Title: Population Level Variation of Transposable Elements in a Maize Diversity Panel

Author Names:
Christine H. O'Connor[*,†], Yinjie Qiu[*], Rafael Della Coletta[*], Patrick J. Monnahan[*,†], Jaclyn M. Noshay[‡], Amanda Gilbert[*], Sarah N. Anderson[§], Suzanne E. McGaugh[†], Nathan M. Springer[‡], Candice N. Hirsch[*]

Institutional Affiliations:
[*]Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108
[†]Department of Ecology, Evolution and Behavior, University of Minnesota, St. Paul, MN 55108
[‡]Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA 50011
[§]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN 55108

Corresponding Author Information:
Candice Hirsch
1991 Upper Buford Circle
St Paul, MN 55108
612-301-9522
cnhirsch@umn.edu

ABSTRACT

Intact transposable elements (TEs) account for 65% of the maize genome and can impact gene function and regulation. Although TEs comprise the majority of the maize genome and affect important phenotypes, genome wide patterns of TE polymorphisms in maize have only been studied in a handful of maize genotypes, due in part to the challenging nature of assessing highly repetitive sequences. We implemented a method to use short read sequencing data from 509 diverse inbred lines to classify the presence/absence of 494,564 non-redundant TEs that were previously annotated in four genome assemblies including B73, Mo17, PH207, and W22. Different orders of TEs (i.e. LTRs, Helitrons, TIRs) had different frequency distributions within the population. Across the different orders, TE family size was negatively correlated with average population frequency of TEs in the family and nested TEs are at lower frequency than non-nested TEs. Age of LTR elements was positively correlated with population frequency. Comparison with SNP data revealed that while a majority of TEs are tagged by nearby SNPs ($r^2 > 0.9$) there are also many TEs in low to moderate linkage disequilibrium with nearby SNPs. This study provides a population scale genome-wide assessment of TE variation in maize, and provides valuable insight on variation in TEs in maize and factors that contribute to this variation.

INTRODUCTION

Transposable elements (TEs) are present in all eukaryotic genomes (BENNETZEN 2000; WICKER *et al.* 2007). In maize, 65% of the genome is made up of intact TEs (JIAO *et al.* 2017), and another 20% is comprised of fragmented TEs (SCHNABLE *et al.* 2009). There are many examples of phenotypic effects of TEs from null mutations, such as maize kernel color (SELINGER AND CHANDLER 2001), white wine grapes (CADLE-DAVIDSON AND OWENS 2008), and color variation in the common morning glory (CLEGG AND DURBIN 2000). TE insertions can also affect positive and negative gene regulatory functions, such as insertion of a long terminal repeat LTR retrotransposon into the promoter region of the Ruby gene in oranges that leads to red fruit flesh of blood oranges (BUTELLI *et al.* 2012), and an LTR retrotransposon that is associated with red skin color in apples (ZHANG *et al.* 2019). TE insertions in maize have also been associated with genes that are upregulated in response to abiotic stress (MAKAREVITCH *et al.* 2015).

TEs are classified into two classes, depending on how they replicate, and from there into superfamilies and families, by sequence similarity (WICKER *et al.* 2007). Class I elements, or retrotransposons, replicate via an RNA intermediate (BENNETZEN 2000; LISCH 2013). Long terminal repeat retrotransposons (LTR) are the most abundant type of retrotransposons in maize (BENNETZEN 2000) and account for over half of the maize genome by sequence length (ANDERSON *et al.* 2019; STITZER *et al.* 2019). Class II elements, or DNA transposable elements, replicate via a DNA intermediate, and the two largest orders are Helitrons and terminal inverted repeat (TIR) elements. TIRs are defined by terminal inverted repeat sequences at both ends of the TE (WICKER *et al.* 2007) and intact TIRs make up around 3% of the maize genome (ANDERSON *et al.* 2019). Helitrons are defined by their 'rolling circle' replication mechanism (LISCH 2013) and intact

Helitrons make up around 4% of the maize genome (ANDERSON *et al.* 2019). TEs are found throughout the maize genome, including near and in genes, and can be anywhere from a few hundred base pairs to >10 kb in length (BENNETZEN 2000), and range in age, from very recent insertions to >2 million years old insertions (STITZER *et al.* 2019).

Studies done on TEs in maize have shown that there is extensive variation in TE insertion presence/absence patterns at specific loci across maize inbreed lines (SANMIGUEL *et al.* 1996; FU AND DOONER 2002; MORGANTE *et al.* 2005; DOONER *et al.* 2019). Early work on the *bronze* locus in multiple maize lines found that different lines differed in not only the gene order and content but also in TE content (FU AND DOONER 2002). More recent work on mutations in the same region found that not only were high mutation rates due to TE insertions, but that different TEs were inserting in different maize lines (DOONER *et al.* 2019). Whole genome analysis of four maize genomes with *de novo* TE annotations revealed extensive TE polymorphism between maize lines on a whole genome scale (ANDERSON *et al.* 2019). On average, about 500 Mb of TE sequence, or ~20% of the maize genome, was variable between the four inbred lines included in the study: B73, Mo17, PH207 and W22. Another 1.6 Gb of TE sequence was only shared between two or three of the lines.

Despite the widespread prevalence of TEs in the maize genome, connection of TE insertions to functional phenotypic variation, and evidence of widespread TE polymorphism between maize genotypes, there have been very few scans of specific TE insertion frequencies in divergent maize populations. The analysis of the frequency of TE insertions can provide insights into the level of variability for TEs and help understand the presence of common and rare TE variants. To understand patterns of TE polymorphism on a genome-wide scale, we utilized short

read sequencing of 509 diverse maize lines to score TE presence/absence of 494,564 non-redundant TEs that were annotated in four reference genome assemblies. This study provides a genome-wide analysis of TE polymorphism frequencies across a large number of genotypes as we continue to try to understand how TEs contribute to phenotypic variation and adaptation within species.

MATERIALS AND METHODS

Whole genome resequencing: A subset of 509 lines from the Wisconsin Diversity Panel (HANSEY *et al.* 2010; HIRSCH *et al.* 2014; MAZAHERI *et al.* 2019) was used for this study (Table S1). For 57 genotypes, available short-read sequence data was downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA; Table S1). These samples ranged in theoretical coverage of 10-55x sequencing depth. For 454 genotypes, plants were grown under greenhouse conditions (27C/24C day/night and 16 h light/8 h dark) with five plants of a single genotype per pot for DNA extractions. Plants were grown in Metro-Mix 300 (Sun Gro Horticulture) with no additional fertilizer. Tissue from the newest leaf of each seedling in the pot was collected and immediately flash frozen in liquid nitrogen at the Vegetative 2 developmental stage. DNA was extracted using a standard cetyltrimethyl ammonium bromide (CTAB) DNA extraction protocol (SAGHAI-MAROOF *et al.* 1984), and treated with 25 uL of RNase at 39°C for 30 minutes. Genomic DNA for each genotype was submitted to Novogene (Novogene Co., Ltd., Beijing, China) for whole genome sequencing with 150 base pair paired end reads generated on a HiSeq X Ten sequencing machine. For each genotype at least 20x theoretical sequencing depth was achieved.

Read Alignment and processing: Quality control analysis of the sequence data was conducted using using fastqc version 0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapter sequence and low-quality base trimming was done using cutadapt version 1.18 (MARTIN 2011) and sickle version 1.33 (https://github.com/najoshi/sickle) both with default parameters. Sequence reads were aligned to the B73 v4 (JIAO *et al.* 2017), Mo17 v1 (SUN *et al.* 2018), PH207 v1 (HIRSCH *et al.* 2016), and W22 v1 (SPRINGER *et al.* 2018) genome assemblies. Alignment of reads was conducted using SpeedSeq version 0.1.2 (CHIANG *et al.* 2015), which efficiently parallelizes bwa mem (LI AND DURBIN 2009), with read groups labelled separately for each FASTQ. Alignments were subsequently filtered to require a minimum mapping quality of 20 using samtools view (LI *et al.* 2009). To assess sample integrity, single nucleotide polymorphisms (SNPs) were called relative to the B73 v4 genome assembly using Platypus v 0.8.1 (RIMMER *et al.* 2014) with default parameters. A subset of 50,000 random bi-allelic SNPs with less than 50% missing data and less than 10% heterozygosity was used to calculate Nei's pairwise genetic distances using StAMPP version 1.5.1 and a neighbor joining tree was generated with nj() in ape version 5.3 all using R version 3.6.2 (R DEVELOPMENT CORE TEAM 2020). Three samples with discordance to known pedigree relationships were removed from downstream analysis to result in the 509 genotypes included in Table S1.

Transposable element identification in the Wisconsin Diversity Panel: Transposable element (TE) annotations for the B73, Mo17, PH207, and W22 genome assemblies were obtained from https://mcstitzer.github.io/maize_TEs. Mean coverage over the genome windows adjacent to

each of the TE boundaries was calculated using multicov within bedtools version 2.29.2 (QUINLAN AND HALL 2010). The windows ranged in size from 1 to 10 bp, and included the internal and external sequences that flanked the start and stop boundaries of each of the annotated TEs for four windows per TE (Figure S1). Coverage down sampling to 20%, 32%, 45%, 66% and 83% of their original coverage levels was performed using samtools view (LI *et al.* 2009). A non-redundant set of TEs with present/absent scores based on whole-genome comparisons of B73, Mo17, PH207 and W22 was used as previously reported (ANDERSON *et al.* 2019) for analysis of the true positive, true absent, false positive, and false absent rates of the short read based present/absent calls.

After assessing accuracy of different TE calling parameters the following criteria were chosen to balance accuracy and the number of ambiguous calls: 1) A TE was called present if the mean coverage over the 10 bp just after the start and just before the end of the TE were both greater than or equal to 8x, 2) A TE was called absent if the mean coverage over the 10 bp just after the start and just before the end of the TE were both less than 7x, and 3) all other TEs were called as ambiguous that did not meet either of these criterion. TEs for which the alignment of short reads to its own genome assembly did not result in a present call were removed from downstream analyses. The threshold for calling a TE as present in the reference genome samples was less stringent than the final selected criteria for other genotypes (greater than 0x coverage), because there was no need to balance the calling a true present versus true absent TE. Method development showed that a low coverage threshold, 1x or 2x coverage, for calling a TE as present had a very high accuracy rate, but that the accuracy of calling a TE as absent was much lower at this low coverage.

This method was applied to the short read alignments across the four genome assemblies for the full set of sequenced lines described above to score a TE as present, absent, or ambiguous in a genotype. If a TE was called as ambiguous in more than 25% of genotypes it was excluded from further analysis. The non-redundant TE dataset from Anderson et al 2019 was used to combine TE population frequencies across homologous TEs from the different genome assemblies. The geometric mean of the population frequency for each TE was calculated between homologous TEs (File S1). TE metadata (e.g. location relative to the nearest gene or family) for the TE in the reference genome it was first identified in was used in downstream analyses using the ordering as previously described (ANDERSON *et al.* 2019).

Information about TE family/superfamily size and age of LTR retrotransposons was obtained from the previously published TE annotation gff files (ANDERSON *et al.* 2019). Gene annotations from B73 version 4 (JIAO *et al.* 2017), Mo17 (SUN *et al.* 2018), PH207 (HIRSCH *et al.* 2016), and W22 (SPRINGER *et al.* 2018) reference genomes were used to identify TE locations relative to genes. All metadata is included in File S1.

The Kolmogorov-Smirnov test was used to test whether the frequency distribution of nested and non-nested TEs was different. This test was implement using the ks.test function in base R version 3.6.2 (R DEVELOPMENT CORE TEAM 2020).

SNP identification and analysis: Joint SNP calling was performed across all genotypes using *freebayes* v1.3.1-17 (GARRISON AND MARTH 2012) for the alignments to the B73 v4 reference genome assembly. Sites with less than 1x average coverage over the whole population or greater than 2x the population level mean coverage were excluded from SNP calling. SNP calls were

quality filtered using GATK (v4.1.2) (VAN DER AUWERA *et al.* 2013) recommended filters as follows: QualByDepth (QD) less than 2, FisherStrand (FS) greater than 60, root mean square mapping quality (MQ) less than 40, MappingQualityRankSum (MQRankSum) less than -12.5, or ReadPosRankSum less than -8. If a SNP failed (e.g. MQ < 40) any one of those filters it was removed from the dataset. Finally, *vcftools* (v 0.1.13) (DANECEK *et al.* 2011) was used to filter sites for a minimum quality score of 30, a minimum allele count of 50 (called in at least 25 homozygotes or 50 heterozygotes) and to filter out genotypes called in fewer than 90% of the individuals in the population. In total, 3,146,253 SNPs remained after all of these filtering steps.

These SNP calls were used to conduct linkage disequilibrium analysis with the TE presence/absence scores described above. For this analysis, TE presence/absence data from only the alignments to B73 reference genome assembly were used. Any SNPs located within TEs were removed. Linkage disequilibrium between SNPs and TEs was calculated using plink v1.90b6.16 (PURCELL et al. 2007) with the --make-founders option to calculate LD among all inbred lines, --allow-extra-chr to calculate LD in extra scaffolds, --ld-window-r2 0 to report r2 in the 0 to 1 range (default is 0.2 to 1), and --ld-window 1000000 --ld-window-kb 1000 to calculate LD within 1mb windows. The SNP in the highest LD with a TE within the 1 Mb window was used for downstream visualization. Principle components analysis of SNPs and TEs were conducted using plink v1.90b6.16 (PURCELL et al. 2007).

Data Availability: All sequence data is available on the NCBI SRA (BioProject PRJNA661271, Table S1). Code for this study is available at https://github.com/HirschLabUMN/TE_variation.

RESULTS AND DISCUSSION

Using short read sequence data to identify TE presence/absence: For this study, we implemented an approach for scoring TE presence/absence variation from short read sequence alignment using the average coverage of windows surrounding the boundaries of previously annotated TEs. Iterations of the approach were evaluated for accuracy using false positive, true positive, false negative, true negative rates. For the accuracy assessment, presence/absence scores defined by previous comparison of TE content generated for four maize genome assembles including B73, Mo17, PH207, and W22 were used as truth (ANDERSON *et al.* 2019). It should be noted that some polymorphisms in this 'truth' set might be wrong for reasons described in the publication from which they were generated. As such, it will not be possible to get 100% accuracy in comparison to this set unless the same miscalls are generated in two independent methods. Still, this set represents a high quality set of TE polymorphisms for which to assess the relative accuracy of different parameters. The following parameters were factorially tested: 1) mean coverage of the windows used to classify a TE as present (2x-8x coverage), 2) size of the windows around the start and end of the TE (1, 3, 5, and 10 bp), and 3) coverage before and after the start and end of each TE (4 sides) versus only coverage just after the start and just before the end of the TE (2 sides) (Figure S1).

Initially, a TE was only classified as present or absent. To be classified as present the mean coverage over all genome fragments used had to be greater than or equal to the coverage threshold. If mean coverage was less than the threshold over any genome fragment at the start or end of the TE in question, the TE was classified as absent. From this initial factorial analysis, coverage threshold was identified as an important factor for TE calling accuracy (Figure 1). In

contrast, window size did not have a major impact on TE calling accuracy, nor did using coverage information from only just after the start and just before the end of a TE versus before and after the start and end of the TE boundaries (Figure 1). It is notable that the TE classification accuracy rate patterns differed in W22, compared to Mo17 and PH207, where the mean genome-wide coverage was substantially lower (6x average mean coverage for W22) versus 28x and 30x for Mo17 and PH207).

Further investigation showed that ~11% of the TEs had inconsistent coverage information, where coverage at one side of the TE was above the coverage threshold but the other side was below the threshold (Figure S2). In the original method, these TEs were called as absent, but 42% in PH207 and 50% in Mo17 are actually present based on the prior genome-wide comparisons. This is a much higher error rate than the overall error rate of ~10%. Based on this observation, a third possible TE classification category of ambiguous was added. With this new category, the criteria for classifying a TE as present did not change (i.e. all windows included in the analysis had to be above the given coverage threshold), but the criteria for classifying a TE as absent was changed to require that the mean coverage over all genome fragments of the TE had to be less than the coverage threshold minus one. All other TEs were classified as ambiguous (Figure S2).

Within our panel of sequenced genotypes, we had a range of observed genome wide mean coverages (Figure S3). To test the effects of coverage on TE classification accuracy rates the Mo17 and PH207 reads were down sampled to 20%, 32%, 45%, 66% and 83% of their original coverage levels. This corresponded to genome-wide mean coverages of ~6x, ~10x, ~15x, ~20x, and ~26x. Reducing mean coverage from 30x to 15x had a neglectable effect on TE classification

accuracy at a given coverage limit threshold (Figure S4), but reducing genome wide mean coverage from 30x to 6x had a noticeable effect on TE classification accuracy (Figure S4). As long as mean coverage was greater than 10x, classification rate accuracy was not affected by sequencing coverage. It is worth noting that the accuracy rates of Mo17 and PH207 down sampled to 6x genome wide coverage were similar to those of W22, which only had 6x mean coverage, which indicates that the down sampling accurately mimicked patterns of coverage for genotypes that have low mean coverage.

From these assessments of accuracy across different parameters a final criterion was identified to balance inaccurate classifications and the rate of ambiguous classifications. This final criterion required the average coverage of the two 10 bp internal flanking windows to be greater than or equal to 8x to be classified as present, and both sides less than 7x average coverage to be classified as absent. All other TEs were classified as ambiguous. Highly polymorphic flanking regions in some haplotypes will result in a higher ambiguous classification rate for particular TEs that could impact downstream data analysis of TE polymorphisms. Thus, TEs with great than 25% ambiguous calls were filtered from further analyses. These criterion and filtering methods provide a set of TE presence/absence calls that allows for population level analysis of TE frequencies on a genome-wide scale in maize. Still, it should be noted, that rare alleles will be under-represented in this set due to the fact that only those TEs that were previously annotated in at least one of four *de novo* assembled maize genomes are included in this study.

TE frequency distribution in a panel of diverse inbred lines: The final criteria for classifying a TE as present or absent was applied to short read sequence data from 509 genotypes of the Wisconsin Diversity Panel (HANSEY *et al.* 2010; HIRSCH *et al.* 2014; MAZAHERI *et al.* 2019) mapped to the B73, Mo17, PH207, and W22 reference genome assemblies. For each annotated TE in these genome assemblies, the presence/absence frequency of the TE in this panel of diverse inbred lines was determined (Figure S5). Genetic relationships based on SNPs and TEs were assessed to further validate the quality of these TE presence/absence calls based on their consistency with other marker types and known pedigree relationships. Principle component analysis of the SNPs and TEs both revealed expected relationships based on previous pedigree information and heterotic group membership (Figure S6). As previously reported, there are a number of redundant (shared) TEs across these four genome assemblies (ANDERSON *et al.* 2019). The population frequency of homologous TEs determined from mapping the short read data to the different genome assemblies were highly correlated (r^2 = 0.93; Figure S7), demonstrating the consistency of this pipeline, and enabling classifications to be combined into a non-redundant set of TEs. For downstream analysis of these redundant TEs, the geometric mean was calculated across frequencies obtained from all genomes for which a homologous TE was present (Figure 2; File S1).

Different orders of TEs have different mechanisms of replication (WICKER *et al.* 2007), and families within these orders have different insertional preferences (SANMIGUEL *et al.* 1998; SULTANA *et al.* 2017; SPRINGER *et al.* 2018) and different effects on DNA methylation and chromatin accessibility (EICHTEN *et al.* 2012; CHOI AND PURUGGANAN 2018; NOSHAY *et al.* 2019). As such, we hypothesized the frequency of TEs in the population would be variable across orders of TEs and

families. For non-nested elements (i.e. those not contained within another TE), a subset of the TEs were present in all or nearly all of the diverse lines included in this study for LTRs, Helitrons, and TIRs (Figure 2; Figure S8). However, for all three orders, there was a substantial number of TEs that were at low (<20%) to moderate (20-80%) frequency in the population. This proportion was particularly high for the LTRs where 52.1% of non-nested TEs were present in low to moderate frequency in the population. These polymorphic TEs have the potential to drive phenotypic variation as has been seen for a number of morphological/developmental (CHUCK *et al.* 2007; STUDER *et al.* 2011) and adaptive traits (YANG *et al.* 2013) in maize. The demonstrated extent of TE variability on a genome-wide scale across a large number of individuals that this study provides is critical in furthering our understanding of the contribution of variable TEs in producing phenotypic variation within species. Given that only TEs that were annotated in at least one of only four genomes were included in this study, it is expected that the proportion of polymorphic TEs at low to moderate frequency would be substantially higher than what is presented here if all TEs in all genotypes in the population were annotated. In contrast, the majority of the high frequency TEs were likely already captured from just these four genome assemblies, and the count of these TEs would remain relatively static if all genomes were to be *de novo* assembled and annotated.

TE Family size is negatively correlated with population frequency: There is a wide range of family sizes for TEs in each of the orders. For LTRs family sizes range from 1 to 16,433, for Helitrons family sizes range from 1 to 7,145, and for TIRs family sizes range from 1 to 10,351 (File S1). We sought to test if there is a relationship between the size of family for which a particular TE is a

member (based on its size in the annotated genome assembly) and its frequency in the population. For this analysis the average frequency of elements within a family was used so that large families do not skew the results. For all three orders there was a significant ($p<0.0001$) negative spearman correlation between family size and average TE frequency of members of the family (LTRs – rho = -0.14; Helitrons – rho = -0.39; TIRs – rho = -0.19). For LTRs there is a large number of elements in very small families (~22% of LTRs in families of less than 10). To test the impact of these very small families on these results, we removed them and the correlation was no longer significant for LTRs, indicating that much of this relationship is driven by small families. These results suggest that TEs present in small TE families are much more likely to exhibit variability in maize populations but it is worth noting that individual members of large families are often polymorphic too.

Age of LTRs is positively correlated with population frequency: A negative relationship between age of TEs and their frequency was previously demonstrated using a limited number (n=4) of genotypes (ANDERSON *et al.* 2019). We sought to test this relationship between age and frequency on our broader set of germplasm. For LTRs the age of an element can be estimated based on the LTR similarity, where younger elements have higher LTR similarity and older elements have lower LTR similarity. For other orders of TEs there are not accurate methods to assess age. Thus, when assessing the relationship between the age of TEs and their frequency in the population the analysis was limited to LTR elements. LTR similarity was negatively correlated with population frequency, which suggests that older LTRs were generally more frequent in the population than younger LTRs (Figure 3). It is worth noting, however, that there are a large number (n=23,506) of

LTRs that are very young (LTR similarity above 95), and are at high frequency (>80%) in the population. These TEs are potentially under selection or linked to other features in the genome under selection that could have driven their rapid rise in frequency in the population.

Nested TEs are less frequent and younger than non-nested TEs: Nested TEs exhibited higher levels of moderate frequency (20-80%) and low frequency (<20%) TEs as compared to the non-nested TEs (Figure 2; Figure S8). In all three orders, the frequency distribution of nested and non-nested TEs was significantly different (KS test, p-value < 2x10-16). This trend towards lower frequency TEs in the nested elements is likely a reflection of the younger age of these elements that have inserted into pre-existing TEs (Figure S9), and therefore do not follow the same distribution as the non-nested TEs. We further explored this difference in frequencies between nested and non-nested elements by plotting the frequency of each nested TE versus the frequency of the TE in which it is nested (Figure 4a). Overall, there was a strong correlation between the nested TEs and the TE into which it was nested (r^2 = 0.578; Figure 4a), which is expected. There was a subset of nested-outer elements that did not fall on the diagonal. Within this subset there was a much great portion that had a higher frequency in the outer element than in the nested element (Figure 4b).This is expected as the outer element must first be present before the nested element is able to insert as a nested element.

Within the set of outer elements that were fixed or nearly fixed in the population (frequency greater than 0.95), there was a continuous range of frequencies for inner elements that likely represent nested insertions with a range of ages. Using LTR similarity as a proxy to LTR age, we tested if there is a relationship between the age of the nested element contained within

fixed elements and their frequency in the population. Indeed, those nested elements that were at lower frequency in the population were younger (higher LTR similarity) than those that were at high frequency in the population (Figure 4c). There were also outer elements in a range of frequencies that had extremely rare nested TEs that are likely very young or deleterious nested insertions.

Many TE insertional polymorphisms are not tagged by SNP markers: There is a limited number of plant species in which a species or population level cataloging of TE presence/absence variation has been conducted at a genome-wide level (QUADRANA *et al.* 2016; STUART *et al.* 2016; DOMINGUEZ *et al.* 2020). As such, the contribution of TEs to phenotypic variation, utility for genomic prediction, and other applications linking genotypes and phenotypes has not been directly evaluated. In many instances, a linked marker has been identified as significant and with fine-mapping it is revealed that the causative variant is in fact a polymorphic TE insertion (e.g. (STUDER *et al.* 2011)). We sought to test the extent to which the extensive number of polymorphic TE insertions that were classified in this study were in linkage disequilibrium (LD) with genome-wide SNP markers to begin to understand the extent to which phenotypic variation that is caused by TEs has or has not been accounted for in previous studies that used only SNP markers. For this analysis, all SNPs within 1Mb plus or minus a TE were evaluated and the SNP that was in the highest LD was recorded. The majority of TEs were in moderate (r^2 0.5-0.9) to high (r^2 >0.9) LD with a nearby SNP marker, but there was a subset that were in low LD (r^2 < 0.5) with all SNPs within a 1Mb window of the TE (Figure 5a). LTRs had the highest portion of TEs in high LD with a SNP (~73%), while Helitrons had ~61% in high LD and TIRs had only ~54% in high LD (Figure 5b).

Of the subset of TEs that had a SNP in high LD (>0.9), the majority of the SNPs were within 200Kb of the TE (Figure 5c). This distance increased when TEs in all levels of LD with SNPs were included (Figure S10). TEs that were at very high or low frequency within the population were generally in low LD with SNPs, and those that were at moderate frequency in the population were in moderate to high LD with SNPs (Figure 5d-f). This trend was consistently observed across orders of TEs (Figure 5d-f). This analysis demonstrates that while TE presence/absence patterns generally reflect maize breeding history (Figure S6), there are TE insertional polymorphisms that are not tagged by SNPs. This lack of LD is generally related to their frequency in the population, and rare TEs variants, that may be of important phenotypic consequence are the ones that are not being tagged. Including TE insertional polymorphism will likely be important in understanding the full complexity of phenotypic trait variation and location adaptation, and developing improved maize varieties in the future.

ACKNOWLEDGMENTS

This work was funded in part by NSF Grant IOS-1546727 to CNH, NSF Grant IOS-1934384 to CNH and NMS, and USDA Grant 2018-67013-27571 to CNH. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper.

FIGURE LEGENDS

Figure 1. TE calling accuracy rates. a) TE calling accuracy rates using 4 windows flanking the inside and outside of the TE boundaries to call a TE as present or absent. B) TE call accuracy rates using the two internal windows (immediately after the start and immediately before the end) of the TE. All TEs were called as either present or absent in this analysis (i.e. no ambiguous call category was implemented).

Figure 2. TE frequency distribution of the non-redundant set of TEs annotated in the B73, Mo17, PH207, and W22 genome assemblies. Only TEs with less than 25% ambiguous calls are included in these plots.

Figure 3. Relationship between TE age and frequency in a population of diverse inbred lines. Age is approximated by LTR similarity as previously reported (ANDERSON *et al.* 2019). Only LTR TEs are included in this plot (n=186,211).

Figure 4. Relationship between population frequency of nested elements and the elements in which they are nested. a) Proportion of genotypes a TE is present in between nested TEs and the TE in which the element is nested. Only low ambiguous TEs were included in this plot. Only TEs from the B73 genome are plotted. b) Distribution of the proportion of genotypes the outer TE is present in minus proportion of genotypes the nested TE is present in. c) Age distributions for nested elements that are nested in TEs that are fixed or nearly fixed (frequency >0.95) in the population. LTR similarity is used as a proxy to age which higher LTR similarity corresponds to a

younger TE and lower LTR similarity corresponds to an older TE. This plot only contains nested LTRs as age estimates are not available for other orders.

Figure 5. Linkage disequilibrium between TEs and SNPs in a panel of diverse inbred lines. a) Linkage disequilibrium (LD) between TEs and the SNP with the highest LD within 1Mb of the middle of the TE. b) Proportion of TEs in high ($r^2 > 0.9$), moderate (r^2 0.5-0.9), and low ($r^2 <0.5$) LD with SNPs within 1 Mb of the middle of the TE. Category is based on the SNP with the highest LD in the window. c) Distance between TEs and the SNP with the highest LD to it for TEs that had a SNP in high ($r^2 > 0.9$) LD. Distance is calculated as the middle of the TE to the SNP. Only SNPs within 1 Mb of a TE were evaluated. d-f) Density plots of population frequencies for TEs in high, moderate, and low LD with SNPs based on the highest LD within 1Mb of the middle of the TE for LTRs (d), Helitrons (e), and TIRs (f).

REFERENCES

Anderson, S. N., M. C. Stitzer, A. B. Brohammer, P. Zhou, J. M. Noshay *et al.*, 2019 Transposable elements contribute to dynamic genome content in maize. Plant J 100**:** 1052-1065.

Bennetzen, J. L., 2000 Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42**:** 251-269.

Butelli, E., C. Licciardello, Y. Zhang, J. Liu, S. Mackay *et al.*, 2012 Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24**:** 1242-1255.

Cadle-Davidson, M. M., and C. L. Owens, 2008 Genomic amplification of the Gret1 retroelement in white-fruited accessions of wild vitis and interspecific hybrids. Theor Appl Genet 116**:** 1079-1094.

Chiang, C., R. M. Layer, G. G. Faust, M. R. Lindberg, D. B. Rose *et al.*, 2015 SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods 12**:** 966-968.

Choi, J. Y., and M. D. Purugganan, 2018 Evolutionary Epigenomics of Retrotransposon-Mediated Methylation Spreading in Rice. Mol Biol Evol 35**:** 365-382.

Chuck, G., A. M. Cigan, K. Saeteurn and S. Hake, 2007 The heterochronic maize mutant Corngrass1 results from overexpression of a tandem microRNA. Nat Genet 39**:** 544-549.

Clegg, M. T., and M. L. Durbin, 2000 Flower color variation: a model for the experimental study of evolution. Proc Natl Acad Sci U S A 97**:** 7016-7023.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27**:** 2156-2158.

Dominguez, M., E. Dugas, M. Benchouaia, B. Leduque, J. M. Jimenez-Gomez *et al.*, 2020 The impact of transposable elements on tomato diversity. Nat Commun 11**:** 4058.

Dooner, H. K., Q. Wang, J. T. Huang, Y. Li, L. He *et al.*, 2019 Spontaneous mutations in maize pollen are frequent in some lines and arise mainly from retrotranspositions and deletions. Proc Natl Acad Sci U S A 116**:** 10734-10743.

Eichten, S. R., N. A. Ellis, I. Makarevitch, C. T. Yeh, J. I. Gent *et al.*, 2012 Spreading of heterochromatin is limited to specific families of maize retrotransposons. PLoS Genet 8**:** e1003127.

Fu, H., and H. K. Dooner, 2002 Intraspecific violation of genetic colinearity and its implications in maize. Proc Natl Acad Sci U S A 99**:** 9573-9578.

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. arXiv arXiv:1207.3907.

Hansey, C. N., J. M. Johnson, R. S. Sekhon, S. M. Kaeppler and N. de Leon, 2010 Genetic Diversity of a Maize Association Population with Restricted Phenology. Crop Sci. 51**:** 704-715.

Hirsch, C. N., J. M. Foerster, J. M. Johnson, R. S. Sekhon, G. Muttoni *et al.*, 2014 Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26**:** 121-135.

Hirsch, C. N., C. D. Hirsch, A. B. Brohammer, M. J. Bowman, I. Soifer *et al.*, 2016 Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. Plant Cell 28**:** 2700-2714.

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

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25**:** 1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25**:** 2078-2079.

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

Makarevitch, I., A. J. Waters, P. T. West, M. Stitzer, C. N. Hirsch *et al.*, 2015 Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet 11**:** e1004915.

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal 17**:** 10-12.

Mazaheri, M., M. Heckwolf, B. Vaillancourt, J. L. Gage, B. Burdo *et al.*, 2019 Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol 19**:** 45.

Morgante, M., S. Brunner, G. Pea, K. Fengler, A. Zuccolo *et al.*, 2005 Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37**:** 997-1002.

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

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

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26**:** 841-842.

R Development Core Team, 2020 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rimmer, A., H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg *et al.*, 2014 Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet 46**:** 912-918.

Saghai-Maroof, M. A., K. M. Soliman, R. A. Jorgensen and R. W. Allard, 1984 Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci U S A 81**:** 8014-8018.

SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima and J. L. Bennetzen, 1998 The paleontology of intergene retrotransposons of maize. Nat Genet 20**:** 43-45.

SanMiguel, P., A. Tikhonov, Y. K. Jin, N. Motchoulskaia, D. Zakharov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science 274**:** 765-768.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei *et al.*, 2009 The B73 maize genome: complexity, diversity, and dynamics. Science 326**:** 1112-1115.

Selinger, D. A., and V. L. Chandler, 2001 B-Bolivia, an allele of the maize b1 gene with variable expression, contains a high copy retrotransposon-related sequence immediately upstream. Plant Physiol 125**:** 1363-1379.

Springer, N. M., S. N. Anderson, C. M. Andorf, K. R. Ahern, F. Bai *et al.*, 2018 The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat Genet 50**:** 1282-1288.

Stitzer, M., S. N. Anderson, N. M. Springer and J. Ross-Ibarra, 2019 The Genomic Ecosystem of Transposable Elements in Maize. bioRxiv https://doi.org/10.1101/559922.

Stuart, T., S. R. Eichten, J. Cahn, Y. V. Karpievitch, J. O. Borevitz *et al.*, 2016 Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. Elife 5.

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

Sultana, T., A. Zamborlini, G. Cristofari and P. Lesage, 2017 Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet 18**:** 292-308.

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

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel *et al.*, 2013 From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43**:** 11 10 11-11 10 33.

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

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

Zhang, L., J. Hu, X. Han, J. Li, Y. Gao *et al.*, 2019 A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nat Commun 10**:** 1494.

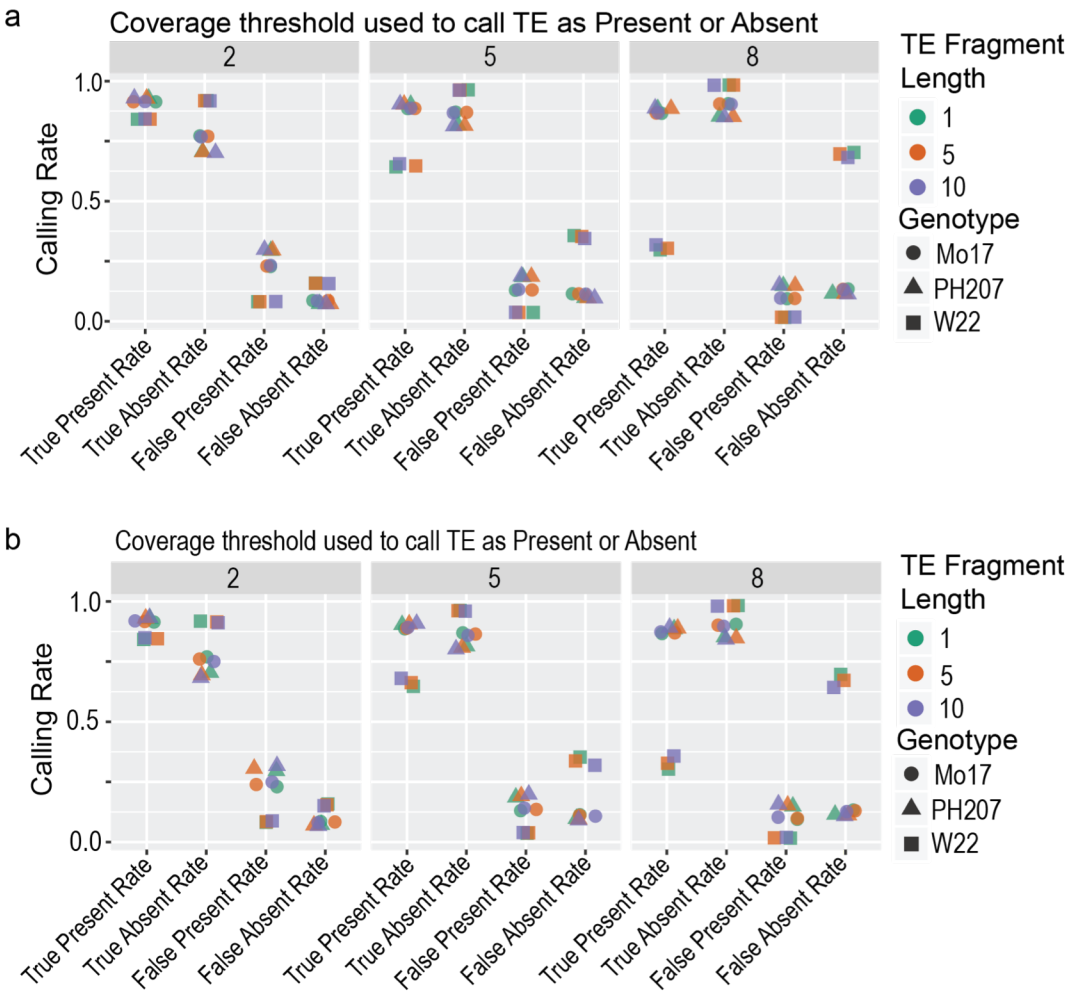

Figure 1. TE calling accuracy rates. a) TE calling accuracy rates using 4 windows flanking the inside and outside of the TE boundaries to call a TE as present or absent. b) TE call accuracy rates using the two internal windows (immediately after the start and immediately before the end) of the TE. All TEs were called as either present or absent in this analysis (i.e. no ambiguous call category was implemented).

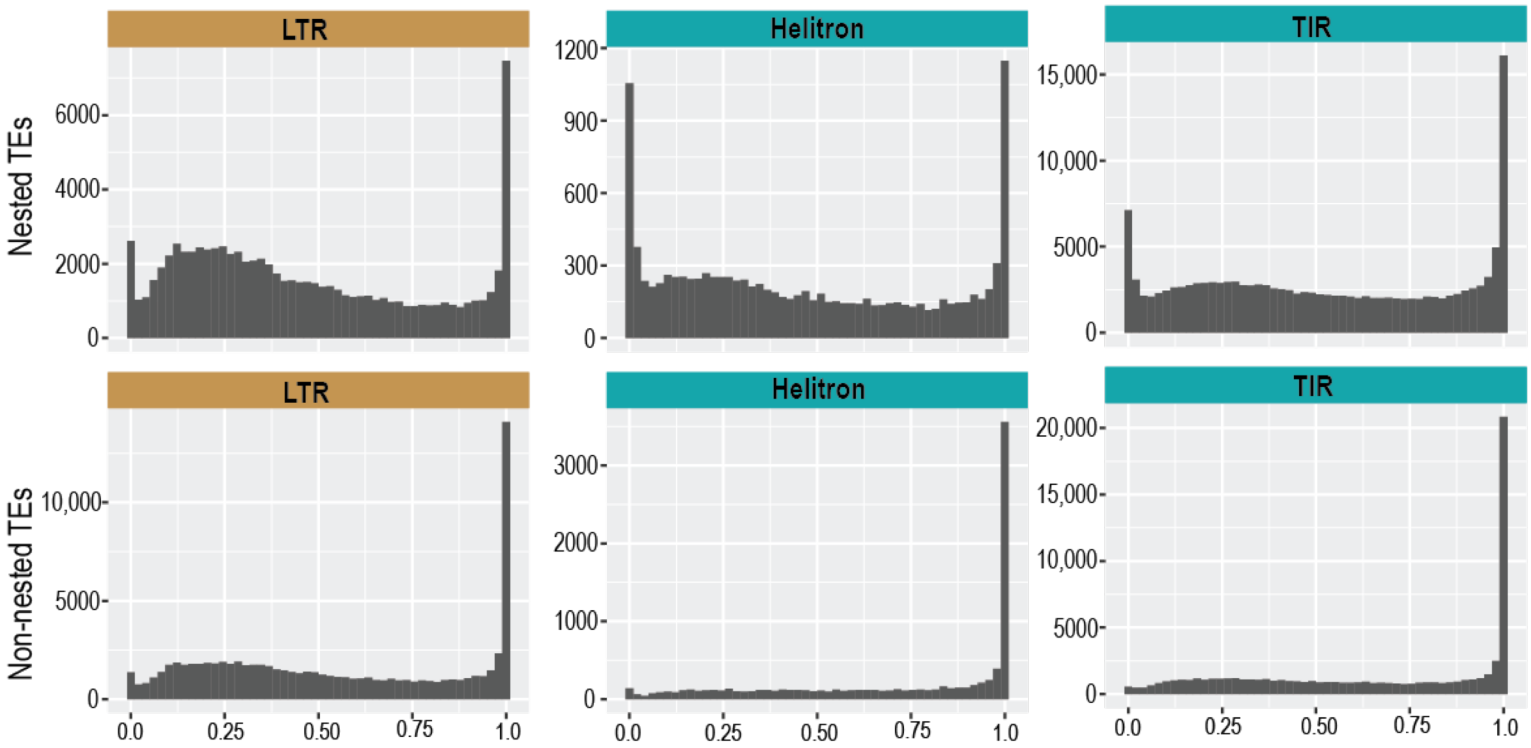

Figure 2. TE frequency distribution of the non-redundant set of TEs annotated in the B73, Mo17, PH207, and W22 genome assemblies. Only TEs with less than 25% ambiguous calls are included in these plots.

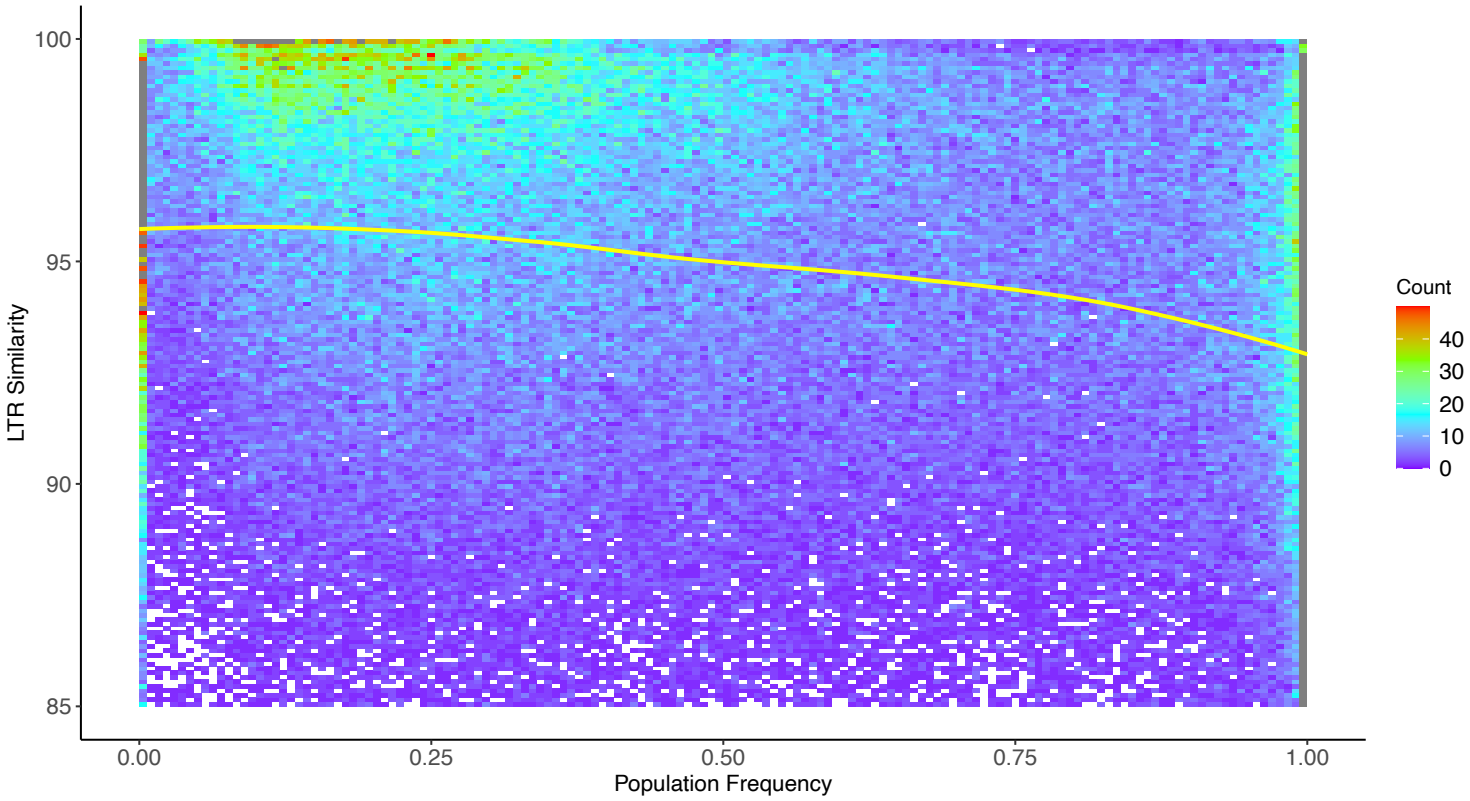

Figure 3. Relationship between TE age and frequency in a population of diverse inbred lines. Age is approximated by LTR similarity as previously reported (Anderson et al., 2019). Only LTR TEs are included in this plot (n=186,211).

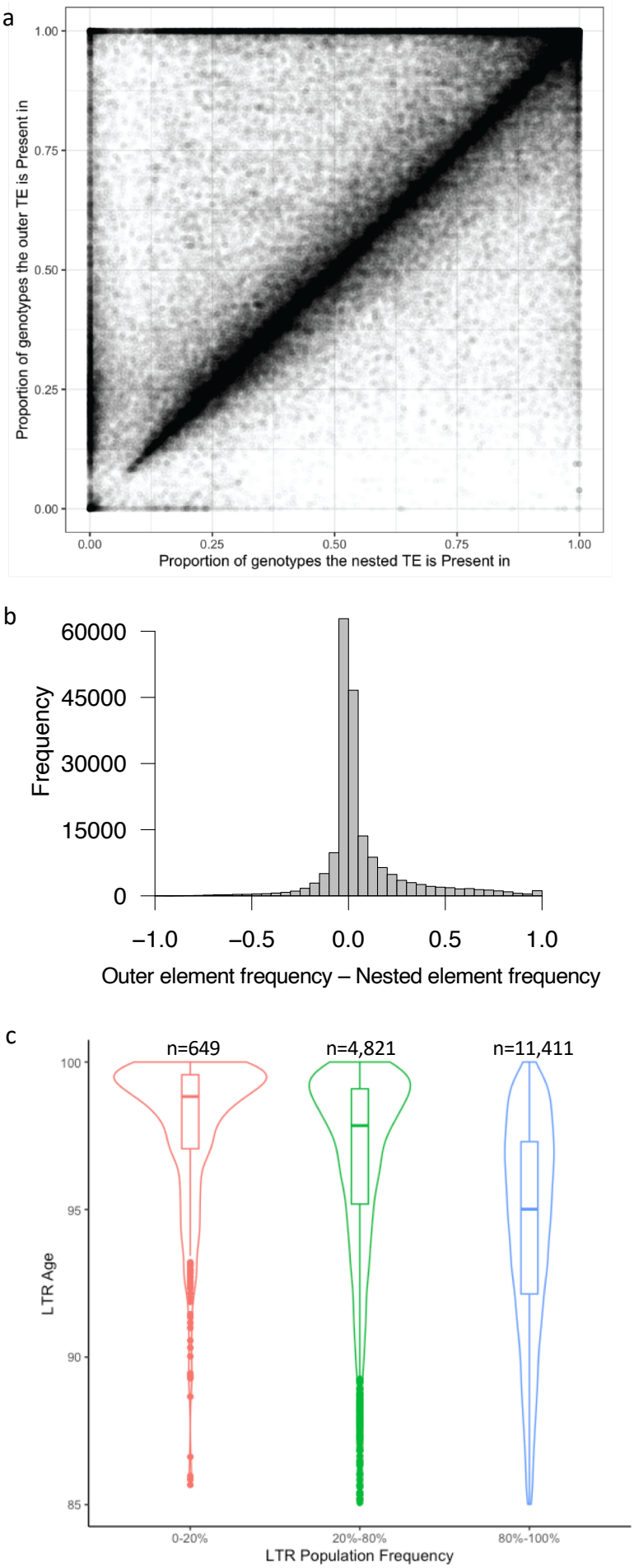

Figure 4. Relationship between population frequency of nested elements and the elements in which they are nested. a) Proportion of genotypes a TE is present in between nested TEs and the TE in which the element is nested. Only low ambiguous TEs were included in this plot. Only TEs from the B73 genome are plotted. b) Distribution of the proportion of genotypes the outer TE is present in minus proportion of genotypes the nested TE is present in. c) Age distributions for nested elements that are nested in TEs that are fixed or nearly fixed (frequency >0.95) in the population. LTR similarity is used as a proxy to age which higher LTR similarity corresponds to a younger TE and lower LTR similarity corresponds to an older TE. This plot only contains nested LTRs as age estimates are not available for other orders.

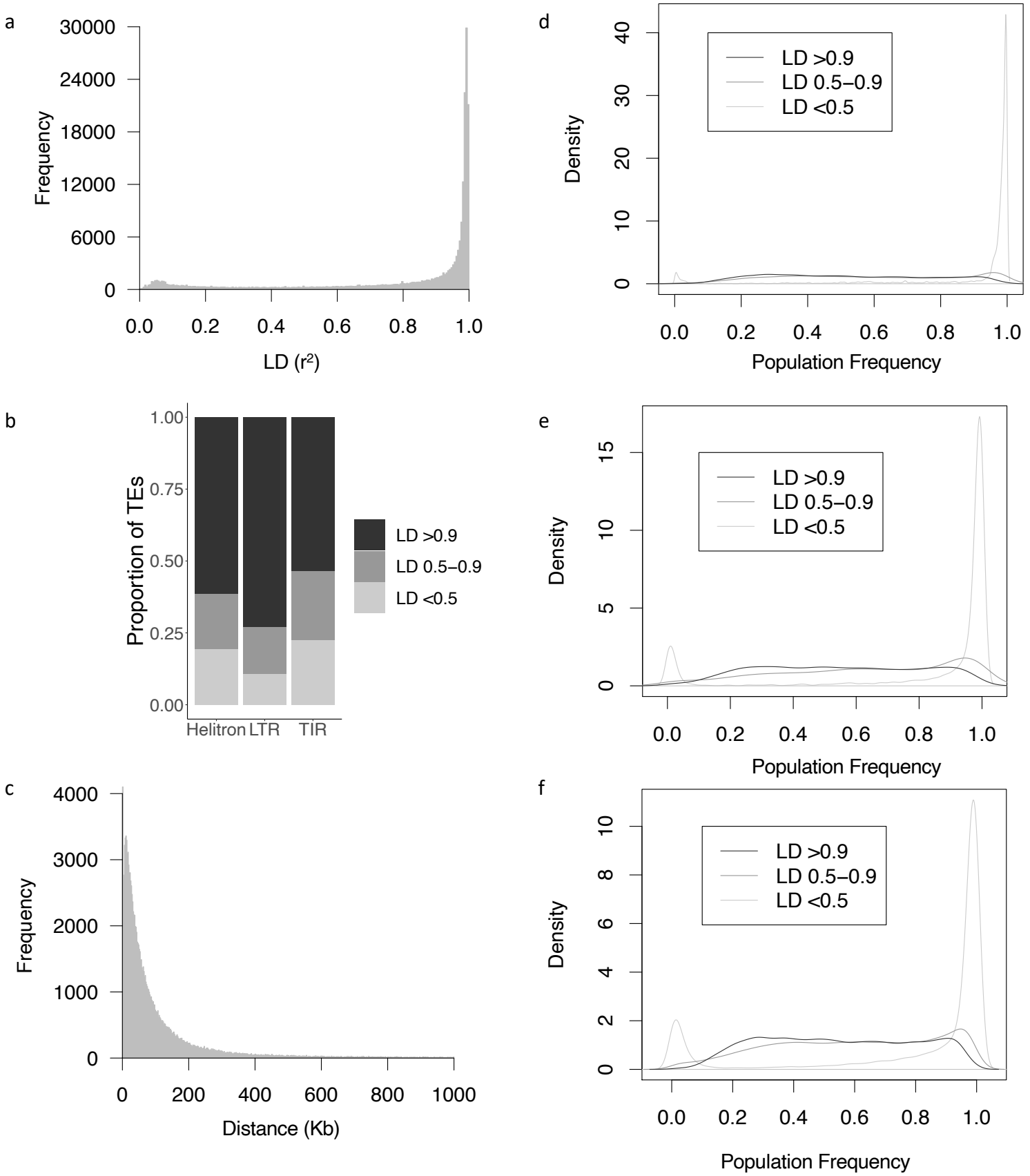

Figure 5. Linkage disequilibrium between TEs and SNPs in a panel of diverse inbred lines. a) Linkage disequilibrium (LD) between TEs and the SNP with the highest LD within 1Mb of the middle of the TE. b) Proportion of TEs in high ($r^2 > 0.9$), moderate (r^2 0.5-0.9), and low ($r^2 < 0.5$) LD with SNPs within 1 Mb of the middle of the TE. Category is based on the SNP with the highest LD in the window. c) Distance between TEs and the SNP with the highest LD to it for TEs that had a SNP in high ($r^2 > 0.9$) LD. Distance is calculated as the middle of the TE to the SNP. Only SNPs within 1 Mb of a TE were evaluated. d-f) Density plots of population frequencies for TEs in high, moderate, and low LD with SNPs based on the highest LD within 1Mb of the middle of the TE for LTRs (d), Helitrons (e), and TIRs (f).

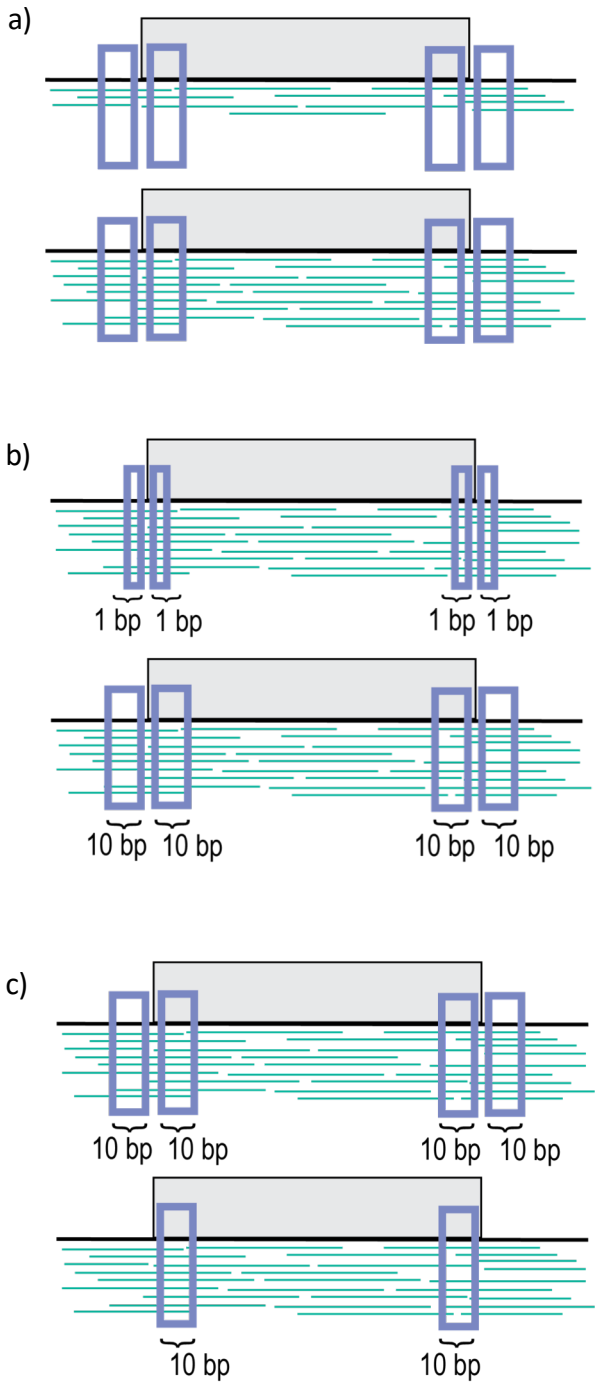

Figure S1. Schematic of factors tested for accuracy of TE calling using overage of short read sequence alignments in windows surrounding annotated transposable elements. a) variable coverage within windows to call a TE as present. b) variable window sizes. c) internal and external flanking windows versus only internal windows.

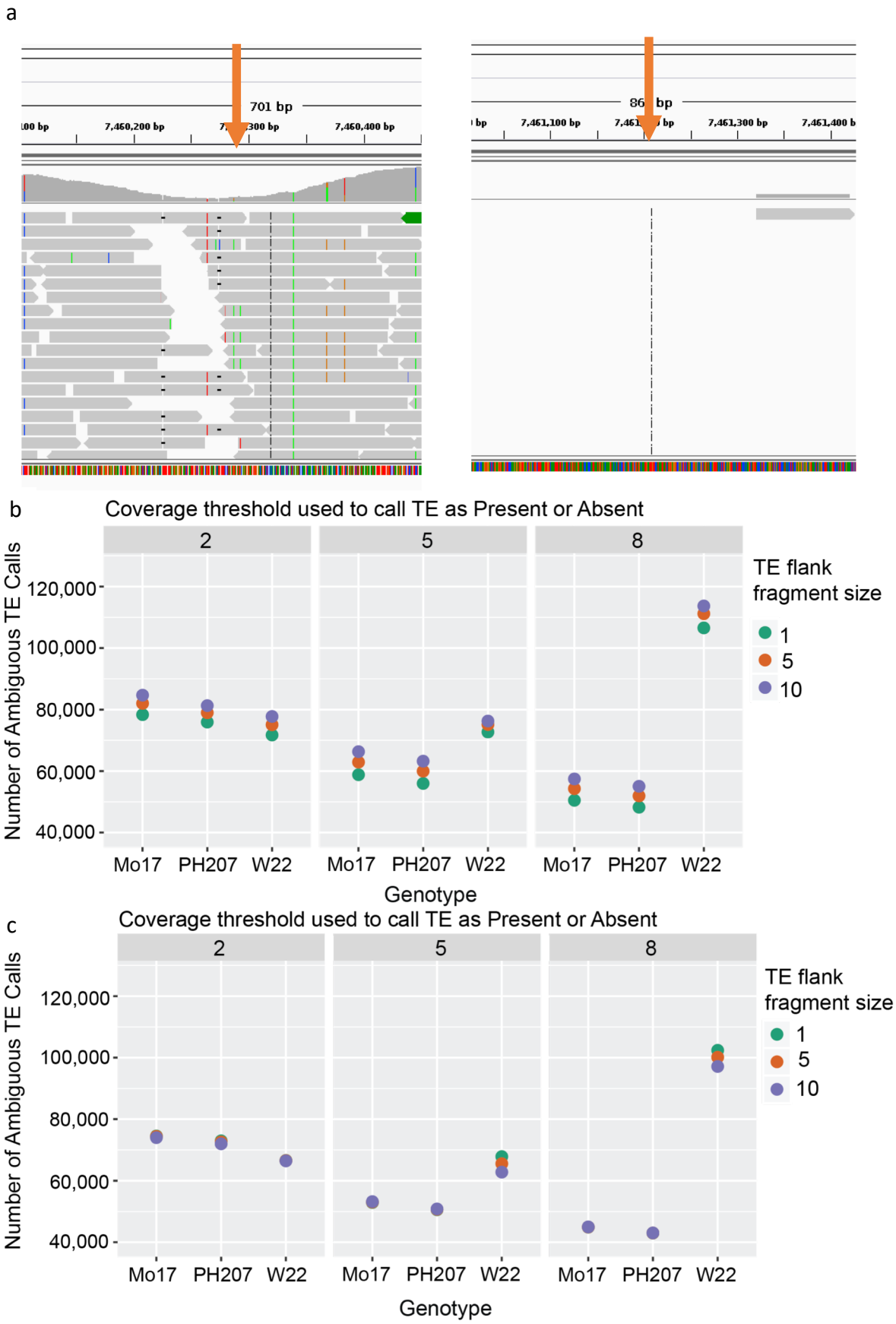

Figure S2. Rate of ambiguous calls at variable coverage thresholds. a) Example showing inconsistent coverage of B73 reads mapped to the Mo17 genome that results in an ambiguous call. b) Number of ambiguous TE calls using four flanking windows for each TE. c) Number of ambiguous TE calls using only the inner two flanking windows.

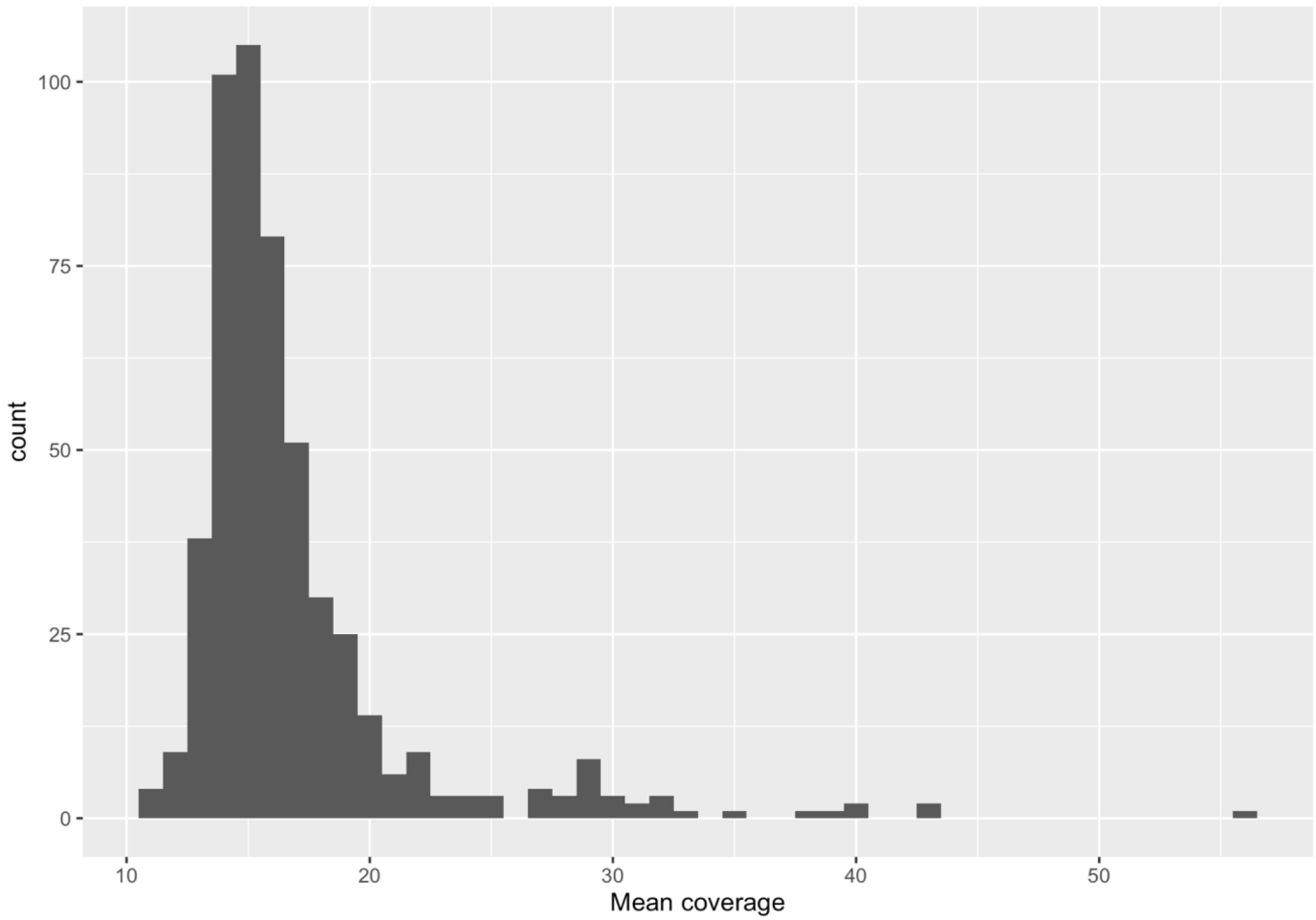

Figure S3. Realized genome-wide mean coverage for samples used in this study.

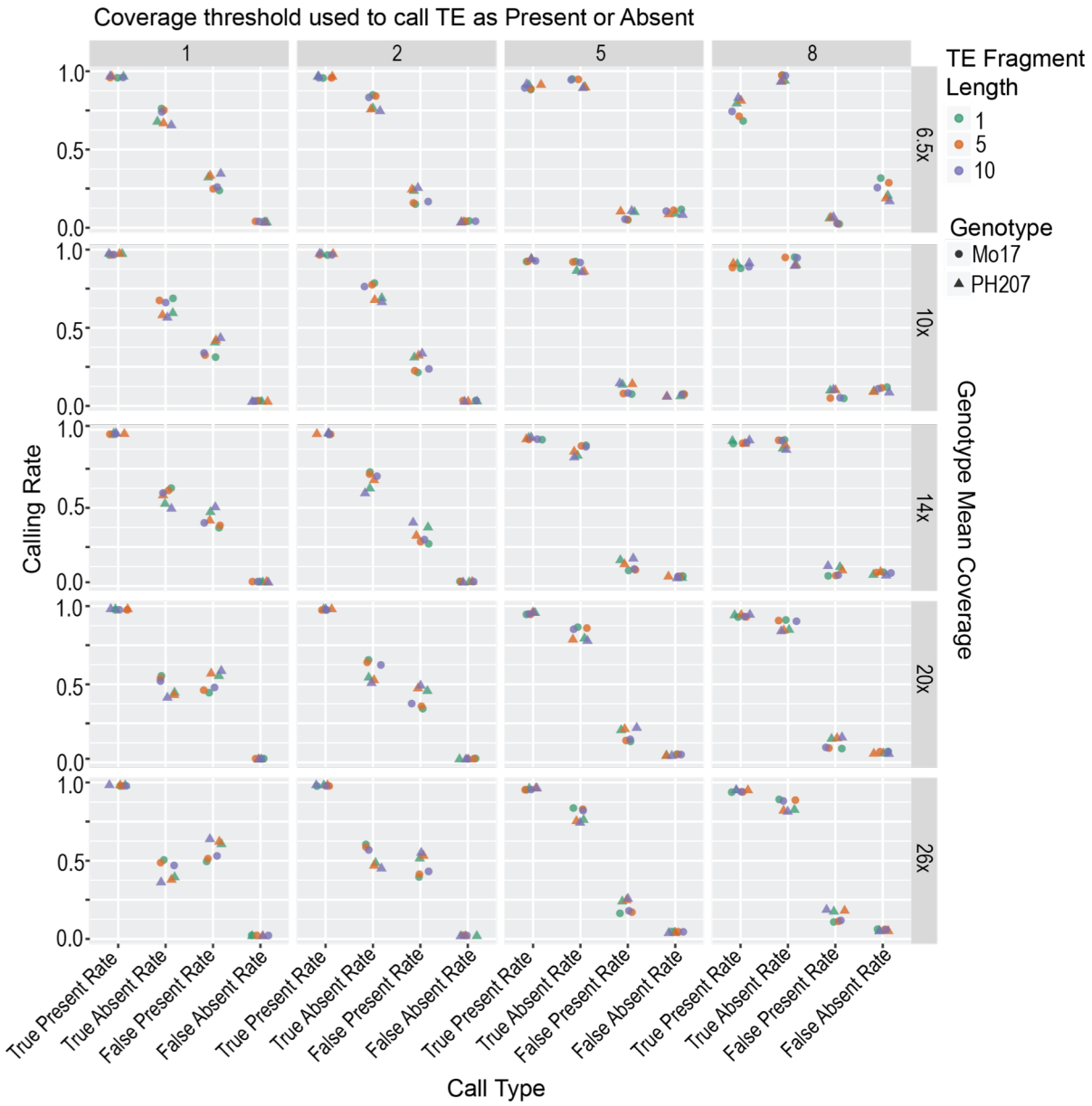

Figure S4. TE calling accuracy for Mo17 and PH207 down sampled to different coverage levels. Coverage over the two internal flanking windows were used to call a TE as Present, Absent, or Ambiguous. Ambiguous calls were exclude from TE calling rates. The majority of samples used in this study have realized coverage between 10-30x.

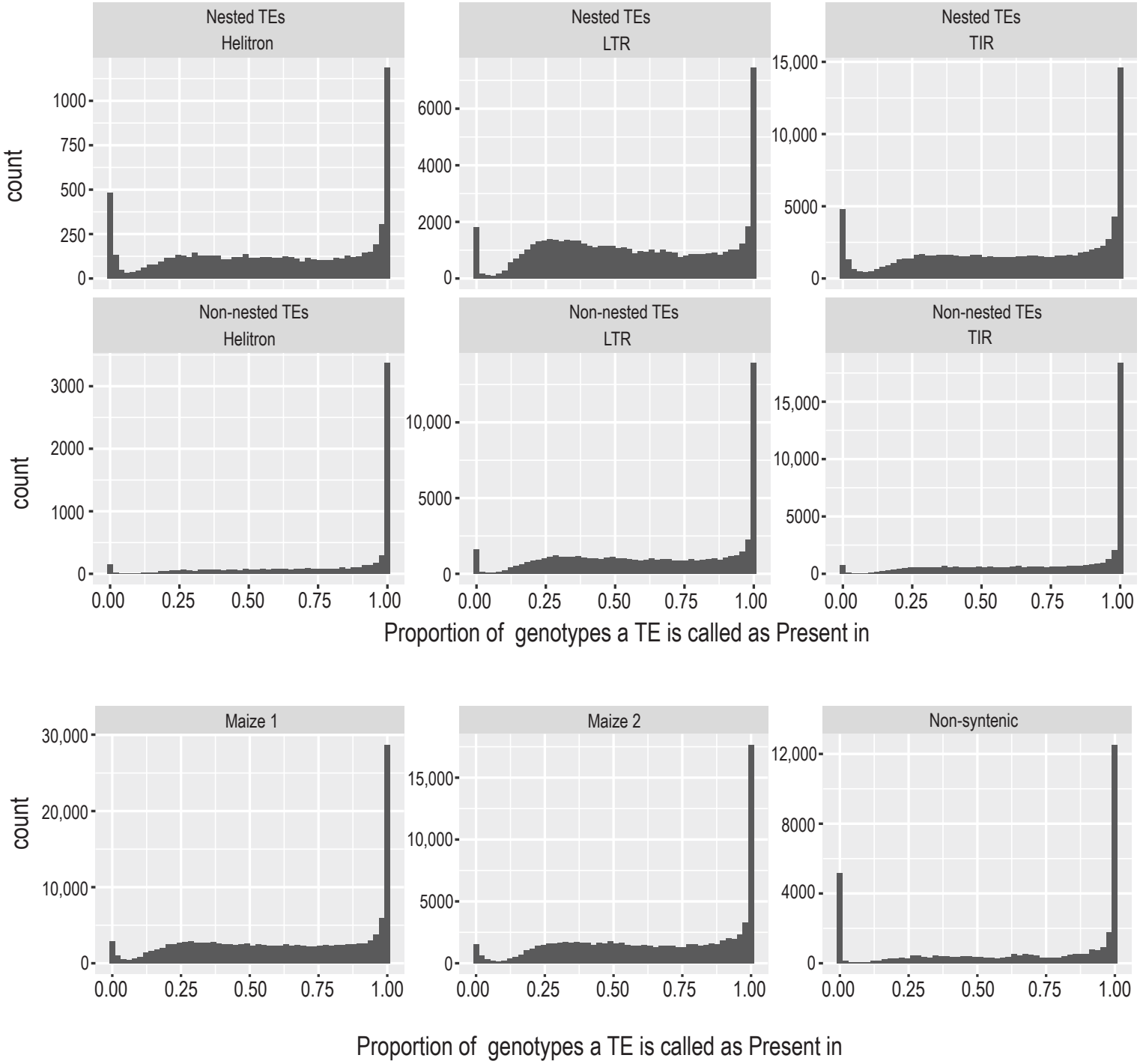

Figure S5. TE frequency distribution in a panel of diverse maize inbred lines. Genotypes include 508 individuals for the Wisconsin Diversity Panel (Hansey et al., 2010). Frequencies are for TEs annotated in the B73 v4 reference genome assembly (Anderson et al., 2019).

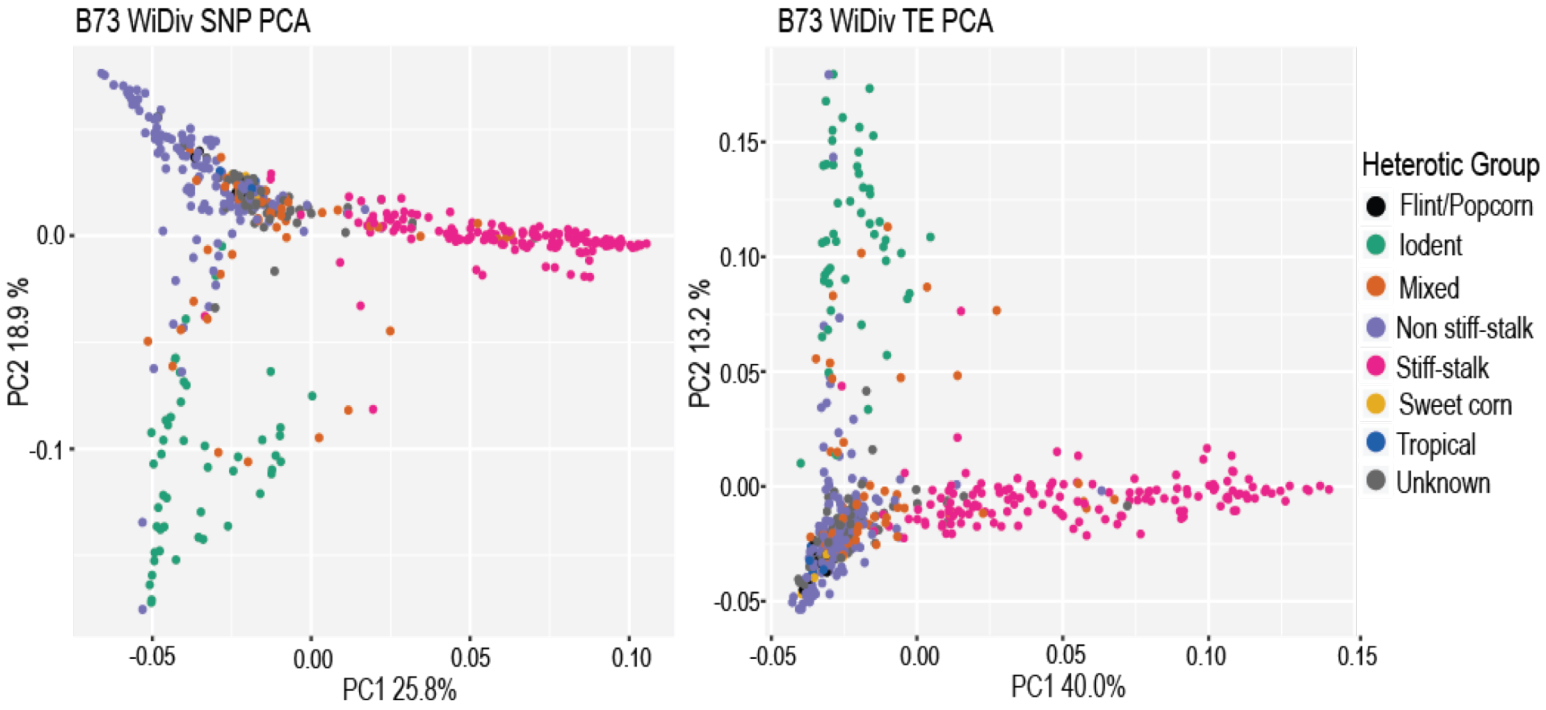

Figure S6. Principle component analysis (PCA) of SNPs (left) and TE presence/absence (right) in a set of diverse maize inbred lines. TEs and SNPs were called relative to the B73 reference genome assembly.

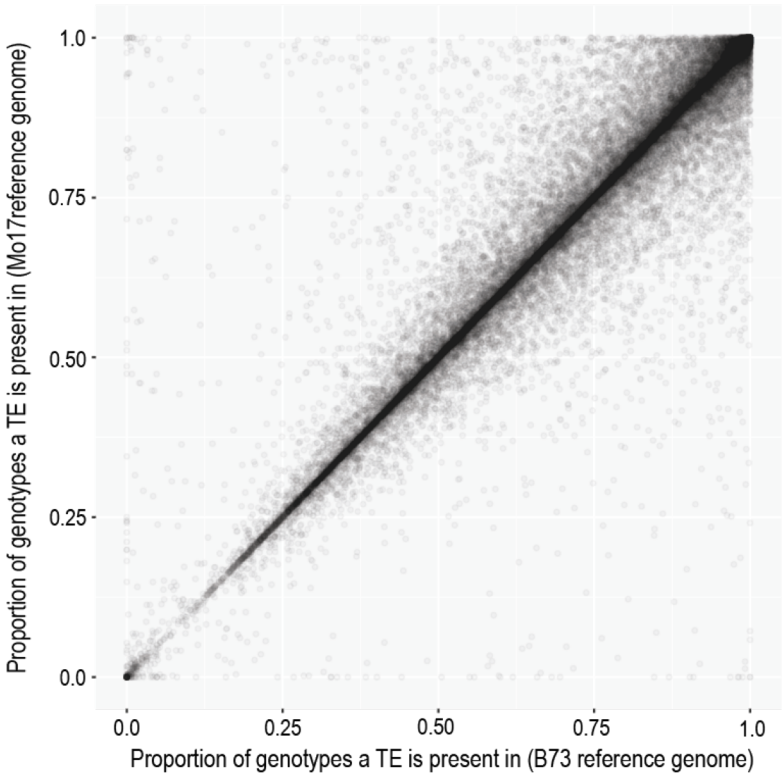

Figure S7. Proportion of genotypes a TE was called as 'Present' in between homologous TEs in B73 and Mo17. Only low ambiguous TEs included.

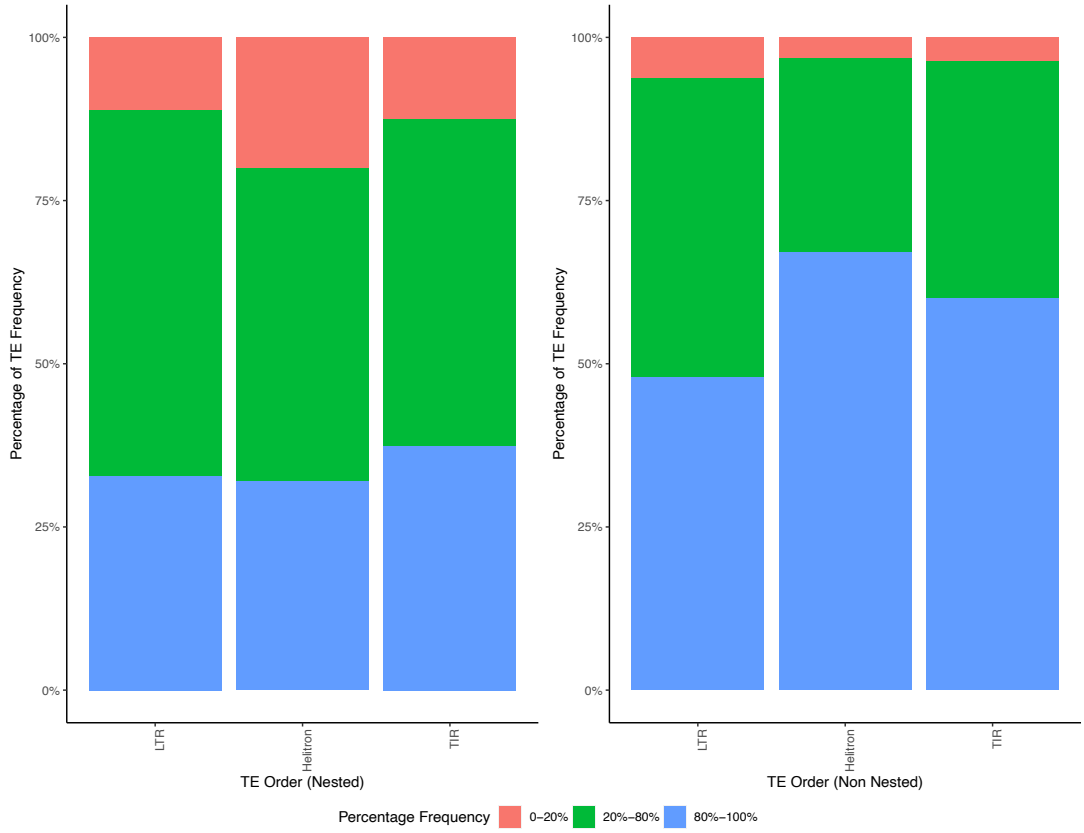

Figure S8. Proportion of low, medium, and high frequency TEs in a panel of diverse inbred lines. Genotypes include 508 individuals for the Wisconsin Diversity Panel (Hansey et al., 2010). Frequencies are for the set of non-redundant TEs across four genome assemblies (Anderson et al., 2019).

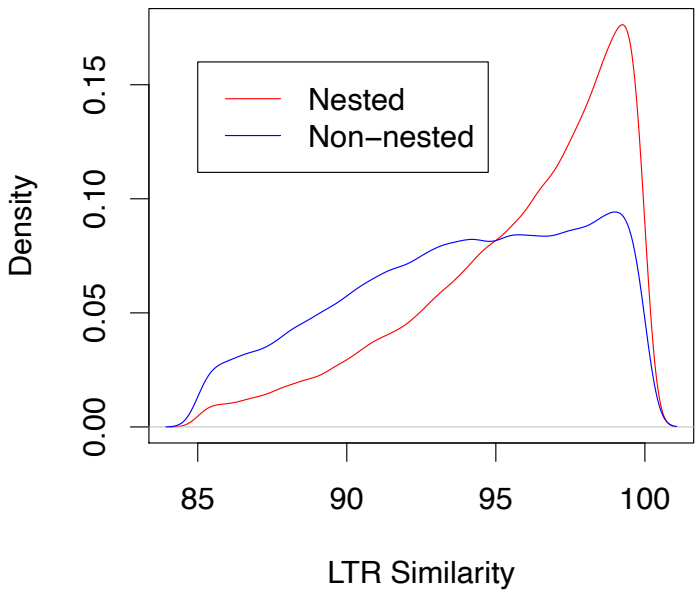

Figure S9. Age distribution of nested and non-nested elements in a non-redundant set of 202,267 LTR elements. Age of elements is approximated based on LTR similarity for each TE.

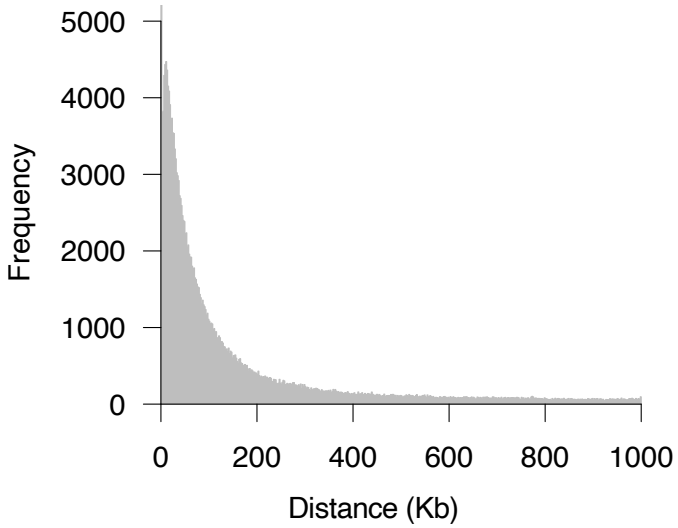

Figure S10. Distance between TEs and the SNP with the highest LD to it. Distance is calculated as the middle of the TE to the SNP. Only SNPs within 1Mb of a TE were evaluated.

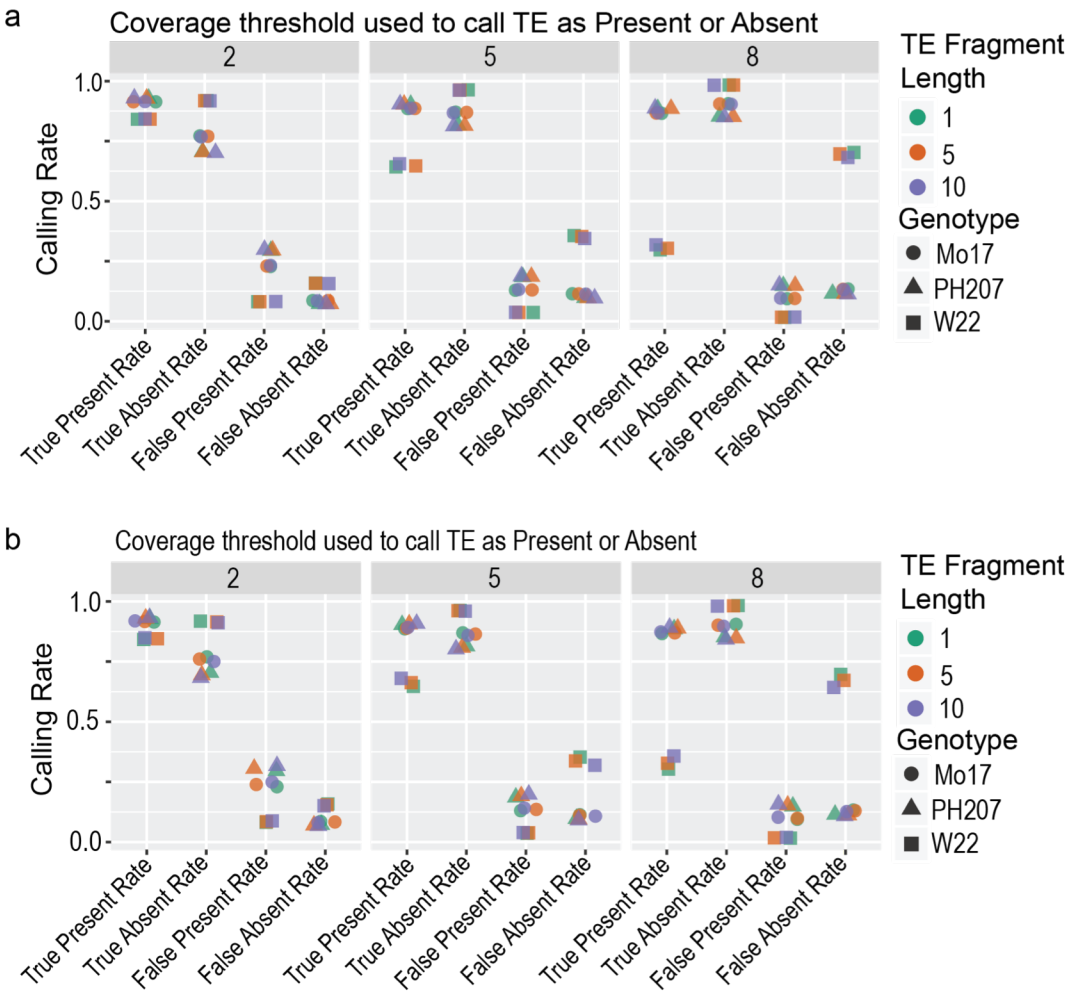

Figure 1. TE calling accuracy rates. a) TE calling accuracy rates using 4 windows flanking the inside and outside of the TE boundaries to call a TE as present or absent. b) TE call accuracy rates using the two internal windows (immediately after the start and immediately before the end) of the TE. All TEs were called as either present or absent in this analysis (i.e. no ambiguous call category was implemented).

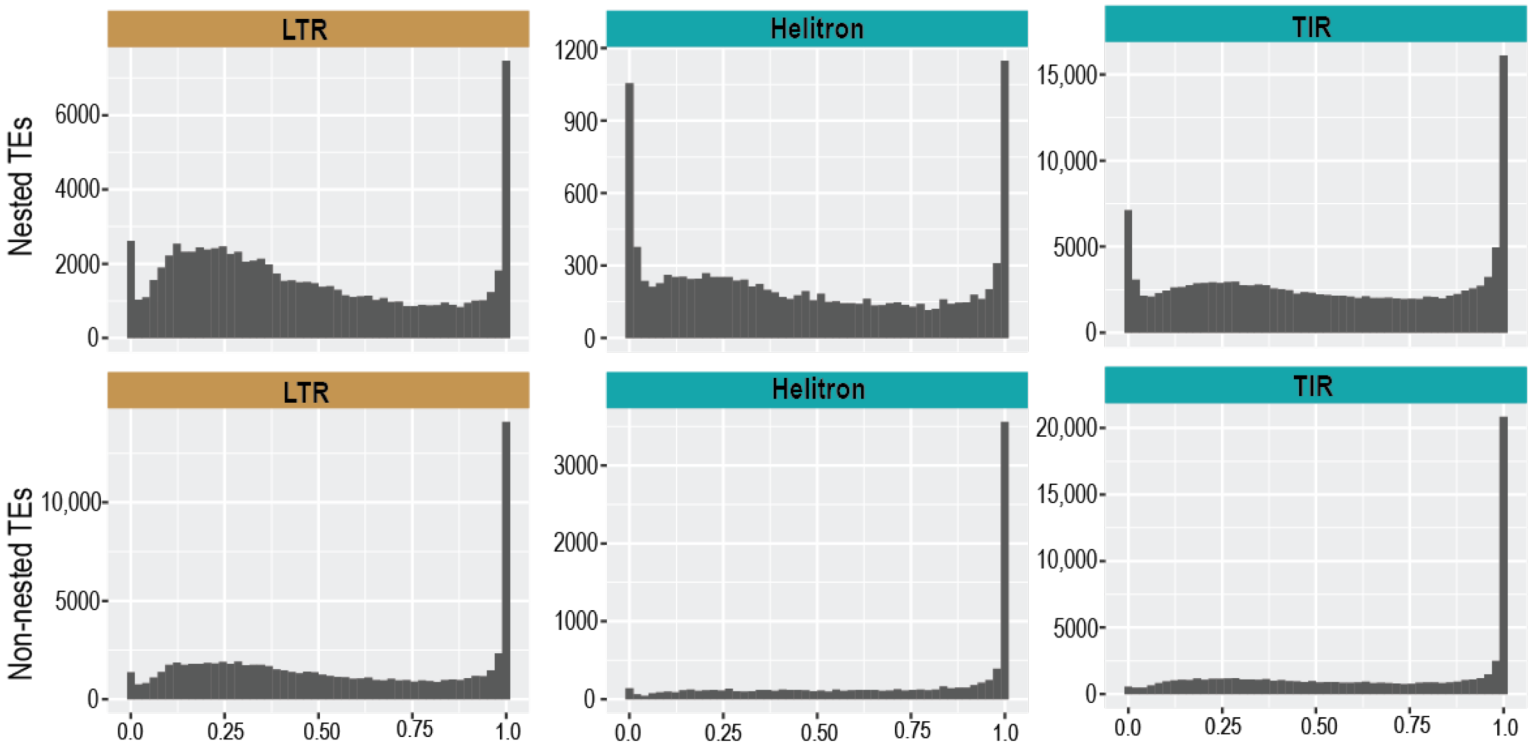

Figure 2. TE frequency distribution of the non-redundant set of TEs annotated in the B73, Mo17, PH207, and W22 genome assemblies. Only TEs with less than 25% ambiguous calls are included in these plots.

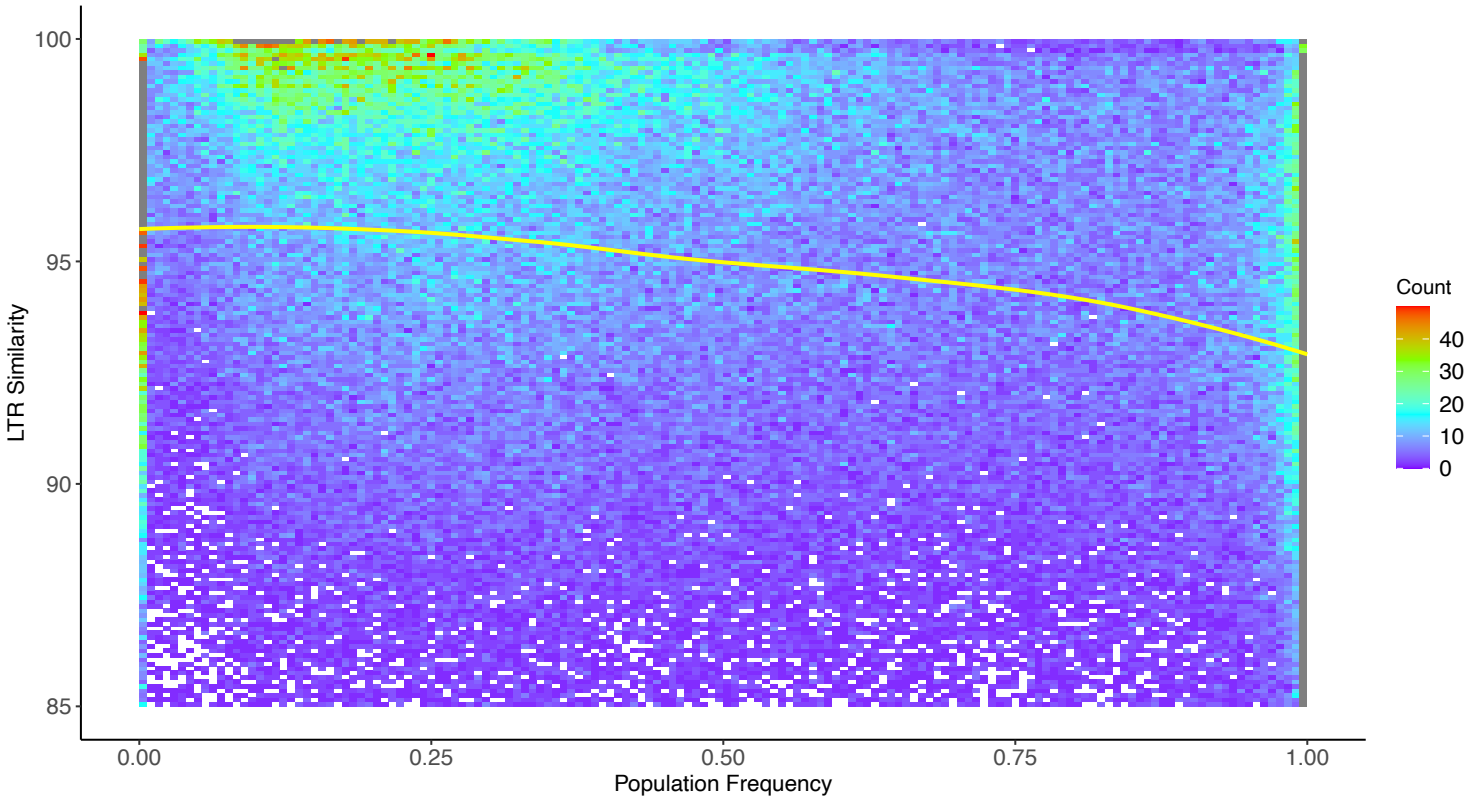

Figure 3. Relationship between TE age and frequency in a population of diverse inbred lines. Age is approximated by LTR similarity as previously reported (Anderson et al., 2019). Only LTR TEs are included in this plot (n=186,211).

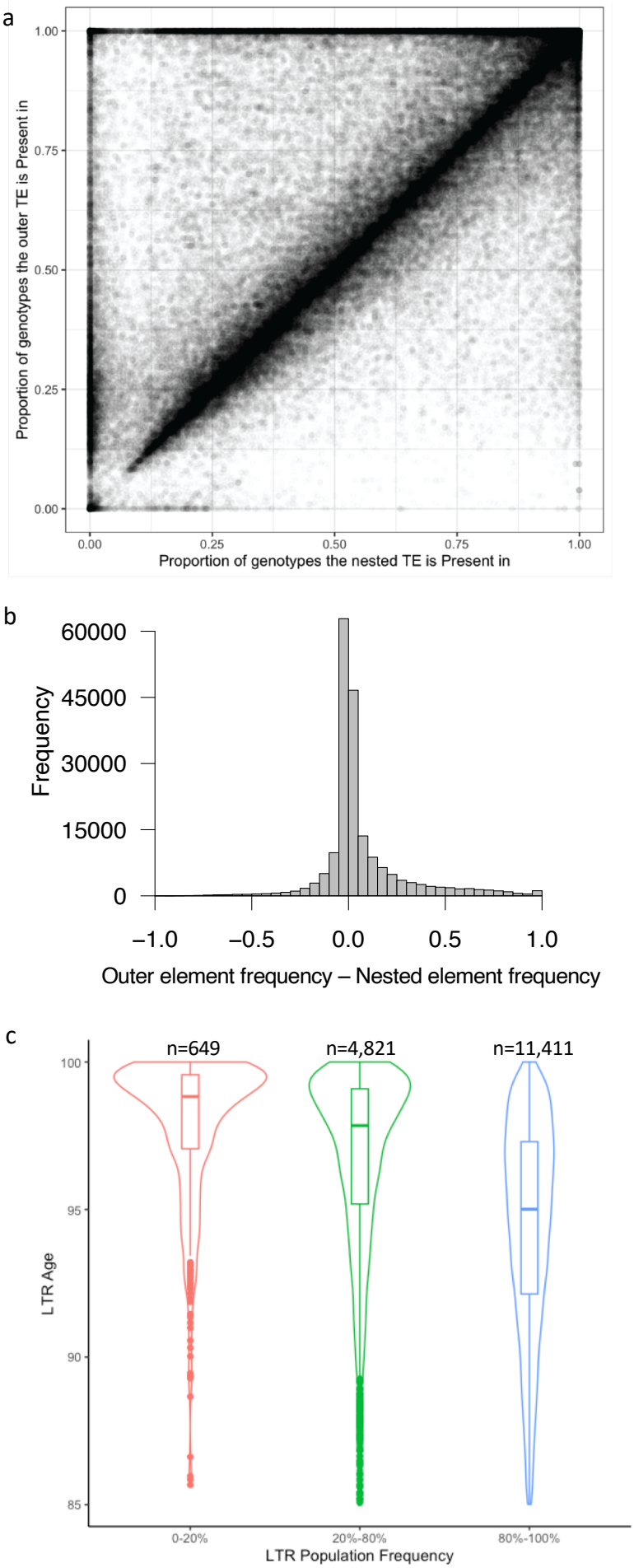

Figure 4. Relationship between population frequency of nested elements and the elements in which they are nested. a) Proportion of genotypes a TE is present in between nested TEs and the TE in which the element is nested. Only low ambiguous TEs were included in this plot. Only TEs from the B73 genome are plotted. b) Distribution of the proportion of genotypes the outer TE is present in minus proportion of genotypes the nested TE is present in. c) Age distributions for nested elements that are nested in TEs that are fixed or nearly fixed (frequency >0.95) in the population. LTR similarity is used as a proxy to age which higher LTR similarity corresponds to a younger TE and lower LTR similarity corresponds to an older TE. This plot only contains nested LTRs as age estimates are not available for other orders.

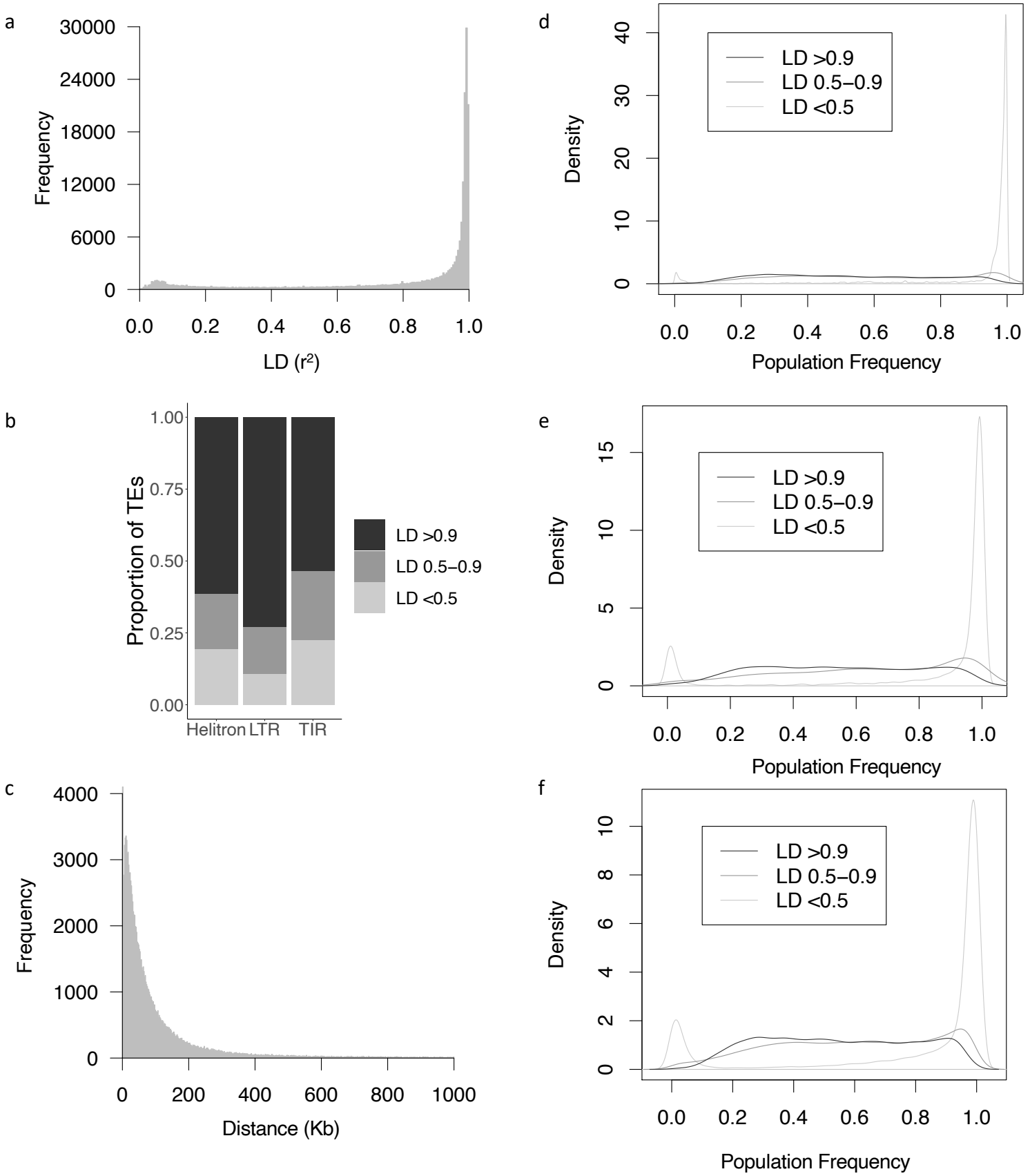

Figure 5. Linkage disequilibrium between TEs and SNPs in a panel of diverse inbred lines. a) Linkage disequilibrium (LD) between TEs and the SNP with the highest LD within 1Mb of the middle of the TE. b) Proportion of TEs in high ($r^2 > 0.9$), moderate (r^2 0.5-0.9), and low ($r^2 < 0.5$) LD with SNPs within 1 Mb of the middle of the TE. Category is based on the SNP with the highest LD in the window. c) Distance between TEs and the SNP with the highest LD to it for TEs that had a SNP in high ($r^2 > 0.9$) LD. Distance is calculated as the middle of the TE to the SNP. Only SNPs within 1 Mb of a TE were evaluated. d-f) Density plots of population frequencies for TEs in high, moderate, and low LD with SNPs based on the highest LD within 1Mb of the middle of the TE for LTRs (d), Helitrons (e), and TIRs (f).

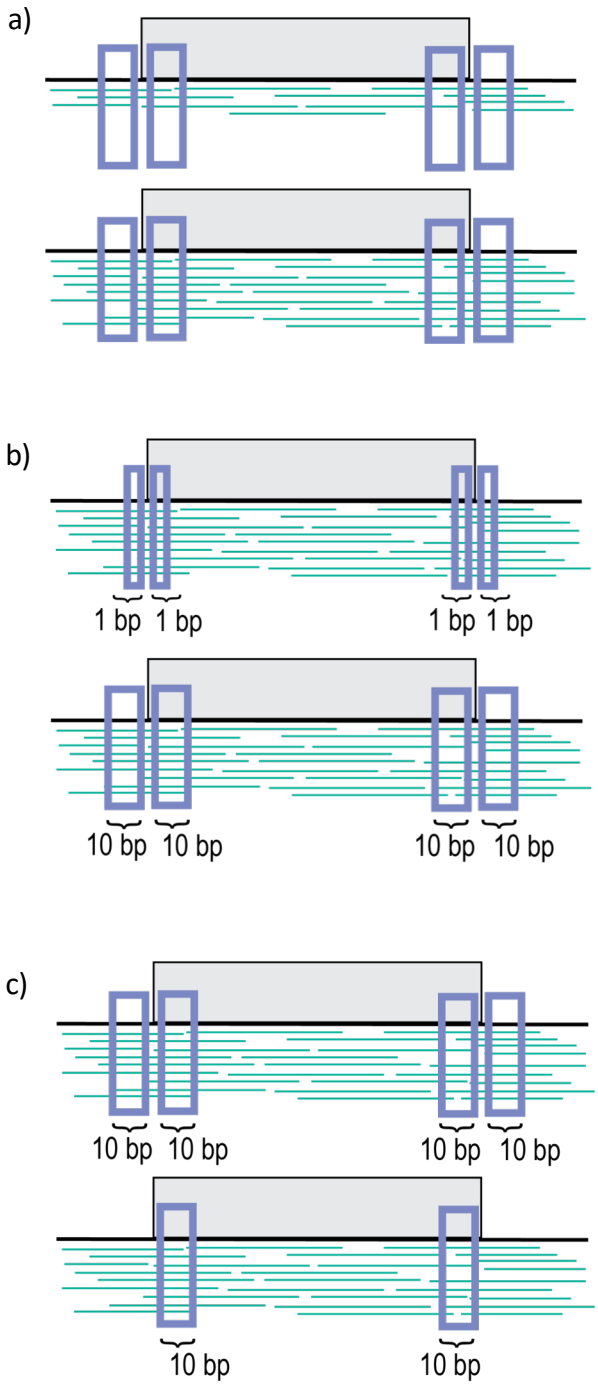

Figure S1. Schematic of factors tested for accuracy of TE calling using overage of short read sequence alignments in windows surrounding annotated transposable elements. a) variable coverage within windows to call a TE as present. b) variable window sizes. c) internal and external flanking windows versus only internal windows.

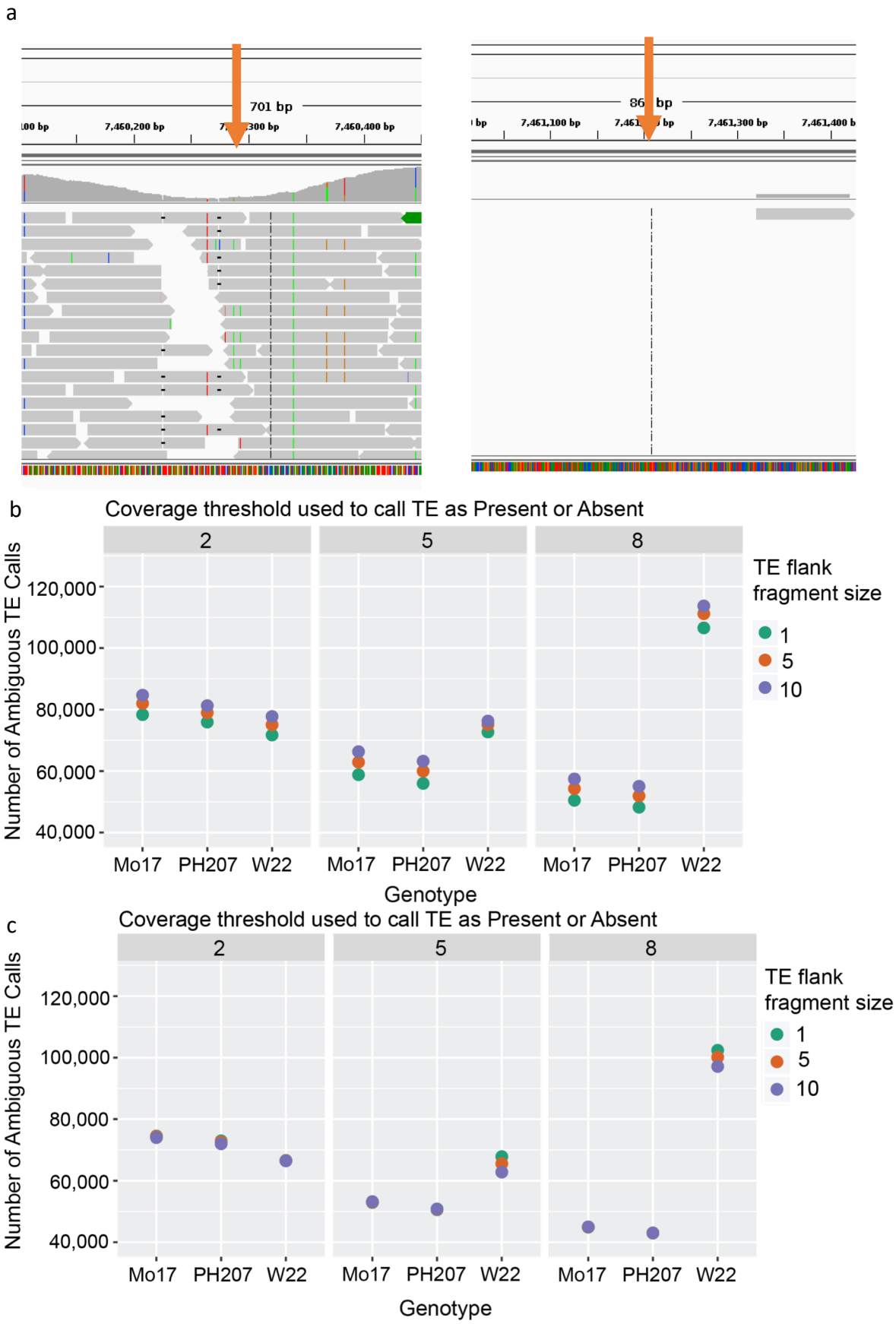

Figure S2. Rate of ambiguous calls at variable coverage thresholds. a) Example showing inconsistent coverage of B73 reads mapped to the Mo17 genome that results in an ambiguous call. b) Number of ambiguous TE calls using four flanking windows for each TE. c) Number of ambiguous TE calls using only the inner two flanking windows.

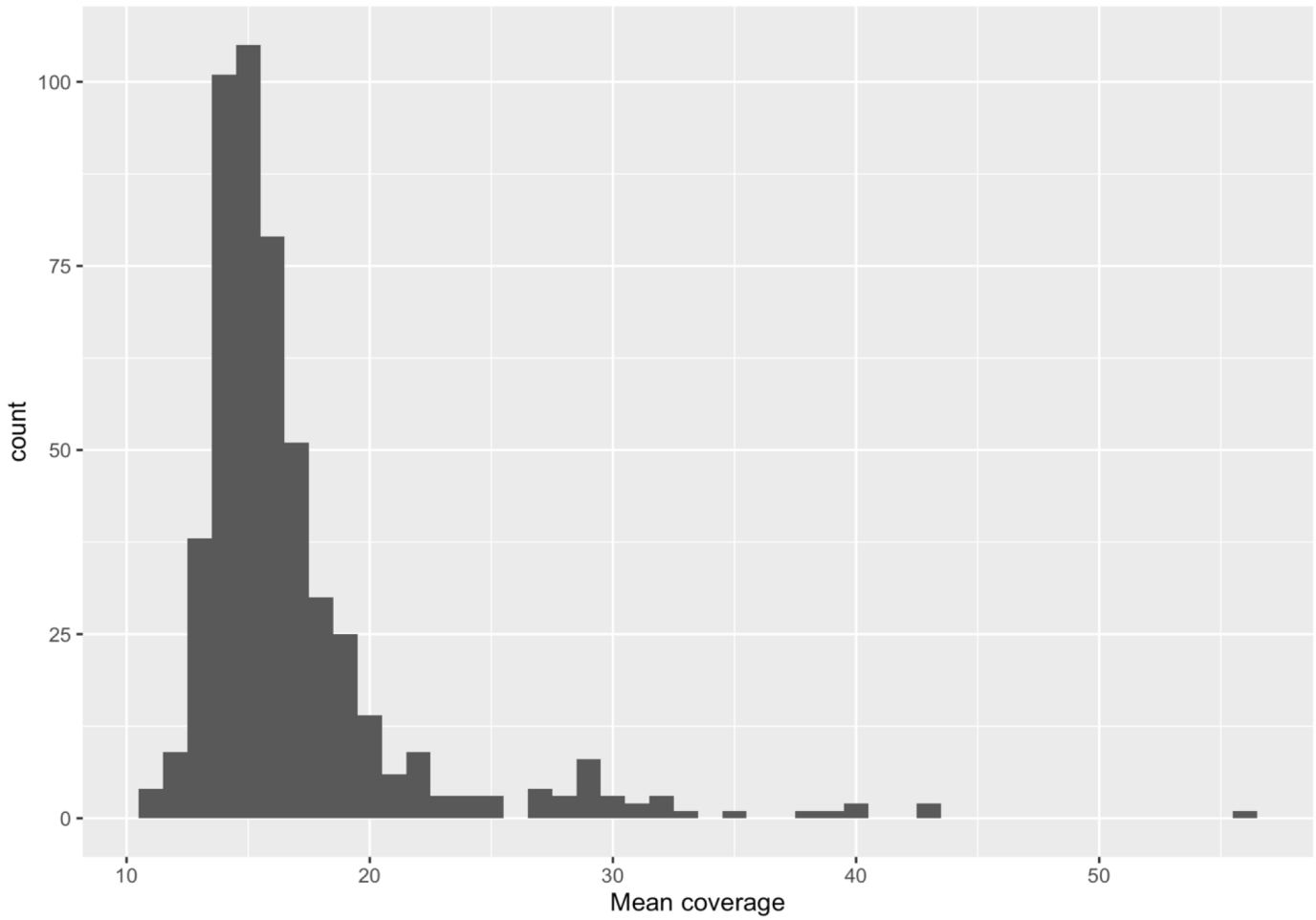

Figure S3. Realized genome-wide mean coverage for samples used in this study.

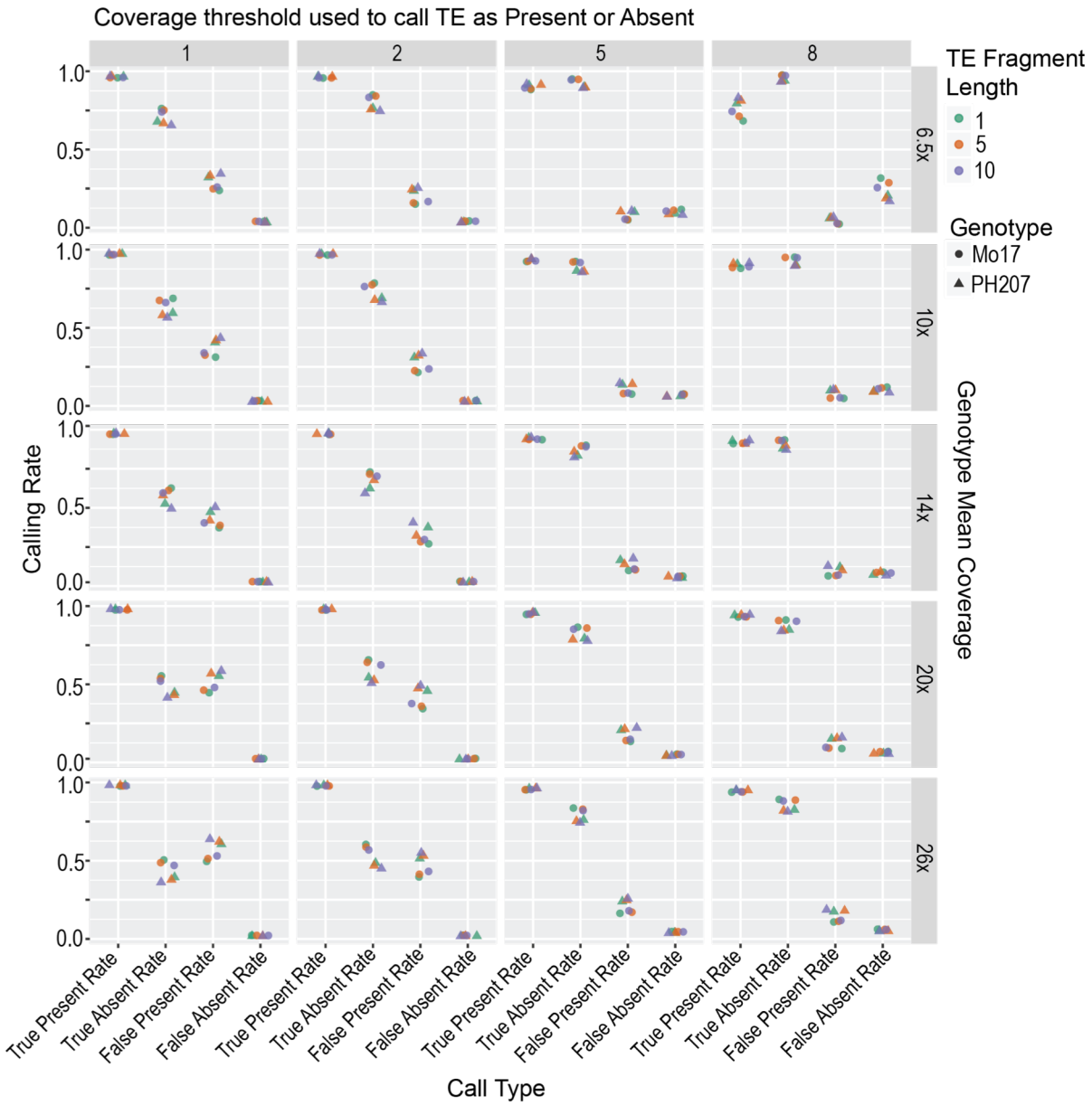

Figure S4. TE calling accuracy for Mo17 and PH207 down sampled to different coverage levels. Coverage over the two internal flanking windows were used to call a TE as Present, Absent, or Ambiguous. Ambiguous calls were exclude from TE calling rates. The majority of samples used in this study have realized coverage between 10-30x.

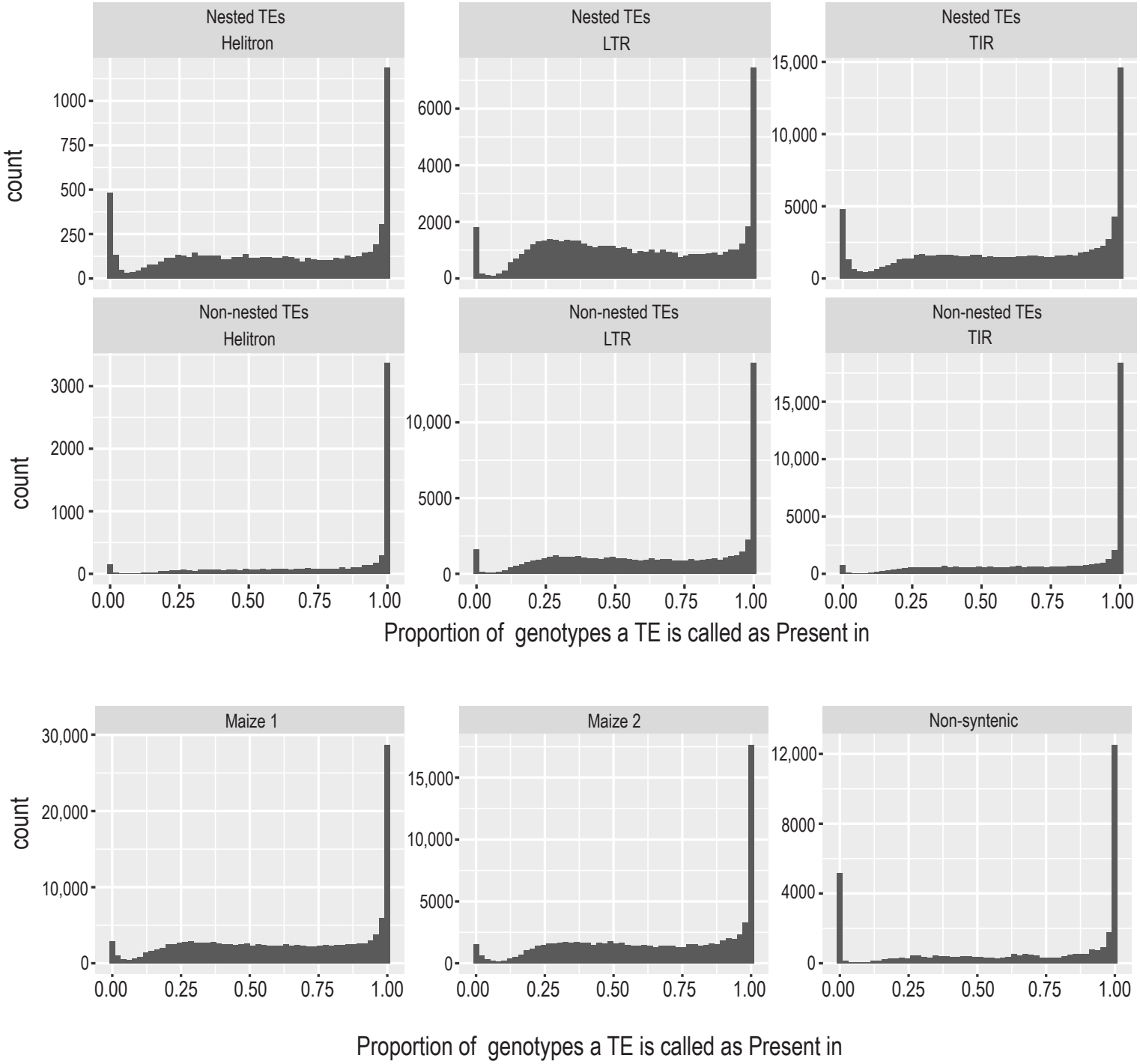

Figure S5. TE frequency distribution in a panel of diverse maize inbred lines. Genotypes include 508 individuals for the Wisconsin Diversity Panel (Hansey et al., 2010). Frequencies are for TEs annotated in the B73 v4 reference genome assembly (Anderson et al., 2019).

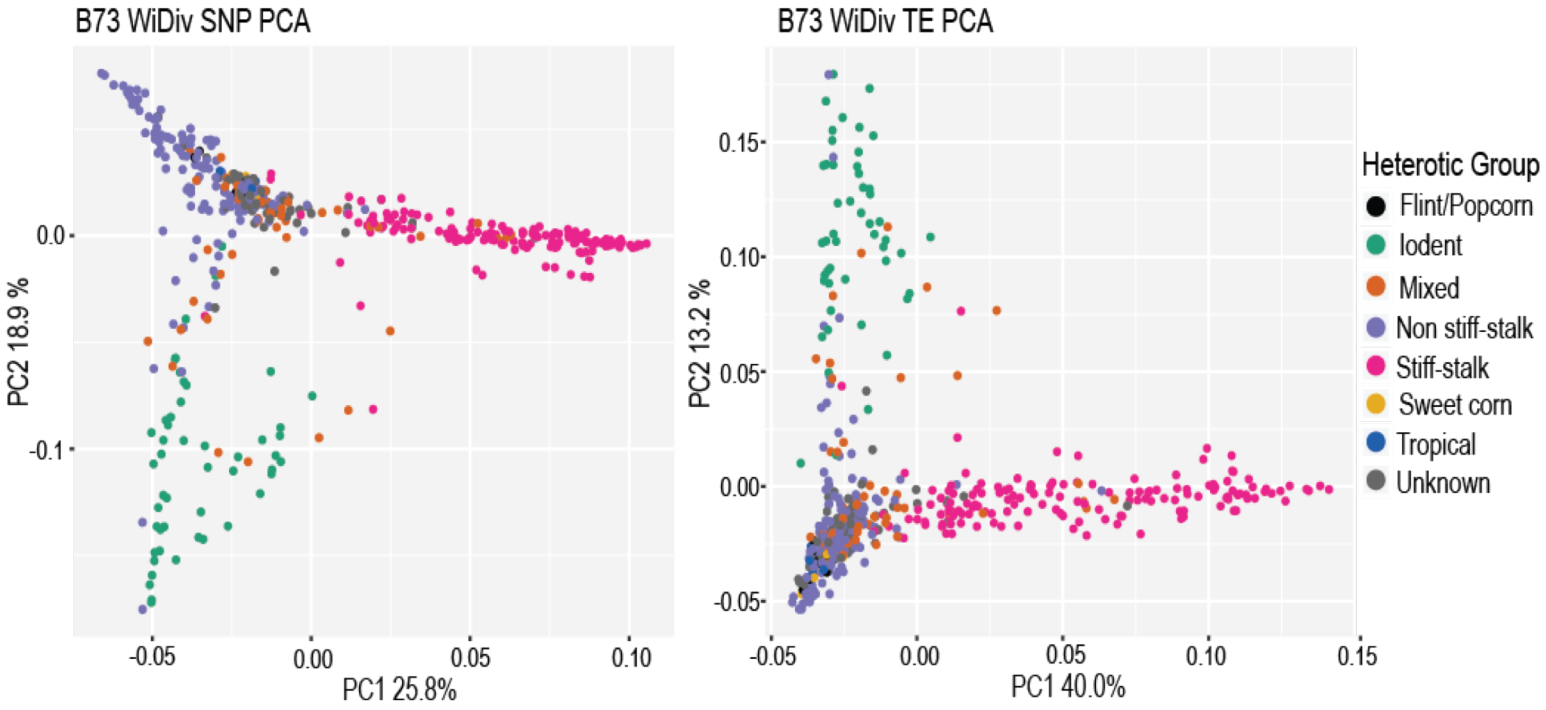

Figure S6. Principle component analysis (PCA) of SNPs (left) and TE presence/absence (right) in a set of diverse maize inbred lines. TEs and SNPs were called relative to the B73 reference genome assembly.

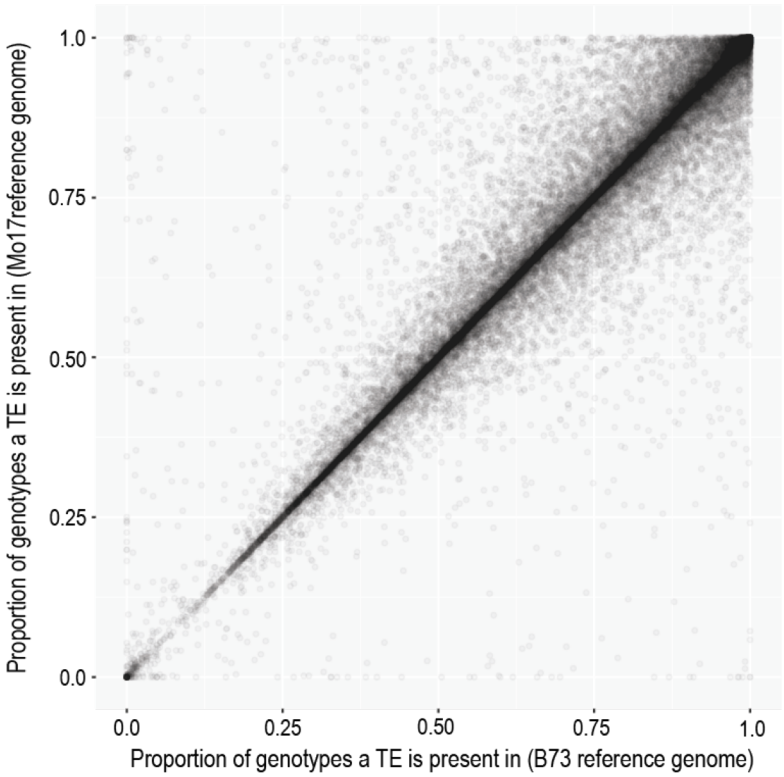

Figure S7. Proportion of genotypes a TE was called as 'Present' in between homologous TEs in B73 and Mo17. Only low ambiguous TEs included.

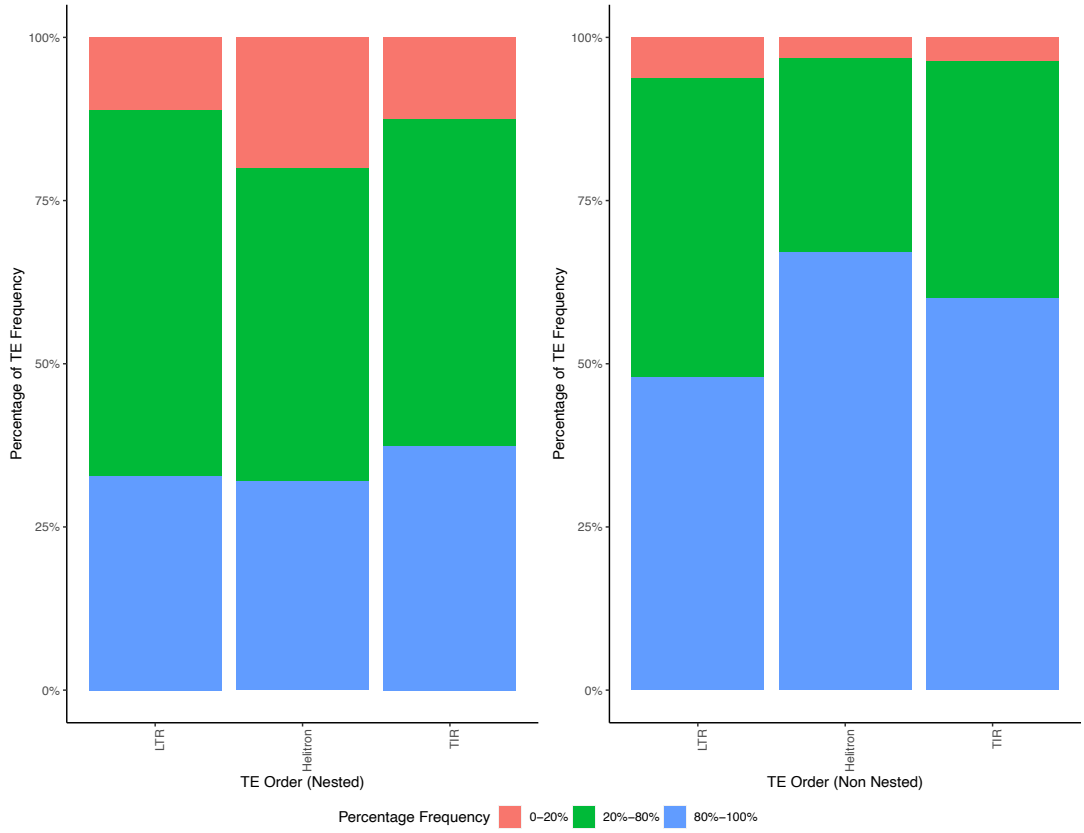

Figure S8. Proportion of low, medium, and high frequency TEs in a panel of diverse inbred lines. Genotypes include 508 individuals for the Wisconsin Diversity Panel (Hansey et al., 2010). Frequencies are for the set of non-redundant TEs across four genome assemblies (Anderson et al., 2019).

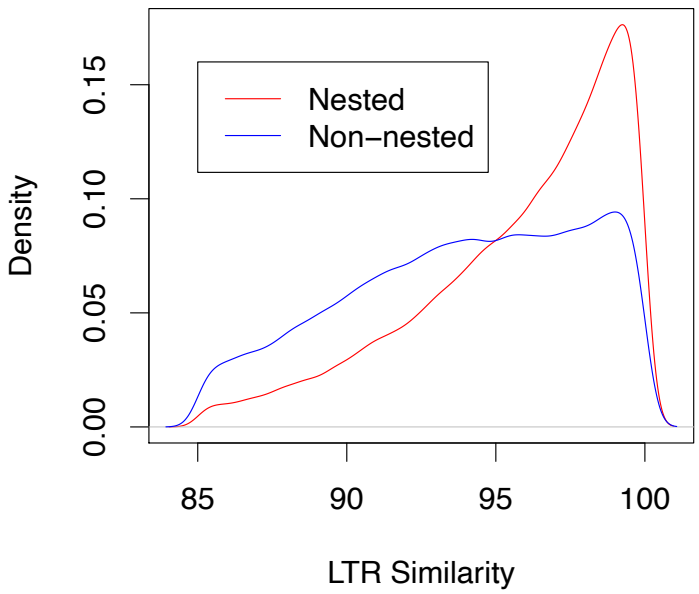

Figure S9. Age distribution of nested and non-nested elements in a non-redundant set of 202,267 LTR elements. Age of elements is approximated based on LTR similarity for each TE.

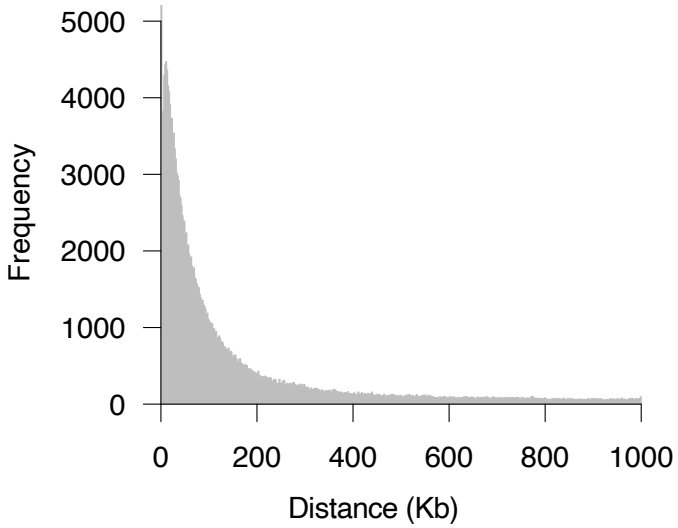

Figure S10. Distance between TEs and the SNP with the highest LD to it. Distance is calculated as the middle of the TE to the SNP. Only SNPs within 1Mb of a TE were evaluated.