BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

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

* Correspondence: slockton@gmail.com
Department of Ecology and Evolutionary Biology, University of California, Irvine, USA

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 *Tc1*-like insertions are higher in selfing *Caenorhabditis elegans* than in outcrossing *C. remanei*. This pattern of diversity suggests less efficacious selection against insertions in the selfing species; indeed, Dolgin *et al*. [38] tentatively conclude that *Tc-1* element insertions are effectively selectively neutral in *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 *Lypersicon* species have generally shown differences in TE complement that are consistent with less efficacious selection against insertions in selfing species. For example, the *Lyt1* element family has higher copy numbers in the selfing members of the genus [27,39], and *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 *Ac*-like elements between selfing *Arabidopsis thaliana* and outcrossing *A. lyrata*. *Ac*-like insertions were slightly more numerous in *A. thaliana* but segregated at significantly lower frequencies in *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 *Ac*-like 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 *A. thaliana* and *A. lyrata*, generating polymorphism profiles from four populations of *A. thaliana* representing seven TE families. We compare these *A. thaliana* data to data gathered from four populations of the outcrossing congener *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 *A. thaliana* with seed obtained from The *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 *A. thaliana* TE polymorphism data for *Ac*-like III (henceforth "*Ac*"); *Helitron Basho* TEs ("*Basho*"); CACTA; *Gypsy*-like ("*Gypsy*"); LINE-like ("LINE"); SINE-like I ("SINE"); and *Tourist*-like MITE ("MITE") elements. These TEs represent three RNA-mediated class I retrotransposons (LINE, SINE and *Gypsy*) and four class II DNA transposons (*Ac, Basho*, MITE and CACTA). The primers used to generate *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 Stubbsand, 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 Φ_{PT}, a statistic analogous to F_{ST} that measures genetic differentiation among populations. For our analyses, we used TE-display bands as genetic markers, and thus Φ_{PT} became a measure of TE display band differentiation. Our analyses were performed with two different packages: GenAlEx 6 [45] was used to compare Φ_{PT} between populations and the R package ade4 [46] was used to calculate Φ_{PT} 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 $\hat{F} = 1 - (\bar{H}_{Obs} / \bar{H}_{Exp})$, where $\hat{F}$ is the estimated inbreeding coefficient, $\bar{H}_{Obs}$ is the average observed heterozygosity per locus [48], and $\bar{H}_{Exp}$ is the average expected heterozygosity, under random mating, calculated by

$$H_{\exp} = 1 - \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{k} p_i^2,$$

where p_i is the frequency of the ith of k alleles, summed over the lth 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

$$z = q^2(1 - \hat{F}) + q\hat{F} \quad (1)$$

[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:

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

where p_i is TE allele frequency of the ith 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 ith 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 ($N_e s$) 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, $N_e s$ 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 θ from SNP data [47], and assuming a mutation rate (μ) of 1.5×10^{-8} [52]. For *A. thaliana*, we also used estimates of θ 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 Φ_{PT} between populations relative to *A. lyrata* [5]. Higher Φ_{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, Φ_{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 Φ_{PT} values mirror geographic distances in some cases. For example, the lowest Φ_{PT} tended to be between the populations

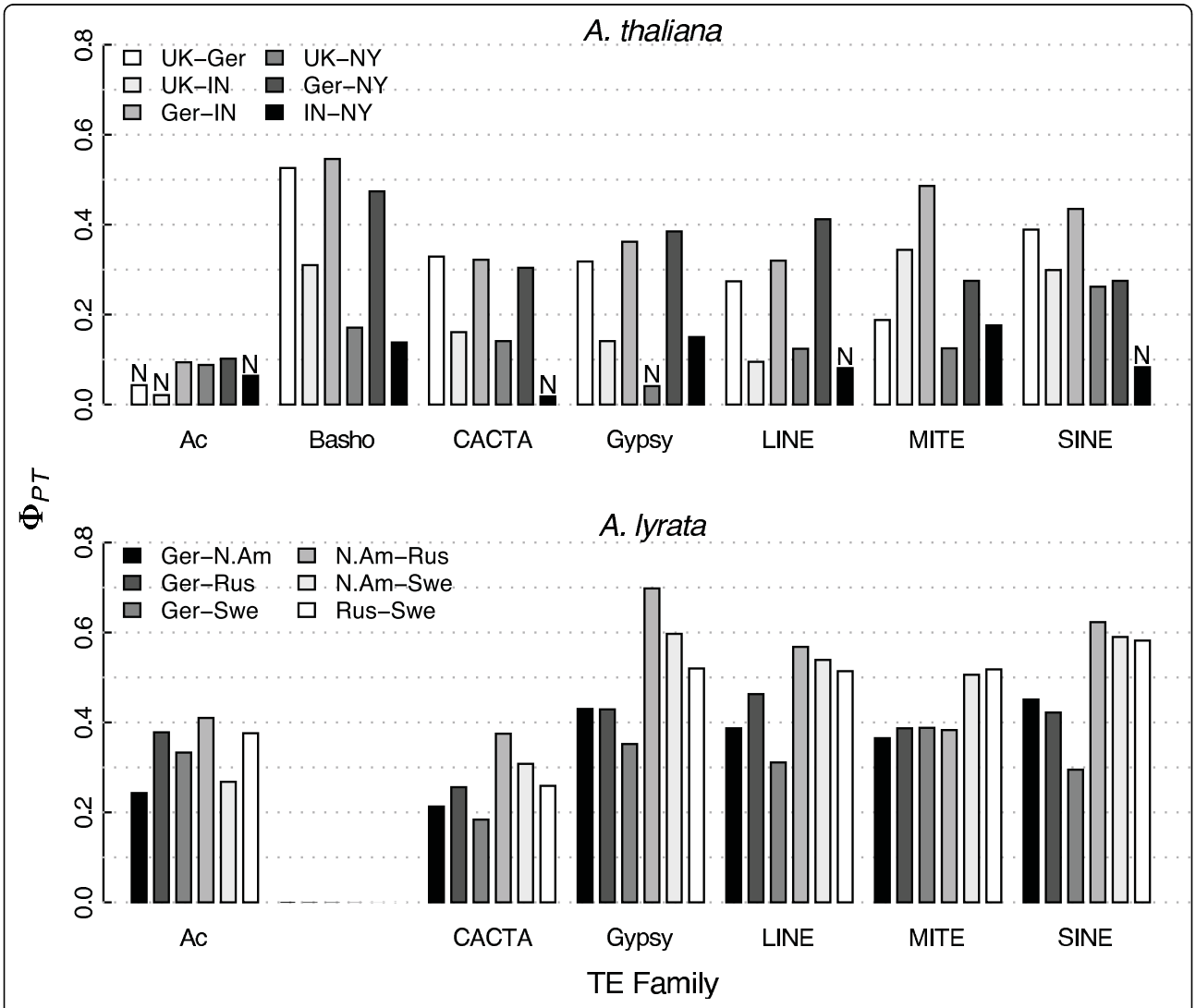

Figure 1 AMOVA Φ_{PT} per TE family in each population pairwise comparison in both species. *A. lyrata* data from [5]. *A. thaliana* populations: UK, Ascot, United Kingdom; Ger, Anholt, Germany; IN, Knox, Indiana, USA; NY, Long Island, New York, USA. *A. lyrata* populations: Ger, Plech, Germany; N.Am, North America; Rus, Russia; Swe, Sweden. "N"s indicate comparisons that show no significant population differentiation ($\Phi_{PT} = 0$, $p > 0.05$).

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 *A. thaliana* populations [53,61]. It is striking, however, that for most TE families we also observe low Φ_{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 Φ_{PT} values is relatively low values for *A. thaliana Ac* elements (Fig. 1), suggesting that *Ac* population dynamics differ from those of the other TEs surveyed. To test this idea more formally, we estimated total Φ_{PT} values among all populations for each TE family, and then compared the observed values to Φ_{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, Φ_{PT} was calculated for each replicate. From this exercise, it is clear that Φ_{PT} from *A. thaliana Ac* is much lower than expected under the null hypothesis (p = 0.003; Fig. 2). In contrast, data from *A. lyrata 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 *A. thaliana 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 *A. thaliana* compared to *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 *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 genome-wide effect of species with regard to the partitioning of TE variation. In addition, variance components graphically demonstrate that *A. thaliana 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

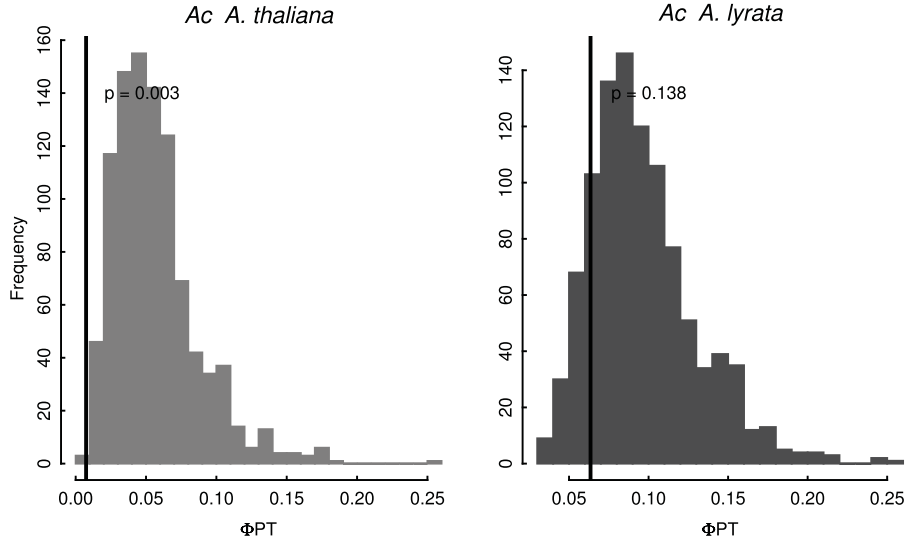

Figure 2 *Ac* AMOVA Φ_{PT} compared to bootstrap replicates subsampled from all TE family TE display bands.

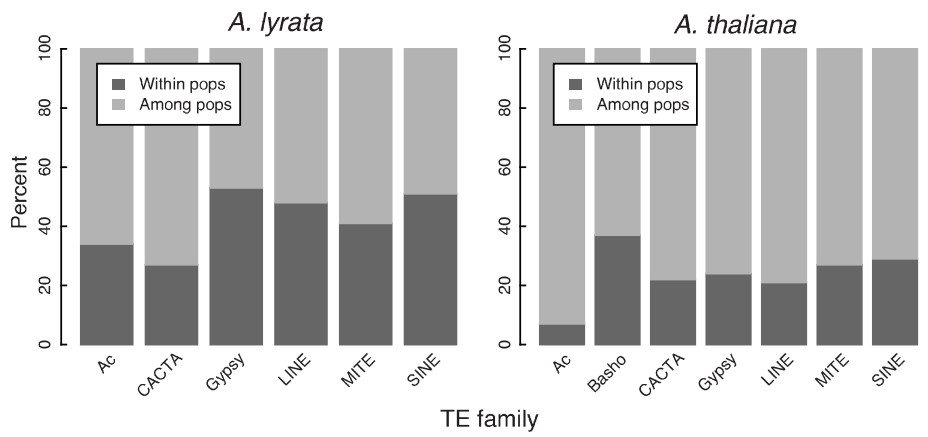

Figure 3 AMOVA percent molecular variation within and among populations ("pops").

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} was 0.061, but the *A. thaliana* median p_{TE} of 0.125 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 $N_e s$ estimates for pooled TEs were not significantly different from zero, and individual TEs pooled across populations also yielded $N_e s$ 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 $N_e s$ point estimate significantly less than zero, at -1.9 (Fig. 5; SI Table 1).

The SFS and $N_e s$ 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 Φ_{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 $N_e s$ 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 $N_e s$ 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],

Table 1 Median per individual TE allele frequency (p_{TE})

	A. lyrata						*A. thaliana*					
	Plech, Germany	N. America	Russia	Sweden	Mean	Standard Deviation	Anholt, Germany	IN, USA	NY, USA	Ascot, UK	Mean	Standard Deviation
Ac	0.221	0.208	0.291	0.31	0.258	0.051	0.192	0.231	0.324	0.221	0.242	0.068
Basho	-	-	-	-	-	-	0.166	0.33	0.351	0.305	0.288	0.102
CACTA	0.159	0.235	0.23	0.184	0.202	0.037	0.298	0.224	0.212	0.256	0.248	0.047
Gypsy	0.247	0.337	0.315	0.344	0.311	0.044	0.289	0.278	0.286	0.269	0.281	0.005
LINE	0.237	0.2	0.15	0.282	0.217	0.056	0.222	0.339	0.329	0.259	0.287	0.065
MITE	0.133	0.164	0.291	0.309	0.224	0.089	0.136	0.311	0.313	0.245	0.251	0.102
SINE	0.23	0.366	0.274	0.28	0.288	0.057	0.249	0.299	0.342	0.308	0.300	0.046

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

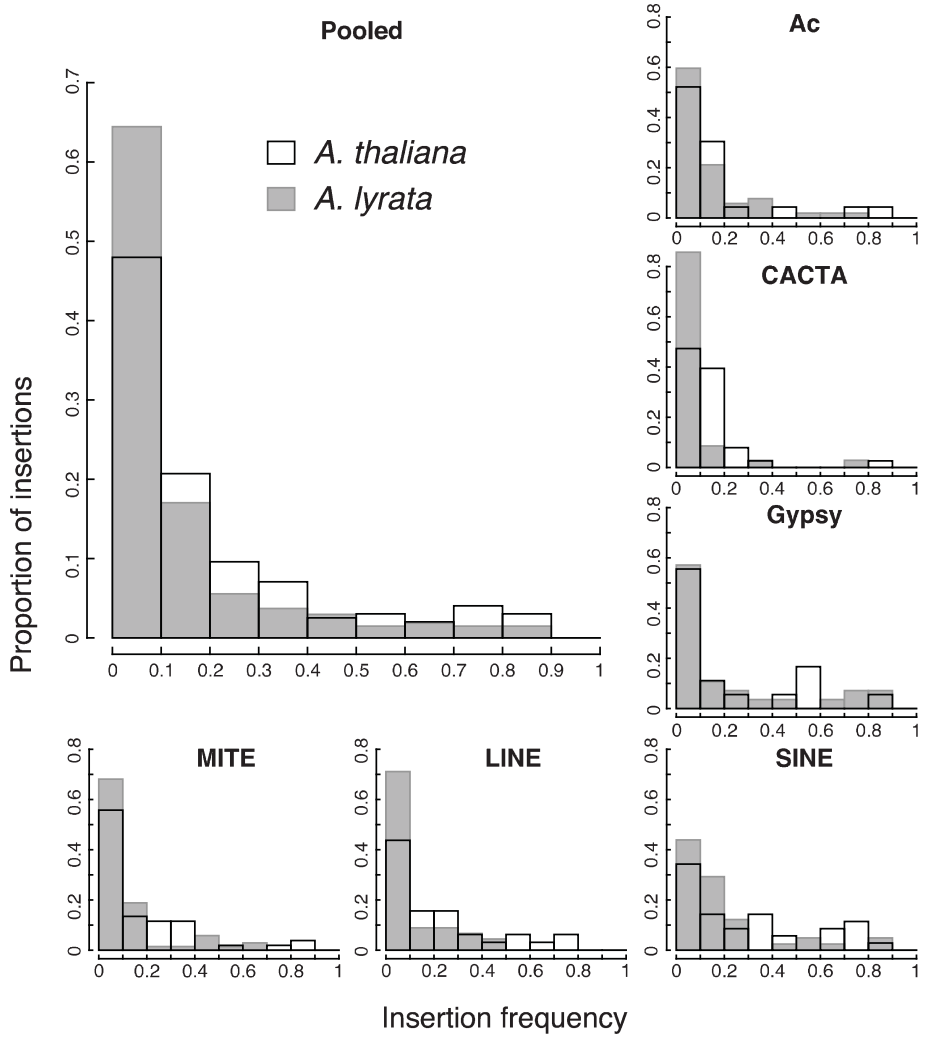

Figure 4 Site Frequency Spectra of all pooled TEs ("Pooled"), and each TE family in species-wide samples of *A. thaliana* (black bordered bars) and *A. lyrata* (grey bars with grey borders).

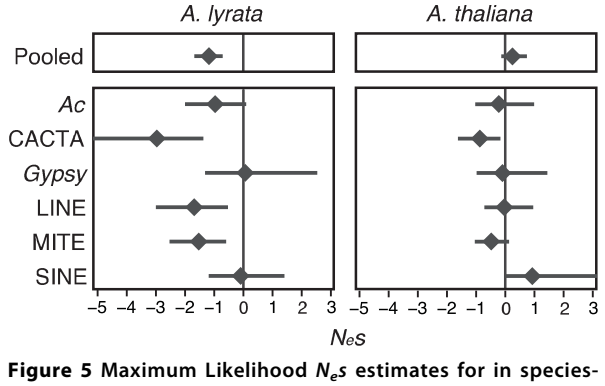

Figure 5 Maximum Likelihood N_eS estimates for in species-wide samples of *A. thaliana* and *A. lyrata* in pooled TEs ("Pooled") and each TE family. Horizontal bars are the 95% Confidence Interval and the diamond represents the ML N_eS point estimate. The vertical line is N_eS = 0.

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

Table 2 Median per individual TE copy number (n_{TE})

	A. lyrata						*A. thaliana*				
	Plech, Germany	North America	Russia	Sweden	Mean	Standard Deviation	Anholt, Germany	IN, USA	NY, USA	Ascot, UK	Mean	Standard Deviation
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	8.6	7.57	0.83	9.92	10.9	9.85	9.9	10.14	0.51
Gypsy	14.85	17.77	18.05	12.83	15.88	2.49	19.96	18.95	15.94	14.94	17.45	2.39
LINE	9.47	14.59	13.89	9.34	11.82	2.81	14.97	12.88	13.86	13.83	13.89	0.85
MITE	12.8	13.76	20.65	28.48	18.92	7.27	9.93	21.85	17.79	17.82	16.85	4.99
SINE	17.1	25.26	17.09	21.94	20.35	3.99	25.94	24.86	27.72	23.87	25.60	1.65
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* n_{TE} sums exclude *Basho* insertions

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 Φ_{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]. Such

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 Φ_{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: $N_e s$ 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]

Acknowledgements
We thank N. Komarova and S. I. Wright for assistance and discussion; R. Gaut for technical assistance; E. Thorhallsdottir, M. Clauss, 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

1. Bennetzen JL: **Transposable element contributions to plant gene and genome evolution.** *Pl Mol Biol* 2000, **42**:251-269.
2. Wessler SR: **Transposable elements and the evolution of eukaryotic genomes.** *Proc Natl Acad Sci USA* 2006, **103(47)**:17600-17601.
3. Zonneveld BJ, Leitch IJ, Bennett MD: **First nuclear DNA amounts in more than 300 angiosperms.** *Ann Bot (Lond)* 2005, **96(2)**:229-244.
4. Cornman RS, Arnold ML: **Phylogeography of *Iris missouriensis* (Iridaceae) based on nuclear and chloroplast markers.** *Mol Ecol* 2007, **16(21)**:4585-4598.
5. Lockton S, Ross-Ibarra J, Gaut BS: **Demography and weak selection drive patterns of transposable element diversity in natural populations of *Arabidopsis lyrata*.** *Proc Natl Acad Sci USA* 2008, **105(37)**:13965-13970.
6. Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF: **Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*.** *Genome Res* 2006, **16(10)**:1252-1261.
7. Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, *et al*: **Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice.** *Genome Res* 2006, **16(10)**:1262-1269.
8. SanMiguel P, Tickhonov A, Jin Y-K, Melake-Berhan A, Springer PS, Edwards KJ, Avramova Z, Bennetzen JL: **Nested retrotransposons in the intergenic regions of the maize genome.** *Science* 1996, **274**:765-768.
9. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796-815.
10. Matzke MA, Matzke AJM: **Polyploid and Transposons.** *TREE* 1998, **13**:241.
11. Ungerer MC, Strakosh SC, Zhen Y: **Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation.** *Curr Biol* 2006, **16(20)**:R872-873.
12. Langley CH, Brookfield JF, Kaplan N: **Transposable Elements in Mendelian Populations. I. a Theory.** *Genetics* 1983, **104(3)**:457-471.
13. Montgomery E, Charlesworth B, Langley CH: **A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*.** *Genet Res* 1987, **49(1)**:31-41.
14. Petrov DA, Hartl DL: **High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups.** *Mol Biol Evol* 1998, **15(3)**:293-302.
15. Charlesworth B, Charlesworth D: **The population dynamics of transposable elements.** *Genet Res* 1983, **42**:1-27.
16. Le Rouzic A, Boutin TS, Capy P: **Long-term evolution of transposable elements.** *Proc Natl Acad Sci USA* 2007, **104(49)**:19375-19380.
17. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite.** *Nature* 1980, **284(5757)**:604-607.
18. Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE: **Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila.** *Mol Biol Evol* 2003, **20(6)**:880-892.
19. Hollister JD, Gaut BS: **Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression.** *Genome Res* 2009, **19(8)**:1419-1428.
20. Biemont C, Tsitrone A, Vieira C, Hoogland C: **Transposable element distribution in Drosophila.** *Genetics* 1997, **147(4)**:1997-1999.
21. Hoogland C, Biemont C: **Chromosomal distribution of transposable elements in *Drosophila melanogaster*: test of the ectopic recombination model for maintenance of insertion site number.** *Genetics* 1996, **144(1)**:197-204.
22. Rizzon C, Marais G, Gouy M, Biemont C: **Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome.** *Genome Reserach* 2002, **12**:400-407.
23. Wright SI, Agrawal N, Bureau TE: **Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*.** *Genome Res* 2003, **13(8)**:1897-1903.
24. Montgomery EA, Huang S-M, Langley CH, Judd BH: **Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution.** *Genetics* 1991, **129**:1085-1098.
25. Brookfield JF, Badge RM: **Population genetics models of transposable elements.** *Genetica* 1997, **100(1-3)**:281-294.
26. Charlesworth B, Langley CH: **The population genetics of *Drosophila* transposable elements.** *Ann Rev Genet* 1989, **23**:251-287.
27. Charlesworth D, Charlesworth B: **Transposable elements in inbreeding and outbreeding populations.** *Genetics* 1995, **140**:415-417.
28. Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B: **On the role of unequal exchange in the containment of transposable element copy number.** *Genet Res* 1988, **52(3)**:223-235.
29. Morgan MT: **Transposable element number in mixed mating populations.** *Genet Res* 2001, **77(3)**:261-275.

30. Wright SI, Schoen DJ: **Transposon dynamics and the breeding system.** *Genetica* 1999, **107(1-3)**:139-148.
31. Charlesworth B, Morgan MT, Charlesworth D: **The effects of deleterious mutations on neutral molecular varaition.** *Genetics* 1993, **134**:1289-1303.
32. Nordborg M: **Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization.** *Genetics* 2000, **154(2)**:923-929.
33. Macpherson JM, Gonzalez J, Witten DM, Davis JC, Rosenberg NA, Hirsh AE, Petrov DA: **Nonadaptive explanations for signatures of partial selective sweeps in Drosophila.** *Mol Biol Evol* 2008, **25(6)**:1025-1042.
34. Dolgin ES, Charlesworth B: **The effects of recombination rate on the distribution and abundance of transposable elements.** *Genetics* 2008, **178(4)**:2169-2177.
35. Hill WG, Robertson A: **Linkage disequilibrium in finite populations.** *Theor Appl Genet* 1968, **38**:226-231.
36. Bachtrog D: **Accumulation of Spock and Worf, two novel non-LTR retrotransposons, on the neo-Y chromosome of *Drosophila miranda*.** *Mol Biol Evol* 2003, **20(2)**:173-181.
37. Bergero R, Forrest A, Charlesworth D: **Active miniature transposons from a plant genome and its nonrecombining Y chromosome.** *Genetics* 2008, **178(2)**:1085-1092.
38. Dolgin ES, Charlesworth B, Cutter AD: **Population frequencies of transposable elements in selfing and outcrossing *Caenorhabditis* nematodes.** *Genet Res* 2008, **90(4)**:317-329.
39. Young RJ, Francis DM, St Clair DA, Taylor BH: **A dispersed family of repetitive DNA sequences exhibits characteristics of a transposable element in the genus *Lycopersicon*.** *Genetics* 1994, **137(2)**:581-588.
40. Tam SM, Causse M, Garchery C, Burck H, Mhiri C, Grandbastien MA: **The distribution of copia-type retrotransposons and the evolutionary history of tomato and related wild species.** *J Evol Biol* 2007, **20(3)**:1056-1072.
41. Flowers JM, Purugganan MD: **The evolution of plant genomes: scaling up from a population perspective.** *Curr Opin Genet Dev* 2008, **18(6)**:565-570.
42. Wright SI, Quang HL, Schoen DJ, Bureau TE: **Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*.** *Genetics* 2001, **158**:1279-1288.
43. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, *et al*: **The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community.** *Nucleic Acids Res* 2003, **31(1)**:224-228.
44. Excoffier L, Smouse P, Quattro J: **Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data.** *Genetics* 1992, **131**:479-491.
45. Peakall R, Smouse PE: **GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research.** *Molecular Ecology Notes* 2006, **6(1)**:288-295.
46. Dray S, Dufour AB: **The ade4 package: implementing the duality diagram for ecologists.** *Journal of Statistical Software* 2007, **22(4)**:1-20.
47. Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, Gos G, Charlesworth D, Gaut BS: **Patterns of polymorphism and demographic history in natural populations of Arabidopsis lyrata.** *PLoS ONE* 2008, **3(6)**: e2411.
48. Weir BS: **Genetic Data Analysis II.** Sunderland, MA: Sinauer Assoc., Inc. 1996.
49. Abbot RJ, Gomez MF: **Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh.** *Heredity* 1989, **62**:411-418.
50. Brown AHD: **Enzyme polymorphism in plant populations.** *tpb* 1979, **15**:1-42.
51. Wright S: **Systems of mating.** *Genetics* 1921, **6**:111-178.
52. Koch MA, Haubold B, Mitchell-Olds T: **Comparative evolutionary analysis of the chalcone synthase and alcohol dehydrogenase loci among different lineages of *Arabidopsis*, *Arabis* and related genera (Brassicaceae).** *Mol Biol Evol* 2000, **17**:1483-1498.
53. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, *et al*: **The pattern of polymorphism in *Arabidopsis thaliana*.** *PLoS Biol* 2005, **3(7)**:e196.
54. DeRose-Wilson LJ, Gaut BS: **Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of *Arabidopsis thaliana* and *Arabidopsis lyrata*.** *BMC Evol Biol* 2007, **7**:66.
55. Leinonen PH, Sandring S, Quilot B, CM J, Mitchell-Olds T, Ågren J, Savolainen O: **Local adaptation in European populations of *Arabidopsis lyrata*.** *American Journal of Botany* 2009, **96**:1129-1137.
56. Muller MH, Leppala J, Savolainen O: **Genome-wide effects of postglacial colonization in *Arabidopsis lyrata*.** *Heredity* 2008, **100(1)**:47-58.
57. Clauss MJ, Mitchell-Olds T: **Population genetic structure of *Arabidopsis lyrata* in Europe.** *Mol Ecol* 2006, **15(10)**:2753-2766.
58. Schmid KJ, Torjek O, Meyer R, Schmuths H, Hoffmann MH, Altmann T: **Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers.** *Theor Appl Genet* 2006, **112(6)**:1104-1114.
59. Bakker EG, Toomajian C, Kreitman M, Bergelson J: **A genome-wide survey of R gene polymorphisms in Arabidopsis.** *Plant Cell* 2006, **18(8)**:1803-1818.
60. Francois O, Blum MG, Jakobsson M, Rosenberg NA: **Demographic history of european populations of *Arabidopsis thaliana*.** *PLoS Genet* 2008, **4(5)**: e1000075.
61. Jorgensen S, Mauricio R: **Neutral genetic variation among wild North American populations of the weedy plant *Arabidopsis thaliana* is not geographically structured.** *Mol Ecol* 2004, **13(11)**:3403-3413.
62. Sharbel TF, Haubold B, Mitchell-Olds T: **Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe.** *Mol Ecol* 2000, **9(12)**:2109-2118.
63. Charlesworth D: **Effects of inbreeding on the genetic diversity of populations.** *Philos Trans R Soc Lond B Biol Sci* 2003, **358(1434)**:1051-1070.
64. Mable BK, Adam A: **Patterns of genetic diversity in outcrossing and selfing populations of *Arabidopsis lyrata*.** *Mol Ecol* 2007, **16(17)**:3565-3580.
65. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123**:585-595.
66. Bustamante CD, Wakeley J, Sawyer S, Hartl DL: **Directional selection and the site-frequency spectrum.** *Genetics* 2001, **159(4)**:1779-1788.
67. Wakeley J, Aliacar N: **Gene genealogies in a metapopulation.** *Genetics* 2001, **159(2)**:893-905.
68. Wright SI, Lauga B, Charlesworth D: **Rates and patterns of molecular evolution in inbred and outbred Arabidopsis.** *Mol Biol Evol* 2002, **19(9)**:1407-1420.
69. Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, Lopez R, Price HJ: **Evolution of genome size in Brassicaceae.** *Annals of Botany* 2005, **95(1)**:229-235.
70. Hollister JD, Gaut BS: **Population and Evolutionary Dynamics of Helitron Transposable Elements in *Arabidopsis thaliana*.** *Mol Biol Evol* 2007, **24(11)**:2515-2524.

doi:10.1186/1471-2148-10-10
Cite this article as: Lockton and Gaut: The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. *BMC Evolutionary Biology* 2010 **10**:10.