Reduced Efficacy of Natural Selection on Codon Usage Bias in Selfing *Arabidopsis* and *Capsella* Species

Suo Qiu[1,2,*,†], Kai Zeng[2,†], Tanja Slotte[3,4,†], Stephen Wright[3,5,*] and Deborah Charlesworth[2]

[1]State Key Laboratory of Biocontrol and Key Laboratory of Gene Engineering of the Ministry of Education, Sun Yat-Sen University, Guangzhou, China

[2]Institute of Evolutionary Biology, School of Biological Sciences, The University of Edinburgh, United Kingdom

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

[4]Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Sweden

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

*Corresponding author: E-mail: kaidy330@gmail.com; stephen.wright@utoronto.ca.

†These authors contributed equally to this work.

Accepted: 3 August 2011

Abstract

Population genetic theory predicts that the efficacy of natural selection in a self-fertilizing species should be lower than its outcrossing relatives because of the reduction in the effective population size (N_e) in the former brought about by inbreeding. However, previous analyses comparing *Arabidopsis thaliana* (selfer) with *A. lyrata* (outcrosser) have not found conclusive support for this prediction. In this study, we addressed this issue by examining silent site polymorphisms (synonymous and intronic), which are expected to be informative about changes in N_e. Two comparisons were made: *A. thaliana* versus *A. lyrata* and *Capsella rubella* (selfer) versus *C. grandiflora* (outcrosser). Extensive polymorphism data sets were obtained by compiling published data from the literature and by sequencing 354 exon loci in *C. rubella* and 89 additional loci in *C. grandiflora*. To extract information from the data effectively for studying these questions, we extended two recently developed models in order to investigate detailed selective differences between synonymous codons, mutational biases, and biased gene conversion (BGC), taking into account the effects of recent changes in population size. We found evidence that selection on synonymous codons is significantly weaker in the selfers compared with the outcrossers and that this difference cannot be fully accounted for by mutational biases or BGC.

Key words: *Arabidopsis*, *Capsella*, natural selection, inbreeding, codon usage bias.

Introduction

Population genetics theory predicts that the effective population size (N_e) of an inbreeding species will be lower than that of an outcrossing species with the same census population size (Wright 1969; Pollak 1987), at both within-population and species-wide levels (Pannell and Charlesworth 2000; Ingvarsson 2002). The smaller N_e in a selfer is attributable to a reduction in the number of independent gametes sampled for reproduction. In addition, the effectiveness of recombination is expected to be lowered in selfers because of increased homozygosity (Conway et al. 1999; Nordborg 2000). The lower effective recombination rate intensifies hitchhiking effects caused by selection at linked sites, including the continual elimination of deleterious mutations (Charlesworth et al. 1993; Kaiser and Charlesworth 2009) or the spread of advantageous ones (Smith and Haigh 1974). Hitchhiking effects further reduce N_e and can also affect a larger proportion of the genome in selfers than in outcrossers.

A reduced N_e implies that the efficacy of selection on nonsynonymous mutations should be reduced in a highly inbred species relative to closely related outcrossing sibling species. This hypothesis can be tested by either comparing nonsynonymous substitutions in inbreeding lineages with lineages that have remained outcrossing or by comparing frequency spectra of such variants in populations of the two kinds of species. Using these approaches, it has been found that the outcrossing crucifer species *Capsella grandiflora* has experienced a greater efficacy of selection on

nonsynonymous mutations than the selfer *Arabidopsis thaliana* (Slotte et al. 2010), consistent with the predicted effects of the difference in mating system. However, in another study, where *A. thaliana* was compared with a closer outcrossing relative, *A. lyrata*, no detectable disparity in the strength of selection was found (Foxe et al. 2008).

Inbreeding species should also experience a lower effectiveness of selection on synonymous codons, an important class of weakly selected sites in genomes (Marais et al. 2004). As has been established in a diversity of taxa, natural selection can favor some codons for an amino acid over the others. These "preferred" codons confer greater efficiency and/or accuracy during the protein translation process, resulting in biases in codon usage (see a recent review by Hershberg and Petrov 2008). However, the selective difference between "preferred" and "unpreferred" codons is weak, usually of the order of $4N_e s \approx 1$, where $2s$ is the selection coefficient against homozygotes for the unpreferred state. With such weak selection, we might expect stronger effects of low N_e on patterns of codon bias than on nonsynonymous substitution rates because weakly selected mutations are most susceptible to fixation if N_e is reduced (McVean and Charlesworth 1999).

Codon bias in *A. thaliana* has been inferred to be much lower than in many other eukaryote species such as *Drosophila* and *Caenorhabditis* (Duret and Mouchiroud 1999; dos Reis and Wernisch 2009), although the large evolutionary distances between these species makes it hard to conclude whether the reduction in *A. thaliana* is a direct consequence of becoming self-fertilizing. However, previous attempts to detect a correlation between mating system and the efficacy of selection on synonymous codons in *A. thaliana* and its close relatives were inconclusive. For example, comparisons of substitutions between preferred and unpreferred codons in *A. thaliana* and *A. lyrata* found no significant difference in selection intensity in these two lineages (Wright et al. 2002, 2007). The failure to detect an effect of changes in mating system could be attributable to insufficient data (in terms of the number of genes available and/or the number of substitutions accumulated between *A. thaliana* and *A. lyrata*) and the fact that the outgroup species used to polarize direction of substitutions were very divergent, reducing the accuracy and power of inferences.

Some questions with respect to selection on synonymous codons in plants thus remain unanswered. 1) Are synonymous codons under recent selection in plant genomes? 2) Do selfing species experience lower effective selection on codon bias? 3) What are the contributions of mutational biases and biased gene conversion (BGC) to patterns of codon bias? For instance, codon preferences in *A. thaliana* correspond well with expectations based on tRNA abundances and wobble rules, and highly expressed genes show a higher frequency of preferred codons, as expected under models of translational accuracy or efficiency (Wright et al. 2004). However, the extent to which polymorphism patterns observed at synonymous sites in *A. thaliana* can be explained by recent selection on synonymous codons is unknown.

In this study, we analyze these questions using extensive protein-coding and intron sequence polymorphism data from *Arabidopsis* and *Capsella* species. The first comparison is between *A. thaliana* (selfer) and *A. lyrata* (outcrosser) with an estimated divergence time of 5–13 My (Koch et al. 2000; Beilstein et al. 2010) and the second is between the outcrosser *C. grandiflora* and its highly selfing relative *C. rubella*, which is estimated to have originated, possibly from a single ancestral *C. grandflora* genotype that lost self-incompatibility, 30,000–50,000 years ago (Guo et al. 2009). In addition to previously published polymorphism data, our analyses used newly obtained polymorphism data from single exons of 354 loci in *C. rubella* and 89 new loci in *C. grandiflora* (making the total number of *C. grandiflora* loci analyzed in the second comparison 346); most of these are orthologous in the two *Capsella* species. Further details of diversity patterns will be presented elsewhere (Slotte T, Hazzouri K, Wright S, unpublished data), but diversity is considerably lower in the selfer (synonymous site diversity, π_s, in *C. grandiflora* and *C. rubella* are 0.017 and 0.004, respectively), similar to the overall diversity difference between the *Arabidopsis* species pair (π_s are 0.018 and 0.006 for *A. lyrata* and *A. thaliana*, respectively). Clearly, the selfing species show evidence of having lower N_e, and therefore, the expectation that they should have less codon usage bias is applicable to both these species pairs.

To extract more information from the data, we have extended the two-allele model developed by Zeng and Charlesworth (2009, 2010a, 2010b), so that detailed selective differences between synonymous codons can be estimated for all the amino acids with 2-fold degenerate codons (referred to as 2-fold amino acids), together with the relevant mutational parameters, taking into account the confounding effects of recent changes in population size. We have also extended a multiallele method (Zeng 2010), so that the effects of mutational biases and BGC can be studied in *A. thaliana*, where signatures of recent population expansion have been found. Using these extended methods, we conclude that selection on synonymous codons is significantly weaker in the inbreeding species compared with the outcrossers and that this difference cannot be fully accounted for by mutational biases or BGC.

Materials and Methods

Models

Two-Allele Model

We analyzed diversity data in order to detect footprints of recent selection and estimate the strength of selection on the variants analyzed. To do this, we developed the following model based on the framework of Zeng and

Charlesworth (2009). Briefly, this model is an extension of the Li–Bulmer model of drift, selection, and reversible mutation between alleles in a diploid Wright–Fisher population (Li 1987; Bulmer 1991). Instead of binning synonymous codons into "preferred" and "unpreferred" classes, we focused on the ten amino acids encoded by 2-fold degenerate codons (abbreviated as 2-fold amino acids; serine was treated as two amino acids, one with four codons and the other with two, referred to as Ser4 and Ser2, respectively). We define the respective fitnesses of the three possible genotypes, assuming a genic selection model, which has been shown to be a good approximation for systems under weak selection (Garcia-Dorado and Caballero 2000). To illustrate the notation, for the example of Phe, an amino acid with two codons, we denote TTT and TTC by $A_{\text{Phe-1}}$ and $A_{\text{Phe-2}}$, respectively. Denoting the selection coefficient for Phe as s_{Phe}, we write the fitness of the $A_{\text{Phe-1}}A_{\text{Phe-1}}$ genotype as $1 - 2s_{\text{Phe}}$; for $A_{\text{Phe-1}}A_{\text{Phe-2}}$, it is $1 - s_{\text{Phe}}$; and for $A_{\text{Phe-2}}A_{\text{Phe-2}}$, it is equal to 1. For each of the nine other amino acids with two codons, we define an individual selection coefficient.

To model mutation, we note that seven of the ten such amino acids have synonymous codons that differ by T versus C (like the situation for $A_{\text{Phe-1}}$ and $A_{\text{Phe-2}}$), whereas three have codons that differ by A versus G. We assumed a mutation rate from T to C of u_{TC} per site per generation and a back-mutation rate of $\kappa_{\text{TC}}u_{\text{TC}}$. We further assumed that these two mutational parameters are shared by all the seven amino acids with T/C-ending codons. Similarly, we define u_{AG} and κ_{AG} for the three amino acids with A/G-ending codons.

To estimate parameters under the above model, it is necessary to combine information from multiple sites across the genome. Consider a sample of n sequences sampled randomly from a population. We can extract all the synonymous sites (fixed or polymorphic) where, for instance, Phe is encoded. Then, we can construct the observed site-frequency spectrum (SFS) for Phe, denoted by $\boldsymbol{D}_{\text{Phe}} = (d_{\text{Phe-0}}, d_{\text{Phe-1}}, \ldots, d_{\text{Phe-}n})$, where $d_{\text{Phe-}i}$ is the number of sites at which $A_{\text{Phe-1}}$ is represented i times in the sample. Similarly, an SFS is constructed for every other 2-fold amino acid.

According to diffusion theory, when the evolutionary forces are weak, in a population with a constant effective size of N_1, the sampling properties of the SFSs in the model are governed by the following scaled parameters: $\gamma_x = 4N_1s_x$, $\gamma_y = 4N_1s_y$ and $\theta_{\text{TC}} = 4N_1u_{\text{TC}}$ and $\theta_{\text{AG}} = 4N_1u_{\text{AG}}$, where $x \in X = \{\text{Phe, Tyr, Cys, His, Asn, Ser2, Asp}\}$ and $y \in Y = \{\text{Gln, Lys, Glu}\}$. With these parameters, for each sample size in the data set, an expected SFS can be generated by iterating a Markov transition matrix [see equations (1) to (6) of Zeng and Charlesworth 2009]. Let Θ represent all the selection and mutational parameters in the model and $\boldsymbol{D}$ represent all the observed SFSs. Assuming that the population is at equilibrium and that the sites evolve independently, we can use the matrix-based approach of Zeng and Charlesworth (2009) to obtain the following ln-likelihood of the sample:

$$L_0(\boldsymbol{D}|\Theta) = \sum_{x\in X} L_0(\boldsymbol{D}_x|\gamma_x, \theta_{\text{TC}}, \kappa_{\text{TC}}) + \sum_{y\in Y} L_0(\boldsymbol{D}_y|\gamma_y, \theta_{\text{AG}}, \kappa_{\text{AG}}), \quad (1)$$

where the L_0 terms on the right-hand side are defined by equation (12) of Zeng and Charlesworth (2009) and $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$ are the observed SFSs for 2-fold amino acids with T/C- or A/G-ending codons, respectively.

To take account of the possibility of a recent change in population size, our model assumes that the population started at equilibrium with an effective population size of N_1. At time zero, the population size changed instantly to N_2 and then remained constant for t generations, until the population was sampled. This highly simplistic demographic model has been widely used and has been shown to be a useful starting point for reducing the false-positive inferences about selection that can be caused by demographic changes in populations (Williamson et al. 2005; Li and Stephan 2006; Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). Two more parameters are needed to characterize this population growth/contraction model: $g = N_2/N_1$ and $\tau = t/(2N_2)$. We can write the ln-likelihood as

$$L_1(\boldsymbol{D}|\Theta) = \sum_{x\in X} L_1(\boldsymbol{D}_x|\gamma_x, \theta_{\text{TC}}, \kappa_{\text{TC}}, g, \tau) + \sum_{y\in Y} L_1(\boldsymbol{D}_y|\gamma_y, \theta_{\text{AG}}, \kappa_{\text{AG}}, g, \tau), \quad (2)$$

where the L_1 terms on the right-hand side are defined by equation (13) of Zeng and Charlesworth (2009). When $t = +\infty$ and/or $g = 1$, the more elaborate model defined by equation (2), denoted by L_1, reduces to the simpler model defined by equation (1), denoted by L_0. We use a chi-squared test with 2 degrees of freedom (df) to distinguish between L_1 and L_0 (Williamson et al. 2005; Zeng and Charlesworth 2009). The L_1 model is computationally demanding; we therefore used the rescaling method described in Haddrill et al. (2011) to speed up the calculation (see eq. 2 therein).

Equations (1) and (2) offer a way to test for the action of natural selection. For example, we can use a likelihood ratio test with 1 df to determine whether the selection coefficient for a particular amino acid, for example, γ_{Phe}, differs significantly from zero, by comparing a reduced model with γ_{Phe} fixed at zero, denoted by $L_0(\boldsymbol{D}|\Theta, \gamma_{\text{Phe}} = 0)$, with the full model $L_0(\boldsymbol{D}|\Theta)$. Note that, although equations (1) and (2) rely on the assumption that sites evolve independently, a recent simulation study has shown that this assumption has little impact on the detection of natural selection by such likelihood ratio tests and on estimation of the mutational bias parameters (Zeng and Charlesworth 2010b).

In the above likelihood-based analyses, we used the simplex algorithm of Press et al. (1992, Chap. 10) to find the parameter values that maximize the expressions in equations (1) and (2). We ran the algorithm multiple times with random starting points to ensure that the global maximum was found. For species where the simpler model of L_0 is not rejected by the likelihood ratio test, we further used Markov chain Monte Carlo (MCMC) to obtain confidence intervals for the γ estimates (due to computational difficulties, we did not carry out an MCMC analysis when L_1 is the preferred model). The details of the MCMC analysis were similar to those described in Zeng (2010, see equations (11) and (12) therein). Briefly, uniform priors were assumed for all parameters, and the Markov chain was constructed using the Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970). To obtain random samples from the posterior distribution, a burn-in of 5×10^5 proposals was run, and samples were then taken every 10^4 proposals. The procedure was repeated several times to check for convergence of the Markov chain. The final results were based on at least 2,500 random samples.

Multiallele Model

As described in the Results section, the data collected from *A. lyrata*, *C. grandiflora*, and *C. rubella* do not favor the nonequilibrium model of L_1 over the equilibrium of L_0. For these sets of data, we also used a multiallele model (Zeng 2010) to test for selective differences between synonymous codons for amino acids with more than two codons (referred to as multifold amino acids) and estimate a mutation rate matrix between the four nucleotides. This model is very parameter rich (with 52 parameters) and computationally demanding. As above, we used the simplex algorithm with multiple random starting points in the search of the maximum likelihood estimates of the parameters. Similarly, confidence intervals were obtained by MCMC which, in this case, involved a burn-in of 1.5×10^6 proposals and a sampling interval of 5×10^4 proposals. The final results were based on at least 3,000 random samples.

For *A. thaliana*, however, there is strong evidence for recent population growth. Although it is, in principle, possible to implement a nonequilibrium version of the multiallele model (see supplementary text of Zeng 2010), so that the synonymous data from this species could also be analyzed by this approach, this is computationally prohibitive. Therefore, for *A. thaliana*, we used the nonequilibrium version of the multiallele model to analyze data only from introns (see below) in order to detect the action of BGC and to estimate a mutation rate matrix between nucleotides.

Data and Sequence Analysis

Because the methods described in the previous section do not require outgroup sequences or polarization of the direction of mutations, we extracted within-species polymorphism data for the four plant species separately. For *A. thaliana*, we obtained data for 780 exons and 821 introns from the polymorphism data published by Nordborg et al. (2005). The samples used in the analysis were taken from the 15 populations sampled from sites scattered across Europe, which represents the native distribution of this species (Beck et al. 2008). The *A. lyrata* polymorphism data were compiled from several studies (Wright et al. 2003, 2006; Ramos-Onsins et al. 2004; Foxe et al. 2008; Kawabe et al. 2008; Qiu S, Hamilton K, Charlesworth D, unpublished data). The plants in these studies were sampled from two to seven natural populations in Europe (supplementary table 1, Supplementary Material online). We excluded North American populations of this species because they were found to have gone through strong population bottlenecks (Wright et al. 2006; Ross-Ibarra et al. 2008). In total, 120 large exons and 41 introns from *A. lyrata* were retained for the analysis.

For *C. grandiflora*, a polymorphism data set containing 257 exons was recently published (Slotte et al. 2010). For the present study, we sequenced an additional 89 exons. Samples taken from the five populations in Greece, which is within the natural range of this species (Guo et al. 2009), were used in the analysis (supplementary table 2, Supplementary Material online). We further produced a large polymorphism data set of 354 exons in the selfer *C. rubella*. These samples were collected from eight populations situated throughout the Mediterranean region (supplementary table 2, Supplementary Material online). Most of these exons (334) are orthologous between *C. grandiflora* and *C. rubella*. These new sequence data were obtained using the same sequencing protocols described in Slotte et al. (2010) and were submitted to GenBank, with accession numbers JN403374–JN406266.

For all four species, we built our final data sets by randomly selecting one allele from each population (note that our analysis does not require information about the phase

Table 1
Summary of the Data Analyzed, and Diversity Statistics for the Study Species

Species	Mating Type	Data	N_{pop}	N_{loci}	$\bar{n}$	L_s	S_s	π_s
Arabidopsis thaliana	Highly selfing	Exon	15	780	13.9	85,239	1,687	0.006
		Intron	15	821	15.0	74,779	1,645	0.006
Arabidopsis lyrata	Outcrossing	Exon	7	120	4.4	16,806	547	0.018
		Intron	7	41	4.9	15,234	366	0.013
Capsella rubella	Highly selfing	Exon	8	354	7.3	60,217	559	0.004
Capsella grandiflora	Outcrossing	Exon	5	346	4.3	58,520	1,807	0.017

NOTE.—One allele was randomly sampled from each population. L_s, the total number of silent sites (synonymous or intronic); $\bar{n}$, mean sample size; N_{loci}, the total number of loci; N_{pop}, the number of populations; S_s, the total number of silent polymorphic sites; π_s, mean silent site diversity.

between variants within a heterozygous individual). A summary of the final data is presented in table 1. It has been shown analytically that this sampling strategy can effectively reduce the potential influence of population structure on population genetic analysis (Wakeley 1999; Wakeley and Aliacar 2001). We also repeated our analysis on *A. thaliana* using a reduced data set constructed by sampling one allele from each of the five "populations" identified by the *STRUCTURE* model (Pritchard et al. 2000; Falush et al. 2003; Nordborg et al. 2005). All our major conclusions remain unchanged, but the power of the statistical tests is much lower. Therefore, we present only results obtained from the larger sample in the Results section.

Data Deposition

The new sequence data obtained in this study were submitted to GenBank, with accession numbers JN403374–JN406266. All sequence alignments used in this study are available from the corresponding authors upon request.

Results

Analyses of 2-Fold Amino Acids

Demographic History

Using the data summarized in table 1, we first test whether there is evidence for recent changes in population size, by comparing the equilibrium model L_0 (eq. 1) with the non-equilibrium model L_1 (eq. 2). Strong evidence of recent population growth was found in *A. thaliana* ($\chi^2 = 21.32$, df = 2, $P = 2.34 \times 10^{-5}$). The estimated ratio of the new population size to that before the expansion was $g = 8.71$, and the time, measured in units of twice the extant effective population size since the expansion is $\tau = 0.0067$. In contrast, the data from all the other species fail to reject the equilibrium model ($P > 0.2$ in all cases). The finding of a recent population expansion in *A. thaliana* is in agreement with previous studies (Sharbel et al. 2000; Nordborg et al. 2002, 2005; Beck et al. 2008). For *C. grandiflora*, a recent study of the frequency distribution of polymorphisms and linkage disequilibrium also suggested a large stable population (Foxe et al. 2009).

For the other two species, the failure to reject L_0 should be interpreted with caution. First, the *A. lyrata* data set contains only 120 gene fragments in samples from two to seven European populations, considerably less than that for either *A. thaliana* or *C. grandiflora* (table 1). This must lower the power to detect recent demographic changes. The same concern applies to *C. rubella*. In this species, although we have roughly the same numbers of genes and populations as in *C. grandiflora*, the data contain less information because the diversity is much lower (table 1). Indeed, bottleneck events have been reported in both *A. lyrata* and *C. rubella* (Wright et al. 2006; Ross-Ibarra et al. 2008; Foxe et al. 2009; Guo et al. 2009), although in the latter species this may be a "selection bottleneck" at the time when a self-compatibility mutation spread in the species. Although the model of Zeng and Charlesworth (2009) could be further extended to incorporate more complex demographic models, this would introduce more parameters. Given the dearth of data in these two species, we did not pursue this possibility. Nonetheless, these inferences of bottleneck events may not be incompatible with the lack of significant difference in the comparison between L_0 and L_1, as they are based on data from different sampling strategies (see the Discussion section below). Furthermore, a recent simulation study has shown that the likelihood ratio tests for detecting selection developed under the Zeng and Charlesworth (2009) method is unlikely to produce false-positive results even when the true demographic history is much more complex than that assumed in the model (Zeng and Charlesworth 2010b). Therefore, the L_0 model should allow useful conclusions to be drawn from the data of *A. lyrata* and *C. rubella*.

Effect of Mating System Differences

Our next analyses test the effect of changes in mating system on codon bias, by comparing the selection coefficients (γ values, see Materials and Methods) for the ten 2-fold amino acids between the inbreeding and outcrossing species (fig. 1). For the seven amino acids with T/C-ending codons, the analysis always infers that C-ending codons are favored over the T-ending ones in the two outcrossing species (i.e., with positive γ values). This is consistent with the preferred codons estimated for these amino acids using multivariate statistical methods or analysis of gene expression (Chiapello et al. 1998; Wright et al. 2004). The selective differences are significant for all seven amino acids with T/C-ending codons in both of the outcrossing species, except for Asp in *A. lyrata* (fig. 1). In the two inbreeding species, the selection coefficients for these seven amino acids are much lower and are significant for only four of them in *A. thaliana* and none in *C. rubella* (fig. 1). For the three amino acids with A/G-ending codons, the G-ending codons are nominally favored by natural selection over the A-ending one in all four species. However, only 2 of the 12 γ values (γ_{Lys} and γ_{Glu} in *A. thaliana*) are marginally significantly different from zero. Therefore, selection on these three amino acids may be very weak, and the observations are likely to be heavily affected by sampling variance, as suggested by the large confidence intervals associated with these estimates (fig. 1).

Comparisons between *A. thaliana* (selfer) and *A. lyrata* (outcrosser) and between *C. rubella* (selfer) and *C. grandiflora* (outcrosser) show that the γ values of the selfer in each pair of species tend to be lower than those of the outcrosser, consistent with the prediction of a lower intensity of selection in selfers. This pattern is particularly clear for the six amino acids that give strong statistical support for selection in both outcrossing species (i.e., Phe, Asn, Tyr, Ser2, His, and Cys; Wilcoxon rank

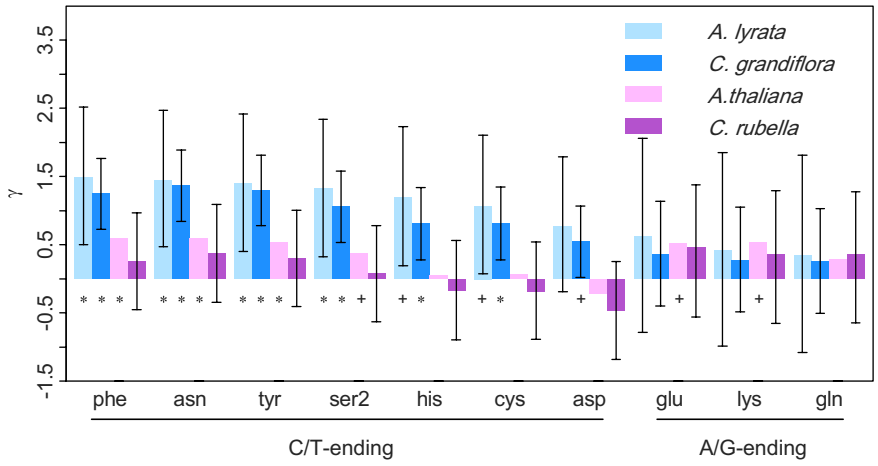

Fig. 1.—Comparisons of the estimated selection coefficient, γ, for the ten amino acids with 2-fold degenerate codons between the inbreeding and outcrossing species. The maximum likelihood estimates of the γ values are shown by the bars. Whiskers are 95% confidence intervals obtained by the MCMC analysis. We use * and + to indicate γ values that are significantly different from zero at $P < 0.01$ and $0.01 < P < 0.05$, respectively, under a likelihood ratio test with 1 df.

sum test for paired samples, $P = 0.03$). Additionally we tested, using likelihood ratios (with 1 df), whether the γ value for each amino acid is the same in a selfer as in its related outcrosser. Between *A. thaliana* and *A. lyrata*, the difference in γ is significant for six of the seven amino acids with T/C-ending codons, Phe ($P = 0.03$), Cys ($P = 0.02$), Asp ($P = 0.03$), His ($P = 0.02$), Asn ($P = 0.03$), and Ser2 ($P = 0.02$) but not for Tyr ($P = 0.06$). In contrast, none of the three amino acids with A/G-ending show detectable between-species difference ($P > 0.15$ for all tests). A very similar pattern is found in the comparison between *C. rubella* and *C. grandiflora*, where all seven amino acids with T/C-ending codons show a significant between-species difference (all $P < 0.034$) but none of the three A/G-ending ones (all $P > 0.55$). We therefore conclude that amino acids with T/C-ending codons, where selection on codon bias is strongest, also most strongly support a reduced efficacy of selection in the two selfers.

To further clarify whether the difference in γ is due to mating system changes, we tested the possibility that the difference is caused by the sets of genes we compared having different gene expression levels, a key determinant of codon bias. We therefore obtained expression data for each gene in the *A. thaliana* and *A. lyrata* data sets in five separate tissues from the *A. thaliana* MPSS Web site (Meyers et al. 2004), using the methods described in Wright et al. (2004). There is no significant difference in expression (Wilcoxon test with the null hypothesis that mean expression levels of *A. thaliana* and *A. lyrata* are equal, $P = 0.28$). For *C. rubella* and *C. grandiflora*, the genes we analyzed are orthologous, so it is reasonable to assume that they have similar expression levels in these closely related species. Put together, we suggest that the difference in γ is probably not caused by differences in gene expression, although more data on gene expression in each species are needed to draw a definitive conclusion.

In spite of the large differences in γ values between different amino acids and much reduced estimated selection strength in the two selfing species, the selection coefficients correlate positively across the four species; for instance, Phe tends to have consistently high γ values across all species. For the seven amino acids with T/C-ending codons, the pairwise correlation coefficients between γ estimates from different species, using Kendall's rank correlation, had *P* values ranging from 0.007 to 0.035. To test the possibility that this might be due to a methodological bias, we applied equation (1) to a *Drosophila melanogaster* polymorphism data set (Shapiro et al. 2007). In contrast to the observation of weaker selection acting on the amino acids with A/G-ending codons in the four plant species, estimates of γ values associated with these amino acids are very high in *D. melanogaster* (supplementary fig. 1, Supplementary Material online), consistent with previous reports (McVean and Vieira 2001; Zeng 2010). Furthermore, for the amino acids with T/C-ending codons, the γ values in *D. melanogaster* are uncorrelated with those obtained from any of the four plant species (Kendall's rank correlation, $P > 0.05$). This suggests that the results in figure 1 are unlikely to be artifactual, and we therefore conclude that the pattern of selection on the ten 2-fold amino acids is fairly well conserved across the four plant species studied here.

Analysis of Multifold Amino Acids

We also applied a multiallele model (Zeng 2010) to protein-coding data in the three species where the observations do not exclude the equilibrium model (see above). The purpose of this analysis is to estimate selective differences between

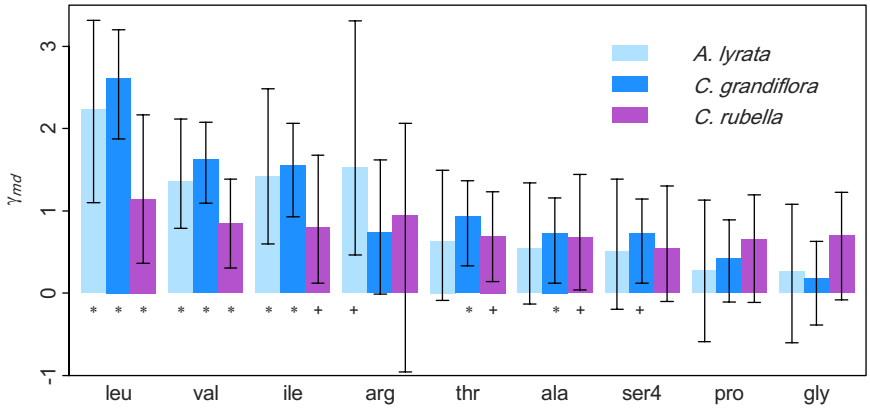

Fig. 2.—Estimates of γ_{md} for multifold amino acids in the three plant species. γ_{md} is the difference in γ between the best and worst codons among those encoding the same amino acid. The bars show the maximum likelihood estimates, and the whiskers are 95% confidence intervals obtained by the MCMC analysis. γ_{md} values that are significantly different from zero at $P < 0.01$ and $0.01 < P < 0.05$ under a likelihood ratio test with 1 df are indicated by * and +, respectively.

codons for amino acids with more than two codons (multifold amino acids), as well as the mutation rates between nucleotides (see below). In this model, every synonymous codon for an amino acid has its own selection coefficient, γ. As a measure of selection intensity, for each amino acid, we calculated the maximum difference in gamma, γ_{md}, defined as the difference in γ between the best and worst codons among those encoding the same amino acid. The outcome of the analysis is given in figure 2.

The first interesting finding is that γ_{md} varies dramatically between different amino acids in the two outcrossing species. In *C. grandiflora*, for example, only six out of the nine multifold amino acids have γ_{md} values significantly different from zero, and in *A. lyrata*, only four have significant γ_{md} values. The inference that selection is consistently undetectable in some multifold amino acids across these two outcrossing species (i.e., for Pro and Gly) is in sharp contrast to the detectable selection in all amino acids in *Drosophila* (McVean and Vieira 2001; Zeng 2010) and may offer an explanation of the overall reduction in selection on synonymous codons observed in several plant species compared with Drosophila (see Discussion).

Second, comparing γ_{md} between the sibling species *C. grandiflora* and *C. rubella*, we found that γ_{md} is significant in both of the species for five amino acids (Leu, Val, Ile, Thr, and Ala). Among these, the larger γ_{md} value is always found in *C. grandiflora*, consistent with a reduced intensity of selection on synonymous mutations in the selfer, *C. rubella* (note that we refrained from comparing the difference between species statistically using the multiallele model because of extreme computational constraints imposed by a very large number of parameters). The difference is greatest for Leu, Val, and Ile, which also yielded large and significant γ_{md} values in *A. lyrata*. In the other multifold amino acids, the γ_{md} values are generally small and do not differ significantly from zero. The larger γ_{md} values in *C. rubella* in some of these families seen in figure 2 are probably due to sampling variance and/or other complications that the model does not take into account (see Discussion).

Preferred Codons in C. grandiflora

Preferred codons have not previously been characterized in *C. grandiflora*. Using the multiallele model, we can define preferred codons as those whose γ value is not significantly different from zero (note that the codon with the highest fitness among those encoding the same amino acid has a γ value of zero). Given the lack of selective difference between codons for some multifold amino acids (Arg, Pro, and Gly; fig. 2), we restricted our attention to those in which γ_{md} differs significantly from zero. We compared the preferred codons in *C. grandiflora* identified by this method with those in *A. thaliana* (Chiapello et al. 1998). Generally, this finds the same preferred codons as in *A. thaliana* (fig. 3), which may be attributable to the fact that *C. grandiflora* is closely related to *A. thaliana*, with a mean synonymous site divergence of 27.7% and a divergence time estimated to be 10–18 My (Koch et al. 2000; Beilstein et al. 2010). However, the preferred codons of *Silene latifolia*, another plant species, which diverged from *A. thaliana* much longer ago (synonymous site divergence is saturated), are also nearly identical to those in *A. thaliana* (Qiu et al. 2010). The conservation of the identity of preferred codons at these different evolutionary time scales may therefore lend further support to the suggestion that the mechanisms underlying selection on codon bias may be quite conserved between plant species (Qiu et al. 2010).

The Mutation Matrix

The multiallele model also produces estimates for the mutation rates between nucleotides for the three species where the L_0 model cannot be excluded. To make a more complete cross-species comparison, we have also implemented

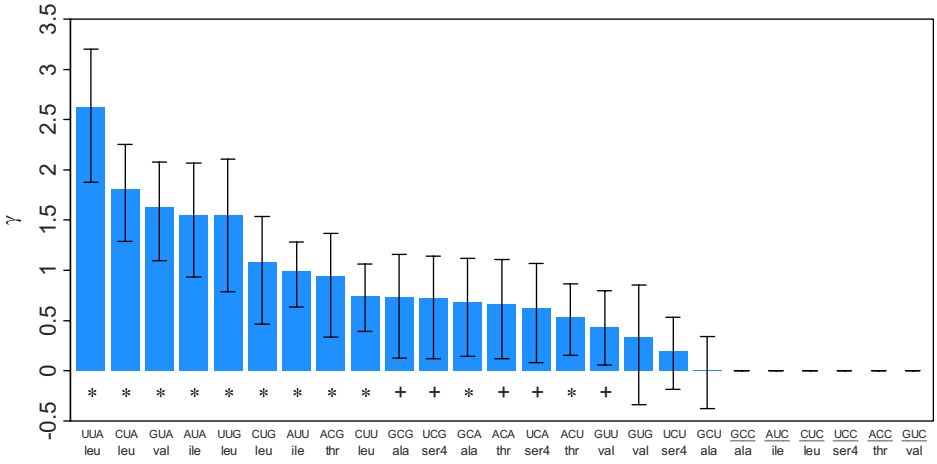

Fig. 3.—Estimates of γ for the codons encoding the multifold amino acids with significant γ_{md} values in *Capsella grandiflora*. The underlined codons are those used as wild type in the analysis (i.e., with the selection coefficient γ fixed at zero). A positive γ value therefore means that the focal codon is selectively unpreferred compared with the wild type (note that this comparison is meaningful only among codons encoding the same amino acid). The underlined codons are also the preferred codons found in *Arabidopsis thaliana* (Chiapello et al. 1998). γ values that are significantly different from zero at $P < 0.01$ and $0.01 < P < 0.05$ under a likelihood ratio test with 1 df are indicated by * and +, respectively.

a nonequilibrium version of the multiallele model (see Materials and Methods) in order to infer mutation rates in *A. thaliana*. Because this nonequilibrium model is computationally extremely demanding, instead of analyzing protein-coding data as in the other three species, we applied it only to the intron polymorphism data set. In addition, mutation matrices were also inferred using the method of Nielsen et al. (2007) for both *Arabidopsis* species. Instead of extracting information from polymorphism data, this method analyzes substitutions between species. Therefore, the two approaches should provide partially independent estimates of the mutation rates.

Encouragingly, the mutation rate estimates obtained by the multiallele model and the model of Nielsen et al. (2007) are very similar for both *A. thaliana* and *A. lyrata*, suggesting that the estimates are reliable (in fig. 4*a*, for *A. thaliana*, $R^2 = 0.981$, $P = 0.0001$; $\tau_{Kendall} = 1$, $P = 0.0085$, and in fig. 4*b*, for *A. lyrata*, $R^2 = 0.928$, $P = 0.002$; $\tau_{Kendall} = 0.733$, $P = 0.06$). Figure 4*c* compares the mutation rate estimates obtained by the multiallele model for all four species (again, due to computational constraints, no between-species comparisons were made statistically using the multiallele model). The relative mutation rates seem to indicate that the changes in mating system have not led to major changes in the mutation machinery. The biggest discrepancies are seen in *C. rubella*. However, as stated above, this is not unexpected, given the lack of information in the data from this species and the complications caused by a possible recent bottleneck (see Discussion). Considering the similarity in the mutation pattern and the fact that the SFS, from which our conclusion regarding the effect of changes in mating system on selection is derived, is only slightly affected by mutational biases (Zeng 2010), we suggest that the differences in codon bias in the different species described above are unlikely due to shifts in mutation patterns.

In *A. thaliana*, mutation rates have also been estimated by a mutation accumulation study (Ossowski et al. 2009). Our estimates using the two model-based methods yield relative mutation rates between different bases in qualitative agreement with those from the mutation accumulation experiment, but the absolute rates are quantitatively different (fig. 4*a*). For instance, all three methods suggest that transitions are more common than transversions and that C:G → T:A mutations occur most frequently. However, the mutation accumulation experiment predicts a much higher proportion of C:G → T:A mutations than either model-based method. This discrepancy is unexpected, because, when applied to *Drosophila* data sets, both model-based methods yielded estimates very similar to those from mutation accumulation experiments (Nielsen et al. 2007; Keightley et al. 2009; Zeng 2010). One possible explanation is that most of the substitutions analyzed in the mutation accumulation study are in intergenic regions, which are likely to include transposable elements and repeated sequences (Ossowski et al. 2009), which in plants are densely methylated (Finnegan et al. 1998; Zhang et al. 2006; Law and Jacobsen 2010). In contrast, our estimates are based on genic regions, where DNA methylation is much less common (Zhang et al. 2006). Hence, the difference shown in figure 4*a* may be caused by different levels of methylation in these two kinds of regions.

Effect of BGC

The intron data in *A. thaliana* and *A. lyrata* (table 1) also allow us to test for the occurrence of BGC caused by biases in the DNA repair machinery (Marais 2003). BGC is expected to operate when recombination occurs in a heterozygote and will cause

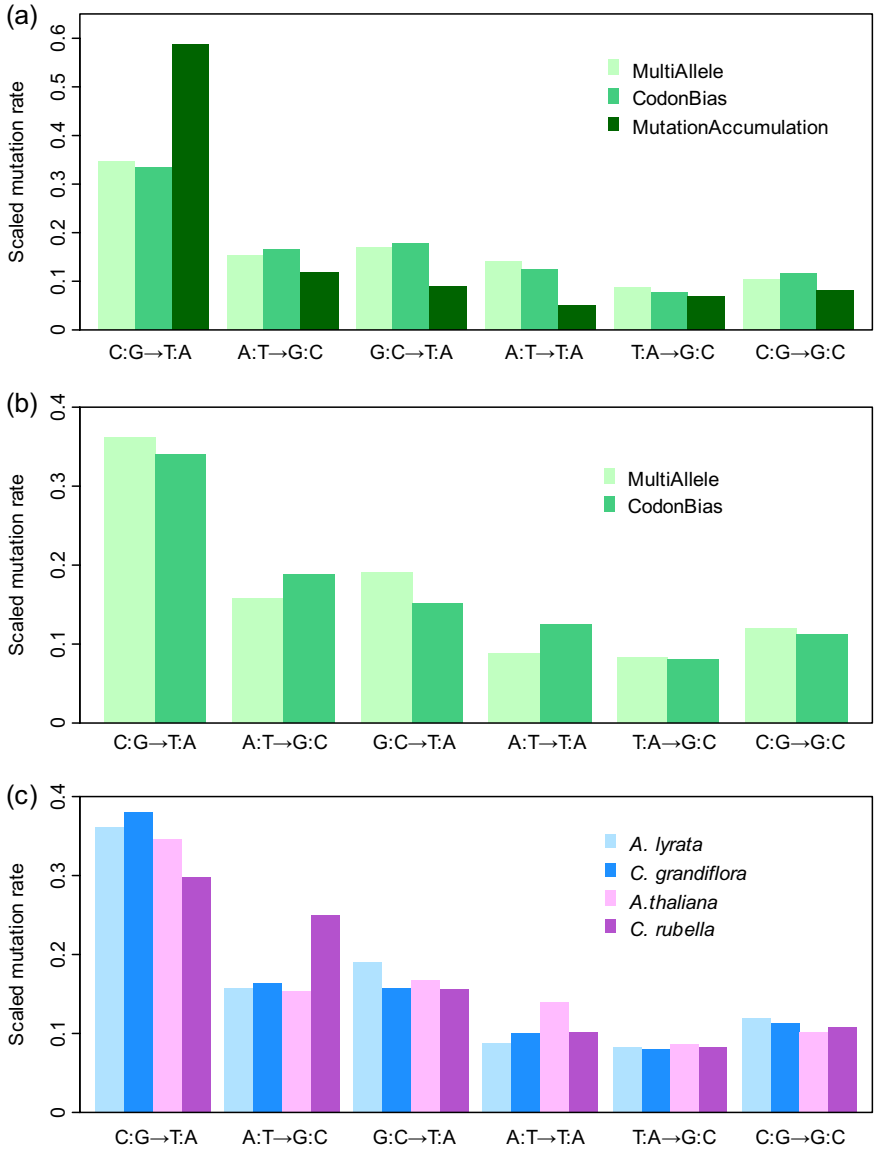

Fig. 4.—Mutation rate estimates. The *x* axis shows the six types of mutations, with the first two being transitions and the others being transversions. The mutation rates are scaled, so that the six estimates obtained from a species add up to unity. (*a*) Mutation rates inferred in *Arabidopsis thaliana* by the two model-based approaches described in the text and the mutation accumulation experiment of Ossowski et al. (2009). (*b*) Mutation rates inferred in *Arabidopsis lyrata* by the two model-based approaches. (*c*) Comparison of the mutation rates of the four plant species inferred by the multiallele approach.

an increase in the fixation probability of G/C nucleotides (Marais 2003). In a finite population, this effect is similar to natural selection favoring G/C over A/T nucleotides (Nagylaki 1983). We used the extended multiallele model to obtain maximum likelihood estimates of the γ values reflecting this process, for the four nucleotides A, T, C, and G in *A. thaliana*. None of the estimated values is significantly different from zero ($P > 0.05$; $\gamma_A = -0.003$, $\gamma_T = -0.026$, $\gamma_C = 0$ and $\gamma_G = -0.118$; the nucleotide with a negative γ value is selectively better than that with $\gamma = 0$). Thus, in agreement with a previous analysis (Marais et al. 2004), this result suggests that BGC is likely to be weak in introns in *A. thaliana*, probably because of the low effective recombination rate due to high homozygosity and the low N_e expected in this highly selfing species (Nordborg and Donnelly 1997; Nordborg 2000).

We also analyzed the intron data in *A. lyrata* in order to understand the general importance of BGC in *Arabidopsis* and to see whether the lack of evidence in *A. thaliana* is due entirely to selfing. Because of the small size of this data set, we employed the original two-allele model of Zeng and Charlesworth (2009), treating A/T as one allele and G/C as the other. The estimated scaled selection coefficient for the

difference between the two alleles is −0.16, meaning that A/T is better than G/C, although this difference is not statistically significant ($P > 0.05$). Thus, BGC also seems to be undetectable in these introns in *A. lyrata*. Additionally, we also failed to detect BGC in our previous analysis of noncoding data from an outcrossing plant *S. latifolia* (Qiu et al. 2010). These results suggest that BGC may be weak in many plants (but see Muyle et al. 2011, forthcoming). However, the importance of BGC in plants deserves more research in the future with the aid of larger data sets from additional outcrossing species and more careful treatment of the possibility that some intron sites may be functionally important (Ko et al. 1998; Rose 2002; Andolfatto 2005). Understanding the effects of BGC, if any, is important for a thorough understanding of patterns of molecular evolution in plants; in other species, such as humans, BGC has significant effects on patterns of molecular evolution (Berglund et al. 2009; Galtier et al. 2009).

Discussion

Demographic Changes

Our finding of clear evidence of recent population expansion in *A. thaliana* supports previous studies that proposed a size increase in the *A. thaliana* population after the last glaciation (Sharbel et al. 2000; Nordborg et al. 2002, 2005; Beck et al. 2008). However, the likelihood surface contains a plateau (data not shown) and we were unable to obtain confidence intervals for g and τ. Fortunately, our conclusions about selection are unaffected by the uncertainties about the details of the expansion history because estimates of the mutation and selection parameters are virtually the same for different values of g and τ on the plateau of the likelihood surface (data not shown).

On the other hand, the fact that the *A. lyrata* data do not reject the equilibrium model seems to be in contradiction to a previous study of sequence polymorphism in this species, which reported evidence of recent population bottlenecks and attributed this to postglacial recolonization (Ross-Ibarra et al. 2008). In that study, the support for bottlenecks was gathered from polymorphism patterns within populations, which probably reflect demographic history at the level of local populations (possibly including selective processes such as local selective sweeps and background selection). In contrast, we adopted a "scattered" sampling strategy, which probably reflects species-wide dynamics and has the additional advantage of reducing the influence of population structure (Wakeley 1999; Wakeley and Aliacar 2001). Hence, the difference between the two studies may be a consequence of differences in demographic history between local populations versus the species as a whole. However, the power of our analysis may be limited by the fact that the *A. lyrata* data set is small and was compiled from several different studies. In the future, the demographic history of *A. lyrata* should be revisited with more data.

Consistent with the results of Foxe et al. (2009), our model also suggests that diversity patterns in *C. grandiflora* are not incompatible with a model with a stable population size. The situation is, however, more complex for *C. rubella*. Although our model does not reject the equilibrium model, other researchers have hypothesized that it may have experienced an extreme bottleneck event about 30,000 years ago when it evolved a high self-fertilization rate (Guo et al. 2009). In particular, extant populations of *C. rubella* may have originated from a single ancestral *C. grandflora* genotype that lost self-incompatibility (Guo et al. 2009). Such an extreme bottleneck would have substantially reduced the level of diversity in *C. rubella*, which is estimated to be only a quarter of the level in *C. grandflora* (table 1). The *C. rubella* sequence data may thus provide little information about the details of the bottleneck event. Furthermore, if the bottleneck was very brief, and the population size has afterward remained very small but fairly stable, the rate of genetic drift would be high. Polymorphic sites inherited from the ancestral population would therefore often become fixed, and consequently, extant polymorphism patterns could resemble those in a stable population. In support of this scenario, Tajima's D value calculated from the synonymous sites in the *C. rubella* sample is close to zero at 0.11, only slightly higher than the average of 0.04 in *C. grandiflora*. Thus, the primary signal of a population bottleneck in *C. rubella* is obtained by examining patterns of shared and unique polymorphisms with *C. grandiflora* (Foxe et al. 2009; Guo et al. 2009), rather than by departures from an equilibrium frequency spectrum in *C. rubella*.

Selfing Species Have Weaker Selection on Codon Usage

Our comparisons of codon bias in both pairs of species suggest that the intensity of selection on synonymous codons has been reduced in the selfing species compared with their outcrossing sibling species. Among the potential causes of the difference, our results show that differences in gene expression between species are unlikely, and, in *Arabidopsis*, there is no evidence that shifts in BGC are playing a role. For the two *Arabidopsis* species, the shift is also unlikely to be a result of changes in mutational parameters because the two different model-based approaches, which obtain estimates using partially nonoverlapping data (within-species polymorphisms vs. between-species substitutions), produce very similar estimates of the mutation matrix. Overall, therefore, the results suggest that a reduction in N_e is a more likely explanation of the observation.

The mutation process could, however, be important in the *Capsella* comparison because of the noticeable difference in the mutation pattern (fig. 4*c*). If the loss of self-incompatibility in *C. rubella* was indeed accompanied by a change in the mutation process, the nonequilibrium dynamics afterward could produce polymorphism patterns that are difficult to distinguish from those under a model with a recent change in

population size (Zeng and Charlesworth 2010a) and may therefore bias the estimates of selection coefficients (Zeng and Charlesworth 2009). However, the fact that the *C. rubella* data do not exclude the equilibrium model as a sufficient explanation of the data suggests that the effects of recent changes in the mutation process, if any, are probably weak and may have a limited impact on the inferences. Furthermore, the observation that diversity is greatly reduced in *C. rubella* (table 1) implies that N_e is lowered. Following a population size contraction, estimates drawn from the equilibrium model tend to overestimate the actual intensity of selection (see figure 4 of Zeng and Charlesworth 2009). Our estimate of a reduction in the efficacy of selection in *C. rubella* may therefore be conservative. However, as discussed above, a clearer resolution of the problem requires more data, especially from noncoding regions, because comparing between synonymous sites and noncoding sites can provide useful information (Kern and Begun 2005; Zeng and Charlesworth 2010a).

In summary, we suggest that, as predicted by population genetic theory, a reduction in N_e in self-fertilizing species seems to be a parsimonious and coherent explanation of the observations of reductions in the efficacy of selection and the observation of lower levels of silent diversity observed in *A. thaliana* and *C. rubella*.

As mentioned in the Introduction, previous studies failed to detect a selective difference by investigating differences in substitution patterns and genome composition (e.g., Wright et al. 2002, 2007; Foxe et al. 2008). The disagreement may be attributable to a difference in power, due to both the power of the methods used in these studies, and also to the small amount of data then available. Moreover, both substitution and the evolution of genome composition are mutation-limited processes and change slowly, so the amount of time since self-fertilization evolved may not be long enough for a clear-cut signal to emerge. Self-compatibility in *A. thaliana* may indeed have evolved quite recently (estimated to be less than 413,000 years ago; Bechsgaard et al. 2006); for European populations, the self-compatibility locus region is uniformly of one haplotype, though mutations have accumulated among sublineages since its spread (Tsuchimatsu et al. 2010). In contrast, polymorphism patterns in population samples are determined by the evolutionary dynamics of the species in the last $\sim N_e$ generations and should therefore be heavily influenced by the existence of self-fertilization. Hence, polymorphism data may offer higher power to detect the effect of selfing on the efficacy of selection. Silent site polymorphisms may be particularly informative because they are probably under weak selection, and weakly selected sites are expected to be most sensitive to changes in N_e (McVean and Charlesworth 1999). This may explain the lack of a significant difference between *A. thaliana* and *A. lyrata* in the intensity of selection on nonsynonymous sites reported by Foxe et al. (2008) because diversity patterns at these sites may be less responsive to changes in N_e (note that the lack of difference may also be attributable to the fact that Foxe et al. investigated within-population diversity patterns and did not use a scattered sampling strategy; see above).

Previous analyses of synonymous variants often classified synonymous codons simply as preferred or unpreferred and assumed that substitutions between these two classes were selected, whereas within-class substitutions were neutral (Wright et al. 2002). However, as shown in figure 3 (see also Zeng 2010), this is an oversimplification, and substitutions between most codons are probably selected. This unrealistic assumption may bias estimates of selection. To test whether this bias occurs, we reanalyzed the *A. thaliana* data set using the original two-allele model of Zeng and Charlesworth (2009). We treated the preferred codons identified by Chiapello et al. (1998) as one allele and all the other codons as the other allele. Just as with the extended model specified by equations (1) and (2), this method found significant evidence of recent population expansion ($P = 9.42 \times 10^{-5}$). However, the maximum likelihood estimate of γ is 0.28, less than half the value of γ_{Phe} in figure 1. The same occurred when we reanalyzed the *C. grandiflora* data. The γ value between preferred and unpreferred codons was estimated to be 0.55, much lower than the largest γ_{md} value of 2.6 for Leu in figure 2.

Varying Intensity of Selection on Codon Usage in Plants

It is unexpected that, even in the two outcrossing species, where natural selection is expected to be most effective, significant selective differences between codons are detectable only for some amino acids (figs. 1 and 2). This may, however, offer an explanation, in addition to the effect of selfing, of the observation that codon bias in *A. thaliana* is much lower than in many other eukaryote species such as Drosophila (Duret and Mouchiroud 1999; dos Reis and Wernisch 2009). For instance, taking the results in supplementary figure 1, Supplementary Material online, at face value, many of the ten 2-fold amino acids seem to have comparable γ values between *A. lyrata*, *C. grandiflora*, and *D. melanogaster*, whereas others, especially those with G/A-ending codons, have considerably lower γ in the two plants. Hence, it is possible that the lower overall extent of codon bias observed in plants is due to reduced selection on some synonymous codons.

Some information about the variation in selection intensity can be gained from the wobble rules of tRNA binding. In several eukaryotic genomes, including *A. thaliana*, it has been found that GNN (where N stands for any of the four nucleotides) tRNAs can pair with both C-ending and T-ending codons, whereas TNN and CNN tRNAs can each pair only with A-ending and G-ending codons, respectively (Percudani 2001; Wright et al. 2004). In *A. thaliana*, only a GNN tRNA exists for each of the 2-fold amino acids with

T/C-ending codons, whereas there are two tRNAs (TNN and CNN) for each of the 2-fold amino acids with A/G-ending codons. Stronger selection on the former set of amino acids relative to the latter may therefore be attributable to competition for a single tRNA type between the T/C-ending codons during translation and to the fact that the tRNA has higher affinity to the C-ending codon due to the Watson and Crick base-matching rule. However, this hypothesis is unlikely to hold universally. For instance, *D. melanogaster* shows a quantitatively very different pattern of varying selection for different amino acids (supplementary fig. 1, Supplementary Material online; McVean and Vieira 2001; Zeng 2010). Further investigations are needed to understand the mechanism that leads to different patterns of codon bias among different organisms.

Supplementary Material

Supplementary figure 1 and tables 1 and 2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank John Paul Foxe and Khaled Hazzouri for assistance with PCR and sequence editing. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (http://www.ecdf.ed.ac.uk/), which is partially supported by the e-Science Data, Information and Knowledge Transformation (eDIKT) initiative (http://www.edikt.org.uk). This work was supported by the National Natural Science Foundation of China (30730008, 40976081 to Suhua Shi); the National Basic Research Program of China (2007CB815701 to Suhua Shi); the National S&T Major Project of China (2009ZX08010-017B to Suhua Shi); the Chang Hung-Ta Science Foundation of Sun Yat-Sen University; a Biomedical Personal Research Fellowship from the Royal Society of Edinburgh and the Caledonian Research Foundation to K.Z.; a Discovery grant from the Natural Sciences and Engineering Research Council to S.W.; and an Early Researcher Award from the Government of Ontario to S.W.

Literature Cited

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

Bechsgaard JS, Castric V, Charlesworth D, Vekemans X, Schierup MH. 2006. The transition to self-compatibility in *Arabidopsis thaliana* and evolution within S-haplotypes over 10 Myr. Mol Biol Evol. 23: 1741–1750.

Beck JB, Schmuths H, Schaal BA. 2008. Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. Mol Ecol. 17:902–915.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 107:18724–18728.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS Biol. 7:e26.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303.

Chiapello H, Lisacek F, Caboche M, Henaut A. 1998. Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. Gene 209:GC1–GC38.

Conway DJ, et al. 1999. High recombination rate in natural populations of *Plasmodium falciparum*. Proc Natl Acad Sci U S A. 96:4506–4511.

dos Reis M, Wernisch L. 2009. Estimating translational selection in eukaryotic genomes. Mol Biol Evol. 26:451–461.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc Natl Acad Sci U S A. 96:4482–4487.

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.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Finnegan EJ, Genger RK, Peacock WJ, Dennis ES. 1998. DNA methylation in plants. Annu Rev Plant Physiol Plant Mol Biol. 49:223–247.

Foxe JP, et al. 2008. Selection on amino acid substitutions in *Arabidopsis*. Mol Biol Evol. 25:1375–1383.

Foxe JP, et al. 2009. Recent speciation associated with the evolution of selfing in *Capsella*. Proc Natl Acad Sci U S A. 106:5241–5245.

Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet. 25:1–5.

Garcia-Dorado A, Caballero A. 2000. On the average coefficient of dominance of deleterious spontaneous mutations. Genetics 155:1991–2001.

Guo YL, et al. 2009. Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. Proc Natl Acad Sci U S A. 106:5246–5251.

Haddrill PR, Zeng K, Charlesworth B. 2011. Determinants of synonymous and nonsynonymous variability in three species of *Drosophila*. Mol Biol Evol. 28:1731–1743.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu Rev Genet. 42:287–299.

Ingvarsson PK. 2002. A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. Evolution 56:2368–2373.

Kaiser VB, Charlesworth B. 2009. The effects of deleterious mutations on evolution in non-recombining genomes. Trends Genet. 25:9–12.

Kawabe A, Forrest A, Wright SI, Charlesworth D. 2008. High DNA sequence diversity in pericentromeric genes of the plant *Arabidopsis lyrata*. Genetics 179:985–995.

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.

Keightley PD, et al. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res. 19:1195–1201.

Kern AD, Begun DJ. 2005. Patterns of polymorphism and divergence from noncoding sequences of *Drosophila melanogaster* and *D. simulans*: evidence for nonequilibrium processes. Mol Biol Evol. 22:51–62.

Ko CH, Brendel V, Taylor RD, Walbot V. 1998. U-richness is a defining feature of plant introns and may function as an intron recognition signal in maize. Plant Mol Biol. 36:573–583.
Koch MA, 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.
Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 11:204–220.
Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2:e166.
Li WH. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol. 24:337–345.
Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. 19:330–338.
Marais G, Charlesworth B, Wright SI. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biol. 5:R45.
McVean GA, Charlesworth B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet Res. 74:145–158.
McVean GA, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics 157:245–257.
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equations of state calculations by fast computing machine. J Chem Phys. 21:1087–1091.
Meyers BC, et al. 2004. *Arabidopsis* MPSS. An online resource for quantitative expression analysis. Plant Physiol. 135:801–813.
Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glemin S. Forthcoming 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). Mol Biol Evol.
Nagylaki T. 1983. Evolution of a finite population under gene conversion. Proc Natl Acad Sci U S A. 80:6278–6281.
Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. Mol Biol Evol. 24:228–235.
Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics 154:923–929.
Nordborg M, Donnelly P. 1997. The coalescent process with selfing. Genetics 146:1185–1195.
Nordborg M, et al. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet. 30:190–193.
Nordborg M, et al. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3:e196.
Ossowski S, et al. 2009. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327:92–94.
Pannell JR, Charlesworth B. 2000. Effects of metapopulation processes on measures of genetic diversity. Philos Trans R Soc Lond B Biol Sci. 355:1851–1864.
Percudani R. 2001. Restricted wobble rules for eukaryotic genomes. Trends Genet. 17:133–135.
Pollak E. 1987. On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics. 117:353–360.
Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. Numerical recipes in C: the art of scientific computing. Cambridge: Cambridge University Press.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.
Qiu S, Bergero R, Zeng K, Charlesworth D. 2010. Patterns of codon usage bias in *Silene latifolia*. Mol Biol Evol. 28:771–780.
Ramos-Onsins SE, Stranger BE, Mitchell-Olds T, Aguade M. 2004. Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. Genetics 166:373–388.
Rose AB. 2002. Requirements for intron-mediated enhancement of gene expression in *Arabidopsis*. RNA. 8:1444–1453.
Ross-Ibarra J, et al. 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. PLoS One. 3:e2411.
Shapiro JA, et al. 2007. Adaptive genic evolution in the *Drosophila* genomes. Proc Natl Acad Sci U S A. 104:2271–2276.
Sharbel TF, Haubold B, Mitchell-Olds T. 2000. Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol Ecol. 9:2109–2118.
Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol Biol Evol. 27:1813–1821.
Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res. 23:23–35.
Tsuchimatsu T, et al. 2010. Evolution of self-compatibility in *Arabidopsis* by a mutation in the male specificity gene. Nature 464:1342–1346.
Wakeley J. 1999. Nonequilibrium migration in human history. Genetics 153:1863–1871.
Wakeley J, Aliacar N. 2001. Gene genealogies in a metapopulation. Genetics 159:893–905.
Williamson SH, et al. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. 102:7882–7887.
Wright S. 1969. Evolution and the genetics of populations. Chicago (IL): University of Chicago Press.
Wright SI, Iorgovan G, Misra S, Mokhtari M. 2007. Neutral evolution of synonymous base composition in the Brassicaceae. J Mol Evol. 64:136–141.
Wright SI, Lauga B, Charlesworth D. 2002. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Mol Biol Evol. 19:1407–1420.
Wright SI, Lauga B, Charlesworth D. 2003. Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Mol Ecol. 12:1247–1263.
Wright SI, Yau CB, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol Biol Evol. 21:1719–1726.
Wright SI, et al. 2006. Testing for effects of recombination rate on nucleotide diversity in natural populations of *Arabidopsis lyrata*. Genetics 174:1421–1430.
Zeng K. 2010. A simple multiallele model and its application to identifying preferred-unpreferred codons using polymorphism data. Mol Biol Evol. 27:1327–1337.
Zeng K, Charlesworth B. 2009. Estimating selection intensity on synonymous codon usage in a nonequilibrium population. Genetics 183:651–662, 651SI-623SI.
Zeng K, Charlesworth B. 2010a. Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. J Mol Evol. 70:116–128.
Zeng K, Charlesworth B. 2010b. The effects of demography and linkage on the estimation of selection and mutation parameters. Genetics 186:1411–1424.
Zhang X, et al. 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell 126:1189–1201.

Associate editor: Brandon Gaut