TECHNICAL ADVANCE

Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize

Rosalind Williams-Carrier1, Nicholas Stiffler1, Susan Belcher1, Tiffany Kroeger1, David B. Stern2, Rita-Ann Monde1,1, Robert Coalter1,1 and Alice Barkan1,*

1Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA, and
2Boyce Thompson Institute for Plant Research, Ithaca, NY 14853, USA

Received 8 February 2010; revised 2 April 2010; accepted 9 April 2010; published online 13 May 2010.
*For correspondence (fax +1 541 346 5891; e-mail abarkan@uoregon.edu).
†Present address: Department of Radiation Oncology, University of Miami, Miller School of Medicine, Miami, FL 33136, USA.
‡Present address: Department of Wildlife and Fisheries Biology, University of California, Davis, CA 95616, USA.

SUMMARY

High-copy transposons have been effectively exploited as mutagens in a variety of organisms. However, their utility for phenotype-driven forward genetics has been hampered by the difficulty of identifying the specific insertions responsible for phenotypes of interest. We describe a new method that can substantially increase the throughput of linking a disrupted gene to a known phenotype in high-copy Mutator (Mu) transposon lines in maize. The approach uses the Illumina platform to obtain sequences flanking Mu elements in pooled, bar-coded DNA samples. Insertion sites are compared among individuals of suitable genotype to identify those that are linked to the mutation of interest. DNA is prepared for sequencing by mechanical shearing, adapter ligation, and selection of DNA fragments harboring Mu flanking sequences by hybridization to a biotinylated oligonucleotide corresponding to the Mu terminal inverted repeat. This method yields dense clusters of sequence reads that tile approximately 400 bp flanking each side of each heritable insertion. The utility of the approach is demonstrated by identifying the causal insertions in four genes whose disruption blocks chloroplast biogenesis at various steps: thylakoid protein targeting (cpSecE), chloroplast gene expression (polynucleotide phosphorylase and PTAC12), and prosthetic group attachment (HCF208/CCB2). This method adds to the tools available for phenotype-driven Mu tagging in maize, and could be adapted for use with other high-copy transposons. A by-product of the approach is the identification of numerous heritable insertions that are unrelated to the targeted phenotype, which can contribute to community insertion resources.

Keywords: transposon tagging, maize, chloroplast, Mutator, Illumina, transposon display.

INTRODUCTION

With the availability of sequence-indexed insertion collections and efficient methods for directed gene silencing, reverse genetic approaches have become the primary means for assigning gene function in model organisms. Nonetheless, phenotype-driven forward genetic screens continue to provide groundbreaking insights into complex biological processes. Chemical, radiation and insertional mutagenesis each offer advantages, but the use of endogenous transposable elements has been particularly important in organisms such as maize, for which map-based cloning and transformation are relatively difficult. The majority of forward-genetic transposon tagging experiments in maize have used the Mutator (Mu) transposon system, whose primary advantage is its high forward mutation rate (Lisch and Jiang, 2009; McCarty and Meeley, 2009; Settles, 2009). However, the large number of Mu transposons in most Mu-active lines detracts from the utility of this system by complicating identification of the insertion underlying the phenotype of interest. We describe a new method that can substantially increase the throughput of
this final step in a Mu tagging experiment, and that is particularly useful for large-scale efforts to link Mu insertions to phenotypes.

We developed this method as a means to more fully exploit a large collection of photosynthesis mutants in maize, the Photosynthesis Mutant Library (PML) (Stern et al., 2004). Mutants in the PML collection were assembled by screening Mu-active maize lines for chlorophyll-deficient mutants. The phenotypes range from albino to slightly pale green, and include virescent, albescent and variegated mutants (Figure S1). These phenotypes provide an easily scored read-out of disrupted chloroplast biogenesis or homeostasis, and reflect primary defects in the synthesis or assembly of subunits of the photosynthetic apparatus, pigment or prosthetic group metabolism, the targeting of proteins to and within the chloroplast, chloroplast gene expression and chloroplast protein turnover. The collection consists of approximately 2000 independently arising mutants culled from approximately 28 000 F1 individuals, and accompanying chloroplast protein and RNA data that elucidate the function of the disrupted gene (Stern et al., 2004). Based on the distribution of alleles recovered to date, the collection is near saturation for genes whose disruption results in one of the phenotypes used to assemble the collection. The PML collection includes many albino mutants with severe plastid ribosome deficiencies, a condition that results in embryo lethality in Arabidopsis (reviewed by Williams and Barkan, 2003; Stern et al., 2004; Asakura and Barkan, 2006; Schmitz-Linneweber et al., 2006; Beick et al., 2008). For this and other reasons, the PML collection is a valuable complement to related mutant collections being developed in Arabidopsis (Myouga et al., 2009; Ajjawi et al., 2010). The PML collection can be used for reverse genetics (e.g. Ostheimer et al., 2003; Schmitz-Linneweber et al., 2006; Watkins et al., 2007; Kroeger et al., 2009; Pfalz et al., 2009) and for phenotype-driven gene discovery (e.g. Voelker and Barkan, 1995b; Voelker et al., 1997; Fisk et al., 1999; Walker et al., 1999; Jenkins and Barkan, 2001; Till et al., 2001). However, phenotypes have been attributed to specific mutations for only a small fraction of the genes that the collection represents. Thus, the PML collection comprises a largely untapped resource for the discovery of genes that are required for the development of photosynthetically competent chloroplasts in maize.

High-throughput sequencing (HTS) technologies offer new avenues for cataloging transposon-flanking sequences. We describe a method based on the Illumina HTS platform that reports the sequence of several hundred base pairs flanking each side of each Mu element in high-copy Mu lines. The ability to multiplex samples in conjunction with a low false negative rate make this method an economical and rapid approach for linking specific Mu insertions to phenotypes of interest.

RESULTS

Overview of approach

Sequencing with the Illumina platform requires ligation of DNA fragments within a specified size range to adapters that provide anchor points for attachment to the sequencing flow cell and for priming the sequencing reactions. Sequence reads are generated from each of approximately 102 DNA fragments in each channel of the flow cell. To obtain genomic DNA fragments that are suitable for Illumina sequencing and that are derived predominantly from Mu flanking sequences, our method ligates adapters to DNA fragments generated by mechanical shearing, and enriches Mu flanking sequences by hybridization to an oligonucleotide that is complementary to the terminal inverted repeat found at the ends of all Mu elements. An outline of the method follows and is summarized in Figure 1.

Shearing of genomic DNA by sonication. Fragments are ligated to modified Illumina adapters to mark samples from different individuals. Several individuals with the mutation of interest are marked with distinct bar codes and analyzed in the same channel to identify insertions that co-segregate with the mutation.

Enrichment of Mu-containing DNA fragments. Mu-containing DNA fragments are enriched by hybridization to a biotinylated 60-mer oligonucleotide corresponding to the end of the Mu terminal inverted repeat (TIR), which is shared by all members of the Mu family.

DNA amplification. Low-cycle PCR using primers matching the adapter termini amplifies sufficient DNA for application to the flow cell.

Generation of sequence reads. Sequence reads are generated from one end of each DNA fragment. The use of shearing to generate these ends, together with the effective enrichment of Mu flanking sequences and the ability to sequence approximately 103 ends per channel, results in the generation of hundreds or thousands of sequence reads that tile approximately 400 bp flanking each side of each Mu element. The broad sampling of flanking sequence ensures detection of insertions despite potential sequence polymorphisms with respect to the reference B73 genome (Schnable et al., 2009).

Data analysis. An informatic pipeline maps sequence reads to the reference maize genome, identifies clusters of reads that mark Mu insertions, and identifies insertions that co-segregate with the mutation of interest. The pipeline reports maize genes mapping near each insertion, and provides functional annotations for the most closely related genes in rice and Arabidopsis. The annotations facilitate the
identification of the most attractive candidates for validation by gene-specific PCR of additional individuals. These steps are discussed in more depth below.

**DNA shearing and adapter ligation**

Total DNA is sonicated to generate fragments averaging approximately 500 bp. These fragments are too small to include both termini of the same Mu insertion, so subsequent PCR steps are not compromised by the need to amplify across intact Mu elements. In conjunction with hybrid selection of fragments containing the terminal 60 nucleotides of Mu, this size range produces fragment ends that tile approximately 400 bp flanking each Mu TIR.

The ends of the DNA fragments are processed to yield blunt-ended fragments with phosphorylated 5' termini and a 3' adenosine extension. DNA fragments are then ligated to modified Illumina adapters (Bentley *et al.*, 2008) that include 3 bp bar codes adjacent to the sequencing primer binding site (Figure 2). DNA fragments of approximately 220–500 bp are gel-purified to remove adapter dimers, which would otherwise reduce the yield of useful sequence data.

**Enrichment of Mu flanking sequences by hybridization to a Mu TIR oligonucleotide**

To enrich DNA fragments containing Mu-gene junctions, the DNA is denatured and annealed to a biotinylated 60-mer oligonucleotide corresponding to the end of the Mu TIR. Results of an early trial suggested that DNA fragments that span the biotin moiety may be recovered inefficiently; therefore, we place the biotin moiety on the Mu internal end of the oligonucleotide (see Figure 1). The hybridized DNA is collected using streptavidin-coupled magnetic beads. Two successive hybrid enrichment steps are performed to ensure that the majority of the sequenced DNA fragments harbor Mu sequences. Low-cycle PCR using primers that bind to the ends of the adapters is used to bulk up the recovered DNA after each hybrid selection, to reduce the impact of subsequent losses from surface adhesion. The number of cycles is minimized to avoid selecting against fragments that are difficult to amplify by PCR. We have found that 12 and 15 cycles of amplification after the first and second selection rounds, respectively, generate sufficient material for subsequent steps while minimizing cycle number. A final gel purification step yields fragments between approximately 220 and 500 bp, the optimal size range for Illumina sequencing.

We initially tested the effectiveness of the enrichment method by cloning and sequencing a sampling of the DNA fragments recovered at the end of this procedure (data not shown). Thirty-six of the 43 sequenced clones included Mu TIR sequences. Low-cycle PCR using primers that bind to the ends of the adapters is used to bulk up the recovered DNA after each hybrid selection, to reduce the impact of subsequent losses from surface adhesion. The number of cycles is minimized to avoid selecting against fragments that are difficult to amplify by PCR. We have found that 12 and 15 cycles of amplification after the first and second selection rounds, respectively, generate sufficient material for subsequent steps while minimizing cycle number. A final gel purification step yields fragments between approximately 220 and 500 bp, the optimal size range for Illumina sequencing.

We initially tested the effectiveness of the enrichment method by cloning and sequencing a sampling of the DNA fragments recovered at the end of this procedure (data not shown). Thirty-six of the 43 sequenced clones included Mu TIR sequences. Low-cycle PCR using primers that bind to the ends of the adapters is used to bulk up the recovered DNA after each hybrid selection, to reduce the impact of subsequent losses from surface adhesion. The number of cycles is minimized to avoid selecting against fragments that are difficult to amplify by PCR. We have found that 12 and 15 cycles of amplification after the first and second selection rounds, respectively, generate sufficient material for subsequent steps while minimizing cycle number. A final gel purification step yields fragments between approximately 220 and 500 bp, the optimal size range for Illumina sequencing.

**Sequencing**

Reagent kits are available for obtaining reads of 36, 54 or 72 nucleotides, with the cost increasing with read length. Paired-end reads would allow unambiguous identification of
reads that are derived from fragments containing Mu, but this adds cost and is not necessary. We have found that the results obtained with the least expensive option (36 nucleotide reads from a single end) are adequate to unambiguously identify most or all heritable Mu insertion sites. However, we prefer to use 54 nucleotide reads because reads that span gene–Mu junctions retain sufficient similarity to the reference genome to allow alignment using generic alignment software.

Our current protocol routinely generates over two million reads that align to a Mu element when DNA from 20 individuals is analyzed in a single channel. These reads arise from the Mu end of Mu–gene junction fragments, and imply a similar number of reads arising from the gene end of junction fragments. Assuming approximately 100 Mu insertions per individual, this corresponds to roughly 1000 reads per insertion, a number that is consistent with our results for validated insertions, as described below.

**Genetic strategy**

To identify insertions that co-segregate with the mutation of interest, insertions are compared among differentially bar-coded DNAs from multiple individuals carrying the mutation. Transposon display approaches involving Southern blotting or AFLP methods are generally designed to identify a single DNA fragment that is very tightly linked to the mutation of interest, which is then cloned and sequenced. However, the Mu Illumina method provides immediate access to Mu flanking sequences without the need for a separate cloning step, which significantly reduces the effort required to sort through candidates. Consequently, we use a minimized genetic strategy that generates several candidate causal insertions. A typical experiment uses DNA from four individuals harboring the mutation of interest, and that are no more closely related than the ‘first cousin’ relationship shown in Figure 3. Generation of this material requires only two generations following the initial mutant isolation (one out-crossing step, followed by self-pollination of the out-crossed progeny to uncover recessive phenotypes). Assuming 50–100 Mu insertions per plant, this approach is expected to identify a handful of insertions that are in common among all analyzed individuals (in addition to the resident Mu TIRs found stably in many maize lines; see below). The nature of the disrupted genes, and the position of the insertions within them, can then be used to prioritize the most attractive candidates for validation by gene-specific PCR of additional individuals. In an ideal situation, two independent Mu-induced alleles will have been identified and can be analyzed in parallel (marked by different bar codes): genes disrupted in both lines are especially strong candidates for validation.

The analysis of homozygous wild-type relatives can be useful to reduce the number of candidate insertions for follow-up. However, the identification of +/+ relatives of Mu-induced mutants is complicated by an epigenetic phenomenon referred to as Mu suppression, which masks the phenotypes of many Mu-induced mutations when Mu activity is silenced (May et al., 2003; Lisch and Jiang, 2009). If the phenotype of interest is known to be expressed when Mu activity is silenced, then comparisons among closely related +/+ and –/– individuals is recommended.
However, we do not routinely incorporate apparent +/- relatives into our analyses because of uncertainty regarding the genotype of such plants.

**Multiplexing**

The sequencing of approximately $10^7$ DNA fragment ends per channel, coupled with the high degree of enrichment for Mu-containing fragments, allows multiplexing to reduce costs. Analysis of DNAs from 20 different individuals in one channel generates ample depth of sequencing (approximately 400–4000 reads per Mu insertion) to unambiguously map most or all heritable Mu insertions. Thus, we typically analyze four individuals representing each of five different mutants in the same channel. Deeper multiplexing should be possible, but we have not tested this. To reduce the labor associated with sample preparation, we also multiplex within each bar code: each channel includes ten bar codes, each of which marks two DNA samples from unrelated mutants. Comparisons among the data returned for each bar code identify insertions that are shared among individuals harboring the same (or allelic) mutations.

**Identifying Mu insertion sites**

Sequence reads are initially placed into bins according to their bar code, which identifies the DNA samples from which they are derived. Approximately half of the Mu-gene junction fragments will be sequenced from the gene end and half from the Mu end. Reads that align to any member of the Mu family are counted, providing a rapid means to estimate the success of the run. The remaining reads are aligned to the reference maize genome using Bowtie (Langmead et al., 2009), with a seed length of 28 nucleotides, a maximum of two mismatches in the seed, and the ‘–best’ parameter, which chooses one alignment from among multiple hits via a sequence quality assessment.

Our analysis pipeline next identifies clusters of reads marking putative Mu insertions. A read cluster is defined as having a minimum number of reads (e.g. 400, but set by the user) that map within boundaries delimited by the first gap between reads of >100 bp. The number of reads flanking each insertion will depend on the degree of multiplexing and the quality of the DNA samples. With our current instrumentation and protocol, heritable Mu insertions in a 20-fold multiplexed channel are reliably identified by clusters of >400 reads spanning a genomic region of 300–1500 bp: all 42 clusters with these features that we have tested by gene-specific PCR have been confirmed as representing heritable Mu insertions (data not shown). There is some variation in the number of reads marking heritable insertions (approximately 400 to several thousand), probably due to bias in the sequencing reactions and/or PCR. Non-heritable insertions resulting from somatic transpositions presumably exist in our samples, which come from Mu-active maize lines. However, somatic insertions contribute negligible signal when the data are analyzed according to these parameters, most probably because such insertions are found in only a small fraction of the cells used for DNA extraction and are thus marked by fewer sequence reads. Maize is highly polymorphic, and our Mu-active lines are derived from diverse backgrounds. Nonetheless, the deep sampling of approximately 400 bp flanking each Mu TIR yields ample read alignments to highlight virtually all heritable Mu insertions (see Discussion below).

The distribution of reads within a cluster can be visualized using various tools, including the Integrative Genomics Viewer (IGV) (http://www.broadinstitute.org/igv). Although not routinely necessary, viewing the underlying reads can be helpful when following up clusters of particular interest. The viewer displays the directionality of each read using arrows of different colors; this highlights the Mu insertion site as the point of convergence between reads arising from the two strands (Figure 1b and Figure S2). The IGV viewer can also display the sequence of individual reads. Precise Mu insertion sites are identified by the presence of Mu TIR sequences in fragments spanning the Mu–gene junction (Figure S2).

We used IGV to visually inspect 50 randomly selected read clusters meeting the criteria outlined above (>400 reads spanning between 300 and 1500 bp on the reference genome, and including at least one read detecting a gene–Mu junction). Forty-seven of these 50 Mu insertions are marked by numerous sequence reads on each side of Mu (data not shown; analogous data are shown in Figure S2 for the four gene identifications discussed below). Because the sequences on the two sides of each insertion are recovered as independent events, this finding demonstrates that the false negative rate for detection of Mu flanking sequences is extremely low. Reads that do not arise from Mu–gene junction fragments do not result in false positives because they rarely cluster in a specific genomic region, and the few that do cluster lack reads spanning a Mu–gene junction.

**Annotating insertion sites**

The analysis pipeline identifies read clusters that meet the criteria described above, and screens them to identify those matching a set of background clusters detected in most DNA samples (Table S1). These presumably represent stable Mu remnants (Settles et al., 2004; Yi et al., 2009), other sequences with similarity to the Mu TIR, or highly repeated sequences that tend to contaminate the preparations. Clusters matching this background set are removed from consideration as candidate genes. The remaining clusters are then annotated using the following automated scheme. The genomic region spanned by each cluster (typically approximately 800 bp) is appended with 100 bp of additional genomic sequence on each side, and used for a BLASTN search of maize, rice and sorghum gene models. The gene model that emerges as the top hit is reported; this is a maize gene except in the rare instance in which a maize gene is not

© 2010 The Authors
recognized (e.g. it may be missing from the draft genome or have a faulty gene model). In such cases, rice and sorghum serve as back-up to capture the relevant gene. A very small fraction of read clusters fail to detect genes in any of these organisms and are therefore not annotated. An output table reports the following information: the position of the cluster in the reference maize genome, the number of reads within the cluster, the number of base pairs spanned by the cluster, the locus ID of the identified maize gene, the locus ID of the most closely related Arabidopsis gene, and InterPro domains and targeting predictions (Predotar and TargetP) (Emanuelsson and von Heijne, 2001; Small et al., 2004) for both the maize and Arabidopsis proteins. In addition, the gene description for the most closely related Arabidopsis gene is imported from the Arabidopsis Information Resource (http://www.arabidopsis.org/). Ready access to these functional annotations is valuable for prioritizing candidates for validation tests. Table S2 provides an example of an output table representing the data for one bar code (i.e. two DNA samples) in a recent 20-fold multiplexed channel.

An automated process compares the Mu insertions among relevant samples to identify those that co-segregate with a particular mutation. Typically, between one and eight insertions co-segregate with the mutation of interest, but this number depends upon the relationships among the individuals analyzed. Where multiple allelic mutants are included in the analysis, the presence of insertions at different sites within the same gene provides immediate strong evidence that this gene is the relevant one. Otherwise, tight linkage between an insertion and phenotype is established subsequently by gene-specific PCR of additional individuals.

Validation: identification of the causal insertions in four chloroplast biogenesis mutants

We initially optimized the procedure by analyzing DNAs harboring two known insertions (data not shown). Since then, we have used four Illumina channels to analyze 80 samples from the PML collection that represent mutations in 15 different ‘unknown’ genes. Twelve of these 15 analyses successfully identified the causal insertions, as validated by gene-specific PCR of multiple alleles or additional individuals. Table 1 summarizes the locus and allele information for four of these identifications; supporting data for these are provided below. The other eight identifications will be the subject of future publications. Three attempts to identify the disrupted gene failed. These failures all involved alleles that we had previously attempted to identify by Southern blot display of Mu insertions. That both the Southern blotting and Mu Illumina approaches failed to detect a causal Mu insertion suggests that these alleles may not be tagged by Mu. Thus, our success rate at identifying causal Mu insertions is at least 80% and possibly higher.

Identification of two genes involved in chloroplast gene expression. Mutations in the PML collection that are found to cause defects in chloroplast RNA metabolism are assigned the temporary prefix ‘crp’ for ‘chloroplast RNA processing’. Mutations in crp4 caused a slight reduction in seedling chlorophyll content (Figure S1) and the accumulation of aberrant, extended chloroplast RNAs from several loci (Figure 4a, and data not shown). Four independent crp4 alleles were analyzed by Mu Illumina. The results revealed a Mu insertion in the gene encoding chloroplast polynucleotide phosphorylase in plants harboring each of these alleles (Figure 4b,c). This identification makes good sense, as polynucleotide phosphorylase functions as a 3¢-5¢ exoribonuclease (reviewed by Bollenbach et al., 2007), and its knockdown or knockout in Arabidopsis results in analogous RNA defects (Walter et al., 2002; Marchive et al., 2009).

Mutations in crp5 cause a pale yellow–green seedling phenotype (Figure S1) and aberrant transcript patterns from several chloroplast loci (Figure 5a, and data not shown). Mu Illumina identified insertions in the gene encoding the maize ortholog of Arabidopsis PTAC12 in each of two crp5 alleles (Figure 5b). Arabidopsis PTAC12 is associated with the chloroplast ‘transcriptionally active chromosome’ (Pfalz

Table 1 Maize genes described in this study. Putative rice orthologs were identified as the top BLASTN hit in a query of rice gene models at http://rice.plantbiology.msu.edu/. The rice gene ID was used to recover the Arabidopsis ortholog from the POGs database (Walker et al., 2007) (http://pogs.uoregon.edu/)

<table>
<thead>
<tr>
<th>Maize gene</th>
<th>Maize gene ID (<a href="http://www.maizesequence.org">http://www.maizesequence.org</a>)</th>
<th>Rice ortholog</th>
<th>Arabidopsis ortholog</th>
<th>Gene product</th>
</tr>
</thead>
<tbody>
<tr>
<td>tha5</td>
<td>GRMZM2G300408</td>
<td>Os08g30830°</td>
<td>AT4G14870</td>
<td>cpSecE</td>
</tr>
<tr>
<td>pet2/ZmHCF208</td>
<td>GRMZM2G087083</td>
<td>Os10g37840</td>
<td>AT5G52110</td>
<td></td>
</tr>
<tr>
<td>crp4</td>
<td>GRMZM2G377761</td>
<td>Os07g07310</td>
<td>AT3G03710</td>
<td>Chloroplast polynucleotide phosphorylase (Walter et al., 2002; Marchive et al., 2009)</td>
</tr>
<tr>
<td>ZmPTAC12</td>
<td>GRMZM2G005938</td>
<td>Os08g09270</td>
<td>AT3G59040</td>
<td>Maize ortholog of Arabidopsis PTAC12 (Pfalz et al., 2006)</td>
</tr>
</tbody>
</table>

°This rice locus had not been recognized as the gene encoding cpSecE due to a faulty gene model.
et al., 2006), but lacks known domains and is of unknown function. The distinctive RNA defects in crp5 mutants were not reported for the corresponding Arabidopsis mutant (Pfalz et al., 2006), and suggest unanticipated functions for this protein. The crp5 gene has been assigned the permanent name ZmPTAC12.

Identification of two genes involved in assembly of the photosynthetic apparatus. Proteins are targeted to the thylakoid lumen via two ancient pathways, the cpSec pathway and the cpTAT pathway (reviewed by Cline and Dabney-Smith, 2008). PML mutants with chloroplast protein profiles similar to those in maize mutants with established defects in these pathways (Voelker and Barkan, 1995b; Voelker et al., 1997; Walker et al., 1999) were assigned the prefix ‘tha’ for ‘thylakoid assembly’. tha5 mutants have a pale-green seedling phenotype (Figure S1), and a protein profile that phenocopies that of tha1 mutants (Voelker and Barkan, 1995b): reduced levels of the core subunits of photosystem I, photosystem II and the cytochrome b6f complex, a reduction in the lumenal protein plastocyanin (a substrate of the cpSec pathway), and increased accumulation of the stromal intermediate of plastocyanin (Figure 6a,b).

tha1 encodes cpSecA (Voelker et al., 1997). Therefore, these results suggest that tha5 also encodes a protein required for the cpSec pathway. Indeed, Mu Illumina analysis of two tha5 mutant alleles identified Mu insertions in the gene encoding cpSecE (Figure 6c), a component of the thylakoid Sec translocon (Schuenemann et al., 1999; Froderberg et al., 2001). cpSecE mutants have not been reported previously in plants. However, disruption of a nuclear gene encoding cpSecY, a partner of cpSecE in the Sec translocon, causes a much more severe phenotype (Roy and Barkan, 1998). Whether these phenotypic differences reflect differences in allele strength or the ability of cpSecY to function to some extent without cpSecE remains to be determined.
The signal from somatic insertions is negligible, even in spanning approximately 800 bp of genomic sequence. Insertion is represented by approximately 400–3000 reads of low rate of false positives and false negatives. Each heritable sequences with maize genes, related genes in other model feasible. An informatic pipeline associates Mu five different genes, and deeper multiplexing is probably Illumina channel currently allows simultaneous analysis of required for the attachment of heme to cytochrome b6 respectively (Cline and Dabney-Smith, 2008). Loss of the mature form of PC in to subunits of photosystem II (D1), photosystem I (PSA-D), the cytochrome b6\_f complex (PET-D) and thylakoid ATP synthase (ATP-A). This protein profile is similar to that reported for tha1 mutants (Voelker and Barkan, 1995b), which have a Mu insertion in the gene encoding cpSecA (Voelker et al., 1997). (b) tha5 mutants have a defect in the translocation of plastocyanin across the thylakoid membrane. Stromal and thylakoid fractions were obtained from chloroplasts isolated from tha5-1 mutants. An aliquot of the thylakoid fraction was incubated with carbonate to remove extrinsic proteins on the stromal face, and then centrifuged to recover membranes (carb pellet). The results show that iPC accumulates outside the thylakoid lumen, as shown previously for tha1 mutants (Voelker and Barkan, 1995b).

(c) Summary of Mu insertion sites. The diagram shows a portion of maize locus GRMZM2G300408, predicted to encode cpSecE. The sequences of the target site duplications are shown below. The Illumina reads that identified these insertions are shown in Figure S2b. These insertion sites were also validated by sequencing Mu–gene junction fragments obtained by gene-specific PCR (data not shown).

Mutants in the PML collection that specifically lack the cytochrome b6\_f complex have been assigned the prefix ‘pet’ for ‘photosynthetic electron transport’. Mutations in pet2 cause a pale-green seedling phenotype (Figure S1), and were shown previously to disrupt accumulation of the cytochrome b6\_f complex at a post-translational step (Voelker and Barkan, 1995a). Mu Illumina analysis of two pet2 alleles (Figure 7) identified insertions in the maize ortholog of the Arabidopsis gene HCF208 and the Chlamydomonas gene CCB2 (Kuras et al., 1997; Lyska et al., 2007). HCF208/CCB2 is required for the attachment of heme to cytochrome b6. The mutant phenotypes reported in Arabidopsis and Chlamydomonas are analogous to that seen in maize. The maize pet2 mutants offer the opportunity to study this gene in a C4 plant.

**DISCUSSION**

**Mu** Illumina analysis provides a rapid means to map heritable Mu insertions to specific genomic sequences with a low rate of false positives and false negatives. Each heritable insertion is represented by approximately 400–3000 reads spanning approximately 800 bp of genomic sequence. The signal from somatic insertions is negligible, even in Mu–active maize lines. Twenty-fold multiplexing in a single Illumina channel currently allows simultaneous analysis of five different genes, and deeper multiplexing is probably feasible. An informatic pipeline associates Mu flanking sequences with maize genes, related genes in other model organisms, and functional annotations, which simplifies identification of the most promising candidates for validation by gene-specific PCR.

Several methods have been reported previously that provide alternatives to Southern blot display for identifying transposon insertions underlying mutant phenotypes (Frey et al., 1998; Van den Broeck et al., 1998; Hanley et al., 2000; Yephremov and Saedler, 2000; Settles et al., 2004, 2007; Wang et al., 2008; Liu et al., 2009; Yi et al., 2009). Mu Illumina analysis adds to this toolkit. The method of choice will vary with the scale of the project, the copy number of the
transposable element, and the instrumentation that is available to the researcher. Mu Illumina analysis is especially suited to large-scale efforts because the cost per gene decreases with the degree of multiplexing. The method incorporates several unique features that are expected to result in a particularly low false negative rate for Mu insertion detection: (i) mechanical shearing to generate the DNA termini from which sequence reads are derived, (ii) minimized use of PCR, (iii) use of a long oligonucleotide primer, and (iv) use of the Illumina platform, which maximizes detection of insertions via its great depth of sequencing. Unlike methods that rely on the ligation of adapters to a fixed restriction site, sequences obtained with this method tile an extended region flanking each insertion. This is especially pertinent in highly polymorphic species like maize to ensure that sufficient sequence is captured near each transposon to obtain reads that align unambiguously with the reference genome. The minimized use of PCR reduces PCR bias, and the hybrid selection strategy captures sequences flanking even defective TIRs that are missing up to approximately 15 bp at their terminus (unpublished observations). Another significant advantage of Mu Illumina analysis, and one that is shared with a related method that uses the 454 platform (Liu et al., 2009), is that the gene disrupted by each insertion is reported immediately, without the need for a separate cloning and sequencing step. The resulting ease of moving from co-segregating insertion to validation makes generation of an extensive pedigree unnecessary. Somatic insertions do not interfere with this approach because the number of sequence reads marking an insertion reflects the prevalence of that insertion in the initial DNA sample. However, should a researcher wish to study somatic events, the parameters for insertion identification could be modified to report clusters with fewer reads.

The Mu Illumina method is enhancing our ability to exploit the PML mutant collection as a means to identify genes involved in chloroplast biogenesis and photosynthesis. The PML collection complements two resources that are under development in Arabidopsis: Chloroplast 2010 (Aijawi et al., 2010) and the Chloroplast Function Database (Myouga et al., 2009). Whereas the Arabidopsis projects are targeting genes that are predicted to encode chloroplast-localized proteins, PML is a phenotype-driven resource. The powerful genomic resources and ease of transformation offered by Arabidopsis are unmatched. However, maize offers other advantages. For example, the large maize seed supports heterotrophic growth for several weeks, facilitating access to tissue from non-photosynthetic mutants. In addition, the ease of generating large quantities of chloroplast extract from maize simplifies the recovery of co-immunoprecipitates for mass spectrometry (e.g. Watkins et al., 2007; Kroeger et al., 2009) and RNA co-immunoprecipitation assays (e.g. Schmitz-Linneweber et al., 2005). Maize has proven to be particularly useful for studying mutations that, either directly or indirectly, cause the loss of plastid ribosomes (Jenkins and Barkan, 2001; Ostheimer et al., 2003; Asakura and Barkan, 2006; Schmitz-Linneweber et al., 2006; Watkins et al., 2007; Beick et al., 2008; Kroeger et al., 2009): Arabidopsis embryogenesis is more sensitive to compromised chloroplast translation than is maize embryogenesis, a difference that is believed to reflect differences in chloroplast gene content (reviewed in Stern et al., 2004; Asakura and Barkan, 2006). Current data support the view that the functions of most chloroplast biogenesis genes are largely conserved in Arabidopsis and maize; thus concurrent study of orthologues in both organisms can be advantageous (Asakura and Barkan, 2006, 2007; Asakura et al., 2008).

We developed the Mu Illumina method for use in phenotype-driven gene discovery. However, a by-product is the identification of other heritable insertions segregating in the same lines. From each PML plant analyzed, we identify approximately 100 read clusters with features characteristic of heritable Mu insertions (see Table S2 for an example). This Mu copy number is higher than that reported for UniformMu lines (Settles et al., 2004; McCarty et al., 2005), and correlates with the very high forward mutation rate in our lines. These ‘bystander’ insertions are available for community use as a reverse-genetic resource via a searchable interface at http://teosinte.uoregon.edu/mu-illumina.

**EXPERIMENTAL PROCEDURES**

**Mu Illumina method**

Illumina data were generated on an Illumina Genome Analyzer Ix at the University of Oregon Core Genomics Facility. Appendix S1 provides a detailed protocol for DNA sample preparation.

**PCR genotyping**

Gene-specific PCR of Mu insertion sites was performed as follows. Two sets of PCR reactions were performed for each insertion: the mutant allele was amplified with a gene specific primer and a degenerate MuTIR primer (5’-GCCCTGTACATCGATCCAGATCCGGAATCC(G/C))); the wild-type allele was amplified with gene specific primers that flank the insertion site, and that are separated by approximately 400–1500 bp. Primers flanking the insertion were designed using Primer 3 (Rozen and Skaletsky, 2000) with the following parameters: Tm between 60 and 68°C, length between 18 and 27 nucleotides, and product size 400–1500 bp. For PCR reactions, we used Phusion polymerase (Finzymes, http://www.finzymes.com) with the ‘GC’ buffer supplied with the enzyme, and the following reaction profile: 98°C for 30 sec, followed by 34 cycles of 98°C for 10 sec, 62°C for 15 sec and 72°C for 30 sec, with a final extension of 72°C for 10 min. The gene-specific primers used here are summarized in Table S3.

**Other methods**

The methods used for immunoblotting and RNA gel blot hybridization have been described previously (Barkan, 1998). The antibodies and chloroplast fractionation protocol used to demonstrate the defect in the cpSec pathway in tha5 mutants were described by Voelker and Barkan (1995b).
ACKNOWLEDGMENTS

We are grateful to Doug Turnbull, Todd Mockler (Oregon State University) and Eric Johnson for useful discussions during the development of this method, and to Doug Turnbull for performing the Illumina sequencing. We also wish to thank Elise Kikis, Jamie Brenchley and Faith Harris for cataloging chloroplast RNA defects in the PML collection, and the many undergraduates in the Barkan laboratory who cataloged protein deficiencies in PML mutants. This work was supported by grants DBI-0421799, DBI-0077756 and DBI-0922560 from the US National Science Foundation.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Seedling phenotypes of mutants represented in the PML collection.

Figure S2. Screen captures showing sequence reads marking the Mu insertion sites in crp5 (ZmPTAC12), tha5 (cpSecE) and pet2 (HCF208/CBB).

Table S1. Background read clusters identified in most DNA samples.

Table S2. Example of output from cluster analysis pipeline.

Table S3. Synthetic oligonucleotides used in this study.

Appendix S1. Detailed protocol for Mu Illumina sample preparation. Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

REFERENCES


Till, B., Schmitz-Linneweber, C., Williams-Carrier, R. and Barkan, A. (2001) CRS1 is a novel group II intron splicing factor that was derived from a domain of ancient origin. RNA, 7, 1227–1238.


