



Figures and figure supplements

Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation

Tim Stuart *et al*

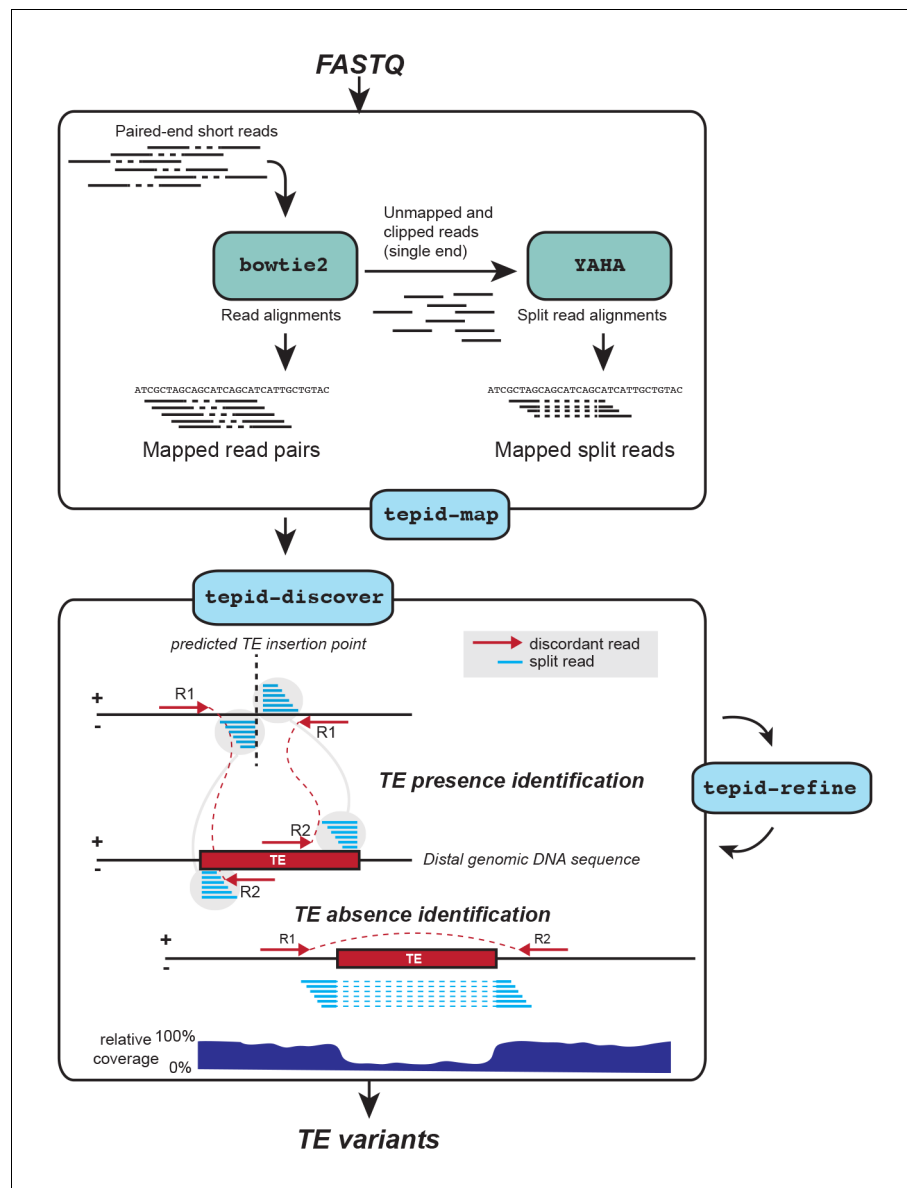


Figure 1. TE variant discovery pipeline. Principle of TE variant discovery using split and discordant read mapping positions. Paired end reads are first mapped to the reference genome using Bowtie2 [Langmead and Salzberg, 2012]. Soft-clipped or unmapped reads are then extracted from the alignment and re-mapped using Yaha, a split read mapper [Faust and Hall, 2012]. All read alignments are then used by TEPIID to discover TE variants relative to the reference genome, in the 'tepid-discover' step. When analyzing groups of related samples, these variants can be further refined using the 'tepid-refine' step, which examines in more detail the genomic regions where there was a TE variant identified in another sample, and calls the same variant for the sample in question using lower read count thresholds as compared to the 'tepid-discover' step, in order to reduce false negative variant calls within a group of related samples.

DOI: [10.7554/eLife.20777.002](https://doi.org/10.7554/eLife.20777.002)

The following source data is available for figure 1:

Source data 1. TE presences in Ler.

DOI: [10.7554/eLife.20777.003](https://doi.org/10.7554/eLife.20777.003)

Source data 2. TE absences in Ler.

DOI: [10.7554/eLife.20777.004](https://doi.org/10.7554/eLife.20777.004)

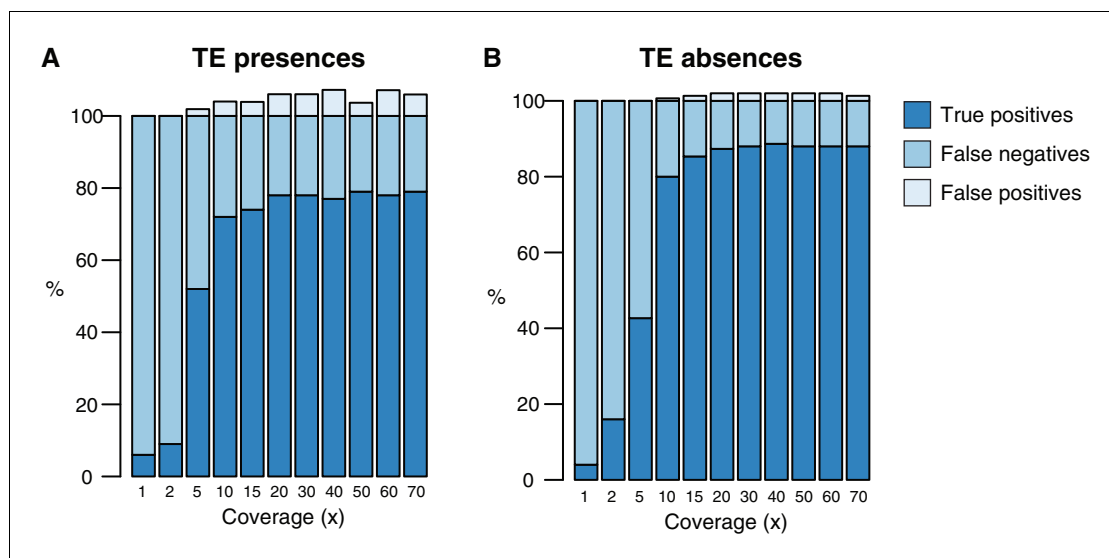


Figure 1—figure supplement 1. Testing of the TEPID pipeline using simulated TE variants in the Arabidopsis Col-0 genome (TAIR10), for a range of sequencing coverage levels. TE presence variants (A) and TE absence variants (B).

DOI: [10.7554/eLife.20777.005](https://doi.org/10.7554/eLife.20777.005)

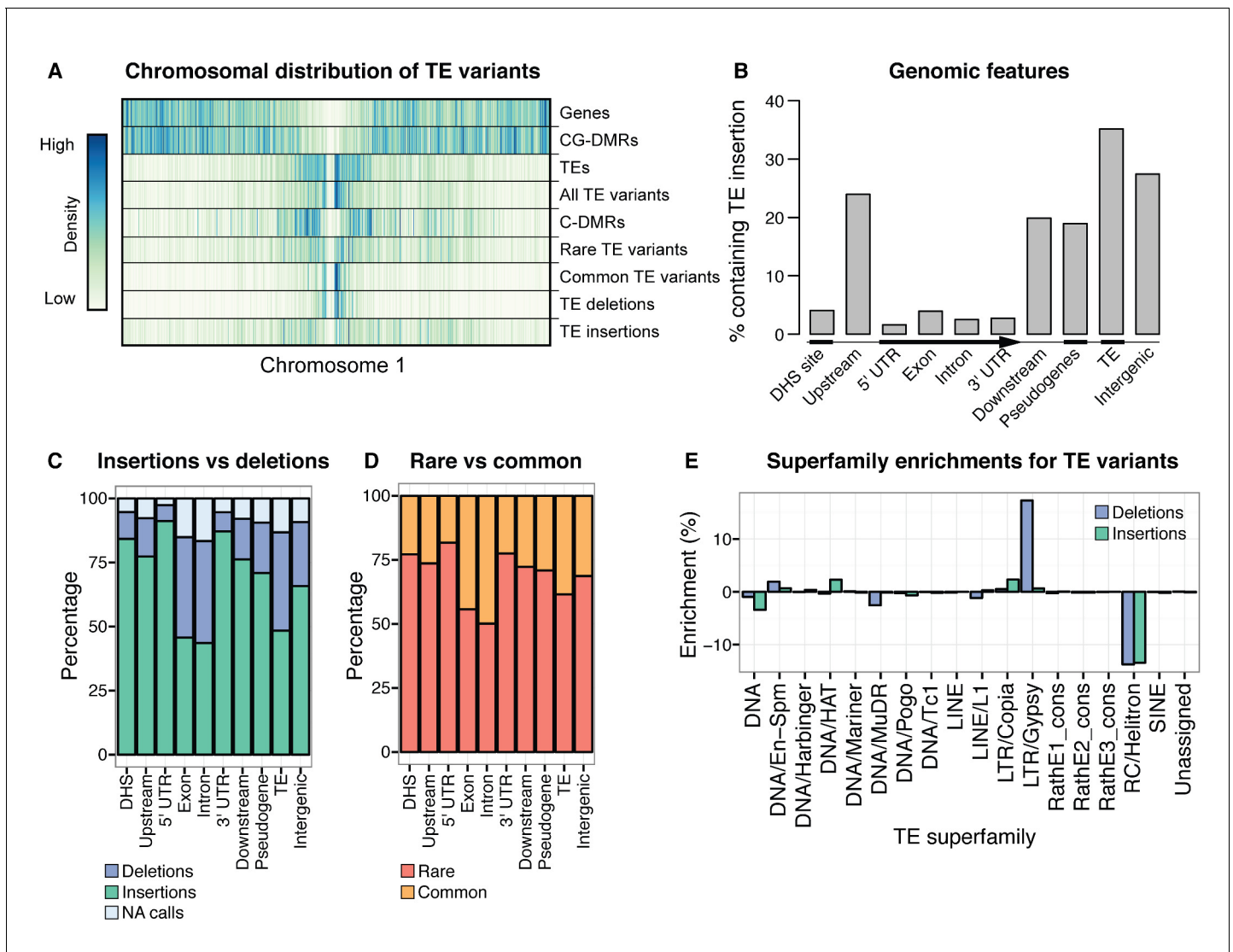


Figure 2. Extensive novel genetic diversity uncovered by TE variant analysis. (A) Distribution of identified TE variants on chromosome 1, with distributions of all Col-0 genes, Col-0 TEs, and population DMRs. (B) Proportion of different genomic features containing one or more TE variants. (C) Proportion of TE variants within each genomic feature classified as deletions or insertions. (D) Proportion of TE variants within each genomic feature classified as rare (<3% MAF) or common (≥3% MAF). (E) Enrichment and depletion of TE variants categorized by TE superfamily compared to the expected frequency due to genomic occurrence.

DOI: [10.7554/eLife.20777.008](https://doi.org/10.7554/eLife.20777.008)

The following source data is available for figure 2:

Source data 1. TE presence variants in all 216 Arabidopsis accessions.

DOI: [10.7554/eLife.20777.009](https://doi.org/10.7554/eLife.20777.009)

Source data 2. TE absence variants in all 216 Arabidopsis accessions.

DOI: [10.7554/eLife.20777.010](https://doi.org/10.7554/eLife.20777.010)

Source data 3. All TE variants.

DOI: [10.7554/eLife.20777.011](https://doi.org/10.7554/eLife.20777.011)

Source data 4. TE family enrichments for TE insertion and TE deletion variants.

DOI: [10.7554/eLife.20777.012](https://doi.org/10.7554/eLife.20777.012)

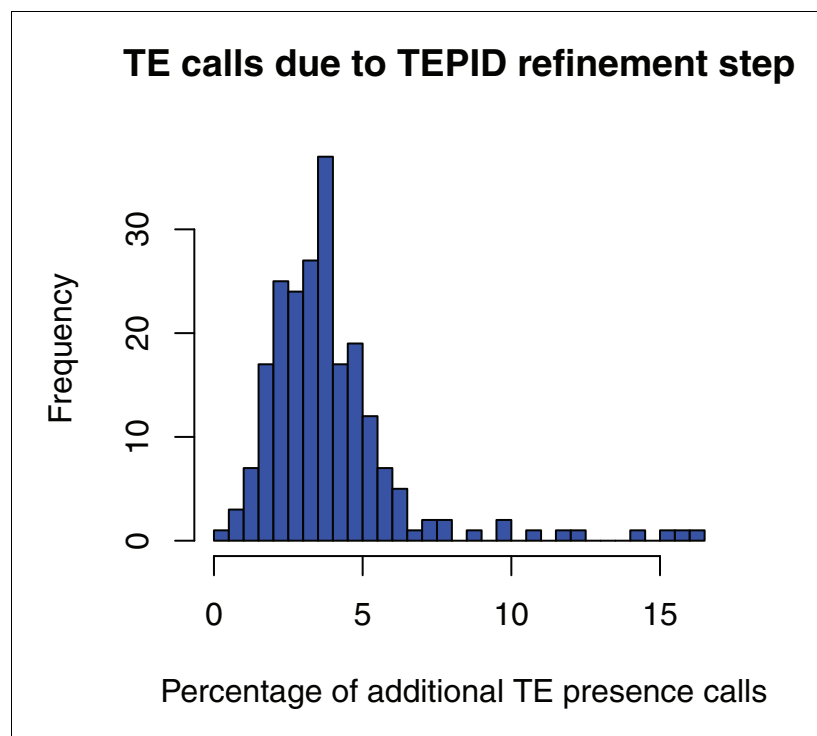


Figure 2—figure supplement 1. Percentage of total TE presence calls that were made due to the TEPID refinement step for each accession in the population.

DOI: [10.7554/eLife.20777.013](https://doi.org/10.7554/eLife.20777.013)

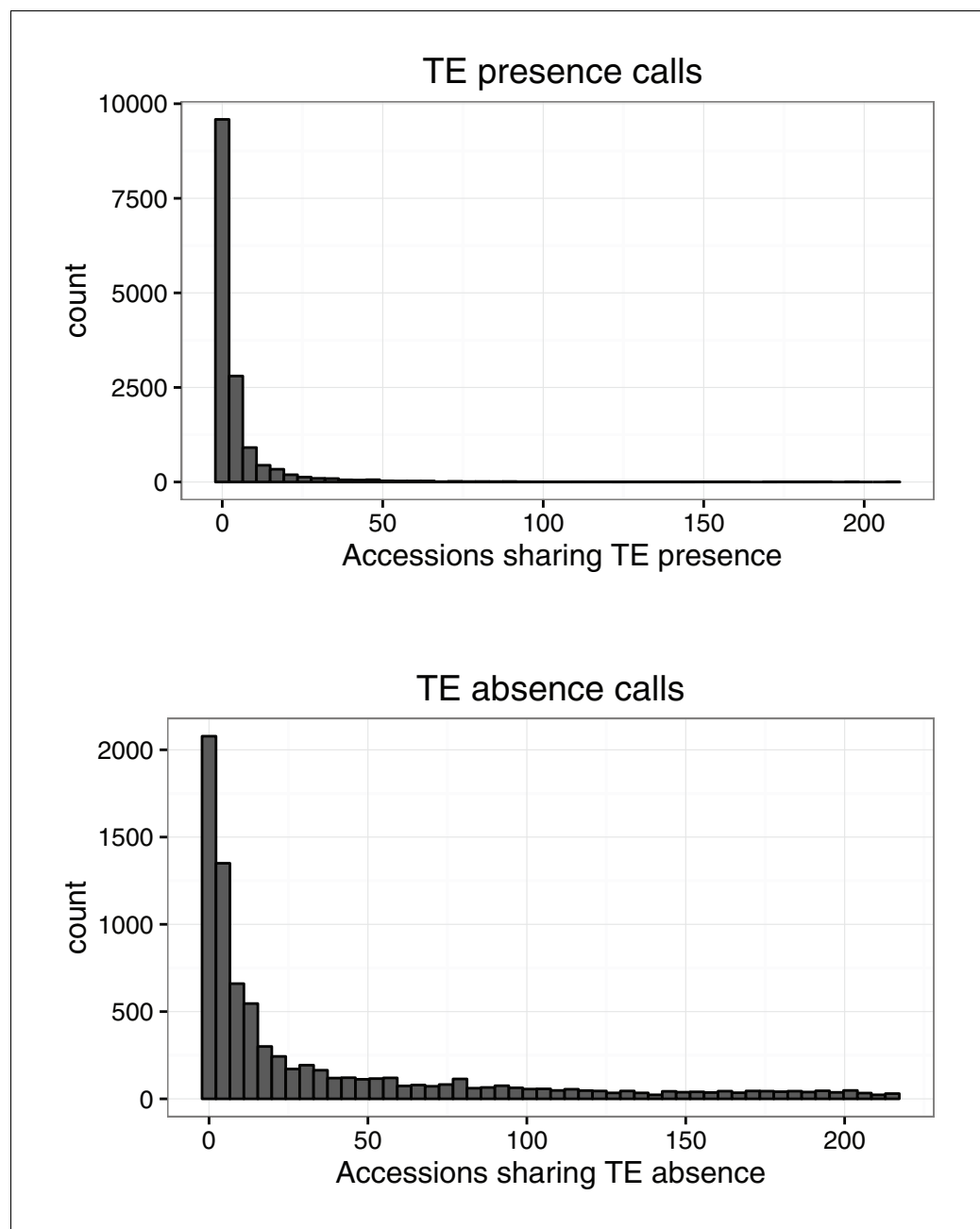


Figure 2—figure supplement 2. Number of accessions sharing TE variants identified by TEPID.

DOI: [10.7554/eLife.20777.014](https://doi.org/10.7554/eLife.20777.014)

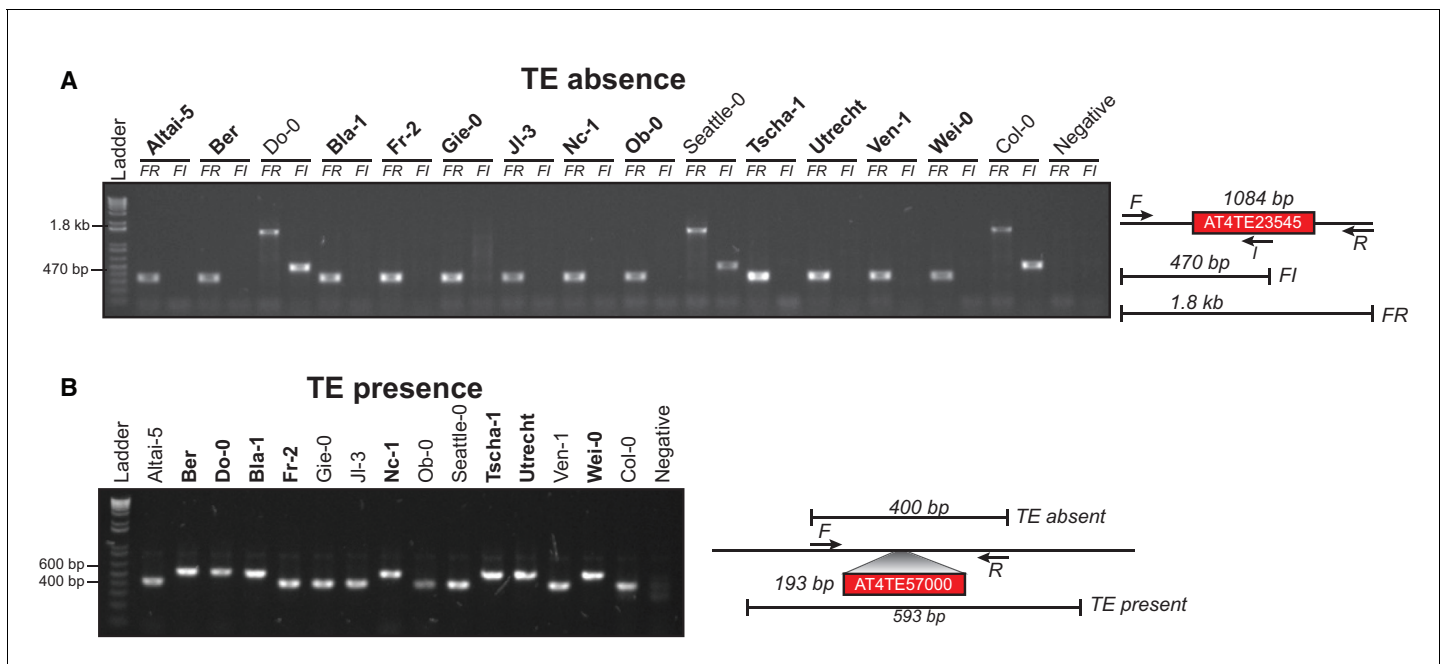


Figure 2—figure supplement 3. Example PCR validations for two TE variants. (A) PCR validations for a TE absence variant. Accessions that were predicted to contain a TE absence are marked in bold. Two primer sets were used; forward (F) and reverse (R) or internal (I). Accessions with a TE absence will not produce the FI band and produce a shorter FR band, with the change in size matching the size of the deleted TE. (B) PCR validations for a TE presence variant. Accessions that were predicted to contain a TE presence are marked in bold. One primer set was used, spanning the TE insertion site. A band shift of approximately 200 bp can be seen, corresponding to the size of the inserted TE.

DOI: [10.7554/eLife.20777.015](https://doi.org/10.7554/eLife.20777.015)

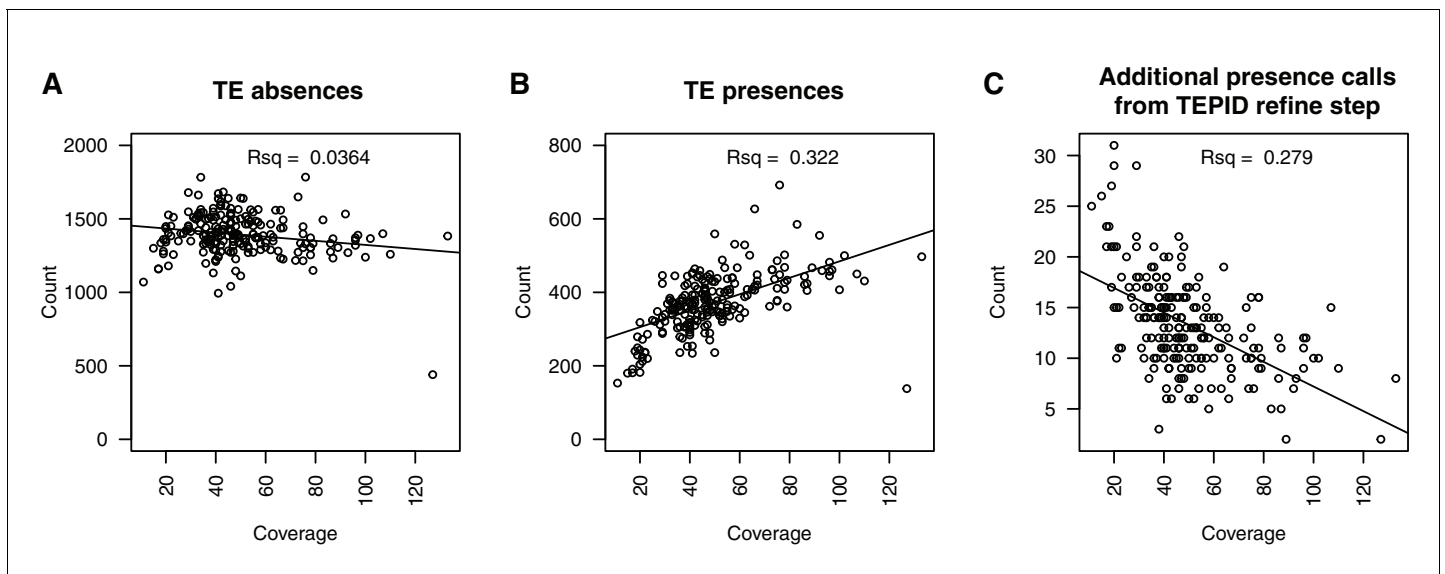


Figure 2—figure supplement 4. Relationship between sequencing depth and number of TE variants discovered in each accession. (A) Number of TE absence variants identified versus the sequencing depth of coverage for each accession. (B) Number of TE presence variants identified versus the sequencing depth of coverage for each accession. (C) Number of additional TE presence calls made due to the TEPID refinement step versus sequencing depth of coverage for all accessions.

DOI: [10.7554/eLife.20777.016](https://doi.org/10.7554/eLife.20777.016)

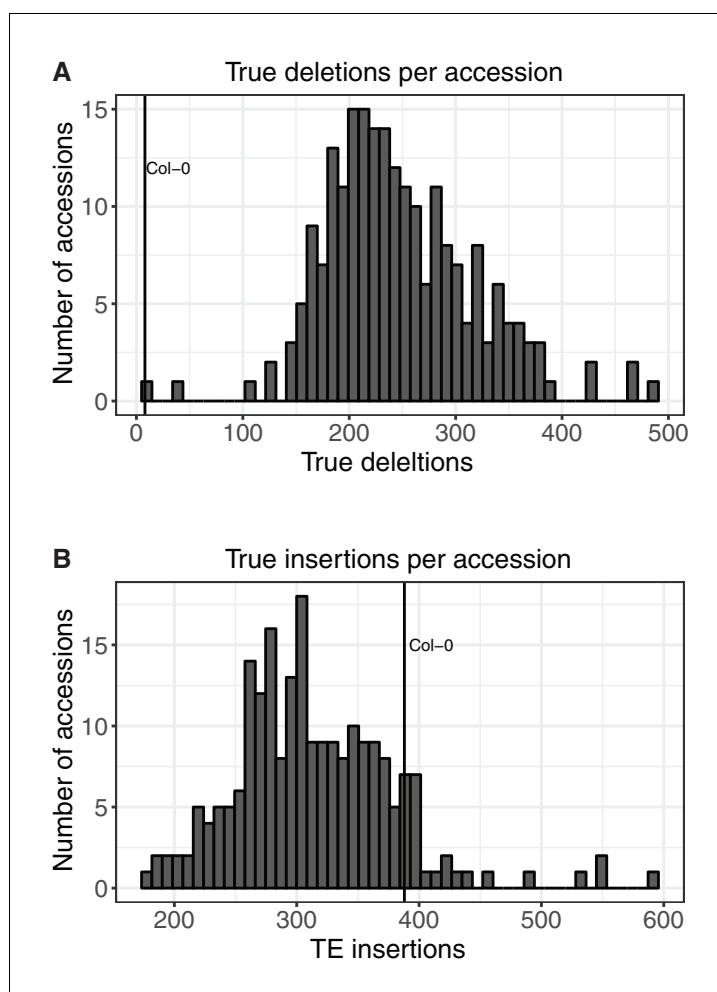


Figure 2—figure supplement 5. Number of TE insertions and TE deletions found in each accession. (A) Number of true TE deletions per accession. (B) Number of true TE insertion per accession.

DOI: [10.7554/eLife.20777.017](https://doi.org/10.7554/eLife.20777.017)

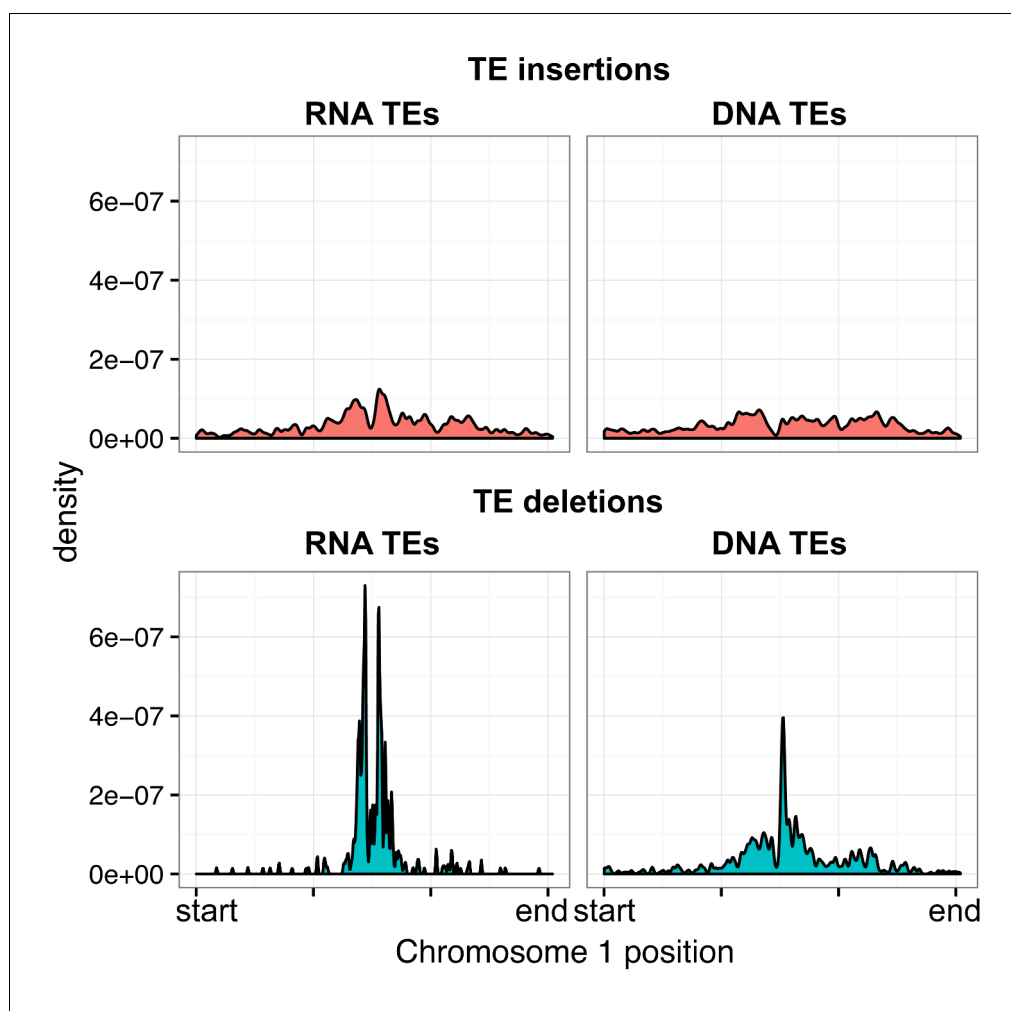


Figure 2—figure supplement 6. Distribution of RNA and DNA transposable elements over chromosome 1, for TE insertions and TE deletions.

DOI: [10.7554/eLife.20777.018](https://doi.org/10.7554/eLife.20777.018)

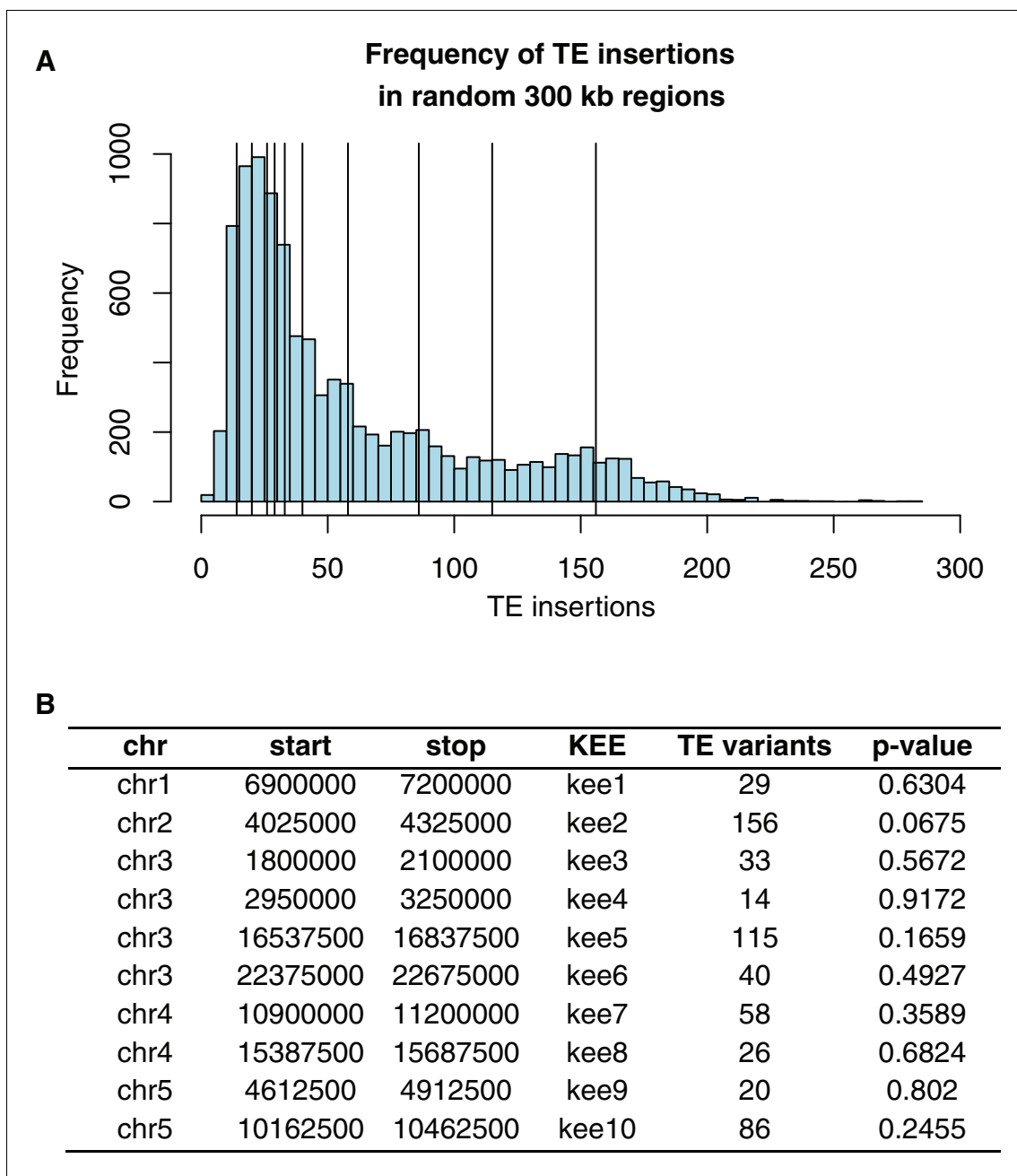


Figure 2—figure supplement 7. Frequency of TE insertion in the *KEE* regions. (A) Number of TE insertion variants within each 300 kb *KNOT ENGAGED ELEMENT* (*KEE*, vertical lines) and the number of TE insertion variants found in 10,000 randomly selected 300 kb windows (histogram). (B) Table showing number of TE insertion variants within each *KEE* region, and the associated p-value determined by resampling 10,000 times.

DOI: [10.7554/eLife.20777.019](https://doi.org/10.7554/eLife.20777.019)

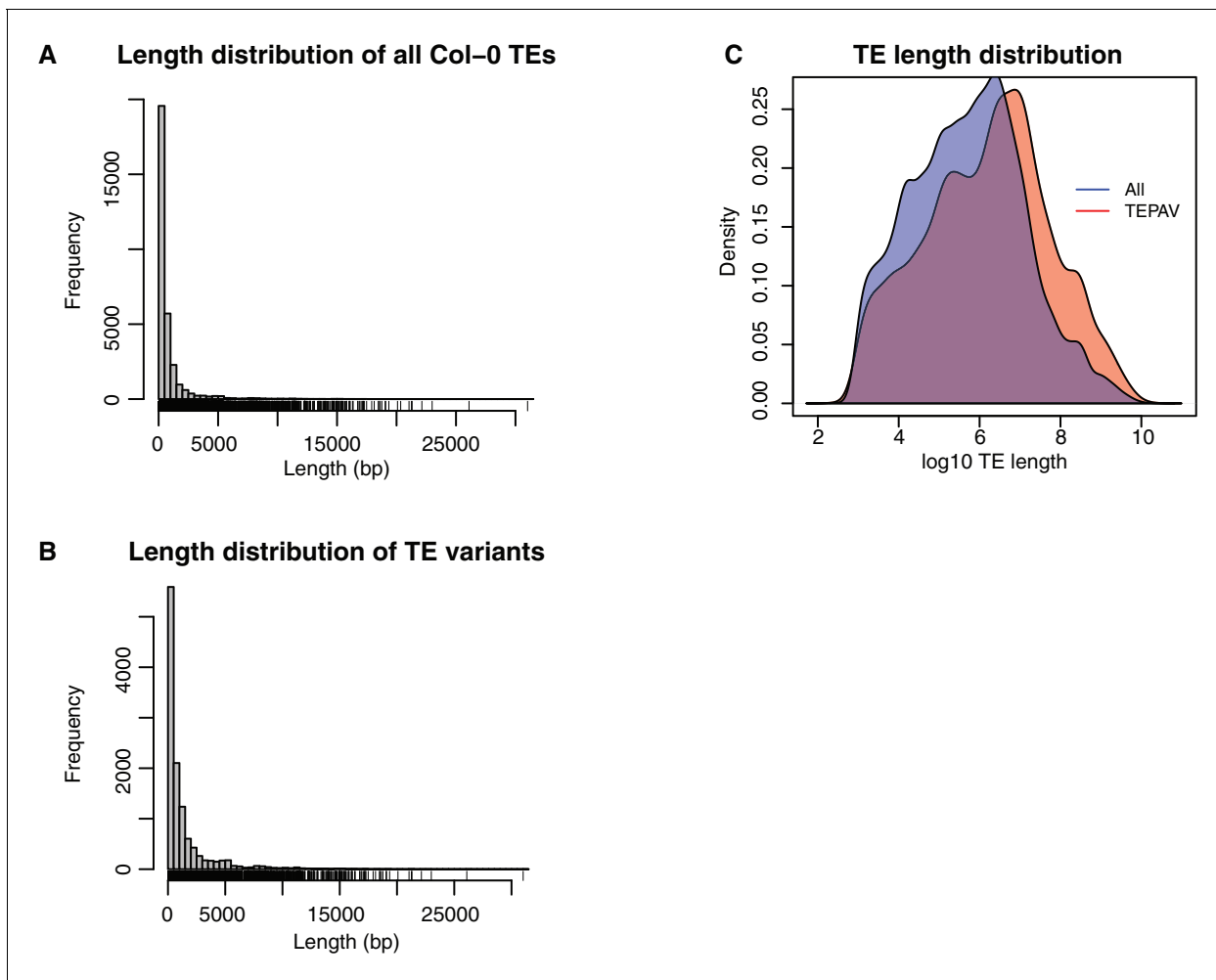


Figure 2—figure supplement 8. Length distribution for all Col-0 TEs and all TE variants. (A) Length distribution for all annotated TEs in the Col-0 reference genome. (B) Length distribution for all TE variants. (C) Density distribution of log₁₀ TE length for all Col-0 TEs (red) and TE variants (blue). DOI: [10.7554/eLife.20777.020](https://doi.org/10.7554/eLife.20777.020)

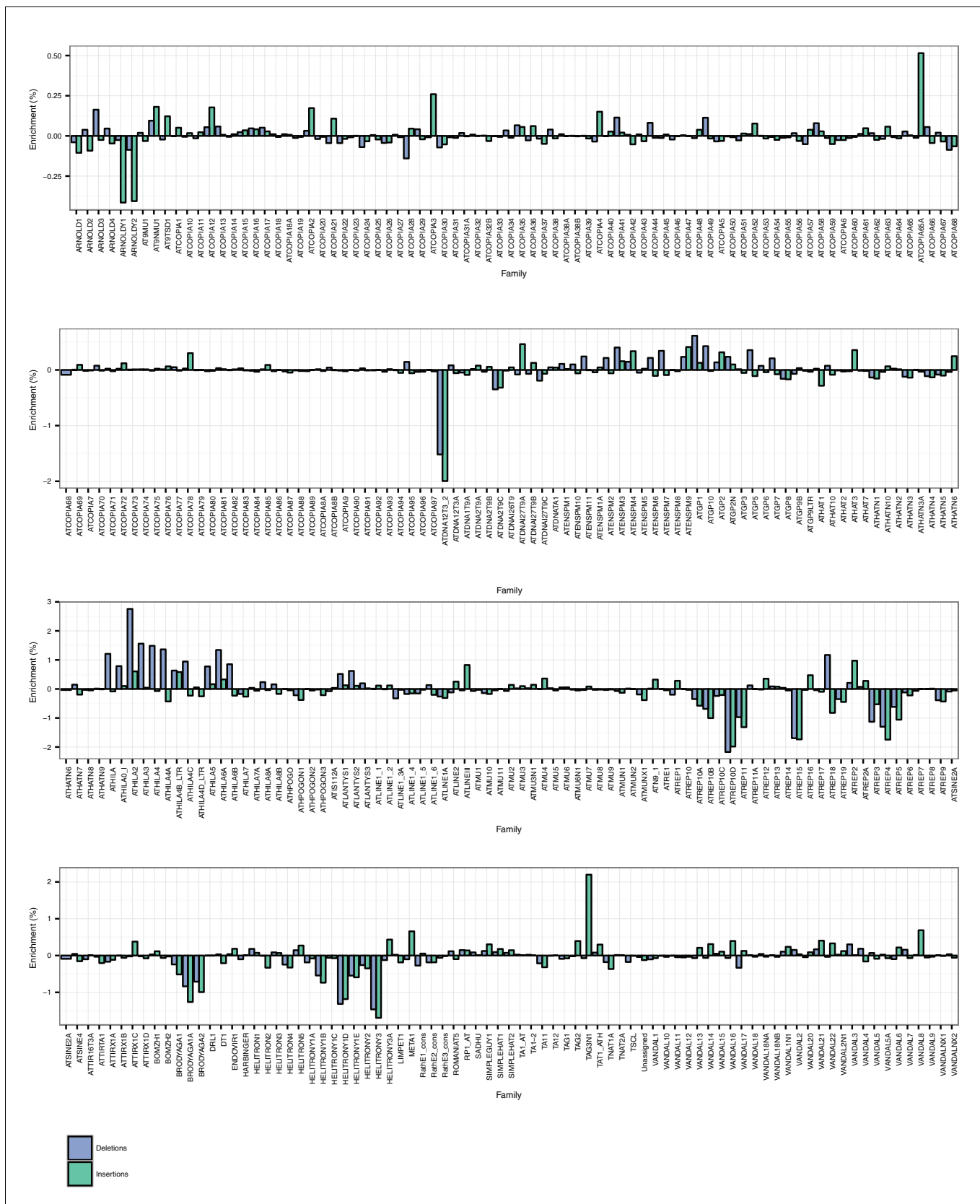


Figure 2—figure supplement 9. TE family enrichments and depletions for TE insertions and TE deletions.

DOI: 10.7554/eLife.20777.021

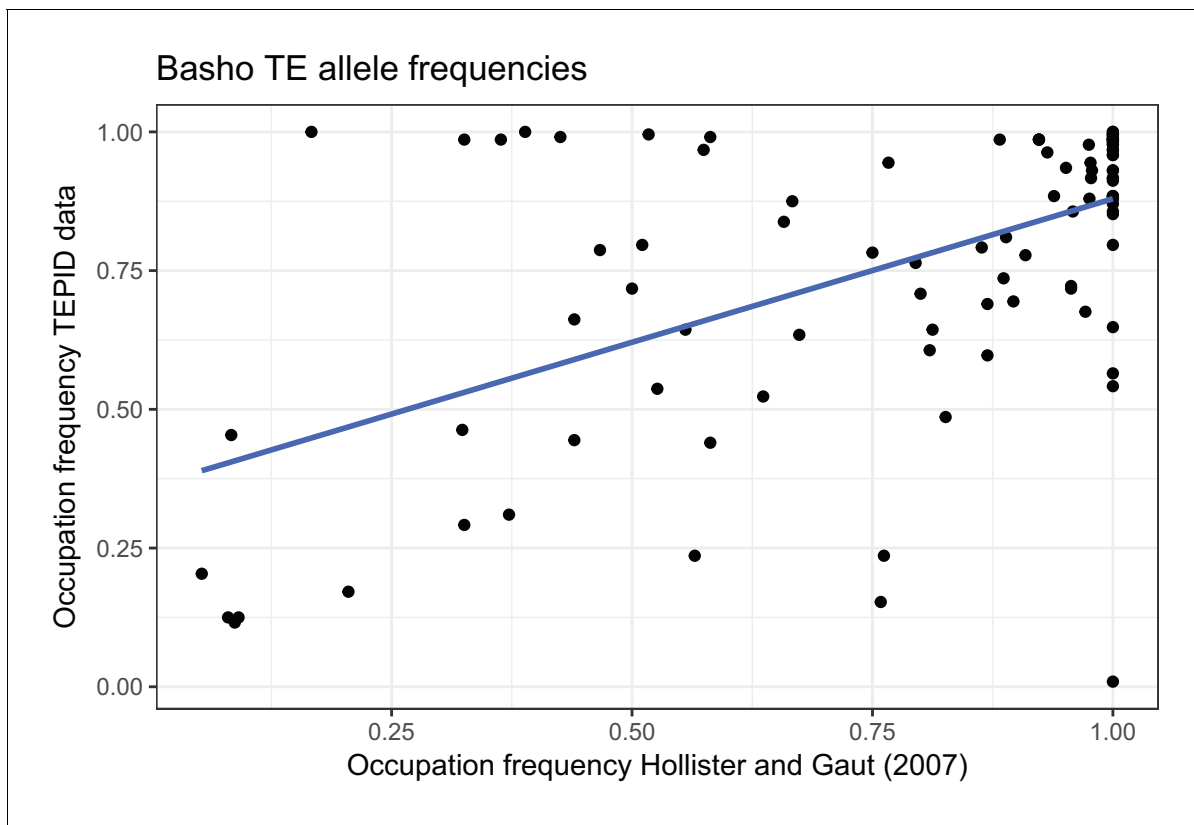


Figure 2—figure supplement 10. TE occupation frequencies for *Basho* TEs previously genotyped by (Hollister and Gaut, 2007).

DOI: [10.7554/eLife.20777.022](https://doi.org/10.7554/eLife.20777.022)

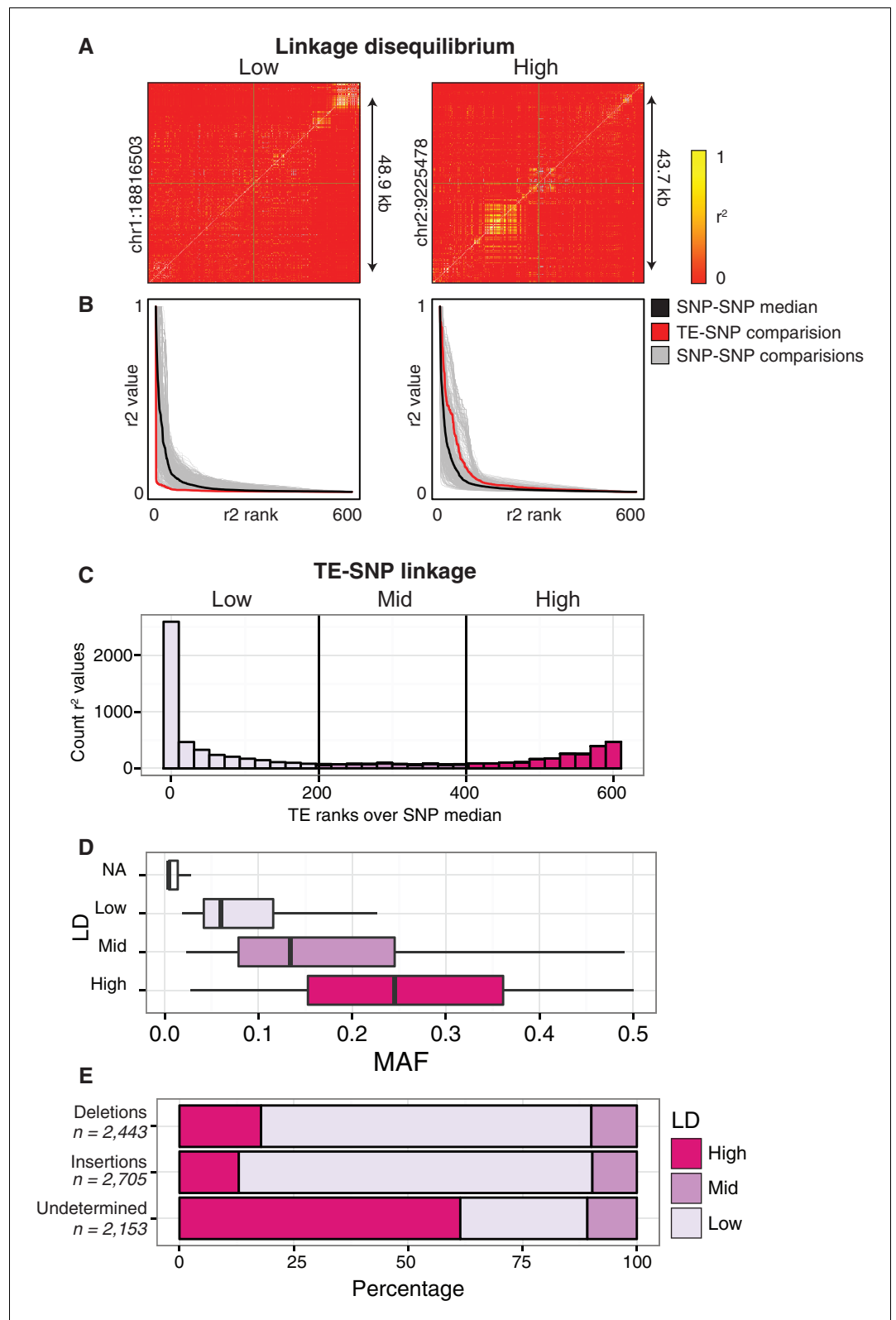


Figure 3. Patterns of TE-SNP linkage. (A) r^2 correlation matrices for individual representative high and low-LD TE variants showing the background level of SNP-SNP linkage. (B) Rank order plots for individual representative high and low-LD TE variants (matching those shown in A). Red line indicates the median r^2 value for each rank across SNP-based values. Blue line indicates r^2 values for TE-SNP comparisons. Grey lines indicate all individual SNP-SNP

Figure 3 continued on next page

Figure 3 continued

comparisons. (C) Histogram of the number of TE r^2 ranks (0-600) that are above the SNP-based median r^2 value for common TE variants. (D) Boxplots showing distribution of minor allele frequencies for each LD category. Boxes represent the interquartile range (IQR) from quartile 1 to quartile 3. Boxplot upper whiskers represent the maximum value, or the upper value of the quartile 3 plus 1.5 times the IQR (whichever is smaller). Boxplot lower whisker represents the minimum value, or the lower value of the quartile 1 minus 1.5 times the IQR (whichever is larger). (E) Proportion of TE insertions, TE deletions, and unclassified TE variants in each LD category.

DOI: [10.7554/eLife.20777.023](https://doi.org/10.7554/eLife.20777.023)

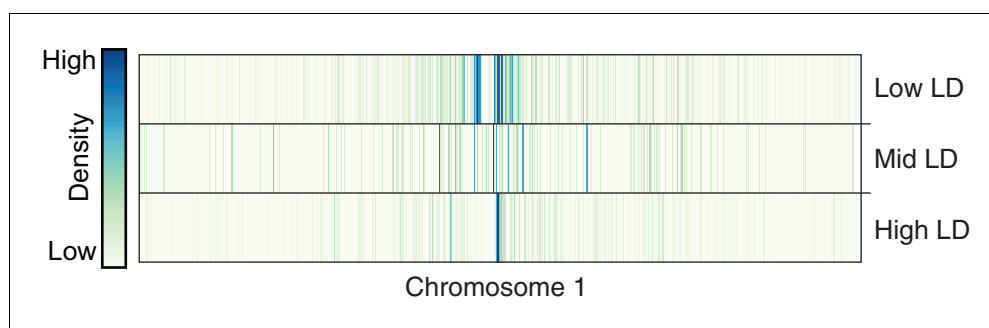


Figure 3—figure supplement 1. Distribution of TE variants across chromosome 1 for each LD category (high, mid, low).

DOI: [10.7554/eLife.20777.024](https://doi.org/10.7554/eLife.20777.024)

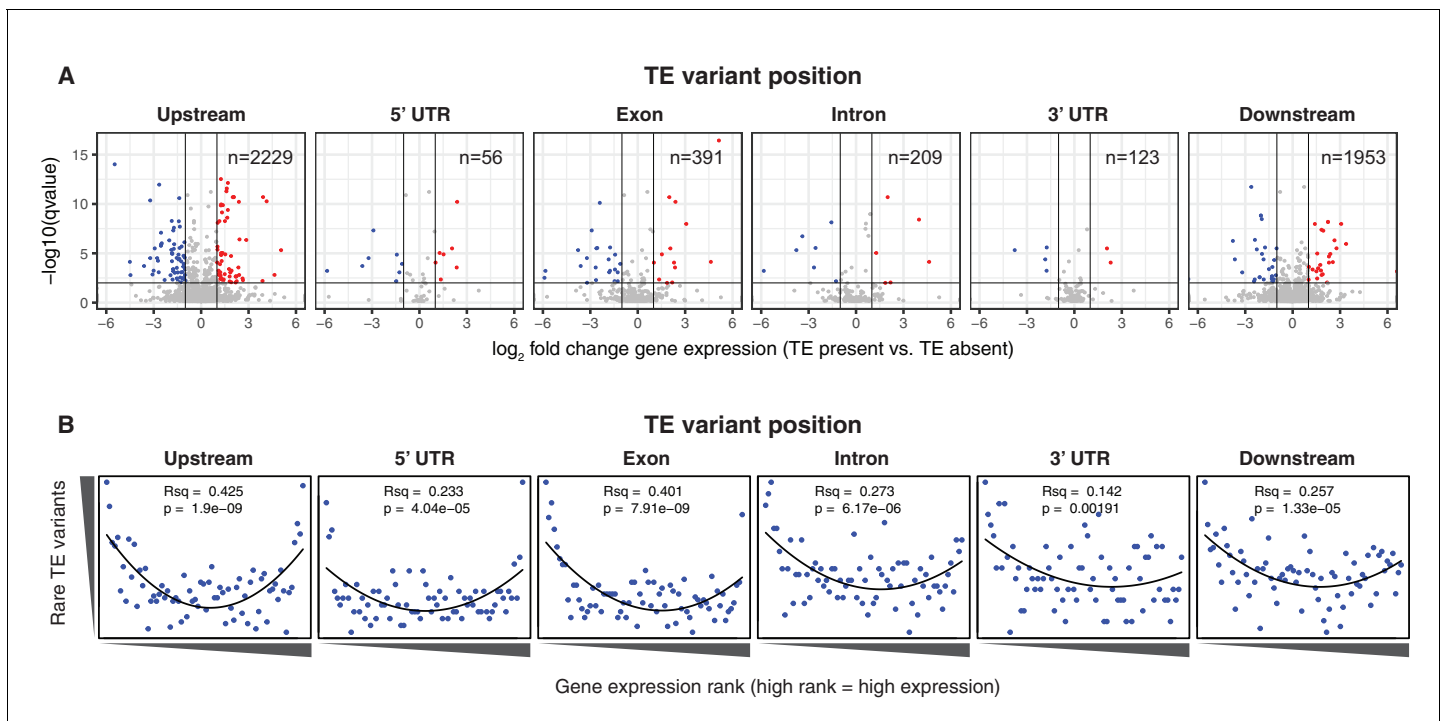


Figure 4. Differential transcript abundance associated with TE variant presence/absence. (A) Transcript abundance differences for genes associated with TE insertion variants at different positions, indicated in the plot titles. Genes with significantly different transcript abundance in accessions with a TE insertion compared to accessions without a TE insertion are colored blue (lower transcript abundance in accessions containing TE insertion) or red (higher transcript abundance in accessions containing TE insertion). Vertical lines indicate ± 2 fold change in FPKM. Horizontal line indicates the 1% false discovery rate. (B) Relationship between rare TE variant counts and gene expression rank. Cumulative number of rare TE variants in equal-sized bins for gene expression ranks, from the lowest-ranked accession (left) to the highest-ranked accession (right). Lines indicate the fit of a quadratic model.

DOI: [10.7554/eLife.20777.025](https://doi.org/10.7554/eLife.20777.025)

The following source data is available for figure 4:

Source data 1. Differentially expressed genes associated with TE presence/absence.

DOI: [10.7554/eLife.20777.026](https://doi.org/10.7554/eLife.20777.026)

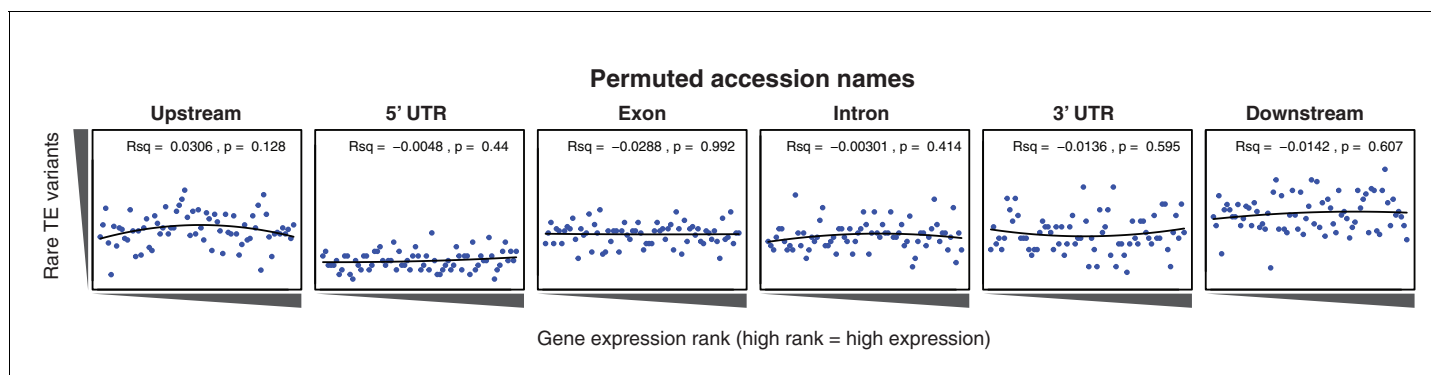


Figure 4—figure supplement 1. Relationship between rare TE variants and gene expression rank as for **Figure 4B** for permuted TE variants.

DOI: [10.7554/eLife.20777.027](https://doi.org/10.7554/eLife.20777.027)

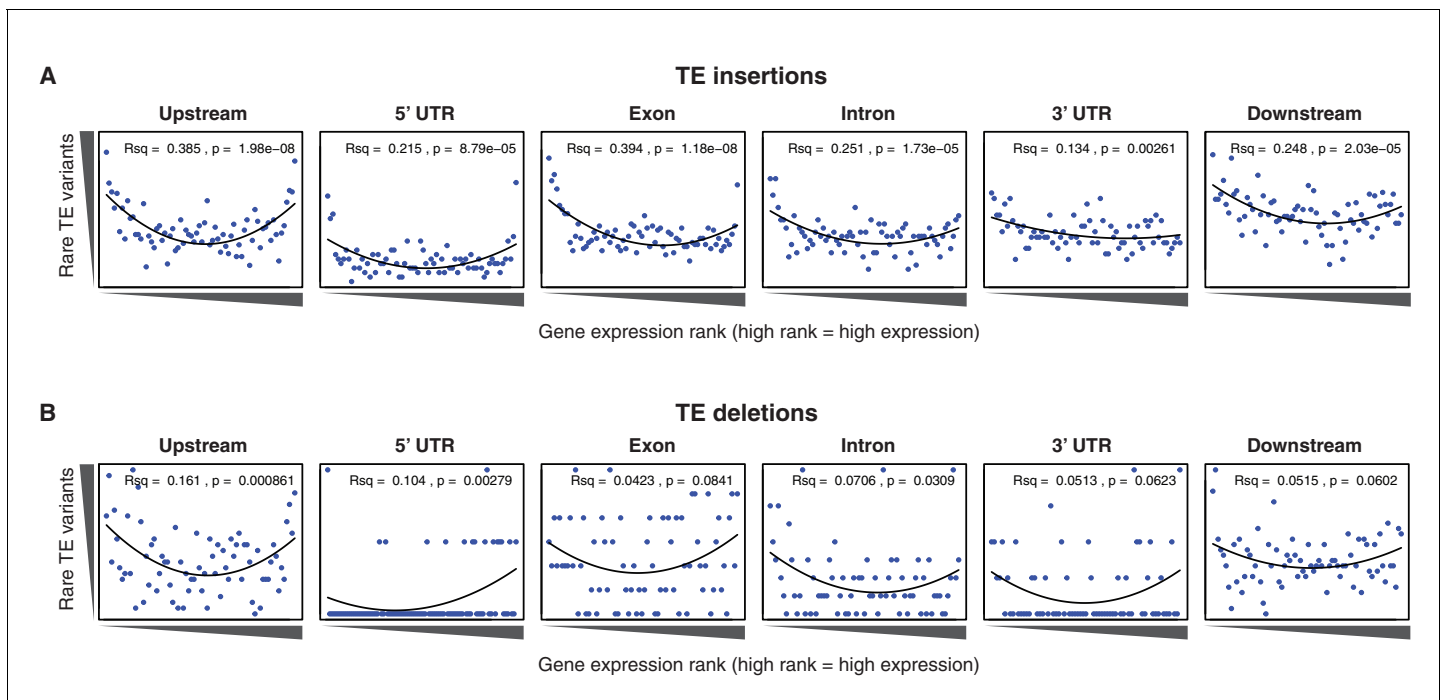


Figure 4—figure supplement 2. Relationship between rare TE variants and gene expression rank as for **Figure 4B** for TE insertions and TE deletions separately. (A) Burden of rare TE insertion variants.(B) Burden of rare TE deletion variants.

DOI: [10.7554/eLife.20777.028](https://doi.org/10.7554/eLife.20777.028)

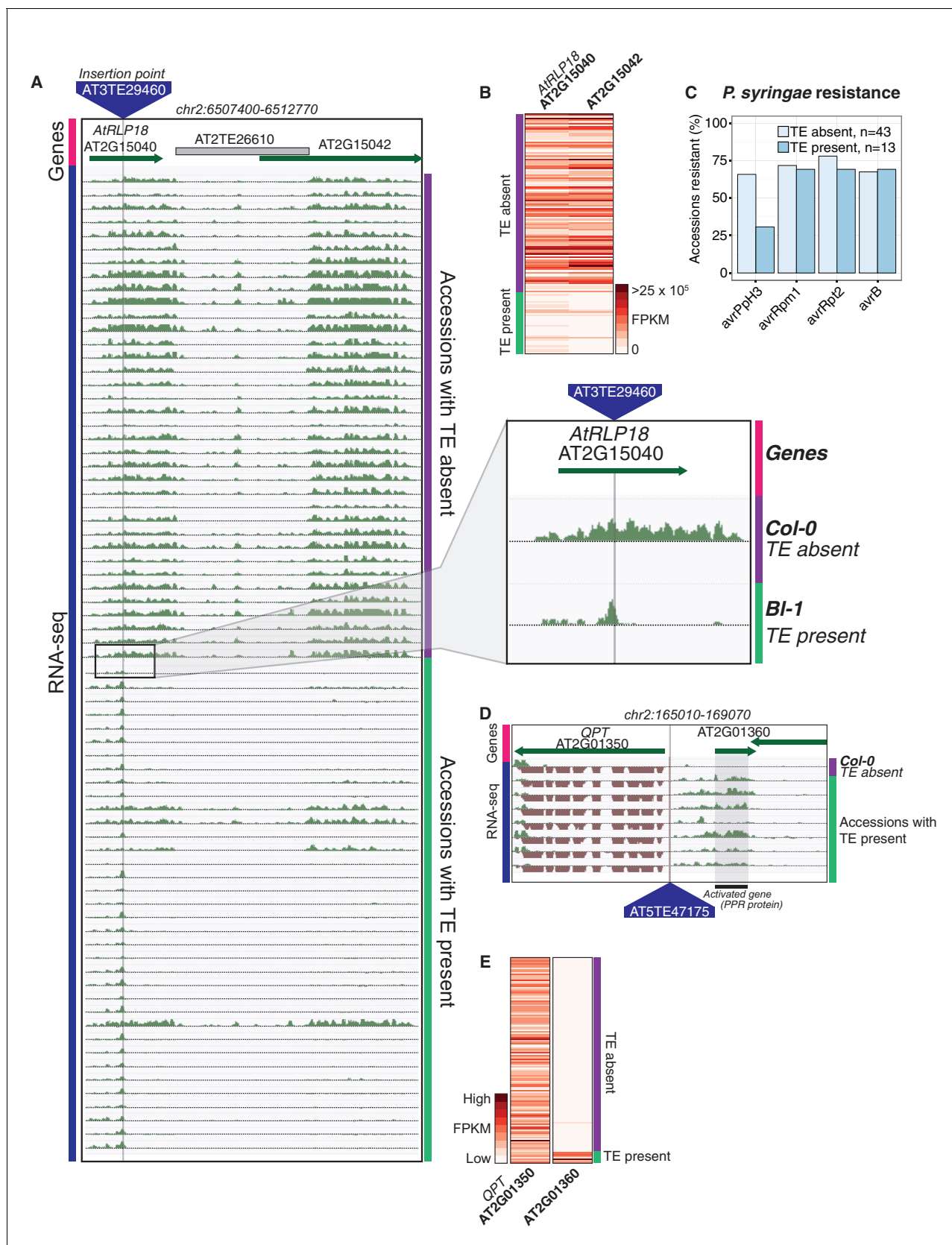


Figure 5. Effects of TE variants on local gene expression. (A) Genome browser representation of RNA-seq data for genes *AtRLP18* (AT2G15040) and a leucine-rich repeat family protein (AT2G15042). All accessions predicted to contain the TE insertion are shown. Inset shows magnified view of the TE

Figure 5 continued on next page

Figure 5 continued

insertion site for two accessions. (B) *AtRLP18* and AT2G15042 RNA-seq FPKM values for all accessions. (C) Percentage of accessions with resistance to *Pseudomonas syringae* transformed with different *avr* genes, for accessions containing or not containing a TE insertion in *AtRLP18*. (D) Genome browser representation of RNA-seq data for a PPR protein-encoding gene (AT2G01360) and *QPT* (AT2G01350), showing transcript abundance for these genes in accessions containing a TE insertion variant in the upstream region of these genes, as well as in Col-0. (E) RNA-seq FPKM values for *QPT* and a gene encoding a PPR protein (AT2G01360), for all accessions. Note that scales are different for the two heatmaps shown in E, due to the higher transcript abundance of *QPT* compared to AT2G01360. Scale maximum for AT2G01350 is 3.1×10^5 , and for AT2G01360 is 5.9×10^4 .

DOI: [10.7554/eLife.20777.029](https://doi.org/10.7554/eLife.20777.029)

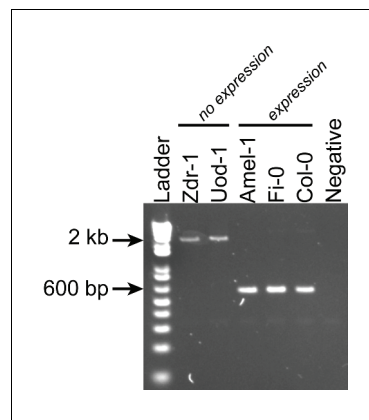


Figure 5—figure supplement 1. PCR validations for a TE insertion within the *AtRLP18* gene. Zdr-1, Uod-1, Amel-1 and Fi-0 were all predicted to contain the TE insertion at this locus, but only Amel-1, Fi-0 and Col-0 expressed the *AtRLP18* gene.

DOI: [10.7554/eLife.20777.030](https://doi.org/10.7554/eLife.20777.030)

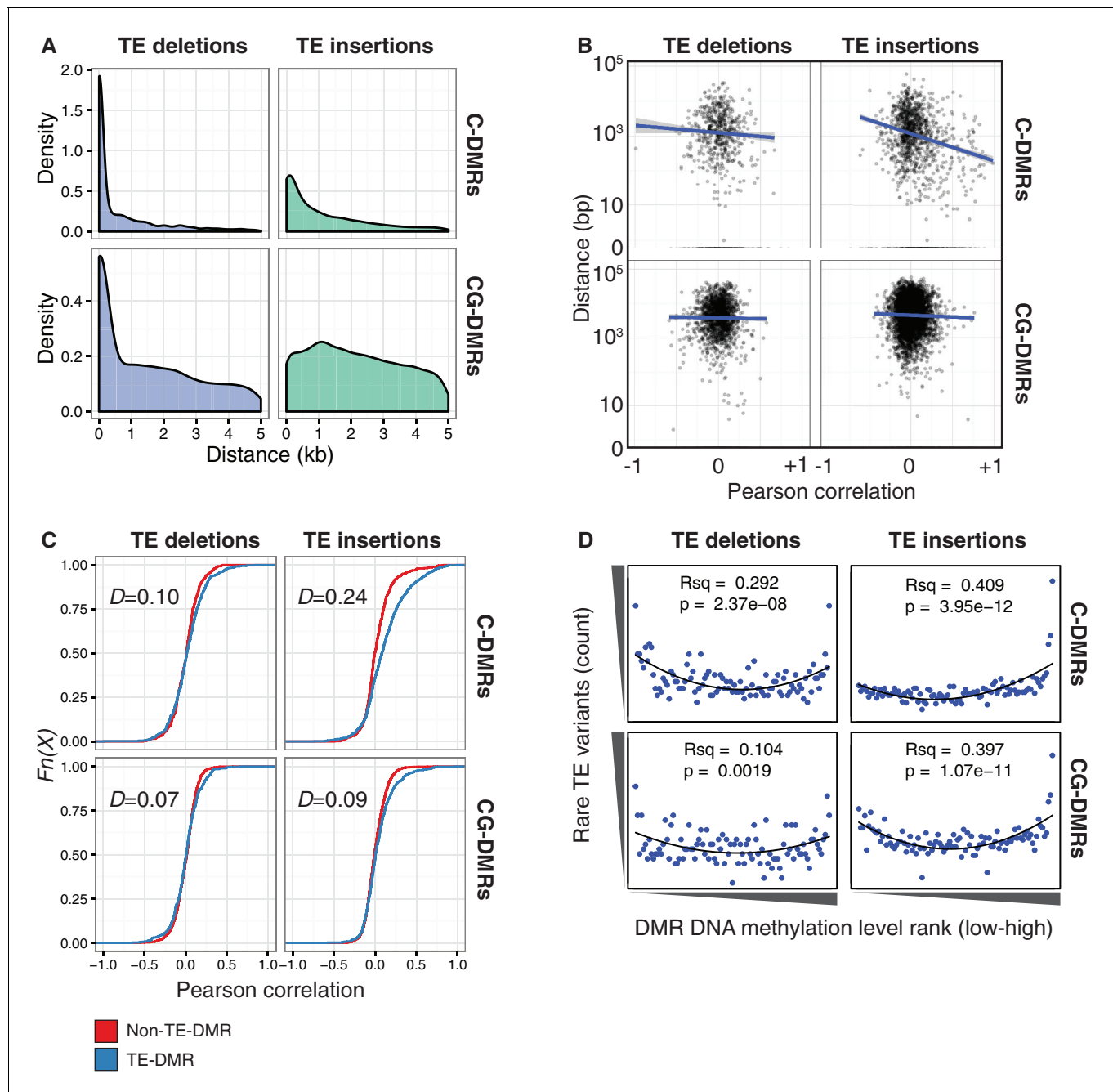


Figure 6. TE variants are associated with nearby DMR methylation levels. (A) Distribution of distances from TE variants to the nearest population DMR, for TE deletions and TE insertions, C-DMRs and CG-DMRs. (B) Pearson correlation between DMR DNA methylation level and TE presence/absence, for all DMRs and their closest TE variant, versus the distance from the DMR to the TE variant (log scale). Blue lines show a linear regression between the correlation coefficients and the log10 distance to the TE variant. (C) Empirical cumulative distribution of Pearson correlation coefficients between TE presence/absence and DMR methylation level for TE insertions, TE deletions, C-DMRs and CG-DMRs. The Kolmogorov–Smirnov statistic is shown in each plot, indicated by D . (D) Relationship between rare TE variant counts and nearby DMR DNA methylation level ranks, for TE insertions, deletions, C-DMRs, and CG-DMRs. Plot shows the cumulative number of rare TE variants in equal-sized bins of DMR methylation level ranks, from the lowest ranked accession (left) to the highest ranked accession (right). Lines indicate the fit of a quadratic model, and the corresponding R^2 and p values are shown in each plot.

DOI: 10.7554/eLife.20777.031

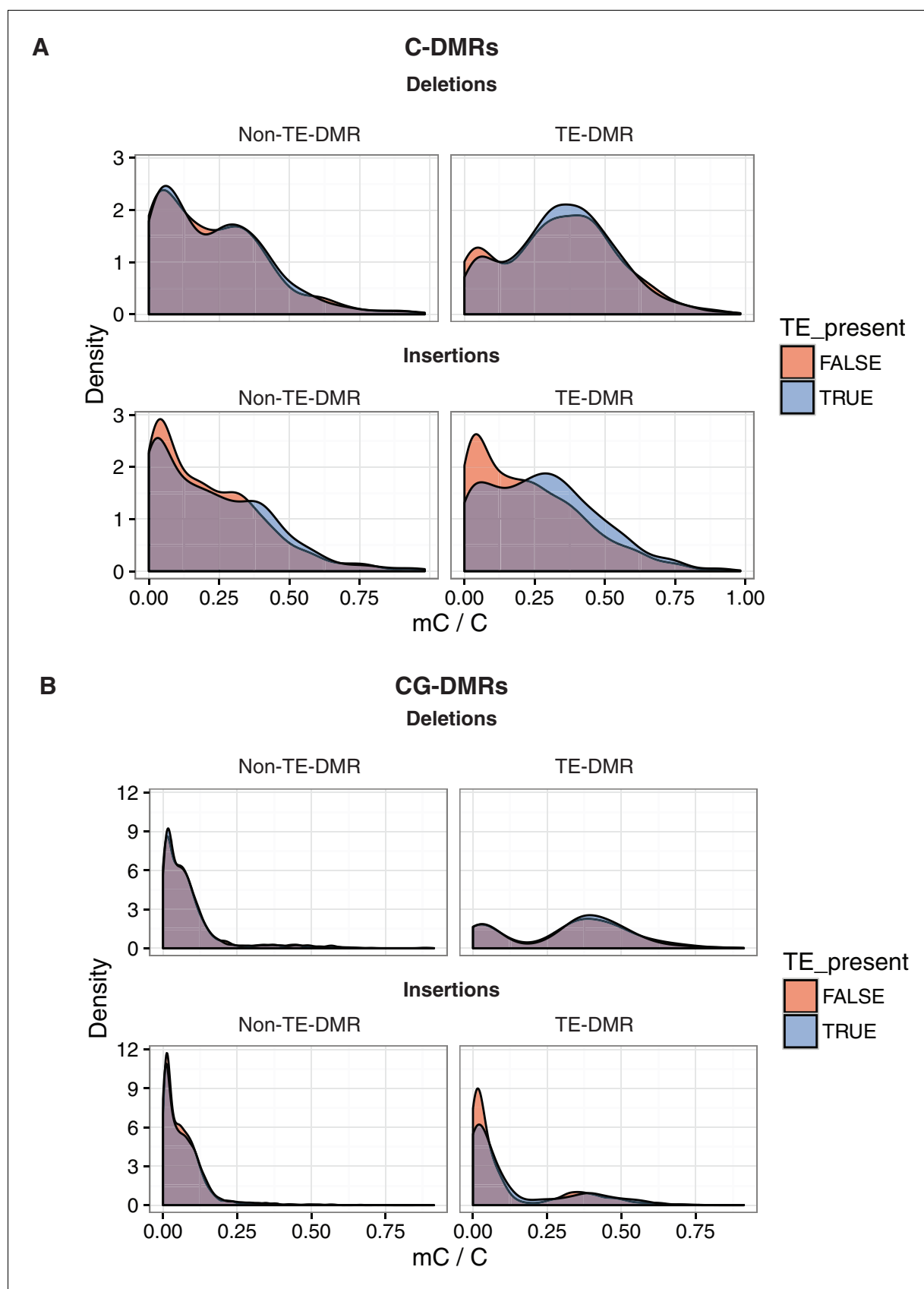


Figure 6—figure supplement 1. DNA methylation levels at DMRs near or far from TE variants. (A) DNA methylation density distribution at C-DMRs within 1 kb of a TE variant (TE-DMRs) or further than 1 kb from a TE variant (non-TE-DMRs), in the presence or absence of the TE, for TE insertions and TE deletions. (B) As for A, for CG-DMRs.

DOI: [10.7554/eLife.20777.032](https://doi.org/10.7554/eLife.20777.032)

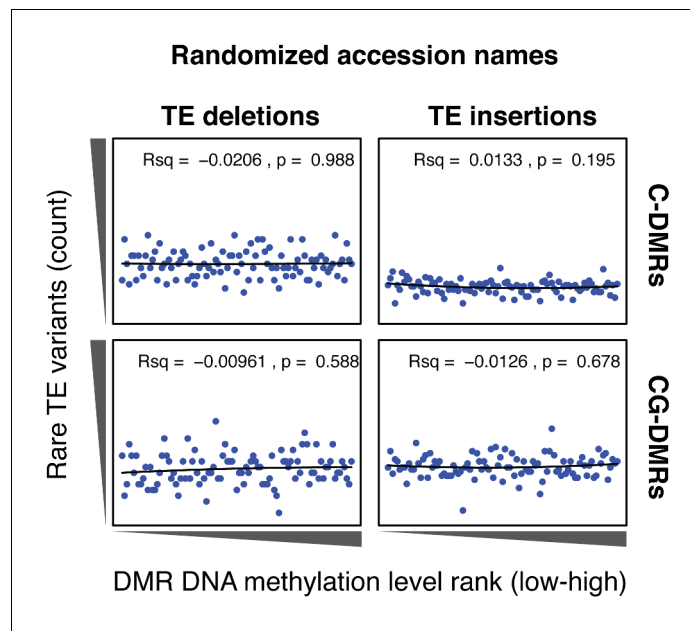


Figure 6—figure supplement 2. Cumulative number DMR methylation level ranks for DMRs near rare TE variants with accessions selected at random. Lines indicate the fit of a quadratic model, and the corresponding R^2 and p values are shown in each plot.

DOI: [10.7554/eLife.20777.033](https://doi.org/10.7554/eLife.20777.033)

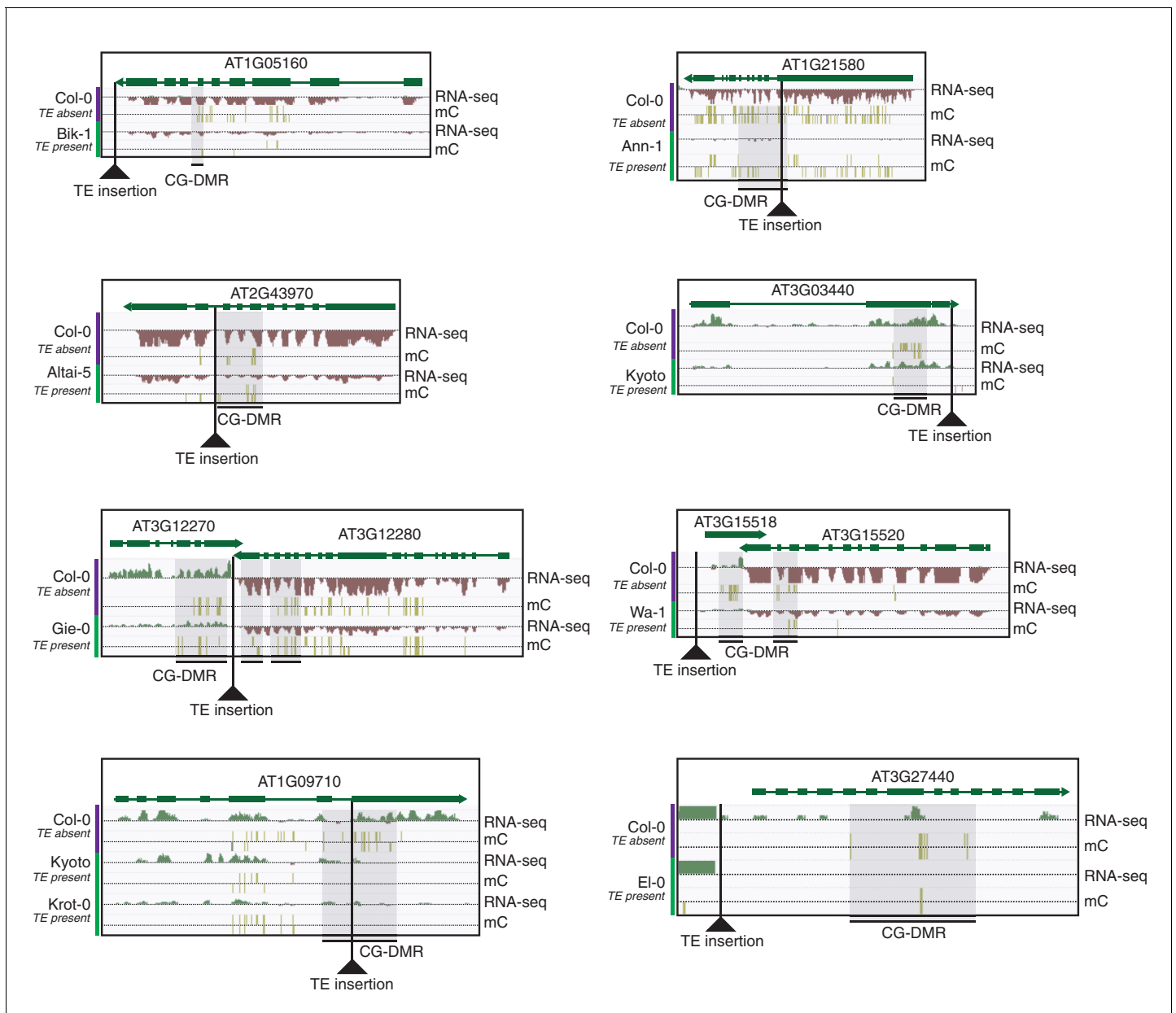


Figure 6—figure supplement 3. Selected examples of TE insertions apparently associated with transcriptional downregulation of nearby genes and loss of gene body CG methylation leading to the formation of a CG-DMR.

DOI: [10.7554/eLife.20777.034](https://doi.org/10.7554/eLife.20777.034)

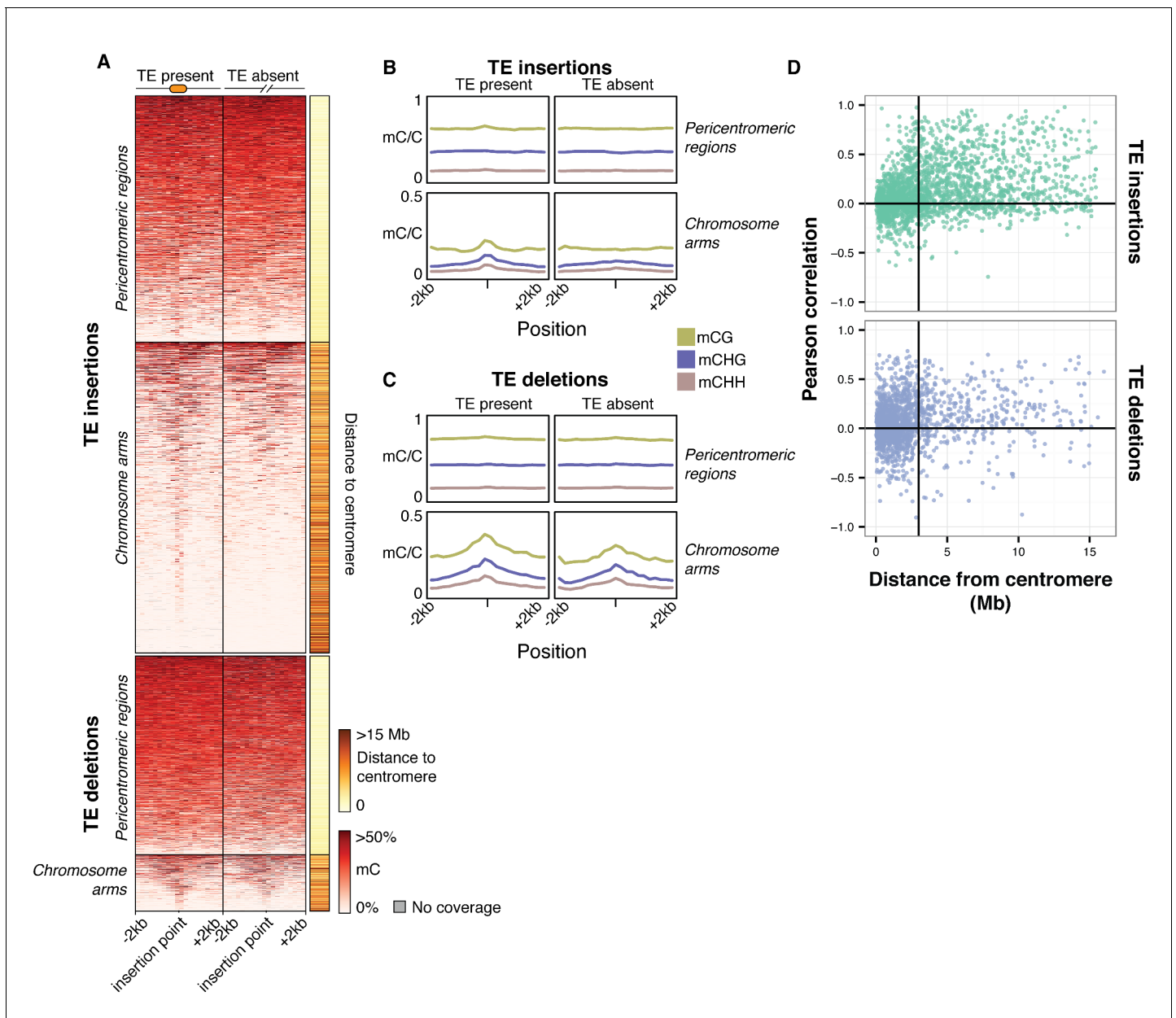


Figure 7. Local patterns of DNA methylation surrounding TE variant sites. (A) DNA methylation levels in 200 bp bins flanking TE variant sites, ± 2 kb from the TE insertion point. TE variants were grouped into pericentromeric variants (< 3 Mb from a centromere) or variants in the chromosome arms (> 3 Mb from a centromere). (B) DNA methylation level in each sequence context for TE insertion sites, ± 2 kb from the TE insertion point. (C) As for B, for TE deletions. (D) Distribution of Pearson correlation coefficients between TE presence/absence and DNA methylation levels in the 200 bp regions flanking TE variant, ordered by distance to the centromere.

DOI: [10.7554/eLife.20777.036](https://doi.org/10.7554/eLife.20777.036)

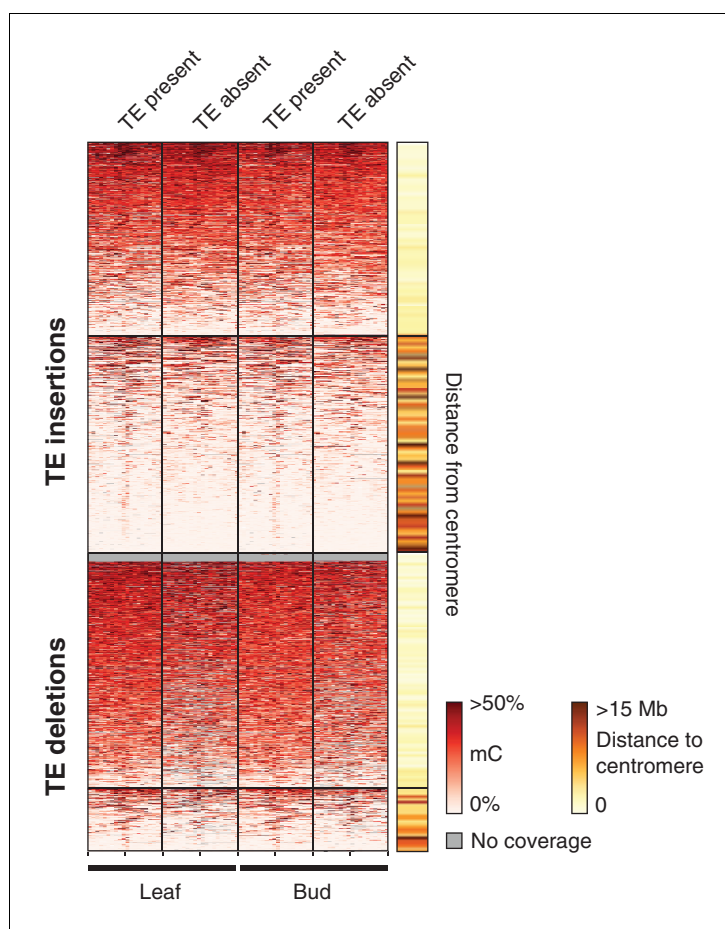


Figure 7—figure supplement 1. DNA methylation levels in 200 bp bins flanking TE variant sites in the 12 accessions with DNA methylation data for both leaf and bud tissue, ± 2 kb from the TE insertion point. TE variants were grouped into pericentromeric variants (< 3 Mb from a centromere) or variants in the chromosome arms (> 3 Mb from a centromere).

DOI: [10.7554/eLife.20777.037](https://doi.org/10.7554/eLife.20777.037)

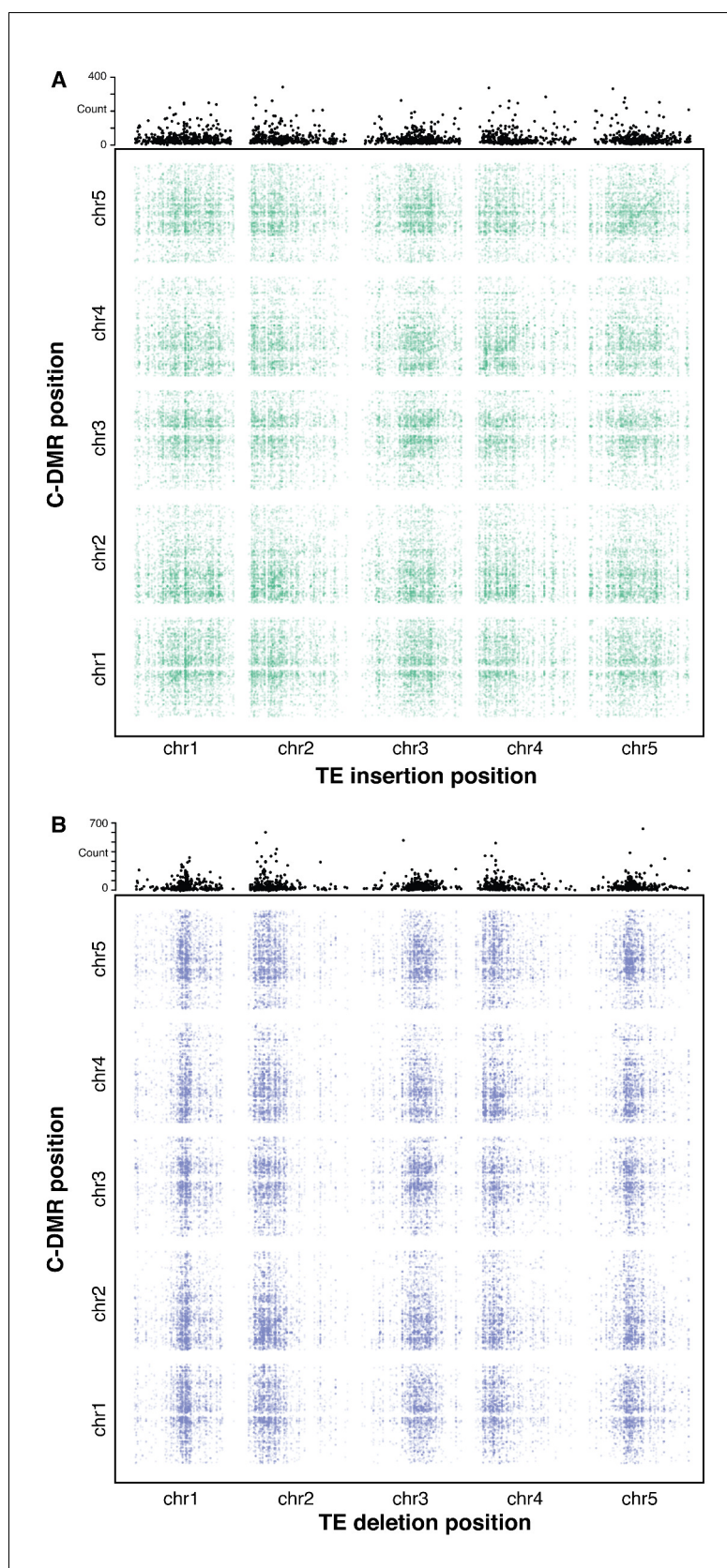


Figure 8. Association scan between TE variants and C-DMR methylation variation. (A) Significant correlations between TE insertions and C-DMR DNA methylation level. Points show correlations between individual TE-DMR

Figure 8 continued on next page

Figure 8 continued

pairs that were more extreme than all 500 permutations of the DMR data. Top plots show the total number of significant correlations for each TE insertion across the whole genome. **(B)** As for **(A)**, for TE deletions.

DOI: [10.7554/eLife.20777.038](https://doi.org/10.7554/eLife.20777.038)