Selection on Accessible Chromatin Regions in *Capsella grandiflora*

Robert Horvath,[1] Emily B. Josephs,[2] Edouard Pesquet,[3] John R. Stinchcombe,[4] Stephen I. Wright,[4] Douglas Scofield,[5] and Tanja Slotte*,[1]

[1]Department of Ecology, Environment and Plant Sciences, Science for Life Laboratory, Stockholm University, Stockholm, Sweden
[2]Department of Plant Biology, Michigan State University, Lansing, MI, USA
[3]Department of Ecology, Environment and Plant Sciences, Stockholm University, Stockholm, Sweden
[4]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada
[5]Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden

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

Associate editor: Michael Purugganan

Abstract

Accurate estimates of genome-wide rates and fitness effects of new mutations are essential for an improved understanding of molecular evolutionary processes. Although eukaryotic genomes generally contain a large noncoding fraction, functional noncoding regions and fitness effects of mutations in such regions are still incompletely characterized. A promising approach to characterize functional noncoding regions relies on identifying accessible chromatin regions (ACRs) tightly associated with regulatory DNA. Here, we applied this approach to identify and estimate selection on ACRs in *Capsella grandiflora*, a crucifer species ideal for population genomic quantification of selection due to its favorable population demography. We describe a population-wide ACR distribution based on ATAC-seq data for leaf samples of 16 individuals from a natural population. We use population genomic methods to estimate fitness effects and proportions of positively selected fixations (α) in ACRs and find that intergenic ACRs harbor a considerable fraction of weakly deleterious new mutations, as well as a significantly higher proportion of strongly deleterious mutations than comparable inaccessible intergenic regions. ACRs are enriched for expression quantitative trait loci (eQTL) and depleted of transposable element insertions, as expected if intergenic ACRs are under selection because they harbor regulatory regions. By integrating empirical identification of intergenic ACRs with analyses of eQTL and population genomic analyses of selection, we demonstrate that intergenic regulatory regions are an important source of nearly neutral mutations. These results improve our understanding of selection on noncoding regions and the role of nearly neutral mutations for evolutionary processes in outcrossing Brassicaceae species.

Key words: **ATAC-sequencing, open chromatin region, gene expression variation, natural selection, functional noncoding sequences, distribution of fitness effects.**

Article

Introduction

Accurate estimates of genome-wide rates and fitness effects of new mutations are essential for an improved understanding of the response of populations to selection, the evolution of mating systems, and the maintenance of quantitative variation (Wright and Andolfatto 2008; Tataru et al. 2017). Quantifying the genomic impact of selection is therefore a major aim in evolutionary genetics (Ohta 1973; Kimura 1983; Gillespie 2004). Indeed, the relative contributions of different classes of mutations to polymorphism and divergence and the impact of selection at linked sites lie at the heart of many debates in population genetics (Ohta 1973; Kimura 1983; Kreitman 1996; Ohta and Gillespie 1996; Gillespie 2004; Kern and Hahn 2018; Jensen et al. 2019, Chen et al. 2020).

For instance, under the strict neutral theory, the vast majority of mutations are assumed to be either strongly deleterious or neutral (Kimura 1983). Most polymorphisms in natural populations are then expected to be neutral, and the rate of evolution is independent of the effective population size (N_e) (Kimura 1983; Ohta 1992). In contrast, under the nearly neutral theory, many new mutations are expected to be under such weak selection that changes in N_e determine whether they are efficiently selected or not (Ohta 1992). The most well-developed versions of the nearly neutral theory mainly consider the role of nearly neutral deleterious mutations (Kreitman 1996), although Ohta included weakly beneficial mutations in her description of the nearly neutral theory (Ohta 1992). Because fitness distributions of new mutations will ultimately impact the expected rate of evolution of a population, especially with varying N_e (Charlesworth and Eyre-Walker 2007), empirical estimates of the genome-wide distribution of fitness effects (DFE) of new mutations are crucial.

Recent empirical studies in plants suggest that nearly neutral deleterious mutations are key for the evolution of coding regions (Hough et al. 2013; Chen et al. 2020), but we currently know much less about selection on noncoding sequences (Haudry et al. 2013; Williamson et al. 2014, Joly-Lopez et al. 2020). The dearth of knowledge of how selection affects noncoding regions is unfortunate, as mutations in noncoding regions could potentially contribute a wealth of nearly neutral mutations (Ohta 2002) and understanding selection on noncoding regions is thus essential for a complete understanding of genome evolution. In contrast to animals, plant genomes harbor a lower percentage of well-characterized functional noncoding sequences and the proportion of noncoding sites under selective constraint in plants is lower than in animals (Haudry et al. 2013; Hough et al. 2013). However, we currently lack a complete understanding of the DFE of mutations in functional noncoding regions in plants. Quantifying the contribution of beneficial as well as deleterious mutations is especially interesting in this regard, as mutations in noncoding *cis*-regulatory regions have long been suggested to be likely to contribute to adaptation (e.g., King and Wilson 1975; Wray 2007; Stern and Orgogozo 2008).

A first step toward an improved understanding of selection on the noncoding part of the genome is the identification and annotation of functional noncoding regions. Comparative genomic approaches such as phylogenetic footprinting can identify highly conserved noncoding sequences (CNS) as candidate functional noncoding regions (Miller et al. 2004) but these approaches are not well suited for detecting functional noncoding regions under positive selection or those under weak purifying selection, because such sequences will not be evolutionarily conserved (Andolfatto 2005). However, thanks to recent progress in sequencing methods, it is now possible to identify potentially functional noncoding sequences not only based on sequence conservation or purely computational methods that identify transcription factor binding sites (e.g., Nguyen and Androulakis 2009), but also based on epigenetic characteristics or genetic associations. For instance, single noncoding loci can be associated with phenotypic traits such as gene expression variation through association mapping or QTL mapping. Such loci are commonly termed expression quantitative trait loci (eQTL) (Gilad et al. 2008; Josephs et al. 2015, 2017, 2020). Combined with polymorphism data, these alternative approaches can help us better characterize selection on functional noncoding sequences in plants (Wright and Andolfatto 2008).

A particularly promising approach to characterizing putatively functional noncoding regions relies on identifying accessible noncoding sequences. These noncoding regions are located in accessible chromatin regions (ACRs), which allow DNA-binding proteins and transcription factors to bind their target sequences. Such regions are thought to be enriched for *cis*-regulatory elements and were found to be good candidate regions for narrowing down functional sequences in maize (Vera et al. 2014; Rodgers-Melnick et al. 2016). ACRs can be identified through approaches such as Micrococcal Nuclease analysis of chromatin structure (MNase-seq) (Zaret 2005) or Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) (Buenrostro et al. 2013; Buenrostro, Wu, Chang, et al. 2015). Such methods have been successfully used to identify ACRs in 13 angiosperm species and revealed a positive correlation between the number of detected ACRs and the number of annotated genes in the genome (Lu et al. 2019). ACRs were conserved throughout evolution and harbored more CNS, but fewer single-nucleotide polymorphisms (SNPs) and transposable elements (TEs) than their flanking sequences, suggesting that ACRs are likely to harbor *cis*-regulatory regions (Lu et al. 2019). In the rice genome, ACRs are under similar levels of constraint than UTRs and promoters, but under weaker constraint than coding sequences or CNSs (Joly-Lopez et al. 2020). However, for most plant genomes, we still lack a complete understanding of the full DFE of mutations in candidate functional noncoding regions. This is unfortunate as weak selection on mutations in these regions can impact patterns of polymorphism and divergence through selective interference (Comeron et al. 2008; Good et al. 2014).

In this study, we use ATAC-seq analyses to identify ACRs and investigate the DFE of intergenic ACRs in a plant species that is especially well suited for this purpose, the outcrossing Brassicaceae species *Capsella grandiflora*. *Capsella grandiflora* has a large and relatively constant effective population size without strong population structure (Foxe et al. 2009; Slotte et al. 2010; St. Onge et al. 2011; Douglas et al. 2015) which makes it ideal for investigating variation in selection across the genome (Steige et al. 2017). Indeed, natural selection was previously reported to be efficient in both coding and CNS in the *C. grandiflora* genome (Slotte et al. 2010; Williamson et al. 2014; Steige et al. 2017). The rapid decay of linkage disequilibrium (within 0.004 cM, or ~1 kb) in *C. grandiflora* (Foxe et al. 2009; Mattila et al. 2019), further minimizes interference from nearby coding sites and facilitates the study of selection patterns on candidate functional noncoding regions in the proximity of genes, such as *cis*-regulatory regions.

Here, we use ATAC-seq analyses to identify ACRs and describe a population distribution of ACRs in a natural population of *C. grandiflora*. We then generate and analyze whole-genome resequencing data from 40 *C. grandiflora* individuals to explore the strength and nature of selection in intergenic ACRs. We achieve this by quantifying the DFE of new mutations and the proportion of positively selected substitutions (α) in different parts of the genome. Finally, we leverage previous results to test for an enrichment of eQTL (Josephs et al. 2015, 2017, 2020) and CNS (Williamson et al. 2014) and a depletion of TEs in ACRs. Our results are important for an improved understanding of the nature, prevalence and strength of selection on noncoding regions, and the role of nearly neutral mutations for evolutionary processes in outcrossing Brassicaceae species.

Results

Open Chromatin Identification and Profile across a *C. grandiflora* Population

We used ATAC-seq to identify ACRs in young and still extending leaves of 16 nine-week-old plants from a single Greek *C. grandiflora* population. Across our 16 individuals, we identified 31,243 distinct ACRs, which together made up 12% of the genome. Out of these, 28,273 (92%) were located on one of the eight main genome scaffolds. Splitting up ACRs into mutually exclusive genic and intergenic ACRs revealed that overall, 25.8% of the ACRs were located in intergenic regions, whereas 74.2% and 60% were located in genes and in coding sequences, respectively. The ACR density was higher within genes and lower in the immediate surroundings of genes (fig. 1*A*). Out of the intergenic ACRs, 34.3% were proximal to genes (<2 kb up- and downstream), and 65.7% were distal (>2 kb) to genes. In concordance with previous findings (e.g., Lu et al. 2019), ACRs located in genes as well as intergenic ACRs exhibited higher GC content than their surrounding regions (supplementary fig. S1, Supplementary Material online).

We further divided ACRs based on their frequency in our sample into unique ACR (uACR), common ACR (cACR), and high-frequency ACR (hACR) (see Materials and Methods). In our population sample, 18.2%, 57.5%, and 24.3% of ACRs were uACRs, cACRs, and hACRs, respectively.

ACRs Are Depleted of CNSs

Some CNSs were previously reported to be involved in gene expression regulation in plants (e.g., Freeling and Subramaniam 2009). To assess whether they are also enriched in intergenic ACRs, as might be expected if ACRs harbor regulatory elements, we investigated the distribution of previously identified CNSs in *C. grandiflora* (Williamson et al. 2014). Interestingly, we observed the opposite pattern where ACRs were more frequent outside of CNSs than within CNSs (fig. 1*B*). Indeed, ACRs were significantly depleted of CNSs (1,000 permutation-based two-sided *P*-value < 0.002, supplementary table S1, Supplementary Material online). Such a significant depletion was also observed in distal and proximal ACRs (1,000 permutation-based two-sided *P*-value < 0.002, supplementary table S1, Supplementary Material online). A lack of CNSs in ACRs might indicate that although intergenic ACRs could be enriched in functional sites and thus under selective constraint, which would be expected to lead to higher sequence conservation, other factors such as elevated mutation rates in ACRs (Monroe et al. 2020) or rapid regulatory site turnover could mitigate this effect and lead to a lack of sequences that can be identified as CNS in ACRs.

Impact of Selection on Intergenic ACRs and Comparable Intergenic Regions

To investigate the impact of selection on mutations in ACRs in our *C. grandiflora* population, we conducted population genomic analyses of whole-genome sequences of 40 individuals (including those used to identify ACRs; mean coverage 34×). We identified a total of 54.9 million sites and 2 million SNPs that could be polarized and thus included when estimating the DFE. We quantified polymorphism and divergence at ACRs, 0-fold degenerate and 4-fold synonymous sites (supplementary table S2, Supplementary Material online) and found that intergenic ACRs had a lower proportion of SNPs than nearby genomic regions (fig. 1*C*) as well as a slightly lower mean nucleotide diversity (π) than intergenic sites in general (supplementary table S2, Supplementary Material online). Reduced diversity did not seem to be a result of locally reduced mutation rates in intergenic ACRs, as intergenic ACRs had a higher mean divergence between *Capsella*

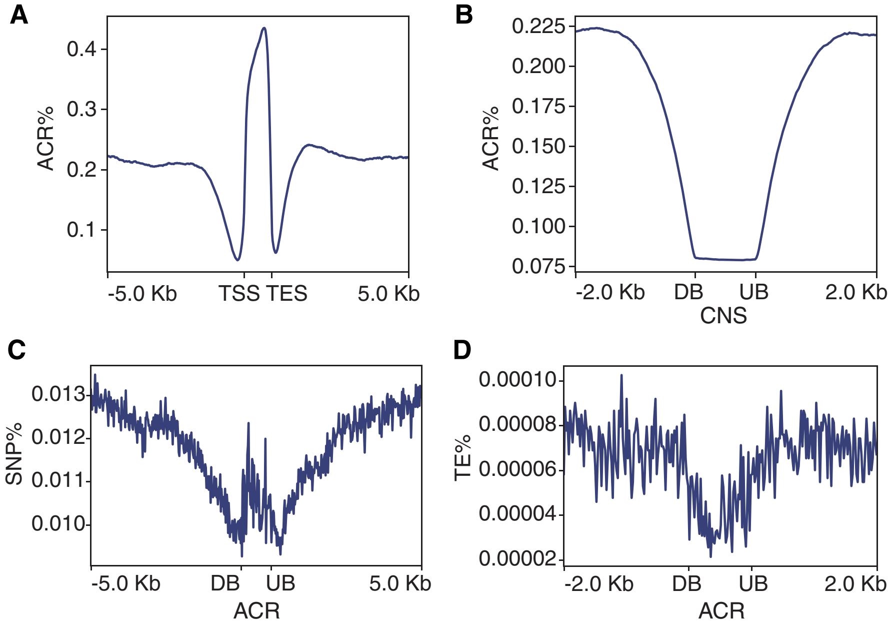

Fig. 1. Genetic profile plots. (*A*) Proportion of ACRs in and around genes (TSS: transcription start site; TES: transcription end site). (*B*) Proportion of ACRs in and around CNSs. (*C*) Proportion of SNPs in and around intergenic ACRs. (*D*) Proportion of TE insertions in and around ACRs. All plots only include mappable regions of the genome (see Materials and Methods) and the inferred boundaries of the CNS and ACR are labeled upstream boundary (UB) and downstream boundary (DB).

and *Arabidopsis* than other intergenic sites (supplementary table S2, Supplementary Material online). Differences in diversity and divergence between intergenic ACRs and other intergenic regions may thus reflect different selective pressures on these regions.

To investigate the impact of positive and negative selection on intergenic ACRs, we estimated the DFE and α in intergenic ACRs. For comparison, we also estimated DFE and α for 0-fold degenerate sites and for a set of proximal and distal intergenic negative control regions (see Methods). For this purpose, we used DFE-alpha v.2.16 (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009; Schneider et al. 2011).

Our DFE analyses of intergenic ACRs revealed that in *C. grandiflora*, approximately 18% (95% CI: 16.8–19.5%) of new mutations in intergenic ACRs have a strongly deleterious fitness effect ($-N_es > 10$). Thus, some mutations in intergenic ACRs are under strong constraint. However, the proportion of new mutations in intergenic ACRs under strong purifying selection is significantly lower than the estimated 75.6% (95% CI: 75.5–75.7%) for new 0-fold degenerate mutations (two-sided *P*-value < 0.02; fig. 2A). Further, 73.1% (95% CI: 71.8–74.2%) of new mutations in intergenic ACRs were neutral and nearly neutral mutations ($1 > -N_es > 0$), which was significantly more than the 19% (95% CI: 18.9–19.1%) of new mutations at 0-fold degenerate sites (two-sided *P*-value $<$ 0.02; fig. 2A). These results suggest that mutations in intergenic ACRs could make a substantial contribution to the genome-wide content of nearly neutral mutations, because we found that nearly neutral mutations could arise at 5.1 million different sites in intergenic ACRs but only at 4 million different 0-fold degenerate sites. Overall, the one- and two-epoch model DFE estimates revealed similar results (fig. 2 and supplementary fig. S2, Supplementary Material online) and the estimated DFE for 0-fold degenerate mutations in this study was similar to previous estimates for *C. grandiflora* (Williamson et al. 2014; Mattila et al. 2019).

Although we observed clear differences in the impact of purifying selection on intergenic ACRs and 0-fold degenerate sites, our estimates of the proportion of mutations fixed by positive selection (α) in intergenic ACRs were negative (supplementary fig. S3, Supplementary Material online), indicating limitations in our ability to reliably estimate α in intergenic ACRs using DFE-alpha. In contrast, α for the 0-fold degenerate sites was 0.32 (95% CI: 0.31–0.33; supplementary fig. S3, Supplementary Material online), in line with previous estimates for *C. grandiflora* (Slotte et al. 2010; Williamson et al. 2014; Josephs et al. 2017; Mattila et al. 2019).

To further disentangle how selection on different intergenic ACRs affects the DFE, and to contrast these findings to nonaccessible intergenic regions, we estimated DFE and α for proximal ACRs, distal ACRs, and control intergenic regions (comparable proximal and distal intergenic regions not overlapping with ACRs or CNSs). Proximal and distal ACRs have a significantly higher proportion of strongly deleterious new mutations than control regions (two-sided *P*-value $<$ 0.02; fig. 2B), indicating that proximal as well as distal ACRs are under stronger purifying selection than comparable intergenic regions. As before, our ability to reliably estimate α in proximal and distal ACRs was limited (supplementary fig. S3, Supplementary Material online).

Finally, we estimated the DFE separately for unique, common, and hACRs to investigate whether regions which are consistently accessible are under different selective constraint than regions that are only sporadically accessible. Although differences were subtle, our analyses suggest that ACRs with higher population frequencies harbor a higher proportion of strongly deleterious mutations than unique and cACRs (two-sided *P*-value $<$ 0.02; fig. 2C). The differences in the proportion of strongly deleterious mutations implies that new mutations in ACRs present at higher population frequencies might be under more efficient purifying selection than those in lower frequency ACRs.

Regions Containing ACRs Are Enriched with eQTL

Our population genomic analyses suggest that intergenic ACRs are under (weak) selection. To test whether ACRs are also more likely to harbor functional noncoding sequences, such as *cis-* or *trans*-regulatory regions, we investigated the overlap between intergenic ACR and previously described *cis-* and *trans*-eQTL in *C. grandiflora* leaves (Josephs et al. 2015, 2020).

In *C. grandiflora*, we found an enrichment of *cis*-eQTL within and in the proximity (500 bp surroundings) of ACRs in and around (5 kb up- and downstream) genes (fig. 3, 3,032, 95% CI based on 1,000 permutations: 2,669–2,814, permutation-based two-sided *P*-value $<$ 0.005, supplementary table S4, Supplementary Material online). There was also a significant enrichment of *cis*-eQTL within and in the proximity of proximal ACRs (1,000 permutation-based two-sided *P*-value $<$ 0.002, supplementary table S4, Supplementary Material online). Similarly, we found an enrichment of *trans*-eQTL within and in the proximity of ACRs at least 5 kb away from genes (1,000 permutation-based two-sided *P*-value $<$ 0.005, supplementary table S5, Supplementary Material online). These results were expected under the assumption that intergenic ACRs harbor functional sites such as *cis-* and *trans*-regulatory regions, because eQTL are expected to be located in or around the sequences regulating the levels of gene transcription. Hence, these results are in line with our interpretation of the DFE, that SNPs in intergenic ACRs are more likely to have fitness consequences than in comparable intergenic regions.

To elucidate the biological function of genes affected by eQTL located in intergenic ACRs, we investigated their gene ontology (GO) annotations using GO Slim terms, cut-down versions of GO providing a broader overview of gene function present. The most represented specific GO Slim biological process terms were “metabolic process” and “biosynthetic process,” respectively (supplementary fig. S4, Supplementary Material online), indicating that many of these genes are involved in basic metabolic processes. The most common specific GO Slim molecular function terms were “binding” and “catalytic activity” (supplementary fig. S4, Supplementary Material online). For cellular component, the most common specific GO Slim term annotation was “cell” and

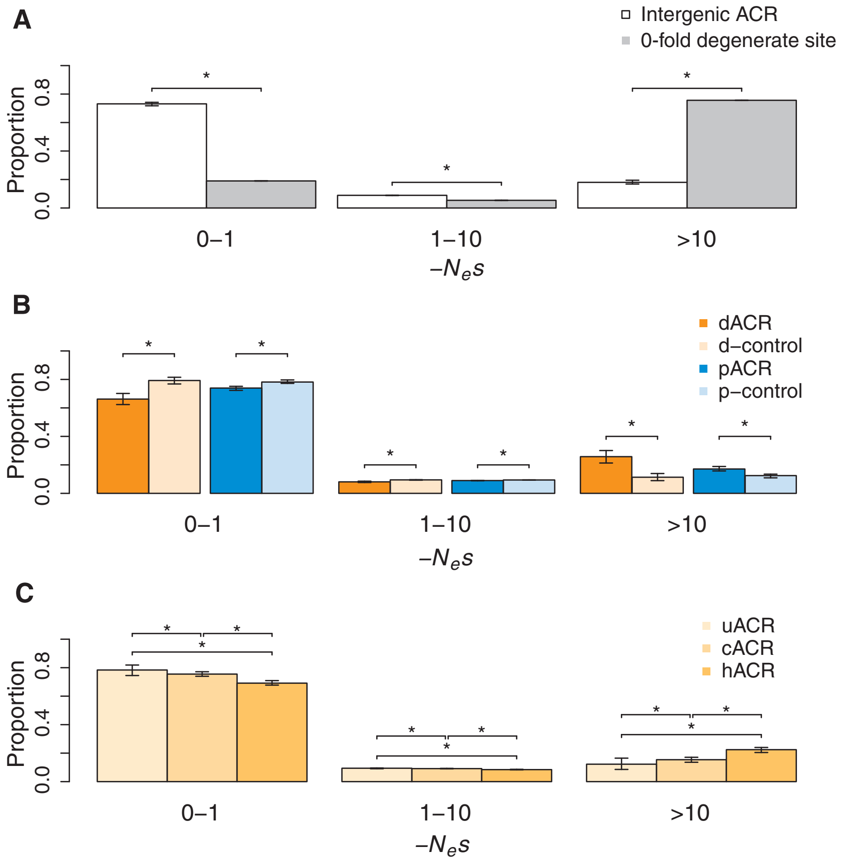

Fig. 2. Estimated distribution of fitness effects of new mutations in intergenic ACRs, intergenic control regions, and 0-fold degenerate sites using a one-epoch demographic model. The estimated DFE was binned based on the mean selective effect ($-N_es$) into three bins: effectively neutral ($0 < -N_es < 1$), intermediate fitness effect ($1 < -N_es \leq 10$), and strongly deleterious ($-N_es > 10$) mutations. (*A*) DFE of intergenic ACRs and 0-fold degenerate sites. (*B*) DFE of distal ACRs (dACRs), distal control regions (d-control), proximal ACRs (pACRs), and proximal control regions (P-control). (*C*) DFE of intergenic ACRs split up into unique (u), common (c), and high-frequency (h) ACRs. Error bars show the 95% CI of each estimate based on 100 bootstrap replicates. Significant differences between estimates are shown by asterisks (Kruskal–Wallis test with a Dunn post hoc test, *P*-value < 0.05).

"intracellular" (supplementary fig. S4, Supplementary Material online). Overall, there were 11 GO terms that were overrepresented relative to the set of all genes analyzed for eQTL (Benjamini–Hochberg adjusted Fisher's exact test *P*-value < 0.05, supplementary table S6, Supplementary Material online). Two of these GO terms in the domain biological process ("response to organonitrogen compound" and "response to chitin") indicate an overrepresentation of genes associated with defense response genes and with nitrogenated nutrients availability, respectively. The molecular function GO terms corresponded to catalytic activity, further indicating an overrepresentation of genes associated to metabolism. Finally, eight GO terms were from the cellular component domain and included terms such as "cytoplasm" and "endomembrane system," as well as other subcellular compartments, further confirming the association with general cellular functions (supplementary table S6, Supplementary Material online). Together, these results suggest that regulatory regions associated with intergenic ACRs mostly regulate genes involved in metabolic and biosynthetic processes, many of which have enzymatic activity or bind to DNA or proteins.

Depletion of TE Insertions in ACRs

TE insertions can interfere with the function of any functional part of the genome, such as genes or *cis*-regulatory regions, in two distinct ways. First, TEs can directly disturb a functional element through insertions into such elements. Second, TEs can trigger a host-induced epigenetic defence mechanism against TE spreading, which involves methylation and histone modifications of the targeted TE sequence, which can spread beyond the borders of the TEs and interfere with nearby functional sequences (Hollister and Gaut 2009). In *C. grandiflora*, varying selective constraint on TEs with distinct effects on gene expression was reported to be the main factor preventing an increase in the copy number of TEs (Uzunović et al. 2019) and, therefore, functional sites and their surroundings are expected to be depleted of TE insertions.

To test whether this was the case in ACRs, we first used PopoolationTE2 v1.10.04 (Kofler et al. 2016) to identify TE insertions in our *C. grandiflora* population. A total of 9,260 different TE insertions were identified in our population from which 1,008 were located in ACRs. TE insertions were less common in ACRs than in their surroundings (fig. 1*D*) and ACRs were significantly depleted of TE insertions (1,000

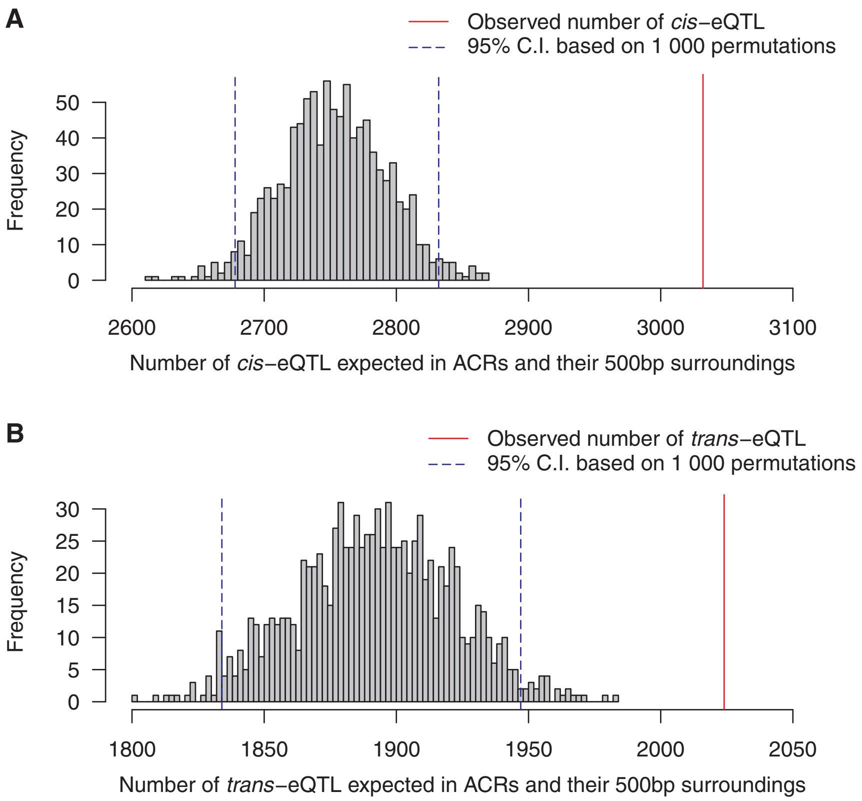

FIG. 3. Comparison between the observed and expected number of *cis*- and *trans*-eQTL located in ACRs and their 500 bp surroundings in *C. grandiflora*. (*A*) The expected number of *cis*-eQTL found within and in the proximity of ACRs located in and around genes (5 kb up- and downstream). (*B*) The expected number of *trans*-eQTL found within and in the proximity of ACRs at least 5 kb away from genes. The expected number of *cis*- and *trans*-eQTL within and in the proximity of ACRs (gray) were based on 1,000 permutations. The observed number of *cis*- and *trans*-eQTL located within and in the proximity of ACRs is indicated by the red lines. The blue dashed lines delimit the 95% confidence interval of the permutation test.

permutation-based two-sided *P*-value < 0.002; supplementary table S7, Supplementary Material online). Similarly, distal and proximal ACRs were significantly depleted of TE insertions (1,000 permutation-based two-sided *P*-value < 0.002; supplementary table S7, Supplementary Material online), indicating that this pattern is not only driven by ACRs located in genes and coding sequences. These results are in line with the hypothesis that if intergenic ACRs harbor functional sites then TEs in such regions might have a negative fitness impacts and should therefore be less numerous (Hollister and Gaut 2009). Alternatively, TEs could directly cause the loss of ACRs by altering chromatin accessibility.

To further investigate whether the observed depletion of TE insertions in ACRs could result from biases in recombination rate or GC content, considering that ACRs had higher GC content than their surroundings (supplementary fig. S1, Supplementary Material online), we split up ACRs based on the recombination rate and GC content. Splitting up ACRs revealed that regardless of the recombination rate, ACRs were significantly depleted of TE insertions (supplementary table S8, Supplementary Material online), but GC content affected the observed relationship between TE insertions in ACRs without any clear pattern (supplementary table S9, Supplementary Material online). These results indicate that although ACRs are overall depleted of TE insertions, other genetic features might also influence how TEs accumulate in ACRs. Further TE-specific features such as insertion biases, which were not accounted for in our analyses, could also contribute to the complexity of this relationship.

Discussion

The Genomic Distribution of ACRs in a *C. grandiflora* Population

In this study, we identified a total of 31,243 ACRs based on an ATAC-seq assay of 16 *C. grandiflora* individuals from a natural population. In line with previous reports in 13 plant species (Lu et al. 2019), ACRs in *C. grandiflora* were located predominantly within genes, represented a small proportion of the total genome and had a higher GC content compared with nearby genomic regions. Similarly to a previous study (Lu et al. 2019), ACRs also harbored a lower number of SNPs than their surroundings. The overall patterns of chromatin accessibility that we observe are broadly comparable to those previously reported for other plant species (Rodgers-Melnick et al. 2016; Alexandre et al. 2018; Maher et al. 2018; Lu et al. 2019).

Our results further demonstrate that accessibility is not a static feature of the genome of a species but shows within-population variation. Intraspecific variation in chromatin accessibility has previously been reported among five geographically and genetically diverse ecotypes of *Arabidopsis thaliana* (Alexandre et al. 2018). Our study extends these findings to show the presence of chromatin accessibility variation within a single natural plant population.

Intergenic ACR Are under Stronger Negative Selection than Comparable Intergenic Regions

To gain a better understanding of the impact of selection on noncoding sequences and the contribution of such regions to the genome-wide DFE of new mutations in plants, we estimated the DFE of new mutations in intergenic ACRs. Estimating the DFE of intergenic ACRs revealed that some SNPs in intergenic ACRs were under negative selection in *C. grandiflora* (fig. 2*A*). The proportion of mutations in intergenic ACRs with strongly deleterious fitness effect was higher than in comparable intergenic regions (fig. 2*B*), indicating that ACRs were more likely to harbor functional noncoding sites than other intergenic regions, as suggested previously (Vera et al. 2014; Rodgers-Melnick et al. 2016). High-frequency ACRs (hACRs) harbored a slightly but significantly higher proportion of strongly deleterious mutations than unique and cACRs (fig. 2*C*), which could be expected if broadly important functional *cis*-regulatory regions were more likely to be located in hACRs than in regions of the genome which are less frequently accessible. However, chromatin accessibility can be variable between cell types and developmental states (Shen et al. 2012; Pajoro et al. 2014; Buenrostro, Wu, Litzenburger, et al. 2015) and, therefore, low-frequency ACRs in whole tissue samples could also represent ACRs in less abundant cell types and/or states.

ACRs Contribute to the Genome-Wide Input of Nearly Neutral Mutations

To investigate the suggestion that functional noncoding regions such as regulatory regions could contribute a wealth of nearly neutral mutations to the overall load of nearly neutral mutations in the genome (e.g., Ohta 2002), we compared the estimated proportion of nearly neutral new mutations at intergenic ACRs to other regions of the genome. Doing so revealed that intergenic ACRs harbor significantly more nearly neutral and neutral mutations than 0-fold degenerate sites (fig. 2*A*). Taking into account differences in the number intergenic ACR and 0-fold degenerate sites in *C. grandiflora*, this would mean that intergenic ACRs contribute roughly 1.3 times more new nearly neutral mutations than 0-fold degenerate sites. Hence, intergenic regions, especially intergenic ACRs, are an important contributor to the genome-wide rate of nearly neutral mutations, which is an important finding because the efficacy of selection on such mutations depends on the effective population size (N_e) (Wright and Andolfatto 2008).

The population genetic methods we used to estimate the DFE require data for a class of sites assumed to be evolving neutrally. In this study, we assumed that 4-fold degenerate sites evolve neutrally when estimating the DFE and α of 0-fold degenerate sites, and for analyses of the DFE at intergenic ACRs we used nearby intergenic regions as control regions. The assumption of neutrality may not strictly hold in either case, especially not for synonymous sites, for which selective constraint has previously been demonstrated in several systems (Chamary and Hurst 2005; Chamary et al. 2006; Eöry et al. 2010; Künstner et al. 2011; Gu et al. 2012; Lawrie et al. 2013; Gossmann et al. 2018). It is possible that 4-fold degenerate sites might also experience some selective constraint in *C. grandiflora* (Williamson et al. 2014). Although this issue cannot be entirely avoided, it might lead to underestimation of the strength of purifying selection and overestimation of α (Eöry et al. 2010; Künstner et al. 2011). However, we consider it unlikely that all differences in the DFE between intergenic ACRs and 0-fold degenerate sites are due to deviations from neutrality at the assumed neutral sites.

How does the inferred purifying selection on intergenic ACRs compare with other noncoding regions? Here, we estimated that in *C. grandiflora* 73.1% of new mutations in intergenic ACRs are effectively neutral. Previous estimates suggest that in *C. grandiflora* 28%, 45%, and 70% of new mutations are effectively neutral in CNS, UTRs and introns, respectively (Williamson et al. 2014). Although the previously estimated proportion of nearly neutral intergenic sites was nearly 100% in *C. grandiflora* (Williamson et al. 2014), here, we show that a small presumably functional subset of the intergenic sites are impacted by higher selective constraint. We show that intergenic ACRs are affected by weaker purifying selection than 0-fold degenerate sites, CNS, and UTRs but the impact of purifying selection on introns and intergenic ACRs seems to be similar in *C. grandiflora*. This result differs somewhat from those in rice, where, using a different analysis approach, the level of constraint was estimated to be similar for ACRs and promoter regions and UTRs (Joly-Lopez et al. 2020). These findings therefore contribute to an improved understanding of the currently relatively limited knowledge of the impact of purifying selection on intergenic ACRs in plant genomes.

Previous estimates of the proportion of effectively neutral mutations in intergenic regions of other plant and animal species suggested that more than 90% of intergenic sites in *A. thaliana* and *A. lyrata* (Haudry et al. 2013) and approximately 50% in *Drosophila* (Andolfatto 2005) are effectively neutral. In mouse, humans, and *Drosophila*, it was estimated that between 67–77% (Kousathanas et al. 2011), 78–100% (Eyre-Walker and Keightley 2009), and 55–66% (Halligan et al. 2004; Halligan and Keightley 2006) of intergenic sites 500 bp up- and downstream of genes are effectively neutral, respectively. Although it is important to bear in mind that differences in the effective population size can influence the proportion of effectively neutral mutations in a population (Ohta 1973; Eyre-Walker and Keightley 2007; Wright and Andolfatto 2008), our results are consistent with the suggestion that a smaller fraction of plant than animal genomes is evolving under selective constraint (Hough et al. 2013).

One surprising finding was that intergenic ACRs were depleted for CNS. In rice, CNS regions were associated with transcription factors and developmental genes that vary in expression across developmental stages (Joly-Lopez et al. 2020). In contrast, ACRs were associated with constitutively expressed genes (Joly-Lopez et al. 2020), which could explain the seeming contradiction between our findings of ACRs being depleted of CNS and a previous report that CNS were enriched in ACRs (Lu et al. 2019). One difference between Lu et al. (2019) and this study, besides the different study systems, is that in Lu et al. (2019) leaf tissue was sampled from very young plants (maximum 10 days old), whereas in our case leaves were sampled from mature (approximately 9-week-old) plants. It is therefore possible that CNSs involved in gene expression regulation during leaf development could be underrepresented in our data set. Our GO analyses showing mainly genes associated with metabolism suggest that this could indeed be the case. Alternatively, elevated mutation rates at ACRs in the Brassicaceae (Monroe et al. 2020) could contribute to the depletion of CNSs in ACRs in our study. These findings further indicate that completely relying on comparative genomic approaches to detect functional sites in *C. grandiflora* will underestimate the amount of such sites.

Evidence for Positive Selection at Intergenic ACRs Is Limited

Cis-regulatory variation has long been considered to be important for adaptation. If ACRs are enriched for *cis*-regulatory elements under positive selection, this should leave a signature of excess divergence relative to expectations given fixation of neutral and weakly deleterious mutations. To test for evidence of positive selection acting on genetic variation in accessible intergenic regions, we estimated the proportion of substitutions fixed through positive selection, α, in intergenic ACRs.

Our α estimates for intergenic regions were mostly negative, but most confidence intervals included values above 0 (supplementary fig. S3, Supplementary Material online), indicating that our ability to detect positive selection in intergenic ACRs was limited. Negative α estimates in plants were previously reported in other studies (e.g., Gossmann et al. 2010) and were hypothesized to be due to ancient demographic events that were not captured by polymorphism data (Booker et al. 2017). Methodological limitations could also limit our ability to estimate α. For example, violation of the neutrality assumption could be a considerable issue, as discussed above. Other sources of potential bias include the requirement of an outgroup for polarization of alleles, and the higher GC content of ACRs, which is generally associated with a higher mutation rate (Coulondre et al. 1978; Fryxell and Moon 2005; DeRose-Wilson and Gaut 2007). Finally, a biological factor that could interfere with our α estimates using DFE-alpha is the potential presence of weakly beneficial mutations in our population. To investigate this issue, we ran polyDFE v2.0. (Tataru et al. 2017; Tataru and Bataillon 2019), which can estimate the contribution of weakly advantageous mutations to the DFE (see Materials and Methods). Although the α and DFE estimates for ACRs were very noisy, especially for proximal and distal ACRs (dACRs), these polyDFE analyses revealed a potential presence of weakly beneficial mutations in intergenic ACRs (supplementary figs. S5 and S6, Supplementary Material online). These results suggest that some sites in intergenic ACRs may be under weak positive selection, but further work is needed to fully elucidate the contribution of positive selection to the evolution of intergenic ACRs.

Enrichment of eQTL and Depletion of TEs Support Functional Importance of ACRs

In line with previous reports of accessible regions being promising candidates for functional regions in maize (Vera et al. 2014; Rodgers-Melnick et al. 2016) and several other plant species, including *Arabidopsis thaliana* (Lu et al. 2019), at least two lines of evidence suggest that in *C. grandiflora* accessible DNA sequences identified through ATAC-seq are also promising candidates for functional regions. First, previously identified *cis-* and *trans*-eQTL in comparable *C. grandiflora* leaf samples were found more frequently in and around accessible DNA sequences indicating that regions containing ACRs are enriched for variants involved in gene expression regulation. An enrichment of loci involved in gene expression regulation in ACRs is in line with the expectation that regulatory elements need to be accessible to fulfil their function (e.g., Li et al. 2011; Jiang 2015) and suggests that variation in ACRs contributes disproportionately to the genetic basis of gene expression variation. Second, we observed a TE depletion in ACRs as well as in intergenic ACRs, consistent with the hypothesis that if intergenic ACRs harbor functional sites, then selection against TE insertions in such regions should remove them from intergenic ACRs. This is in line with the results of Uzunović et al. (2019) who concluded that gene expression alterations caused by TE insertions results in negative selection against TEs. Overall these results are consistent with those in other plant species, where TEs were also depleted in ACRs (e.g., Lu et al. 2019).

Conclusions

Our results show that intergenic ACRs in *C. grandiflora* are under stronger purifying selection than other intergenic regions and suggest that ACRs harbor functional sites, such as regulatory elements. In line with this hypothesis, we find an enrichment of eQTL as well as a depletion of TE insertions in intergenic ACRs. Despite these findings, selective constraint on intergenic ACRs still seems to be insufficient to cause sequence conservation. Overall, purifying selection on intergenic ACRs is mostly weak, similar to levels seen at intronic sites in *C. grandiflora*. We also find little evidence for recurrent strong positive selection at intergenic ACRs, although we cannot completely rule out an impact of weak positive selection on patterns of polymorphism. Our study is among the first to estimate the DFE of mutations at ACRs in a plant genome. By doing so, we can directly compare contribution of mutations in noncoding regulatory regions and at nonsynonymous sites to the overall load of nearly neutral mutations in the genome. We conclude that such mutations are expected to contribute roughly 1.3 times more to the overall load compared with mutations at 0-fold degenerate nonsynonymous sites. Given the rapid decay of linkage disequilibrium in *C. grandiflora* and our evidence for weak selection on ACRs, our results suggest that intergenic sites in this species evolve mostly in agreement with the nearly neutral theory (Ohta 1973, 1992).

Materials and Methods

Plant Material

Field-collected *C. grandiflora* seeds from a single population in Greece (Mikro Papingo: longitude: 39.98; latitude: 20.77) were surface sterilized and germinated on half strength Murashige-Skoog medium without b5 vitamins (Sigma-Aldrich, Missouri, USA). A total of 40 plants were grown in a growth chamber (16 h light/23°C; light intensity: 110 μE m^{-2} s^{-1} ; 8 h dark/ 20°C; 70% maximum humidity) in potting soil. For ATAC-seq, young and still extending leaves were sampled within the first 2 h after the start of the light period, from 16 nine-week-old plants at roughly identical developmental stage before the production of flower buds started. Leaf samples for DNA-seq were collected approximately 2 weeks later from all 40 plants.

ATAC Sequencing

Nuclei extraction was performed immediately after harvesting the samples following the "Purification of Total Nuclei Using Sucrose Sedimentation" protocol from Bajic et al. (2018), without using Triton X-100 detergent and performing additional sieving first with a 100-µm nylon cell strainer before sieving again with a 70-µm nylon cell strainer as described by Bajic et al. (2018). A final sieving with a 70-µm nylon cell strainer was also added at the end of the protocol to remove any remaining cell debris. To perform an Assay for Transposase-Accessible Chromatin with high throughput sequencing (ATAC-seq; Buenrostro et al. 2013; Buenrostro, Wu, Chang, et al. 2015), the Nextera XT kit (Illumina, San Diego, USA) was used for the library preparation performed by the National Genomics Infrastructure in Stockholm following Bajic et al. (2018), but excluding the initial extension step in the PCR following tagmentation. The resulting single library was sequenced (paired-end 151-bp reads) on the Illumina NovaSeq 6000 by the National Genomics Infrastructure in Stockholm. To investigate the reproducibility of the sites detected as transposase hypersensitive sites (THSs) through ATAC-seq analyses, samples from each of the 16 plants were divided into three technical replicates after the nuclei extraction but before the ATAC-seq library preparation. To assess the quality of the ATAC-seq data, we compared the overlap of signals among technical replicates (supplementary fig. S7, Supplementary Material online) and inspected the location of the detected ACRs across the genome.

Transposase Hypersensitive Site and ACR Identification

The ATAC-seq reads were trimmed with Trimmomatic 0.36 (Bolger et al. 2014) and mapped with BWA mem 0.7.17 (Li 2013) to the v1.0 *Capsella rubella* reference genome (Slotte et al. 2013) using default settings. PCR duplicates were removed with the Picard toolkit 2.10.3 (http://broadinstitute.github.io/picard/; last accessed June 2019) and reads with mapping quality below 30 were removed with SAMtools 1.9 (Li et al. 2009; Li 2011), which resulted in an average coverage per sample between 7× and 26.6× (supplementary table S10, Supplementary Material online). THSs were defined as the peak regions called with MACS2 using the -q 0.05 setting (Zhang et al. 2008). MACS2 called between 3,480 and 27,599 THS peaks per sample (supplementary table S11, Supplementary Material online) and peaks within each sample which were within 150 bp from each other were merged together following previous studies (Rodgers-Melnick et al. 2016; Maher et al. 2018). For the population-wide ACR assessment, only THS peaks found in at least two of the three technical replicates were considered as ACR in an individual. Following our peak merging strategy within samples, THS peaks found in different replicates from a single individual were considered identical if they were within 150 bp from each other. Overall, the pairwise comparison of ATAC-seq samples from different plants to two technical replicates revealed a significantly lower overlap between the THS detected in two different individuals than between two technical replicates (*t*-test, df = 49.9, two-sided *P*-value < 0.005). In this study, only ACRs found on one of the eight main scaffolds of the *C. rubella* v1.0 reference genome were included in the downstream analyses, which represented 92% of all the detected ACRs. Furthermore, ACRs present in one individual (<7%) and with no overlap with any other ACRs present in other individuals were labeled uACR, ACRs present in 2 to 13individuals (>10% and <85%) were labeled cACR, ACRs present in at least 14 individuals (>85%) were labeled hACR. Here, ACR present in multiple individuals were only considered identical if they had at least one overlapping base pair. ACRs located in the proximity of genes (2 kb up- and downstream), where most *cis*-regulatory regions can be expected, were labeled as proximal ACR (pACR) and ACRs more than 2 kb away from genes were labeled ACR) following Lu et al. (2019).

Identification of Mappable Regions of the Genome

To be able to compare intergenic ACRs to other nonaccessible intergenic regions of the genome, we first identified the mappable regions of the genome to which potential ATAC-seq reads can be mapped and, therefore, can serve as negative control regions. To do so, 100-bp long reads were generated based on the reference genome using the create-reads-for-te-sequences.py script from PopoolationTE2 v1.10.04 (Kofler et al. 2016). The simulated reads were mapped back to the reference using the same pipeline as used for the ATAC-seq data and the genomic regions with mapped reads were then defined as mappable regions following Lu et al. (2019).

DNA Sequencing

For whole-genome resequencing, DNA was extracted from leaf tissue sampled in a similar way to the ATAC-seq samples (i.e., young and still extending leaves) from the 16 individuals used in the ATAC-seq and from an additional 24 individuals from the same population using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). For DNA library preparation, we used the TruSeq Nano DNA sample preparation kit (Illumina, San Diego, USA) and sequencing (paired-end 150-bp reads, average coverage of 34x per individual) on the Illumina

HiSeqX was performed by the Swedish National Genomics Infrastructure (SNP&SEQ Technology Platform in Uppsala).

Single-Nucleotide Polymorphisms Identification

Whole-genome resequencing reads were trimmed with Trimmomatic 0.36 (Bolger et al. 2014) and mapped with BWA mem 0.7.17 (Li 2013) to the v1.0 *C. rubella* reference genome (Slotte et al. 2013) using the default settings. The mapped reads were then sorted with SAMtools 1.9 (Li et al. 2009) and PCR duplicates were removed with the Picard toolkit 1.118 (http://broadinstitute.github.io/picard). SNPs and invariant sites were called using GATK 4.1.1 (McKenna et al. 2010) with the HaplotypeCaller and GenotypeGVCFs tools and filtered with SelectVariants. For each call site, we requested the following criteria to be met: QD $<$ 5.0; FS $>$ 20.0; SOR $>$ 3.0; MQ $<$ 50.0; $-2.5 >$ MQRankSum $>$ 2.5 and $-2.0 >$ ReadPosRankSum $>$ 2.0. Additionally, we removed sites in each sample with a coverage below 5 and above 200 and sites with more than 20% of missing data in our population were also removed from the analyses. Finally, only sites on the eight main scaffolds of the v.1.0 *C. rubella* reference genome were included in the downstream analyses.

Inferring the Site-Frequency Spectrum

Unfolded site-frequency spectrum (SFS) were generated using est-sfs (Keightley and Jackson 2018) to polarize the filtered sites called by GATK. The polarization was done based on a *C. rubella*, *A. thaliana* and *A. lyrata* whole-genome alignment generated by Steige et al. (2017). Est-sfs was run using *A. thaliana* and *A. lyrata* as the two outgroups and a Jukes-Cantor DNA substitution model (Jukes and Cantor 1969; Keightley and Jackson 2018). The number of divergent sites for each class of sites was estimated through parsimony using the *C. rubella*, *A. thaliana*, and *A. lyrata* whole-genome alignment (Steige et al. 2017).

Quantifying Selection

Per-site nucleotide diversity (π) and Tajima's D in regions of interest were estimated using VCFtools 0.1.15 (Danecek et al. 2011) and we obtained per-site estimates of Watterson's theta (w) based on the number of variant, total sites called and the sample size in each region of interest (Watterson 1975). Divergence between *Capsella* and *Arabidopsis* was estimated based on a *C. rubella*, *A. thaliana*, and *A. lyrata* whole-genome alignment generated by Steige et al. (2017) followed by a Jukes–Cantor correction (Jukes and Cantor 1969). We identified 4-fold and 0-fold degenerate sites as in Steige et al. (2017).

The DFE of new mutations and the proportion of fixed substitutions through positive selection (α) in each region of interest was estimated using DFE-alpha v.2.16 (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009; Schneider et al. 2011). To account for the effect of demography on the estimates of DFE and α, DFE-alpha requires a SFS for putatively neutrally evolving sites (Keightley and Eyre-Walker 2007). Therefore, a set of comparable presumably neutrally evolving sites were defined as follows: 4-fold degenerate sites were considered neutral while estimating DFE and α for 0-fold degenerate sites, whereas when running DFE-alpha on intergenic ACRs, intergenic sites located within 1 kb next to the ACR were considered to be comparable neutral sites. All neutral intergenic sites used in the DFE-alpha analyses were required to be within the mappable regions of the genome, more than 500 bp away from any CNS and not within any other ACR. Here, we defined neutral intergenic sites close to our sites of interests to avoid biases due to differences between the two groups of sites such as different recombination rates, mutation rates, or strong differences in GC content. Unfolded SFS and 100 SFS bootstrap replicates were generated as described above and used as input for DFE-alpha. For each estimate and bootstrap, the run with the highest likelihood out of five independent runs was retained and 95% confidence intervals for the estimated parameters were obtained from 100 bootstrap replicates. We used the one- and two-epoch demographic models implemented in DFE-alpha to account for demographic effects. Results generated when running DFE-alpha with a two-epoch demographic model had higher Akaike Information Criterion (AIC), but the 95% CI of the AIC from the one- and two-epoch models were largely overlapping (supplementary table S3, Supplementary Material online). Therefore, the simplest model (one-epoch model), which assumes a constant population size, was chosen. An inferred constant population size is consistent with previous inference of demographic history in *C. grandiflora* (Foxe et al. 2009; Slotte et al. 2010; St. Onge et al. 2011; Douglas et al. 2015; Mattila et al. 2019).

We also obtained a second estimate of DFE and α using polyDFE v2.0 (Tataru et al. 2017; Tataru and Bataillon 2019). The set of presumably neutrally evolving sites were defined as above. PolyDFE was run as described by Tataru and Bataillon (2020) and all estimates were based on a model averaging approach that included all models (A, B, C, D). For full description of the procedure, see Supplementary Methods, Supplementary Material online. All polyDFE runs were run with and without using the divergence data. The results generated by polyDFE were analyzed and bootstrap-based 95% confidence intervals of the observed DFE and α were generated by generating 100-bootstrap replicate estimates following Tataru and Bataillon (2020).

Intergenic Negative Control Regions

To contrast the observed genetic patterns in pACRs and dACRs to comparable proximal and distal negative control intergenic regions, a set of proximal (2 kb up- and downstream of genes) and distal (more than 2 kb away from genes) intergenic regions were selected using shuffle in BEDtools (Quinlan and Hall 2010). The include option of BEDtools was used to only include regions of the genome in the proximal and distal intergenic negative control regions that were not in genes, ACRs or CNSs. These regions were then labeled as proximal and distal control regions (p-control and d-control) and DFE-alpha was run on these negative controls as described above.

eQTL and CNS Data

In this study, we intersected our ACR with two data sets of previously identified *cis*- and *trans*-eQTL in leaf tissues from another Greek *C. grandiflora* population from the same region (Josephs et al. 2015, 2017, 2020). In these studies, RNA samples were collected from two to three young full-grown leaves from approximately 5-week-old plants during the dark period (Josephs et al. 2015, 2017, 2020). The eQTL data were generated from leaves at a comparable developmental stage to the samples used in this study. The *cis*-eQTL data set analyzed included 4,233 eQTL in genes and 2,233 proximal eQTL (2 kb up- and downstream of genes) located on one of the eight main scaffolds. The *trans*-eQTL data set included 3,686 distal eQTL (at least 5 kb away from genes) located on one of the eight main scaffolds. We also analyzed CNS identified in *C. grandiflora* by Williamson et al. (2014). The CNS data set included 95,182 CNSs located on one of the eight main scaffolds, with a median length of 38 bp (length distribution: 12–658 bp).

Transposable Element Detection

TE insertions in the *C. grandiflora* population were identified with PopoolationTE2 v1.10.04 (Kofler et al. 2016) following the recommended workflow running each individual sample separately. Briefly, a TE-merged reference was generated based on the *C. rubella* TE library published by Slotte et al. (2013) using RepeatMasker 4.0.8 (Smit et al. 2013–2015), as in Horvath and Slotte (2017). The DNA-seq data were mapped to the TE-merged reference with bwa bwasw 0.7.8 (Li and Durbin 2010) and PopoolationTE2 was run with the joint setting and the following requirements for each detected TE insertions: a minimum mapping quality of 15; a maximum proportion of other TE insertions and structural variants of 50% as well as a minimum coverage of 2 for each individual.

Association between ACRs and eQTL, CNS, and TEs

To investigate whether eQTL were associated with ACRs in *C. grandiflora*, we used a permutation approach. Only ACRs in and around genes (5 kb up- and downstream) were included when testing for an association between *cis*-eQTL and ACRs, because the *cis*-eQTL data only included SNPs in these regions (Josephs et al. 2015). When testing for an association between *trans*-eQTL and ACRs only distal *trans*-eQTL and distal ACRs (at least 5 kb away from genes) were included. The observed number of *cis*- and *trans*-eQTL in ACRs was obtained by counting the number of observed *cis*- and *trans*-eQTL overlapping with ACRs. We then permuted *cis*- and *trans*-eQTL designations among all SNPs investigated by Josephs et al. (2015, 2017, 2020) 1,000 times, and for each permutation we recorded the number of random *cis*- and *trans*-eQTL SNPs overlapping with ACRs. To elucidate the biological function of genes affected by eQTL found in intergenic ACRs, a GO term enrichment test was performed on genes affected by eQTL in dACRs and pACRs and their 500 bp surroundings using the topGO R package (Alexa and Rahnenfuhrer 2021). To investigate whether CNSs and TEs were associated with ACRs, the ACRs of interest were shuffled using BEDtools (Quinlan and Hall 2010) to generate 1,000 random samples of the same number of sites within the mappable regions of interest. Two-sided *P*-values were calculated based on the results of the permutation tests. ACR, eQTL, CNS, TE, SNP, and GC content density plots were generated and smoothed over 20-bp windows using deepTools v3.1.0 (Ramírez et al. 2016).

Supplementary Material

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

Acknowledgments

The authors thank Aurélie Desamoré and Benjamin Laenen for seed collection, and Aurélie Desamoré, Sara Mehrabi, and Aleksandra Losvik for their help in the wet lab. The authors acknowledge support from the National Genomics Infrastructure in Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. DNA sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. The computations and data handling were enabled by resources in projects SNIC 2017/7-174, SNIC 2017/7-175, and SNIC 2020/2-27 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. This work was supported by grants from the Swedish Research Council and the Erik Philip-Sörensen foundation awarded to T.S., and a grant from the Science for Life Laboratory, Swedish Biodiversity Program to T.S. The Swedish Biodiversity Program has been made available by support from the Knut and Alice Wallenberg Foundation.

Data Availability

All whole-genome sequences and ATAC-seq data generated in this study have been uploaded to the European Nucleotide Archive ENA under study accession number PRJEB39830.

References

Alexa A, Rahnenfuhrer J. 2021. topGO: enrichment analysis for gene ontology. R package version 2.44.0. Saarbrücken, Germany: Max Planck Institut für Informatik.

Alexandre CM, Urton JR, Jean-Baptiste K, Huddleston J, Dorrity MW, Cuperus JT, Sullivan AM, Bemm F, Jolic D, Arsovski AA, et al. 2018. Complex relationships between chromatin accessibility, sequence divergence, and gene expression in *Arabidopsis thaliana*. *Mol Biol Evol*. 35(4):837–854.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437(7062):1149–1152.

Bajic M, Maher KA, Deal RB. 2018. Identification of open chromatin regions in plant genomes using ATAC-Seq. *Methods Mol Biol*. 1675:183–201.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.

Booker TR, Jackson BC, Keightley PD. 2017. Detecting positive selection in the genome. *BMC Biol*. 15(1):98.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 10(12):1213–1218.

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Prot Mol Biol*. 109:21–29.

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523(7561):486–490.

Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol*. 6(9):R75.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*. 7(2):98–108.

Charlesworth J, Eyre-Walker A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proc Natl Acad Sci U S A*. 104(43):16992–16997.

Chen J, Glémin S, Lascoux M. 2020. From drift to draft: how much do beneficial mutations actually contribute to predictions of Ohta's slightly deleterious model of molecular evolution? *Genetics* 214(4):1005–1018.

Comeron JM, Williford A, Kliman RM. 2008. The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity* 100(1):19–31.

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* 274(5673):775–780.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.

DeRose-Wilson LJ, Gaut BS. 2007. Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of *Arabidopsis thaliana* and *Arabidopsis lyrata*. *BMC Evol Biol*. 7(1):66.

Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, Arunkumar R, Ågren JA, Hazzouri KM, Wang W, et al. 2015. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proc Natl Acad Sci U S A*. 112(9):2806–2811.

Eöry L, Halligan DL, Keightley PD. 2010. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol Biol Evol*. 27(1):177–192.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet*. 8(8):610–618.

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(9):2097–2108.

Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. 2009. Recent speciation associated with the evolution of selfing in *Capsella*. *Proc Natl Acad Sci U S A*. 106(13):5241–5245.

Freeling M, Subramaniam S. 2009. Conserved noncoding sequences (CNSs) in higher plants. *Curr Opin Plant Biol*. 12(2):126–132.

Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol Biol Evol*. 22(3):650–658.

Gilad Y, Rifkin SA, Pritchard JK. 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet*. 24(8):408–415.

Gillespie JH. 2004. Population genetics: a concise guide. 2nd ed. Baltimore (MD): The Johns Hopkins University Press.

Good BH, Walczak AM, Neher RA, Desai MM. 2014. Genetic diversity in the interference selection limit. *PLoS Genet*. 10(3):e1004222.

Gossmann TI, Bockwoldt M, Diringer L, Schwarz F, Schumann VF. 2018. Evidence for strong fixation bias at 4-fold degenerate sites across genes in the great tit genome. *Frontiers Ecol Evol*. 6(203):1–10.

Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol*. 27(8):1822–1832.

Gu W, Wang X, Zhai C, Xie X, Zhou T. 2012. Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. *Mol Biol Evol*. 29(10):3037–3044.

Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res*. 14(2):273–279.

Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res*. 16(7):875–884.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet*. 45(8):891–898.

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res*. 19(8):1419–1428.

Horvath R, Slotte T. 2017. The role of small RNA-based epigenetic silencing for purifying selection on transposable elements in *Capsella grandiflora*. *Genome Biol Evol*. 9(10):2911–2920.

Hough J, Williamson RJ, Wright SI. 2013. Patterns of Selection in Plant Genomes. *Annu Rev Ecol Evol Syst*. 44(1):31–49.

Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, Charlesworth B. 2019. The importance of the neutral theory in 1968 and 50 years on: a response to Kern and Hahn 2018. *Evolution* 73(1):111–114.

Jiang J. 2015. The "dark matter" in the plant genomes: non-coding and unannotated DNA sequences associated with open chromatin. *Curr Opin Plant Biol*. 24(24):17–23.

Joly-Lopez Z, Platts AE, Gulko B, Choi JY, Groen SC, Zhong X, Siepel A, Purugganan MD. 2020. An inferred fitness consequence map of the rice genome. *Nat Plants*. 6(2):119–130.

Josephs EB, Lee YW, Stinchcombe JR, Wright SI. 2015. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *Proc Natl Acad Sci U S A*. 112(50):15390–15395.

Josephs EB, Lee YW, Wood CW, Schoen DJ, Wright SI, Stinchcombe JR. 2020. The evolutionary forces shaping *cis* and *trans* regulation of gene expression within a population of outcrossing plants. *Mol Biol Evol*. 37(8):2386–2393.

Josephs EB, Wright SI, Stinchcombe JR, Schoen DJ. 2017. The relationship between selection, network connectivity, and regulatory variation within a population of *Capsella grandiflora*. *Genome Biol Evol*. 9(4):1099–1109.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 32–132.

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(4):2251–2261.

Keightley PD, Jackson BC. 2018. Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics* 209(3):897–906.

Kern AD, Hahn MW. 2018. The neutral theory in light of natural selection. *Mol Biol Evol*. 35(6):1366–1371.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

King M, Wilson A. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188(4184):107–116.

Kofler R, Gómez-Sánchez D, Schlötterer C. 2016. PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. *Mol Biol Evol*. 33(10):2759–2764.

Kousathanas A, Oliver F, Halligan DL, Keightley PD. 2011. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Mol Biol Evol*. 28(3):1183–1191.

Kreitman M. 1996. The neutral theory is dead. Long live the neutral theory. *Bioessays* 18(8):678–683.

Künstner A, Nabholz B, Ellegren H. 2011. Significant selective constraint at 4-fold degenerate sites in the avian genome and its consequence for detection of positive selection. *Genome Biol Evol*. 3:1381–1389.

Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet*. 9(5):e1003527.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21):2987–2993.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv Preprint ArXiv: https://arxiv.org/abs/1303.3997. Accessed April 7, 2021.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Li XY, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD. 2011. The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biol*. 12(4):R34.

Lu Z, Marand AP, Ricci WA, Ethridge CL, Zhang X, Schmitz RJ. 2019. The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat Plants*. 5(12):1250–1259.

Maher KA, Bajic M, Kajala K, Reynoso M, Pauluzzi G, West DA, Zumstein K, Woodhouse M, Bubb K, Dorrity MW, et al. 2018. Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *Plant Cell*. 30(1):15–36.

Mattila TM, Laenen B, Horvath R, Hämälä T, Savolainen O, Slotte T. 2019. Impact of demography on linked selection in two outcrossing Brassicaceae species. *Ecol Evol*. 9(17):9532–9545.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(9):1297–1303.

Miller W, Makova K, Nekrutenko A, Hardison R. 2004. Comparative genomics. *Annu Rev Genomics Hum Genet*. 5:15–56.

Monroe JG, Srikant T, Carbonell-Bejerano P, Exposito-Alonso M, Weng M-L, Rutter MT, Fenster CB, Weigel D. 2020. Mutation bias shapes gene evolution in *Arabidopsis thaliana*. *bioRxiv*. doi: https://doi.org/10.1101/2929.06.17.156752.

Nguyen TT, Androulakis IP. 2009. Recent advances in the computational discovery of transcription factor binding sites. *Algorithms* 2(1):582–605.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246(5428):96–98.

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst*. 23(1):263–286.

Ohta T. 2002. Near-neutrality in evolution of genes and gene regulation. *Proc Natl Acad Sci U S A*. 99(25):16134–16137.

Ohta T, Gillespie J. 1996. Development of neutral and nearly neutral theories. *Theor Popul Biol*. 49(2):128–142.

Pajoro A, Madrigal P, Muiño JM, Matus JT, Jin J, Mecchia MA, Debernardi JM, Palatnik JF, Balazadeh S, Arif M, et al. 2014. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biol*. 15(3):R41.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 44(W1):W160–W165.

Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. 2016. Open chromatin reveals the functional maize genome. *Proc Natl Acad Sci U S A*. 113(22):E3177–E3184.

Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics* 189(4):1427–1437.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. 2012. A map of the *cis*-regulatory sequences in the mouse genome. *Nature* 488(7409):116–120.

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(8):1813–1821.

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo YL, Steige K, Platts AE, Escobar JS, Newman LK, et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet*. 45(7):831–835.

Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker. Open-4.0. Available from: http://www.repeatmasker.org. Accessed April 7, 2021.

St. Onge KR, Källman T, Slotte T, Lascoux M, Palmé AE. 2011. Contrasting demographic history and population structure in *Capsella rubella* and *Capsella grandiflora*, two closely related species with different mating systems. *Mol Ecol*. 20(16):3306–3320.

Steige KA, Laenen B, Reimegård J, Scofield DG, Slotte T. 2017. Genomic analysis reveals major determinants of *cis*- regulatory variation in *Capsella grandiflora*. *Proc Natl Acad Sci U S A*. 114(5):1087–1092.

Stern DL, Orgogozo V. 2008. The loci of evolution: how predictable is genetic evolution? *Evolution* 62(9):2155–2177.

Tataru P, Bataillon T. 2019. polyDFEv2.0: testing for invariance of the distribution of fitness effects within and across species. *Bioinformatics* 35(16):2868–2869.

Tataru P, Bataillon T. 2020. polyDFE: inferring the distribution of fitness effects and properties of beneficial mutations from polymorphism data. *Methods Mol Biol*. 2090:125–146.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics* 207(3):1103–1119.

Uzunović J, Josephs EB, Stinchcombe JR, Wright SI, Parsch J. 2019. Transposable elements are important contributors to standing variation in gene expression in *Capsella grandiflora*. *Mol Biol Evol*. 36(8):1734–1745.

Vera DL, Madzima TF, Labonne JD, Alam MP, Hoffman GG, Girimurugan SB, Zhang J, McGinnis KM, Dennis JH, Bass HW. 2014. Differential nuclease sensitivity profiling of chromatin reveals biochemical footprints coupled to gene expression and functional DNA elements in maize. *Plant Cell*. 26(10):3883–3893.

Watterson GA. 1975. On the number of segregating sites in genetic models without recombination. *Theor Pop Biol*. 7(2):256–276.

Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLoS Genet*. 10(9):e1004622.

Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet*. 8(3):206–216.

Wright SI, Andolfatto P. 2008. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annu Rev Ecol Evol Syst*. 39(1):193–213.

Zaret K. 2005. Micrococcal nuclease analysis of chromatin structure. *Curr Prot Mol Biol*. 1–17.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 9(9):R137.