Published March 13, 2015

ORIGINAL RESEARCH

HtStuf: High-Throughput Sequencing to Locate Unknown DNA Junction Fragments

Lisa B. Kanizay,* Thomas B. Jacobs, Kevin Gillespie, Jade A. Newsome, Brittany N. Spaid, and Wayne A. Parrott

Abstract

Advances in high-throughput sequencing have led to many new technologies for assessing genomes and population diversity. In spite of this, inexpensive and technically simple methods for efficiently pinpointing the location of transgenes and other specific sequences in large genomes are lacking. Here we report the development of a modified TA cloning and Illumina sequencing method called high-throughput sequencing to locate unmapped DNA fragments (HtStuf). Transgenic insertion sites were identified and confirmed in nine out of 10 transgenic soybean [*Glycine max* (L.) Merr.] lines, and major rearrangements of the transgene were detected in these lines. Additionally this method was used to map insertions of the introduced DNA transposon, *mPing*, in four T6 lines derived from a single event. Fifteen of the *mPing* insertion sites were validated with polymerase chain reaction. Together, these data demonstrate the simplicity and effectiveness of this novel sequencing method.

THE REDUCED COST and increased power of high-throughput sequencing has allowed many genomes to be resequenced as a means of assessing genome diversity, identifying novel genes or quantitative trait loci (QTLs) (Huang et al., 2009; Hufford et al., 2012), and determining transgene integration sites (Kovalic et al., 2012; Ming et al., 2008). While whole-genome sequencing (WGS) and resequencing of genomes have proven beneficial for these purposes, they are cost prohibitive for a large number of samples, and the analysis of large datasets requires bioinformatics expertise. Additionally, resequencing genomes requires a reference genome for comparison. Certain questions are better addressed by sequencing smaller, defined regions with techniques such as genotyping-by-sequencing (GBS) (Elshire et al., 2011). When a sequence of interest inserts randomly into a genome by transformation or transposition, it is often essential to find its location without wasting resources by resequencing unmodified regions. Identifying the junction site of the inserted DNA (location in the genome immediately flanking the insertion) can aid in mapping promoter and enhancer traps, mutagenized genes, T-DNA insertions, or transposon tags and can improve the general understanding of the transformation process.

In addition to WGS, several polymerase chain reaction (PCR)-based methods exist to capture junction fragments

Published in The Plant Genome 8
doi: 10.3835/plantgenome2014.10.0070

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

T.B. Jacobs and W.A. Parrott, Institute for Plant Breeding, Genetics and Genomics, Univ. of Georgia, Athens, GA 30602. W.A. Parrott, Dep. of Crop and Soil Sciences, Univ. of Georgia, Athens, GA 30602. L.B. Kanizay, T.B. Jacobs, K. Gillespie, J.A. Newsome, B.N. Spaid, and W.A. Parrott, Center for Applied Genetic Technologies, Univ. of Georgia, Athens, GA 30602. L.B. Kanizay and T.B. Jacobs contributed equally to this manuscript. Received 22 Oct. 2014. Accepted 15 Dec. 2014. *Corresponding author (lkanizay@gmail.com).

Abbreviations: GBS, genotype-by sequencing; GOI, gene-of-interest; GSP, Gene specific primer; *hpt*, hygromycin phosphotransferase; QTL, Quantitative trait loci; TD, transposon display; WGS, whole-genome sequencing.

(summarized in Leoni et al., 2011), generally referred to as genome walking. These include thermal asymmetric interlaced-PCR (Liu and Whittier, 1995), TOPO vector-ligation PCR (Orcheski and Davis, 2010), transposon display (TD) (Van den Broeck et al., 1998), digestion-ligation-amplification (Liu et al., 2013), and inverse PCR (Ochman et al., 1988). Thermal asymmetric interlaced-PCR and similar methods require the use of random adapters and high melting temperature primers and can lead to false positives caused by nonspecific amplification, making it technically difficult to target a wide range of templates. Genome walking kits such as APAgene GOLD (Bio S&T Inc.) and Universal GenomeWalker 2.0 (Clontech) are fairly expensive and have stringent requirements that are not amenable to high-throughput procedures.

The PCR-based addition of Illumina adapters and barcodes has been used to generate transposon-specific sequencing libraries in maize (*Zea mays* L.) (McCarty et al., 2005). The maize Uniform*Mu* project uses adaptor ligation followed by PCR addition of barcodes and multiplexing for Illumina sequencing. This method is effective at identifying the location of the endogenous *Mu* transposons in maize. The generation and analysis of the sequence data generated in the Uniform*Mu* project is inhibited by the abundance of *Mu* elements in maize but aided by the ability to use phenotypic markers for detecting transposon activity. Since not all species have the same capacity for genetics as maize—or a plethora of well-characterized endogenous transposons—we aimed to develop a simplified sequencing strategy for the identification and mapping of foreign transposons and transgenes.

TOPO vector-ligation PCR uses a cloning vector in place of linear adapters, from which nested PCR primers can be designed to flank and amplify the insert (Orcheski and Davis, 2010). This procedure is advantageous since it reduces nonspecific amplification. In addition, nested universal primer sites exist in the vector for sequencing purposes, and the required materials are present in most molecular biology labs. The previously published TOPO vector-ligation protocol used cloning and Sanger sequencing to identify junction fragments (Orcheski and Davis, 2010). It should be possible to take advantage of the cleaner amplification provided by the vector-ligation PCR strategy and use the amplification products as templates for high-throughput Illumina sequencing.

The goal of the method presented here was to specifically sequence portions of the soybean [*Glycine max* (L.) Merr.] genome directly flanking either transgene or introduced transposon insertion sites. To increase throughput and depth of coverage, the method uses Illumina sequencing and TA cloning. The modified TOPO vector-ligation PCR yields site-specific amplicons that can be directly sequenced with Illumina technology, but with the correct primer sequences, any sequencing technology could be used. Ten transgenic lines were examined to determine the efficacy of the high-throughput sequencing to locate unknown DNA junction fragments (HtStuf) method in mapping single- or low-copy transgenes. Additionally, four previously developed lines containing the *mPing* DNA transposon were analyzed to determine the efficacy of HtStuf in mapping short, high-copy sequences. *mPing* is a small, 430-bp element originally identified in rice (*Oryza sativa* L.) (Naito et al., 2006) and previously transformed into soybean with the goal of generating mutations in soybean genes (Hancock et al., 2011). Here we show that HtStuf is useful in mapping long transgene cassettes as well as short, multicopy sequences in the soybean genome.

Materials and Methods

Plant Materials

Somatic embryos of soybean cultivar Williams82 were biolistically transformed as previously described (Hancock et al., 2011). The DNA used for bombardment was a gel-extracted 5081 bp *Pac*I (NEB) linear fragment, containing a hygromycin phosphotransferase (*hpt*) gene (for selection) under the control of the Ubi-3 *Solanum tuberosum* promoter (StUbiP) and terminator (StUbiT) (Garbarino and Belknap, 1994) and gene-of-interest (GOI) cassette driven by the *Glycine max* ubiquitin (GmUbi) promoter (Hernandez-Garcia et al., 2009) and *Pisum sativum* rubisco (rbcS) terminator (An et al., 1985) (Supplemental Fig. S1). DNA for library preparation was extracted from young leaves of T0 plants in the greenhouse. Ten events were analyzed.

To determine the zygosity of T1 plants from Event 16, the Invader assay (Hologic Corp.) was run on a Synergy 2 plate reader (BioTek Instruments, Inc.) according to manufacturer instructions, except using a single reaction for each plant. The assay contains a probe that produces fluorescence when bound to the *hpt* gene that is used to quantify the relative abundance of the target sequence in a genomic sample.

The four transgenic soybean lines (in cultivar Jack) that were used for *mPing* sequence analysis are T6 lines derived from a previously reported event, pPing 2-9 (Hancock et al., 2011).

Library Preparation and Primer Design

For an overview of the library preparation method see Fig. 1. DNA was collected from young soybean leaves and extracted using a modified cetyltrimethylammonium bromide (CTAB) protocol (Murray and Thompson, 1980). DNA quality was checked on a 1% agarose gel. Approximately 100 to 1000 ng of genomic DNA were fragmented to 1 to 5 kilobase (kb) with DNA Fragmentase (New England Biolabs) in 10-μL reactions, according to manufacturer instructions. Digestion times varied depending on the size of the initial DNA sample. Samples were digested for 30 min, which resulted in an average fragment size of 1 kb in most cases. After fragmentation, DNA samples were cleaned using Zymo Clean and Concentrator kit (Zymo Research) and eluted with 11 μL of 10 mM Tris-HCl. Eluted samples were run on a bioanalyzer with a high-sensitivity DNA

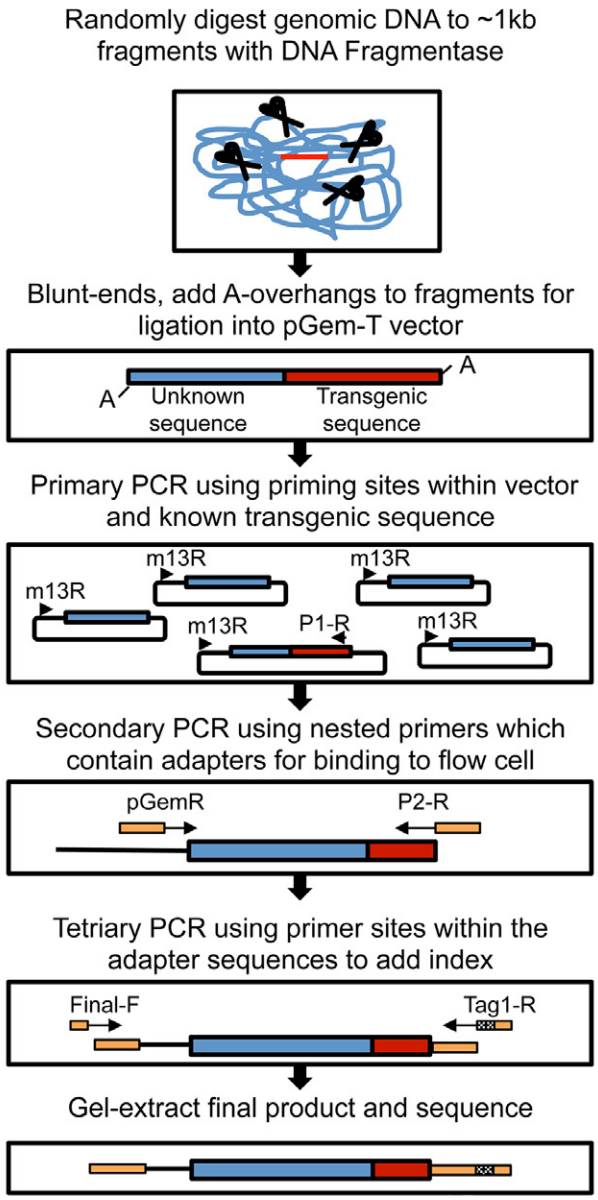

Figure 1. Library generation overview. Genomic DNA is first fragmented into ~1 kb pieces and ligated to the pGEM T-vector. The amplicon libraries are produced by a series of nested-polymerase chain reactions (PCRs) followed by gel extraction.

chip (Agilent Technologies) or on a Fragment Analyzer (Advanced Analytical) to confirm proper fragmentation.

Fragmented samples contain overhangs that need to be removed before A-tailing and ligation. The overhangs were removed using a T4 polymerase reaction (per reaction: 2 μL 10× NEB Buffer2, 2 μL 10× BSA, 1 μL 2mM dNTPs, 0.2 μL T4 DNA Polymerase, 3 μL DNA, 11.8 μL water, and incubated at 12°C for 15 min), which was stopped by immediately cleaning with Zymo columns and eluted with 8 μL 10 mM Tris-HCl. Samples were A-tailed and ligated to a pGEM-T Easy Vector System (Promega Corp.) according to manufacturer instructions. A 4°C overnight ligation was used. The ligation was diluted 1:10, and 1 μL was used as a template for primary PCR. Gene specific primer (GSP) 1 was used with the universal M13 reverse primer for primary PCR with KAPA 2X HiFi HotStart ReadyMix (KAPA Biosystems) (per reaction: 5 μL 2× ReadyMix, 0.3 μL of each 10 μM primer, 1 μL diluted ligation, water to 10 μL) with the following conditions: 95°C 3 min; 30 cycles (98°C 20 s, 60°C 15 s, 72°C 1 minute); 72°C 5 min; hold at 12°C. The PCR products were visualized on a 1% agarose, Tris-borate-EDTA (TBE) gel. A smear with a large product at approximately 3 kb was often observed (Supplemental Fig. S2). The primer sequences used in library preparation are in Supplemental Table S1.

The primary PCR products were then diluted 1:100, and 1 μL was used in a secondary touchdown-PCR reaction with the conditions 95°C 3 min; 10 cycles (98°C 20 s, 70 to 60°C 15 s (–1°C per cycle), 72°C 1 minute); 20 cycles (98°C 20 s, 60°C 15 s, 72°C 1 minute); 72°C 5 min; hold at 12°C) with nested primers GSP2 and pGEM reverse. The secondary primers contain 5′ tails that are used to start adding on the Illumina adaptor sequences. A touchdown-PCR method was used to ensure specific amplification. The PCR products were visualized on a 1% agarose, TBE gel. Multiple banding patterns and smears were typically observed. The PCR products were obvious at this point and greater than 250 bp in length; shorter fragments were likely primer dimers.

Secondary PCR products were diluted 1:100. At this point, if there were multiple amplicons per sample, all secondary reactions were pooled within a sample (this was the case for the samples from the 10 transgene events). The tertiary PCR primers bind the tails that were added during the secondary PCR and amplify any secondary product (including primer dimers). The tertiary adds 6-nt indexing barcodes (Faircloth and Glenn, 2012) to the samples and produces a final PCR product with complete Illumina TruSeq-style adapters. The tertiary cycle conditions are as follows: 95°C 3 min; 10 to 13 cycles (98°C 20 s, 60°C 15 s, 72°C 1 minute); 72°C 5 min; hold at 12°C. An aliquot of 5 μL of tertiary PCR products were visualized on 1% TBE agarose gels to check amplification.

Tertiary PCR products were then pooled, run on a 1% agarose, Tris-acetate EDTA + cytidine gel, and 500 to 1000 bp molecules were gel-extracted. The gel extraction was performed with a Zymo gel extraction kit (Zymo Research), and the final libraries were eluted in 10 μL of 10 mM Tris-HCl. Libraries were quantified with the qPCR KAPA library quantification kit (KAPA Biosystems) according to manufacturer instructions, and reactions were run on a LightCycler480II (Roche). Libraries were also run on a bioanalyzer with a high-sensitivity chip (Agilent Technologies) to ensure the correct size of library fragments. Libraries were then prepared and run on a MiSeq (Illumina Inc.) according to manufacturer instructions. For transgene analysis, a paired-end, 500-cycle sequencing run was used. For transposon analysis, single-end, 250-cycle sequencing runs were used. Raw reads were de-multiplexed with the 6-nt indexes using the MiSeq Reporter software, version 2.3.32 (Illumina Inc.).

Sequence Analysis and Flanking Sequence Confirmation

Transgene Insertion Mapping

All sequences were analyzed using the commercial software Geneious 7 (Biomatters Ltd., 2013), a user-friendly, graphical-user-interface-based bioinformatics software package. Fastq files were imported into Geneious where they were trimmed for quality (error probability limit set to 5%). Then the primer and adaptor sequences were removed (IllR_pGEM from Read1 and IllF from Read2). Read1 reads were then sorted by amplicon, using the GSP2 primers as barcodes in the separate-reads-by-barcode function under the sequence menu, with one mismatch allowed. The reads were paired and de novo assembled into contigs with the following settings specified: do not merge contigs when there is a variant with coverage over approximately 6, merge homopolymer variants, do not allow gaps, minimum overlap 25, no minimum overlap identity, word length 24, index word length 14, ignore words repeated more than 100 times, reanalyze threshold 16, maximum mismatches per read 5%, and maximum ambiguity 4.

The de novo assemblies were first mapped to the linear, 5081 bp, transgene (Supplemental Fig. S1) using the default settings under the medium-sensitivity mapping option in Geneious. Of the contigs that mapped to the transgene, only those with at least 100 raw reads were considered for additional analysis. These assemblies were examined by eye in the assembly viewer, where contigs with large stretches of mismatches to the reference (≥20 bp) could quickly be identified. These mismatched portions were usually found on the ends of the contigs and were BLASTed to the NCBI nr database (www.ncbi.nlm.nih.gov/) and specifically to the soybean genome (*Glycine max* V1.1) in Phytozome V9.1 (www.phytozome.net/). Contigs at least 50 bases in length that matched the soybean genome with 90% identity were considered putative flanking sequences.

To verify the identified insertion loci, reverse primers were designed to the putative flanking sequences and were used in PCR with the respective GSP1 forward primers. Four templates were used for each primer: T0 DNA, nontransformed DNA, sequencing library, and no-template control. The sequencing library was used as a positive control. An example can be seen in Supplemental Fig. S3. The PCR reaction was as follows: 5 μL 2xApex Master Mix (Genesee Scientific), 0.3 μL 10 μM primers, 1 μL template, 3.4 μL water, using the following conditions: 95°C 3 min; 32 cycles (95°C 15 s, 60°C 15 s, 72°C 30 s); 72°C 5 min; hold at 12°C. The PCR products were visualized on a 1% agarose, TBE gel. Only primer sets with amplification in the T0 and sequencing library were considered positive (Supplemental Fig. S3). The PCR products were purified and Sanger sequenced to ensure PCR products were from the expected DNA sequence.

Transposon Insertion Mapping

Individuals from four additional transgenic lines that contain the *mPing* transposon were sequenced twice: once as individuals and again in pools of four to six siblings so that more samples could be processed at once. All sequences were processed using Geneious 7. The 3′ Illumina adaptor was removed from all sequences using the trim-primer function. Then, cleaned reads were examined for the presence of *mPing*. To separate *mPing*-containing sequences from background, reads containing the last 55 bp from the 3′ or 5′ end of *mPing* were filtered using the separate-reads-by-barcode function, with one mismatch allowed. This function both separates out the *mPing*-containing reads and removed this portion of *mPing* from the genomic portion of the read. These *mPing*-containing reads were mapped back to the reference soybean genome with the following custom sensitivity settings: no fine tuning, maximum gap size of 50 with no more than 15% of the read having gaps, a word length of 20 nt and index word length of 12 nt, a maximum of 30% mismatches per read, maximum ambiguity of 4, and allowing read mapping to repeat regions. More stringent mapping parameters were also used, and the results were the same. The sequences were mapped to a concatenation of the 20 soybean chromosomes. The reads were also mapped to each chromosome individually so that the specific location of each *mPing* insertion could be determined (Table 1). Insertions were manually BLASTed to the soybean genome (V1.2 and V2.0) to verify mapping results. General stats for the number of reads generated and mapped can be found in Supplemental Table S2.

To validate the *mPing* insertions that mapped in Geneious 7, primer pairs were designed 100 to 300 bp upstream and downstream of each putative *mPing* insertion site using Primer3 Plus (Rozen and Skaletsky 2000). Primers were tested first with the DNA from the individuals in which the insertion was sequenced. If these results were positive, the primers were used in a broader range of DNA samples from siblings of the sequenced lines. The PCR reaction was as follows: 5 μL 2xApex Master Mix (Genesee Scientific), 0.2 μL 10 μM primers, 1 μL template, 3.6 μL water, using the following conditions: 95°C 3 min; 34 cycles (95°C 30 s, 55–60°C 20 s, 72°C 1 minute); 72°C 5 min; hold at 12°C. The PCR products were visualized on a 1% agarose, TBE gel. Primer sequences for validated insertion sites are publicly available from SoyBase.org.

Results

Genome Fragmenting and Library Construction for Transgene Analysis

The 10 events analyzed here were generated by biolistic transformation of cultured soybean embryos with a linearized vector. The hygromycin-resistant events were PCR screened for presence of the gene of interest. Positive events were selected, and regenerated into whole plants. An initial attempt to identify genomic flanking

Table 1. Mapped *mPing* insertions in 14 individuals from four lineages of a single event.

PCR validated	Gmax v1.1 location	Gmax v2.0 location	Event 2-9 B2													
			3-47-2-3†				19-6-16-2†				16-9-9-5†			32-13-A-11†		
			4	5	6	7	13	14	15	18	4	5	6	8	10	11
not tested	01:1429049	01:1424207	X	X	X	X	X	X	X	X				X	X	X
yes††	01:2606076	01:2622091	X		X	X	X	X			X		X		X	X
not tested	01:4222382	01:4244346								X		X				
no	02:10936581	02:11029144													X	
no	03:39420017	03:37402214	X	X	X	X										
yes	4:5309363	04:5375178		X			X	X	X		X	X	X		X	X
yes	4:49160246	04:52305882		X	X	X	X				X	X	X	X	X	X
no	05:8103550	05:1020942													X	
no	05:4822311	05:6537162	X										X			
yes††	05:36268911	05:36551688			X	X	X	X	X	X	X	X		X		X
yes	05:38666410	05:41682411											X			
yes	06:1589315	06:1607130	X		X	X										
yes	08:5836631	08:5843843	X			X					X					X
yes	08:14637964	08:14568353	X		X	X	X						X			
yes††	08:44669303	08:45103227			X		X					X				
yes	10:6392637	10:6420224	X			X					X	X				
no	10:38503885	10:39052083												X	X	X
yes††	10:40875417	10:41422677	X	X	X	X	X			X	X		X	X		X
yes	11:1057718	11:1066579	X	X	X	X	X	X	X	X						
yes	11:36804300	11:32347400									X			X	X	X
yes††	12:35682591	12:35656855	X	X	X	X	X				X	X	X		X	X
yes	15:43604507	15:44363652										X	X			

† The 14 individuals are from one of these four lines derived from Event 2-9 B2.

†† Insertion that had been previously identified.

sequences with the HtStuf method used two amplicons from the ends of the linearized vector with the expectation that flanking sequences would be adjacent to the ends of the vector. However, only vector rearrangements were detected (data not shown), indicating that there was a complex integration of the linear fragments, which was likely caused by vector breakage and ligation of the vector fragments during the transformation process (Svitashev et al., 2002). Therefore, it was reasoned that primers designed to capture amplicons throughout the entire length of the transgene would be needed (Supplemental Fig. S4). DNA molecules greater than 1 kb in length do not efficiently bind to Illumina flow cells, therefore amplicon length cannot exceed 1 kb. To ensure adequate coverage, partially overlapping amplicons were designed 500 to 800 bp apart, resulting in 14 amplicons. The goal was to produce sequence reads with enough overlap with one another so that they could be efficiently assembled in silico. Primers were not designed to the GOI, as this sequence will be variable between different vectors, and the objective was to create a method that could be used across different experiments. With 14 amplicons per transgenic event, and 10 events to examine, the use a high-throughput sequencing method was essential to provide sequence data at a sufficient depth to identify genomic flanking sequences.

To generate sequence-specific PCR libraries compatible with Illumina sequencing, genomic DNA was fragmented to 200 bp to 2 kb, with a peak concentrated around 1 kb. A Covaris sonicator (Covaris Inc.) was used initially for mechanical sheering, but this method proved to be too costly and required too much input DNA (results not shown). Therefore, DNA Fragmentase (New England Biolabs), an enzyme mix that randomly cuts DNA, was used. A digestion time of 30 min at 37°C proved to be effective for concentrations ranging from 20 to 400 ng ul^{-1} (data not shown). There was no attempt to normalize or equalize the amount of DNA from each sample.

The pGEM-T Easy Vector System (Promega Corp.) was used to facilitate the generation of sequencing library templates. The use of a TA cloning system as opposed to linear adapters or blunt ligation, should limit the formation of adaptor concatemers and empty-vector ligation products (Orcheski and Davis, 2010). Since the entire genome was fragmented and ligated into the vector, a series of nested PCRs was used to specifically amplify the regions of interest, that is, transgenes or transposons and their flanking sequences. This was done using two sets of nested primer pairs in which one primer binds to the transgene and its pair binds to the vector backbone (Fig. 1).

The conditions used for primary and secondary amplification were a compromise between specificity and broad amplification. By using a nontouchdown protocol in the primary amplification then a more specific nested, touchdown protocol in the secondary amplification, the desired junction fragments were sufficiently amplified.

After the nontouchdown primary amplification, a large product at ~3 kb (presumably the cloning vector) and a smear of smaller products were observed in all DNA samples, indicating broad amplification of potential junction fragments (Supplemental Fig. S2). After the secondary amplification, samples with the amplicons-of-interest had the most PCR products—typically a smear with some banding— indicating true junction fragments were preferentially amplified (Supplemental Fig. S2). When a touchdown protocol was used for primary amplification, few samples produced products in the secondary amplification, and when a nontouchdown protocol was used for secondary amplification, nonspecific amplification was observed in the nontransformed control sample (data not shown). Both of these were undesirable for the production of junction fragment amplicons.

Since the 14 individual amplicons from the same transgenic event were produced with unique primers to different regions of the transgene (Supplemental Fig. S4), amplicons from the same transgenic event were pooled after the secondary PCR. The tertiary reaction is a low-cycle PCR to add on the necessary Illumina adapters and sequencing barcodes and resulted in a smear for all samples (Supplemental Fig. S1). The 14 sequences were later separated computationally into individual PCR amplicons using their unique primer sequences. The barcoded, tertiary PCRs from each of the 10 events were pooled and loaded into a single well for gel extraction. When many sequencing barcodes are used (>20), it could be useful to pool tertiary PCR products, column-purify to concentrate, and then load the concentrated samples into a single well. After gel purifying, samples were ready for sequencing. No attempt was made to normalize the amount of DNA sequenced from each amplicon.

Transgene Sequencing and Mapping

Illumina sequencing of the PCR libraries from the 10 unique transgenic events resulted in 2,364,016 paired-end reads (read1 + read2). There was an even distribution of the reads (8.4–10.6%) across the 10 events, even though individual input DNA was not normalized (Supplemental Table S3). The reads were also evenly distributed across amplicons within each event. Twelve amplicons produced an average of 6 to 9% each of the total reads (Supplemental Table S4), which was the expectation for evenly distributed reads. The amplicons from Stubi262F and Stubi389R made up only 2.4 and 5% of the sequenced reads, respectively. These two underrepresented amplicons likely did not amplify as efficiently as the other 12; however, this did not prevent the detection of flanking sequences (Supplemental Table S4). These results demonstrate that even without a DNA normalization step, this library preparation method results in reads that are evenly distributed between amplicons and individual events. Avoiding a DNA normalization step saves a considerable amount of time and resources. Fifty-six percent of the reads were assembled into contigs and considered usable (Supplemental Table S3). This result does suggest that the library preparation method could be improved to increase the number of usable reads; regardless, the reads used were sufficient for identifying flanking sequences.

From the 10 events, 90 putative flanking sequences were identified (4–16 flanking sequences per event) and 20 were confirmed (Supplemental Table S3). The false positives were likely due to artifacts generated during library preparation and were easily removed during PCR validation. The 20 confirmed flanking sequences were found in 10 of the 14 amplicons (Supplemental Table S4) and across the entire vector sequence (Supplemental Fig. S4). None of the flanking sequences were identified in more than one amplicon or event, indicating that all 14 amplicons were required to identify flanking sequences in these events and a subset would be insufficient. These results demonstrate that a flanking sequence can be found at any position along the vector DNA and validate the necessity of capturing amplicons along the entire length of the vector. At the junction sequences, microhomologies between the transgene, genomic flanking sequences and palindromic sequences were observed. Similar transgene integration patterns have been observed in biolistically-transformed oat (*Avena sativa* L.) (Svitashev et al., 2002). Such complex integration patterns would be difficult, if not impossible, to dissect with Southern blot technology, and would be expensive and time consuming to identify with Sanger sequencing. The ease of library construction and the depth of sequence coverage provide relatively inexpensive and straightforward analysis of transgene integration.

Three of the transgenic events had pairs of confirmed flanking sequences that mapped to the same chromosome but lay several kilobases to 1 megabase apart. For example, Event 8 (E8) has four flanking sequences: two are mapped to chromosome two, 8 kb apart; and two to chromosome 14, 406 kb apart (Supplemental Table S5). Similar results can be seen for Events 9 and 13. Interestingly, Event 31 has three confirmed flanking sequences that map to the chloroplast or chromosome 9 as well as two that map to chromosome 15. These data are reminiscent of the complex arrangements of transgenes interspersed with chloroplast (Aragao et al., 2013) or genomic (Jackson et al., 2001; Svitashev et al., 2002; Svitashev and Somers, 2001) sequences observed in other transgene mapping experiments.

Segregation and Zygosity Analysis of Transgenes in Plants

Segregation analysis was performed on T1 individuals derived from five events to determine the linkage of the transgenes with their confirmed flanking sequences. Event 8 has three segregating units and, surprisingly, the two flanking sequences that mapped to chromosome 14, only 406 kb apart (<1 cM), segregated independently (Supplemental Table S6). This unexpected result may be due to chromosome 14 sequences being incorporated into new, unidentified loci during the transgene integration process or due to improper reference genome assembly. The five

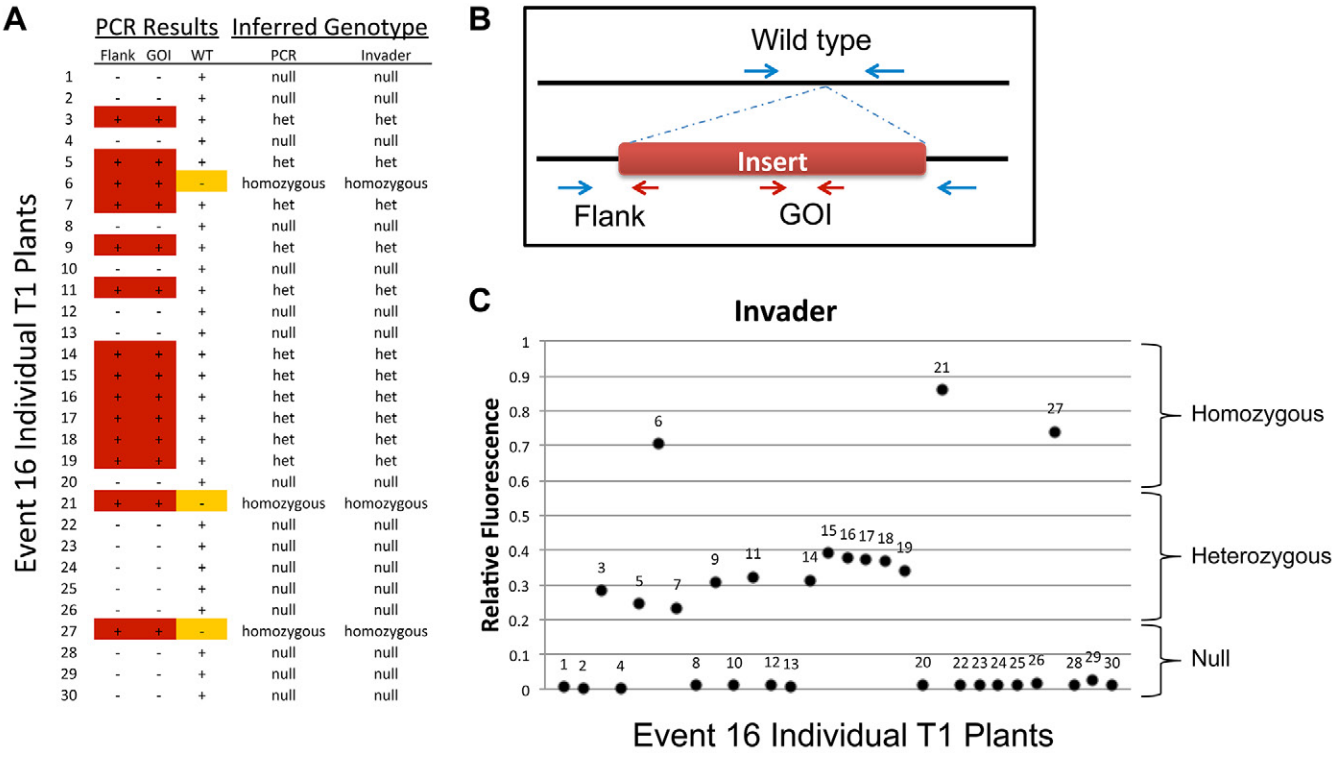

Figure 2. Segregation and zygosity check for Event ^#6 individuals. Polymerase chain reaction (PCR) (A) and Invader assays (C) were performed and genotype was inferred for 30 individuals from Event 16 (A). The primers used were designed within and to flank the transgene insert (B). Individual plants positive for the flank and gene-of-interest (GOI) amplicons are indicated with a plus sign (+) and colored red. Individual plants negative for the wild-type (WT) amplicon (homozygotes) are indicated with a minus sign (–) and colored gold. Invader assay results perfectly correlate with the PCR data (A, C).

confirmed flanking sequences in Event 31 (E31) all form one segregating unit. A second, un-identified insertion is present in E31 as four of the T1 plants were positive for the GOI and negative for the five flanking sequences (Supplemental Table S7). A single segregating unit was observed for three events (E13, E16, and E36). Together, these results indicate that while vector rearrangements frequently occur within an insertion site during biolistic transformation, the transgenic insertions still segregate as single units most of the time (three out of five events).

Determining the zygosity of segregating progeny, and identifying homozygous lines, is an important step in transgene analysis. To this end, the Invader assay is routinely used to identify homozygous plants. However, Invader only measures zygosity of certain selectable marker genes. While this is useful for events that segregate as a complete unit, it is not useful when different components segregate independently. Initially, it was reasoned that if the location of the transgene insertion was known, it would be possible to design primers to test for the presence or absence of the insertion. Such a PCR test was developed for Event 16, whereby two primers were designed to flank the insertion on chromosome 20 (Fig. 2B). Only wild-type sequences are successfully amplified with these primers, as the transgenic insertion is too large or complex for the PCR conditions. In this case, a negative result would indicate a homozygous line. Three T1 plants (6, 21, and 27) tested negative with the wild-type primer set (Fig. 2A). These three plants were positive for the flanking sequence and GOI, and the Invader results also identified the same individuals as homozygous (Fig. 2C). The two methods had perfect correlation in identifying heterozygotes and null segregants, illustrating that knowledge of the genomic insertion site (as identified using the HtStuf method) can be used to determine the zygosity of transgenic plants.

mPing Sequencing and Mapping

The DNA transposon *mPing* was previously transformed into soybean with the goal of generating mutations in soybean genes (Hancock et al., 2011). The HtStuf method was tested on four T6 lines derived from a single event as a way to rapidly map *mPing* insertions in a large population. Libraries from *mPing*-containing lines were generated as for the transgene mapping but with two major differences. First, just one set of primers was used (as opposed to 14) at each of the nested PCRs (a pGEM primer and a *mPing* primer). Second, since the same set of primers was used to amplify putative *mPing* insertion sites in all samples, each sample was kept separate until after the Illumina sequencing barcodes were added in the tertiary amplification step (secondary PCR products were not pooled).

Sequencing 15 individuals from the four lines produced over 600,000 reads. These reads were processed using Geneious 7, and approximately 220,000 *mPing*-containing reads were identified (~37% of total reads).

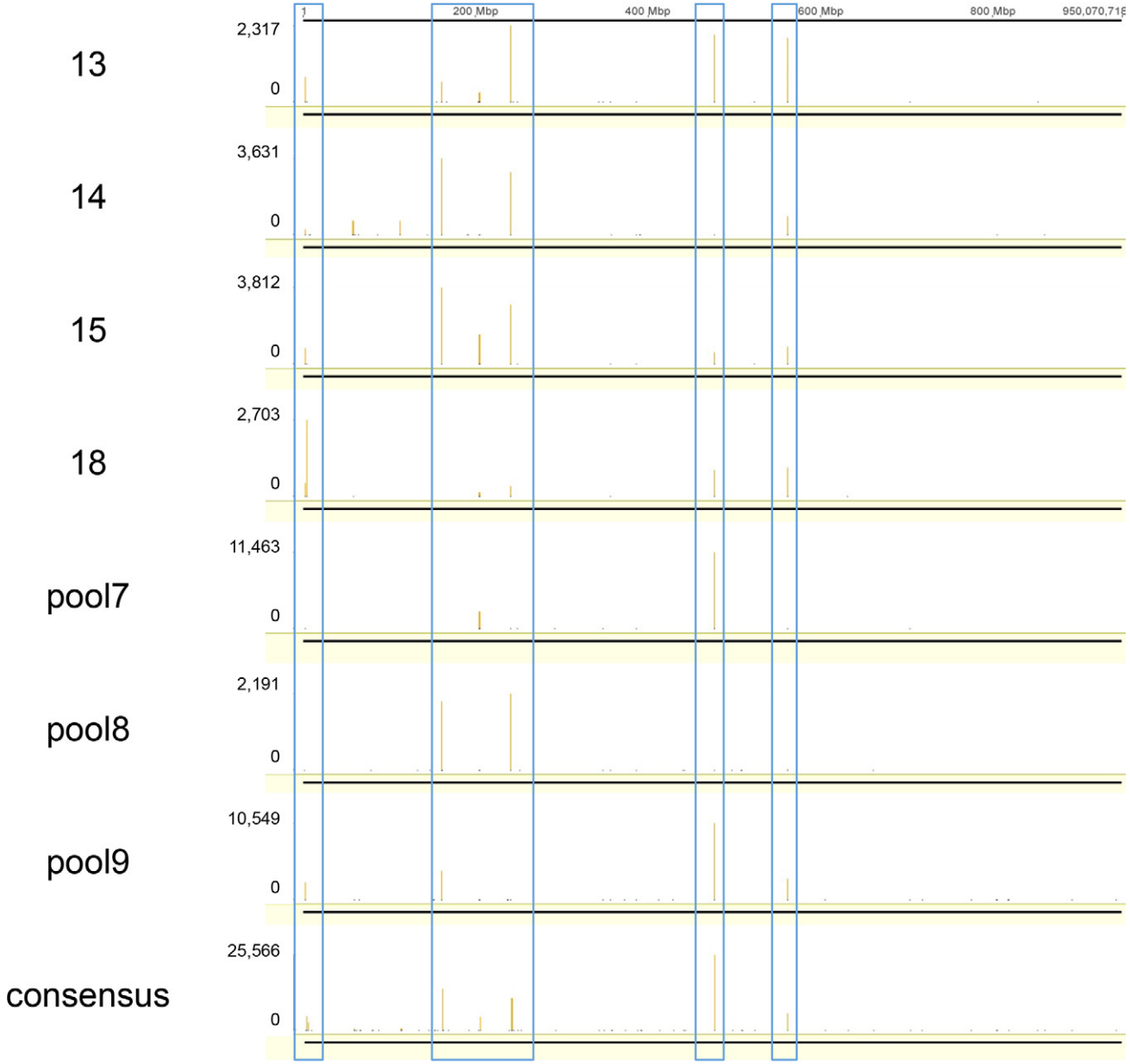

Figure 3. Coverage graphs of *mPing*-containing reads mapped to the soybean genome. Individuals 13, 14, 15, and 18 are siblings from the same line. Pools 7, 8, and 9 are also from this line and contain five to six individuals. Insertions that are germinal (shared between individuals) clearly show up in the consensus graph (blue rectangles). Somatic insertions show up in individual graphs (e.g., shorter orange bars in 14).

Of these, 170,030 (that contained the flanking sequence information) were mapped to the soybean reference genome. Based on shared insertions and coverage of mapped loci, seven loci stood out above background levels (Fig. 3). To expand the analysis, 24 pools containing four to six DNA samples were processed into libraries and sequenced. These sequences contained the same 15 initially sequenced individuals as well as 84 additional individuals from the same four lines. More than 620,000 *mPing*-containing reads were mapped to 22 unique locations in the genome, and 15 were PCR-validated (Table 1).

The PCR-validated insertions were shared between different individuals, indicating that they are germinal insertions, that is, occurring in the previous generation's germ line (Table 1), rather than in somatic tissues that does not contribute to gamete formation. In fact, many of these insertions were shared between individuals from different lines, indicating they occurred at least two generations prior. Five insertions that were previously validated by cloning of TD products (Hancock et al., 2011; Hancock, personal communication, 2013) were identified here as well, providing further support of the validity and robustness of this method.

Transposon display is a modified amplified fragment length polymorphism technique (Van den Broeck et al., 1998) that requires running a polyacrylamide gel, extracting bands or cloning, and then sequencing the bands or clones (Supplemental Fig. S5) (Hancock et al., 2011). To

initially identify the five insertions validated here, a TD reaction was cloned and 96 colonies were Sanger sequenced (Hancock, personal communication, 2013). The use of TD to identify and validate all 15 *mPing* insertions identified here would be costly and time prohibitive (for a cost comparison, see Supplemental Fig. S5). Moreover, the use of restriction enzymes for the initial production of genomic DNA fragments in TD can result in sampling biases.

Segregation Analysis of *mPing* Insertions

The individuals analyzed here were T6 lines, and 15 germinal insertions were validated with PCR, indicating a germinal transposition rate of slightly more than two germinal *mPing* insertions per generation, which is in agreement with previously reported events (Hancock et al., 2011). Five of the 15 PCR validated insertions were identified as being shared in all four lines and seven of them were shared between two and three lines (Table 1). There were three insertions found in only one line, that is, unique to a line. It is also clear from examining the results that most of the insertions are not homozygous, as they are segregating within each line (Table 1).

Discussion

The bioinformatics analysis of the data produced using HtStuf is straightforward and can be done in user-friendly software such as Geneious, as was done here. Unlike the work with mapping *Mu* transposon insertions in maize, the samples here were not tagged with multiplexing identifications. This largely simplified the separation of individually barcoded samples in our analysis, which can be done in MiSeq Reporter (Illumina, Inc.). Constructing libraries with known priming sites allowed for the trimming and isolation of transgene- or *mPing*-containing sequences with the tools available on Geneious. If multiplexing were desired, primers could be modified to incorporate additional levels of barcoding and tools could be developed to process the data; however, we were able to process and analyze 14 amplicons from each of 10 transgenic events (140 amplicons) plus 99 *mPing* individuals using the methods described here.

This sequencing method is robust and accurate; flanking sequences were confirmed in nine out of 10 transgenic events and 15 individuals from four *mPing* lines. The power of this approach comes from the combination of random fragmentation, the ease of TA cloning, and the use of next-generation sequencing technology, which allows pooling many amplicons and samples for simultaneous sequencing. Typical gene-walking experiments clone and then Sanger-sequence PCR products to identify flanking sequences (Leoni et al., 2011). With the complex integration pattern observed with the transgenic events, or with the large numbers of individuals in a transposon mutagenesis screen, such an approach would be tedious, time-consuming, and expensive. Instead, HtStuf allows the processing and analysis of genomic DNA to yield validated insertion sites in 1 wk. Furthermore, the molecular and informatics techniques used here are straight forward enough that student workers have been trained to independently generate sequencing libraries and analyze the results.

An added benefit identified during our analysis is the ability to bypass typical DNA-normalization steps. Such procedures are tedious and time-consuming. Given the emphasis on DNA normalization in most library preparation protocols, the even coverage observed here is fortuitous. This would suggest that at least with the type of amplicon sequencing employed here, added DNA normalization schemes are simply not necessary. Additional amplicon-sequencing experiments not reported here have shown that a relatively even read coverage can be obtained from a range of starting genomic DNA concentrations. We speculate that this may be due to amplifying our first two sets of PCRs past the linear growth phase.

One limitation with this method is the amount of background PCR and ligation artifacts that were sequenced. Only 22% (20 out of 90) of the putative transgene flanking sequences were shown to be real insertions. While this may seem like a low frequency, the PCR screening process to verify putative flanking sequences can be accomplished in relatively little time. The use of Illumina sequencing generated more than enough useful sequence data. One possible way to improve the overall specificity of transgene mapping is to sequence multiple individuals from the same event, as was shown with the *mPing* mapping. All PCR-confirmed *mPing* insertions were mapped in multiple individuals. Insertions in single individuals are either artifacts or, more likely, new, somatic insertions. Since Illumina sequencing yields a large number of reads, and the sequencing libraries are relatively simple to prepare, making two or more libraries per transgenic event to improve specificity would not be a burden.

For transgenic mapping, one advantage of this method over previously published studies that use WGS (Kovalic et al., 2012; Ming et al., 2008) is the enrichment of the transgenic sequences. Here, 56% of the sequenced reads could be assembled into contigs and used for mapping to the transgene. If WGS were used, assuming that most transgenes are 10 kb in size, most crop genomes are at least 1 Gb, and the sequencing reads are paired-end and 100 to 300 bp in length, then only 0.2 to 0.6% of the sequencing reads will contain transgenic bases. Using WGS for the identification of transgenic insertions may be better suited to evaluating high-value transgenic lines to ensure other DNA modifications or insertions were not made during the transformation process.

While we used this method for the identification of transgene and transposon flanking sequences, there are additional applications for this technology. In nonsequenced or poorly sequenced genomes, this genome walking method can be used to close gaps in contigs for genome assembly. Degenerate primers could also be designed to amplify sequences in related organisms to capture sequences from large gene families. In cases of gene amplification, as seen with the

5-enolpyruvylshikimate-3-phosphate synthase gene in glyphosate resistant weeds (Gaines et al., 2010), this technique could be used to identify unique insertion sites. It should be possible to modify the primer sequences and digestion time to work with any system, including long-read sequencing technology offered by Pacific Biosciences.

The ability to sequence specific, unmapped DNA loci with a high-throughput technology is useful for the characterization of transgenic plants. The data presented demonstrates that HtStuf is a quick and reliable method for determining the flanking sequences of transgenes and transposons in the soybean genome. The effectiveness of the technology is based on the combined use of a modified TOPO vector-ligation PCR method with the power of Illumina sequencing. Sequence data can be generated and analyzed quickly, and flanking sequences are identified in nearly all sequenced individuals. This sequencing technology is not limited to soybean and should be applicable in any other species where flanking sequences need to be known.

Acknowledgments

This work was supported by the United Soybean Board, State and Federal monies allocated to the Georgia Agricultural Experiment Stations, and the National Science Foundation (Award ID, 1127083). A special thanks to Dr. R. Kelly Dawe (University of Georgia) and Dr. C. Nathan Hancock (University of South Carolina–Aiken) for critical reading and editing of the manuscript.

References

An, G., B.D. Watson, S. Stachel, M.P. Gordon, and E.W. Nester. 1985. New cloning vehicles for transformation of higher plants. EMBO J. 4:277–284.

Aragao, F.J.L., E. Nogueira, M.L.P. Tinoco, and J.C. Faria. 2013. Molecular characterization of the first commercial transgenic common bean immune to the *Bean golden mosaic virus*. J. Biotechnol. 166:42–50. doi:10.1016/j.jbiotec.2013.04.009

Biomatters, Ltd. 2013. Geneious R7. Biomatters, Ltd., Auckland, NZ.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:10. doi:10.1371/journal.pone.0019379

Faircloth, B.C., and T.C. Glenn. 2012. Not all sequence tags are created equal: Designing and validating sequence identification tags robust to indels. PLoS ONE 7:11. doi:10.1371/journal.pone.0042543

Gaines, T.A., W.L. Zhang, D.F. Wang, B. Bukun, S.T. Chisholm, D.L. Shaner, S.J. Nissen, W.L. Patzoldt, P.J. Tranel, A.S. Culpepper, T.L. Grey, T.M. Webster, W.K. Vencill, R.D. Sammons, J.M. Jiang, C. Preston, J.E. Leach, and P. Westra. 2010. Gene amplification confers glyphosate resistance in *Amaranthus palmeri*. Proc. Natl. Acad. Sci. USA 107:1029–1034. doi:10.1073/pnas.0906649107

Garbarino, J.E., and W.R. Belknap. 1994. Isolation of a ubiquitin-ribosomal protein gene (*ubi3*) from potato and expression of its promoter in transgenic plants. Plant Mol. Biol. 24:119–127. doi:10.1007/BF00040579

Hancock, C.N., F. Zhang, K. Floyd, A.O. Richardson, P. LaFayette, D. Tucker, S.R. Wessler, and W.A. Parrott. 2011. The rice miniature inverted repeat transposable element *mPing* is an effective insertional mutagen in soybean. Plant Physiol. 157:552–562. doi:10.1104/pp.111.181206

Hernandez-Garcia, C.M., A.P. Martinelli, R.A. Bouchard, and J.J. Finer. 2009. A soybean (*Glycine max*) polyubiquitin promoter gives strong constitutive expression in transgenic soybean. Plant Cell Rep. 28:837–849. doi:10.1007/s00299-009-0681-7

Huang, X.H., Q. Feng, Q. Qian, Q. Zhao, L. Wang, A.H. Wang, J.P. Guan, D.L. Fan, Q.J. Weng, T. Huang, G.J. Dong, T. Sang, and B. Han. 2009. High-throughput genotyping by whole-genome resequencing. Genome Res. 19:1068–1076. doi:10.1101/gr.089516.108

Hufford, M.B., X. Xu, J. van Heerwaarden, T. Pyhajarvi, J.M. Chia, R.A. Cartwright, R.J. Elshire, J.C. Glaubitz, K.E. Guill, S.M. Kaeppler, J.S. Lai, P.L. Morrell, L.M. Shannon, C. Song, N.M. Springer, R.A. Swanson-Wagner, P. Tiffin, J. Wang, G.Y. Zhang, J. Doebley, M.D. McMullen, D. Ware, E.S. Buckler, S. Yang, and J. Ross-Ibarra. 2012. Comparative population genomics of maize domestication and improvement. Nat. Genet. 44:808–811. doi:10.1038/ng.2309

Jackson, S.A., P. Zhang, W.P. Chen, R.L. Phillips, B. Friebe, S. Muthukrishnan, and B.S. Gill. 2001. High-resolution structural analysis of biolistic transgene integration into the genome of wheat. Theor. Appl. Genet. 103:56–62. doi:10.1007/s001220100608

Kovalic, D., C. Garnaat, L. Guo, Y.P. Yan, J. Groat, A. Silvanovich, L. Ralston, M.Y. Huang, Q. Tian, A. Christian, N. Cheikh, J. Hjelle, S. Padgette, and G. Bannon. 2012. The use of next generation sequencing and junction sequence analysis bioinformatics to achieve molecular characterization of crops improved through modern biotechnology. Plant Gen. 5:149–163. doi:10.3835/plantgenome2012.10.0026

Leoni, C., M. Volpicella, F. De Leo, R. Gallerani, and L.R. Ceci. 2011. Genome walking in eukaryotes. FEBS J. 278:3953–3977. doi:10.1111/j.1742-4658.2011.08307.x

Liu, S.Z., A.P. Hsia, and P.S. Schnable. 2013. Digestion-ligation-amplification (DLA): A simple genome walking method to amplify unknown sequences flanking *mutator* (*mu*) transposons and thereby facilitate gene cloning. In: T. Peterson, editor, Plant transposable elements: Methods and protocols (Methods in molecular biology). Humana Press, New York. p. 167–176.

Liu, Y.G., and R.F. Whittier. 1995. Thermal asymmetric interlaced PCR: Automatable amplification and sequence of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25:674–681. doi:10.1016/0888-7543(95)80010-J

McCarty, D.R., A.M. Settles, M. Suzuki, B.C. Tan, S. Latshaw, T. Porch, K. Robin, J. Baier, W. Avigne, J.S. Lai, J. Messing, K.E. Koch, and L.C. Hannah. 2005. Steady-state transposon mutagenesis in inbred maize. Plant J. 44:52–61. doi:10.1111/j.1365-313X.2005.02509.x

Ming, R., S.B. Hou, Y. Feng, Q.Y. Yu, A. Dionne-Laporte, J.H. Saw, P. Senin, W. Wang, B.V. Ly, K.L.T. Lewis, S.L. Salzberg, L. Feng, M.R. Jones, R.L. Skelton, J.E. Murray, C.X. Chen, W.B. Qian, J.G. Shen, P. Du, M. Eustice, E. Tong, H.B. Tang, E. Lyons, R.E. Paull, T.P. Michael, K. Wall, D.W. Rice, H. Albert, M.L. Wang, Y.J. Zhu, M. Schatz, N. Nagarajan, R.A. Acob, P.Z. Guan, A. Blas, C.M. Wai, C.M. Ackerman, Y. Ren, C. Liu, J.M. Wang, J.P. Wang, J.K. Na, E.V. Shakirov, B. Haas, J. Thimmapuram, D. Nelson, X.Y. Wang, J.E. Bowers, A.R. Gschwend, A.L. Delcher, R. Singh, J.Y. Suzuki, S. Tripathi, K. Neupane, H.R. Wei, B. Irikura, M. Paidi, N. Jiang, W.L. Zhang, G. Presting, A. Windsor, R. Navajas-Perez, M.J. Torres, F.A. Feltus, B. Porter, Y.J. Li, A.M. Burroughs, M.C. Luo, L. Liu, D.A. Christopher, S.M. Mount, P.H. Moore, T. Sugimura, J.M. Jiang, M.A. Schuler, V. Friedman, T. Mitchell-Olds, D.E. Shippen, C.W. dePamphilis, J.D. Palmer, M. Freeling, A.H. Paterson, D. Gonsalves, L. Wang, and M. Alam. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991–996. doi:10.1038/nature06856

Murray, M.G., and W.F. Thompson. 1980. Rapid isolation of high molecular-weight plant DNA. Nucleic Acids Res. 8:4321–4325. doi:10.1093/nar/8.19.4321

Naito, K., E. Cho, G.J. Yang, M.A. Campbell, K. Yano, Y. Okumoto, T. Tanisaka, and S.R. Wessler. 2006. Dramatic amplification of a rice transposable element during recent domestication. Proc. Natl. Acad. Sci. USA 103:17620–17625. doi:10.1073/pnas.0605421103

Ochman, H., A.S. Gerber, and D.L. Hartl. 1988. Genetic applications of an inverse polymerase chain-reaction. Genetics 120:621–623.

Orcheski, B.B., and T.M. Davis. 2010. An enhanced method for sequence walking and paralog mining: TOPO® Vector-Ligation PCR. BMC Res. Notes 3:61. doi:10.1186/1756-0500-3-61

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132:365–386. doi:10.1385/1-59259-192-2:365

Svitashev, S.K., W.P. Pawlowski, I. Makarevitch, D.W. Plank, and D.A. Somers. 2002. Complex transgene locus structures implicate multiple mechanisms for plant transgene rearrangement. Plant J. 32:433–445. doi:10.1046/j.1365-313X.2002.01433.x

Svitashev, S.K., and D.A. Somers. 2001. Genomic interspersions determine the size and complexity of transgene loci in transgenic plants produced by microprojectile bombardment. Genome 44:691–697. doi:10.1139/g01-040

Van den Broeck, D., T. Maes, M. Sauer, J. Zethof, P. De Keukeleire, M. D'Hauw, M. Van Montagu, and T. Gerats. 1998. Transposon display identifies individual transposable elements in high copy number lines. Plant J. 13:121–129. doi:10.1046/j.1365-313X.1998.00004.x