OXFORD
G3 Genes | Genomes | Genetics

G3, 2021, 11(7), jkab138
DOI: 10.1093/g3journal/jkab138
Advance Access Publication Date: 27 April 2021
Investigation

A capture-based assay for detection and characterization of transposon polymorphisms in maize

Minqi Li,[1] Jaclyn M. Noshay ,[2] Xiaoxiao Dong,[1] Nathan M. Springer ,[2] and Qing Li[1,]*

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China
[2]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN 55108, USA

*Corresponding author: National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Shizishan Street No 1, Hongshan District, Wuhan 430070, China. Email: qingli@mail.hzau.edu.cn

Abstract

Transposons can create allelic diversity that affects gene expression and phenotypic diversity. The detection of transposon polymorphisms at a genome-wide scale across a large population is difficult. Here, we developed a targeted sequencing approach to monitor transposon polymorphisms of interest. This approach can interrogate the presence or absence of transposons reliably across various genotypes. Using this approach, we genotyped a set of 965 transposon-related presence/absence polymorphisms in a diverse panel of 16 maize (*Zea mays* L.) inbred lines that are representative of the major maize breeding groups. About 70% of the selected regions can be effectively assayed in each genotype. The consistency between the capture-based assay and PCR-based assay are 98.6% based on analysis of 24 randomly selected transposon polymorphisms. By integrating the transposon polymorphisms data with gene expression data, ∼18% of the assayed transposon polymorphisms were found to be associated with variable gene expression levels. A detailed analysis of 18 polymorphisms in a larger association panel confirmed the effects of 10 polymorphisms, with one of them having a stronger association with expression than nearby SNP markers. The effects of seven polymorphisms were tested using a luciferase-based expression assay, and one was confirmed. Together, this study demonstrates that the targeted sequencing assay is an effective way to explore transposon function in a high-throughput manner.

Keywords: transposon; targeted sequencing; gene expression; maize

Introduction

Transposons were first discovered by Barbara McClintock in maize (Fedoroff 2012). They are segments of DNA sequence that can change position within a genome. There are at least two major classes of transposons, retrotransposons and DNA transposons, with each having many subclasses. The two classes transpose via different intermediates; retrotransposons use an RNA intermediate while DNA transposons use a DNA intermediate (Wicker *et al.* 2007).

Transposons account for ∼85% of the maize genome (Schnable *et al.* 2009), and are distributed throughout the chromosomes, including the heterochromatic centromeres and the euchromatic arms. In the B73 reference genome nearly 50% of all genes are located within 1 kb of an annotated transposon (Anderson *et al.* 2019), suggesting a potential role of transposon in regulating gene expression. Consistently, there are a growing set of examples in which the causative basis for quantitative trait loci (QTL) underlying natural phenotypic diversity was found to be transposon insertions that influence expression of nearby genes. The traits that are affected by transposons span a range of diverse biological processes; including flowering time (Salvi *et al.* 2007; Huang *et al.* 2018), drought tolerance (Mao *et al.* 2015; Wang *et al.* 2016), disease resistance (Yang *et al.* 2013), ear development (Liu *et al.* 2015; Jia *et al.* 2020), and branching (Studer *et al.* 2011). Transposons can cause large genetic variations and have a complex interplay with the epigenome (Noshay *et al.* 2019). They can either repress or enhance gene expression (Slotkin and Martienssen 2007). Though initially regarded as "junk DNA," transposons are becoming recognized as an important potential source of regulatory elements (Feschotte 2008; Lisch 2013).

Transposon polymorphisms can be detected by PCR-based assays but this is difficult for large transposons or terminal inverted repeat sequences, both of which can lead to PCR amplification failure. Transposons can also be detected by next-generation sequencing, but mapping of sequencing reads to reference genomes can be complex due to the complicated organization and repetitive nature of transposons. Therefore, many bioinformatics pipelines have been developed specifically to detect the presence/absence polymorphism of transposons (Quadrana *et al.* 2016; Stuart *et al.* 2016). Those methods usually rely on the paired-reads that are mapped discordantly or the reads that can be successfully mapped to the reference genome by splitting or by soft-clipping. These methods have been applied in *Arabidopsis* and tomato, and the number of detected polymorphisms depends on the sequencing depth and the genome organization (Quadrana *et al.* 2016; Stuart *et al.* 2016; Domínguez *et al.* 2020). Several studies have developed sequence capture-based assays

Received: October 5, 2020. **Accepted:** April 15, 2021

to detect polymorphisms of the transposon families that are of interest (Baillie *et al.* 2011; Quadrana *et al.* 2016, 2019). In these methods, the capture probes are designed to match the termini of transposons of interest (Supplementary Figure S1). These probes can enrich regions flanking transposons and can detect insertion of transposons based on sequencing reads that span both the transposons and the flanking regions. This approach is powerful for detecting novel insertions that are not present in the reference genome. Therefore, this has been applied in studies to detect new transposons insertions due to recent transposition events (Baillie *et al.* 2011; Quadrana *et al.* 2019). While the presence of reads spanning transposon and flanking regions indicates transposon insertions, the absence of such reads does not imply transposon absence since failure in capture can also result in lack of junction-spanning reads. Therefore, these methods have limited power in the detection of the true absence of transposons.

In maize, large-scale characterization of transposon function is limited due to the lack of an effective high-throughput assay that can interrogate transposon polymorphisms in a large population. Based on the reported case studies of transposons in gene regulation and the wide occurrence of transposons in maize genome, it is worth developing effective assays to systematically study transposon function in maize. Here, we provided a capture-based approach that can detect the presence/absence variation of transposons across a large panel of population. We characterized and verified the function of several transposons in gene expression by integrating with transcriptome data, suggesting the potential to identify transposons that have a role in phenotypic diversity.

Materials and methods

Materials

The 16 inbred lines that are used for targeted sequencing of the interested regions are selected from a panel of >500 inbred lines that are collected all over the world (Yang *et al.* 2011). They represent the three maize groups, including tropical/semitropical (TST), stiff stalk (SS), and non-stiff stalk (NSS). Seedling leaves were used for DNA extraction using standard CTAB method. The genetic distance of these lines were constructed based on 1,060,926 SNPs using the SNPhylo package (Lee *et al.* 2014).

Insertion/deletion identification and annotation

We hypothesized that transposons might create sequence variation that can lead to differential gene expression. To test this idea, we used genes with *cis*-expression QTL (eQTL) detected by Li *et al.* (2013). We retained genes for which the *cis*-eQTL can explain >50% of the variance in gene expression. The 5000 bp upstream of transcriptional start site of these genes were obtained for the two parental genotypes, B73 and Mo17. Genes with missing sequences in any one of the two genotypes were not used. A pairwise alignment was performed using MUMmer (Marçais *et al.* 2018), and Insertion/deletion (InDels) were identified using NCBI BLAST+ (Cock *et al.* 2015). The relative position of the InDels in the promoter sequences were recorded and were used to derive the position of the InDels in the B73 or Mo17 reference genomes. The InDels were annotated using RepeatMasker (http://www.repeatmasker.org, last accessed: 2019/12/23) and those that are covered by transposon (named as TE-InDel) were retained. In some cases, multiple transposons were identified within the insertion sequence, likely due to nested transposon insertions. In such circumstances, the transposon superfamily of the TE-InDel was based on the transposons that covered the longest sequences of the insertion. Only InDels that are with a size of 100–5000 bp were retained since many of the transposons that have been reported to associate with expression and phenotypic changes are within this size range (Salvi *et al.* 2007; Studer *et al.* 2011; Yang *et al.* 2013; Liu *et al.* 2015; Mao *et al.* 2015; Wang *et al.* 2016; Huang *et al.* 2018; Jia *et al.* 2020).

Probe design and sequencing

The 200 bp flanking sequences of the TE-InDels were identified on each side, and were blasted against the entire B73 genome. Only the set of TE-InDels that contain unique sequences on both sides were used for design of capture probes in order to minimize the possibility to capture off-target regions. The 200 bp flanking sequences on both sides were used to design capture probes which are of sizes of 50 bp. The end of probe that is close to the TE-InDels were required to be within 10 bp of the insertion boundaries to maximize our probability of finding junction reads that are effective for InDel identification (Supplementary Figure S1). In few cases where qualified probes cannot be designed, the distance between the probes to the insertion boundary was allowed to be within 50 bp.

To prepare the DNA sequencing libraries, 200 ng genomic DNA of each line was sheared by Biorupter (Diagenode, Belgium) to acquire 150–200 bp fragments. The ends of DNA fragment were repaired and Illumina adaptor was added (Fast Library Prep Kit, iGeneTech, Beijing, China). The libraries were captured with AIExome Enrichment Kit V1 (iGeneTech, Beijing, China) and sequenced on Illumina platform (Illumina, San Diego, CA, USA) with 150 bp paired-end reads. The probe synthesis, library preparation, and sequencing were performed by iGenetech (Beijing, China).

Alignment and InDel calling

The sequencing reads were aligned to both the B73 and Mo17 reference genomes using Tophat2 (Kim *et al.* 2013). Each of the mapped reads was then subjected to the following process to determine whether it spans the InDel boundary (Supplementary Figure S2). First, the position of the insertion site in the two reference genomes B73 and Mo17 was determined. For the line with the insertion, both the left and right positions were recorded. For the line without the insertion, only one single position was recorded. Next, we extended the position of the insertion site to each direction for 10 bp and recorded the new positions covered by the region (a length of 20 bp, named as "20-bp region"). The position of each mapped read was then compared with the new positions. Supplementary Figure S2 illustrates the situation of how insertion or deletion was determined. In cases where a line (with unknown insertion or deletion) is aligned to the reference genome with the insertion, a read was determined to span the insertion boundary if the "20-bp region" is completely within the region that the read was mapped. A presence of insertion was called if both the left and right boundaries were covered by at least one read. An absence of insertion was called if no reads were found to cover both the left and right boundaries. Missing data (FALSE in the figure) were called if only one boundary was found to have covered reads. In cases where a line was aligned to the reference genome without the insertion, an absence of insertion is called if a read spanning the boundary was identified. A presence of insertion is called if no read spanning the boundary as identified. Finally, the results from the two alignments (one to the reference with the insertion and one to the genome without the insertion) were combined. Only consistent calls between the two alignments were kept.

PCR-based genotyping

Two kinds of PCR with different purposes were carried out, one to validate the TE-InDel calling results based on targeted sequencing, and one to explore the function of TE-InDel in a large panel of 140 lines. For the first purpose, a set of 24 TE-InDels were randomly selected and PCR amplified in the same 16 lines that are used for targeted sequencing. For the second purpose, a set of 18 TE-InDels were randomly selected from the set that are significantly associated with gene expression based on *t*-test of the 16 lines. All Primers were designed to span the InDels (Supplementary Table S1). The PCR product was run on the agrose gel, and missing data were recorded as NA.

Association between TE-InDels and gene expression

To test association between TE-InDel and gene expression in the panel of 16 lines, we first choose a set of TE-InDels that can be tested. We require the TE-InDel to have at least 4 genotypes with data for both the TE-InDel calling and expression. We also require each of the two groups with or without the insertion must have at least two genotypes. Expression data was from kernels 15 days after pollination. A *t*-test was performed for each TE-InDel using the R function *t*-test. A cut-off of $P < 0.05$ was considered significant. The mean expression level of the insertion group was compared with that of the deletion group, bigger value in the insertion group suggests a positive effect of the insertion on gene expression and smaller value indicates a negative effect.

In the large panel of 140 lines, an association analysis controlling for both population structure and individual relatedness was used to test the effect of the TE-InDel on gene expression using TASSEL (Bradbury *et al.* 2007). We also included the SNPs that are located on the 5 kb flanking regions. These SNPs were identified in a previous study (Liu *et al.* 2017). A *P*-value of less than 0.05 was considered significant.

Luciferase-based assay

To test the effect of TE-InDels on gene expression, a dual-Luciferase (LUC) transient expression assay was performed using maize protoplasts. The basic vector (35Smini_pGreen_luc_0800) contains the firefly *LUC* gene that is driven by a minimal sequence from the 35S promoter (mini35S). The same vector also contains a renilla *LUC* (*REN*) gene driven by a strong 35S promoter, which serves as an internal control to ensure similar transformation efficiency among comparisons. The TE-InDel sequences, not including any flanking sequences, were placed before the mini35S of the basic vector. Two types of vectors with the TEs orienting in opposite direction relative to the *LUC* gene was constructed. The vector containing the transposon (TE+) together with the basic vector (which serves as the control) was transformed into the maize mesophyll protoplasts that are prepared from B73 seedlings leaves grown in the darkness for about 10 days. The plasmids were transformed into the mesophyll protoplasts using polyethylene glycol-mediated transformation (Yoo *et al.* 2007). Both LUC and REN activities were measured using the dual-LUC reporter assay system (Yeasen, Shanghai and Promega, Madison, WI, USA). At least five biological replicates were performed for this assay. Relative LUC activity was calculated by normalizing the LUC activity to the REN activity.

Data availability

All the data generated in this study are available at NCBI with the accession number: SRP285835. The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables. Supplemental Material available at figshare: https://doi.org/10.25387/g3.14346113.

Results

The capture-based assay can detect transposon polymorphisms robustly

We were interested in exploring the hypothesis that transposon polymorphisms might create expression differences between alleles in maize. To test this hypothesis, we focused on transposons that are located within the promoter regions of genes with *cis*-eQTL in the recombinant inbred line (RIL) population derived from B73 and Mo17 (Li *et al.* 2013). This set of genes was used since high-quality genome sequences were available for the two parental haplotypes to allow identification of promoter polymorphisms (Schnable *et al.* 2009; Sun *et al.* 2018). The regions 5000 bp upstream of the gene were used to identify transposon-related Insertion/deletions (TE-InDel). In total, 965 TE-InDels were identified for 758 genes (Figure 1, Supplementary Table S2). About 78% of genes have one InDel in the 5000 bp upstream regions, and the rest have two or more InDels (Supplementary Figure S3A). The proportion of transposon-related sequence within the TE-InDels varies from 3.86% to 100%, with 879 (91%) having >50% of the insertion sequence classified as a transposon (Supplementary Figure S3B). The 965 TE-InDels include 454 that have an insertion in B73 and 511 that have an insertion in Mo17. We also included 83 InDels in which the insertion does not contain any sequence annotated as a transposon (non-TE InDels), resulting in a total of 1048 InDels.

A capture-based sequencing platform was developed to genotype these InDels across multiple genotypes (Supplementary Figure S1). Probes were designed on the two flanking sides and were used to enrich the target regions from libraries that are prepared for next-generation sequencing. We first assayed these InDels in B73 and Mo17 because the availability of the reference genomes for these two genotypes allows the development of criteria for classifying InDels (Supplementary Figure S2) and the evaluation of the feasibility of this capture-based assay. The capture assays resulted in ∼13 and ∼25 M reads for B73 and Mo17, respectively (Table 1). The mapping rate for B73 and Mo17 to reference genome (B73) is 90.4% and 38.8%, respectively. The lower mapping rate of Mo17 is consistent with the fact that Mo17 is polymorphic at these target regions from B73. Between 10.0% and 59.9% of the mapped reads are located within the target regions, indicating the enrichment is effective. In fact, 87.9% (Mo17) and 100.0% (B73) of the target regions have read coverage (Table 1).

We developed criteria to classify the presence/absence of each insertion (Supplementary Figure S2, see details in Materials and Methods). Figure 1B shows an example of alignment of three libraries from three genotypes (B73, Mo17, and B110) to either B73 or Mo17 reference genome. In this example, the insertion is present in B73 and absent in Mo17. Alignment of the reads from B73 to the B73 reference genome exhibits coverage that continues over the junction of the InDel while Mo17 reads align to the flanking regions but do not cover the insertion. In contrast, alignment of the Mo17 reads to the Mo17 reference genome confirms continuous alignment of reads over the InDel junction but in B73 there is a lack of coverage over the InDel junction (Figure 1B). The third genotype (B110) is predicted to contain the insertion as it

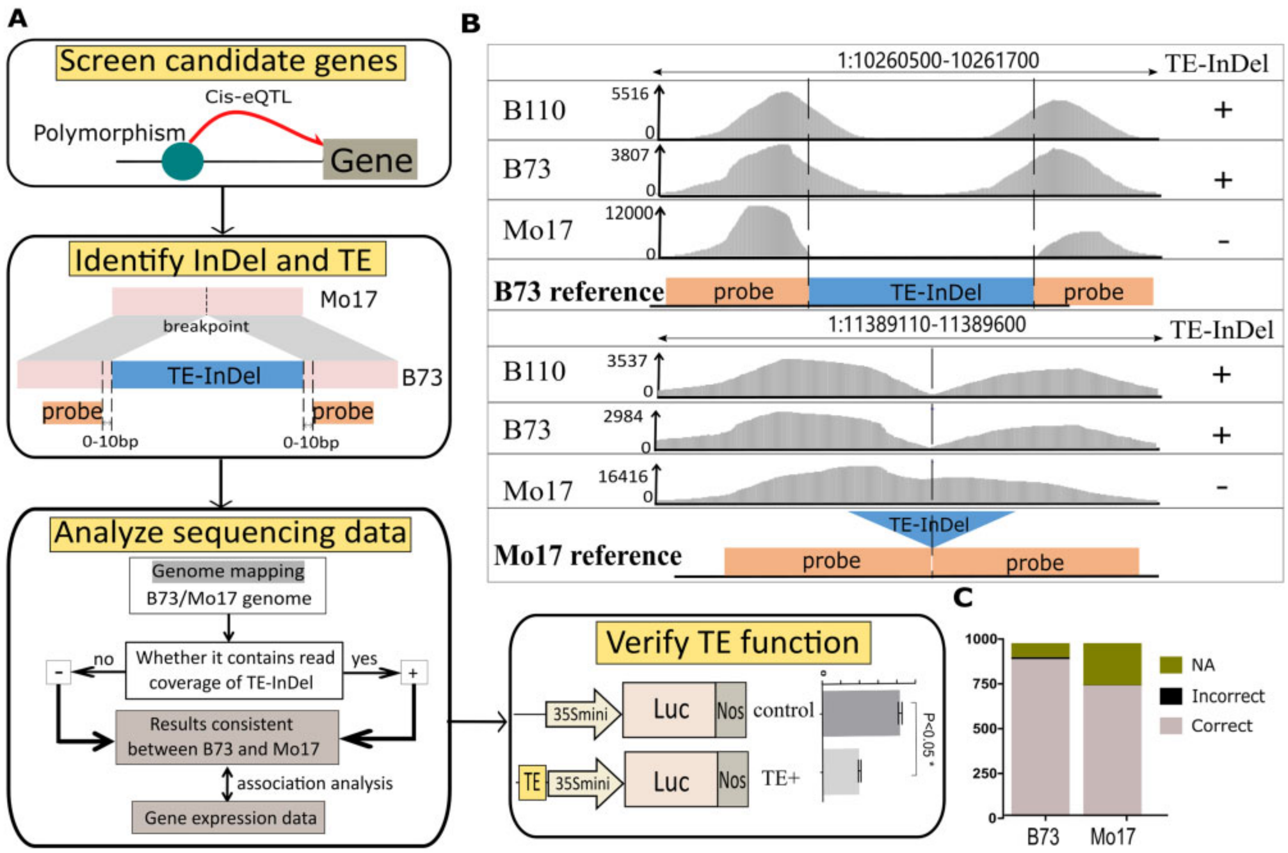

Figure 1 Robustness of the capture-based assay in transposon detection. (A) The workflow of this study. There are four major steps such as selecting candidate genes, identifying and annotating transposon insertions, characterizing transposon polymorphisms in diverse maize lines, and verifying transposon function. (B) An IGV view of alignments of three sequencing libraries to either B73 genome with an insertion (upper panel) or Mo17 genome without an insertion (lower panel). The classification result for each library is shown on the right of each panel. The vertical dashed lines show the location of the InDel junction. (C) Comparison of the insertion presence/absence classifications from the targeted assay with known presence/absence results based on the reference sequence. NA, missing data.

Table 1 Summary on alignment to maize B73 genome

Genotypes	Raw reads	Mapped reads	On-target reads	Average coverage	Regions coverage > 0
B73	12,548,382	11,341,596 (90.4%)	6,792,119 (59.9%)	506.09	965 (100.0%)
Mo17	25,147,250	9,751,882 (38.8%)	971,629 (10.0%)	70.73	848 (87.9%)
B110	13,224,482	9,151,529 (69.2%)	5,637,324 (61.6%)	418.28	955 (99.0%)
By4960	8,861,600	4,626,230 (52.2%)	1,407,021 (30.4%)	102.81	951 (98.5%)
By804	10,127,444	4,818,905 (47.6%)	2,321,220 (48.2%)	170.49	939 (97.3%)
CI7	12,947,424	6,165,547 (47.6%)	3,240,742 (52.6%)	237.64	959 (99.4%)
CIMBL105	11,426,934	4,982,199 (43.6%)	2,231,517 (44.8%)	162.81	882 (91.4%)
CIMBL114	9,311,788	3,957,129 (42.5%)	1,409,165 (35.6%)	104.44	844 (87.5%)
CIMBL145	10,481,814	7,028,876 (67.1%)	4,155,910 (59.1%)	308.96	965 (100.0%)
Dan340	34,015,224	20,126,710 (59.2%)	5,365,614 (26.7%)	399.39	962 (99.7%)
GEMS21	11,279,860	8,708,811 (77.2%)	4,991,779 (57.3%)	369.23	965 (100.0%)
HZS	10,649,782	4,780,612 (44.9%)	2,363,116 (49.4%)	177.19	940 (97.4%)
K22	8,968,980	4,559,151 (50.8%)	2,201,676 (48.3%)	163.95	926 (96.0%)
TY5	12,931,076	6,585,927 (50.9%)	2,753,636 (41.8%)	204.76	951 (98.5%)
Zheng30	16,332,382	7,440,071 (45.6%)	3,684,958 (49.5%)	271.52	958 (99.3%)
Zong31	16,613,548	7,928,667 (47.7%)	4,492,919 (56.7%)	330.22	960 (99.5%)

exhibits a pattern highly similar to B73 (Figure 1B). In this example, alignment to either B73 or Mo17 gives consistent classification. However, there are also cases where the two alignments give conflicting classifications. In these cases, simultaneous use of alignments to both references (B73 and Mo17) to retain only consistent classification eliminates these potentially incorrect calls (Supplementary Figure S4).

We proceeded to classify the InDels based on the alignment to both B73 and Mo17. A total of 887 (887/965, 91.9%) TE-InDels can be assigned for B73 capture data (Figure 1C), including 421 insertions and 466 deletions. When these classifications were compared to the expected result based on the B73 reference genome sequence, 12 are inconsistent, including 3 presence classifications and 9 absence classifications in the targeted assay. For

Mo17, 730 out of the 965 (75.6%) TE-InDels can be classified. The lower classification rate compared to B73 (730 *vs* 887) is probably related to lower depth of read that can be mapped near the target regions (Table 1; Supplementary Table S3). Almost all of the classifications (728/730) from the targeted sequencing assay are consistent with the results from the Mo17 reference genome (Figure 1C). When comparing the TE- and non-TE-InDels, the classification rate, and the consistency rate are largely similar (Supplementary Figure S4). This suggests that the targeted sequencing assay is applicable and reliable in classifying transposon-related insertions or deletions. The following analyses will focus on the TE-InDels.

Detection of transposon polymorphisms in a panel of diverse maize lines

The capture-based approach to classifying the presence/absence of insertion sequences was then applied to a diverse panel of maize inbred lines. Fourteen lines representing three maize populations were selected (Supplementary Table S3). On average, 49.1% of the reads can be mapped to the genome and an average of 41.8% of the mapped reads are within the target regions (Table 1, Supplementary Table S3). For each line, there are nearly 700 TE-InDels (72.5%) that can be classified (Figure 2A). The number of classified TE-InDels in each line is more correlated with the number of mapped reads at the target regions than with overall mapped reads (Supplementary Figure S5), suggesting some variation in target enrichment among the captures. While the number of TE-InDels that can be classified tends to be higher in the reference B73 and Mo17 genomes, there are relatively similar numbers of classifications across the broad diversity in this panel (Figure 2A and Supplementary Figure S6). Absence of the insertion is more easily detected than presence (Figure 2A), probably due to additional sequence variation that can occur near the insertion site. Nearly three-quarters (721/965) of the TE-InDels can be called in at least 10 of the 16 assayed lines (Figure 2B), suggesting that most of the TE-InDels can be classified in a diverse population with a low rate of missing data.

To assess the accuracy of the TE-InDel calls based on the targeted sequencing approach, a subset of 24 randomly selected TE-InDels were genotyped using a PCR assay. For the 384 potential data points (24 × 16), 339 and 322 can be obtained using the PCR and the sequencing assay, respectively (Figure 2C). In total, 288 data points can be assayed in both methods, with 98.6% (284/288) of the calls being consistent (Figure 2C). For the four data points that are inconsistent, all of them are presence of the insertion in the PCR results (Supplementary Table S4). This provides evidence that the capture-based approach is quite reliable.

Transposon polymorphisms are associated with variation in gene expression

The set of TE-InDels assessed in this study were identified near genes with *cis*-eQTL (Li *et al.* 2013). Thus, we evaluated whether they were correlated with gene expression levels by using expression data from a previous study (Liu *et al.* 2016). For a subset of 844 of these TE-InDels, there are at least two genotypes with the insertion and two genotypes without the insertion (Figure 3A). A comparison of the expression levels for haplotypes with and without the insertion reveals that 152 (18%) exhibit a significant expression change and these are largely (74%) in the direction consistent with the *cis*-eQTL. Among the significant changes, the presence of the transposon is often (62%) associated with higher expression for the nearby gene (Figure 3A).

To assess the role of these TE-InDels in gene expression in a larger population, a subset of 18 TE-InDels with significant effects in the panel of 16 genotypes were chosen to assay their association with gene expression in a panel of 140 inbred lines using PCR-based classification of the insertion (Figure 3B, Supplementary Table S5). After performing an association analysis controlling for population structure and individual relatedness, 10 of the 18 TE-InDels exhibit a significant association with gene expression in a consistent direction with the original eQTL results. Among these 10 significant effects there are five examples in which the TE insertion is associated with higher expression and five examples in which the presence of the TE is associated with lower expression.

To explore whether the TE-InDel itself is the causative polymorphism for the differences in gene expression, or it is mere in linkage disequilibrium with nearby causative polymorphisms, an association analysis using both the TE-InDels and the SNPs located within 5 kb on either side of the TE-InDels (Liu *et al.* 2016) was performed. The TE-InDel is the most significant marker for one out of the 10 TE-InDels that can be assayed (Supplementary

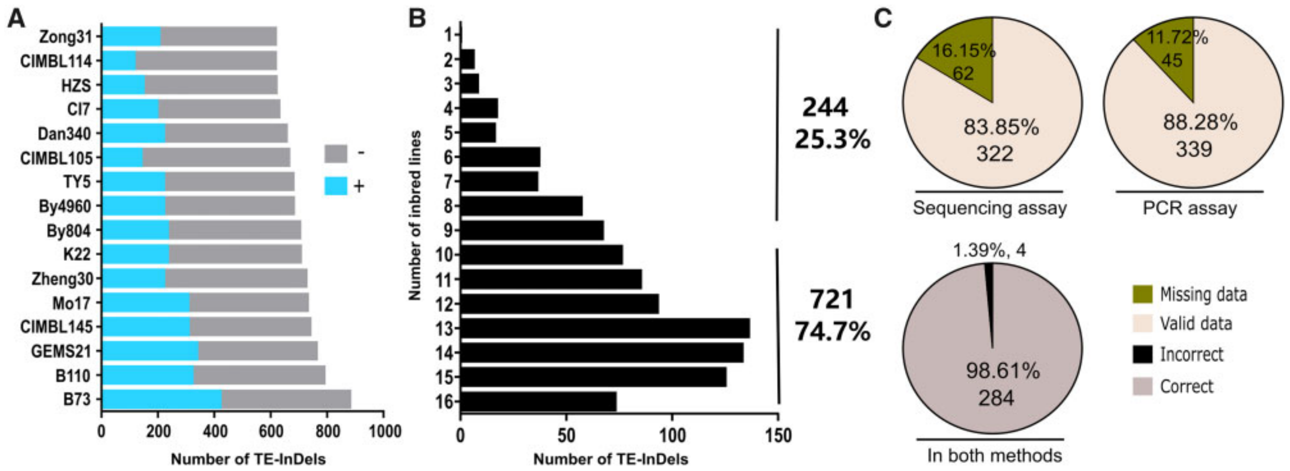

Figure 2 Detection of transposons in diverse inbred lines. (A) Number of TE-InDels that are detected in each inbred line. The "−" and "+" indicate absence and presence of insertions, respectively. (B) For each TE-InDel the number of genotypes with a presence/absence classification was determined and the distribution of the number of genotypes with classification is shown. On the right, the number of TE-InDels with classification in at least 10 or less than 10 genotypes is shown. (C) Comparison between PCR-based assay and the targeted sequencing assay for 24 randomly selected TE-InDels. The upper two charts show the proportion of genotype × TE-InDels that were classified in each assay and the lower chart shows the concordance of the calls for classifications in the two assays.

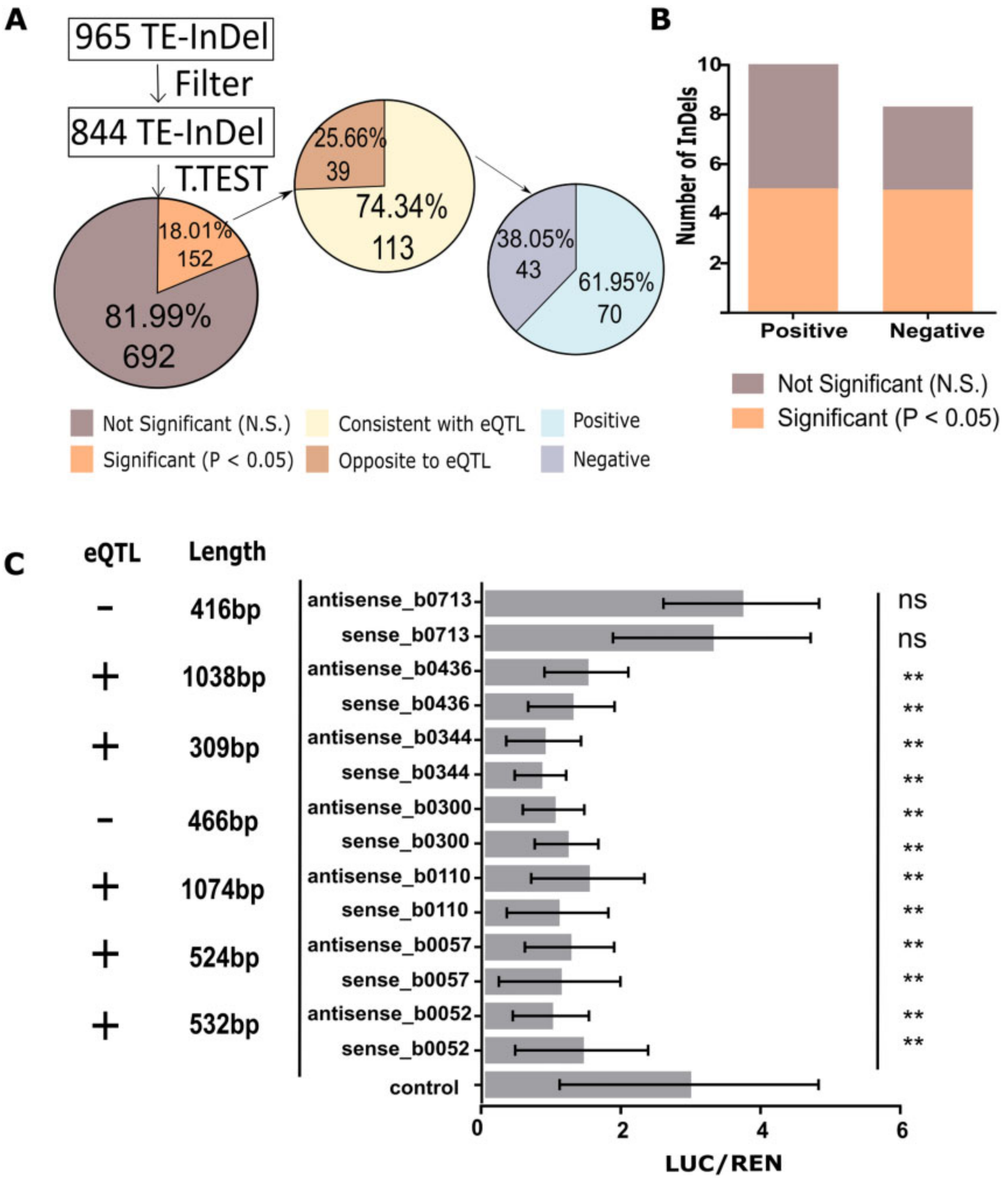

Figure 3 Associations between InDels and gene expression. (A) Summary of association of TE-InDels with gene expression. (B) The association between TE-InDels and gene expression can be detected in a large panel of maize inbred lines. (C) LUC-based validation of the association between TE-InDels and gene expression. $^{**}P < 0.01$.

Table S6, Supplementary Figure S7). There is an additional TE-InDel that is the second most significant marker (Supplementary Figure S7).

To provide evidence of the causative effect the TE-InDels on gene expression, the dual-LUC reporter assay was employed. In this assay, the expression of LUC was driven by minimal 35S promoter sequences. The expression of LUC was compared between two vectors differing only in the presence of the TE-InDel insertion sequence (Figure 1A). A difference in LUC expression would be observed if the TE-InDel insertion sequence can cause an expression difference in the protoplast cells used for this assay. Seven TE-InDels that are significant in the panel of 16 genotypes (including 5 that are significant in the panel of 140) were assessed (Supplementary Table S6). This set included 5 positive associations in which the presence of the TE is associated with higher expression of the nearby gene and two negative associations. While six of the seven TEs tested are associated with a significant change in LUC expression, only one (b0300) exhibits significant effects with the expected direction of function (Figure 3C). This association can be verified when the TE is present in either sense or antisense orientation relative to the LUC gene. The observation that only negative association could be verified could be caused by various reasons. One possibility is the absence of transcription factors or chromatin context that is necessary for the functional impact of the transposon insertion. Together these results suggest a diverse role of transposon in gene expression.

Features of transposons that are significantly associated with gene expression

The insertion sequences for the 113 TE-InDels with significant associations in a consistent direction in this study and the eQTL study (Figure 3A) were further assessed for any common trends (Supplementary Figure S8). In general, larger TE-InDels are slightly more likely to have significant associations with gene expression (Supplementary Figure S8A). TE-InDels that are located closer to transcriptional start sites are more likely to be associated with gene expression (Supplementary Figure S8B). In fact, when we only consider TE-InDels that are located within 2 kb upstream of the gene, we find a slight increase in the percentage of TE-InDels significantly associated with gene expression (Supplementary Figure S8C). We find examples of both DNA transposons and retrotransposons with significant associations with gene expression in both positive and negative directions (Supplementary Figure S8D). The analysis of super-families of transposons finds the Gypsy retrotranpsosons and the Helitron and hAT DNA transposons have a slightly higher probability with gene expression than other types of transposons (Supplementary Figure S8D).

The chromatin characteristics of the insertion sequences for TE-InDels were examined using previously generated chromatin accessibility (Ricci *et al.* 2019) or DNA methylation (Liang *et al.* 2021) data. We first characterized the presence of accessible chromatin regions (ACRs) within TE-InDel insertion sequences or at insertion sites. A subset of the B73 TE-InDel insertions (14/64) contains an ACR in B73 seedling leaf tissue, which may suggest regions of these insertions are playing a role in influencing expression of the nearby gene. ACRs within Mo17 TE-InDel insertion sequences could not be assessed as we did not have chromatin accessibility data for Mo17. However, we could assess whether the TE-InDel insertion in Mo17 inserted within a region that is accessible in the B73 insertion site. We found that 9 of the 49 Mo17 TE-InDel insertions with significant associations with gene expression are within B73 ACRs. We proceeded to assess the presence of unmethylated regions (UMRs) within the B73 or Mo17 TE-InDel insertion sequences. There are many examples of TE-InDel sequences in both B73 and Mo17 that contain a UMR with generally similar frequencies for TE-InDels with and without significant associations with gene expression (Supplementary Figure S9). Although there is no enrichment for UMRs within TE-InDels significant for eQTL association, the frequency of these TE-InDels with UMRs is higher than the genome-wide frequency of TEs containing UMRs. It is possible that other chromatin features around the TEs may have an effect on the expression of nearby genes, and suggests the presence of a complex interplay between TEs and chromatin environments in gene regulation.

Discussion

In this study, we developed a capture-based assay for genotyping the presence/absence of TE-InDels and showed that this assay can be employed to determine whether insertion is present or absent across diverse genotypes. This approach is efficient for classifying TE-InDels with varying sizes and allows the detection of many TE-InDels at population level with high reliability and low missing rate. In theory, this assay can be applied to any loci for which probes can be designed. One situation that can complicate the use of this assay is the fact that the size and sequence of a specific insertion can vary across different genotypes. This will lead to lower mapping rate and missing classifications especially for the TE-InDels where abundant sequence variations around the insertion boundary are present. This problem could be minimized by aligning the sequencing reads to more reference genomes which are becoming available in maize (Jiao *et al.* 2017; Springer *et al.* 2018; Sun *et al.* 2018; Yang *et al.* 2019; Haberer *et al.* 2020).

In combination with other-omics data, such as transcriptome, this assay allows interrogation of transposon function in a high-throughput manner. We found that 18% of the tested TE-InDels are associated with variation in gene expression levels using just a single tissue for monitoring gene expression levels. Targeted analysis for several of the TE-InDels in a large panel suggests that in some cases the insertion sequence has the most power than nearby SNPs to detect expression variation. This suggests that a subset of TE insertions is not well tagged by nearby SNPs and that high-quality presence/absence calls for these insertions could enable detection of novel significant associations with gene expression (or other traits) compared to previous SNP-based analyses (Fuentes *et al.* 2019; Akakpo *et al.* 2020). This is consistent with previous findings in tomato which shows that only 1 out of the 31 significant associations between transposon insertion polymorphisms and traits could be identified by SNPs (Domínguez *et al.* 2020). Moreover, the effect size of the transposon polymorphisms is generally much larger than that of SNPs, suggesting an important role of transposons in phenotypic diversity.

Our findings that TE insertions can be associated with changes in expression of nearby genes agree with previous reports that transposons can create allelic variations in gene expression (Mao *et al.* 2015; Wang *et al.* 2016). It is known that some transposons can only affect gene expression under certain environmental conditions. For example, the DNA transposon located in the promoter region of *ZmNAC111* is not associated with gene expression under normal conditions, and is associated with gene expression under drought treatment (Mao *et al.* 2015). Thus, it is possible that more TE-InDels can be associated with gene expression if expression data from more tissues or growth conditions were included for this analysis. In our analysis, we also found that TE insertions exhibit a mixture of positive and negative influences on the expression of nearby genes. This is consistent with previous findings (Domínguez *et al.* 2020). TEs may disrupt existing *cis*-regulatory elements or provide novel *cis*-regulatory elements and these can have diverse influences on the expression level for nearby genes. Together with our results suggest a complex interplay between the presence of TEs and nearby gene expression.

Funding

This work was supported by the National Natural Science Foundation of China (No: 31871635) and the Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (Program No.2016RC012).

Author contributions

Q.L. designed this study. M.L. and X.D. conducted the experiments and data analysis. J.M.N. contributed data analyses. Q.L., N.M.S., and M.L. wrote and revised the manuscript.

Conflicts of interest

None declared.

Literature cited

Akakpo R, Carpentier MC, Ie Hsing Y, Panaud O. 2020. The impact of transposable elements on the structure, evolution and function of the rice genome. New Phytol. 226:44–49.

Anderson SN, Stitzer MC, Brohammer AB, Zhou P, Noshay JM, *et al.* 2019. Transposable elements contribute to dynamic genome content in maize. Plant J. 100:1052–1065.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, *et al.* 2011. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 479:534–537.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, *et al.* 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 23:2633–2635.

Cock PJ, Chilton JM, Grüning B, Johnson JE, Soranzo N. 2015. NCBI BLAST+ integrated into Galaxy. Gigascience. 4:39.

Domínguez M, Dugas E, Benchouaia M, Leduque B, Jiménez-Gómez JM, *et al.* 2020. The impact of transposable elements on tomato diversity. Nat Commun. 11:4058.

Fedoroff NV. 2012. McClintock's challenge in the 21st century. Proc Natl Acad Sci USA. 109:20200–20203.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 9:397–405.

Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, *et al.* 2019. Structural variants in 3000 rice genomes. Genome Res. 29:870–880.

Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, *et al.* 2020. European maize genomes highlight intraspecies variation in repeat and gene content. Nat Genet. 52:950–957.

Huang C, Sun H, Xu D, Chen Q, Liang Y, *et al.* 2018. ZmCCT9 enhances maize adaptation to higher latitudes. Proc Natl Acad Sci USA. 115:E334–E341.

Jia H, Li M, Li W, Liu L, Jian Y, *et al.* 2020. A serine/threonine protein kinase encoding gene KERNEL NUMBER PER ROW6 regulates maize grain yield. Nat Commun. 11:988.

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, *et al.* 2017. Improved maize reference genome with single-molecule technologies. Nature. 546:524–527.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, *et al.* 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36.

Lee TH, Guo H, Wang X, Kim C, Paterson AH. 2014. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics. 15:162.

Li L, Petsch K, Shimizu R, Liu S, Xu WW, *et al.* 2013. Mendelian and non-Mendelian regulation of gene expression in maize. PLoS Genet. 9:e1003202.

Liang Z, Anderson SN, Noshay JM, Crisp PA, Enders TA, *et al.* 2021. Genetic and epigenetic contributions to variation in transposable element expression responses to abiotic stress in maize. Plant Physiol. doi: 10.1093/plphys/kiab07.

Lisch D. 2013. How important are transposons for plant evolution? Nat Rev Genet. 14:49–61.

Liu H, Wang F, Xiao Y, Tian Z, Wen W, *et al.* 2016. MODEM: multi-omics data envelopment and mining in maize. Database (Oxford). 2016:baw117.

Liu H, Luo X, Niu L, Xiao Y, Chen L, *et al.* 2017. Distant eQTLs and non--coding sequences play critical roles in regulating gene expression and quantitative trait variation in maize. Mol Plant. 10:414–426.

Liu L, Du Y, Shen X, Li M, Sun W, *et al.* 2015. KRN4 controls quantitative variation in maize kernel row number. PLoS Genet. 11: e1005670.

Mao H, Wang H, Liu S, Li Z, Yang X, *et al.* 2015. A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. Nat Commun. 6:8326.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, *et al.* 2018. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 14:e1005944.

Noshay JM, Anderson SN, Zhou P, Ji L, Ricci W, *et al.* 2019. Monitoring the interplay between transposable element families and DNA methylation in maize. PLoS Genet. 15:e1008291.

Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, *et al.* 2016. The *Arabidopsis thaliana* mobilome and its impact at the species level. Elife. 5:e15716.

Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui MA, *et al.* 2019. Transposition favors the generation of large effect mutations that may facilitate rapid adaption. Nat Commun. 10:3421.

Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, *et al.* 2019. Widespread long-range cis-regulatory elements in the maize genome. Nat Plants. 5:1237–1249.

Salvi S, Sponza G, Morgante M, Tomes D, Niu X, *et al.* 2007. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci USA. 104:11376–11381.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. Science. 326:1112–1115.

Sun S, Zhou Y, Chen J, Shi J, Zhao H, *et al.* 2018. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet. 50:1289–1295.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 8:272–285.

Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, *et al.* 2018. The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat Genet. 50:1282–1288.

Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, *et al.* 2016. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. Elife. 5:e20777.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J. 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. Nat Genet. 43:1160–1163.

Wang X, Wang H, Liu S, Ferjani A, Li J, *et al.* 2016. Genetic variation in ZmVPP1 contributes to drought tolerance in maize seedlings. Nat Genet. 48:1233–1241.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, *et al.* 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.

Yang N, Liu J, Gao Q, Gui S, Chen L, *et al.* 2019. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet. 51:1052–1059.

Yang Q, Li Z, Li W, Ku L, Wang C, *et al.* 2013. CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proc Natl Acad Sci USA. 110:16969–16974.

Yang X, Gao S, Xu S, Zhang Z, Boddupalli M, *et al.* 2011. Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. Mol Breeding. 28:511–526.

Yoo SD, Cho YH, Sheen J. 2007. Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nat Protoc. 2:1565–1572.

Communicating editor: S. Pearce