCTGATCCCCATCAAGCTCTAGCTAGAGC | 1036

Query 121 | GGCCCGCTTATCAAGCTTCTGCAGGTCCTGCTGGATGGAAGCT | 164
           | ............................................................................... |
Sbjct 1037 | GGCCCGCTTATCAAGCTTCTGCAGGTCCTGCTGGATGGAAGCT | 1080

Full JSC-B Consensus versus *In planta* sequence

Score = 317.7 bits (160),  Expect = 2E-90
Identities = 160/160 (100),  Gaps = 0/160 (0)
Strand = Plus/Plus

Query 1  | GATCGTAATTTGTCGTTTTATCAAAATGTACTTTCATTTTATAATAACGCTGCGGACATC | 60
         | ............................................................................... |
Sbjct 5221 | GATCGTAATTTGTCGTTTTATCAAAATGTACTTTCATTTTATAATAACGCTGCGGACATC | 5280

Query 61 | TACATTTTTGAATTGAAAAAAAATTGGTAATTACTCAGTTGTCAATGTCGGTCTCCGGTG | 120
         | ............................................................................... |
Sbjct 5281 | TACATTTTTGAATTGAAAAAAAATTGGTAATTACTCAGTTGTCAATGTCGGTCTCCGGTG | 5340

Query 121 | GCGCCGTCGTCGCCGCGGAATGTGAGGAGGAACCACCG | 160
           | ............................................................................... |
Sbjct 5341 | GCGCCGTCGTCGCCGCGGAATGTGAGGAGGAACCACCG | 5380

Panel C.
Panel A. Schematic Diagram and Coordinates of T-DNA I and T-DNA II arrangement of PV-GMPQ/HT4404 in the transformation plasmid and the MON87704 event.
Panel B. Diagram of T-DNA I and T-DNA II on MON87704 and Alignments of Junction Sequences Consensi with in planta locus specific sequence