Published in final edited form as:

Mol Ecol. 2010 August ; 19(16): 3364–3379. doi:10.1111/j.1365-294X.2010.04761.x.

Population history in *Arabidopsis halleri* using multilocus analysis

Andrew J. Heidel^{*1}, Sebastian E. Ramos-Onsins^{*2}, Wei-Kuang Wang³, Tzen-Yuh Chiang³, and Thomas Mitchell-Olds⁴

¹Department of Genetics and Evolution, Max Planck Institute for Chemical Ecology, 07745 Jena, Germany

²Department Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain

³Department of Life Sciences, Cheng-Kung University, Tainan 701, Taiwan

⁴Department of Biology, Duke University, Durham, NC 27708

Abstract

A. halleri is a psuedometallophyte with a patchy distribution in Europe and is often spread by human activity. To determine the population history and whether this history is consistent with potential human effects we surveyed nucleotide variation using 24 loci from 12 individuals in a large *A. halleri* population. The means of total and silent nucleotide variation (θ_W) are within the range expected for the species. The population genetic neutrality tests Tajima's D and Wall's B had significant composite results rejecting panmixia, and Approximate Bayesian Computation analysis revealed that a subdivision model better explained the variation than the standard neutral model, refugia (or admixture), bottleneck, or change of population size models. A categorical regression analysis further supports the subdivision model, and under the subdivision model the neutrality tests are no longer significant. The best support was for two source populations, a situation consistent with the mixing of two populations possibly mediated by human activity. This scenario might limit the genetic diversity and adaptive potential of the population. The nonneutral population variation described here should be considered in bioinformatic searches for adaptation.

Keywords

Arabidopsis halleri; population genetics; population history; multilocus analysis

Introduction

Anthropomorphic influences on ecosystems are of great current importance. The human influence on population genetic structure is clear for severe bottlenecks caused by loss of habitat, harvesting or hunting leading to reductions in population sizes, however the effects on population genetic structure may be less clear for other phenomena. Human-caused introductions are common, and while in general migration is a well-studied phenomenon, anthropomorphic introductions often create long distance dispersal on a level that would rarely happen otherwise. These introductions, from one or multiple source populations, may

Corresponding author: Andrew Heidel Genome Analysis Group, Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany
aheidel@fli-leibniz.de fax: +49 3641 656255.

present address: Genome Analysis Group, Fritz Lipmann Institute, 07745 Jena, Germany

*co-first authors

cause a less obvious and more complex population structure that may in turn influence the potential adaptation in the new location (Lavergne & Molofsky 2007). However, it can be difficult to untangle human effects from other non-human influences on population structure.

Species within the heavily studied genus *Arabidopsis* have been subject to human forces. The most studied species, *A. thaliana*, is a weed, and population genetic studies suggest that it has been spread recently throughout the world as a side effect of agriculture, especially in North America (Bergelson *et al.* 1998; Nordborg *et al.* 2005; Platt *et al.* 2010). The genetic and bioinformatic resources created for *A. thaliana* can also be used in the related *Arabidopsis* species *A. lyrata* and *A. halleri* that have habitats, a breeding system (resulting from insect pollination) and a perennial life history that differ from *A. thaliana*. Since the genetic and bioinformatic resources created for *A. thaliana* can be used in these two related species (Mitchell-Olds 2001), variation in more species can be determined without intensive *de novo* construction of new genetic resources. *A. halleri* diverged from *A. thaliana* ~5 MYA (Koch *et al.* 2000) and is native to central Europe, although a subspecies, *A. halleri* subsp. *gemmaefera*, is native to East Asia. *A. halleri* is a psuedometallophyte (Bert *et al.* 2000), growing in a number of locations polluted with heavy metals from human industry. This ability has influenced its recent population history (Pauwels *et al.* 2005). The genome size (C value) of *A. halleri* is 2.21-fold larger than *A. thaliana* (Oyama *et al.* 2008), and accordingly a study found that the region surrounding the *FLC* locus in *A. halleri* was 27% larger than the corresponding region in *A. thaliana* (Sanyal & Jackson 2006). Variation in *A. halleri* has been studied using microsatellites (Clauss *et al.* 2002; Van Rossum *et al.* 2004), RFLPs in cpDNA (Pauwels *et al.* 2005) and using sequence level variation (Kawabe & Miyashita 2003; Miyashita 2001; Ramos-Onsins *et al.* 2004).

Genetic variation and structure is influenced by a number of natural factors including selection, mutation rate, migration and habitat size. Additionally, variation is influenced by human activity that causes introductions and creates or destroys habitat. Since variation from one factor can complicate evaluation of another factor, methods have developed to disentangle such multiple factors. The use of multiple loci to determine variation reduces the effect of selection and drift at any one individual locus; analysis of multiple loci has been used with allozymes (Shaw 1995) and is now common for sequence level variation (Ramos-Onsins *et al.* 2004; Schmid *et al.* 2005; Wilder & Hollocher 2003; Wright *et al.* 2003). In addition to using multiple loci, it has been shown that it is important to know the background level of variation because it can affect tests of selection (Schmid *et al.* 2005).

Ramos-Onsins *et al.* (2004) performed the largest sequence-level variation study to date in *A. halleri*, studying six genes in the phenylpropanoid pathway and two genes in the glucosinolate pathway from multiple populations. While sampling from many populations can allow a biogeographical history to be elucidated, (Sharbel *et al.* 2000) heavier sampling from within one population and determining the variation within an individual population is critical to investigate local adaptation. Using a larger sample of 24 random loci, 19 of which contain coding regions, we performed a sequence-level analysis of genetic variation in a population of *A. halleri* that possibly was recently established by human activity (Clauss *et al.* 2002) in central Germany. We estimate the genome-wide variation and determine if the variation fits a panmictic model. Next, we combine standard population genetic tests based on SNP frequencies (Tajima's D, Fay and Wu, etc.) with tests for population events (admixture, subdivision, migration) and determine if the results are consistent with human-assisted migration and the potential for tests of selection.

Methods

Species sampling

Plants were collected from a large population of *A. halleri* ($N > 1000$) previously described by Clauss *et al.* (2002) in southern Thuringia (Stutenkamm), Germany. Samples were collected haphazardly with a minimum of 2m between samples to avoid sampling clones since *A. halleri* can propagate clonally. Additionally, samples were collected throughout the entire area of the population. DNA was extracted from 12 plants as described by Clauss *et al.* (2002).

PCR amplification and sequencing

The loci (Table 1) used are a subset of the group of randomly selected loci sequenced previously in *A. thaliana* (Schmid *et al.* 2005). The criteria for using this particular subset was that the primers functioned reliably in *A. halleri*. Therefore, the loci are biased towards regions that are conserved between *A. thaliana* and *A. halleri*, but we believe this fact should not affect our analyses since the loci should still be unlinked. Additional support that the loci are unlinked is that the loci in the subset are distributed across all five *A. thaliana* chromosomes. The loci were amplified by PCR using Thermoprime Taq (Abgene) with 10–20 ng template DNA per 50 μ L reaction and primers given in Supplemental Table 1. All PCR products were purified from an agarose gel using a gel purification kit (Viogene) and cloned into a pGEM-T easy cloning vector (Promega). For each PCR reaction, five independent clones were randomly selected and sequenced. Although sequence errors are likely rare and no recombinant haplotypes were observed in this study, artifact and undetectable singleton errors might compromise the results of the project. Therefore, we have performed the analysis using only the most frequent haplotype per individual to ensure at least 3 clones per haplotype. Thus, we ensure high quality sequence data despite discarding half of the data.

Cloned sequences were sequenced on an ABI 377XL automated sequencer using universal T7 forward and SP6 reverse primers located in the pGEM-T vector termination sites. Sequences could not be obtained for one individual at locus GOLM19, so there were only sequences from 11 individuals at that locus. All sequences from this article have been deposited in GenBank with accession numbers: GU124840-GU125389 and HM125990-HM126013. In the GenBank files, the last number in the name of each sequence indicates the haplotype, and the haplotype labeled “1” is the most frequent haplotype used in all analyses.

Analyses

For each locus in *A. halleri*, the likely orthologous *A. thaliana* locus was identified for use in analyses (Table 1). For the locus *AtIII6* $\times$ 5, there are two potential *A. thaliana* loci, so both were used. In this case, almost no differences in the divergence level were observed using either of the two copies. *At3est41* contains a polymorphic frameshift relative to the *A. thaliana* sequence and may not be a functional gene, but we have treated it as a normal gene since it has not accumulated a large number of nonsynonymous mutations. The codon with the frameshift was eliminated from the analysis on coding regions.

The 24 loci are only loosely linked to each other in *A. thaliana*, so they are treated as independent loci. In absence of physical proximity between loci, the detection of linkage disequilibrium between them may indicate population structure in the studied sample, especially if the disequilibrium is related to the geographical location or breeding patterns. Therefore we calculated linkage disequilibrium among loci using LIAN software (<http://adenine.biz.fh-weiheinstephan.de/cgibin/lian/lian.cgi.pl>). We used the software

STRUCTURE 2.2 (Pritchard *et al.* 2000) to detect population subdivision in the sample using the admixture model and allowing mixed ancestries of the individuals, as recommended in the manual of the program. We also assumed that the allele frequencies are correlated among populations. We performed up to 20 independent runs using a burnin of 10000 plus 10000 values for MCMC after burnin, but also additional runs were performed using 50000 burnin plus 50000 and 90000 burnin plus 10000 to confirm results. We discarded those runs than had very divergent averages or variances in relation to the rest for the same value of K. Evanno's method (Evanno *et al.* 2005) was used to establish the most plausible value of K given the averages and the variances of the different chains. The Bayesian rule was also used to detect the best model following the software manual.

The level of nucleotide variation for total, silent, synonymous and nonsynonymous positions was calculated with π (Tajima 1983) and θ (Watterson 1975). Haplotype diversity (Depaulis & Veuille 1998), the neutrality tests Tajima's *D* (Tajima 1989), Fu's *F_s* (Fu 1997), Fay and Wu's *H* normalized (Fay & Wu 2000; Zeng *et al.* 2006), *R₂* (Ramos-Onsins & Rozas 2002), *ZnA* (Rozas *et al.* 2001), Wall's *B* (Wall 1999) and HKA (Hudson *et al.* 1987) were calculated for silent positions. The HKA test was corrected for multiple hits using the Jukes and Cantor method (Jukes & Cantor 1969) and using *A. thaliana* as the outgroup species. Synonymous and nonsynonymous positions were estimated with the Nei and Gojobori method (Nei & Gojobori 1986). These statistics give information from the frequency spectrum of the variants as well as the linkage disequilibrium among polymorphisms. Nucleotide divergence was calculated and corrected for multiple hits with the Jukes and Cantor method (Nei & Gojobori 1986). All these analyses were calculated with MANVa (Ramos-Onsins, Windsor and Mitchell-Olds, available from authors). The minimum number of recombination events was also calculated (Hudson & Kaplan 1985).

Codon bias was estimated with ENC (Wright, 1990) using DnaSP (Rozas *et al.* 2003). Differences in the level of variability in the synonymous versus non-coding regions were contrasted using two methods: 1) calculating a likelihood ratio test (which assumes the standard neutral model (SNM)) and calculating the levels of variation assuming both have the same level of variation (null hypothesis) or assuming different level of variation for synonymous and non-coding positions (alternative model, see (Schmid *et al.* 2005)) and 2) a non-parametric bootstrap analysis, resampling the positions with replacement and calculating for each replicate the levels of variation and the differences in the levels of variation between synonymous and non-coding regions. McDonald and Kreitman tests (McDonald & Kreitman 1991) were used to detect differences between polymorphism and divergence at synonymous versus nonsynonymous positions. Only codons with a maximum of two variants were considered in the analysis. Also, we calculated for each locus the proportion of adaptive substitutions $\alpha = (DsPn)/(DnPs)$, where *Ds* and *Ps* are the synonymous (assuming synonymous as neutral) divergence and polymorphism, respectively, and *Dn* and *Pn* are the divergence and polymorphism for the nonsynonymous (or non-coding) region (Andolfatto 2005; Smith & Eyre-Walker 2002). We also calculated α from the total number of polymorphisms and divergence obtained from the sum of all loci (Haddrill *et al.* 2008). We used a non-parametric bootstrap (resampling the nucleotide positions with replacement) to estimate the confidence intervals using 1000 replicates. Results should be taken with caution since significant values might be produced by changes in population size; changes in population size might fix slightly deleterious mutations in a higher proportion and consequently increase the value of α (Eyre-Walker 2002). Furthermore, α and their confidence intervals were re-calculated excluding singletons from nonsynonymous (non-coding) regions to exclude the putative slightly deleterious mutations from the analysis (Andolfatto 2005), or including only those loci with more than five synonymous polymorphic variants (Smith & Eyre-Walker 2002). No significant correlation between the ratio *Pn/Ps* and the level of variation for synonymous and for nonsynonymous

(non-coding) positions was observed (using non-parametric bootstrap for estimating the confidence intervals). All these analyses have been performed using the R environment (<http://www.r-project.org/>).

Coalescent simulations were performed with *mlcoalsim* v. 1.91b (Ramos-Onsins & Mitchell-Olds 2007) (available from authors); *mlcoalsim* calculates neutrality tests and other statistics of interest for a range of demographic models and allows prior distributions for most parameters. Multiple hits were allowed in the coalescent simulations (considering the transition/transversion ratio), but these positions were not included in the calculation of the statistics.

We used the Approximate Bayesian Computation (ABC) analysis (Beaumont *et al.* 2002) to infer the best evolutionary parameters for the studied data in five different models: Standard Neutral Model (SNM, or Wright-Fisher model, which assumes panmixia, discrete generations, stationary population size with no selection neither migration), change of population size, bottleneck, admixture of populations and subdivision. We assumed the same value of θ and R for the entire genome, and the remaining parameters (Figure 1 shows the parameters included in each model) were also the same for all loci. The prior distributions used for these models are natural log-uniform distributions with the following limits: θ , from 0.001 to 0.06 per nucleotide; R , from 0.0001 to 0.1 per nucleotide. In the Change of population size model, the ancestral population size N_A ranged from 0.1 to 10 (all N are relative to current $N=1$) and the time to the sudden change of population size, T_C , from 0.001 to 0.4 N generations. In the Bottleneck model, the population size during the bottleneck, N_A , ranged from 0.001 to 1, the ancestral population size, N_A , from 0.01 to 100 and the time to the sudden change of population size, T_C , from 0.001 to 1 N generations. In the Admixture model the ancestral population size, N_A , ranged from 0.1 to 10, the time to the split in two refugia, T_C , from 0.001 to 1 N generations, and the population size for refugia populations, N_1 and N_2 , from 0.001 to 1 (the frequency of the contribution of each refugia to the current population is assumed the same). In the Subdivision model, the population size for the second population, N_2 , ranged from 0.01 to 100, and migration population parameter, $M=4Nm$, from 0.01 to 10. For each model 500,000 coalescent simulations were run to infer the posterior distribution of the parameters. The best 1000 iterations (the tolerance was 1/500) were chosen to calculate a linear regression with the observed summary statistics. We used scripts in R, downloaded from the M. A. Beaumont web page (<http://www.rubic.rdg.ac.uk/~mab/>) to calculate the posterior distributions and the highest probability density values using the ABC analysis. One of the main pitfalls of the ABC approach is that this method is sensitive to the statistics chosen (Beaumont 2008). In our approach, we calculated for each model the correlation matrix for a wide number of statistics; then we selected those statistics that were less correlated (we chose an arbitrary value of the square of the correlation less than 0.5625). The selected statistics are the following: the average and variances (for all loci) of Tajima's D , Fu's F_s , Ramos-Onsins and Rozas' R_2 , Wall's B , Fay and Wu's H normalized and the average of Watterson's θ . Other combinations used the same statistics but replaced Rozas's ZnA by Wall's B and/or replaced R_2 by the haplotype diversity. The use of different combination of statistics gave varying results values, although mainly in the same direction. Posterior probabilities for each studied scenario were calculated by applying a categorical regression (Beaumont 2006) using the simulations and the tolerance value indicated above, by means of R scripts provided by Beaumont in his web page. The algorithm uses the best iterations (the closest to the observed data) considering all models together. The proportion of iterations coming from each model was counted to calculate a raw posterior probability. Next, we used the linear regression option to obtain the final posterior probabilities.

By using the Rejection Algorithm provided in the Beaumont code, we obtained the list of the best 1000 iterations per model. *P*-values for the statistics of interest were obtained individually. We also calculated the *P*-value for the composite probability of each statistic for all loci, or for the composite probability of a single locus for a number of summary statistics. The composite probability was calculated as the sum of the natural logarithm of the probabilities (for both tails together, see details of the calculation in Ramos-Onsins *et al.* 2008). Probabilities were corrected for multiple testing using the false discovery rate (FDR) method implemented in the *p.adjust* function in R (Benjamini & Hochberg 1995).

We also analyzed a model that assumes a different θ and *R* value for each locus. We used a rejection algorithm for each locus, where only those iterations with the number of segregating sites (*S*) and the number of minimum recombination events (*R_m*) equal to the observed values were chosen. For each of the 10⁵ effective iterations calculated here, and for a given set of common parameters (i.e., in case of the Subdivision model, *M* and *N₂*), up to 1000 runs were performed with different values of θ and *R* (following each a prior distribution) until finding coincidence of the *R_m* and *S* with the observed values. Thus, to achieve 10⁵ effective iterations, the program might calculate up to 10⁸ iterations. Next, we used the ABC analysis to estimate the posterior distribution of the parameters using a tolerance of 1/100. In the ABC analysis, the summary statistic Watterson's θ mean (here uninformative because all values are equal) was replaced by the nucleotide diversity (π) mean.

Results

In this work, given the difficulties in accurately obtaining both alleles from each individual, we have only used one to ensure high quality data (see Methods). Since we are interested in the levels and patterns for the entire population, it is valid to take only one allele per individual. Using one allele, we determined the estimates for total, silent, synonymous, nonsynonymous and non-coding polymorphisms, as well as haplotype diversity and the nucleotide divergence (Tables 2 and 3). Polymorphism data for each locus as well as the classification of the variants as synonymous or non-synonymous are shown in Supplemental Tables 2–25.

Multilocus Variation

Nucleotide variation (θ_W) in the 24 loci has a mean per locus value of 0.0078. There is a high variance in the levels of nucleotide diversity between loci; three loci have higher than 2% nucleotide diversity at silent positions while others have low diversity. The differences in the estimates of silent diversity between loci could be explained by differential mutational rates per locus, but also could be explained by demographic processes that increase the variance in the lengths of the sample genealogies (see below). Note that the number of minimum recombinant events is very high in two loci (*At3est41* and *AtV27*), an occurrence that can not be easily explained by an overall high recombination rate. Gene conversion or other processes that increase the recombination rate in these regions might have occurred. At the multilocus level, the differences in the ratio of silent polymorphism versus divergence among loci cannot be simply explained by the SNM, as indicated by the HKA test ($HKA = 65.248$, $P\text{-value} = 6.44 \times 10^{-6}$) suggesting population structure or heterogeneous forces on the loci. The level of nonsynonymous polymorphisms is approximately five times lower than for synonymous, suggesting constraint in nonsynonymous positions as expected. Variability at non-coding regions is similar to those in synonymous regions, although slightly lower. A Likelihood Ratio Test assuming the Standard Neutral Model and also a non-parametric bootstrap analysis showed that the differences between nonsynonymous and non-coding variation (compared each locus pair-pair) were nonsignificant ($P\text{-value} \gg 0.05$, see Supplemental Figure 1). We did not observe any evidence of bias in the codon usage, as

indicated by the elevated values of effective number of codons (ENC, three loci have $50 > \text{ENC} > 42$, the rest more than 50) (Wright 1990).

We performed a McDonald-Kreitman test (McDonald & Kreitman 1991) for each locus separately and for all loci together (Table 4) for both synonymous versus nonsynonymous and synonymous versus noncoding. The MK test for synonymous versus nonsynonymous, revealed that only the locus *AtIV20* was statistically significant (using a correction for multiple testing). The high number of nonsynonymous polymorphisms in relation to divergence indicates a strong fixation pressure, suggesting that these polymorphisms are under purifying selection. The proportion of adaptive substitutions α at the total positions suggests that many polymorphisms at nonsynonymous positions are slightly deleterious, which are usually at low frequencies in the population. Removing singletons from the analysis increases the value of α , which is expected when slightly deleterious mutations are removed. The value of α was not significantly different from zero (Supplemental Figure 2), suggesting that positive selection is not (or poorly) acting on these set of genes (or we do not have enough power to detect it). The MK test for the comparison of synonymous versus non-coding did not show any significant values. The α values were slightly negative but non-significant.

Tests for Population Histories

We used STRUCTURE v 2.2 (Pritchard *et al.* 2000) to detect structural patterns in the sampled population using a single allele per individual. STRUCTURE assigns probabilistically the samples to one or more virtual populations (or to several if the genotypes are mixed), assuming that there is no linkage disequilibrium within groups and considering all loci (Figure 2). We performed seven independent runs for each of the models with K (number of populations) =1 to K=12. Evanno's method (Evanno *et al.* 2005) found that K=2 was the best value (Figure 2A). Also, using a Bayesian rule proposed in the STRUCTURE manual, we concluded that the most supported value was K=2 (P-value > 0.999 for K=2 in relation to the rest), suggesting that there were two recent source populations.

We used the ABC analysis (Beaumont *et al.* 2002) to infer the parameters for a set of five possible evolutionary models that may explain the patterns of diversity (Figure 1). The SNM is analyzed by default as the null model. The Admixture and Subdivision models could explain the relative high level of linkage disequilibrium. The ABC analysis calculates the posterior distribution for each of the parameters in a model, and employing the same tools, it is possible to calculate the posterior distribution values among models using a categorical regression analysis. Table 5 shows the posterior probabilities calculated among models using this approach. The model with the highest probability was the model of subdivision with two populations (average $P \sim 0.75$). Using this approach, the second best is the Refugia model with two refugia (average $P \sim 0.14$) followed by the SNM and bottleneck. Although the posterior values do not definitely accept one of the models over the other models, the subdivision model has a much higher probability than the rest. We have no proof that a more complex model (e.g., more than two populations in subdivision or in the refugia model) would change the obtained results. Nevertheless, analyses using STRUCTURE seems to support a model where two effective populations may affect the sample.

The models used above assume a single θ and R for all studied loci, which is not necessarily true. On the other hand, assuming a different θ and R for each locus may excessively overestimate the variance over these parameters, and the resulting inference analyses may be unrealistic. We have recalculated the distribution of the posterior parameters for the model of subdivision assuming a different θ and R for each locus to check differences between assumptions. We effectively observe differences in the estimation of parameters although

we find overlapping between the distribution parameters. A more complete inference analysis under this assumption requires very large computation effort and has not been performed.

The posterior distributions for the subdivision model ($K=2$) are shown in Figures 3 and 4. The posterior distribution for the effective population size of the second theoretical population (N_2) was poorly estimated, as we do not have appropriate statistics that can obtain this information (we have no data from the other theoretical population, so the estimation of this parameter is difficult). The migration parameter ($M=4Nm$) was better estimated. This model does not include the value zero in the migration parameter, but the posterior distribution drops (at a natural logarithmic scale) for smaller values of migration and for higher. The highest posterior distribution values (hpd) for the first combination of statistics shown in Table 5 are 0.0113 for θ and 0.0060 for R per nucleotide under the SNM (assuming a single θ and R); for the Subdivision model, the hpd value for θ per nucleotide was 0.0072, 0.0373 for R per nucleotide, 0.0281 for M and 0.356 for N_2 . For the subdivision model assuming different θ and R for each locus, the hpd was 2.121 for M and 8.685 for N_2 . The posterior distributions for M and N_2 are not equal under different assumptions, although some overlapped (see Figures 3 and 4). In both Subdivision models (using different assumptions) we see that intermediate values of M are more represented than low or high (i.e., the values of M expected if the SNM was true), suggesting that the SNM can not explain the observed data. This result is also in the same direction as the results shown using STRUCTURE (Figure 2).

Supporting the above results, we calculated the probability of the summary statistics and composite probabilities in relation to the observed data (Table 6). There are significant values for the composite probabilities of Tajima's D , Wall's B and Rozas' ZA , suggesting that the SNM might not explain properly the patterns of variability, with the Wall's B and Rozas' ZA results especially suggesting patterns referring to linkage disequilibrium are not consistent with the SNM. We have also contrasted the observed data against the subdivision model and none of the summary statistics were significant (results not shown).

We attempted to find loci that significantly depart from the chosen Subdivision model. We calculated the composite probability for a set of statistics in each locus independently (Table 6). We did not find any significant value using the composite probability or looking at each statistic individually, suggesting that these loci are not affected by other possibly selective forces in a significant way, or we do not have enough power to detect them with the present data.

Discussion

Level of variation

The total and silent pairwise nucleotide variation of 0.0082 and 0.0130 respectively is similar to a previous analysis of eight loci in *A. halleri* that found total and silent pairwise nucleotide variation of 0.0081 and 0.0150 respectively (Ramos-Onsins *et al.* 2004). This similarity is not necessarily expected because the previous study used multiple individuals from 3–5 populations in the Czech Republic and France while we used only a single German population. It might be expected that between population variation is greater between two distant groups of populations. However, the French populations are rather recently established by human activity and share chloroplast haplotypes with distant populations (Pauwels *et al.* 2005). Therefore, a potential explanation for the lack of differences for within and between population variation might be recent establishment by human activity from a few sources. As a result, there has not been enough time for substantial divergence to occur between populations.

The only other studies of sequence variation in *A. halleri* are in the subspecies *A. halleri* ssp. *gemmifera*. One group of studies using up to 15 Japanese populations, found the pairwise variation in four genes ranged from 0.0032 to 0.0053 (Kawabe & Miyashita 2002; Kawabe & Miyashita 2003; Miyashita 2001). The other study used 9 individuals from a wider range in China, Taiwan and Japan and found pairwise variation in 3 genes ranged from 0.0016 to 0.0144 (Wang *et al.* 2007). This level of variation is within the range found in our 24 loci although with a lower mean. The set of studies using the 15 Japanese populations found that the variation between populations is the same or lower than our analyses within a single population, again raising the question why multiple populations would be less diverse than our single-population sample. The population history of *A. halleri* ssp. *gemmifera* is not well known so it cannot be ascertained whether any of the Japanese populations are anthropomorphic in origin. If the populations are recently established by human activity, the same explanation may also apply to the Japanese populations. Alternatively if the populations are nearby to each other they could have all have been recently established by nonhuman events.

Population history

A negative Tajima's *D* and a significant Wall's *B* reject the SNM, suggesting either different types of selection or demographic events. Demographic events are much more likely given that silent sites are analyzed. The significant HKA test between silent polymorphism and divergence is further evidence against the SNM. The STRUCTURE results suggest that the population originated from two older populations and the ABC analysis is consistent with this result, supporting a subdivision model with two populations. When the neutrality tests are recalculated under a subdivision model, no significant deviation from neutrality is found, again supporting this model. A subdivision scenario is consistent with what is known about the population history. The population is thought to have been established approximately 1,000 years ago from mining in the area (Clauss *et al.* 2002). An interesting alternative hypothesis to a simple demographic scenario suggests that the population could have formed from the introduction of two populations of human-carried founders. The combination of two populations could have also occurred at another mining location previously and then been transported to the current location. A lack of gene flow allowing subdivision is also likely in the species due to the species growing in isolated patches. The Bottleneck and Expansion/Contraction models previously have been used to explain similar patterns including a negative Tajima's *D* and Fay and Wu's *H* in another European plant, the tree *Populus tremula* (Ingvarsson 2008) although the samples were from multiple populations. The higher dispersal ability of *P. tremula* and not being restricted to isolated populations may prevent subdivision from occurring as easily as in *A. halleri*.

Studies of selection in human influenced populations

A study of *A. halleri* from 5 populations in the Czech Republic and France showed an excess of high frequency mutations that were likely caused by introgression from other *Arabidopsis* species (Ramos-Onsins *et al.* 2004). Introgression also could have contributed to the large number of nonsynonymous mutations in our study. The McDonald-Kreitman results could suggest that one locus, *AtIV20*, might be under positive selection, but the high number of nonsynonymous polymorphisms relative to divergence suggests that the locus merely contains many low frequency deleterious polymorphisms under purifying selection. Introgression may have introduced some of these polymorphisms, and they may have persisted to this point because not enough time has elapsed to eliminate weakly deleterious polymorphisms. This locus is predicted to be involved in RNA processing, and there is no *a priori* reason to expect positive selection.

Long-distance dispersal from human activities followed by restriction to small, isolated populations appears to be common in *A. halleri*, and has been detected from chloroplast haplotypes (Pauwels *et al.* 2005). Within a single population our sequence data was able to find a finer level of variation. Many plant species are transported by human activities, but the restriction of *A. halleri* to small patches is less common and unlike many weedy species. This type of event is likely to limit the genetic diversity of the population and thereby the potential for adaptation and positive selection.

There is an interest in determining the genes required for metal-tolerance in plants such as *A. halleri* (Pauwels *et al.* 2008) and some have been determined (Hanikenne *et al.* 2008; Talke *et al.* 2006). Although the genes identified so far may exert their effects by regulatory changes, additional genes could be found by searching for positively selected genes. However, a departure from a neutral model can bias or at least complicate tests of selection since patterns of variation that appear to be caused by selection may be caused instead by demographic processes or introgression. Given that panmixia was not found here or in any other study of *A. halleri* to date, future findings of an excess of rare alleles and a lack of intermediate frequency alleles at a locus (which normally might be considered evidence of non-neutral evolution), should be interpreted with caution. Furthermore, the variable divergence values may make it difficult to assess MK and similar tests because a large number of fixed variants may overwhelm any signal of selection in many loci. The rare polymorphisms can be excluded to increase the statistical power, however this may not always be helpful (Charlesworth & Eyre-Walker 2008). Finally, introgression may further hide adaptation, so identification of polymorphisms in closely-related species may be necessary to distinguish between adaptive mutations and introgressed polymorphisms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Maria Clauss for assistance with the source population and three anonymous reviewers for their helpful comments on the manuscript. We acknowledge the Max Planck Society for financial support. S.E.R.O. is a recipient of a Ramón y Cajal position (MICINN), and this work was funded by Grants CGL2009-0934 (MICINN, Spain) to S.E.R.O. T.M.O. was supported by grants from NIH (R01 GM086496) and NSF (EF-0723447).

References

- Akaike H. New Look at Statistical-Model Identification. *IEEE Transactions on Automatic Control*. 1974; AC19:716–723.
- Andolfatto P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*. 2005; 437:1149–1152. [PubMed: 16237443]
- Beaumont, MA. Joint determination of topology, divergence time and immigration in population trees, in simulations, genetics and human prehistory. McDonald Institute for Archaeological Research; Cambridge, UK: 2006.
- Beaumont MA, Zhang WY, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics*. 2002; 162:2025–2035. [PubMed: 12524368]
- Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological*. 1995; 57:289–300.
- Bergelson J, Stahl E, Dudek S, Kreitman M. Genetic variation within and among populations of *Arabidopsis thaliana*. *Genetics*. 1998; 148:1311–1323. [PubMed: 9539444]

- Bert V, Macnair MR, de Laguerie P, Saumitou-Laprade P, Petit D. Zinc tolerance and accumulation in metalicolous and nonmetallicolous populations of *Arabidopsis halleri* (Brassicaceae). *New Phytologist*. 2000; 146:225–233.
- Charlesworth J, Eyre-Walker A. The McDonald-Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution*. 2008; 25:1007–1015. [PubMed: 18195052]
- Clauss MJ, Cobban H, Mitchell-Olds T. Cross-species microsatellite markers for elucidating differences in population genetic structure in *Arabidopsis* and *Arabis* (Brassicaceae). *Mol. Ecol*. 2002; 11:591–601. [PubMed: 11918792]
- Depaulis F, Veuille M. Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution*. 1998; 15:1788–1790. [PubMed: 9917213]
- Eyre-Walker A. Changing effective population size and the McDonald-Kreitman test. *Genetics*. 2002; 162:2017–2024. [PubMed: 12524367]
- Evanno GS, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol*. 2005; 14:2611–2620. [PubMed: 15969739]
- Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. *Genetics*. 2000; 155:1405–1413. [PubMed: 10880498]
- Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*. 1997; 147:915–925. [PubMed: 9335623]
- Hadrill PR, Bachtrog D, Andolfatto P. Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Molecular Biology and Evolution*. 2008; 25:1825–1834. [PubMed: 18515263]
- Hamilton G, Currat M, Ray N, et al. Bayesian estimation of recent migration rates after a spatial expansion. *Genetics*. 2005; 170:409–417. [PubMed: 15716508]
- Hanikenne M, Talke IN, Haydon MJ, et al. Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. *Nature*. 2008; 453:391–U344. [PubMed: 18425111]
- Hudson RR, Boos DD, Kaplan NL. A Statistical Test for Detecting Geographic Subdivision. *Molecular Biology and Evolution*. 1992a; 9:138–151. [PubMed: 1552836]
- Hudson RR, Kaplan NL. Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA-Sequences. *Genetics*. 1985; 111:147–164. [PubMed: 4029609]
- Hudson RR, Kreitman M, Aguade M. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics*. 1987; 116:153–159. [PubMed: 3110004]
- Hudson RR, Slatkin M, Maddison WP. Estimation of Levels of Gene Flow from DNA-Sequence Data. *Genetics*. 1992b; 132:583–589. [PubMed: 1427045]
- Ingvarsson PK. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics*. 2008; 180:329–340. [PubMed: 18716330]
- Jukes, TH.; Cantor, CR. Evolution of protein molecules. In: MH, N., editor. *Mammalian protein metabolism*. Vol. III. Academic Press; New York: 1969. p. 21-132.
- Kawabe A, Miyashita NT. DNA variation in the acidic chitinase locus (ChiA) region in *Arabis gemmifera* and its related species. *Genes & Genetic Systems*. 2002; 77:167–175. [PubMed: 12207038]
- Kawabe A, Miyashita NT. DNA polymorphism in active gene and psuedogene of the cytosolic phosphoglucose isomerase (*PgiC*) loci in *Arabidopsis halleri* ssp. *gemmifera*. *Molecular Biology and Evolution*. 2003; 20:1043–1050. [PubMed: 12716977]
- Koch MA, Haubold B, Mitchell-Olds T. Comparative evolutionary analysis of Chalcone Synthase and Alcohol Dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Molecular Biology and Evolution*. 2000; 17:1483–1498. [PubMed: 11018155]
- Lavergne S, Molofsky J. Increased genetic variation and evolutionary potential drive the success of an invasive grass. *Proceedings of the National Academy of Sciences of the United States of America*. 2007; 104:3883–3888. [PubMed: 17360447]
- Llaurens V, Castric V, Austerlitz F, Vekemans X. High paternal diversity in the self-incompatible herb *Arabidopsis halleri* despite clonal reproduction and spatially restricted pollen dispersal. *Molecular Ecology*. 2008; 17:1577–1588. [PubMed: 18266621]

- McDonald JH, Kreitman M. Adaptive Protein Evolution at the *Adh* Locus in *Drosophila*. *Nature*. 1991; 351:652–654. [PubMed: 1904993]
- Mitchell-Olds T. *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. *Trends Ecol. Evol.* 2001; 16:693–700.
- Miyashita NT. DNA variation in the 5' upstream region of the *Adh* locus of the wild plants *Arabidopsis thaliana* and *Arabis gemmifera*. *Molecular Biology and Evolution*. 2001; 18:164–171. [PubMed: 11158375]
- Nei M, Gojobori T. Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions. *Molecular Biology and Evolution*. 1986; 3:418–426. [PubMed: 3444411]
- Nordborg M, Hu TT, Ishino Y, et al. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology*. 2005; 3:1289–1299.
- Oyama RK, Clauss MJ, Formanova N, et al. The shrunken genome of *Arabidopsis thaliana*. *Plant Systematics and Evolution*. 2008; 273:257–271.
- Pauwels M, Saumitou-Laprade P, Holl AC, Petit D, Bonnin I. Multiple origin of metalicolous populations of the psuedometallophyte *Arabidopsis halleri* (Brassicaceae) in central Europe: the cpDNA testimony. *Molecular Ecology*. 2005; 14:4403–4414. [PubMed: 16313601]
- Pauwels M, Willems G, Roosens N, Frerot H, Saumitou-Laprade P. Merging methods in molecular and ecological genetics to study the adaptation of plants to anthropogenic metal-polluted sites: implications for phytoremediation. *Molecular Ecology*. 2008; 17:108–119. [PubMed: 17784915]
- Platt A, Horton M, Huang Y, et al. The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics*. 2010; 6(2):e1000843. [PubMed: 20169178]
- Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000; 155:945–959. [PubMed: 10835412]
- Ramos-Onsins SE, Mitchell-Olds T. mlcoalsim: Multilocus Coalescent Simulations. *Evolutionary Bioinformatics*. 2007; 2:41–44.
- Ramos-Onsins SE, Rozas J. Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution*. 2002; 19:2092–2100. [PubMed: 12446801]
- Ramos-Onsins SE, Stranger BE, Mitchell-Olds T. Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. *Genetics*. 2004; 166:373–388. [PubMed: 15020431]
- Rozas J, Gullaud M, Blandin G, Aguade M. DNA variation at the *rp49* gene region of *Drosophila simulans*: Evolutionary inferences from an unusual haplotype structure. *Genetics*. 2001; 158:1147–1155. [PubMed: 11454763]
- Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*. 2003; 19:2496–2497. [PubMed: 14668244]
- Sanyal A, Jackson SA. Comparative genomics reveals expansion of the FLC region in the genus *Arabidopsis*. *Molecular Genetics and Genomics*. 2006; 275:26–34. [PubMed: 16341708]
- Schmid KJ, Ramos-Onsins SE, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics*. 2005; 169:1601–1615. [PubMed: 15654111]
- Sharbel TF, Haubold B, Mitchell-Olds T. Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Molecular Ecology*. 2000; 9:2109–2118. [PubMed: 11123622]
- Shaw AJ. Genetic biogeography of the rare copper-moss, *Scopelophila-cataractae* (Pottiaceae). *Plant Systematics and Evolution*. 1995; 197:43–58.
- Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*. 1986; 73:751–754.
- Smith NGC, Eyre-Walker A. Adaptive protein evolution in *Drosophila*. *Nature*. 2002; 415:1022–1024. [PubMed: 11875568]
- Tajima F. Evolutionary Relationship of DNA-Sequences in Finite Populations. *Genetics*. 1983; 105:437–460. [PubMed: 6628982]

- Tajima F. Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*. 1989; 123:585–595. [PubMed: 2513255]
- Talke IN, Hanikenne M, Kramer U. Zinc-dependent global transcriptional control, transcriptional deregulation, and higher gene copy number for genes in metal homeostasis of the hyperaccumulator *Arabidopsis halleri*. *Plant Physiology*. 2006; 142:148–167. [PubMed: 16844841]
- Van Rossum F, Bonnini I, Féart S, et al. Spatial genetic structure within a metallicolous population of *Arabidopsis halleri*, a clonal, self-incompatible and heavy-metal-tolerant species. *Molecular Ecology*. 2004; 13:2959–2967. [PubMed: 15367112]
- Wall JD. Recombination and the power of statistical tests of neutrality. *Genetical Research*. 1999; 74:65–79.
- Wang WK, Schaal BA, Chiou YM, et al. Diverse selective modes among orthologs/paralogs of the chalcone synthase (Chs) gene family of *Arabidopsis thaliana* and its relative *A-halleri* ssp *gemmifera*. *Molecular Phylogenetics and Evolution*. 2007; 44:503–520. [PubMed: 17611127]
- Watterson GA. Number of Segregating Sites in Genetic Models without Recombination. *Theoretical Population Biology*. 1975; 7:256–276. [PubMed: 1145509]
- Wilder JA, Hollocher H. Recent radiation of endemic Caribbean *Drosophila* of the dunni subgroup inferred from multilocus DNA sequence variation. *Evolution*. 2003; 57:2566–2579. [PubMed: 14686532]
- Wright F. The Effective Number of Codons Used in a Gene. *Gene*. 1990; 87:23–29. [PubMed: 2110097]
- Wright SI, Lauga B, Charlesworth D. Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. *Molecular Ecology*. 2003; 12:1247–1263. [PubMed: 12694288]
- Zeng K, Fu YX, Shi SH, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*. 2006; 174:1431–1439. [PubMed: 16951063]

Figure 1.

Evolutionary models used and their respective parameters. Models are described from the present to the past. *SNM*: the Standard Neutral Model with parameters $4N\mu$ and $4Nr$, that are common parameters for all loci. These two parameters are present in all other models.

Change N_e : the model changes the population size instantaneously at time T_c (relative value from the beginning that is 1). *Bottleneck*: a single population shrinks at time T_c to a size N_b during $T_b=0.01$ (this value is arbitrarily assigned, as N_b and T_b are correlated, Fay and Wu, 1999). After T_b , the population changes to size N_A . *Admixture*: A single population splits into more after T_c . During $T_b=0.01$ the refugia became isolated with population sizes N_1 and N_2 . After T_b , the population merges into one single population with size N_A . *Subdivision*: Two populations have a migration defined as $4Nm$. The second population has a population size N_2 .

Figure 2. Results from STRUCTURE software. A. Average and Standard deviation of the Likelihood values versus the number of groups (K) assumed. B. Evanno ΔK values versus K indicate that the model with $K=2$ is the most supported. C. Assignment of individuals to different populations. The vertical bars indicate the percentage of the sample belonging to one or the other population (marked with black or white).

Figure 3. Posterior distributions (white histogram and the black continuous line) and prior distributions (discontinuous line) for the four parameters of the subdivision model assuming a single θ and R per nucleotide for all loci. The x-axis is on a natural logarithmic scale.

Figure 4. Posterior distributions (white histogram and the black continuous line) and prior distributions (discontinuous line) for the two common parameters (for all loci) of the subdivision model assuming different θ and R per nucleotide and for each locus. The x-axis is on a natural logarithmic scale.

Table 1

Description of loci

Locus	Related <i>A. thaliana</i> Locus	Description
1. <i>At1est10</i>	<i>At1g32440</i>	chloroplast pyruvate kinase beta subunit
2. <i>At1est13</i>	<i>At1g79560</i>	chloroplast FtsH protease
3. <i>At1est8</i>	<i>At1g70330</i>	adenosine transporter
4. <i>At3est41</i>	<i>At3g03780</i>	cytosolic methionine synthase
5. <i>At3est47</i>	<i>At3g20040</i>	putative hexokinase
6. <i>At3est62</i>	<i>At3g13180</i>	sun family protein
7. <i>At4est79</i>	<i>At4g26630</i>	GTP binding/RNA binding protein
8. <i>AtI31</i>	<i>At1g71696</i>	putative carboxypeptidase
9. <i>AtI6</i>	<i>At1g17240</i>	leucine-rich repeat family protein
10. <i>AtIII18</i> ×5	<i>At3g48510</i>	unknown
11. <i>AtIII19</i> ×5	<i>At3g51070</i>	dehydration-responsive protein-related
12. <i>AtIII6</i> ×5	<i>At3g16440</i> and <i>At3g16450</i>	Myrosinase-binding protein and jacalin lectine family protein
13. <i>AtIV11</i>	<i>At4g09680</i>	unknown
14. <i>AtIV20</i>	<i>At4g31200</i>	SWAP/surp domain-containing protein
15. <i>AtV11</i>	<i>At5g24240</i>	phosphatidylinositol 3- and 4-kinase family
16. <i>AtV26</i>	unknown	noncoding
17. <i>AtV27</i>	<i>At5g49880</i>	mitotic checkpoint family protein
18. <i>AtV8</i>	<i>At5g16570</i>	cytosolic glutamine synthetase
19. <i>AtV9</i>	<i>At5g19120</i>	aspartic-type endopeptidase/pepsin A
20. <i>GOLM19</i>	unknown	noncoding
21. <i>GOLM23</i>	unknown	noncoding
22. <i>GOLM25</i>	<i>At1g60490</i>	phosphatidylinositol 3-kinase
23. <i>PCP17C2</i>	<i>At2g02100</i>	protease inhibitor
24. <i>PCP6C1</i>	<i>At1g68570</i>	chloroplast nitrite transporter

TABLE 2

Levels of haplotype and nucleotide variation at total and at silent positions

Locus	Total positions								Silent (synonymous plus non-coding) positions							
	#sites	S ^a	θ _{wat} ^b	θ _{Taj} ^c	K _{JC} ^d	R _m ^e	H _d ^f		#sites	S ^a	θ _{wat} ^b	θ _{Taj} ^c	K _{JC} ^d	R _m ^e	H _d ^f	
<i>At1est10</i>	552	47	0.0282	0.0377	0.104	2	0.848		324.00	44	0.0450	0.0627	0.186	2	0.848	
<i>At1est13</i>	587	7	0.0039	0.0031	0.041	0	0.318		293.17	7	0.0079	0.0063	0.085	0	0.318	
<i>At1est8</i>	431	9	0.0069	0.0071	0.073	0	0.864		99.82	6	0.0199	0.0257	0.234	0	0.848	
<i>At3est41</i>	492	21	0.0141	0.0181	0.053	10	1.000		229.81	12	0.0173	0.0237	0.095	7	1.000	
<i>At3est47</i>	465	5	0.0036	0.0029	0.074	0	0.803		285.62	4	0.0046	0.0041	0.116	0	0.742	
<i>At3est62</i>	488	31	0.0210	0.0267	0.045	1	0.803		331.47	30	0.0300	0.0377	0.066	1	0.803	
<i>At4est79</i>	510	1	0.0006	0.0003	0.043	0	0.167		407.78	1	0.0008	0.0004	0.051	0	0.167	
<i>At131</i>	487	1	0.0007	0.0006	0.043	0	0.303		270.01	1	0.0012	0.0011	0.066	0	0.303	
<i>At6</i>	721	8	0.0037	0.0018	0.076	0	0.576		164.29	2	0.0040	0.0020	0.149	0	0.318	
<i>AtIII18×5</i>	509	9	0.0059	0.0063	0.032	1	0.848		126.45	1	0.0026	0.0032	0.068	0	0.409	
<i>AtIII19×5</i>	497	10	0.0067	0.0041	0.044	0	0.758		343.05	10	0.0097	0.0060	0.065	0	0.758	
<i>AtIII6×5</i>	341	3	0.0029	0.0029	0.117	0	0.712		341.00	3	0.0029	0.0029	0.117	0	0.712	
<i>AtIV11</i>	723	5	0.0023	0.0024	0.074	2	0.879		227.24	2	0.0029	0.0021	0.115	0	0.439	
<i>AtIV20</i>	671	8	0.0039	0.0040	0.034	0	0.909		149.54	1	0.0022	0.0011	0.124	0	0.167	
<i>AtV11</i>	506	3	0.0020	0.0016	0.053	0	0.682		176.99	2	0.0037	0.0037	0.141	0	0.591	
<i>AtV26</i>	508	1	0.0007	0.0003	0.041	0	0.167		508.00	1	0.0007	0.0003	0.041	0	0.167	
<i>AtV27</i>	493	50	0.0336	0.0312	0.066	7	0.939		231.56	39	0.0558	0.0535	0.116	6	0.848	
<i>AtV8</i>	393	6	0.0051	0.0038	0.043	0	0.712		247.86	5	0.0067	0.0053	0.064	0	0.712	
<i>AtV9</i>	449	12	0.0089	0.0111	0.079	3	0.970		108.78	5	0.0152	0.0231	0.217	1	0.773	
<i>GOLM19</i>	574	3	0.0018	0.0010	0.093	0	0.345		574.00	3	0.0018	0.0010	0.093	0	0.345	
<i>GOLM23</i>	562	12	0.0071	0.0064	0.077	0	0.833		562.00	12	0.0071	0.0064	0.077	0	0.833	
<i>GOLM25</i>	552	15	0.0090	0.0111	0.044	0	0.818		270.12	12	0.0147	0.0209	0.092	0	0.621	
<i>PCP17C2</i>	586	5	0.0028	0.0014	0.074	0	0.318		453.19	5	0.0037	0.0018	0.092	0	0.318	
<i>PCP6C1</i>	581	20	0.0114	0.0113	0.063	1	0.939		314.00	15	0.0158	0.0175	0.114	1	0.924	
Average	528.25	12.17	0.0078	0.0082	0.0619	1.1250	0.6880		293.32	9.29	0.0115	0.0130	0.1077	0.7500	0.5819	
Variance			0.0001	0.0001	0.0005	6.0272	0.0712				0.0002	0.0003	0.0024	3.4130	0.0720	

^aSegregating sites

- ^b Watterson's theta (1975) estimate per site.
- ^c Average number of pairwise differences (Tajima, 1983) per site.
- ^d Nucleotide divergence per site with *A. thaliana* using Jukes and Cantor (1969) correction.
- ^e Minimum number of recombinant events (Hudson and Kaplan, 1985)
- ^f Haplotype diversity (Depaulis and Veuille, 1998)

TABLE 3
Levels of haplotype and nucleotide variation at synonymous, non-synonymous and non-coding positions

Locus	Synonymous positions						Non-synonymous positions						Non-coding positions					
	#sites	S ^a	θ _{Wat} ^b	θ _{Taj} ^c	K _{IC} ^d	K _{IC} ^d	#sites	S ^a	θ _{Wat} ^b	θ _{Taj} ^c	K _{IC} ^d	K _{IC} ^d	#sites	S ^a	θ _{Wat} ^b	θ _{Taj} ^c	K _{IC} ^d	K _{IC} ^d
<i>At1esf10</i>	67.00	17	0.0840	0.1230	0.2212	0.2212	224.00	3	0.0044	0.0022	0.0011	0.0011	257.00	27	0.0348	0.0469	0.1769	0.1769
<i>At1esf13</i>	90.17	0	0.0000	0.0000	0.0340	0.0340	293.83	0	0.0000	0.0000	0.0000	0.0000	203.00	7	0.0114	0.0091	0.1094	0.1094
<i>At1esf8</i>	99.82	6	0.0199	0.0257	0.2342	0.2342	323.18	2	0.0020	0.0010	0.0289	0.0289	-	-	-	-	-	-
<i>At3esf41</i>	79.81	6	0.0249	0.0344	0.0552	0.0552	244.19	4	0.0054	0.0076	0.0124	0.0124	150.00	6	0.0132	0.0180	0.1177	0.1177
<i>At3esf47</i>	49.62	0	0.0000	0.0000	0.0853	0.0853	175.38	1	0.0019	0.0010	0.0120	0.0120	236.00	4	0.0056	0.0049	0.1221	0.1221
<i>At3esf62</i>	49.47	4	0.0268	0.0358	0.0687	0.0687	154.53	1	0.0021	0.0035	0.0032	0.0032	282.00	26	0.0305	0.0380	0.0657	0.0657
<i>At4esf79</i>	28.78	0	0.0000	0.0000	0.0729	0.0729	100.22	0	0.0000	0.0000	0.0100	0.0100	379.00	1	0.0009	0.0004	0.0493	0.0493
<i>At431</i>	58.01	0	0.0000	0.0000	0.0723	0.0723	214.99	0	0.0000	0.0000	0.0141	0.0141	212.00	1	0.0016	0.0014	0.0648	0.0648
<i>At6</i>	164.29	2	0.0040	0.0020	0.1487	0.1487	549.71	4	0.0024	0.0012	0.0586	0.0586	-	-	-	-	-	-
<i>AtIII18×5</i>	126.45	1	0.0026	0.0032	0.0683	0.0683	374.55	7	0.0062	0.0064	0.0208	0.0208	-	-	-	-	-	-
<i>AtIII19×5</i>	53.05	1	0.0062	0.0031	0.1244	0.1244	150.95	0	0.0000	0.0000	0.0000	0.0000	290.00	9	0.0103	0.0065	0.0542	0.0542
<i>AtIII6×5</i>	-	-	-	-	-	-	-	-	-	-	-	-	341.00	3	0.0029	0.0029	0.1175	0.1175
<i>AtIV11</i>	153.24	2	0.0043	0.0031	0.1143	0.1143	488.76	2	0.0014	0.0017	0.0587	0.0587	74.00	0	0.0000	0.0000	0.1167	0.1167
<i>AtV20</i>	149.54	1	0.0022	0.0011	0.1239	0.1239	516.46	7	0.0045	0.0049	0.0104	0.0104	-	-	-	-	-	-
<i>AtV11</i>	88.99	0	0.0000	0.0000	0.1909	0.1909	325.01	1	0.0010	0.0005	0.0064	0.0064	88.00	2	0.0075	0.0074	0.0937	0.0937
<i>AtV26</i>	-	-	-	-	-	-	-	-	-	-	-	-	508.00	1	0.0007	0.0003	0.0406	0.0406
<i>AtV27</i>	69.56	12	0.0571	0.0444	0.1277	0.1277	257.44	12	0.0154	0.0122	0.0247	0.0247	162.00	27	0.0552	0.0574	0.1107	0.1107
<i>AtV8</i>	43.86	1	0.0076	0.0093	0.1441	0.1441	139.14	0	0.0000	0.0000	0.0000	0.0000	204.00	4	0.0065	0.0045	0.0481	0.0481
<i>AtV9</i>	108.78	5	0.0152	0.0231	0.2170	0.2170	335.22	5	0.0049	0.0043	0.0396	0.0396	-	-	-	-	-	-
<i>GOLM19</i>	-	-	-	-	-	-	-	-	-	-	-	-	574.00	3	0.0018	0.0010	0.0928	0.0928
<i>GOLM23</i>	-	-	-	-	-	-	-	-	-	-	-	-	562.00	12	0.0071	0.0064	0.0772	0.0772
<i>GOLM25</i>	87.12	4	0.0152	0.0202	0.0936	0.0936	278.88	3	0.0036	0.0018	0.0009	0.0009	183.00	8	0.0145	0.0212	0.0919	0.0919
<i>PCP17C2</i>	37.19	0	0.0000	0.0000	0.1482	0.1482	127.81	0	0.0000	0.0000	0.0158	0.0158	416.00	5	0.0040	0.0020	0.0868	0.0868
<i>PCP6C1</i>	88.00	3	0.0113	0.0098	0.1447	0.1447	266.00	5	0.0062	0.0042	0.0098	0.0098	226.00	12	0.0176	0.0204	0.1030	0.1030
Average	84.64	3.25	0.0141	0.0169	0.1245	0.1245	277.01	2.85	0.0031	0.0026	0.0164	0.0164	281.42	8.32	0.0119	0.0131	0.0915	0.0915
Variance			0.0005	0.0008	0.0033	0.0033			0.0000	0.0000	0.0003	0.0003			0.0002	0.0003	0.0011	0.0011

-: Not applicable

^aSegregating sites

^bWatterson's theta (1975) estimate per site.

^cAverage number of pairwise differences (Tajima, 1983) per site.

^dNucleotide divergence per site with *A. thaliana* using Jukes and Cantor (1969) correction.

TABLE 4

McDonald and Kreitman test and the proportion of adaptive substitutions

Locus	Polymorphisms				Polym (no singletons)				Fixed divergence ^d				MK test and α (syn vs nsyn)				MK test and γ (syn vs ncod)			
	syn ^b	nsyn	ncod	sil	syn	nsyn	ncod	sil	syn	nsyn	ncod	sil	P _{total} ^c	P _{No single}	α Total ^d	α No single	P _{Total}	P _{No single}	α Total	α No single
<i>At1est10</i>	17	3	27	44	16	0	23	39	5.77	0.00	32.55	38.39	1.0000	1.0000	-	-	0.0264	0.0223	0.72	0.75
<i>At1est13</i>	0	0	7	7	0	0	7	7	3.07	0.00	17.20	20.10	1.0000	1.0000	-	-	0.5453	0.5453	-	-
<i>At1est8</i>	6	2	-	6	6	0	-	6	19.79	9.17	-	19.79	1.0000	0.3037	0.28	1.00	-	-	-	-
<i>At3est41</i>	6	4	6	12	6	4	5	11	2.07	1.01	15.13	17.04	1.0000	1.0000	-0.37	-0.37	0.0382	0.0299	0.86	0.89
<i>At3est47</i>	0	1	4	4	0	0	3	3	4.23	2.02	27.09	31.29	0.4286	1.0000	-	-	1.0000	1.0000	-	-
<i>At3est62</i>	4	1	26	30	4	1	25	29	2.09	0.00	6.27	8.36	1.0000	1.0000	-	-	0.5870	0.5913	-1.17	-1.09
<i>At4est79</i>	0	0	1	1	0	0	0	0	2.10	1.01	18.60	20.69	1.0000	1.0000	-	-	1.0000	1.0000	-	-
<i>At431</i>	0	0	1	1	0	0	1	1	4.20	3.03	13.57	17.76	1.0000	1.0000	-	-	1.0000	1.0000	-	-
<i>At6</i>	2	4	-	2	0	2	-	0	24.25	30.15	-	24.25	0.6892	0.5013	-0.61	-	-	-	-	-
<i>AtIII18×5</i>	1	7	-	1	1	4	-	1	8.37	6.08	-	8.37	0.0743	0.3034	-8.63	-4.50	-	-	-	-
<i>AtIII19×5</i>	1	0	9	10	0	0	1	1	6.51	0.00	14.51	20.87	1.0000	1.0000	-	-	0.3803	1.0000	-3.04	-
<i>AtIII6×5</i>	-	-	3	3	-	-	2	2	-	-	4.32	4.32	-	-	-	-	-	-	-	-
<i>AtV11</i>	2	2	0	2	1	2	0	1	17.25	28.07	8.64	25.89	0.6359	1.0000	0.39	-0.23	1.0000	1.0000	1.00	1.00
<i>AtV20</i>	1	7	-	1	0	4	-	0	18.44	3.02	0.00	18.44	0.0005	0.0028	-41.74	-	-	-	-	-
<i>AtV11</i>	0	1	2	2	0	0	1	1	16.99	2.01	7.45	24.13	0.1500	1.0000	-	-	0.1108	0.3200	-	-
<i>AtV26</i>	-	-	1	1	-	-	0	0	-	-	20.55	20.55	-	-	-	-	-	-	-	-
<i>AtV27</i>	12	12	27	39	6	6	19	25	4.35	4.07	4.30	7.55	1.0000	1.0000	-0.07	-0.07	0.4164	0.2054	-1.27	-2.20
<i>AtV8</i>	1	0	4	5	1	0	1	2	5.50	0.00	9.29	13.56	1.0000	1.0000	-	-	1.0000	1.0000	-1.37	0.41
<i>AtV9</i>	5	5	-	5	5	2	-	5	20.73	12.32	-	20.73	0.4809	1.0000	-0.68	0.33	-	-	-	-
<i>GOLM19</i>	-	-	3	3	-	-	1	1	-	-	52.09	52.09	-	-	-	-	-	-	-	-
<i>GOLM23</i>	-	-	12	12	-	-	6	6	-	-	41.04	39.99	-	-	-	-	-	-	-	-
<i>GOLM25</i>	4	3	8	12	3	0	7	10	6.38	0.00	12.75	18.07	0.1923	1.0000	-	-	1.0000	1.0000	0.00	-0.17
<i>PCP17C2</i>	0	0	5	5	0	0	2	2	5.51	2.02	33.89	39.31	1.0000	1.0000	-	-	1.0000	1.0000	-	-
<i>PCP6C1</i>	3	5	12	15	1	2	8	9	12.10	2.01	18.19	30.19	0.0524	0.1206	-9.01	-11.02	0.3145	0.2252	-1.66	-4.32
Total	65	57	158	223	50	27	112	162	189.71	105.98	357.43	541.75	0.0469	1.0000	-0.57	0.03	0.1513	0.3954	-0.29	-0.19

When probabilities for each locus were corrected for multiple testing (Benjamini and Hochberg, 1995), only *AtV20* at P_{total} was significant.

-: Not applicable

- ^a Corrected with Jukes and Cantor (1969)
- ^b $\text{syn:cSynonymous, nsyn: Non-synonymous, ncod: Non-coding, sil: Silent}$
- ^c Probabilities calculated with the Fisher exact test
- ^d $\alpha = 1 - (\text{Polym(NSyn)} * \text{Fix(Syn)}) / (\text{Polym(Syn)} * \text{Fix(NSyn)})$

TABLE 5

Posterior Probabilities from a Hierarchical Regression Analysis in the five evolutionary models studied (assuming a unique θ and R for all loci)

Models	P_1^b	P_2^c	P_3^d	P_4^e
Subdivision	0.789	0.626	0.817	0.783
Admixture	0.098	0.200	0.091	0.153
Standard Neutral	0.060	0.027	0.065	0.029
Bottleneck	0.043	0.144	0.021	0.030
Change of Ne	0.009	0.003	0.006	0.006

^a Number of parameters for each model

^b Statistics used: Tajima's D, Fu's Fs, R2, Wall's B, Fay and Wu's H normalized and Watterson's θ

^c Statistics used: Tajima's D, Fu's Fs, Haplotype diversity, Wall's B, Fay and Wu's H normalized and Watterson's θ

^d Statistics used: Tajima's D, Fu's Fs, R2, Rozas' et al Z_A, Fay and Wu's H normalized and Watterson's θ

^e Statistics used: Tajima's D, Fu's Fs, Haplotype diversity, Rozas' et al. Z_A, Fay and Wu's H normalized and Watterson's θ

TABLE 6
Neutrality tests at silent (synonymous plus non-coding) positions and their significance under the SNM with different assumptions

Locus	D ^a	SNM ₁ ^b	SNM _{all} ^c	F _s ^d	SNM ₁	SNM _{all}	H _{if} ^e	SNM ₁	SNM _{all}	Z _{af} ^g	SNM ₁	SNM _{all}	B ^h	SNM ₁	SNM _{all}	C _{st} ⁱ	SNM ₁	SNM _{all}
<i>At1est10</i>	1.7983			2.7120			-0.2043			0.7611			0.6512			30.056		
<i>At1est13</i>	-0.7926			2.0983			-4.0776			0.9091			0.8333			16.020		
<i>At1est8</i>	1.1014			-0.6936			-1.1336			0.2626			0.0000			7.217		
<i>At3est41</i>	1.5495		NA	-7.7245		NA	0.3605		NA	0.2030		NA	0.0000		NA	13.526		NA
<i>At3est47</i>	-0.4182			-0.2893			-1.0575			0.3610			0.3333			4.875		
<i>At3est62</i>	1.1557			3.9733			-0.6269			0.6643			0.4828			21.168		
<i>At4est79</i>	-1.1405			-0.4757			0.3382		x	x		x	x		x	10.684		
<i>At131</i>	-0.1989			0.2974			0.5411		x	x		x	x		x	4.424		
<i>At6</i>	-1.4534			-1.3248			0.4493			0.0083			0.0000			4.838		
<i>AtIII8x5</i>	0.5405			0.7348			0.6088			-			-			7.643		
<i>AtIII9x5</i>	-1.5738			-0.1658			0.6489			0.6786			0.6667			12.141		
<i>AtIII6x5</i>	-0.0283			-0.6519			-0.4511			0.1919			0.0000			1.894		
<i>AtIV11</i>	-0.8487			-0.7246			0.5841			0.0182			0.0000			4.063		
<i>AtIV20</i>	-1.1495			-0.4757			0.3382		x	x		x	x		x	6.627		
<i>AtV11</i>	-0.0455			-0.1374			-0.4942			0.0455			0.0000			7.399		
<i>AtV26</i>	-1.1495			-0.4757			0.3382		x	x		x	x		x	8.529		
<i>AtV27</i>	-0.1834			1.3006			-1.7323			0.4459			0.2895			19.132		*
<i>AtV8</i>	-0.7315			-0.0250			-0.0733			0.5960			0.5000			7.208		
<i>AtV9</i>	1.9135			1.4996			-0.3420			0.6036			0.2500			19.669		
<i>GOLM19</i>	-1.6000			-0.5370			-1.4828			0.5050			0.5000			14.343		
<i>GOLM23</i>	-0.4047			-0.8773			0.9323			0.4120			0.3636			12.860		
<i>GOLM25</i>	1.7578			6.1635			-0.2983			0.6600			0.6364			28.371		
<i>PCP17C2</i>	-1.8309			0.3249			-2.3206			0.5041			0.5000			13.586		
<i>PCP6C1</i>	0.4475			-0.8037			-0.2269			0.4012			0.3571			13.642		
<i>C_{st}^j</i>	63.7554	*	*	46.5902			35.5494	*	*	47.4876	*	*	45.7716	*	*			
Average	-0.1369	-	-	0.1551	-	-	-0.3909	-	-	0.4332	-	-	0.3349	-	-			
Variance	1.3783	-	-	5.8695	-	-	1.3071	-	-	0.0675	-	-	0.0743	-	-			

Probabilities for the matrix of individual statistics were corrected for multiple testing using FDR.

Probabilities: blank: $P > 0.05$, *: $P \leq 0.05$

x': Not applicable

-.: Probability for the averages and variances are not calculated (the composite probability gives the information about the total data).

NA: Non-available result because the level of recombination of the locus *At3es41* is extremely improbable under the given assumptions (Minimum number of recombinations, $R_{M=7}$)

^aTajima's D test (Tajima, 1989).

^b Assuming a single θ and R for all loci.

^c Assuming different θ and R for each locus.

^d Fu's F_s test (Fu, 1997).

^e Ramos-Onsins and Rozas' R_2 test (Ramos-Onsins and Rozas, 2002).

^f Fay and Wu's H normalized test (Fay and Wu 2000, Zeng et al.2006).

^g Rozas et al. Z_A statistic (Rozas et al. 2001).

^h Wall's B test (Wall, 1999).

ⁱ Composite probability combining the following statistics: Tajima's D , Fu's F_s , R_2 , Wall's B , Fay and Wu's H normalized and Watterson's theta. Probabilities corrected for multiple testing using FDR (Benjamini and Hochberg, 1995).

^j Composite probability combining the given statistic of each locus. Probabilities corrected for multiple testing using FDR.