**Supplemental Material for A capture-based assay for high-throughput detection and characterization of transposons in maize**

**Figure S1 Comparison between the TE-capture developed by Quadrana et al. (2016, 2019, left panel) and in this study (right panel).** The differences include: (1) the position of probes relative to TE: the probes were on the TE in the study by Quadrana et al. (2016, 2019) and were on the flanking regions in this study; (2) the purpose by Quadrana et al. (2016, 2019) is to detect new insertions of the interested TE in unknown positions, and the purpose of this study is to detect the presence/absence of a known TE across multiple genotypes.

**Figure S2 A schematic to illustrate the strategy to classify the presence or absence of InDels.** LB and RB indicate the position of left and right boundary of the InDel, respectively. Read\_LB and Read\_RB indicate the left and right mapping positions of the sequencing reads. Yes, satisfy the judgment conditions; No, does not satisfy the judgment conditions; +, presence of insertion; -, absence of insertion; FALSE, missing data.

**Figure S3 Selection of the TE-InDels.** (A) The distribution of number of TE-InDels in the gene promoters is shown. Most genes have a single TE-InDel but several have 2 or more TE-InDels. (B) For each insertion sequence classified as a TE-InDel we determined the proportion of the insertion sequence that is annotated as transposon-related. The distribution of the proportion of the insertion sequence that is TE sequence is shown.

**Figure S4 Comparison of TE-InDel calls between targeted sequencing assay and the reference genome.** (A) The sequence data from the B73 sequence captures was aligned to the B73, Mo17 or both references (BM) and each TE-InDel was classified as NA (missing data or low coverage), correct (based on expectations from the reference genome) or incorrect, - (absence of insertions) or + (presence of insertions). (B) A similar analysis was performed for the Mo17 capture data. (C) Same as (A) but for non-TE-InDels in B73 capture data. (D) Same as (B) but for non-TE-InDels in Mo17 capture data.

**Figure S5 The relationship between number of TE-InDel classifications and mapped reads.** For each of the 16 genotypes, we determined the total number of mapped reads (A) or the number of reads that are mapped to the target regions (B) for alignments to the B73 reference (blue) or the Mo17 reference (orange). We assessed the correlation between the number of reads and the number of TE-InDels that could be classified in each genotype.

**Figure S6 Genetic relatedness of the 16 inbred lines.** The bar plot on the right indicates the number of TE-InDels that are identified in each line. The - and + indicates the presence or absence of TE-InDels, respectively.

**Figure S7 Manhattan plot to show the association between gene expression and TE-InDel and its nearby SNPs**

**Figure S8 Features of TE-InDels with significant association with gene expression.** (A) The proportion of TE-InDel that is associated with gene expression for different size classes. (B) The proportion of significant associations for TE-InDels with varying distance from the transcriptional start site. (C) The proportion of TE-InDels that are associated with gene expression for genes where the TE-InDel is located within 2kb/5kb upstream of the gene. (D) The proportion of significant associations for different types of transposons. Total represents number of TE-InDels that have data in at least 2 lines with insertion and 2 lines without insertion.

**Figure S9 Chromatin characteristics of TE-InDels**. Percent of TE-InDels with an ACR (A) or UMR (B) in the sequence coordinates in the corresponding genome for those TE-InDels defined as having a significant eQTL association in a consistent direction (significant), not significant (ns), and the overall genome-wide frequency (genome).

**Table S1 Primers used in this study**

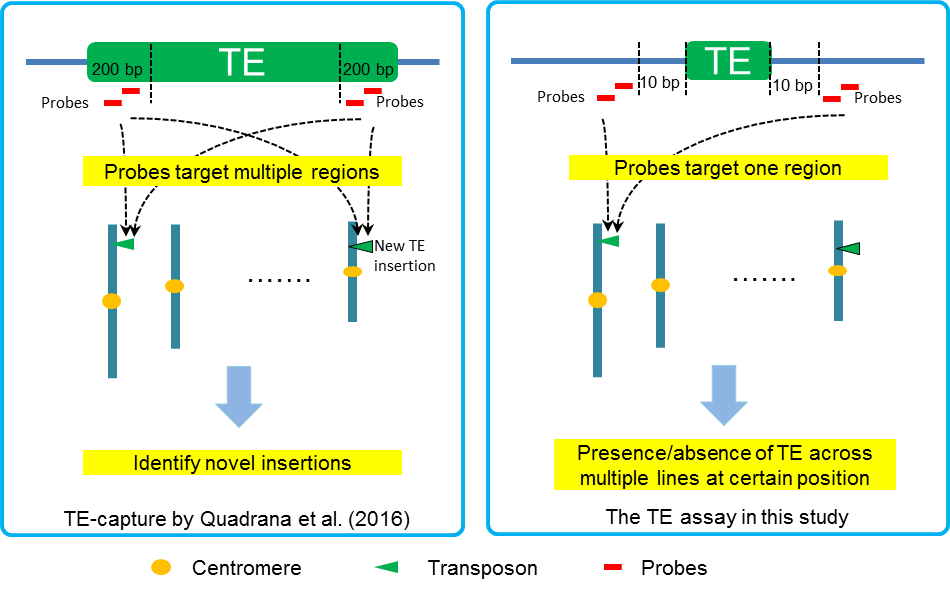
**Table S2 Information of InDels in this study**

**Table S3 Summary on alignment to maize Mo17 genome**

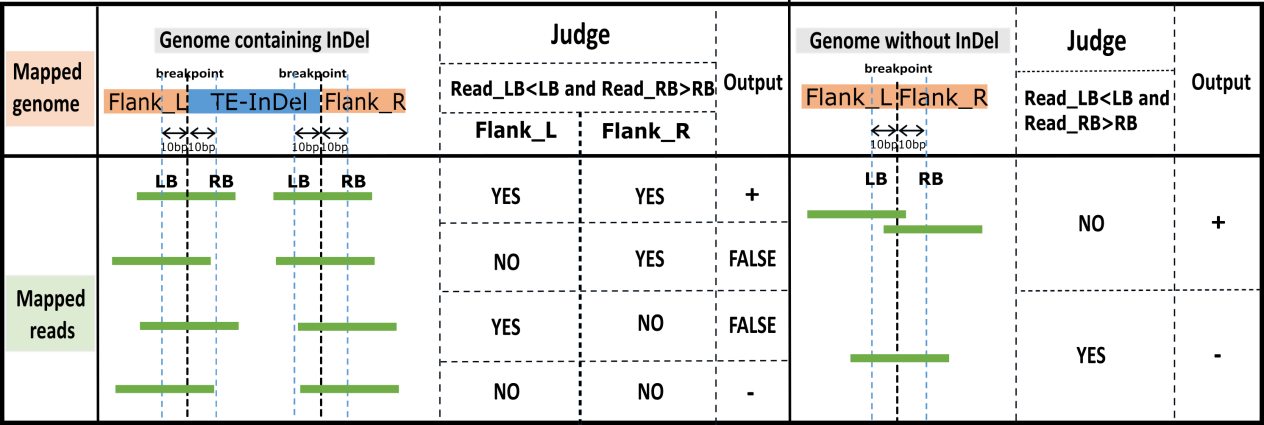
**Table S4 InDel calls in a panel of 16 lines**

**Table S5 PCR-based genotyping of a set of InDels in more lines**

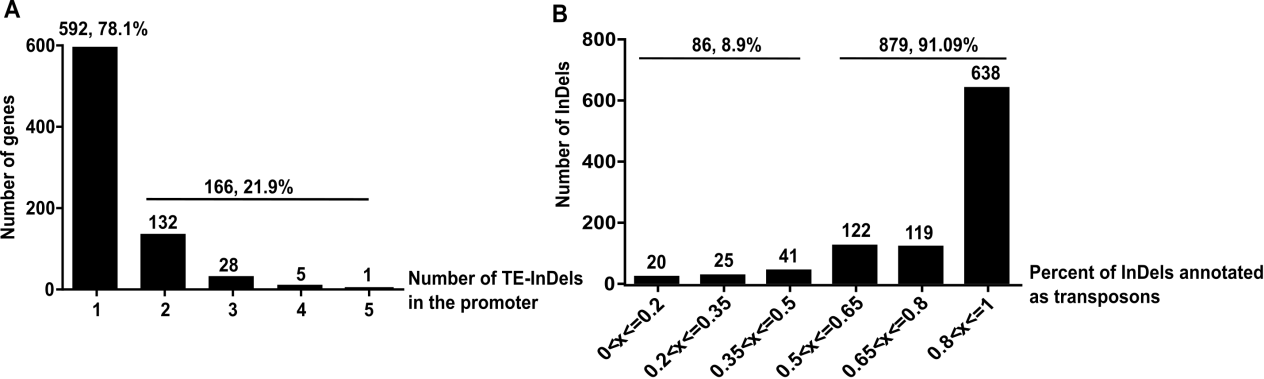
**Table S6 Association analysis for TE-InDels and the SNPs nearby**

****

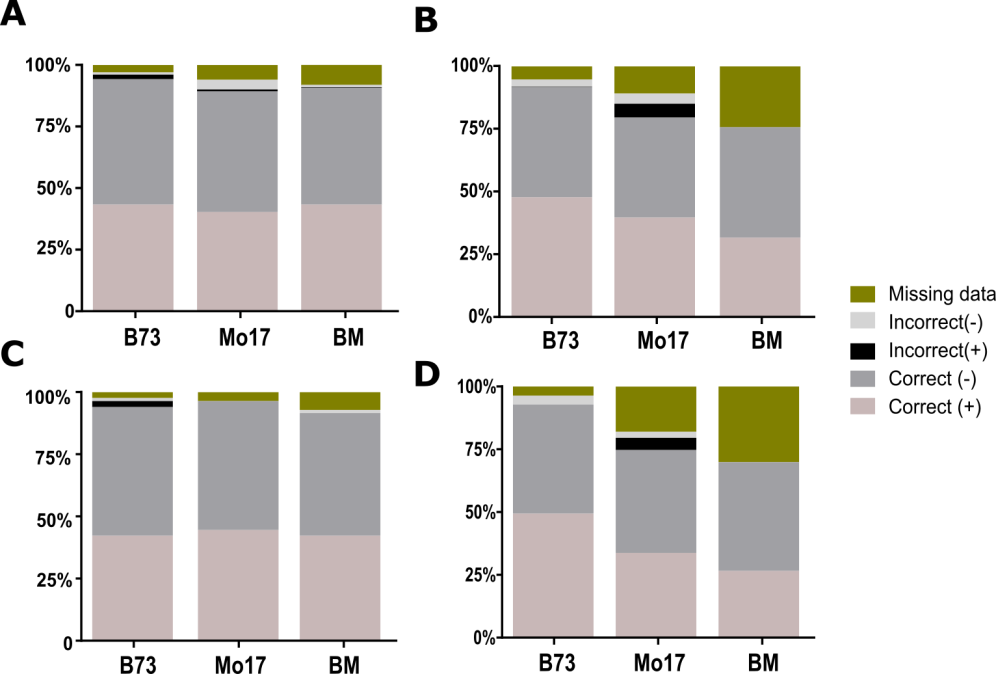
**Figure S1 Comparison between the TE-capture developed by Quadrana et al. (2016, 2019, left panel) and in this study (right panel).** The differences include: (1) the position of probes relative to TE: the probes were on the TE in the study by Quadrana et al. (2016, 2019) and were on the flanking regions in this study; (2) the purpose by Quadrana et al. (2016, 2019) is to detect new insertions of the interested TE in unknown positions, and the purpose of this study is to detect the presence/absence of a known TE across multiple genotypes.



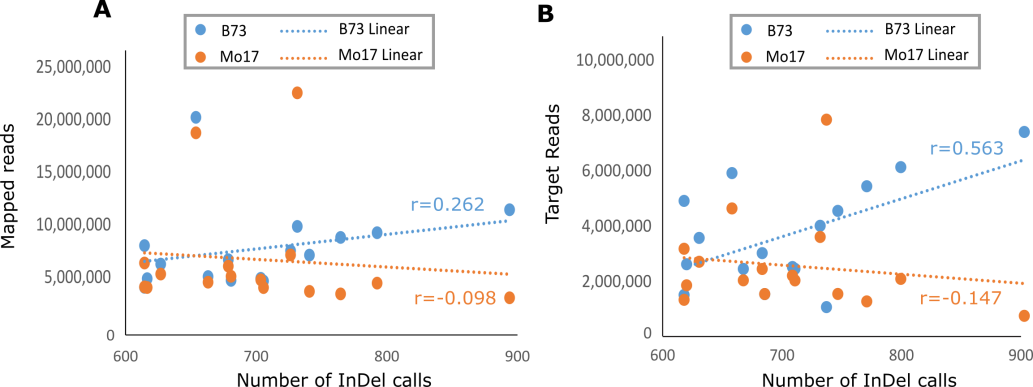
**Figure S2 A schematic to illustrate the strategy to classify the presence or absence of InDels.** LB and RB indicate the position of left and right boundary of the InDel, respectively. Read\_LB and Read\_RB indicate the left and right mapping positions of the sequencing reads. Yes, satisfy the judgment conditions; No, does not satisfy the judgment conditions; +, presence of insertion; -, absence of insertion; FALSE, missing data.

****

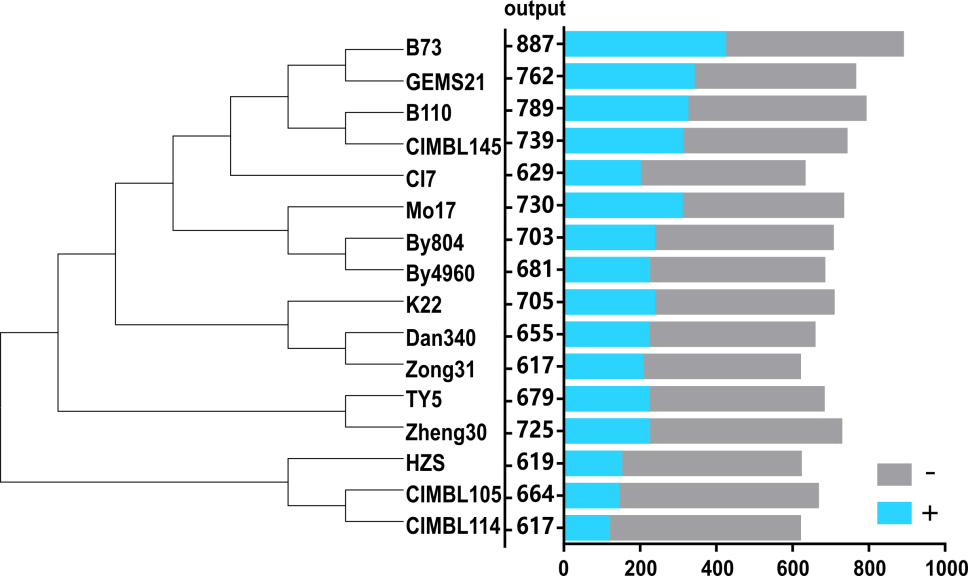
**Figure S3 Selection of the TE-InDels.** (A) The distribution of number of TE-InDels in the gene promoters is shown. Most genes have a single TE-InDel but several have 2 or more TE-InDels. (B) For each insertion sequence classified as a TE-InDel we determined the proportion of the insertion sequence that is annotated as transposon-related. The distribution of the proportion of the insertion sequence that is TE sequence is shown.

****

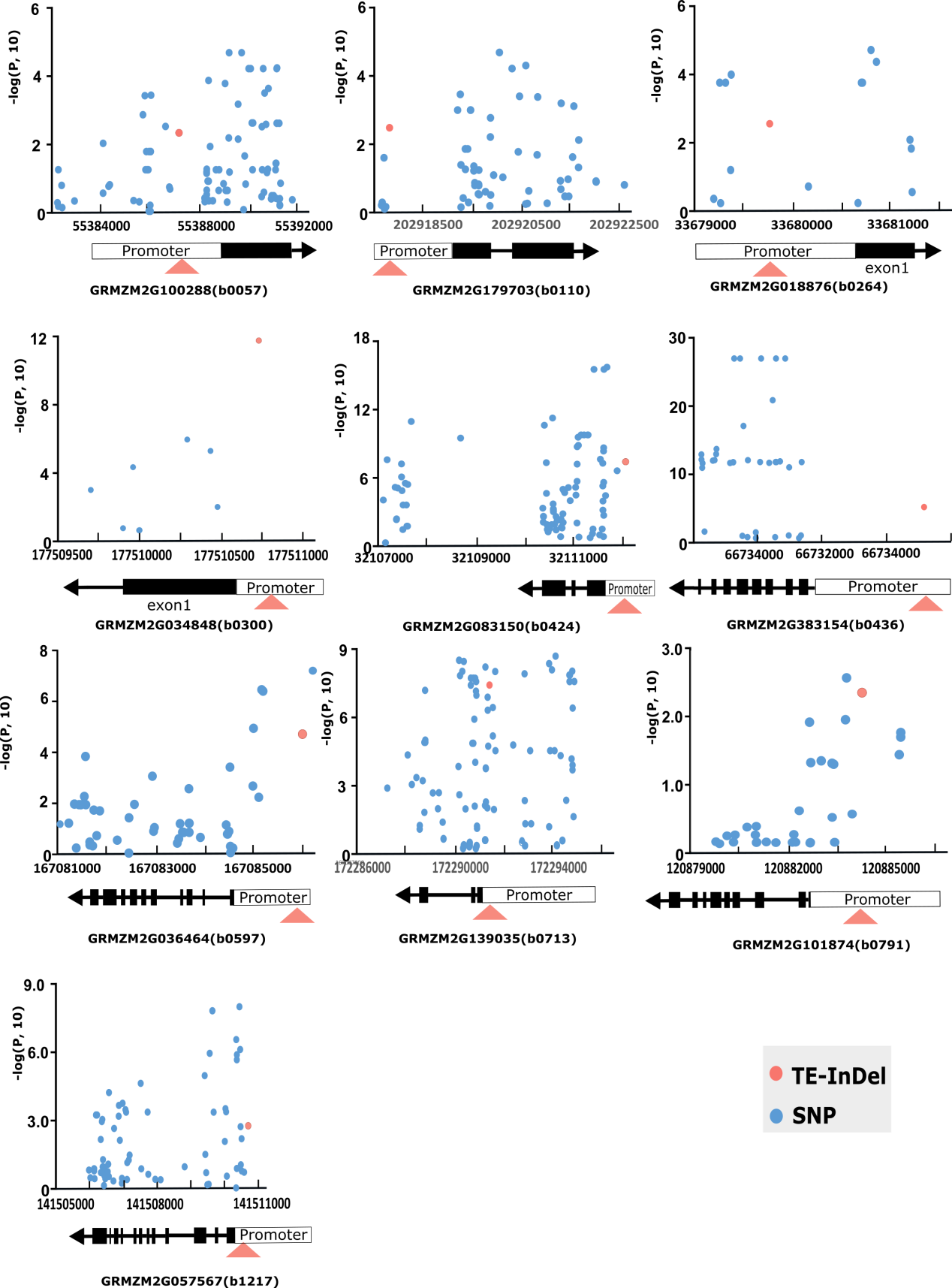
**Figure S4 Comparison of TE-InDel calls between targeted sequencing assay and the reference genome.** (A) The sequence data from the B73 sequence captures was aligned to the B73, Mo17 or both references (BM) and each TE-InDel was classified as NA (missing data or low coverage), correct (based on expectations from the reference genome) or incorrect, - (absence of insertions) or + (presence of insertions). (B) A similar analysis was performed for the Mo17 capture data. (C) Same as (A) but for non-TE-InDels in B73 capture data. (D) Same as (B) but for non-TE-InDels in Mo17 capture data.

****

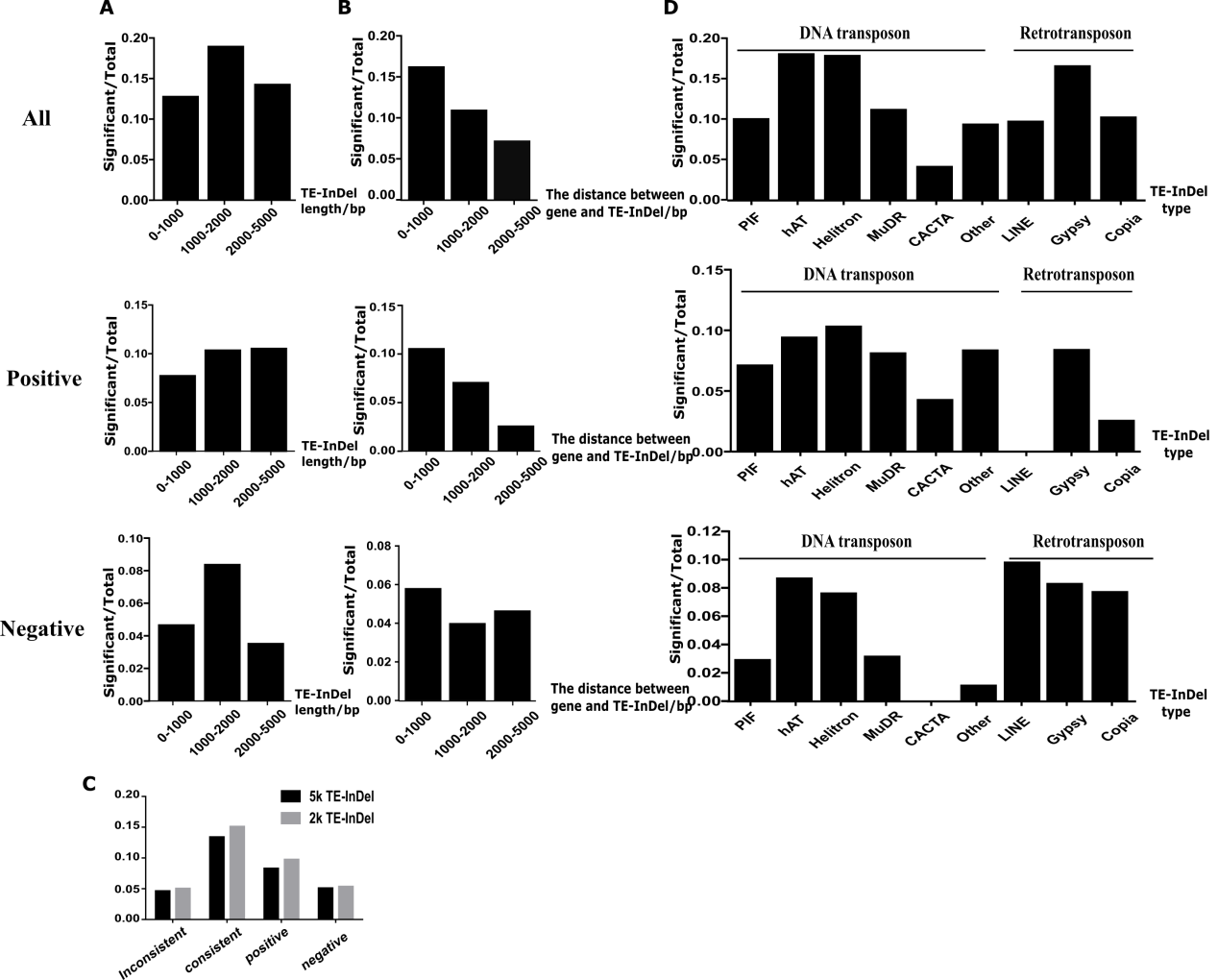
**Figure S5 The relationship between number of TE-InDel classifications and mapped reads.** For each of the 16 genotypes, we determined the total number of mapped reads (A) or the number of reads that are mapped to the target regions (B) for alignments to the B73 reference (blue) or the Mo17 reference (orange). We assessed the correlation between the number of reads and the number of TE-InDels that could be classified in each genotype.

****

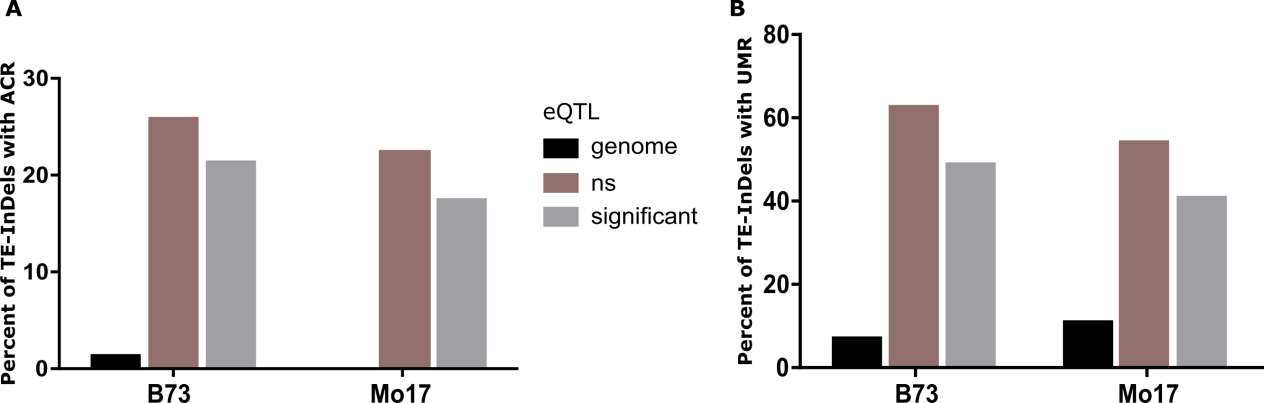
**Figure S6 Genetic relatedness of the 16 inbred lines.** The bar plot on the right indicates the number of TE-InDels that are identified in each line. The – and + indicates the presence or absence of TE-InDels, respectively.

****

**Figure S7 Manhattan plot to show the association between gene expression and TE-InDel and its nearby SNPs**

****

**Figure S8 Features of TE-InDels with significant association with gene expression.** (A) The proportion of TE-InDel that is associated with gene expression for different size classes. (B) The proportion of significant associations for TE-InDels with varying distance from the transcriptional start site. (C) The proportion of TE-InDels that are associated with gene expression for genes where the TE-InDel is located within 2kb/5kb upstream of the gene. (D) The proportion of significant associations for different types of transposons. Total represents number of TE-InDels that have data in at least 2 lines with insertion and 2 lines without insertion.



**Figure S9 Chromatin characteristics of TE-InDels**. Percent of TE-InDels with an ACR (A) or UMR (B) in the sequence coordinates in the corresponding genome for those TE-InDels defined as having a significant eQTL association in a consistent direction (significant), not significant (ns), and the overall genome-wide frequency (genome).