Higher Rates of Protein Evolution in the Self-Fertilizing Plant *Arabidopsis thaliana* than in the Out-Crossers *Arabidopsis lyrata and Arabidopsis halleri*

Bryan L. Payne and David Alvarez-Ponce*

Department of Biology, University of Nevada, Reno

*Corresponding author: E-mail: dap@unr.edu.

Accepted: March 5, 2018

Abstract

The common transition from out-crossing to self-fertilization in plants decreases effective population size. This is expected to result in a reduced efficacy of natural selection and in increased rates of protein evolution in selfing plants compared with their outcrossing congeners. Prior analyses, based on a very limited number of genes, detected no differences between the rates of protein evolution in the selfing *Arabidopsis thaliana* compared with the out-crosser *Arabidopsis lyrata*. Here, we reevaluate this trend using the complete genomes of *A. thaliana*, *A. lyrata*, *Arabidopsis halleri*, and the outgroups *Capsella rubella* and *Thellungiella parvula*. Our analyses indicate slightly but measurably higher nonsynonymous divergences (d_N), synonymous divergences (d_S) and d_N/d_S ratios in *A. thaliana* compared with the other *Arabidopsis* species, indicating that purifying selection is indeed less efficacious in *A. thaliana*.

Key words: rates of evolution, d_N/d_S, selfing, out-crossing, *Arabidopsis*.

In plants, the transition from out-crossing to self-fertilization is quite common, and is generally seen as a dead-end due to accumulation of deleterious mutations (Stebbins 1957). Population genetics theory predicts that selfing organisms will have a lower effective population size (N_e) than their out-crossing congeners with the same population size (Pollak 1987). Reduced N_e is expected to result in a reduced efficacy of natural selection (Charlesworth et al. 1993; Charlesworth and Wright 2001), thus allowing the fixation of slightly deleterious mutations (Ohta 1973). As a result, selfing organisms are expected to exhibit accelerated rates of protein evolution (Kimura 1983; Charlesworth and Wright 2001; Glémin 2007) and less codon usage bias (Qiu et al. 2011). These predictions are supported by some empirical evidence: natural selection is reduced in selfing species of the family Triticeae (Haudry et al. 2008; Escobar et al. 2010) and the genera *Capsella* (Qiu et al. 2011; Slotte et al. 2013), *Eichhornia* (Ness et al. 2012), *Collinsia* (Hazzouri et al. 2013), and *Mimulus* (Brandvain et al. 2014) (for review, see Hough et al. [2013] and Shimizu and Tsuchimatsu [2015]). In addition, an analysis of polymorphism data for a number of plant species revealed a weak increased in the nonsynonymous to synonymous polymorphism ratio (π_a/π_s) of selfers (Glémin et al. 2006).

The plant *Arabidopsis thaliana* is thought to have shifted to self-fertilization 150,000–1,000,000 years ago (Charlesworth and Vekemans 2005; Bechsgaard et al. 2006; Tang et al. 2007; Tsuchimatsu et al. 2010; Shimizu and Tsuchimatsu 2015; Durvasula et al. 2017; Tsuchimatsu et al. 2017). In agreement with the predicted reduction in the efficacy of natural selection, this species exhibits less codon bias than the out-crosser *Arabidopsis lyrata* (Qiu et al. 2011). However, analysis of 16 genes did not detect significantly higher rates of protein evolution in *A. thaliana* compared with *A. lyrata* (Wright et al. 2002). In addition, comparison of 13 pairs of orthologous genes in these two species revealed no differences in the ratios of nonsynonymous to synonymous polymorphisms or in the ratios of nonsynonymous to synonymous fixations (Foxe et al. 2008). A comparison of 675 *A. thaliana* and 73 *A. lyrata* nonorthologous genes found higher ratios of nonsynonymous to synonymous polymorphisms and higher ratios of nonsynonymous to synonymous fixations in *A. thaliana* (Foxe et al. 2008); however, these results may have been affected by biases in the data set—for example, seven of the *A. lyrata* genes were chosen due to their high levels of expression, and highly expressed genes tend to evolve under strong purifying selection (Pál et al. 2001; Drummond et al. 2005).

These analyses, in any case, were limited by the very small amount of genomic information available at the time. Here, we revisit the prediction that *A. thaliana* should exhibit faster rates of protein evolution than *A. lyrata* or than *Arabidopsis halleri* taking advantage of the now completely sequenced genomes of *A. thaliana* (The Arabidopsis Genome Initiative 2000), *A. lyrata* (Hu et al. 2011), *A. halleri* (Briskine et al. 2017) and the outgroup *Capsella rubella* (Slotte et al. 2013). *A. thaliana* diverged 6–13 Ma ago from the *A. lyrata*/*A. halleri* clade (Beilstein et al. 2010; Hohmann et al. 2015) and 8–14 Ma ago from *C. rubella* (Koch and Kiefer 2005; Hohmann et al. 2015) (fig. 1).

For each *C. rubella* gene, we identified the most likely orthologs in *A. thaliana* and *A. lyrata*. For each of the 18,107 identified trios, protein sequences were aligned, and the resulting alignments were used to guide the alignment of the corresponding coding sequences (CDSs). To reduce the impact of annotation errors, we removed all alignments for which >5% of positions included gaps. For each of the resulting 12,994 alignments, PAML (free-ratios model; Yang 2007) was used to estimate the nonsynonymous divergence (d_N), synonymous divergence (d_S) and the nonsynonymous to synonymous divergence ratio ($\omega = d_N/d_S$) in each of the branches of the phylogeny (fig. 1). The ratio d_N/d_S is expected to be lower than 1 when nonsynonymous mutations are under purifying selection (with values closer to 0 indicating stronger selection), equal to 1 when protein sequences evolve neutrally, and higher than 1 for genes under positive selection (for review, see Alvarez-Ponce [2014]).

In the *A. thaliana* branch, the median of the values estimated by the free-ratios model were $d_N = 0.0108$, $d_S = 0.0757$, and $d_N/d_S = 0.1427$, and the mean values were $d_N = 0.0133$, $d_S = 0.0805$, and $d_N/d_S = 0.1865$ In the *A. lyrata* branch, the median values were $d_N = 0.0085$, $d_S = 0.0612$, and $d_N/d_S = 0.1389$, and the mean values were $d_N = 0.0107$, $d_S = 0.0667$, and $d_N/d_S = 0.1880$ (supplementary table S1, Supplementary Material online and fig. 2). A Mann–Whitney *U* test showed significant differences in the d_N ($P = 1.964 \times 10^{-119}$), d_S ($P < 10^{-300}$), and d_N/d_S ($P = 0.0127$) of both species. In 8,572 of the cases, d_N was higher in *A. thaliana* than in *A. lyrata*, and in 4,396 of the cases d_N was higher in *A. lyrata*, indicating that rates of protein sequence evolution are often higher in *A. thaliana* (binomial test, $P = 1.20 \times 10^{-324}$). In 8,938 of the cases, d_S was higher in *A. thaliana*, and in 4,055 of the cases d_S was higher in *A. lyrata*, indicating faster rates of evolution of synonymous sites in *A. thaliana* (binomial test, $P = 4.94 \times 10^{-324}$); these results are consistent with prior studies reporting higher mutation rates in *A. thaliana* (Yang et al. 2013). In 6,625 of the cases, d_N/d_S was higher in *A. thaliana*, and in 6,161 of the cases d_N/d_S was higher in *A. lyrata*, indicating that purifying selection on protein sequences is often less effective in *A. thaliana* (binomial test, $P = 4.22 \times 10^{-5}$). Differences were stronger when analyses were restricted to genes that are highly expressed in *A. thaliana* (supplementary table S2, Supplementary Material online).

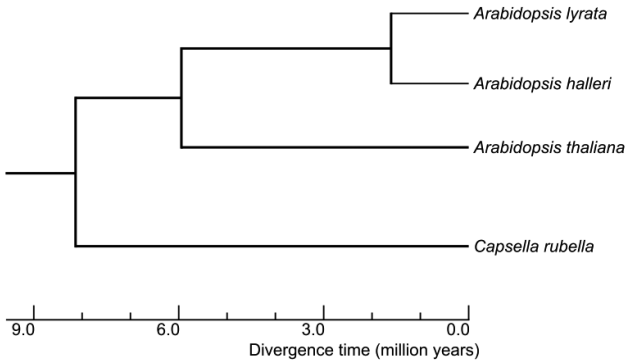

Fig. 1.—Phylogenetic relationships among the species used in the current study. The tree topology and divergence times were obtained from Hohmann et al. (2015).

For each alignment, Tajima's relative rate test (Tajima 1993) was used to contrast whether the number of substitutions accumulated in *A. thaliana* and *A. lyrata* was significantly different. Statistically significant differences were detected in 1,363 and 1,333 genes for synonymous and nonsynonymous sites, respectively. Of the 1,363 genes with significant differences in synonymous rates of evolution, there were more unique synonymous changes in *A. thaliana* in 1,222 genes compared with 141 genes where *A. lyrata* had more unique synonymous changes. Of the 1,333 genes with an asymmetry in the number of nonsynonymous sites, *A. thaliana* and *A. lyrata* had more unique changes in 1,077 and 256 cases, respectively (table 1).

For each of the 12,994 alignments, we compared the likelihood of the free-ratios model (in which each of the three branches exhibits an independent d_N/d_S ratio) versus that of a two-ratios model (one d_N/d_S ratio for *A. thaliana* and *A. lyrata*, and another for *C. rubella*). The free-ratios model fitted the data significantly better in 907 of the alignments (likelihood ratio test, $P < 0.05$), indicating that the d_N/d_S ratio is significantly different in *A. thaliana* and *A. lyrata*. In 477 of the 907 cases where the free-ratio model fit better than the two-ratios model, d_N/d_S was higher for *A. thaliana*, and in 430 of the cases d_N/d_S was higher for *A. lyrata*; these numbers were not significantly different from the 50%:50% (453.5:453.5) expected by chance (binomial test, $P = 0.166$).

Given that *A. thaliana* and *A. lyrata* are very closely related, some gene alignments may not contain sufficient information (in terms of number of substitutions) to accurately infer the strength of purifying selection acting on each branch. In order to increase the power of our analyses, we combined all 12,994 alignments into a single concatenome containing 17.8 million base pairs and repeated our analyses on it. The *A. thaliana* lineage exhibited higher d_N, d_S and d_N/d_S values (0.0127, 0.0759, and 0.1671, respectively) (supplementary table S3, Supplementary Material online) than the *A. lyrata*

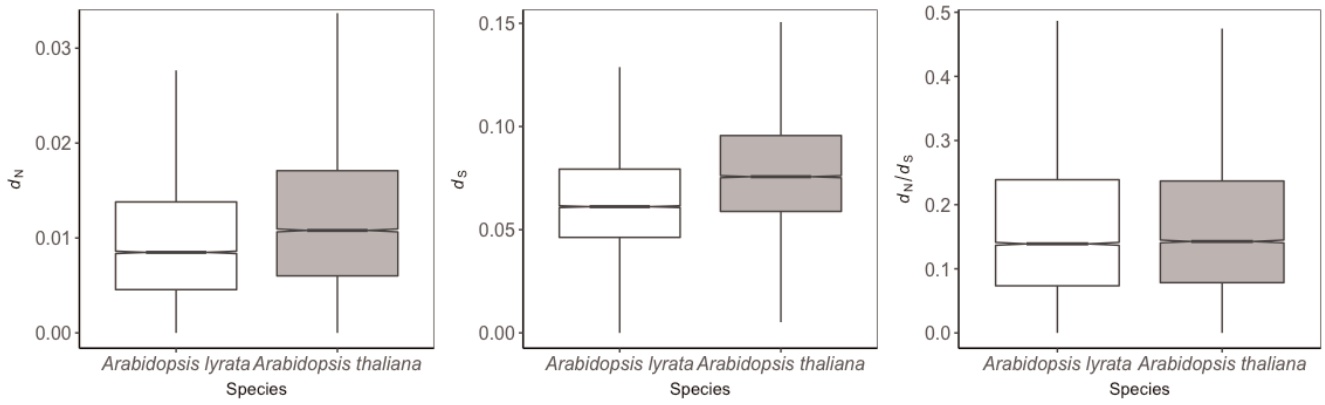

Fig. 2.—Distribution of d_N, d_S, and d_N/d_S values in the *A. thaliana* and *A. lyrata* branches. Values above the 90th percentile are not represented.

Table 1
Tajima's Relative Rate Tests

	All Substitutions	Synonymous Substitutions	Nonsynonymous Substitutions
Unique substitutions in *A. thaliana*	413,215	249,860	163,355
Unique substitutions in *A. lyrata*	343,062	205,266	137,578
Genes where *A. thaliana* had more substitutions	9203	8701	7575
Genes where *A. lyrata* had more substitutions	3207	3424	4102
Genes where $P < 0.05$	2008	1363	1333
Genes where $P < 0.05$ and *A. thaliana* had more substitutions	1824	1222	1077
Genes where $P < 0.05$ and *A. lyrata* had more substitutions	184	141	256
χ^2 value for concatenome	6507.5	4369.4	2208.0
P value for concatenome	$\ll$0.001***	$\ll$0.001***	$\ll$0.001***

***$P < 0.001$.

branch (0.0102, 0.0622, and 0.1644). These values are comparable to the mean values resulting from analysis of individual alignments. The free-ratios model fitted the data significantly better than the two-ratios model ($2\Delta\ell = -10.213$, $P = 0.0014$), indicating that d_N/d_S is significantly higher in *A. thaliana*, even though the differences are small. Tajima's relative rate test (Tajima 1993) revealed an excess of synonymous and nonsynonymous changes in *A. thaliana* compared with *A. lyrata* ($\chi^2 = 4369.4$ and 2207.0, $P \ll 0.001$ and $P \ll 0.001$, respectively). The *A. thaliana* concatenome contained 249,860 synonymous 163,355 nonsynonymous substitutions that were not present in *A. lyrata*, and the *A. lyrata* concatenome contained 205,266 unique synonymous substitutions and 137,578 unique nonsynonymous substitutions.

It is expected that the evolution of selfing in *A. thaliana* may have resulted in pseudogenization of, or at least relaxation of purifying selection in, genes involved in out-crossing. If these represent a sufficiently large number of genes, this effect alone, rather than a reduction of N_e, might conceivably explain the higher average rates of protein evolution observed in *A. thaliana*. To discard this possibility, we repeated our analyses separately for genes of different functional categories. For all 23 KOG categories represented in the data set, the number of genes with higher d_N and d_S values in *A. thaliana* was significantly higher than the number of genes with higher d_N and d_S values in *A. lyrata*. For 19 of the categories, the number of genes for which d_N/d_S was higher in *A. thaliana* was higher than the number of genes for which d_N/d_S was higher in *A. lyrata*. For only three categories there were more genes with a higher d_N/d_S in *A. lyrata* (binomial test, $P = 0.0009$; table 2). These results indicate that the higher d_N, d_S and d_N/d_S values observed in *A. thaliana* represent a generalized trend, not specific to certain functional categories.

Throughout the current work we have reported the comparison of the *A. thaliana* reference genome from the TAIR 10 release, a composite genome from 11 Columbia ecotype (Col-0) individuals, with that of *A. lyrata*, using *C. rubella* as outgroup. Nonetheless, equivalent results were obtained using another 18 *A. thaliana* accessions instead of the reference one (supplementary tables S4 and S5, Supplementary Material online), using *A. halleri* (Briskine et al. 2017) instead of *A. lyrata* (supplementary tables S6–S8, Supplementary Material online) or using the outcrossing and more distantly related *Thellungiella parvula* (Dassanayake et al. 2011) as outgroup instead of the selfing and closely related *C. rubella* (supplementary tables S9 and S10, Supplementary Material online).

Table 2

Analyses of Evolutionary Rates in Different KOG Categories

Category[a]	Genes with Higher d_N in *A. thaliana*	Genes with Higher d_N in *A. lyrata*	Genes with Higher d_S in *A. thaliana*	Genes with Higher d_S in *A. lyrata*	Genes with Higher d_N/d_S in in *A. thaliana*	Genes with Higher d_N/d_S in in *A. lyrata*	d_N *P* Value[b]	d_S *P* Value[b]	d_N/d_S *P* Value[b]
A	245	99	248	97	175	163	2.03×10^{-15}***	2.2×10^{-16}***	0.550
B	77	32	78	32	62	45	1.94×10^{-5}***	1.36×10^{-5}***	0.122
C	256	138	257	137	209	174	2.89×10^{-9}***	1.53×10^{-9}***	0.082
D	121	50	124	48	85	84	5.64×10^{-8}***	6.13×10^{-9}***	1.000
E	189	113	216	86	152	150	1.44×10^{-5}***	4.63×10^{-14}***	0.954
F	60	32	60	32	53	36	0.005**	0.0046**	0.089
G	500	308	566	245	389	422	1.47×10^{-11}***	$<2.2 \times 10^{-16}$***	0.261
H	116	74	130	60	94	94	0.003**	4.16×10^{-7}***	1.000
I	225	125	243	107	184	166	9.99×10^{-8}***	2.69×10^{-13}***	0.364
J	214	114	201	127	178	135	3.63×10^{-8}***	5.20×10^{-5}***	0.017*
K	697	362	720	344	541	511	$<2.2 \times 10^{-16}$***	$<2.2 \times 10^{-16}$***	0.371
L	166	68	161	73	119	113	1.23×10^{-10}***	8.70×10^{-9}***	0.743
M	104	67	121	50	78	92	0.006**	5.64×10^{-8}***	0.319
O	710	317	743	287	507	494	$<2.2 \times 10 - 16$***	$<2.2 \times 10^{-16}$***	0.704
P	340	167	367	140	262	243	1.22×10^{-14}***	$<2.2 \times 10^{-16}$***	0.423
Q	227	115	241	101	189	153	1.40×10^{-9}***	2.45×10^{-14}***	0.058
S	2344	1213	2405	1156	1784	1751	$<2.2 \times 10^{-16}$***	$<2.2 \times 10^{-16}$***	0.590
T	673	335	736	272	517	479	$<2.2 \times 10^{-16}$***	$<2.2 \times 10^{-16}$***	0.241
U	353	182	389	150	269	241	1.24×10^{-13}***	$<2.2 \times 10^{-16}$***	0.232
V	52	28	54	26	44	35	0.001**	0.002**	0.368
W	48	26	51	23	34	40	0.014*	0.001**	0.561
Y	22	3	19	6	17	8	1.57×10^{-4}***	0.015*	0.108
Z	159	58	155	63	119	92	4.64×10^{-12}***	3.94×10^{-10}***	0.073

[a]Category functions: A, RNA processing and modifications; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control, cell division, chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination, and repair; M, cell wall, cell membrane and envelope biogenesis; O, posttranslational modification; P, inorganic ion transport and metabolism; Q, secondary metabolite biosynthesis; S, function unknown; T, signal transduction; U, intracellular trafficking, secretion, and vesicular transport; V, defense mechanisms; W, extracellular structures; Y, nuclear structure; Z, cytoskeleton (Tatusov et al. 2003).

[b]*P*-values determined using a binomial test comparing the total number of genes where d_N/d_S was higher in *A. thaliana* and in *A. lyrata*.

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

In summary, all our genome-wide analyses converge at showing that, as expected from the reduced N_e due to selfing, proteins evolved faster in *A. thaliana* than in *A. lyrata* or *A. halleri*. Such protein sequence evolution acceleration is likely due to the combination of faster mutation rates in *A. thaliana* (supported by high d_S values and by prior results; Yang et al. 2013) and by a weaker efficacy of natural selection on nonsynonymous mutations (supported by high d_N/d_S ratios). Prior analyses based on a handful of orthologous genes failed to detect differences in d_N and d_N/d_S between *A. thaliana* and *A. lyrata*, most likely because of limited statistical power (Wright et al. 2002; Foxe et al. 2008). Indeed, the differences that we detected are subtle, consistent with the fact that *A. thaliana* has been selfing for a relatively short amount of time (150,000–1,000,000 years; Charlesworth and Vekemans 2005; Bechsgaard et al. 2006; Tang et al. 2007, Durvasula et al 2017) compared with the time of divergence between *A. thaliana* and the *A. lyrata*/*A. halleri* clade (7–13 Myr; Beilstein et al. 2010; Hohmann et al. 2015). Our analyses have compared the patterns of evolution of the *A. thaliana* lineage (the branch connecting *A. thaliana* and the most recent common ancestor of *A. thaliana* and *A. lyrata*) and the *A. lyrata* and *A. halleri* lineages (the branches connecting *A. lyrata* or *A. halleri* and the most recent common ancestor of *A. thaliana* and *A. lyrata*), and plants in the *A. thaliana* lineage have been selfing for only 1–17% of the length of the branch.

In addition to the recent transition to selfing of *A. thaliana*, other scenarios may account for the small magnitude of the differences observed between the rates of protein evolution of *A. thaliana* and *A. lyrata*. First, most proteins are under strong purifying selection in both species, in agreement with prior observations (Wright et al. 2002; Foxe et al. 2008. Yang and Gaut 2011), thus hindering the detection of strong differences. Second, selfing increases homozygosity, thus exposing recessive alleles to selection, which can reduce rates of protein evolution (see Glémin 2007). Last, population genetics analyses indicate that the N_e of *A. lyrata* may have also been reduced within the last 100,000 years (Mattila et al. 2017); this might have increased the rates of protein evolution in this species, thus attenuating the differences between *A. thaliana* and *A. lyrata*.

Finally, it should be noted that the fast rates of protein evolution observed in *A. thaliana* might be due to peculiarities of the biology of this species other than selfing. In particular, *A. thaliana* switched to an annual life history, whereas *A. lyrata* is perennial. Annual plants tend to evolve faster than perennial plants (Smith and Donoghue 2008; Gaut et al. 2011; Lanfear et al. 2013), which might account for the higher rates of synonymous evolution observed in *A. thaliana*. However, annual plants exhibit lower nonsynonymous to synonymous polymorphism ratios (Chen et al. 2017), and thus the annual life history of *A. thaliana* may not explain the observed d_N/d_S ratios observed in this species.

Materials and Methods

For each *C. rubella* gene, the longest encoded protein was chosen for analysis and orthologs in *A. thaliana* and *A. lyrata* were identified using a best reciprocal hit approach (BLASTP, *E*-value $< 10^{-10}$). Only genes for which orthologs could be identified in both *Arabidopsis* species were retained. Trios of orthologous protein sequences were aligned using PRANK v.140603 (Löytynoja and Goldman 2005), and the resulting alignments were used to guide the alignments of the CDSs using an in-house script. Alignments which contained <5% gaps were retained for analyses. For each alignment, the codeml program of PAML v. 4.9 (Yang 2007) was used to estimate d_N, d_S and d_N/d_S in each of the three branches (free-ratios model) and in the *A. thaliana*/*A. lyrata* branch and the *C. rubella* branch separately (two-ratios model). Values of d_N/d_S above ten were removed from mean calculations, in order to prevent the bias introduced by these outliers, which represent artifacts due to the presence of very few mutations in the relevant lineages. The fit of both nested models was compared using a likelihood ratio test, assuming that twice the difference between the log-likelihoods of both models ($2\Delta\ell$) follow a χ^2 distribution with one degree of freedom (Huelsenbeck and Crandall 1997). Tajima's relative rate tests (Tajima 1993) were conducted using in-house scripts. *A. thaliana* genes were classified into different eukaryotic orthologous groups (KOG) categories using the eggNOG database v4.5.1 (Huerta-Cepas et al. 2015). Data for the 18 accessions of *A. thaliana* were obtained from the 1000 genomes project (Gan et al. 2011). *A. thaliana* gene expression data were obtained from Schmid et al. (2005) and processed as in Alvarez-Ponce and Fares (2012). All our alignments and scripts are available upon request.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

DAP dedicates this paper to the memory of Mario A. Fares. We are grateful to Julio Rozas for helpful discussion. This work was supported by funds from the University of Nevada, Reno.

Literature Cited

Alvarez-Ponce D. 2014. Why proteins evolve at different rates: the determinants of proteins' rates of evolution. In: Fares MA, editor. Natural selection: methods and applications. London: CRC press. p. 126–178.

Alvarez-Ponce D, Fares MA. 2012. Evolutionary rate and duplicability in the *Arabidopsis thaliana* protein–protein interaction network. Genome Biol Evol. 4(12):1263–1274.

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(9):1741–1750.

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(43):18724–18728.
Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. 2014. Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. PLoS Genet. 10(6):e1004410.
Briskine RV, et al. 2017. Genome assembly and annotation of *Arabidopsis halleri*, a model for heavy metal hyperaccumulation and evolutionary ecology. Mol Ecol Resour. 17(5):1025–1036.
Charlesworth B, Morgan MT, harlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134(4):1289–1303.
Charlesworth D, Vekemans X. 2005. How and when did *Arabidopsis thaliana* become highly self-fertilising. BioEssays 27(5):472–476.
Charlesworth D, Wright SI. 2001. Breeding systems and genome evolution. Curr Opin Genet Dev. 11(6):685–690.
Chen J, Glémin S, Lascoux M. 2017. genetic diversity and the efficacy of purifying selection across plant and animal species. Mol Biol Evol. 34(6):1417–1428.
Dassanayake M, et al. 2011. The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet. 43(9):913–918.
Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102(40):14338–14343.
Durvasula A, et al. 2017. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 114(20):5213–5218.
Escobar JS, et al. 2010. An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). Evolution 64(10):2855–2872.
Foxe JP, et al. 2008. Selection on amino acid substitutions in *Arabidopsis*. Mol Biol Evol. 25(7):1375–1383.
Gan X, et al. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477(7365):419–423.
Gaut B, Yang L, Takuno S, Eguiarte LE. 2011. The patterns and causes of variation in plant nucleotide substitution rates. Annu Rev Ecol Evol Syst. 42(1):245–266.
Glémin S, Baxin E, Charlesworth D. 2006. Impact of mating systems on patterns of sequence polymorphism in flowering plants. Proc Biol Sci. 273(1604):3011–3019.
Glémin S. 2007. Mating systems and the efficacy of selection at the molecular level. Genetics 177(2):905–916.
Haudry A, et al. 2008. Mating system and recombination affect molecular evolution in four Triticeae species. Genet Res. 90(1):97–109.
Hazzouri KM, et al. 2013. Comparative population genomics in *Collinsia* sister species reveals evidence for reduced effective population size, relaxed selection and evolution of biased gene conversion with an ongoing mating shift. Evolution 67(5):1263–1278.
Hohmann N, Wolf EM, Lysak MA, Koch MA. 2015. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. Plant Cell 27(10):2770–2784.
Hough J, Williamson RJ, Wright SI. 2013. Patterns of selection in plant genomes. Annu Rev Ecol Evol Syst. 44(1):31–49.
Hu TT, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 43(5):476–481.
Huelsenbeck JP, Crandall KA. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu Rev Ecol Syst. 28(1):437–466.
Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, et al. 2015. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44:D286–D293.
Johnston MO, et al. 2009. Correlations among fertility components can maintain mixed mating in plants. Am Nat. 173(1):1–11.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
Koch MA, Kiefer M. 2005. Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps of three diploid species—*Capsella rubella*, *Arabidopsis lyrata subsp. petraea*, and *A. thaliana*. Am J Bot. 92(4):761–767.
Lanfear R, et al. 2013. Taller plants have lower rates of molecular evolution. Nat Commun. 4:1879.
Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A. 102(30):10557–10562.
Mattila TM, Tyrmi J, Pyhäjärvi T, Savolainen O. 2017. Genome-wide analysis of colonization history and concomitant selection in *Arabidopsis lyrata*. Mol Biol Evol. 34(10):2665–2677.
Ness RW, Siol M, Barrett SC. 2012. Genomic consequences of transitions from cross-to self-fertilization on the efficacy of selection in three independently derived selfing plants. BMC Genomics 13:611.
Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246(5428):96–98.
Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158(2):927–931.
Pollak E. 1987. On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics 117(2):353–360.
Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D. 2011. Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. Genome Biol Evol. 3:868–880.
Schmid M, et al. 2005. A gene expression map of *Arabidopsis thaliana* development. Nat Genet. 37(5):501–506.
Shimizu KK, Tsuchimatsu T. 2015. Evolution of selfing: recurrent patterns in molecular adaptation. Annu Rev Ecol Evol Syst. 46(1):593–622.
Slotte T, et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet. 45(7):831–835.
Smith A, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322(5898):86–89.
Stebbins GL. 1957. Self fertilization and population variability in the higher plants. Am Nat. 91(861):337–354.
Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135(2):599.
Tang C, et al. 2007. The evolution of selfing in *Arabidopsis thaliana*. Science 317(5841):1070–1072.
Tatusov RL, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.
The Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814):796–815.
Tsuchimatsu T, et al. 2010. Evolution of self-compatibility in *Arabidopsis* by a mutation in the male specificity gene. Nature 464(7293):1342–1346.
Tsuchimatsu T, et al. 2017. Patterns of polymorphism at the self-incompatibility locus in 1,083 *Arabidopsis thaliana* genomes. Mol Biol Evol. 34:1878–1889.
Wright SI, Lauga B, Charlesworth D. 2002. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Mol Biol Evol 19(9):1407–1420.
Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. Mol Biol Evol. 28(8):2359–2369.
Yang YF, Zhu T, Niu DK. 2013. Association of intron loss with high mutation rate in *Arabidopsis*: implications for genome size evolution. Genome Biol Evol. 5(4):723–733.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.

Associate editor: Laura Rose