Genome Wide Analyses Reveal Little Evidence for Adaptive Evolution in Many Plant Species

Toni I. Gossmann,1 Bao-Hua Song,2 Aaron J. Windsor,2 Thomas Mitchell-Olds,2 Christopher J. Dixon,3 Maxim V. Kapralov,3 Dmitry A. Filatov,3 and Adam Eyre-Walker∗,1

1Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Brighton, United Kingdom
2Institute for Genome Sciences and Policy, Department of Biology, Duke University
3Department of Plant Sciences, University of Oxford, Oxford, United Kingdom

*Corresponding author: E-mail: a.c.eyre-walker@sussex.ac.uk.
Associate editor: Naoki Takebayashi

Abstract

The relative contribution of advantageous and neutral mutations to the evolutionary process is a central problem in evolutionary biology. Current estimates suggest that whereas Drosophila, mice, and bacteria have undergone extensive adaptive evolution, hominids show little or no evidence of adaptive evolution in protein-coding sequences. This may be a consequence of differences in effective population size. To study the matter further, we have investigated whether plants show evidence of adaptive evolution using an extension of the McDonald–Kreitman test that explicitly models slightly deleterious mutations by estimating the distribution of fitness effects of new mutations. We apply this method to data from nine pairs of species. Altogether more than 2,400 loci with an average length of ≈280 nucleotides were analyzed. We observe very similar results in all species; we find little evidence of adaptive amino acid substitution in any comparison except sunflowers. This may be because many plant species have modest effective population sizes.

Key words: adaptive evolution, distribution of fitness effects, plants, effective population size, McDonald-Kreitman test.

Introduction

The contribution of adaptive evolution relative to genetic drift is a fundamental problem in molecular evolution (Kimura 1983; Gillespie 1991). Several methods to estimate the fraction of adaptive substitutions, α, have been developed based on the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991) that contrast polymorphism and divergence between selectively and neutrally evolving sites (Charlesworth 1994; Fay et al. 2001; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Welch 2006; Boyko et al. 2008; Eyre-Walker and Keightley 2009). These methods have been applied to a variety of species. Estimates in Drosophila (Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Welch 2006; Bachtrog 2008) and rodents (Halligan et al. 2010) suggest that ≈50% of all amino acid substitutions have been fixed as a consequence of adaptive evolution and for microorganisms estimates may be even higher (Charlesworth and Eyre-Walker 2006; Liti et al. 2009). However, although analyses of DNA sequence diversity show signs of some adaptive evolution (Fay et al. 2001; Zhang and Li 2005), overall the level of adaptive evolution in hominids appears to be very low (Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005; Boyko et al. 2008; Eyre-Walker and Keightley 2009). The contrast between hominids and other animals has led to the suggestion that effective population size may be an important determinant of the rate of adaptive evolution because hominids typically have low effective population sizes in contrast to rodents, insects, and bacteria (Eyre-Walker et al. 2002; Fraser et al. 2007; Halligan et al. 2010). However, some caution should be exercised because the level of adaptive evolution is typically measured as the proportion of substitutions that are adaptive and this depends both on the numbers of substitutions that are effectively neutral and the number that are advantageous. It has been shown that the number of effectively neutral mutations is negatively correlated to the effective population size in many species (Woolfit and Bromham 2003, 2005; Popadin et al. 2007; Moran et al. 2008; Piganeau and Eyre-Walker 2009). Hence, the correlation between proportion of substitutions that are adaptive and Ne may be a consequence of the correlation between the proportion of effectively neutral mutations and Ne and may not reflect any change in the absolute rate of adaptive evolution.

The rate of adaptive evolution has also been studied in plants. On a genome-wide scale, previous studies in Arabidopsis thaliana have shown little evidence for adaptive evolution (Bustamante et al. 2002; Barrier et al. 2003; Schmid et al. 2005). This was attributed to the high frequency of inbreeding in A. thaliana and the reduction in effective population size that this caused. However, the outcrossing species A. lyrata (Barnaud et al. 2008; Foxe et al. 2008), the partially outcrossing cultivated tropical grass Sorghum bicolor (Hamblin et al. 2006), and the mainly outcrossing Zea species (Bijlsma et al. 1986; Ross-Ibarra et al. 2009) also show little evidence of positive selection. Instead, all these species show evidence of slightly deleterious
mutations (SDMs) segregating. In contrast, to the pattern in other plant species, Strasburg et al. (2009), Ingvarsson (2010), and Slotte et al. (2010) have recently estimated that ≈75%, ≈30%, and ≈40% of fixed amino acid differences were driven by adaptive substitutions in sunflowers, aspen trees, and some Brassicaceae species, respectively.

There is evidence that SDMs contribute to variation in many populations (Akashi 1999; Cargill et al. 1999; Fay et al. 2002; Hughes 2005; Charlesworth and Eyre-Walker 2006). They are subject to weak negative selection (N_e ≈ 1), segregate at lower frequencies than neutral mutations, and contribute proportionally more to polymorphism than to divergence when compared with neutral mutations. The presence of SDMs in the analyses of Arabidopsis sp., S. bicolor, and Zea sp. may explain why there is so little apparent positive selection because SDMs are expected to bias the estimate of adaptive evolution downward if population sizes are either stationary or contracting—they can lead to an underestimate if population sizes have expanded (McDonald and Kreitman 1991; Eyre-Walker 2002). Here, we apply a method to estimate the proportion of adaptive substitutions that controls for the effects of SDMs by estimating the distribution of fitness effects (DFE) of new mutations (Eyre-Walker and Keightley 2009). It estimates the DFE from the polymorphic data and predicts the expected number of substitutions originating from neutral and SDMs. If the observed number of substitutions is greater than the expectation inferred from the DFE, it can be attributed to advantageous substitutions, yielding an estimate of α. We present a new parameterization of this method that allows us to estimate the rate of adaptive substitution relative to the rate of synonymous substitution, thereby allowing us to explore whether the rate of adaptive evolution depends on N_e independent of the effects of N_e on the number of effectively neutral substitutions.

We also apply the method of Fay et al. (2001), who suggested controlling for SDMs by removing low-frequency mutations from the analysis. We apply these methods to 11 data sets covering the divergence between nine pairs of largely independent species. We find little evidence of adaptive amino acid substitution in any comparison except sunflowers. Moreover, we estimate the effective population size and find that some of the plant species analyzed have large effective population sizes, suggesting that other factors may be more important than population size in determining the rate of adaptive evolution.

Materials and Methods

Sequence Data


The annotated protein-coding genome of A. thaliana was obtained from TAIR (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release). The annotated P. trichocarpa and S. bicolor genomes were obtained from JGI (http://genome.jpi-psf.org/PoPt1_1, Tuskan et al. 2006 and http://genome.jgi-psf.org/Sorbi1, Paterson et al. 2009). Predicted coding sequences of Z. mays were obtained from http://magi.plantgenomics.iastate.edu/downloadAll.html (Fu et al. 2005) and http://ftp.maizesequence.org/release-3b.50/sequences/. The rice genome was downloaded from ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/O_sativa/annotation_dbs/pseudomolecules/version_6.0/all_dir/. We analyzed the polymorphism data for each of the 11 plant species. Additionally, for some data sets, information about populations were available; five populations of A. thaliana (Nordborg et al. 2005), six populations of A. lyrata (Ross-Ibarra et al. 2008), three domesticated subspecies of O. sativa, one domesticated subspecies (Z. mays mays) and one wild subspecies (Z. mays parviglumis) of Z. mays, two populations of B. stricta, three populations of S. globosa, and three populations of P. balsamifera (Keller et al. 2010).

Preparation of the Data

Sequences were aligned using ClustalW using default parameter values (Thompson et al. 1994). Coding regions were assigned using protein-coding genomic data or, if given, taken from the GenBank input files. The outgroup was assigned using the best Blast (Altschul et al. 1990) hit or, if given, taken from the polymorphism data set. We only used polymorphism data for which we could assign an outgroup sequence with the exception of A. thaliana for which no outgroup data are currently available (the A. lyrata sequence is nearly complete but unpublished). For all analyses, the number of synonymous sites served as the neutral standard. For computational reasons, the method to estimate the DFE needs all sites to have been sampled in the same number of chromosomes for each species; because some loci had been sampled in more individuals than others and other loci had missing data, we reduced the data set to a common number of chromosomes by randomly sampling the polymorphisms at each site without replacement. The numbers of synonymous and nonsynonymous sites and substitutions were computed using the F3x4 model implemented in PAML (Yang 1997) in which codon frequencies are estimated from the nucleotide frequencies at the three codon positions. The proportion of sites
Table 1. Summary of data sets used for the analyses. Number of coding sites (Sites) and number of segregating sites (Segregating) are from the polymorphic data. Average divergence between the species pairs at silent sites (\(d_s\)) and nucleotide diversity for synonymous sites (\(\pi_s\)). Sample size (Alleles) is constant for each species for computational reasons.

<table>
<thead>
<tr>
<th>Plant family</th>
<th>Ingroup</th>
<th>Population</th>
<th>Outgroup</th>
<th>Alleles</th>
<th>Loci</th>
<th>Sites</th>
<th>Segregating</th>
<th>(d_s)</th>
<th>(\pi_s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brassicaceae</td>
<td>Arabidopsis lyrata</td>
<td>A. thaliana</td>
<td>None</td>
<td>24</td>
<td>72</td>
<td>22,739</td>
<td>610</td>
<td>0.15</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>United States</td>
<td>14</td>
<td>51</td>
<td>14,263</td>
<td>191</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Russia</td>
<td>14</td>
<td>47</td>
<td>16,401</td>
<td>74</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Iceland</td>
<td>16</td>
<td>48</td>
<td>17,332</td>
<td>93</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Canada</td>
<td>16</td>
<td>51</td>
<td>19,458</td>
<td>103</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sweden</td>
<td>8</td>
<td>62</td>
<td>20,609</td>
<td>123</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A. thaliana</td>
<td></td>
<td>None</td>
<td>24</td>
<td>919</td>
<td>260,737</td>
<td>3,401</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>United States</td>
<td>8</td>
<td>664</td>
<td>167,990</td>
<td>888</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>North Sweden</td>
<td>8</td>
<td>742</td>
<td>191,492</td>
<td>902</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>South Sweden</td>
<td>8</td>
<td>771</td>
<td>202,016</td>
<td>1,480</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Central Europe</td>
<td>8</td>
<td>789</td>
<td>211,398</td>
<td>1,430</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>England</td>
<td>8</td>
<td>816</td>
<td>222,405</td>
<td>1,252</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boechera stricta</td>
<td>A. thaliana</td>
<td></td>
<td>North</td>
<td>16</td>
<td>130</td>
<td>38,589</td>
<td>111</td>
<td>0.21</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>South</td>
<td>10</td>
<td>130</td>
<td>39,800</td>
<td>123</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poaceae</td>
<td>Sorghum bicolor(^a)</td>
<td>S. propinquum</td>
<td>14</td>
<td>135</td>
<td>30,241</td>
<td>129</td>
<td>0.03</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oryza rufipogon(^a)</td>
<td>S. bicolor(^a)</td>
<td>18</td>
<td>73</td>
<td>18,176</td>
<td>121</td>
<td>0.78</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>O. sativa(^a)</td>
<td>S. bicolor(^a)</td>
<td>18</td>
<td>73</td>
<td>18,197</td>
<td>68</td>
<td>0.78</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temperate(^b)</td>
<td></td>
<td>18</td>
<td>73</td>
<td>18,163</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tropical(^b)</td>
<td></td>
<td>18</td>
<td>71</td>
<td>17,618</td>
<td>85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Indica(^a)</td>
<td></td>
<td>18</td>
<td>73</td>
<td>18,197</td>
<td>68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zea mays(^a)</td>
<td>S. bicolor(^a)</td>
<td>8</td>
<td>478</td>
<td>74,155</td>
<td>939</td>
<td>0.015</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>spp. mays(^a)</td>
<td></td>
<td>10</td>
<td>378</td>
<td>59,944</td>
<td>1,241</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>spp. parviglumis</td>
<td></td>
<td>8</td>
<td>478</td>
<td>74,155</td>
<td>939</td>
<td>0.015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salicaceae</td>
<td>Populus tremula</td>
<td>P. trichocarpa</td>
<td>24</td>
<td>69</td>
<td>32,255</td>
<td>578</td>
<td>0.05</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P. balsamifera</td>
<td>P. trichocarpa</td>
<td>24</td>
<td>508</td>
<td>174,855</td>
<td>1,446</td>
<td>0.01</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Central</td>
<td>P. trichocarpa</td>
<td>14</td>
<td>501</td>
<td>172,046</td>
<td>1,108</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>East</td>
<td>P. trichocarpa</td>
<td>6</td>
<td>519</td>
<td>178,473</td>
<td>643</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>North</td>
<td>P. trichocarpa</td>
<td>10</td>
<td>508</td>
<td>178,743</td>
<td>919</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asteraceae</td>
<td>Helianthus annuus</td>
<td>H. petiolaris</td>
<td>14</td>
<td>54</td>
<td>7,183</td>
<td>259</td>
<td>0.05</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H. petiolaris</td>
<td>H. annuus</td>
<td>10</td>
<td>54</td>
<td>7,231</td>
<td>175</td>
<td>0.05</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>Caryophyllaceae</td>
<td>Schiedea globosa</td>
<td>S. adamantis</td>
<td>24</td>
<td>23</td>
<td>8,030</td>
<td>189</td>
<td>0.02</td>
<td>0.014</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Includes domesticated species.

Estimated by PAML were also used to compute the numbers of synonymous and nonsynonymous sites for the polymorphism data. For each species or population, data were summed across all genes, although analyses run on the unsummed data gave similar results (results not presented). Statistics concerning numbers of loci, numbers of sites as well as polymorphic sites are shown in table 1.

The DFE of new mutations and the proportion of adaptive substitutions (\(\alpha\)) were estimated using the method II of Eyre-Walker and Keightley (2009), which accounts for the segregation and fixation of SDMs. To estimate the rate of adaptive evolution in a manner that is independent of the number of effectively neutral substitutions, we reparameterize the method as follows. In the original formulation, the expected numbers of synonymous, \(D_s\), and nonsynonymous, \(D_n\), substitutions were as follows:

\[
D_s = 2utl_s
\]
\[
D_n = 2utl_n \int D(S)P(S) dS \bigg/ 1 - \alpha
\]

where \(L_s\) and \(L_n\) are the numbers of synonymous and nonsynonymous sites, \(u\) is the nucleotide mutation rate, and \(t\) is the time of divergence. \(D(S)\) is the distribution of \(S\), the strength of selection (multiplied by four times the effective population size), \(P(S)\) is \(Ne\) times the probability of fixation, and \(\alpha\) is the proportion of substitutions that are adaptive. We can reparameterize the expression for \(D_n\) as follows:

\[
D_n = 2utl_n \left( \int D(S)P(S) dS + \omega_s \right)
\]

where \(\omega_s\) is the ratio of the rate of adaptive nonsynonymous substitution to synonymous substitution. This is similar to the parameterizations given by Bierne and Eyre-Walker (2004) and Obbard et al. (2009).

Levels of polymorphism were quantified using Watterson’s estimator of \(4N_e \mu\) (\(\theta_W\)) (Watterson 1975) and nucleotide diversity, \(\pi\) (Tajima 1983) for synonymous sites. Our estimates of \(\pi\) and \(\theta_W\) differ slightly from previous published estimates due to the fact that we have excluded some data and alleles from our analysis. We estimated the
effective population size using the level of synonymous site diversity and dividing this by an estimate of the mutation rate per generation. Estimates of the synonymous mutation rate per year and generation times were taken from the literature: For species A. thaliana, A. lyrata, and B. stricta, we assumed, as other have done, the mutation rate of \( \mu = 1.5 \times 10^{-8} \) per site per year as estimated in the Brassicaceae (Koch et al. 2000) and a generation time of 1 year for the annual A. thaliana (Koornneef et al. 2004) and 2 years for the perennials B. stricta (Dobes et al. 2004) and A. lyrata (Ross-Ibarra et al. 2008). For the genera Oryza, Sorghum, Helianthus, and Zea, we assumed the mutation rate to be \( \mu = 1.0 \times 10^{-8} \) per site per year and a generation time of 15 years (Koch et al. 2000; Tuskan et al. 2006; Ingvarsson 2008b); and for Schiedea, we assumed a mutation rate of \( \mu = 1.9 \times 10^{-8} \) per site per year and a generation time of 5 years (Filatov and Burke 2004; Wallace et al. 2009).

Simulations
To investigate the effects of population structure on our estimates of adaptive evolution, we performed forward population genetic simulations using SFS_CODE (Hernandez 2008). We investigated two scenarios. In the first, we have migration between two populations that divided sometime in the past and have continued to exchange migrants, but we only sample from one of the populations; and in the second, we sample from both populations equally. To simulate these two scenarios, we divided an ancestral population into three equal sized populations of 500 individuals 20,000 generations in the past; with a mutation rate of \( 4N_e \mu = 5 \times 10^{-6} \) per site per generation, this would give an expected divergence at neutral sites of 20%, which is similar to that seen in some of our data sets. We allowed one of the populations to evolve independently; this was the outgroup. The other two populations exchanged migrants at various rates of \( 4N_e m \) ranging from 0.01 to 10. All populations were subjected to mutation at \( 4N_e \mu = 0.01 \) at 100 uncoupled loci with a length of 1,002 nucleotides each. Each simulation was repeated 100 times. Because SFS_CODE and convertSFS_CODE do not allow population mixing or the sampling of two populations, we allowed the two populations to exchange migrants at a very high rate (\( 4N_e m = 250 \)) for the last five generations of the simulation (supplementary fig. S1, Supplementary Material online).

Results
Data
To estimate the rate of adaptive substitution in plant protein-coding sequences, we compiled polymorphism data from 11 species and aligned 10 of these to outgroups to analyze the divergence between 9 pairs of species. The data sets range dramatically in size from 23 to 919 loci per species and 6 to 24 sequences per gene (table 1). Note that we treat O. sativa and O. rufipogon as the same species in our analysis, O. sativa being the domesticate of O. rufipogon. For most genera, we have access to polymorphism data from wild populations (A. lyrata, A. thaliana, B. stricta, P. tremula, P. balsamifera, O. rufipogon, H. annuus, H. petiolaris, and S. globosa); for one species from both wild and domesticated populations (Z. mays) and for the species S. bicolor and O. sativa, we only have data from domesticated populations. We use the polymorphism data of 11 species to estimate the DFE of new mutations and use this to further estimate the proportion and relative rates of adaptive substitutions between species. Our comparisons for Boechera, P. balsamifera, Schiedea, and Oryza are the first large-scale investigations of adaptive evolution in these groups. The available outgroup data only allow us to estimate the rate of adaptive substitution between nine pairs of species. We include the polymorphism data from A. thaliana as a comparison for the results obtained from A. lyrata; note also that the divergences between the Brassicaceae species and the Poaceae species are not entirely independent because the same outgroup (A. thaliana and S. bicolor, respectively) is used for different species. Based on average \( d_e \) values (table 1), A. thaliana and A. lyrata share \( \approx 30\% \) of their divergence with B. stricta and O. rufipogon and Z. mays share \( \approx 22\% \) of their divergence with S. bicolor.

Distribution of Effects of New Mutations
Species-wide (e.g., ignoring population information within species) estimates of the DFE show a remarkably consistent picture among the investigated species (fig. 1) with the exception of B. stricta, P. balsamifera, and S. globosa. For all species, the largest proportion of mutations have \( N_e s > 100 \) and hence are strongly deleterious and for most species less than 25% of amino acid–changing mutations behave as effectively neutral (\( 0 < N_e s < 1 \)). In contrast to this, B. stricta, P. balsamifera, and S. globosa show an excess of neutral mutations (\( > 25\% \)) and a decrease of strongly deleterious mutations (\( < 55\% \)). The confidence intervals vary between species due to the varying sample sizes. Our estimates of the DFE are not greatly affected by the removal of singletons, suggesting that sequencing errors are not an issue (supplementary fig. S2, Supplementary Material online). We additionally estimated the DFE from individual wild populations of Arabidopsis, Populus, Boechera, and Schiedea. For all populations of A. thaliana and A. lyrata (supplementary fig. S3, Supplementary Material online), the distributions between and within these species are fairly similar to each other, although the confidence intervals for A. lyrata are larger because of the smaller data set. In all the investigated Arabidopsis populations more than 50% of the mutations are strongly deleterious (\( N_e s > 100 \)) and more than 18% behave as effectively neutral (\( 0 < N_e s < 1 \)). In contrast, populations of B. stricta and S. globosa show differences in their DFE, whereas populations of P. balsamifera are remarkably similar to each other (supplementary fig. S4, Supplementary Material online). Additionally, we find some differences between the DFE estimated from the wild and domesticated populations of rice (supplementary
The DFE in 11 plant species. Estimates are given for the proportion of mutations in four different $N_e s$ ranges and SEs. For *Z. mays*, the results are shown for *Z. mays* spp. *parviglumis*. Fig. S5, Supplementary Material online). Both domesticated varieties of *O. japonica* show a higher proportion of effectively neutral mutations than *O. rufipogon*, which may reflect the lower effective population of the domesticated varieties; the DFE for the domesticated populations of *O. japonica* are more like that found in *B. stricta*, *P. balsamifera*, and *S. globosa*. In contrast, *O. indica* shows a very different DFE both to other *Oryza* species and to all other plant species, with many mutations inferred to be slightly or mildly deleterious ($1 < N_e s < 100$). However, the confidence intervals are large. The DFE inferred from domesticated *Z. mays* is similar to the wild population. These results suggest that the method to estimate the DFE copes well with the different demographic histories these populations have experienced, except possibly during some domestication events.

Adaptive Substitutions

Surprisingly, we find little evidence of adaptive evolution in any of the species we have considered (supplementary fig. 2, Supplementary Material online), except in one sunflower comparison, despite the fact that we have controlled for the effects of SDMs, which tend to bias estimates of adaptive evolution downward. The estimates of $\alpha$ are positive in the comparisons involving *Populus* species and *S. globosa*, but these are less than 15% and nonsignificant; the estimate of $\alpha$ using the polymorphism data from *H. petiolaris* is significantly positive. All other comparisons are negative, with the estimate using *B. stricta* being significantly negative, but not after correction for multiple tests. We obtain similar results for estimates of $\alpha$ for 23 subpopulations of seven species (supplementary table S1, Supplementary Material online) with all positive estimates being nonsignificant.

The mean estimate of $\alpha$ across all ten species-wide comparisons is $-0.18$ (standard error [SE] = 0.47). This is not significantly different from zero, suggesting that plants generally go through very little adaptive evolution in their protein-coding sequences; however, it should be appreciated that we have overestimated the SE because of the non-independence between some of the data sets. Similar results are obtained if we estimate $\omega_a$, the rate of adaptive nonsynonymous substitution relative to the rate of synonymous substitution (fig. 2b). Although the method, which we have used to control for SDMs in the estimation of $\alpha$, appears to be robust (Eyre-Walker and Keightley 2009), we also estimated the proportion of adaptive substitutions using a simple method based on the MK test, which just uses the numbers of synonymous and nonsynonymous polymorphisms and substitution summed across genes (Fay et al. 2001) rather than the full site frequency spectrum. Estimates of $\alpha$ are negative in all species except *H. petiolaris* and *S. globosa* when we consider all polymorphisms (supplementary table S2, Supplementary Material online). Furthermore, the $\alpha$ estimates for the *Populus* species become positive if we remove polymorphisms below 15% to control for the effects of SDMs (supplementary table S2, Supplementary Material online). However, all positive estimates based on the method of Fay et al. (2001) are nonsignificant.

Discussion

We have estimated the proportion of nonsynonymous substitutions that are due to positive adaptive evolution in a range of plant species using an extension of the MK test. Surprisingly, we find little evidence of adaptive evolution in the plant species we have investigated; in four of the comparisons, the estimates of $\alpha$, the proportion of substitutions driven by positive selection, are greater than zero, but for three of them, the estimates are less than 15% and not significantly different from zero. Similar results are obtained when
we estimate \( \omega_a \), the rate of adaptive amino acid substitution relative to the rate of synonymous substitution. The low estimated rate of adaptive evolution is consistent with previous estimates in *Arabidopsis* sp., *Sorghum*, and *Zea* (Bustamante et al. 2002; Hamblin et al. 2006; Foxe et al. 2008; Ross-Ibarra et al. 2009); however, in these previous analyses, it was not possible to exclude the possibility that the low estimates of adaptive evolution were simply a consequence of the segregation of SDMs, which tend to bias estimates of adaptive evolution downward (Fay et al. 2001; Charlesworth and Eyre-Walker 2008). In our analyses, we have used a method that explicitly models SDMs and takes these into account. Our estimate of \( \alpha \) in the *Helianthus* species is also broadly consistent with the estimate of 75% obtained by Strasburg et al. (2009). Using the same data, we find significant evidence of adaptive evolution when we consider the divergence between *H. annuus* and *H. petiolaris* using the polymorphism data from *H. petiolaris* and a nonsignificant estimate of \( \alpha \) when we use the polymorphism data from *H. annuus*. In both cases, our estimate of \( \alpha \) is lower than that obtained by Strasburg et al. (2009), but there are several methodological differences between the analyses so this is perhaps not surprising and the confidence intervals on all estimates are large. Our estimate of \( \alpha \) is also lower than the \( \approx 30\% \) estimate obtained by Ingvarsson (2010) for the divergence between *P. tremula* and *P. trichocarpa* using a similar data set but the method of Bierne and Eyre-Walker (2004); this seems to be largely due to the fact that we had to reduce those genes for which Ingvarsson had 38 haploid genomes to 24 so that all genes would have the same number of sampled chromosomes—this is a limitation of the Eyre-Walker and Keightley (2009) method used here. If we apply the method of Bierne and Eyre-Walker (2004) to our reduced data set removing polymorphisms below a frequency of 15%, as Ingvarsson did, we estimate \( \alpha \) to be 0.1303 (95% confidence intervals: \(-0.2447\) to 0.3928), which is similar to the estimate obtained using the Eyre-Walker and Keightley (2009) method.

There are several reasons why the estimated rate of adaptive evolution might be low in plants. First, limited evidence in animals suggests that the proportion of adaptive substitutions may depend on the effective population size; primates appear to have relatively low rates of adaptive evolution and also low effective population sizes, whereas mice, Drosophila, and enteric bacteria have high rates of adaptive evolution (Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004; Welch 2006; Charlesworth and Eyre-Walker 2006; Bachtrog 2008) and also large effective population sizes (Andolfatto 2001; Charlesworth and Eyre-Walker 2006; Piganeau and Eyre-Walker 2009). There is some evidence of this relationship in plants as well, with species such as *Capsella grandiflora* and *H. petiolaris*, which have large effective population sizes (Strasburg and Rieseberg 2008; Foxe et al. 2009), showing high rates of adaptive evolution (Strasburg et al. 2009; Slotte et al. 2010). This relationship between population size and the rate of adaptive evolution is expected if the rate of adaptive evolution is limited by the supply of advantageous mutations because species with large populations produce advantageous mutations at a higher rate, and selection is also more effective on those that are weakly selected. To investigate whether plants have low effective population sizes like primates, we estimated the effective population size in our plant species using estimates of \( \theta_W \) from the loci involved in our analyses and estimates of the mutation rate per generation and generation time from the literature (see Materials and Methods). These estimates should be treated with caution because the mutation rate has not been directly estimated in plants, and actual demographic processes presumably deviate from the assumptions of these models.

The effective population sizes vary quite substantially between our species from *B. stricta* with an effective population size of just more than 25,000 to *H. annuus* with an \( N_e \) of more than 800,000, but most species have relatively modest population sizes of tens of thousands to just more than 100,000 (fig. 3). Our estimates agree closely with those previous estimates for *O. rufipogon*, but our estimate is lower than the previous estimate from *P. tremula*; Ingvarsson (2008b) used the mode \( \theta_W \) value and estimated \( N_e \) to be 118,000,
which is roughly two times larger than our estimate. As expected, our estimates for Helianthus species are midway between the estimates of the ancestral and derived population sizes for these species, which have undergone a recent population size expansion (Strasburg and Rieseberg 2008). The population-specific estimates of $N_e$ (supplementary table S1, Supplementary Material online) are similar to the overall estimate of $N_e$ in A. thaliana, B. stricta, P. balsamifera, and S. globosa and the estimates for the domesticated populations of rice and maize are substantially lower than for their wild relatives. In contrast to the other wild populations, in A. lyrata, the overall $N_e$ is considerably greater than the population-specific estimates except the German population, reflecting the high level of population structure in this species (Ross-Ibarra et al. 2008). It therefore seems that most of these plants generally have higher effective population sizes than humans and chimpanzees, which have $N_e$ values of 10,000–30,000 (Eyre-Walker et al. 2002). Zea mays has an effective population size of 500,000, which is of similar magnitude to that of rodents (Mus domesticus = 160,000, Eyre-Walker et al. 2002 and M. castaneus = 500,000 Halligan et al. 2010), species for which $\alpha$ is estimated to be 60%, and yet, it shows no evidence of adaptive evolution. Only the sunflower species show some evidence of adaptive evolution in our analysis and these do indeed have the largest effective population sizes; also C. grandiflora shows evidence of adaptive evolution (Slotte et al. 2010) and a large effective population size (Foxxe et al. 2009). These results therefore suggest that the rate of adaptive evolution might be correlated to effective population size but that many plants do not undergo high rates because they have relatively small population sizes.

Although there appears to be no relationship between the current estimate of the effective population size and the rate of adaptive evolution in plants, the low estimates of $\alpha$ and $\omega_s$ could be an artifact of a contracting population. If there are SDMs segregating and current population sizes are smaller than during the divergence between the species being considered, then $\alpha$ (and $\omega_s$) will be underestimated (McDonald and Kreitman 1991; Eyre-Walker 2002) because some SDMs that currently segregate would not have segregated or become fixed in the past. The extent of the bias depends on the magnitude of the difference in effective population size between the polymorphism and divergence phases, the DFE of new mutations, and the true level of adaptive evolution (Eyre-Walker 2002; Eyre-Walker and Keightley 2009). However, contracting effective population size seems an unlikely explanation for the patterns we observe in plants because we see little evidence of adaptive evolution across nine comparisons. In contrast, an expansion in population size might explain why sunflowers show evidence of adaptive evolution because they have recently undergone population expansion (Strasburg and Rieseberg 2008) and expansion can lead to an overestimate of adaptive evolution (McDonald and Kreitman 1991; Eyre-Walker 2002; Eyre-Walker and Keightley 2009).

A third possible explanation for the low levels of adaptive evolution in plants is population structure. Population structure can reduce the probability that an advantageous mutation will become fixed if migration rates are low because of the increased influence of drift within subdivided populations and the added influence of population extinction (Charlesworth et al. 2003; Whitlock 2003). Unfortunately, we do not have information about population structure for most of the species we have considered; A. thaliana and A. lyrata show contrasting levels of structure, so it is unclear what level of structure there was during the divergence of these species. Population structure could affect our estimates in two ways. First, it is possible that population structure could lead to biases in our method to estimate the DFE and hence the rate of adaptive evolution; this seems unlikely because our method gives similar estimates of the DFE for both the total population and each sub-population of Arabidopsis. Furthermore, it gives similar estimates to the method of Fay et al. (2001) when rare single nucleotide polymorphisms are excluded. However, to investigate the matter further, we performed a series of simulations in which we had two subpopulations with varying levels of migration between them (supplementary fig. S1a and b, Supplementary Material online). Our method successfully estimates the proportion of adaptive substitutions irrespective of the level of migration (supplementary table S3, Supplementary Material online) if the polymorphism data are sampled from one sub-population only. The method of Eyre-Walker and Keightley (2009) overestimated $\alpha$ if the level of migration was low between the two subpopulations and the polymorphism data included samples from both subpopulations. This is due to an excess of fixed differences between the two subpopulations, which will be treated as polymorphisms in a common data set and hence strongly affect the site frequency spectrum.

We may have also underestimated $\alpha$ because recent gene duplications are probably underrepresented in our data because of problems of identifying orthologs. Plant genomes contain high frequencies of duplicated genes (Arabidopsis Genome Initiative, 2000; The International
Little Evidence for Adaptive Evolution in Plant Species · doi:10.1093/molbev/msq079

Rice Genome Sequencing Project 2005; Tuskan et al. 2006; Jaillon et al. 2007) due to polyploidization (Blanc and Wolfe 2004; Soltis PS and Soltis DE 2009) and tandem duplications (Rizzon et al. 2006). These have been shown to have higher levels of adaptive evolution than single copy genes, at least in mammals (Han et al. 2009). Thus, we may be missing much of the adaptive evolution that goes in plants simply because it occurs in the divergence of paralogs, which we have not sampled. There may also be a bias toward conserved genes, particularly in the more divergent taxa, because of the need to be able to correctly identify the ortholog if conserved genes undergo less adaptive evolution, then $\alpha$ will be underestimated.

In the MK framework, it is assumed that one category of mutations, in this analysis the synonymous mutations, are neutral. However, there is evidence in at least in Populus species that selection acts on synonymous codon use (Ingvarsson 2008a). Predicting the effects of selection, on the putatively neutral sites, on the estimation of the DFE and rates of adaptive evolution is not straightforward and deserves more investigation. If nonsynonymous polymorphisms are on average more deleterious than synonymous mutations, then the MK test and methods that estimate $\alpha$ from it are expected to be conservative if populations sizes are stationary, this is because the nonsynonymous polymorphisms are less likely to be fixed than the synonymous polymorphisms. With nonstationary population size, the predictions become complex. The fact that our method estimates a proportion of mutations to be slightly deleterious ($1 < N_{e}s < 10$) in all species suggests that nonsynonymous polymorphisms are on average more deleterious than their synonymous counterparts. However, it is possible that selection at synonymous sites along with demographic changes is inducing an underestimate of $\alpha$.

One potential problem with some of our comparisons is the very low level of divergence between the ingroup and outgroup species. It is intriguing that all three of the species pairs for which $\alpha$ is positive have a low level of synonymous divergence relative to nucleotide diversity: H. petiolaris (polymorphism)–H. annuus (outgroup) = 1.7, S. globosa–S. adamantis = 1.1, P. tremula–P. trichocarpa = 4.7, and P. balsamifera–P. trichocarpa = 2.1. Only H. annuus–H. petiolaris and S. bicolor–S. propinquum have comparable ratios; all other comparisons have a ratio greater than 10. This may lead to either an overestimation of $\alpha$ because the fixation time becomes important; advantageous and SDMs go to fixation faster than neutral mutations if they are due to fix (Kimura 1983) so there may be a time after the divergence of species where sites subject to selection are more likely to show fixed differences than neutral sites because the mutations are more likely to have already spread to fixation (Bierne and Eyre-Walker 2004). The application of the MK framework to closely related species requires more investigation. Although some of the species are very closely related to each other, rice is highly divergent from its outgroup sorghum; this could lead to an overestimate of $\alpha$ because we might expect synonymous sites to become saturated. There is no evidence of this effect in our data because the estimate of $\alpha$ for the divergence between rice and sorghum is negative.

The lack of adaptive evolution in plants is particularly surprising in some genera, such as Schiedea. Schiedea is a Hawaiian endemic that adapted to a wide range of ecological conditions (from cool rainforest to arid desert-like conditions of coastal cliffs) and evolved extremely different morphology (ranging from vines to bushes) and reproductive biology (dioecy, gynodioecy, and hermaphrodites) (Wagner et al. 2005). Given fairly low divergence between species in the genus ($K_s < 4\%$), all the morphological and ecological diversity in Schiedea has evolved surprisingly rapidly. Thus, one might expect that species in the genus evolved under fairly strong selection, which should be detectable in Schiedea genes. Indeed, phylogeny-based maximum likelihood analysis revealed that positive selection is more widespread in Schiedea genes compared with mainland plant groups (Kapralov VK, Votintseva AA, Filatov DA, unpublished data). Thus, it is somewhat surprising that this rapidly evolving Hawaiian endemic does not show significantly positive $\alpha$. However, in the current paper, the Schiedea data set is the smallest of all, and adding more genes might provide evidence of positive selection in Schiedea genes. On the other hand, the current analysis is restricted to population-level evolutionary processes in only one of the species. Thus, the approach presented in the current paper cannot detect positive selection that might have acted during the early history of the genus while it adapted to diverse ecological conditions.

Although the influence of effective population size on the rate of adaptive evolution in plants is unclear, there is a clear effect of $N_{e}$ on the DFE; species with small $N_{e}$, such as B. stricta, P. balsamifera, and S. globosa, tend to have a relatively high proportion of mutations that are effectively neutral; there is a significant negative correlation between $P_{n}/P_{s}$ and $N_{e}$, even when we take into account the nonindependence between these two statistics (fig. 4: $r = -0.75,$

![FIG. 4](image-url) The correlation between $P_{n}/P_{s}$ and $N_{e}$. $P_{n}/P_{s}$ and $N_{e}$ were estimated independently from two estimates of $P$, that were obtained by splitting $P$ into two independent halves according to the method suggested by Piganeau and Eyre-Walker (2009).
p = 7 × 10⁻³, Spearman rank test) using the method of Piganeau and Eyre-Walker (2009) (note we do not add one to denominator, as suggested by Piganeau and Eyre-Walker because P₁ is sufficiently large that this correction makes no difference). In contrast to the proportion of mutations that are effectively neutral, the proportion of mutations that are slightly deleterious, 1 < Nₑₛ < 10, and to a lesser extent the proportion of mutations that are moderately deleterious, 10 < Nₑₛ < 100, do not vary very much between species.

Conclusions
We have estimated the proportion of nonsynonymous substitutions that are a consequence of positive adaptive evolution between nine pairs of plant species; we find little evidence of adaptive evolution in any of them. This is in striking contrast to Drosophila, bacteria and mice in which rates of adaptive evolution have been estimated to be substantial. The low estimate is unlikely to be a methodological artifact, and it does not appear due to low effective population size because some plants appear to have population sizes that rival that of mice. It therefore seems that the rate of adaptive evolution may be determined by other reasons as much as by effective population size and plants may have an outstanding role.

Supplementary Material
Supplementary tables S1–S3 and figures S1–S5 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
We are grateful for the access to the maize genome data, which were produced by the Genome Sequencing Center at Washington University School of Medicine in St Louis. We thank Matt Olson and Peter Tiffin for making the P. balsamifera data rapidly and readily available for the analysis. T.I.C. was financially supported by the John Maynard Smith studentship. The work was supported by the John Maynard Smith, National Science Foundation (award EF-0723447), the National Institute of Health (award R01 GM086496), Natural Environment Research Council UK, Duke University, the University of Sussex, and the Max Planck Society.

References


