Transposable elements are important contributors to standing variation in gene expression in *Capsella grandiflora*

Jasmina Uzunović[1*], Emily B. Josephs[2,3], John R. Stinchcombe[1], Stephen I. Wright[1]

[1]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Canada

[2]Department of Evolution and Ecology, University of California, Davis, California, USA

[3]Center for Population Biology, University of California, Davis, California, USA.

* Corresponding author

E-mail: jasmina.uzunovic@mail.utoronto.ca

Abstract

Transposable elements (TEs) make up a significant portion of eukaryotic genomes, and thus are important drivers of genome evolution. However, the evolutionary forces controlling TE copy number and the extent to which TEs affect phenotypic variation on a genome-wide scale are still unclear. We characterised TE insertion polymorphism and its effects on gene expression in 124 whole genome sequences from a single population of *Capsella grandiflora*. The frequency of insertions was negatively correlated with distance to genes, as well as density of conserved non-coding elements, suggesting that the negative effects of TEs on gene regulation are important in limiting their abundance. Rare TE variants strongly influence gene expression variation, predominantly through downregulation. In contrast, rare single nucleotide polymorphisms (SNPs) contribute equally to up- and down-regulation, but have a weaker effect. Taken together, these results imply that TEs are a significant contributor to gene expression variation and can be more likely than rare SNPs to cause extreme changes in gene expression.

Author Summary

Transposable elements (TEs), mobile DNA elements with the ability to excise from the genome and reinsert in new locations, are important components of genomic diversity. Due to their abundance and mobility, TEs play an influential role in genomic evolution, often deleterious. Here we show that TEs in a population of the plant *Capsella grandiflora* are most deleterious when they insert in genic and regulatory regions. We find that TEs indeed are associated with unusual levels of gene expression, predominantly decreased expression. Furthermore, this effect is stronger than the association of single nucleotide polymorphisms with

gene expression variation, highlighting the importance of TE contribution to the maintenance of expression variation.

Introduction

Transposable elements are one of the most abundant and pervasive types of selfish genetic element, making up over 80% of the genome in some species [1]. They experience selective pressure to increase their own numbers, and at least in some species their abundance is thought to reflect a balance between their rate of increase through transposition and the action of selection on the host due to their deleterious fitness effects [2,3]. Although the population dynamics of TEs have been characterized in a handful of model systems, our understanding of the main selective factors controlling TE abundance, and the relative importance of TEs vs. other sources of genetic variation (such as SNPs) in maintaining phenotypic variation remains limited.

Population studies of TEs in the compact genomes of *Drosophila melanogaster*, *Arabidopsis thaliana*, and *Caenorhabditis elegans* have shown that most euchromatic insertions are at very low population frequency [4–8]. A bias towards rare insertions suggests that pervasive purifying selection is acting against TEs in these species. There are two main hypotheses for why TEs are subject to purifying selection [3,9]. First, ectopic recombination between TEs on non-homologous chromosomes or distinct locations on the same chromosome can lead to aberrant chromosomal rearrangements, which are costly for fitness [10,11]. Alternatively, insertions in or near genic regions can negatively affect gene function, regulation, and expression [12].

Studies in *D. melanogaster* populations provide evidence of selection acting against both ectopic recombination between TEs [4] and the deleterious effects of gene disruption [13,14]. In

67 *A. thaliana,* distance to genes seems to be more important than ectopic recombination in
68 mediating selection against new insertions, consistent with the gene disruption hypothesis
69 playing a primary role [6,15]. TE abundance is negatively correlated with gene density,
70 suggesting TEs are harmful when they insert near protein-coding genes. Similarly, in *C. elegans*
71 TEs are enriched in regions characterised by low gene density and high recombination rate,
72 suggesting proximity to genes is a more important factor in TE survival as opposed to ectopic
73 recombination [8]. However, both *A. thaliana* and *C. elegans* are highly self-fertilizing, and
74 elevated homozygosity may minimize the effects of ectopic recombination [15–17]. Thus, while
75 we have preliminary evidence about the relative importance of ectopic recombination and gene
76 disruption in the regulation of TE abundance, it is difficult to generalize about their relative
77 importance with so few studies, especially in outcrossing species.

78 Given the evidence for selective pressure against gene disruption in a number of systems
79 [6,14,18], TEs may be a significant source of genetic variation for fitness-related phenotypes,
80 and can have a major impact on the evolution of the rest of the genome, especially functional
81 elements such as genes or conserved regulatory regions. Active TEs initiate mutations with
82 diverse molecular effects as they move around the genome. Most of these are deleterious, such as
83 disrupting genes or gene regulatory elements [19], causing structural rearrangements, and
84 silencing or decreasing expression of nearby genes through the spread of host-induced epigenetic
85 silencing of TEs [6,7,14,20,21]. However, TEs can also contribute putatively adaptive mutations
86 such as gene duplications, new gene regulatory elements, and selective regulation of gene
87 expression under stressful conditions [22–24]. Despite our knowledge about the many specific
88 ways in which different TEs can impact genes, there are few studies examining the extent to
89 which TEs as a class contribute to genetic and phenotypic variation.

90 A genome-wide survey of eQTL loci from a single large population of the self-
91 incompatible outcrossing plant *Capsella grandiflora,* shows around a third of genes are
92 associated with local eQTLs (those present within 5 kb of the gene) [25]. The variants associated
93 with gene expression are rarer than expected under neutrality, and their frequency is negatively
94 correlated with the size of their effect on gene expression [25], consistent with purifying
95 selection playing a large role in genetic variation for gene expression in *C. grandiflora*.
96 However, Josephs and colleagues [25] looked solely at SNPs, which are only a subset of the total
97 genetic variation that exists, and this system provides an excellent opportunity to explore the
98 relative role of 'classical mutation' vs. the importance of transposable element activity in
99 maintaining genetic variation in a natural population.
100 In this study, we examined TE polymorphism and its effects on expression variation
101 using the same population studied by Josephs et al. We identify insertions across all individuals
102 and explore the relationship between TE population frequency, rates of crossing over, and
103 density of coding sites and non-coding conserved sequences. We test the effect of TEs on gene
104 expression variation using genome-wide expression data. We then performed the same tests on
105 rare SNPs to compare the relative role of the two types of genetic variants.
106

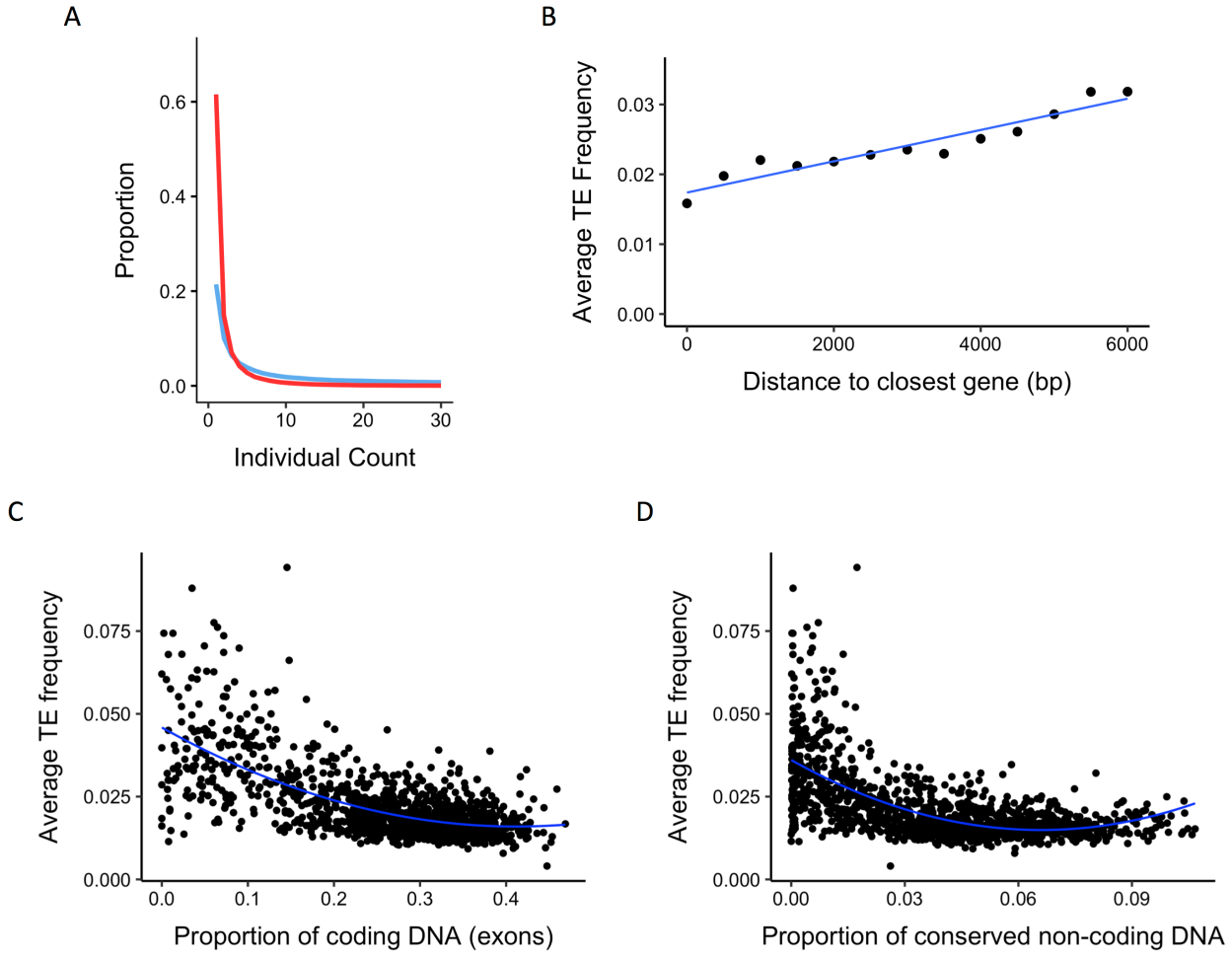
107 **Results and discussion**

108 We combined high-confidence TE calls from 124 *C. grandiflora* plants using the
109 RelocaTE2 pipeline [26]. In total, we found 95,740 TEs in the population with an average of
110 2468 insertions per individual and a median of 2675 insertions per individual (range of 497 to
111 3131). In a survey of 216 *A. thaliana* accessions, there were 23,095 unique (non-reference) TE
112 variants (with a mean of 1657 per accession) [7]. The higher abundance of TEs in the *C.*

113 *grandiflora* population could reflect higher TE activity in outcrossers vs selfers leading to more
114 low frequency insertions, since the *A. thaliana* study did not include insertions shared with the
115 reference genome) [27,28].

116 The frequency spectrum of TEs in the population shows a major skew towards low-
117 frequency insertions compared to 4-fold degenerate SNPs, with a strong enrichment of insertions
118 present in only one individual (Fig 1A). The trend of mostly low-frequency insertions is
119 consistent with observations from other organisms and indicates there is strong negative
120 selection acting against TEs [5–8,29,30].

121



122

Fig 1. Patterns of TE insertion frequency from population data. (A) Frequency spectrum of TEs (red) and 4-fold SNPs (blue). (B) Relationship between TE frequency and distance to the closest gene. The first point represents the average frequency of all TEs which are within a protein-coding gene. Subsequent points average the frequency of all TEs in 500 bp windows from the closest gene. The blue line shows a linear model fit through the data. $r = 0.94$, linear coefficient has a $p < 0.0001$. C and D Relationship between average TE frequency and proportion of (C) coding DNA and (D) conserved non-coding sequence (CNS) in 100 kb windows across the genome. The blue curve shows a quadratic model fit. Spearman's $\rho = -0.562$ in (C) and -0.619 in (D).

Selection against gene disruption regulates TE numbers

There are two main hypotheses for why TEs are deleterious in the genome: (i) they can cause ectopic recombination between two non-homologous insertions or (ii) their insertion can disrupt gene function or expression. In the first case, TEs should be more deleterious in areas of high recombination rate, under the assumption that ectopic recombination rate mirrors the homologous crossover rate. We used estimates of recombination rate in bins of 100 kb across the genome and compared to the average TE population frequency of all insertions present in that bin (S5 Fig). There is no significant correlation between TE frequency and recombination rate across the genome, suggesting ectopic recombination does not play a large role in outcrossing *C. grandiflora*, consistent with results from the highly selfing *A. thaliana* [15]. Ectopic recombination is expected to be of greater importance in outcrossing species due to their higher heterozygosity; a TE is more likely to recombine non-homologously if it does not have a homologous partner to pair with. However, there is no evidence for a larger role of ectopic

146 recombination in this outcrossing species than in a related selfing species. While finer-scale
147 estimates of recombination might reveal a significant effect and it is possible that ectopic
148 recombination rates are less correlated with rates of crossing-over than is often assumed, the
149 previously observed lack of an effect of recombination on TE abundance in *A. thaliana* [15]
150 cannot be attributed to its selfing mating system.

151 If TEs are deleterious because they disrupt gene function, then insertions should be most
152 deleterious, and consequently at lowest population frequency, when they are located in or near
153 genes. Plotting TE frequency against distance from the closest gene, in bins of 500 bp, reveals a
154 significant positive correlation (Fig 1B; r = 0.94), suggesting that TEs closer to genes experience
155 stronger purifying selection. This result is in line with a recent publication in the same species
156 [31].

157 Another way to test the gene disruption hypothesis is to plot the relationship between
158 average TE frequency and coding site density in 100 kb windows across the genome. In testing
159 this, we found a significant negative correlation, suggesting that TEs in regions of high gene
160 density are kept at low population frequency (Fig 1C; Spearman's ρ: -0.562). We also evaluated
161 TE frequency relationship to conserved non-coding sequences (CNSs), using CNSs conserved
162 across 9 Brassicaceae species, many of which show evidence of being regulatory elements [32].
163 Plotting TE frequency vs CNS density also leads to a strong negative correlation, implying TEs
164 face selective pressures due to negative effects of inserting into or near regulatory elements (Fig
165 1D; Spearman's ρ: -0.619). There is an obvious auto-correlation between coding site density and
166 CNS density as regulatory elements will be found close to genes. To check that the relationship
167 of CNS density and TE frequency cannot be fully explained by the correlation between genes
168 and CNSs, we built two quadratic regression models chosen because they showed the best fit to

169 the data. In the first model, TE frequency is predicted only by proportion of coding sites while in
170 the second, the proportion of CNSs is added as an additional predictor (the latter model is shown
171 in Table 1; adjusted $R^2 = 0.502$). Using an F-test on the two models showed the model with
172 coding sites and CNSs explains significantly more variation than the model with coding sites
173 alone ($p < 0.0001$). These results suggest that important TE effects on genes extend not only to
174 direct gene disruption and methylation-mediated gene silencing, but also to disruption of
175 regulatory sequence. The effect of TEs on CNSs could be due to direct disruption of enhancers,
176 methylation of regulatory sequence caused by TE silencing, or spatial disruption of gene
177 regulation caused by insertions between a gene and its regulatory region leading to aberrant
178 expression.

179 **Table 1. Quadratic regression model of average TE frequency.**

Model: Average TE frequency ~ proportion of coding sequences (CSs) + (proportion of coding sequences)^2 + proportion of CNSs + (proportion of CNSs)^2		
Coefficient	**Estimate**	**P-value**
Proportion of coding sequences	-0.100	< 0.0001
(Proportion of coding sequences)^2	0.126	< 0.0001
Proportion of CNSs	-0.236	< 0.0001
(Proportion of CNSs)^2	1.399	0.000381
Adjusted R^2	0.502	< 0.0001

180

181 **TEs contribute significantly to gene expression variation**

182 We have shown that TEs contribute to sequence variation, mostly as rare mutations that
183 are under stronger purifying selection near genes and regulatory regions. Given this, we next
184 investigated the extent to which TE insertions alter gene expression and contribute to standing
185 genetic variation. One approach for testing rare variants is to ask whether individuals with
186 extreme gene expression are often associated with TEs near their genes, or whether there is a
187 greater burden of rare variants at the phenotypic extremes. Individual genes are sorted into
188 expression rank bins, and then the total number of all local rare variants is plotted for all genes in
189 all individuals. If rare variants cause a change in gene expression, either an increase or decrease,
190 then this will result in a parabolic shape [33].

191 To facilitate later comparison with SNP data, we used 109 individuals which passed TE,
192 expression and SNP quality filters. Using all TEs with a MAF < 3% and within 500 bp of a gene,
193 we found that individuals with more TEs near a gene are more likely to have extreme expression
194 values, and TEs appear to have a strong effect on decreasing gene expression, and a weaker
195 effect of increasing expression at the extremes (Fig 2A; $R^2 = 0.450$). As a null control, we
196 permuted mutation count and expression rank; as expected, permutations lead to no significant
197 correlation, demonstrating the significance of our observed effect (S6 Fig). This general result
198 suggests that TEs do contribute to expression variation and that TE silencing and gene disruption
199 effects may be more common than TE-mediated upregulation. This result is in contrast with a
200 recent analysis in *Arabidopsis thaliana*, where TE insertions driving increased expression
201 appeared to be as prevalent as expression-decreasing effects ([6], but see [7]).

202

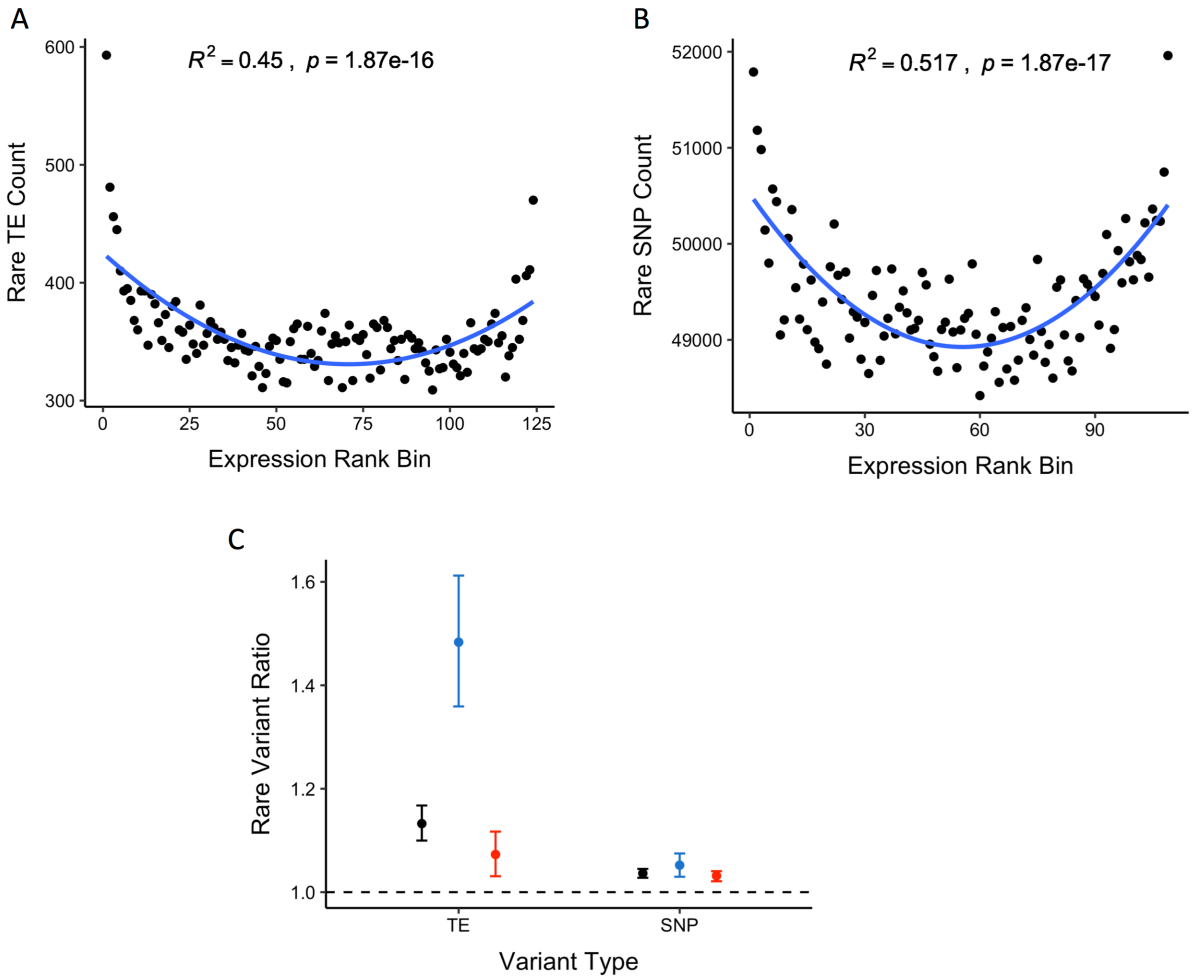


Fig 2. Relationship between (A) TE and (B) SNP count and gene expression rank bin for all rare variants (MAF < 3%) within 500 bp of a gene. C) Ratio of average number of rare variants in expression outliers relative to non-outliers for TEs and SNPs. 'Upper' refers to expression-increasing outliers, and 'lower' refers to expression-decreasing outliers. We ran 1000 permutations of the data to obtain normal distributions of the ratio statistic, and in all cases the observed statistic fell outside of the permuted distribution values.

While TEs appear to strongly downregulate gene expression, it is difficult from such an analysis to directly quantify the strength of the effect. We used a complementary statistical

213 approach by calculating the average number of TE insertions across genes in individuals that are
214 expression outliers relative to the average number in non-expression outliers. The ratio value for
215 TEs is 1.13, and expression outlier individuals have significantly more TEs than non-outlier
216 individuals (Fig 2C; 95% quantile confidence interval: 1.10-1.17 from 1000 bootstraps). When
217 considering only downregulated expression outliers, the ratio of TEs in outliers vs non-outliers is
218 a much higher value of 1.48, consistent with the pattern observed in the burden of rare variants
219 test (Fig 2C; 95% quantile confidence interval: 1.36-1.61 from 1000 bootstraps). These results
220 suggest that the standing genetic variation contributed by TEs has a significant effect on
221 disrupting gene expression, pushing individuals to extreme expression values for a trait which is
222 under stabilizing selection and that TE insertions are more likely to decrease expression than
223 increase it.

224 Breaking down the results by TE type reveals large differences in the direction and type
225 of effect that different TE types have on gene expression. Long terminal repeat (LTR) elements,
226 which are large retrotransposons, show the strongest effect on decreasing gene expression (R^2 =
227 0.471) while other TE types tested show significant but less pronounced effects (Fig 3, S7 Fig).
228 LTRs are the most abundant and the most methylated TE type in most plant genomes [34],
229 including *C. grandiflora*. Methylated TEs, which have been silenced by the host genome to
230 prevent their transcription and transposition, can have stronger effects on nearby gene
231 expression, either through the spread of methylation past the TE or by making the chromatin less
232 accessible to transcription factors and transcriptional machinery.
233

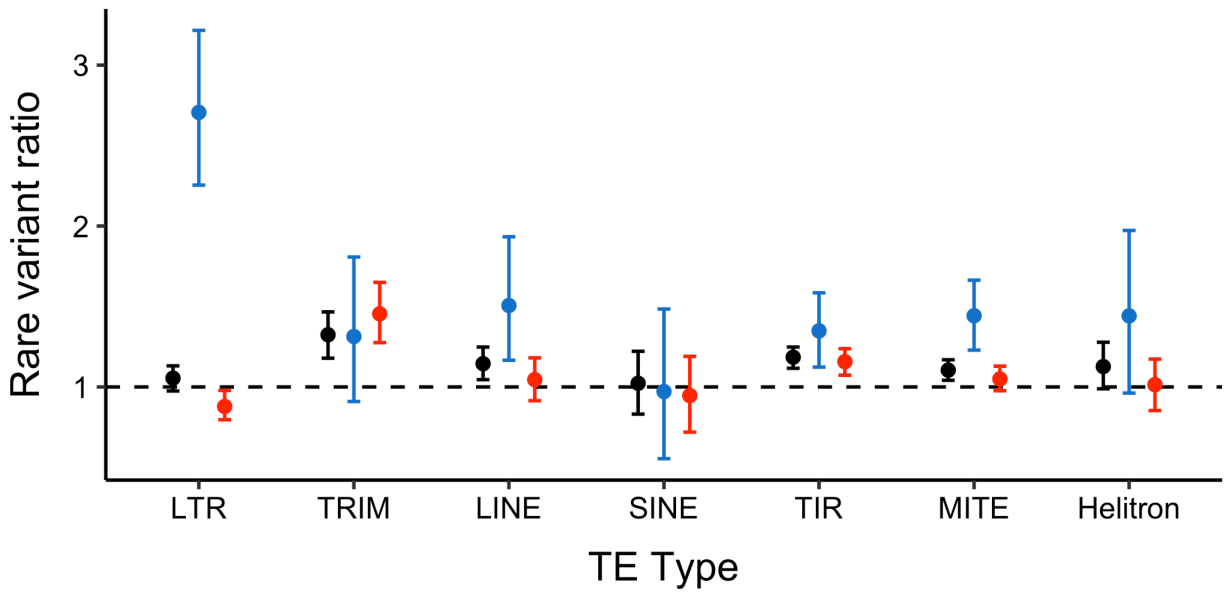
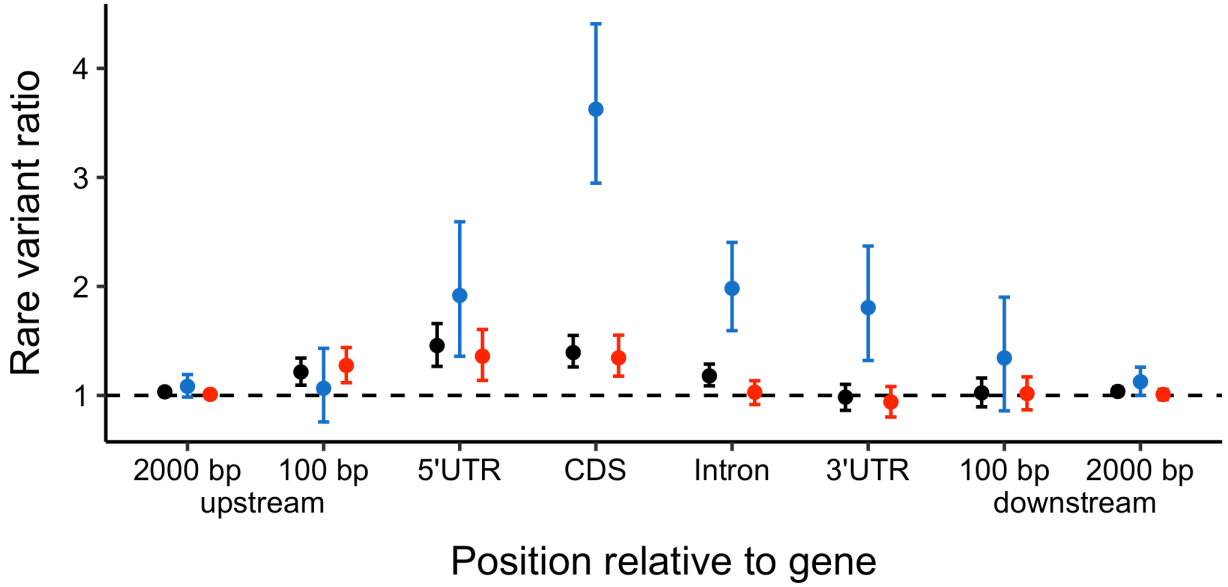


Fig 3. Ratio of average number of rare variants in expression outliers relative to non-outliers for different TE types. Black – all outliers; blue – expression-decreasing outliers; red – expression-increasing outliers.

Genic insertions cause the strongest expression decrease

If TEs affect gene expression by disrupting gene regulatory regions, then there should be a strong effect caused by TEs upstream of the gene. However, the strongest correlation with extreme gene expression occurred when considering TEs inserted into genes (S8 Fig; R^2 = 0.524). Upstream TEs had a lower overall effect; interestingly they were only significantly over-enriched in increased expression outliers, with no significant effect in decreased expression outliers (Fig 4). This is somewhat similar to observations in *A. thaliana*, where upstream insertions showed a more symmetrical effect on expression than insertions inside or downstream of genes [7]. The upstream effect was only significant for TEs within 100 bp of the transcription start site (TSS) (Fig 4, S9 Fig), consistent with previous results suggesting that the greatest abundance of upstream conserved non-coding sites were 200 bp away from the TSS [32].

250 Similarly, upstream eQTLs were most abundant within 1kb from the TSS (closer intervals were
251 not tested) [25]. One possible explanation for the weaker and more symmetrical overall effect of
252 insertions upstream of genes is that a major effect of TE insertions into regulatory sequence is an
253 effect on expression noise, rather than a specific downregulation. Recent work on de novo
254 mutational effects in regulatory regions in yeast [35] suggest many more mutations have an
255 effect on the variance in gene expression than on the expression mean. TE effects could be
256 similar in regulatory regions, driving a pattern of purifying selection associated with increased
257 variance in average expression.
258



259
260 **Fig 4. Ratio of average number of rare variants in expression outliers relative to non-**
261 **outliers for TEs upstream of, in, and downstream of genes.** The up- and downstream
262 categories are non-overlapping. Black – all outliers; blue – expression-decreasing outliers; red –
263 expression-increasing outliers.
264

265 Intragenic TEs appear to cause the strongest shift in expression, so we further categorised
266 by the genic region of insertion. TEs in all genic regions were significantly over-enriched in low
267 expression gene outliers, with the strongest effect from insertions into coding regions (Fig 4, S10
268 Fig). In contrast, only TEs into the 5'UTR and coding sites were also significantly over-enriched
269 in high expression gene outliers. The 5'UTR contains regulatory sites, so TE insertions may be
270 contributing extra regulatory sites leading to increased transcription of the whole gene. TEs are
271 known to be a large source of promoter and regulatory regions which can be co-opted by the host
272 genome and contribute to the evolution of regulatory networks [36,37]. The large effect of TEs
273 in coding regions on decreasing gene expression could be due to faster decay of mRNAs with
274 large mutations, for example by the nonsense-mediated decay (NMD) pathway [38]. However,
275 the strong enrichment of TEs in low expression outliers when they insert into coding sites could
276 be driven by a technical artifact where reads with TEs are dropped due to poor mapping (see
277 Methods). When considering gene expression broken down into individual exons, there were no
278 significant differences in enrichment of TEs in exons upstream of the insertion, downstream of
279 the insertion and the exon containing the insertion (S4 Fig). Therefore there is not evidence for
280 technical artifacts which could lead to misinterpretation of results. Furthermore, TEs appear to
281 equally reduce expression of all exons, suggesting that TEs are not causing early transcription
282 breaks. The impact of TEs in introns could be driven by targeted silencing of insertions in introns
283 [39].
284

285 **TE insertions show stronger enrichment in expression outliers than SNPs**

286 Josephs and colleagues [25] previously showed that a third of genes had an eQTL in this
287 population of *C. grandiflora*, and that there is a significant negative correlation between eQTL

288 MAF and effect size, suggesting that purifying selection is mostly responsible for maintaining
289 genetic variation for gene expression. However, they tested SNPs with a MAF > 5%, and so the
290 contribution of rare variants to gene expression variation is still unclear.

291 We explored the effect of rare SNPs with a MAF < 3% on gene expression. Overall, these
292 variants are present in much larger number, and are equally likely to cause gene up- or down-
293 regulation (Fig 2B; $R^2 = 0.517$). This is in contrast with a recent study in maize, where SNP
294 variants upstream of genes tended to show greater evidence of downregulation [40]. When
295 comparing the ratio of variants in expression outliers vs non-outliers, SNPs had a smaller effect
296 than TEs (1.04 vs 1.13, respectively; Fig 2C). As a class of mutation, TEs are 9% more over-
297 represented in outliers than SNPs. In low expression outliers, there is a considerably greater gap
298 between TE and SNP over-representation of 43% (1.48 vs 1.05 ratio values, respectively; Fig
299 2C). Both tests of the burden of rare variants on gene expression are genome-wide, made by
300 summing across all genes and all individuals in the population. As such, the results highlight the
301 global impact that both TEs and SNPs have on gene expression variation. The outlier test used a
302 stringent cut-off of 4 standard deviations away from the mean to define outliers, so it may be that
303 TEs are on average more likely to cause a large change in expression and are over-enriched in
304 this test. SNPs, on the other hand, may be on average more likely to cause a small change in
305 expression and are more important in explaining small variation in expression. Therefore,
306 individual TEs may be strong candidates for having large effects on expression (large-effect
307 QTL), but considering their numbers SNPs are also likely important, with smaller individual
308 effects on average.

309 A major challenge when comparing SNPs to TEs is detection power. Calling non-
310 reference TEs with short-read data is difficult, especially in outcrossing species where many

311 insertions are heterozygous, leading to more false positives and false negatives than when calling
312 SNPs [41]. These power issues could lead to biases in our population frequency estimates. There
313 may be many more rare TEs in the population which the pipeline is not detecting, however, in
314 this case our results would be undervaluing the true effect sizes and are therefore conservative
315 estimates of the impact of rare TEs. The outlier expression test controls for this problem because
316 it calculates a ratio of TE numbers between two groups, both of which have equal detection
317 power, allowing for the comparison of TE and SNP effect size independent of the ability to call
318 these variants.
319

320 **Conclusions**

321 This study explored the relationship between TEs and the rest of the genome in a
322 population of *C. grandiflora*. We find that selection against disruption of gene function and
323 regulation plays an important role in limiting TE copy number, but we see no evidence for
324 selection against ectopic recombination. TE insertions in or near genes strongly downregulate
325 expression and to a smaller extent upregulate expression, with the effect differing by the type of
326 TE and the genic region into which it has transposed. In comparison, SNPs have an overall
327 weaker effect and are equally likely to up- and down-regulate expression. These results highlight
328 the important contribution of TEs to standing genetic and phenotypic variation.
329

330 ## Methods

331 **Study organism**

332 The self-incompatible, outcrossing plant *Capsella grandiflora*, is a good model for using
333 population genomic approaches to make inferences about natural selection due to its low

334 population structure and large effective population size ($N_e \approx 600{,}000$) [42–44]. The species is
335 a member of the Brassicaceae family and is native to Northern Greece and Southern Albania.
336 There is evidence of prevalent purifying and positive selection on SNPs in *C. grandiflora*
337 [43,45].
338
339 **Genomic data and expression**
340 The DNA and RNA data have been previously described [25,46]. Briefly, plants were
341 sampled from a single population located near Monodendri, Greece (population Cg-9 in [44]).
342 Seeds were collected and grown in the University of Toronto glasshouses and random crosses
343 were performed. The offspring of these crosses were grown in the University of Toronto growth
344 chambers, where leaf tissue was collected and used for DNA and RNA extraction. Whole
345 genome sequencing was done using 100 cycles of paired-end sequencing in a Hiseq 2000 with
346 Truseq libraries (Illumina), resulting in 100 bp paired-end reads. RNA was sequenced at the
347 Genome Quebec Innovation Centre in an Illumina HiSeq resulting in 100 bp paired-end or
348 single-ended reads. For consistency, one read was randomly chosen among the paired end reads
349 and they were subsequently treated as single ended. The reference genome used for mapping
350 reads is *Capsella rubella*, a close relative of *C. grandiflora*, which diverged approximately
351 200,000 years ago [28,42].
352 RNA reads were mapped to an exon-only reference generated from the *C. rubella*
353 reference genome [25,28] using Stampy 1.0.21 with default settings [47]. Expression level was
354 measured with the HTSeq.scripts.count feature from HTSeq [48] and normalized for sequencing
355 depth by dividing read counts by median read count of the entire sample. Genes with a median

356 expression level below five reads per sample before normalization were removed from the
357 analyses.
358
359 **Calling TEs**
360 We built a new TE library to use with the detection pipeline, with a focus on TEs that
361 show evidence of being recently active. Using pindel [49], we found large deletions (>150 bp) in
362 30 of the *C. grandiflora* short-read sequenced individuals relative to the *C. rubella* reference
363 genome. These represent candidate TE insertions that occured in *C. rubella* since they in the
364 reference genome at those locations but absent from those locations in the *C. grandiflora* sample.
365 The PASTEClassifier pipeline [50] allowed us to classify the sequences obtained from the *C.*
366 *rubella* reference into TE order based on the hierarchical TE classification system [51]. Any
367 sequences that remained unclassified were removed.
368 We used the RelocaTE2 [26] pipeline to detect both reference and non-reference TE
369 insertions in 124 *C. grandiflora* individuals from our population. Briefly, RelocaTE2 is based on
370 a combined split read and discordant read pairs mapping approach. All fastq files were
371 subsampled down to 70X coverage to avoid elevated TE calls in individuals with higher
372 coverage, and 23 individuals with below 70X coverage were removed. RelocaTE2 was run using
373 bowtie2 as the TE aligner, using the custom-built TE library described above, and the *C. rubella*
374 reference genome with default settings. We used bwa [52] and samtools [53] to align short reads
375 to the reference genome, and then passed these alignments to RelocaTE2 to genotype TE
376 insertions as heterozygous or homozygous. Heterozygous TE insertions will have reads aligning
377 to the reference over the insertion site because one of the homeologs does not have a TE.

378 To create a population level table of TE insertions and their genotype (homozygous
379 absent, heterozygous, or homozygous present), we merged all the insertions from the RelocaTE2
380 'high-confidence' labelled files, which were insertions supported by at least one junction read
381 from both sides. Any insertions called between individuals which were within 50 bp of one
382 another and were of the same TE classification type were considered as the same insertion.
383 While RelocaTE2 can genotype insertions as homozygous or heterozygous, there is some
384 evidence from the TE population frequency spectrum that these calls are not always accurate.
385 When using allele counts (0, 1, or 2 for absent, heterozygous, homozygous) to make a frequency
386 spectrum, there is a staggered pattern which goes away when examining the spectrum using
387 individual count (0 or 1 for presence or absence; see S1 Fig and Fig 1). This suggests that
388 homozygosity is being over-called. For all rare variant gene expression analyses (see below), we
389 used individual counts, effectively ignoring homozygous vs heterozygous calls, however results
390 were qualitatively the same when performed with allele counts (S2 Fig).
391
392 **TE validation**
393 Validating RelocateTE2's ability to find TEs is important, especially in outbred
394 heterozygous population data where only half of the reads will support a heterozygous TE
395 insertion. RelocaTE2 was previously tested on a rice strain of 20-fold genome coverage of 100
396 bp paired-end Illumina short reads aligned to the rice genome, and resulted in 93% specificity
397 (the percent of calls which were true positives) [26]. The *C. grandiflora* data should have a
398 comparable level of specificity because it has higher coverage than the rice data previously
399 tested. Nevertheless, we used an independent bioinformatics test to estimate the true-positive rate
400 (the proportion of new TEs called by RelocaTE2 that are actually TEs). We ran the SPAdes de

401 novo assembler [54] on all TE calls on scaffold 1 for five individuals, chosen for their high
402 coverage to increase the chance of success of the de novo assembler. The high coverage
403 individuals should not bias the true-positive rate estimate because all individuals were
404 subsampled to the same coverage (70X) for TE calling. We assembled all reads mapping to the
405 region 1000 bp upstream and downstream of the TE insertion call and then aligned the de novo
406 assembly contigs to the TE library using BLAST+ on default settings [55]. If we identified
407 significant hits to the TE library, we considered the TE call to be a true positive.
408 The average true positive value across the 5 individuals was 80% for TEs labelled as
409 high-confidence by RelocaTE2 (see table in Appendix for breakdown by individual). The true
410 positive value for non-high-confidence TEs was 60%. For subsequent analyses, we therefore
411 focused on the high-confidence calls. Analyses considering both high-confidence and low-
412 confidence calls led to similar results, but with lower significance, consistent with the greater
413 noise contributed by the lower-confidence calls (S3 Figure).
414
415 **Gene and genome annotation**
416 TE insertions into genes were categorised as exonic, intronic, 3' UTR or 5' UTR based
417 on the *C. rubella* genome annotation [28]. Similarly, 4-fold SNPs were identified for comparison
418 with TE site frequency spectrum. Conserved non-coding sites (CNSs) are determined through
419 alignment of 9 Brassicaceae species to the *C. rubella* genome as described in [32,45].
420 Recombination rate data comes from fitting a third-order polynomial crossover events from an
421 F2 genetic map derived from a cross between a *C. rubella* and *C. grandiflora* individual [28,56].
422 We ran quadratic multiple linear regression models predicting TE frequency from the
423 proportion of coding sites alone, or both coding sites and CNSs using the R lm function [57]. We

424 used the anova function to test whether the model with two predictor variables explained more
425 variation than the simpler model.
426
427 **SNP dataset**
428 Initial SNP calling and filtering was previously described [25,46]. From the 124
429 individuals used in TE detection, 15 were dropped from further analyses due to an excess of NA
430 SNP genotype calls (over 50,000 NA calls/individual). Any SNPs where more than 9 individuals
431 had an NA call were also dropped. The cutoffs were empirically determined as large outliers
432 when plotting histograms of NA calls per individual and NA calls per SNP. All other NA calls
433 were assumed to be homozygous for the reference allele; given the focus on rare variants, this
434 assumption is unlikely to have a strong effect on results.
435
436 **Burden of rare variants test**
437 The statistical test we used for determining the effect of rare variants (both TEs and
438 SNPs) on gene expression was first described by Zhao and colleagues [33]. Briefly, the test
439 estimates a genome-wide correlation between total rare variant count and gene expression
440 relative to the rest of the population, testing whether individuals with more extreme expression
441 values for a certain gene have more rare genetic variants located near that gene. Individuals are
442 sorted into expression rank bins and the number of rare variants located in and near a gene is
443 summed across all genes in the genome, for each expression rank bin. If there is a burden of rare
444 variants, and having more of them pushes individuals to the extremes of expression (relative to
445 the rest of the population), there will be a quadratic relationship between rare variant count and

446 expression rank. Note that each gene is assessed independently, such that at individual genes,
447 different individuals can end up in the tails of the distribution.
448 The rare variant burden plots were created using TEs and SNPs with a minor allele
449 frequency (MAF) less than 3% from all 109 individuals. For all plots, we used TEs and SNPs
450 which were within a gene or 500 bp upstream or downstream. We identified a small subset of
451 genes where some individuals showed extreme numbers of rare SNPs, possibly reflecting
452 alignment errors or paralogous sequences. Any genes where at least one individual had over 40
453 rare SNPs were removed (874 genes dropped in total).
454 For plots which considered only a single genic region (5'UTR, exon, intron, or 3'UTR),
455 there may be some bias due to genes where UTRs were not annotated. Note that given the
456 residual uncertainty of exact TE insertion locations (approx. 50 bp), this means that there will be
457 some noise associated with annotation of insertion position.
458
459 **Expression outlier test**
460 We developed an alternative method to test the effect of rare burdens on gene expression
461 which allows for a more direct comparison between the effect of TEs and SNPs. Again, we used
462 all variants with a MAF < 3% from the 109 individuals which passed both TE and SNP filtering
463 and which were located within a gene or 500 bp away.
464 The test involves determining individuals which are expression outliers, and comparing
465 their number of rare variants with those of non-outlier individuals, repeated for each gene.
466 Outliers were called as 4 standard deviations away from the mean, where mean and standard
467 deviation were calculated using the expression values of individuals between the 1st and 3rd
468 quartile. Estimating the mean and standard deviation from individuals between the 1st and 3rd

469 quartiles reduces bias in these estimates caused by large expression value outliers and non-
470 symmetrical distributions. As this method underestimates variation (relative to the total sample),
471 an extreme cut-off of 4 standard deviations (SD) was used to call outliers. When testing without
472 discrimination between upper and lower outliers, all genes were used except for 11, in which all
473 individuals were within 4 SD of the mean (total of 18874 genes). When distinguishing between
474 upper and lower outliers, we only considered genes where the mean was at least 4 SD above 0, to
475 ensure it was possible to have lower outliers. After removing genes where all individuals were
476 within 4 SD of the mean, there were 10290 genes tested for upper outliers and 14849 genes
477 tested for lower outliers.

478 For every gene, we computed the number of rare variants per individual to control for
479 differences in number of outliers and non-outliers among and between genes, and checked for a
480 significant difference by running a Wilcoxon signed rank test using the R wilcox.test function for
481 both TEs and SNPs [57]. To determine a per-variant effect size, we calculated the ratio of the
482 mean number of rare variants per outlier individual to the mean number of rare variants per non-
483 outlier individual across all genes. This ratio expresses how many more variants there are per
484 outlier vs non-outlier. Taking the difference between the TE ratio and SNP ratio gives a
485 quantitative measure of how much more likely the average TE is to cause an expression outlier
486 vs the average SNP.

487 We ran 1000 bootstraps of the ratio using the R boot() package in order to calculate 95%
488 quantile confidence intervals [58]. 1000 permutations were also run to build a null distribution,
489 where the outlier vs non-outlier label was switched with a 50% probability for every gene.

490

491 **Expression outlier by exon**

492 We performed a slightly altered version of the expression outlier test in order to control
493 for a possible technical artifact which could lead to the appearance of TE insertions in exons
494 decreasing gene expression. RNA reads that include a TE may be dropped in filtering due to
495 poor mapping, leading to an appearance of reduced overall transcript expression. To test this, we
496 divided total gene expression into individual exon expression. For each gene with a TE insertion
497 in at least one individual in the population, we created three categories: average expression of all
498 exons upstream of the TE insertion, expression of the exon with the TE insertion, and average
499 expression of all exons downstream of the TE insertion. If there was a technical artifact, and
500 RNA reads with TEs were dropped due to poor mapping, we would expect that only exons with
501 TEs had TEs associated with low expression outliers. Upstream and downstream exons wouldn't
502 show TE enrichment in low expression outliers, because their expression would not be affected
503 by poor mapping.
504

505 **Acknowledgements**

506 We thank Aneil Agrawal for key recommendations for statistical methods and Rob Ness
507 for important suggestions on data validation. We thank Young Wha Lee and Niroshini
508 Epitawalage for assistance collecting data. Thomas Bureau, Mathieu Blanchette, Daniel Schoen,
509 Paul Harrison, Alan Moses, Adrian Platts, and Eef Harmsen contributed to the design and
510 implementation of the Value-directed Evolutionary Genomics Initiative (VEGI).
511

512 **References**

513 1. Hurst GD, Werren JH. The role of selfish genetic elements in eukaryotic evolution. Nat Rev
514 Genet. 2001;2: 597–606.

515 2. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution.
516 Nature. 1980;284: 601–603.
517 3. Charlesworth B, Charlesworth D. The population dynamics of transposable elements. Genet
518 Res. 1983;42: 1–27.
519 4. Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, González J. Population genomics of
520 transposable elements in *Drosophila melanogaster*. Mol Biol Evol. 2011;28: 1633–1644.
521 5. Cridland JM, Macdonald SJ, Long AD, Thornton KR. Abundance and distribution of
522 transposable elements in two *Drosophila* QTL mapping resources. Mol Biol Evol. 2013;30:
523 2311–2327.
524 6. Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA,
525 et al. The *Arabidopsis thaliana* mobilome and its impact at the species level. Elife. 2016;5:
526 6919–6931.
527 7. Stuart T, Eichten S, Cahn J, Karpievitch Y, Borevitz J, Lister R. Population scale mapping
528 of transposable element diversity reveals links to gene regulation and epigenomic variation.
529 Elife. 2016;5: e20777.
530 8. Laricchia KM, Zdraljevic S, Cook DE, Andersen EC. Natural variation in the distribution
531 and abundance of transposable elements across the *Caenorhabditis elegans* species. Mol
532 Biol Evol. 2017;34: 2187–2202.
533 9. Barrón MG, Fiston-Lavier A-S, Petrov DA, González J. Population genomics of
534 transposable elements in *Drosophila*. Annu Rev Genet. 2014;48: 561–581.
535 10. Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B. On the role of unequal
536 exchange in the containment of transposable element copy number. Genet Res. 1988;52:
537 223–235.

538 11. Charlesworth B, Lapid A, Canada D. The distribution of transposable elements within and
539 between chromosomes in a population of *Drosophila melanogaster*. II. Inferences on the
540 nature of selection against elements. Genet Res. 1992;60: 115–130.
541 12. Bennetzen JL, Wang H. The contributions of transposable elements to the structure,
542 function, and evolution of plant genomes. Annu Rev Plant Biol. 2014;65: 505–530.
543 13. Lee YCG. The role of piRNA-mediated epigenetic silencing in the population dynamics of
544 transposable elements in *Drosophila melanogaster*. PLoS Genet. 2015;11: e1005269.
545 14. Lee YCG, Karpen GH. Pervasive epigenetic effects of *Drosophila* euchromatic transposable
546 elements impact their evolution. Elife. 2017;6: e25762.
547 15. Wright SI, Agrawal N, Bureau TE. Effects of recombination rate and gene density on
548 transposable element distributions in *Arabidopsis thaliana*. Genome Res. 2003;13: 1897–
549 1903.
550 16. Charlesworth D, Charlesworth B. Transposable elements in inbreeding and outbreeding
551 populations. Genetics. 1995;140: 415–417.
552 17. Wright SI, Schoen DJ. Transposon dynamics and the breeding system. Genetica. 1999;107:
553 139–148.
554 18. Laricchia KM, Zdraljevic S, Cook DE, Andersen EC. Natural variation in the distribution
555 and abundance of transposable elements across the *Caenorhabditis elegans* species. Mol
556 Biol Evol. 2017;34: 2187–2202.
557 19. Cridland JM, Thornton KR, Long AD. Gene expression variation in *Drosophila*
558 *melanogaster* due to rare transposable element insertion alleles of large effect. Genetics.
559 2015;199: 85–93.
560 20. Hollister JD, Smith LM, Guo Y-L, Ott F, Weigel D, Gaut BS. Transposable elements and

561 small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and
562 *Arabidopsis lyrata*. Proc Natl Acad Sci U S A. 2011;108: 2322–2327.
563 21. Wang X, Weigel D, Smith LM. Transposon variants and their effects on gene expression in
564 *Arabidopsis*. PLoS Genet. 2013;9: e1003255.
565 22. González J, Lenkov K, Lipatov M, Macpherson JM, Petrov DA. High rate of recent
566 transposable element-induced adaptation in *Drosophila melanogaster*. PLoS Biol. 2008;6:
567 2109–2129.
568 23. Emera D, Casola C, Lynch VJ, Wildman DE, Agnew D, Wagner GP. Convergent evolution
569 of endometrial prolactin expression in primates, mice, and elephants through the
570 independent recruitment of transposable elements. Mol Biol Evol. 2012;29: 239–247.
571 24. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, et al.
572 Transposable elements contribute to activation of maize genes in response to abiotic stress.
573 PLoS Genet. 2015;11: e1004915.
574 25. Josephs EB, Lee YW, Stinchcombe JR, Wright SI. Association mapping reveals the role of
575 purifying selection in the maintenance of genomic variation in gene expression. Proc Natl
576 Acad Sci U S A. 2015;112: 15390–15395.
577 26. Chen J, Wrightsman TR, Wessler SR, Stajich JE. RelocaTE2: a high resolution transposable
578 element insertion site mapping tool for population resequencing. PeerJ. 2017;5: e2942.
579 27. Dolgin ES, Charlesworth B, Cutter AD. Population frequencies of transposable elements in
580 selfing and outcrossing *Caenorhabditis* nematodes. Genet Res. 2008;90: 317–329.
581 28. Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, et al. The *Capsella*
582 *rubella* genome and the genomic consequences of rapid mating system evolution. Nat
583 Genet. 2013;45: 831–835.

584 29. Tian Z, Zhao M, She M, Du J, Cannon SB, Liu X, et al. Genome-wide characterization of
585 nonreference transposons reveals evolutionary propensities of transposons in soybean. Plant
586 Cell. 2012;24: 4422–4436.
587 30. Novick PA, Smith JD, Floumanhaft M, Ray DA, Stéphane B. The evolution and diversity of
588 DNA transposons in the genome of the lizard *Anolis carolinensis.* Genome Biol Evol.
589 2011;3: 1–14.
590 31. Horvath R, Slotte T. The role of small RNA-based epigenetic silencing for purifying
591 selection on transposable elements in *Capsella grandiflora*. Genome Biol Evol. 2017;9:
592 2911–2920.
593 32. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over
594 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat
595 Genet. 2013;45: 891–898.
596 33. Zhao J, Akinsanmi I, Arafat D, Cradick TJ, Lee CM, Banskota S, et al. A burden of rare
597 variants associated with extremes of gene expression in human peripheral blood. Am J Hum
598 Genet. 2016;98: 299–309.
599 34. Ahmed I, Sarazin A, Bowler C, Colot V, Quesneville H. Genome-wide evidence for local
600 DNA methylation spreading from small RNA-targeted sequences in *Arabidopsis*. Nucleic
601 Acids Res. 2011;39: 6919–6931.
602 35. Metzger BPH, Duveau F, Yuan DC, Tryban S, Yang B, Wittkopp PJ. Contrasting
603 Frequencies and Effects of cis- and trans-Regulatory Mutations Affecting Gene Expression.
604 Mol Biol Evol. 2016;33: 1131–1146.
605 36. Hirsch CD, Springer NM. Transposable element influences on gene expression in plants.
606 Biochimica Biophys Acta. 2016;1860: 157-165

607 37. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from
608 conflicts to benefits. Nat Rev Genet. 2017;18: 71–86.
609 38. Garneau NL, Wilusz J, Wilusz CJ. The highways and byways of mRNA decay. Nat Rev
610 Mol Cell Biol. 2007;8: 113–126.
611 39. Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, et al. Selective silencing of
612 euchromatic L1s revealed by genome-wide screens for L1 regulators. Nature. 2018;553:
613 228–232.
614 40. Kremling KAG, Chen SY, Su MH, Lepak MK, Romay MC, Swarts, KL, et al.
615 Dysregulation of expression correlates with rare-allele burden and fitness loss in maize.
616 Nature. 2018;555: 520–523.
617 41. Ewing AD. Transposable element detection from whole genome sequence data. Mob DNA.
618 2015;6: 24.
619 42. Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. Recent speciation associated
620 with the evolution of selfing in *Capsella*. Proc Natl Acad Sci U S A. 2009;106: 5241–5245.
621 43. Slotte T, Foxe JP, Hazzouri KM, Wright SI. Genome-wide evidence for efficient positive
622 and purifying selection in *Capsella grandiflora*, a plant species with a large effective
623 population size. Mol Biol Evol. 2010;27: 1813–1821.
624 44. St. Onge KR, Källman T, Slotte T, Lascoux M, Palmé AE. Contrasting demographic history
625 and population structure in *Capsella rubella* and *Capsella grandiflora*, two closely related
626 species with different mating systems. Mol Ecol. 2011;20: 3306–3320.
627 45. Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, et al.
628 Evidence for widespread positive and negative selection in coding and conserved noncoding
629 regions of *Capsella grandiflora*. PLoS Genet. 2014;10: e1004622.

630 46. Josephs EB, Wright SI, Stinchcombe JR, Schoen DJ. The relationship between selection,
631 network connectivity, and regulatory variation within a population of *Capsella grandiflora*.
632 Genome Biol Evol. 2017;9: 1099–1109.
633 47. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of
634 Illumina sequence reads. Genome Res. 2011;21: 936–939.
635 48. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput
636 sequencing data. Bioinformatics. 2015;31: 166–169.
637 49. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect
638 break points of large deletions and medium sized insertions from paired-end short reads.
639 Bioinformatics. 2009;25: 2865–2871.
640 50. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: an
641 automatic transposable element classification tool. PLoS One. 2014;9: e91929.
642 51. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified
643 classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8: 973–
644 982.
645 52. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
646 Bioinformatics. 2009;25: 1754–1760.
647 53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
648 Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.
649 54. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a
650 new genome assembly algorithm and its applications to single-cell sequencing. J Comput
651 Biol. 2012;19: 455–477.
652 55. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:

653 architecture and applications. BMC Bioinformatics. 2009;10: 421.
654 56. Slotte T, Hazzouri KM, Stern D, Andolfatto P, Wright SI. Genetic architecture and adaptive
655 significance of the selfing syndrome in *Capsella*. Evolution. 2012;66: 1360–1374.
656 57. R Core Team. R: A language and environment for statistical computing. R Foundation for
657 Statistical Computing, Vienna, Austria. 2016.
658 58. Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-20. 2017.
659

660 ## Supporting Information

661 **S1 Fig. The TE population frequency spectrum made using allele counts.**

662 **S2 Fig. Relationship between TE count and gene expression rank bin for all rare TEs**
663 **within 500 bp of a gene.** TE genotype was considered, where heterozygous TEs were counted
664 once and homozygous TEs were counted twice.

665 **S3 Fig. Relationship between TE count and gene expression rank bin for all rare TEs,**
666 **including low-confidence calls, within 500 bp of a gene.**

667 **S4 Fig. Ratio of average number of rare variants in expression-decreasing outliers relative**
668 **to non-outliers for TEs.** Expression is calculated separately for each exon. 'Up' refers to
669 average expression of all exons upstream of where the TE inserted. 'In' refers to the exon where
670 the TE inserted. 'Down' refers to average expression of all exons downstream of where the TE
671 inserted.

672 **S5 Fig. Relationship between average TE frequency and average cross-over rate in 100 kb**
673 **bins across the genome (Pearson's r = 0.039).**

674 **S6 Fig. Relationship between permuted TE count and gene expression rank bin for all rare**
675 **TEs within 500 bp of a gene.**

676 **S7 Fig. Relationship between TE count and gene expression rank bin for different TE**
677 **types.**
678 **S8 Fig. Relationship between TE count and gene expression rank bin divided by TE**
679 **insertion location relative to gene.** Upstream/downstream refers to 500 bp away from gene
680 annotation start/end, respectively.
681 **S9 Fig. Ratio of average number of rare variants in expression outliers relative to non-**
682 **outliers for TEs inserted upstream of gene transcription start site.** Upstream intervals are
683 non-overlapping. 'Upper' refers to expression-increasing outliers, and 'lower' refers to
684 expression-decreasing outliers.
685 **S10 Fig. (A-D) Relationship between TE count and gene expression rank bin divided by**
686 **genic region of the insertion.**
687 **S1 Table. TE validation statistics showing the proportion of false positives, as determined**
688 **through an independent bioinformatics approach, for five tested individuals.**
689 **S1 File. All TE calls used in analyses across the population.**