Genomics Data 6 (2015) 175–181

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

CONTRAILS: A tool for rapid identification of transgene integration sites in complex, repetitive genomes using low-coverage paired-end sequencing

Kevin C. Lambirth [a], Adam M. Whaley [b], Jessica A. Schlueter [b], Kenneth L. Bost [a], Kenneth J. Piller [a,*]

[a] Department of Biological Sciences, University of North Carolina at Charlotte, Charlotte, NC 28223, United States
[b] Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, United States

ARTICLE INFO

Article history:
Received 27 August 2015
Accepted 2 September 2015
Available online 8 September 2015

Keywords:
Transfer DNA
Insertion
Transformation
Junction sequences
Next generation sequencing
Agrobacterium

ABSTRACT

Transgenic crops have become a staple in modern agriculture, and are typically characterized using a variety of molecular techniques involving proteomics and metabolomics. Characterization of the transgene insertion site is of great interest, as disruptions, deletions, and genomic location can affect product selection and fitness, and identification of these regions and their integrity is required for regulatory agencies. Here, we present CONTRAILS (Characterization of Transgene Insertion Locations with Sequencing), a straightforward, rapid and reproducible method for the identification of transgene insertion sites in highly complex and repetitive genomes using low coverage paired-end Illumina sequencing and traditional PCR. This pipeline requires little to no troubleshooting and is not restricted to any genome type, allowing use for many molecular applications. Using whole genome sequencing of in-house transgenic *Glycine max*, a legume with a highly repetitive and complex genome, we used CONTRAILS to successfully identify the location of a single T-DNA insertion to single base resolution.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Over the past two decades transgenic crops and foods have become integrated into worldwide agriculture, greatly increasing yields and easing cultivation labors through value added traits. *Agrobacterium*-mediated transformation and particle bombardment are common methods for creating crops to achieve this advancement. Current understanding indicates that transfer DNA (T-DNA) integration into the host's genome is a random process that has been reviewed extensively [1–4]. Characterization of integration sites is of great interest, particularly if the host is to be deregulated for human consumption or for commercial applications to assess potential pleiotropic effects resulting from transformation and evaluate the potential for inadvertent mutagenesis [5–7].

T-DNA inserts have been reported in both transcriptionally active and repressed regions of chromatin [1,2,4,8,9]. Additionally, in some instances T-DNA sequences have been detected within host endogenous genes, including promoter and regulatory regions [4,10]. Transgenic plants containing multiple copies of T-DNA sequences have also been reported, and these complex events can lead to silencing of the gene of interest [11] emphasizing the favorable selection of simple, single T-DNA insertion events. Single insertion events in transgenic plants can be generated through multi-generation propagation and are traditionally screened for complexity using Southern blots. While Southern blotting has been proven to be a reliable method for identifying copy numbers, no information regarding T-DNA insertion orientation, random DNA insertions or deletions at the insertion site, or the genomic location of the insert is revealed using this method. Furthermore, Southern blots can require extensive troubleshooting, may require radioactive materials, and can produce ambiguous results if the restriction enzymes exhibit star activity or digested genomic DNA products containing the transgene are similar in size. Thus, many alternative methods to estimate T-DNA copy number have been utilized but aren't without certain shortcomings.

Quantitative PCR analyses of transgene expression levels can be correlated with transgene copy numbers [12–14], although results from these methods are not always reliable due to other factors that could alter transgene expression independent of zygosity, such as gene silencing and truncation. Visualization methods such as Fluorescent In-Situ Hybridization (FISH) have been implemented for years to identify insertion regions on specific chromosomes [15–18], however this is a relatively expensive visual technique and confers no information about the

Abbreviations: FISH, Fluorescent In-situ Hybridization; hTG, human thyroglobulin; IGB, Integrated Genome Browser; NGS, Next-Generation Sequencing; T-DNA, Transfer DNA.

* Corresponding author at: 9201 University City Blvd, Woodward Hall Room 377, Charlotte, NC 28223, United States.
E-mail addresses: kclambirth@uncc.edu (K.C. Lambirth), awhaley9@uncc.edu (A.M. Whaley), jschluet@uncc.edu (J.A. Schlueter), klbost@uncc.edu (K.L. Bost), kjpiller@uncc.edu (K.J. Piller).

http://dx.doi.org/10.1016/j.gdata.2015.09.001
2213-5960/© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

surrounding sequence of the insertion region, or if tandem insertions have occurred. FISH must be coupled with targeted PCR amplification of sequences spanning the observed integration region, followed by sequencing to identify more precise integration points.

PCR techniques designed for transposon characterization, such as splinkerette PCR and inverse PCR [19–21], can reveal detailed integration information and have proven accurate for transgene insertion characterization due to reliance on sequence specific initiation. Consequently, the presence of multiple or complex insertions, truncated transgene sequences, and highly repetitive genomes of host organisms can: a) prevent adequate detection, b) generate non-specific products, or c) fail to amplify products if primer targets are missing. Specialty restriction enzymes may also be required depending on the T-DNA fragment sequence (e.g.: methylation sensitivity, star activity), and a larger amount of genomic DNA is needed in order to visually verify digestion and ligation at each step. Genome walking has been employed effectively with universal primers [22], however as with the other PCR-based techniques, highly complex insertion events and repetitive genomic regions can potentially confound the results. In addition, larger T-DNA insertion sequences (e.g.: >10 kb) are difficult to fully amplify in their entirety due to the limits of traditional polymerase activity; specialized polymerase varieties for longer amplification are available, but are more expensive than traditional polymerase, are subject to PCR-based assay complications, and can only extend amplification reliably to ~20–30 kb.

In order to address these limitations, many groups have utilized next-generation sequencing (NGS) to identify and validate transgene insertion events [23–27]. Within the past 10 years, sequencing costs have been significantly reduced, while throughput and efficiency have greatly increased. NGS has already proven to be a reliable and accurate method for rapid identification of transposon insertion locations [28]. In addition, further analyses may be conducted on the resulting stored datasets in future genomic studies, such as genome-wide single nucleotide polymorphism (SNP) profiling, updated gene models and fusions, and complete sequencing of the transgene fragment for verification of the insert's integrity. Recently, several reports have successfully used NGS to identify transgene insertion locations in various organisms [24,25,29], even at relatively low coverage (2–5×). The short turn-around time, coupled with the absence of a need for pre-experimental troubleshooting makes this a very attractive and cost-effective option for reliably identifying random transgene insertions. Furthermore, reference genomes for many species have been fully sequenced and are available for use, removing the need for complete genome de novo assembly of the resulting sequencing reads. This allows effective use of short read sequences in large and complex genomes, as efficient and accurate algorithms for such large de novo assemblies do not currently exist.

Here, we present and demonstrate CONTRAILS (Characterization of Transgene Insertion Locations with Sequencing): a pipeline using existing bioinformatics tools and paired-end Illumina next-generation genomic sequencing to identify and characterize transgene insertion locations in the highly complex and repetitive genome of the legume *Glycine max* (Fig. 1). Paired-end reads spanning the T-DNA insertion junction allow for one read to map to the reference genome, and the other to map to the transgene sequence. Using short insert (≤500 b.p.) paired-end reads allows the user to narrow the insertion site to a genomic region of 500 b.p. or less, provided assembly is assisted with an established reference genome. In some cases, it is possible for a single read to span both genomic and T-DNA sequences at the transgene insertion junction, giving immediate confirmation of insert location and neighboring sequences at single base resolution. However if this is not achieved, the matched paired-end reads will disclose the location well within conventional PCR amplification range for rapid characterization of the T-DNA junction sites. Using this technique, we have identified and characterized a single T-DNA insert site in a transgenic line expressing recombinant hTG protein [30] to single-base resolution. These results are consistent with previous Southern blot and western blot screens, confirming the findings of the NGS analysis. Using this pipeline in conjunction with event-specific PCR assays, we were able to fully characterize flanking genomic sequences surrounding the T-DNA location.

2. Methods

2.1. Genomic DNA extraction and preparation

Whole-seed genomic DNA was extracted from chips of cotyledon tissue using a Maxwell 16 instrument and DNA extraction kit (Promega, Madison WI). Extracts were cleaned by phenol–chloroform and precipitated with 100% ethanol. DNA concentrations and purity were assessed with a Nanodrop 2000 spectrophotometer (Thermo Scientific, Waltham MA) and agarose gels to ensure optimal quality and concentration (260/280 absorbance ratio 1.8–2.0, greater than 1 μg total DNA).

2.2. Illumina HiSeq 2000 library preparation, sequencing, and quality control

Library generation was conducted at the David H. Murdock Research Institute genomics department according to the Illumina (San Diego, CA) HiSeq protocol, generating reported insert sizes of 350 b.p. after quality control analysis. Paired-end sequencing was conducted on the Illumina HiSeq 2000 system. The soy sample ST77-KP2 characterized in this study was one of two pooled soy samples on a single lane sequenced to ~5× theoretical genome-wide coverage with 100 base-pair reads. Low-quality reads were filtered out using in-house Illumina software and validated with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc), showing the remaining read basecall quality scores all greater than 30.

2.3. Reference genome construction and read alignment

The soybean reference genome sequence version 2.75 was obtained from Phytozome [31] and amended with an extra chromosome scaffold containing the T-DNA sequence located between the left and right border repeat regions (Fig. 2) [30]. Paired sequence reads from the previously described seed genomic DNA sequencing were aligned to the constructed reference using Bowtie (ver. 2.2.1) [32] with parameters –*un-conc* to specify discordant read output. Default Bowtie search methods were used with zero allowed mismatches to limit ambiguous alignments due to the abundance of highly repetitive and homologous endogenous sequences, and in global mode to not trim read ends to enhance alignment scores.

2.4. Identification of the transgene insertion site

Fragments in which one read aligned to known genomic reference sequence and the other read aligned to T-DNA sequence were flagged and separated from reads that aligned strictly to the known soybean reference sequence. Each enriched discordant read sequence was aligned against both the *G. max* reference genome using the "refseq_genomic" function in BLAST [33] and the T-DNA sequence, and matching mates were selected for further characterization. Reads matching the endogenous 7S glycinin promoter were detected in the filtered output and were excluded as illegitimate insertion sites. The genomic read furthest upstream and downstream from the T-DNA read pairs was selected for PCR amplification of the insert junctions to ensure that the anticipated fragment was included within the selected genomic region.

2.5. Validation of T-DNA insertion via PCR

Primers were designed to generate an amplicon that spans the genomic region and into both the right and left border sequences: genomic right border forward (5′-AGGATGACCCGACATGTCTCTTAG-3′), T-DNA right border reverse (5′-CAAATGAAGGGCATGGATCCTGC-3′), T-DNA left border forward (5′-CGGTTTGCGTATTGGCTAGAGC-3′), and genomic

Download English Version:

https://daneshyari.com/en/article/2821989

Download Persian Version:

https://daneshyari.com/article/2821989

Daneshyari.com