Genome-Wide Evidence for Efficient Positive and Purifying Selection in *Capsella grandiflora*, a Plant Species with a Large Effective Population Size

Tanja Slotte,*†,[1] John Paul Foxe,[2] Khaled Michel Hazzouri,[1] and Stephen I. Wright[1,3]

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

[2]Department of Biology, York University, Toronto, Ontario, Canada

[3]Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario, Canada

†Present address: Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden.

***Corresponding author:** E-mail: tanja.slotte@ebc.uu.se.

Associate editor: Naoki Takebayashi

Abstract

Recent studies comparing genome-wide polymorphism and divergence in Drosophila have found evidence for a surprisingly high proportion of adaptive amino acid fixations, but results for other taxa are mixed. In particular, few studies have found convincing evidence for adaptive amino acid substitution in plants. To assess the generality of this finding, we have sequenced 257 loci in the outcrossing crucifer *Capsella grandiflora*, which has a large effective population size and low population structure. Using a new method that jointly infers selective and demographic effects, we estimate that 40% of amino acid substitutions were fixed by positive selection in this species, and we also infer a low proportion of slightly deleterious amino acid mutations. We contrast these estimates with those for a similar data set from the closely related *Arabidopsis thaliana* and find significantly higher rates of adaptive evolution and fewer nearly neutral mutations in *C. grandiflora*. In agreement with results for other taxa, genes involved in reproduction show the strongest evidence for positive selection in *C. grandiflora*. Taken together, these results imply that both positive and purifying selection are more effective in *C. grandiflora* than in *A. thaliana*, consistent with the contrasting demographic history and effective population sizes of these species.

Key words: adaptive evolution, alpha, distribution of fitness effects, McDonald–Kreitman test, nearly neutral theory, site frequency, spectrum.

Introduction

Clarifying the contribution of neutral, beneficial, and deleterious changes to DNA diversity and evolution is of central interest in evolutionary genetics. Over the last several decades, the predominant model has been the neutral theory, first proposed by Kimura (1968), which assumes that the majority of changes that we observe as polymorphism or divergence in DNA sequences are selectively neutral. Although alternative models that allocate a greater role to positive selection have been proposed (Gillespie 2000, 2001), the neutral theory has dominated research in molecular evolution and population genetics, and it has contributed useful null models for testing selection.

Recently, results on the prevalence of adaptive evolution in Drosophila have begun to call the general validity of the neutral theory into question (reviewed in Hahn 2008; Sella et al. 2009). Indeed, several studies have found evidence for a surprisingly high proportion of adaptive amino acid substitutions of about 40–50% in Drosophila (Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Andolfatto 2005; Welch 2006; Begun et al. 2007; Eyre-Walker and Keightley 2009; but see Sawyer et al. 2003). These estimates were obtained using extensions of the classic McDonald–Kreitman (MK) test (McDonald and Kreitman 1991) that use data on polymorphism and divergence in order to estimate the proportion of adaptive substitutions between lineages, α (Charlesworth 1994; Fay et al. 2001; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Eyre-Walker and Keightley 2009). Based on these results and evidence for a genome-wide effect of positive selection on neutral diversity, it has been proposed that population geneticists should abandon the neutral theory in favor of models that incorporate recurrent positive selection (Hahn 2008).

Whereas data for Drosophila show consistently high proportions of adaptive fixations, results for other taxa are mixed. In humans, there is evidence for a general excess of nonsynonymous polymorphism indicative of the prevalence of slightly deleterious mutations, and estimates of α are low, ranging from zero (Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005) up to 10–20% (Boyko et al. 2008). An excess of low-frequency nonsynonymous polymorphism is also seen in *Saccharomyces* (Doniger et al. 2008; Liti et al. 2009), whereas the fraction of adaptive nonsynonymous substitutions in *Escherichia coli* appears to be extremely high (>50%; Charlesworth and Eyre-Walker 2006). A recent study on the house mouse subspecies *Mus musculus castaneus* also found evidence for a high proportion of adaptive substitutions (57%; Halligan et al. 2010), whereas intermediate values of α have been estimated for chicken (~20%; Axelsson and Ellegren 2009) and in the long-lived outcrossing forest

tree *Populus tremula* (~30%; Ingvarsson 2010). Apart from the recent results for *Populus*, widespread adaptive amino acid substitution seems to be rare in plant taxa (Gossmann T, Song B-H, Windsor AJ, Mitchell-Olds T, Eyre-Walker A, unpublished data). For instance, the selfing plant *Arabidopsis thaliana* exhibits an excess of nonsynonymous polymorphism and shows little evidence for adaptive substitutions (Bustamante et al. 2002; Nordborg et al. 2005; Kim et al. 2007), and similar results were found for its close outcrossing relative *A. lyrata* (Foxe et al. 2008).

It is of considerable interest to understand why the proportion of adaptive substitutions and slightly deleterious polymorphism differs so much between taxa (Sella et al. 2009). One general possibility is that differences in effective population size are influencing the substitution rate of beneficial mutations and the efficacy of both positive and negative selection (Ohta 1973). In small populations, the rate of adaptive substitution is reduced, slightly deleterious mutations can fix by random genetic drift, and the level of deleterious polymorphism can be increased; differences in effective population size may thus drive substantial differences in patterns of selected polymorphism and divergence across species (Ohta 1973, 1992; Kimura 1983). The relationship between the rate of adaptive substitutions and effective population size may not be linear, however, and more complicated models including multiple loci, linkage, and/or environmental fluctuations often predict a nonlinear relationship where the rate of adaptive substitution levels off with increasing population size (Gillespie 2000, 2004; for an overview, see Lynch 2007).

A number of studies have found a correlation between population size and levels of constraint on protein sequences, consistent with the idea that the efficacy of purifying selection varies with the effective population size (Akashi 1996; Eyre-Walker 2002; Lindblad-Toh et al. 2005; Woolfit and Bromham 2005; Popadin et al. 2007; Wright and Andolfatto 2008; Ellegren 2009). It is also conceivable that the efficacy of positive selection and thereby the proportion of adaptive substitutions vary with the effective population size (Eyre-Walker 2006; Betancourt et al. 2009; Sella et al. 2009; but see Bachtrog 2008). In agreement with this, several large-scale genomic studies have found evidence for differences in the efficacy of selection across genomic regions that differ in their recombination rates and are expected to experience different effective population sizes (Kliman and Hey 1993; Presgraves 2005; Haddrill et al. 2007; Betancourt et al. 2009; but see Bullaughey et al. 2009). The contrast between estimates of α for humans with those for *Drosophila* and *M. musculus castaneus* is consistent with a population size effect as both *Drosophila* and *M. musculus castaneus* have considerably larger effective population size than humans and also show higher α (Eyre-Walker and Keightley 2009; Halligan et al. 2010). Likewise, a high estimate of α is obtained for *E. coli*, which has an extremely large effective population size (Charlesworth and Eyre-Walker 2006).

Another possible factor influencing adaptive substitution is population structure. Analytical work suggests that population structure affects the rate of adaptation mainly through its effects on the effective population size, but population subdivision can also hinder the fixation of beneficial mutations under models where offspring production is constant for all demes (Whitlock 2003). Indeed, there is strong population differentiation in several species that appear to have a low proportion of adaptive substitutions (*A. thaliana*, Nordborg et al. 2005; Bakker et al. 2006; Ostrowski et al. 2006; Schmid et al. 2006; François et al. 2008; *A. lyrata*, Muller et al. 2008; Ross-Ibarra et al. 2008; *Saccharomyces cerevisiae* and *S. paradoxus*, Liti et al. 2009). So far, however, there are not enough data points to draw any firm conclusions about the factors that govern the proportion of adaptive substitutions, and the compared organisms differ in numerous other biological aspects than their effective population size and population structure. Furthermore, because the effect of population size on substitution rates depends on the distribution of fitness effects (DFE) of new mutations, it would be preferable to assess this distribution as well as the proportion of adaptive fixations (Woolfit 2009).

Here, we sequenced 257 loci across the genome of *Capsella grandiflora*, an annual crucifer that is closely related to *A. thaliana* (divergence time ~10 Ma; Koch et al. 2000). *Capsella grandiflora* is an obligate outcrosser with evidence for a long-term stable and large effective population size (N_e ~500,000) and little population structure (Foxe et al. 2009). We therefore expect natural selection to be highly efficient in this species. Here, we estimate the proportion of nonsynonymous substitutions fixed by positive selection, as well as the distribution of deleterious fitness effects of new nonsynonymous mutations, using a new method that jointly estimates and corrects for demographic changes (Eyre-Walker and Keightley 2009).

We contrast our estimates for *C. grandiflora* with those for a similar data set from *A. thaliana* and find evidence for both a significantly higher rate of adaptive nonsynonymous fixations in *C. grandiflora* and a significantly lower proportion of nearly neutral nonsynonymous mutations in *C. grandiflora*. Altogether, these results imply that both positive and purifying selection are more efficient in *C. grandiflora* than in *A. thaliana*, in line with expectations given the demographic history and effective population sizes of these species.

Materials and Methods

Data

We designed primers in exonic regions conserved between *A. thaliana* and *Brassica rapa* using Primer3 (Rozen and Skaletsky 2000), without regard to gene function. As targets for primer design, we extracted exons that were at least 800 bp long and had a Blast hit to a *B. rapa* GSS sequence from *A. thaliana* Col pseudochromosomes (GenBank accession numbers: NC_003070.6, NC_003071.4, NC_003074.5, NC_003075.4, and NC_003076.5). A total of 527 primer pairs were designed using this strategy. In addition, we designed 65 primer pairs to amplify exonic regions in candidate genes for flower development, based on Gene Ontology annotation (GO:0048444, GO:0048451,

GO:0048446, GO:0048498, GO:00110093, GO:0048441, GO:0048433, and GO:0048497).

We extracted genomic DNA from six *C. grandiflora* individuals (12 haploid genomes) from five populations (supplementary text and supplementary table S1, Supplementary Material online), one *C. rubella* individual, and one individual of the outgroup species *Neslia paniculata*, using a DNeasy kit (Qiagen, Hilden, Germany). Polymerase chain reaction (PCR) amplifications were performed in 25-μl volumes, and PCR products were sequenced on an ABI 3730 sequencer at the Genome Quebec Innovation Centre (McGill University, Canada). Sequences were basecalled using phred v. 0.020425.c (Ewing and Green 1998; Ewing et al. 1998) and trimmed to yield an error rate of $\leq$0.05 for each strand and full overlap of sequences on both strands. Heterozygous single nucleotide polymorphisms were scored using the "call secondary peaks" option in Sequencher v. 4.6 (Gene Codes Corporation, Ann Arbor, MI) or using PolyPhred (Nickerson et al. 1997; Rieder et al. 1998) as implemented in CodonCode Aligner v. 2.0.6 (CodonCode Corporation, Dedham, MA). We inspected chromatograms manually to verify basecalls at all heterozygous positions and discarded loci exhibiting double peaks in the selfing *C. rubella* to avoid inclusion of highly similar duplicate loci. In total, we obtained reliable *C. grandiflora* sequences for 347 loci. For each locus, we used BlastN to check that we had amplified the intended gene and extracted the homologous *A. thaliana* region for annotation. We aligned nucleotide sequences based on their translated amino acid sequences using T-COFFEE (Notredame et al. 2000). Codons that had positions with alignment gaps or missing data were removed prior to analyses. Here, we present analyses of 257 loci for which we obtained a minimum of eight haploid *C. grandiflora* sequences as well as a sequence for the outgroup species *N. paniculata*. Primer sequences are given in supplementary table S3, Supplementary Material online. A total of 29 of these loci are candidate loci for flower development, but all analyses presented are based on the full data set as estimates of α for candidate and background loci were not significantly different (supplementary text, Supplementary Material online). Sequences have been submitted to GenBank under accession numbers GU947896-GU949534.

For analyses of *A. thaliana* polymorphism and divergence to *A. lyrata*, we used data from Nordborg et al. (2005) with alignments to *A. lyrata* from Foxe et al. (2008). We filtered this data set to obtain a scattered sample representing worldwide diversity in *A. thaliana* by selecting 20 individuals from different populations (supplementary text, Supplementary Material online) and excluding loci of <60-bp aligned length after removing codons with missing data. The analyzed *A. thaliana* data set contained a total of 483 loci.

Sequence Analysis

We obtained estimates of the distribution of deleterious fitness effects of new nonsynonymous mutations and the proportion of nonsynonymous divergence driven by positive selection (α) using the method of Eyre-Walker and Keightley (2009) (http://homepages.ed.ac.uk/eang33/) with confidence intervals (CIs) and standard errors derived from 200 bootstrap replicates with resampling over loci. For these analyses, synonymous sites were assumed to evolve neutrally, and we obtained estimates of the DFE and α for nonsynonymous sites. Estimates of the number of nonsynonymous and synonymous sites and divergence differences were obtained by the method of Goldman and Yang (1994) using the Codeml program in PAML (version 4.2b; Yang 2007). We assumed a transition/transversion bias and estimated codon frequencies from average nucleotide frequencies at the three-codon positions (F3 $\times$ 4). Synonymous and nonsynonymous site frequency spectra and summary statistics were obtained using Polymorphorama, a perl script written by D. Bachtrog and P. Andolfatto (Bachtrog and Andolfatto 2006). Summary statistics of polymorphism and divergence at each locus are found in supplementary tables S3 and S4 (Supplementary Material online), respectively.

We tested whether α and properties of the DFE differed between *C. grandiflora* and *A. thaliana* by comparing estimates from 200 bootstrap replicates for each species, as in Keightley and Eyre-Walker (2007). Additional estimates of α were obtained using the maximum likelihood method of Bierne and Eyre-Walker (2004), allowing constraint, neutral diversity, and neutral divergence to vary across loci. For these analyses, we excluded polymorphisms segregating at $\leq$15% and used fixed Jukes–Cantor–corrected divergence differences obtained from Polymorphorama. We controlled for differences between data sets in the level of constraint by repeating all analyses of the DFE and α on a subset of 362 *A. thaliana* loci, which did not differ from those sequenced in *C. grandiflora* in terms of d_N/d_S between *A. thaliana* and *A. lyrata* (supplementary text and supplementary tables S5l, Supplementary Materia online). As all results remained qualitatively unaltered, we report analyses based on the full data set in the main text.

To test for selection on individual loci, we conducted single-locus likelihood ratio tests in MK test 2.0 (Welch 2006; Obbard et al. 2009). For these tests, we compared the fit of a three-parameter model with α set to zero with a four-parameter model where α was estimated using a likelihood ratio test. We applied false discovery rate (FDR) correction (Benjamini and Hochberg 1995) to correct for multiple tests and summarized the MK tables using the neutrality index (NI), defined as $(P_N/D_N)/(P_S/D_S)$ (Rand and Kann 1996). Following Li et al. (2008), we plot $-$log10(NI) after adding pseudocounts of 1 to each cell of the MK table to assure the NI was defined. We inspected the *A. thaliana* annotations for all genes that were significant at $\leq$5% nominal significance level. Although the details of gene function and regulation may differ between Capsella and *A. thaliana*, the high coding sequence similarity and large-scale similarity in gene content between these species (Boivin et al. 2004) suggest that the *A. thaliana* gene annotation can be informative about gene function in Capsella.

Table 1. Summary Statistics for the Two Analyzed Data Sets. The Total Number of Synonymous and Nonsynonymous Sites N_{sites}, Polymorphisms (S), and Divergence (K) to the Outgroup Are Given.

Species	Site Class	N_{sites}	S	K^a	θ_W	π	D_{Tajima}
Capsella grandiflora	Synonymous	32,169	1,686	5,353	0.0228 (0.0179)	0.0230 (0.0185)	0.009 (0.7623)
	Nonsynonymous	94,122	491	1,715	0.0019 (0.0025)	0.0018 (0.0024)	−0.3449 (0.8701)
Arabidopsis thaliana	Synonymous	36,716	943	5,853	0.0077 (0.0103)	0.0072 (0.0117)	−0.2504 (1.1043)
	Nonsynonymous	109,207	804	3,172	0.0020 (0.0032)	0.0017 (0.0032)	−0.4922 (0.9822)

NOTE.—Means of θ_W, π, and Tajima's D (D_{Tajima}) for each site class are listed followed by standard deviations in parentheses.
[a] Total divergence estimated using PAML.

Results

Polymorphism and Divergence

For this study, we sequenced 257 exonic regions in *C. grandiflora* and the outgroup species *N. paniculata*. Neslia was chosen as an outgroup because of its appropriate phylogenetic distance and large-scale chromosomal collinearity with Capsella (Lysak et al. 2006). In addition, we analyzed data for 483 exonic regions sequenced in *A. thaliana* as part of a large-scale assessment of polymorphism (Nordborg et al. 2005) and aligned to *A. lyrata* in Foxe et al. (2008).

The *C. grandiflora* data set contained a total of 126,291 sites, of which 2,177 were polymorphic in *C. grandiflora* and 7,068 exhibited divergence differences to *N. paniculata* (table 1). *Capsella grandiflora* harbored substantial synonymous diversity (mean θ_W and mean $\pi = \sim$2.3%; table 1) and the mean Tajima's D at synonymous sites was close to zero (table 1). At nonsynonymous sites, polymorphism levels were reduced by about an order of magnitude (mean $\theta_W = 0.19\%$ and mean $\pi = 0.18\%$; table 1) and mean Tajima's D was slightly negative for these sites, indicating a shift in the site frequency spectrum toward low-frequency polymorphisms (fig. 1).

The *A. thaliana* data set analyzed here contained a total of 145,923 sites, with 1,747 intraspecific polymorphisms and 8,998 divergence differences to *A. lyrata*. Synonymous diversity was considerably lower than in *C. grandiflora* (mean $\theta_W = 0.8\%$ and mean $\pi = 0.7\%$; table 1), although levels of nonsynonymous polymorphism were similar to those in *C. grandiflora* (mean $\theta_W = 0.2\%$ and mean $\pi = 0.18\%$; table 1). Tajima's D was negative for both synonymous and nonsynonymous polymorphisms, but the shift toward low-frequency polymorphisms was more pronounced for nonsynonymous sites (table 1).

Overall, the average ratio of nonsynonymous to synonymous substitution d_N/d_S was higher between *A. thaliana* and *A. lyrata* than between *C. grandiflora* and *N. paniculata* (0.22 vs. 0.13), although d_S was similar for both comparisons (*C. grandiflora* − *N. paniculata* $d_S = 0.18$ and *A. thaliana* − *A. lyrata* $d_S = 0.16$). However, the difference between *C. grandiflora* and *A. thaliana* in the ratio of nonsynonymous to synonymous polymorphism was much more pronounced than the difference in d_N/d_S (*A. thaliana* $P_N/P_S = 0.85$ and *C. grandiflora* $P_N/P_S = 0.29$; Fisher's exact test $P < 2.2 \times 10^{-16}$; table 1 and fig. 1).

Distribution of Deleterious Fitness Effects and Proportion of Adaptive Substitutions

We estimated properties of the DFE and α using a new method that jointly infers and corrects for demographic changes (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). This method models the deleterious fitness effects of new mutations in the selected class (here nonsynonymous mutations) by a gamma distribution with shape parameter β. Following Eyre-Walker and Keightley (2009), we compare estimates of the DFE of nonsynonymous mutations for *C. grandiflora* and *A. thaliana* by assessing the percentage of mutations falling into different ranges in terms of $N_e s$.

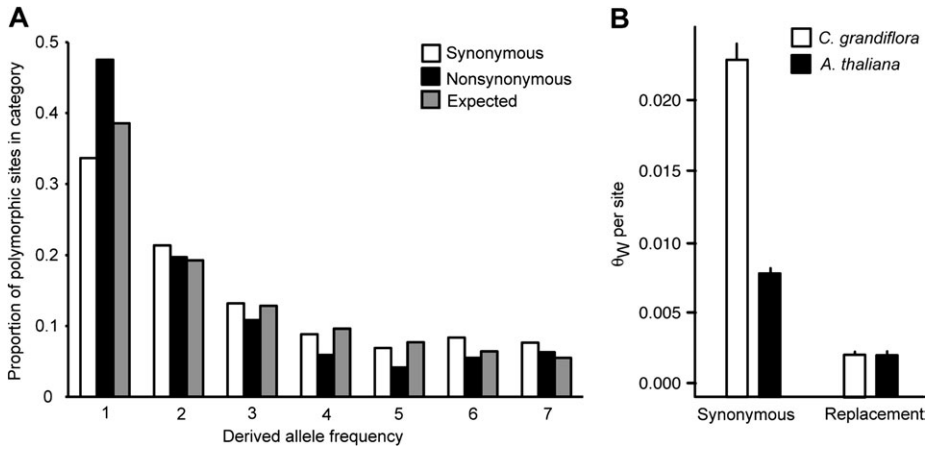

FIG. 1. (*A*) Expected and observed synonymous and nonsynonymous site frequency spectra for *Capsella grandiflora*. (*B*) Polymorphism levels (±standard error across loci) at synonymous and nonsynonymous sites in *C. grandiflora* and *Arabidopsis thaliana*.

Table 2. Estimates of the Proportion of Adaptive Nonsynonymous Substitutions (α) and the Percentage of New Deleterious Nonsynonymous Mutations Falling Within Different $N_e s$ Ranges in *Capsella grandiflora* and *Arabidopsis thaliana*.

	Percentage of Mutations in $N_e s$ Range			
Species	0–1	1–10	>10	α
C. grandiflora	7 (1–9)	7 (4–21)	86 (77–89)	0.40 (0.21 to 0.89)
A. thaliana	20 (16–25)	14 (7–21)	66 (59–71)	−0.08 (−0.46 to 0.15)

NOTE.—Point estimates are given, followed by 95% CIs in parentheses.

Table 3. *P* Values for One-Sided Tests of Significance of Differences in the DFE and α between *Capsella grandiflora* and *Arabidopsis thaliana*, Based on 200 Pairs of Bootstrap Replicates.

Estimate	Null Hypothesis	*P* Value
Effectively neutral mutations ($N_e s$ 0–1)	Cg ≥ At	<0.005
Strongly deleterious mutations ($N_e s$ >10)	Cg ≤ At	<0.005
α	Cg ≤ At	<0.005
β	Cg ≤ At	0.200

NOTE.—We tested the null hypothesis that the proportion of effectively neutral mutations for *C. grandiflora* was greater than or equal to that of *A. thaliana*. For strongly deleterious mutations, adaptive substitutions, and the shape parameter β, the null hypothesis was that values for *C. grandiflora* were smaller than or equal to those for *A. thaliana*. Values reported in the text are two times these *P* values.

Based on this method, we find that *C. grandiflora* and *A. thaliana* differ markedly in terms of the proportion of nonsynonymous mutations that are nearly neutral. In *C. grandiflora*, our estimates suggest that there are relatively few effectively neutral nonsynonymous mutations; only ~7% fall into this category ($N_e s$ = 0–1; table 2). In contrast, in *A. thaliana*, ~20% of new nonsynonymous mutations appear effectively neutral, and the difference between these proportions is significant (tables 2 and 3). Furthermore, in *C. grandiflora*, 86% of new nonsynonymous mutations are so strongly selected against that they very rarely fix ($N_e s >$ 10, following Eyre-Walker and Keightley (2009); table 2), whereas this proportion is significantly lower, 66% in *A. thaliana* (95% CI: 59–71%; tables 2 and 3).

Our estimates of the proportion of adaptive substitutions also differ significantly between *C. grandiflora* and *A. thaliana* (tables 2 and 3). The point estimate of the fraction of nonsynonymous substitutions fixed by positive selection between *C. grandiflora* and *N. paniculata* is 40%, and although the CI is wide (95% CI: 21–89%), it does not encompass zero. In contrast, α is not significantly different from zero for *A. thaliana* (table 2).

If the set of loci that we analyzed in *A. thaliana* were less functionally constrained than those in the *C. grandiflora* data set, then this could result in underestimates of α and an enrichment of nearly neutral mutations in *A. thaliana*. To assess this possibility, we examined d_N/d_S between *A. thaliana* and *A. lyrata* for a subset of the loci analyzed in *C. grandiflora* (supplementary text, Supplementary Material online). There was a weak but significant difference in d_N/d_S between *A. thaliana* and *A. lyrata* between the 483 loci analyzed in *A. thaliana* and loci we sequenced in *C. grandiflora*, which may indicate a difference in constraint across sets of loci. However, controlling for the difference in d_N/d_S, we obtain an even lower estimate of α for *A. thaliana*, suggesting that the above comparison is conservative (supplementary text and supplementary table S5, Supplementary Material online). Furthermore, we still infer a lower proportion of nearly neutral mutations and a higher proportion of strongly selected mutations in *C. grandiflora*, suggesting that differences in gene sets are not driving the result (supplementary text and supplementary table S5, Supplementary Material online).

Methods to estimate α that rely on summing counts of polymorphisms and divergence differences across loci can be biased if the effective population size varies across the genome and is correlated with the degree of constraint, and modified estimators that are insensitive to this effect have therefore been devised (Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Welch 2006). For comparison, we therefore estimated α using the maximum likelihood method of Bierne and Eyre-Walker (2004), which is insensitive to biases resulting from pooling counts across loci. After excluding low-frequency (≤15%) polymorphisms, this method gives a slightly lower point estimate of α for *C. grandiflora*, about 32% (95% CI: 19–43%), and the estimate for *A. thaliana* is still negative and nonsignificant (95% CI: −37% to 10%, point estimate −9%). Thus, these estimates are in good agreement with those obtained using the method of Eyre-Walker and Keightley (2009).

Tests for Selection on Individual Loci

We conducted MK tests on each individual locus in *C. grandiflora* and *A. thaliana*. We summarized the MK tables using the NI, defined as $(P_N/D_N)/(P_S/D_S)$ (Rand and Kann 1996). Over all tested loci, there was a shift toward higher values of the negative log of the NI (−log10(NI); following Li et al. 2008), suggesting a greater prevalence of positive selection in *C. grandiflora* (fig. 2). At a nominal 5% significance level, there was an excess of loci exhibiting a significant deviation from neutrality for *C. grandiflora* (20 significant tests vs. 13 expected). A total of 17 of these were in the direction of positive selection (fig.2; supplementary table S6, Supplementary Material online). In contrast, for *A. thaliana*, there were fewer significant genes than expected (20 significant tests vs. 24 expected; fig. 2; supplementary table S5, Supplementary Material online), and 15 of these showed evidence of purifying selection.

Eight loci in the *C. grandiflora* data set were significant at an FDR of ≤10%, whereas no *A. thaliana* loci were significant after multiple test correction. The *A. thaliana* homologs of four of these *C. grandiflora* loci, including the only two loci significant at ≤5% FDR, are associated with plant reproduction. Three of these loci encode proteins that are likely involved in pollen development (AT3G52080/CHX28, AT4G20050/QRT3 and AT4G14080/MEE48; fig. 2), whereas the fourth, AT3G46510, encodes a protein that interacts with a class of receptor kinases similar to the S-locus

A

C. grandiflora			*A. thaliana*		
Locus Tag	Type of selection	P-value	Locus Tag	Type of selection	P-value
AT3G52080**	+	0.00002	AT3G55830 AT3G55840	-	0.00255
AT3G46510**	+	0.00034	AT2G25590	+	0.00564
AT4G20050**	+	0.00065	AT2G26850	-	0.00628
AT2G45220**	+	0.00084	AT5G01110	-	0.01495
AT4G14080*	+	0.00159	AT1G55540	-	0.01636
AT1G07740*	-	0.00194	AT1G61710	-	0.01653
AT4G37770*	-	0.00243	AT3G01830 AT3G01840	+	0.02371
AT4G10710*	+	0.00274	AT1G78955	-	0.02402
AT4G25360	+	0.00623	AT2G40830	-	0.02538
AT5G39500	+	0.01241	AT3G55090	-	0.02612
AT1G18900	+	0.01388	AT1G74070	-	0.03002
AT1G13980	+	0.01732	AT1G04550	-	0.03130
AT2G46020	+	0.02196	AT1G24490	+	0.03266
AT5G19640	+	0.02559	AT5G24400	-	0.03290
AT1G12570	+	0.02573	AT5G13250	-	0.03910
AT5G23450	+	0.02678	AT4G23040	-	0.04149
AT2G23450	+	0.02776	AT2G17060	+	0.04376
AT4G02460	-	0.03325	AT1G22650	+	0.04603
AT3G23410	+	0.04830	AT5G45280	-	0.04712
AT3G24180	+	0.04887	AT4G26200	-	0.04878

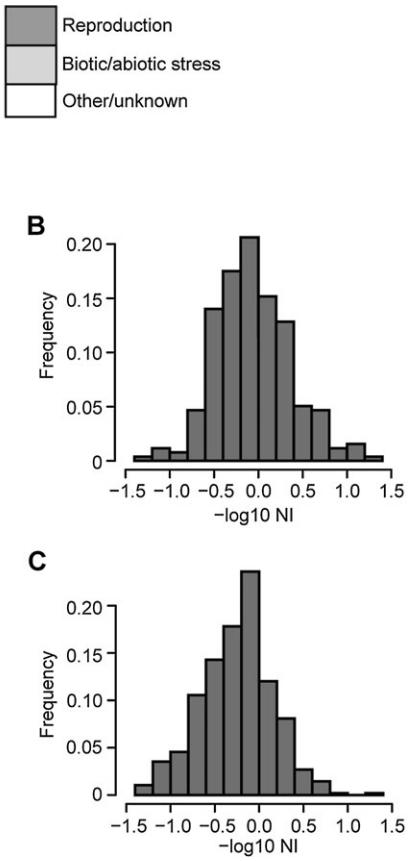

Fig. 2. (*A*) *P* values of single-locus likelihood ratio tests for selection and the direction of deviation (loci with deviations in direction of positive vs. purifying selection are marked with a + or − sign, respectively) for *Capsella grandiflora* (left) and *Arabidopsis thaliana* (right). Values for loci with significant deviations from neutrality at ≤5% nominal significance level (significance at ≤10% FDR is indicated by an asterisk) are shown, and loci involved in reproduction or biotic/abiotic stress are highlighted in dark or light gray, respectively. (*B*) Histogram of −log10(NI) for *C. grandiflora* and (*C*) *A. thaliana*.

receptor kinase that is involved in the self-incompatibility response (SD1 receptor kinases; Samuel et al. 2008).

Discussion

Our analyses indicate that a substantial proportion, about 40%, of amino acid substitutions in *C. grandiflora* has been fixed by positive selection. This contrasts markedly with a lack of evidence for adaptive fixations in *A. thaliana*. Furthermore, a significantly lower proportion of new nonsynonymous mutations appear to be nearly neutral in *C. grandiflora* than in *A. thaliana*, and conversely, a higher proportion of new nonsynonymous mutations are very strongly selected against in *C. grandiflora*. Thus, both positive and purifying selection seem to be more efficient in *C. grandiflora* than in *A. thaliana*.

A difference in the efficacy of selection between *C. grandiflora* and *A. thaliana* is consistent with contrasts in historical demographics and effective population sizes between these species. In particular, *C. grandiflora* is mainly distributed in a former glacial refugial area and appears to have had a stable and large effective population size with little geographical structure (Foxe et al. 2009), whereas *A. thaliana* exhibits lower levels of genetic variation and stronger population differentiation and has undergone a relatively recent and possibly human-mediated spread (Nordborg et al. 2005; François et al. 2008; Platt et al. 2010). The effective population size affects the rate of fixation of adaptive mutations under models where the rate of adaptation is limited by the appearance of new positively selected alleles (Gillespie 1994), and strong population structure can hinder the fixation of such alleles (Whitlock 2003). A higher proportion of slightly deleterious nonsynonymous mutations in *A. thaliana* than in *C. grandiflora* is consistent with theoretical predictions from the nearly neutral theory, according to which species with a smaller effective population size should harbor a greater proportion of sites that are effectively neutral if a substantial proportion of mutations have small selection coefficients (Ohta 1992). Given our estimates, the number of slightly deleterious nonsynonymous substitutions may be three times higher in *A. thaliana* than in *C. grandiflora*.

One possible limitation of our approach is that it does not account for local adaptation; inferences based on the MK test in global samples can only detect positive selection causing species-wide fixation. Although our results clearly suggest a higher global rate of adaptive evolution in *C. grandiflora*, we cannot rule out the possibility of

extensive local adaptation in *A. thaliana*. If local adaptation were more frequent in *A. thaliana* than in *C. grandiflora*, we may be underestimating rates of adaptive substitution and overestimating the slightly deleterious fraction in *A. thaliana*. However, recent analysis of within- and between-population samples in *A. lyrata* implies that weak purifying selection on amino acids predominates in both local and global samples, suggesting that high rates of local positive selection on amino acids may be unlikely (Foxe et al. 2008). Furthermore, rampant genome-wide positive selection on amino acids within populations may be unlikely in *A. thaliana*, given its recent expansion and low average within-population diversity (Bakker et al. 2006; François et al. 2008), although more investigation of the extent of local adaptation is clearly needed.

Several methodological factors can also lead to biased estimates of α. First, the presence of slightly deleterious nonsynonymous mutations can lead to underestimation of α because these mutations contribute to polymorphism but not substantially to divergence. Second, the assumed neutrality of synonymous sites could be violated in the presence of strong and extensive codon bias. Third, population size expansions can result in overestimates of α because slightly deleterious mutations can fix by drift when the effective population size is small (Eyre-Walker 2002). Segregating slightly deleterious mutations are not a major concern in our case because the method we used to estimate α explicitly accounts and corrects for this effect (Eyre-Walker and Keightley 2009). With respect to codon bias, little is known in Capsella, but in Arabidopsis, codon bias is generally weak and primarily restricted to a small subset of very highly expressed genes (Wright et al. 2004), and it is therefore unlikely to have a major effect on our results. Finally, our comparison is likely to be conservative in the presence of population size changes because *A. thaliana* appears to have undergone a recent population expansion (François et al. 2008), and our α estimate for *A. thaliana* (which was negative and nonsignificant) could actually be an overestimate. The DFE-alpha method accounts for a recent population size change when estimating the distribution of deleterious fitness effects; however, as with other current MK-based methods, its α estimates are still sensitive to long-term population size changes. For *C. grandiflora*, the method of Eyre-Walker and Keightley (2009) does not infer a recent change in population size, in agreement with divergence population genetic analyses that suggest a stable and large effective population size in *C. grandiflora* (Foxe et al. 2009). Thus, we do not expect our estimate of the proportion of adaptive nonsynonymous substitutions in *C. grandiflora* to be strongly affected by this source of bias.

When contrasting estimates of α and the DFE between *C. grandiflora* and *A. thaliana*, we have implicitly assumed that the sets of loci sequenced in the two species are equivalent. If the two data sets differ in their levels of constraint, this may affect our estimates and confound conclusions about between-species differences. It has also been shown that if the effective population size varies across the genome and is negatively correlated with the degree of constraint, MK-based methods that sum counts of polymorphism and divergence changes across loci can be biased (Smith and Eyre-Walker 2002). However, after controlling for constraint, our results are qualitatively unchanged, and it therefore seems unlikely that a difference in constraint constitutes a major source of bias. Furthermore, we obtain very similar estimates of α using the maximum likelihood method of Bierne and Eyre-Walker (2004), which is insensitive to biases due to summing counts across loci (Bierne and Eyre-Walker 2004). It therefore seems unlikely that the difference between *C. grandiflora* and *A. thaliana* in the proportion of adaptive fixations is a methodological artifact.

Based on annotation data, three of the four genes showing strong signs of positive selection in *C. grandiflora* are likely to function in plant reproduction. Due to power limitations, we did not conduct a formal test of overrepresentation of specific Gene Ontology categories. Nevertheless, the fact that several of the individual genes that show the strongest evidence for positive selection seem to be involved in reproduction is consistent with findings from other taxa. Indeed, rampant positive selection and rapid divergence have often been found in proteins involved in reproduction and defense, most likely as a result of sexual conflict and sperm/pollen competition in the case of reproductive proteins and coevolutionary arms races in the case of immunity genes (e.g., Fiebig et al. 2004; Schein et al. 2004; Drosophila 12 Genomes Consortium 2007; Kosiol et al. 2008; Clark et al. 2009; Obbard et al. 2009). This result highlights the possibility of variation in selection across the genome and emphasizes the importance of exploring the likelihood of models with different selection parameters (e.g., Obbard et al. 2009).

Here, we have shown that the efficacy of both purifying and positive selection differs between two closely related annual plant species that differ in terms of population size and structure. Our results are parallel to those seen in a recent contrast between the house mouse subspecies *M. musculus castaneus* and humans, which also differ markedly in their effective population size (Halligan et al. 2010). The proportion of adaptive fixations in *C. grandiflora* seems to be of a similar magnitude as that in *P. tremula*, a tree species with a large effective population size and low levels of population structure (Ingvarsson 2010), whereas our results for *A. thaliana* agree with the previous findings of an excess of onsynonymous polymorphism in this species (e.g., Nordborg et al. 2005; Kim et al. 2007) as well as with patterns seen in the close outcrossing relative *A. lyrata* which is also strongly structured (Foxe et al. 2008). Our results are consistent with theoretical predictions on the efficacy of both positive and purifying selection, and this study constitutes one of the first to find convincing evidence for a high proportion of adaptive fixations in an annual plant. Future studies should aim to explore the effects of population structure and local adaptation on adaptive evolution and our ability to detect it.

Supplementary Material

Supplementary text and tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Adam Eyre-Walker for sharing unpublished work and for helpful discussion, Aneil Agrawal for helpful discussion, and Arne Strid and Kit Tan for helpful advice on *C. grandiflora* sampling. This research was supported by an Early Researcher Award (Government of Ontario) and a Natural Sciences and Engineering Research Council of Canada Discovery Grant to S.I.W.

References

Akashi H. 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* 144:1297–1307.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437:1149–1152.

Axelsson E, Ellegren H. 2009. Quantification of adaptive evolution of genes expressed in avian brain and the population size effect on the efficacy of selection. *Mol Biol Evol*. 26:1073–1079.

Bachtrog D. 2008. Similar rates of protein adaptation in *Drosophila miranda* and *D. melanogaster*, two species with different current effective population sizes. *BMC Evol Biol*. 8:334.

Bachtrog D, Andolfatto P. 2006. Selection, recombination and demographic history in *Drosophila miranda*. *Genetics* 174: 2045–2059.

Bakker EG, Stahl EA, Toomajian C, Nordborg M, Kreitman M, Bergelson J. 2006. Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. *Mol Ecol*. 15:1405–1418.

Begun DJ, Holloway AK, Stevens K, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*. 5:e310.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol*. 57:289–300.

Betancourt AJ, Welch JJ, Charlesworth B. 2009. Reduced effectiveness of selection caused by a lack of recombination. *Curr Biol*. 19:655–660.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol Biol Evol*. 21:1350–1360.

Boivin K, Acarkan A, Mbulu RS, Clarenz O, Schmidt R. 2004. The Arabidopsis genome sequence as a tool for genome analysis in Brassicaceae. A comparison of the Arabidopsis and *Capsella rubella* genomes. *Plant Physiol*. 135:735–744.

Boyko AR, Williamson SH, Indap AR, et al. (14 co-authors). 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*. 4:e1000083.

Bullaughey K, Przeworski M, Coop G. 2009. No effect of recombination on the efficacy of natural selection in primates. *Genome Res*. 18:544–554.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in *Arabidopsis*. *Nature* 416:531–534.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res*. 63:213–227.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol*. 23:1348–1356.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

Clark NL, Gasper J, Sekino M, Springer SA, Aquadro CF, Swanson WJ. 2009. Coevolution of interacting fertilization proteins. *PLoS Genet*. 5:e1000570.

Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang SP, Fay JC. 2008. A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet*. 4:e1000183.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450:203–218.

Ellegren H. 2009. A selection model of molecular evolution incorporating the effective population size. *Evolution* 63:301–305.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 8:186–194.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 8:175–185.

Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics* 162:2017–2024.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol*. 21:569–575.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol*. 26:2097–2108.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.

Fiebig A, Kimport R, Preuss D. 2004. Comparisons of pollen coat genes across Brassicaceae species reveal rapid evolution by repeat expansion and diversification. *Proc Natl Acad Sci U S A*. 101:3286–3291.

Foxe JP, Dar VU, Zheng H, Nordborg M, Gaut BS, Wright SI. 2008. Selection on amino acid substitutions in *Arabidopsis*. *Mol Biol Evol*. 25:1375–1383.

Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. 2009. Recent speciation associated with the evolution of selfing in *Capsella*. *Proc Natl Acad Sci U S A*. 106:5241–5245.

François O, Blum MG, Jakobsson M, Rosenberg NA. 2008. Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet*. 4:e1000075.

Gillespie JH. 1994. The causes of molecular evolution. Oxford: Oxford University Press.

Gillespie JH. 2000. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* 155:909–919.

Gillespie JH. 2001. Is the population size of a species relevant to its evolution? *Evolution* 55:2161–2169.

Gillespie JH. 2004. Why k = 4 *Nus* is silly. In: Singh RS, Uyenoyama MK, editors. The evolution of population biology. Cambridge: Cambridge University Press. p. 178–192.

Goldman N, Yang ZH. 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. *Mol Biol Evol*. 11:725–736.

Haddrill PR, Halligan DL, Tomaras D, Charleworth B. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol*. 8:R18.

Hahn MW. 2008. Toward a selection theory of molecular evolution. *Evolution* 62:255–265.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet*. 6:e1000825.

Ingvarsson PK. 2010. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. *Mol Biol Evol*. 27:650–660.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and

population demography based on nucleotide polymorphism frequencies. *Genetics* 177:2251–2261.
Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet*. 39:1151–1155.
Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–626.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
Kliman RM, Hey J. 1993. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol Biol Evol*. 10:1239–1258.
Koch M, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol Biol Evol*. 17:1483–1498.
Kosiol C, Vinar T. 2008. da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet*. 4:e1000144.
Li YF, Costello JC, Holloway AK, Hahn MW. 2008. "Reverse ecology" and the power of population genomics. *Evolution* 62:2984–2994.
Lindblad-Toh K, Wade CM, Mikkelsen TS, et al. (47 co-authors). 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803–819.
Liti G, Carter DM, Moses AM, et al. (26 co-authors). 2009. Population genomics of domestic and wild yeasts. *Nature* 458:337–341.
Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.
Lysak MA, Berr A, Pecinka A, Schmidt R, McBreen K, Schubert I. 2006. Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proc Natl Acad Sci U S A*. 103:5224–5229.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in Drosophila. *Nature* 351:652–654.
Muller MH, Leppälä J, Savolainen O. 2008. Genome-wide effects of postglacial colonization in *Arabidopsis lyrata*. *Heredity* 100:47–58.
Nickerson DA, Tobe VO, Taylor SL. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res*. 25:2745–2751.
Nordborg M, Hu TT, Ishino Y, et al. (24 co-authors). 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol*. 7:e196.
Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 302:205–217.
Obbard DJ, Welch JJ, Kim KW, Jiggins FM. 2009. Quantifying adaptive evolution in the Drosophila immune system. *PLoS Genet*. 5:e1000698.
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98.
Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst*. 23:263–286.
Ostrowski MF, David J, Santoni S, et al. (11 co-authors). 2006. Evidence for a large-scale population structure among accessions of *Arabidopsis thaliana*: possible causes and consequences for the distribution of linkage disequilibrium. *Mol Ecol*. 15: 1507–1517.
Platt A, Horton M, Huang YS, et al. (23 co-authors). 2010. The scale of population structure in Arabidopsis thaliana. *PLoS Genet*. 6:e1000843.
Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc Natl Acad Sci U S A*. 104:13390–13395.
Presgraves DC. 2005. Recombination enhances protein adaptation in *Drosophila melanogaster*. *Curr Biol*. 15:1651–1656.
Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol Biol Evol*. 13:735–748.
Rieder MJ, Taylor SL, Tobe VO, Nickerson DA. 1998. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res*. 26:967–973.
Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, Gos G, Charlesworth D, Gaut BS. 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS One*. 6:e2411.
Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*. 132:365–386.
Samuel MA, Mudgil Y, Salt JN, Delmas F, Ramachandran S, Chilelli A, Goring DR. 2008. Interactions between the *S*-domain receptor kinases and AtPUB-ARM E3 ubiquitin ligases suggest a conserved signaling pathway in Arabidopsis. *Plant Physiol*. 147:2084–2095.
Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL. 2003. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J Mol Evol*. 57(1 Suppl): S154–S164.
Schein M, Yang Z, Mitchell-Olds T, Schmid KJ. 2004. Rapid evolution of a pollen-specific oleosin-like gene family from *Arabidopsis thaliana* and closely related species. *Mol Biol Evol*. 21:659–669.
Schmid KJ, Törjék O, Meyer R, Schmuths H, Hoffmann MH, Altmann T. 2006. Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers. *Theor Appl Genet*. 112:1104–1114.
Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the Drosophila genome? *PLoS Genet*. 5:e1000495.
Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415:1022–1024.
Welch JJ. 2006. Estimating the genomewide rate of adaptive protein evolution in Drosophila. *Genetics* 173:821–837.
Whitlock MC. 2003. Fixation probability and time in subdivided populations. *Genetics* 164:767–779.
Woolfit M. 2009. Effective population size and the rate and pattern of nucleotide substitutions. *Biol Lett*. 5:417–420.
Woolfit M, Bromham L. 2005. Population size and molecular evolution on islands. *Proc Biol Sci*. 272:2277–2282.
Wright SI, Andolfatto P. 2008. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annu Rev Ecol Evol Syst*. 39:193–213.
Wright SI, Yau CBK, Looseley M, Meyers B. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Mol Biol Evol*. 21:1719–1726.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.
Zhang LQ, Li WH. 2005. Human SNPs reveal no evidence of frequent positive selection. *Mol Biol Evol*. 22:2504–2507.