Title
The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata

Permalink
https://escholarship.org/uc/item/0gg2f58v

Authors
Lockton, Steven
Gaut, Brandon S

Publication Date
2010-01-12

License
CC BY 4.0

Peer reviewed
The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*

Steven Lockton*, Brandon S Gaut

**Abstract**

**Background:** Transposable Elements (TEs) make up the majority of plant genomes, and thus understanding TE evolutionary dynamics is key to understanding plant genome evolution. Plant reproductive systems are diverse and mating type variation is one factor among many hypothesized to influence TE evolutionary dynamics. Here, we collected a large TE-display data set in self-fertilizing *Arabidopsis thaliana*, and compared it to data gathered in outcrossing *Arabidopsis lyrata*. We analyzed seven TE families in four natural populations of each species to tease apart the effects of mating system, demography, transposition, and selection in determining patterns of TE diversity.

**Results:** Measures of TE band differentiation were largely consistent across TE families. However, patterns of diversity in *A. thaliana* Ac elements differed significantly from that other TEs, perhaps signaling a lack of recent transposition. Across TE families, we estimated higher allele frequencies and lower selection coefficients on *A. thaliana* TE insertions relative to *A. lyrata* TE insertions.

**Conclusions:** The differences in TE distributions between the two *Arabidopsis* species represents a synthesis of evolutionary forces that include the transposition dynamics of individual TE families and the demographic histories of populations. There are also species-specific differences that could be attributed to the effects of mating system, including higher overall allele frequencies in the selfing lineage and a greater proportion of among population TE diversity in the outcrossing lineage.

**Background**

Transposable elements (TEs) are prevalent in plant genomes [1] and ubiquitous among eukaryotes [2]. Although TEs comprise most of an average plant genome [3], their content varies markedly among populations [4,5] and species [6,7]. For example, TEs make up ~70% of the maize genome [8] but just 10% of the *Arabidopsis thaliana* genome [9]. Moreover, TEs can accrue rapidly after polyploid and hybrid speciation events [10,11]. These observations raise questions about the evolutionary forces that govern the distribution of TEs within plant genomes.

Population genetics has the potential to illuminate these forces, but our understanding of the population genetics of TEs has been based primarily on studies of *Drosophila melanogaster*. These studies have revealed that there are far fewer TE insertions in the *D. melanogaster* genome than possible insertion sites [12,13] and that insertions tend to be at low population frequencies [12-14]. Both observations suggest that the spread of TEs is countered by natural selection [15-18]. However, the precise nature of selection against insertions is unclear. Some insertions may disrupt gene products or gene expression [19]. Purifying selection against these deleterious insertions could be the driving force that governs observed TE distributions [15,19-23]. Another possibility is that TEs facilitate deleterious chromosome rearrangements through non-homologous (or ectopic) recombination [18,24-28].

The mating system of host species is likely to be an important factor that shapes TE dynamics [27,29,30]. For example, in highly homozygous selfing species most
TEs have a paired homologous allelic partner, which reduces the probability of an ectopic recombination event [27,29,30]. If selection against TEs is primarily mediated by these ectopic events, then selfing species are predicted to have less efficacious selection and higher TE copy numbers than outcrossing species. Conversely, the deleterious effects of recessive TE insertions are expected to be stronger in a homozygous selfer, which may result in more efficacious selection and fewer TEs in selfers [20,29,30]. Thus, the effect of breeding system is difficult to predict precisely, but simulations of TE population dynamics provide evidence to support the possibility that both ectopic recombination and deleterious insertions will lead to differences in TE accumulation between selfers and outcrossers [29,30].

Mating system influences the efficacy of selection against TEs in at least two other ways: First, the effective population size \( N_e \) in a selfing species is expected to be half that of an otherwise identical outcrosser [31,32]. Population size has a direct effect on the efficacy of selection, because efficacy is reflected in the compound parameter \( N_e s \), where \( s \) is the strength of selection. It is thus not surprising that empirical studies suggest that shifts in \( N_e \) over time influence the number and frequency of TEs [5,33]. Second, inbreeding reduces the effective recombination rate, which may lead to the accumulation of weakly deleterious TE insertions [34] via Hill-Robertson effects [35]. Observations that TEs accumulate on non-recombining sex chromosomes support this conjecture [36,37].

Despite predictions that TE population dynamics may differ markedly between selfing and outcrossing species, comparative data are quite rare. Recently, however, Dolgin et al. [38] documented that population frequencies of \( \text{Tc1-like} \) insertions are higher in selfing \( \text{Caenorhabditis elegans} \) than in outcrossing \( \text{C. remanei} \). This pattern of diversity suggests less efficacious selection against insertions in the selfing species; indeed, Dolgin et al. [38] tentatively conclude that \( \text{Tc-1} \) element insertions are effectively selectively neutral in \( \text{C. elegans} \).

Plants are particularly well suited for inter-species comparisons of TE population dynamics because of broad diversity in mating systems. Studies of selfing and cultivated \( \text{Lypersicon} \) species have generally shown differences in TE complement that are consistent with less efficacious selection against insertions in selfing species. For example, the \( \text{Lyt1} \) element family has higher copy numbers in the selfing members of the genus [27,39], and \( \text{copia-like} \) insertions are generally found at higher population frequencies in selfers [40]. In perhaps the best known study TE diversity between plant species with contrasting mating systems [41], Wright et al. [42] compared population diversity of \( \text{Ac-like} \) elements between selfing \( \text{Arabidopsis thaliana} \) and outcrossing \( \text{A. lyrata} \). \( \text{Ac-like} \) insertions were slightly more numerous in \( \text{A. thaliana} \) but segregated at significantly lower frequencies in \( \text{A. lyrata} \), consistent again with reduced efficacy of natural selection against insertions in the selfing lineage.

Although the limited data gathered to date suggests that selection against TEs is less efficacious in selfing lineages, it is difficult to determine whether extant patterns of TE diversity are due to the effects of selection or complicated by other factors that may differ between species, such as demographic history and transposition dynamics [42]. How might one discriminate among these factors? One approach is to increase sampling to multiple TE families and multiple populations. If patterns of TE diversity vary across element families, transposition dynamics may play a major role in explaining differences between species like those observed for \( \text{Aclike} \) elements [42]. In contrast, if diversity patterns are consistent across TE families, forces that affect whole genomes (such as demography and breeding system) may be the primary determinants of TE diversity. Here we extend the study of Wright et al. [42] to contrast TE population genetics between \( \text{A. thaliana} \) and \( \text{A. lyrata} \), generating polymorphism profiles from four populations of \( \text{A. thaliana} \) representing seven TE families. We compare these \( \text{A. thaliana} \) data to data gathered from four populations of the outcrossing congener \( \text{A. lyrata} \) [5]. By contrasting TE frequencies and patterns across species, populations, and TE families, we gain insight into the relative roles of transposition, demography, and breeding system in shaping TE diversity.

**Methods**

We sampled four populations of \( \text{A. thaliana} \) with seed obtained from The \( \text{Arabidopsis Information Resource (TAIR)} \) [43]. The sample included 12 individuals from Ascot, U.K. (TAIR seed stock numbers CS22220-CS22235), 12 from Anholt, Germany (CS22313-CS22324), 12 from Knox, Indiana, USA (CS22401-CS22412), and 11 individuals from Cold Spring Harbor, New York (CS22419-CS22430). Plants were grown in a growth chamber for eight weeks, and DNA was extracted from leaf material. Our TE display procedure followed [5], including the extensive technical replication, to produce \( \text{A. thaliana} \) TE polymorphism data for \( \text{Ac-like} \) III (henceforth “\( \text{Ac} \)”; \( \text{Helitron Basho} \) TEs (“\( \text{Basho} \)”; \( \text{CACTA} \); \( \text{Gypsy-like } (\text{“Gypsy”)}; \text{LINE-like } (\text{“LINE”)}; \text{SINE-like I } (\text{“SINE”)}; \text{and } \text{Tourist-like } \text{MITE} \) (“\( \text{MITE}) \) elements. These TEs represent three RNA-mediated class I retrotransposons (LINE, SINE and \( \text{Gypsy} \) and four class II DNA transposons (\( \text{Ac}, \text{Basho}, \text{MITE} \) and CACTA). The primers used to generate \( \text{Ac} \) TE-display data were identical to those used by Wright et al. [42]. We also utilized the TE display data from
[5], encompassing 44 individuals from four natural A. lyrata populations; 11 individuals from Plech, Germany, 12 from Karhumäki, Russia, 12 from North America, and nine from Stubbsan, Sweden.

Molecular Analysis of Variance
To measure levels of population differentiation in our A. thaliana sample, we performed a Molecular Analysis of Variance (AMOVA) [44]. We focused on \( F_{ST} \), a statistic analogous to \( F_{X} \) that measures genetic differentiation among populations. For our analyses, we used TE-display bands as genetic markers, and thus \( F_{ST} \) became a measure of TE display band differentiation. Our analyses were performed with two different packages: GenAlEx 6 [45] was used to compare \( F_{ST} \) between populations and the R package ade4 [46] was used to calculate \( F_{ST} \) among all populations).

Allele frequencies and copy numbers
We compared our A. thaliana TE-display data to A. lyrata data by estimating TE allele frequencies and copy numbers in both species. We used estimates of the inbreeding coefficient (\( F \)) to estimate TE allele frequencies from dominant TE display data. For each A. lyrata population, \( F \) was estimated independently using (SNP) data in 77 loci [47] by \( \hat{F} = 1 - (\frac{\hat{F}_{OBS}}{\hat{F}_{EXP}}) \), where \( \hat{F} \) is the estimated inbreeding coefficient, \( \hat{F}_{OBS} \) is the average observed heterozygosity per locus [48], and \( \hat{F}_{EXP} \) is the average expected heterozygosity, under random mating, calculated by

\[
H_{exp} = 1 - \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} p_i^2
\]

where \( p_i \) is the frequency of the \( i \)th of \( k \) alleles, summed over the \( i \)th of \( m \) SNP loci [48].

Ross-Ibarra et al. [47] sampled the same German, Russian, and Swedish populations, and \( F \) was estimated directly for these populations. However, they sampled two North American populations (Ontario, Canada and Indiana, USA) that were combined to yield our North American sample. We thus average \( \hat{F} \) between these two populations to procure an estimate of \( F \) for our North American sample. To estimate \( F \) in A. thaliana, we assumed the proportion of selfing (\( S \)) in A. thaliana was 0.99 [49] and estimated \( F = 0.98 \) from the relation \( F = S/(2 - S) \) [50].

Given estimates of \( F \), we estimated \( p_{TE} \), the TE allele frequency, using

\[
z = q^2(1 - \hat{F}) + q\hat{F}
\]

[51], where \( z \) is the observed frequency of the null TE display band (\( i.e. \), 1 - the population frequency of the dominant TE band), \( q \) is the estimated null TE allele frequency, and \( q = 1 - p_{TE} \). We calculated allele frequency estimates both within populations and across entire species’ samples.

We calculated \( n_{TE} \), the expected TE copy number of an individual, as:

\[
n_{TE} = \sum_{i=1}^{m} \left[ \frac{2(\hat{p}_i^2 + \hat{p}_i(1-\hat{p}_i)\hat{F}) + 2\hat{p}_i(1-\hat{p}_i) - 2\hat{p}_i(1-\hat{p}_i)\hat{F}}{\hat{p}_i^2 + \hat{p}_i(1-\hat{p}_i)\hat{F} + 2\hat{p}_i(1-\hat{p}_i) - 2\hat{p}_i(1-\hat{p}_i)\hat{F}} \right] \times I_i \] (2)

where \( \hat{p}_i \) is TE allele frequency of the \( i \)th locus, summing over \( m \) TE loci. \( I_i \) is an indicator variable, where \( I_i = 1 \) when a TE band is present, and \( I_i = 0 \) when a TE band is absent, at the \( i \)th locus in a given individual. Bands fixed within our sample were included in our calculations of allele frequencies and \( n_{TE} \).

Estimation of selection coefficients
We used the Maximum Likelihood (ML) approach of Petrov et al. [18], with modifications introduced by Lockton et al. [5], to estimate the population-selection coefficients (\( N_s, s \)) from our TE display data. Lockton et al. [5] modified the method to correct for ascertainment biases inherent in TE-display data and also to employ \( \hat{F} \). In this method, \( N_s, s \) is compound parameter, but following Petrov et al. [18] we assume values for \( N_e \) based on nucleotide polymorphism data. We used the point estimates of \( N_e \) inferred from demographic modeling of the same four A. lyrata populations - i.e., Germany \( N_e = 136,000 \); North America \( N_e = 11,000 \); Russia \( N_e = 12,000 \); and Sweden \( N_e = 12,000 \) (Ross-Ibarra et al., 2008). Species-wide A. lyrata \( N_e \) was calculated to be 250,000 by using estimates of \( \theta \) from SNP data [47], and assuming a mutation rate (\( \mu \)) of 1.5 \times 10^{-8} [52]. For A. thaliana, we also used estimates of \( \theta \) from SNP diversity data [53] to estimate \( N_e \), assuming \( \mu = 1.5 \times 10^{-8} \) [52]. Species-wide \( N_e \) was estimated to be 125,000; the UK population was 98,500; Germany, 83,000; and both New York and Indiana was 71,000. However, results did not differ qualitatively when \( N_e \) was assumed to be 125,000 in each A. thaliana population (data not shown).

Results & Discussion
TE Display Bands
We identified 267 TE display bands in A. thaliana across seven TE families. To compare, in A. lyrata, 274 bands were amplified in six TE families [5]. Of the six TE families shared between species (Ac, CACTA, Gypsy, LINE, MITE, and SINE) there were more TE bands in outcrossing A. lyrata (\( n = 274 \)) than in the self-fertilizing A. thaliana (\( n = 210 \)). A sample of the bands amplified using TE family-specific primers were cloned, sequenced, and identified: 95% (20/21) of the A. thaliana bands were
successfully identified as TEs belonging to their respective families (data not shown). The single unidentified *A. thaliana* Ac band showed strong sequence similarity to an “unknown protein” (BLASTn e-value: 5e-102) in the *A. thaliana* genome sequence.

We readily identified 57 Basho bands in *A. thaliana*, but few strong bands in *A. lyrata*. The putative Basho bands that were amplified, cloned, and sequenced from *A. lyrata* could not be identified in TE databases using BLAST. Because of the uncertainty of the *A. lyrata* Basho data, they were not included in additional analyses. These empirical results are consistent with previous studies suggesting that some Basho subfamilies may be absent from *A. lyrata* [54].

**Molecular Analysis of Variance**

We utilized an AMOVA to examine *A. thaliana* band differentiation between populations for each TE family (Fig. 1). Overall, *A. thaliana* tends to have lower levels of $\Phi_{PT}$ between populations relative to *A. lyrata* [5]. Higher $\Phi_{PT}$ values for *A. lyrata* are consistent with its more disjunct distribution [55,56], its high nucleotide diversity [47], and its relatively large and stable populations [57]. Nonetheless, $\Phi_{PT}$ values between *A. thaliana* populations are typically significantly > 0, as might be expected of a species that is increasingly recognized as having considerable population structure [58-60]. The $\Phi_{PT}$ values mirror geographic distances in some cases. For example, the lowest $\Phi_{PT}$ tended to be between the populations

![Figure 1 AMOVA $\Phi_{PT}$ per TE family in each population pairwise comparison in both species. *A. lyrata* data from [5]. *A. thaliana* populations: UK, Ascot, United Kingdom; Ger, Anholt, Germany; IN, Knox, Indiana, USA; NY, Long Island, New York, USA. *A. lyrata* populations: Ger, Plech, Germany; N.Am, North America; Rus, Russia; Swe, Sweden. "N"s indicate comparisons that show no significant population differentiation ($\Phi_{PT} = 0$, $p > 0.05$).](image)
geographically closest to each other (Fig. 1; Indiana, USA, and Long Island, NY, USA; non-significant in 4/7 TE families), which may reflect low genetic structure among human-commensal North American \textit{A. thaliana} populations [53,61]. It is striking, however, that for most TE families we also observe low $\Phi_{PT}$ between the UK and the US population samples (Fig. 1). Indeed, our UK sample appeared to have less TE band differentiation vs. each of the U.S. samples than UK vs. Germany (Fig. 1), even though the two European populations are closer geographically. The striking difference between the UK and German populations could reflect, in part, patterns of migration across Europe [60] and particularly the possibility of admixture in Central Europe from separate glacial refugia [58,62]. To our knowledge genetic similarity between US and UK populations has not been noted previously.

Arguably the most striking aspect of pairwise $\Phi_{PT}$ values is relatively low values for \textit{A. thaliana} \textit{Ac} elements (Fig. 1), suggesting that \textit{Ac} population dynamics differ from those of the other TEs surveyed. To test this idea more formally, we estimated total $\Phi_{PT}$ values among all populations for each TE family, and then compared the observed values to $\Phi_{PT}$ from bootstrapped replicates. The bootstrap samples were based, first, on combining bands across TEs, under the null hypothesis that all TE families are representative of a homogeneous evolutionary process. Then, for each TE family, bootstrap replicates mimicked the observed number of bands from each population and, finally, $\Phi_{PT}$ was calculated for each replicate. From this exercise, it is clear that $\Phi_{PT}$ from \textit{A. thaliana} \textit{Ac} is much lower than expected under the null hypothesis ($p = 0.003$; Fig. 2). In contrast, data from \textit{A. lyrata} \textit{Ac} elements did not reject the null hypothesis of homogeneity (Fig. 2), nor did TE data from any other TE family in either species after multiple-test correction (data not shown). Thus, population genetic information does vary among TE families, with \textit{A. thaliana} \textit{Ac} an obvious outlier.

We also estimated variance components for each TE family in both species using AMOVA (Fig. 3). If breeding system has an appreciable effect on TE diversity, a selfer should exhibit less TE band diversity within each population than among populations, because inbreeding leads to populations with low genetic diversity [63]. Our data are consistent with this prediction: Among-population variation was proportionally higher in \textit{A. thaliana} compared to \textit{A. lyrata} across all TE families (two-tailed sign test, $p = 0.03$; Fig. 3). One must be careful about interpreting these results, however, as differences in sampling could contribute to apparent differences between species. Indeed, our \textit{A. thaliana} TE-display suggests that our within-population variation is a smaller component than found in a previous population study based on combined microsatellite and SNP data [59]. Nonetheless, the partitioning of variation is consistent across TE families, and does suggest some genomewide effect of species with regard to the partitioning of TE variation. In addition, variance components graphically demonstrate that \textit{A. thaliana} \textit{Ac} elements differ from other elements with regard to the distribution of diversity (Fig. 3).

**TE insertion frequencies and the strength of selection**

AMOVA utilizes TE-display bands, but allele frequencies are often more helpful for evaluating evolutionary dynamics. For TE-display data, a band from an inbred species is more likely to represent a homozygous locus.
than bands in an outcrossing species. We therefore used independent estimates of the inbreeding coefficient \((F)\) to estimate allele frequencies \((p_{TE})\) from our TE band data (see Methods). This method intrinsically corrects for possibilities that \(A.\ lyrata\) may not always be obligately outcrossing (e.g., [64]) but does assume a constant rate of selfing in \(A.\ thaliana\). With \(p_{TE}\) estimates, we can examine site frequency spectra (SFS), which form the basis for inferring the strength of selection [65,66]. We combined data across populations to construct species-wide samples. Species-wide, and across all TE families, the \(A.\ lyrata\) median \(p_{TE}\) of 0.061 was substantially lower for a constant rate of selfing in \(A.\ thaliana\). With \(p_{TE}\) estimates, we can examine site frequency spectra (SFS), which form the basis for inferring the strength of selection [65,66]. We combined data across populations to construct species-wide samples. Species-wide, and across all TE families, the \(A.\ lyrata\) median \(p_{TE}\) of 0.061 was substantially higher (Table 1; Wilcoxon rank sum test, \(p = 1.28 \times 10^{-6}\)). In addition to lower median frequencies, the \(A.\ lyrata\) SFS showed a skew towards lower frequency insertions in \(A.\ lyrata\) relative to \(A.\ thaliana\) (Fig. 4). This skew is evident not only for the SFS pooled among TE families, but also for most individual TEs (Fig. 4). The standard interpretation of a left-skewed SFS is that purifying selection acts on deleterious variants, limiting their population frequencies. Thus, the skew in \(A.\ lyrata\) relative to \(A.\ thaliana\) is consistent with stronger selection acting on TEs in \(A.\ lyrata\), as concluded by Wright et al. [42] for \(Ac\) elements alone.

If we assume a transposition-selection equilibrium, the strength and direction of selection can be estimated from the SFS using an ML framework [18]. The ML approach, as implemented here, incorporates information about inbreeding \(F\) into the model [5]. Applying this approach, \(A.\ thaliana\)’s \(N_{S}\) estimates for pooled TEs were not significantly different from zero, and individual TEs pooled across populations also yielded \(N_{S}\) estimates very close to zero (Fig. 5; SI Table 1). The corresponding estimates were lower in \(A.\ lyrata\) [5], and the total sample of TEs yielded an \(N_{S}\) point estimate significantly less than zero, at -1.9 (Fig. 5; SI Table 1).

The SFS and \(N_{S}\) results are consistent with a species-wide reduction in the strength of selection in the selfing species compared to the outcrossing species. This lends superficial support either to mechanisms of selection (such as ectopic recombination) that are hypothesized to be more prevalent in an outcrosser, or to complicating factors in selfers (such as interference due to linkage or smaller population sizes) that slow selection [18,20,22,26]. However, these ‘species-wide’ results could also be an artifact of combining samples across populations. Under this line of reasoning, the skewed SFS in \(A.\ lyrata\) relative to \(A.\ thaliana\) may come from combining data from relatively more diverged populations (as measured by \(F_{PT}\); Fig. 1) that share little variation. The pooled sample from highly diverged populations would thus consist of predominantly low frequency variants.

To address this issue, we estimated \(N_{S}\) separately for each of the populations. The estimates for each \(A.\ lyrata\) population were slightly greater than zero in all populations except Germany (Additional file 1: Table S1). In contrast, the \(A.\ thaliana\) per-population estimates are slightly negative for two population samples (IN, USA, -0.044; UK, -0.702), slightly positive for a third (NY, USA, 0.169), and undefined (not estimable) for the fourth population (Germany; Additional file 1: Table S1). In this context, it is also important to remember that the models used to estimate \(N_{S}\) values assume constant population sizes, and selection-transposition equilibrium [18,33]. As noted previously [5], positive estimates may be misleading because they reflect demographic forces (presumably population bottlenecks during colonization) in the history of individual populations more than selective strength. Many of these assumptions are probably not valid for \(A.\ lyrata\) populations [47].
but the impact of these assumptions on *A. thaliana* data are less clear. If, for example, *A. thaliana* follows particular kinds of metapopulation dynamics [67], then the approach may be reasonable.

These considerations make it difficult to determine whether there really is a systematic difference in SFS between the outcrosser and the inbreeder. However, *A. lyrata* does trend toward lower average allele frequencies. For example, averaging $p_{TE}$ within populations and taking a grand average across populations, we find that *A. lyrata* has a grand average $p_{TE}$ of 0.250 (sd ± 0.033) and *A. thaliana* has slightly higher grand average of 0.268 (sd ± 0.030). Similar trends are produced by taking the average of medians across populations: *A. lyrata*: 0.189 ± 0.050; *A. thaliana*: 0.198 ± 0.048. Thus, in considering individual populations - as opposed to ‘species-wide’ samples - there is a slight trend toward lower population frequencies of TEs in *A. lyrata*, consistent with the notion that selection against TEs is stronger in the outcrosser. However, the effect is much muted relative to the species-wide sample, suggesting that population differentiation and demography contributes to some of the differences between pooled, species-wide samples.

**Copy number estimates**

The final way in which we compare TEs among populations and between species is by estimating copy number, $n_{TE}$ (see Methods). If selection is more effective in a heterozygous outcrosser, $n_{TE}$ is predicted to be lower in *A. lyrata* than in *A. thaliana*. Per-population $n_{TE}$, summed over all six shared TE families (Table 2), were not significantly different between species (Wilcoxon rank sum test, $p = 0.649$). Among *A. thaliana* TE families, median $n_{TE}$ varied per population, but $n_{TE}$ varied among populations more in *A. lyrata* (Table 2), perhaps again reflecting relatively higher divergence among *A. lyrata* populations (Fig. 1). In *A. lyrata*, as previously noted [5], there is a trend towards lower $n_{TE}$ for each TE family in the German population vs. Russia, Sweden, and North America, but no similar clear pattern emerged in the *A. thaliana* data (Table 2).

We observed higher $Ac$ copy numbers in *A. lyrata* than in *A. thaliana* in our population-level samples, as measured by the number of TE loci amplified (*A. thaliana n = 25 vs. A. lyrata n = 54) or by $n_{TE}$ (mean per population $n_{TE}$ 12.44 vs. 21.97; Table 2). These results directly contradict those of Wright et al. [42], who detected more copies of $Ac$ in *A. thaliana* than in *A. lyrata*. Although our $Ac$ bands were amplified using the primers of Wright et al. [42], the sampling strategies differed markedly between studies. The samples of Wright et al. were “species-wide” but uneven, in that 15 populations were represented by a single individual but four populations were represented by $> 6$
individuals [42]. We believe the contrasting results between studies highlight the effect that sampling can have on subsequent inferences.

Overall, our study is like previous studies in that we detect apparent allele frequency differences between an outcrossing and an inbreeding species but no systematic differences in $n_{TE}$ [38,40,42]. However, several features of our data must be kept in mind: First, the PCR primers were designed from *A. thaliana* genomic sequence, causing a potential ascertainment bias between species. While this bias should not cause difficulties for frequency estimates - which are conditioned on observing a band at an insertion site - this bias could lead to an underestimate of the number of insertion sites in *A. lyrata*. Thus, extrapolating from allele frequencies ($p_{TE}$) to copy number ($n_{TE}$) could lead to a systematic underestimate of copy number in *A. lyrata*. Second, TE display

![Figure 4 Site Frequency Spectra](image)

![Figure 5 Maximum Likelihood Ns estimates](image)
bands represent TE sequence found on a limited range of band sizes (~50 to 1000 bases); if there are general differences in TE sizes between species then copy number comparisons may be inaccurate. There is reason to believe that this would trend toward an underestimate in A. lyrata: for example, early comparisons of gene structure between the congeners suggest that introns are generally larger in A. lyrata [68] and may contribute to the 1.5-fold difference in genome size between the two species [69]. Finally, it is very important to remember that the TE-display protocol amplifies TEs that represent clades or subfamilies of TE families and not entire families. For example, in A. thaliana we amplified 52 different MITE elements from the Tourist-like subfamily, whereas at least 818 MITEs have been found in the A. thaliana genome sequence [9].

Conclusions
The motivation for this study was to determine whether observed differences in the frequency and population dynamics of TEs can be attributed to species-wide effects, which presumably reflect differences between outcrossing and selfing mating systems, or are better attributed to factors like transposition dynamics and demographic history that may also differ between species. Our study is unique in that we sampled multiple TE families and multiple populations to compare population dynamics between a selfer (A. thaliana) and an outcrosser (A. lyrata).

Our results indicate that patterns of genetic diversity are heterogeneous across two of the seven surveyed TE families. Unlike other elements, Bashos were amplifiable within A. thaliana but apparently absent from A. lyrata. These observations are consistent with molecular evolutionary analyses that suggest recent bursts of Basho insertions within A. thaliana [70] and an apparent lack of some Basho subfamilies from A. lyrata [54]. Ac element diversity also differed substantially from other element families, exhibiting low levels of TE band diversity (Fig. 1) and statistically low values of $\Phi_{PT}$ (Fig. 2) within A. thaliana. These Ac observations could be consistent either with a lack of recent transposition or particularly strong selection targeting new insertions. In any case, our Basho and Ac results clearly demonstrate that TE dynamics can vary between species and among TEs. They also underline the importance of sampling multiple TE families to make robust inferences about TE dynamics. Although few analyses of TE have studied more than one TE family (albeit in a single species and ignoring between population variation, e.g., [18]), virtually all previous population genetic analyses of TE diversity and mating systems have analyzed data from a single TE family and generalized about mating system dynamics from this single observation [38,40,42].

<table>
<thead>
<tr>
<th>Table 2 Median per individual TE copy number (n)</th>
<th>A. lyrata</th>
<th>A. thaliana</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plech, Germany</td>
<td>12.97</td>
<td>7.76</td>
</tr>
<tr>
<td>North Russia</td>
<td>17.27</td>
<td>20.44</td>
</tr>
<tr>
<td>Sweden</td>
<td>21.97</td>
<td>20.18</td>
</tr>
<tr>
<td>Mean</td>
<td>13.93</td>
<td>24.4</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>3.87</td>
<td>4.56</td>
</tr>
</tbody>
</table>

Ac 17.37 23.89 20.44 26.18 21.97 3.87 13.93 11.95 11.92 11.97 12.44 0.99
Basho ----- 29.96 40.72 35.76 33.81 35.06 4.48
CACTA 7.19 7.83 6.67 8.66 7.57 8.83 9.92 10.9 11.92 11.96 12.49
Sum* 78.78 103.1 96.79 107.37 96.51 12.59 94.65 101.39 97.08 92.33 96.36 3.87

* A. thaliana nTE sums exclude Basho insertions.
generalizations inherently assume that a single TE family represents the TE complement within a genome, and this may be a poor assumption.

Our data also clearly demonstrate that demographic history shapes TE diversity, because pairwise comparisons involving geographically closer populations often have lower $\Phi_{PT}$ values (Fig. 1). In *A. lyrata*, demographic events perturb selection-transposition equilibria and influence the distribution and frequency of TEs [5,33,47]. Presumably demographic events play a similar role in *A. thaliana*, although the magnitude of these effects is difficult to estimate with the present data.

Finally, some aspects of the data cannot be easily attributed to demography or transposition and thus may reflect differences due to host mating system. These include: the apportionment of diversity within and between populations (Fig. 3); an SFS that provides a consistently higher signal of higher TE population frequencies in *A. thaliana* at both species-wide (Fig. 4) and population scales; and apparent differences in selection coefficients (Fig. 5). Generally, our results contribute to a growing empirical literature that suggests reduced efficacy of selection against TE insertions within selfing lineages [38,40,42], but many questions remain to be addressed about the generality of this observation across taxa and the relative importance of the mechanisms (e.g., ectopic recombination, reduced population sizes, lower effective recombination rates) that may contribute to this effect.

**Additional file 1: Ns maximum likelihood estimates**

Maximum likelihood estimates of the strength of selection, including 95% confidence intervals, for TEs pooled across populations and across TE families in both species. Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-10-10-S1.XLS](http://www.biomedcentral.com/content/supplementary/1471-2148-10-10-S1.XLS)

**Acknowledgements**

We thank N. Komarova and S. I. Wright for assistance and discussion; R. Gaut for technical assistance; E. Thorhillscottitr, M. Claus, O. Savolainen, and B. Mable for seed material; and two anonymous reviewers for helpful comments. The work was supported by NSF grants DEB-0426166 and DEB-0723860 to B.S.G.

**Authors’ contributions**

SL carried out the molecular biology work, analyzed the data, and drafted the manuscript. BSG designed the study, assisted with data analysis, and helped draft the manuscript. Both authors read and approved the final manuscript.

**Received:** 14 July 2009  
**Accepted:** 12 January 2010  
**Published:** 12 January 2010

**References**


Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Popul Biol 1968; 3(7):100(1).


Submit your next manuscript to BioMed Central and take full advantage of:

• Convenient online submission
• Thorough peer review
• No space constraints or color figure charges
• Immediate publication on acceptance
• Inclusion in PubMed, CAS, Scopus and Google Scholar
• Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit