The impact of natural selection on the distribution of *cis*-regulatory variation across the genome of an outcrossing plant

Kim A. Steige[1,§], Benjamin Laenen[2,§], Johan Reimegård[3], Douglas G. Scofield[1,4], Tanja Slotte[1,2,*]

[1]Dept. of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyv. 18D, 75236 Uppsala, SWEDEN

[2]Science for Life Laboratory, Dept. of Ecology, Environment and Plant Sciences, Stockholm University, Lilla Frescati, SE-10691 Stockholm, SWEDEN

[3]Science for Life Laboratory, Dept. of Cell and Molecular Biology, Uppsala University, Box 596, 75124 Uppsala, SWEDEN

[4]Uppsala Multidisciplinary Center for Advanced Computational Science, Department of Information Technology, Uppsala University, Box 137, Uppsala 751 05, SWEDEN

[§]These authors contributed equally

*Corresponding author: Tanja Slotte; Tanja.Slotte@su.se

Abstract

Understanding the causes of gene expression variation is of major importance for many areas of biology. While *cis*-regulatory changes have long been suggested to be particularly important for adaptation, our understanding of what determines *cis*-regulatory variation remains limited in most species. Here, we have investigated the prevalence, selective importance, and genomic correlates of *cis*-regulatory variation in the outcrossing crucifer species *Capsella grandiflora*. We identify genes with *cis*-regulatory variation through analyses of allele-specific expression (ASE) in deep transcriptome sequencing data from flower buds and leaves, and use population genomic analyses of high-coverage whole genome resequencing data from both a range-wide sample and a natural population to quantify the impact of positive and purifying selection on these genes. Our results show that in *C. grandiflora*, *cis*-regulatory variation is pervasive, affecting an average of 35% of genes within individual plants. Genes harboring *cis*-regulatory variation are (1) under weaker purifying selection, (2) significantly more likely to harbor nearby transposable element (TE) insertions, and (3) undergo lower rates of adaptive substitutions in comparison to other genes. Using a linear model, we identified ASE as the strongest factor contributing to purifying selection when considered alongside several other commonly used contributing factors. In turn, the main genomic correlates of *cis*-regulatory variation are presence of nearby TE insertions and gene expression level; notably, the signal of relaxed positive and purifying selection on genes with ASE remains after controlling for expression level. Our results suggest that variation in the intensity of selection across the genome is a major determinant of the presence of intraspecific *cis*-regulatory variation in this outcrossing plant species.

Keywords

cis-regulatory changes; allele-specific expression; purifying selection; positive selection; transposable elements; Shepherd's Purse

Introduction

Understanding the causes of regulatory variation is of major importance for many areas of biology and medicine (Albert and Kruglyak 2015). Changes in *cis*-regulatory elements, such as promoters or enhancers, that affect the expression of a focal gene, have long been suggested to be particularly important for adaptation (King and Wilson 1975; Carroll 2000; Wray 2007; Carroll 2008; Wittkopp and Kalay 2012; but see Hoekstra and Coyne 2007). However, in most species we still have a limited understanding of the distribution and degree of *cis*-regulatory variation across the genome and the relative importance of genome-scale evolutionary forces in shaping these patterns.

Due to the development of methods for high-throughput measurement of gene expression, we can now identify *cis*-regulatory variation on a transcriptome-wide scale. This can be done by mapping local expression quantitative trait loci (eQTL), which are likely to be enriched for *cis*-acting regulatory variants, or by directly identifying genes with *cis*-regulatory variation via analysis of allele-specific expression (ASE), because significant allele-specific differences in expression must be due to differences in linked *cis*-regulatory regions (Pastinen 2010; Fraser 2011). Using these methods, ample *cis*-regulatory variation has been identified in many species, including humans (Schadt et al. 2003; Cheung et al. 2005; Stranger et al. 2007; Veyrieras et al. 2008; Pickrell et al. 2010; Lappalainen et al. 2011; Stranger et al. 2012), mice (Doss et al. 2005; Crowley et al. 2015), *Drosophila* (Wittkopp et al. 2008; Massouras et al. 2012), yeast (Brem et al. 2002; Ronald et al. 2005; Skelly et al. 2011), *Caenorhabditis* (Rockman et al. 2010), maize (Stupar and Springer 2006), *Capsella* (Josephs et al 2015), and *Arabidopsis* (Zhang et al. 2011; Lowry et al. 2013).

While most studies thus far have focused on describing the location of variants associated with expression variation in relation to transcription start or end sites, a few have gone farther by identifying other features associated with the presence of ASE or local eQTL. In both yeast (Ronald et al. 2005), *Arabidopsis* (Zhang et al. 2011; Lowry et al. 2013) and *Caenorhabditis* (Rockman et al. 2010), genes with local eQTL are located in primarily in regions with elevated levels of polymorphism. An elegant study in *C. elegans* showed that this was likely because genes with *cis*-regulatory variation are less affected by purifying selection in the form of background selection (Rockman et al. 2010). In line with this, genes with *cis*-regulatory variation were predominantly located in chromosome arms with increased rates of recombination (Rockman et al. 2010). Thus, genome-wide variation in purifying selection can sometimes be more important than gene-specific selective or mutational effects for shaping *cis*-regulatory variation. Similar patterns have been observed in *Arabidopsis thaliana* (Lowry et al. 2013) and it has been suggested that *cis*-regulatory SNPs exhibit a signature of relaxed purifying selection in this selfing species (Zhang et al. 2011). However, in most species, we know little about the impact of positive and purifying selection on genes with empirically-identified *cis*-regulatory variation. Moreover, in outcrossing species, theory predicts that background selection should not have as large an impact on patterns of genomic and regulatory variation as in selfing species such as *C. elegans* and *A. thaliana* (Slotte 2014).

In this study, we have investigated the genomic distribution and selective forces acting on *cis*-regulatory variation in the outcrossing crucifer species *Capsella grandiflora*. This species is an obligate outcrosser with a sporophytic self-incompatibility system similar to that of *Arabidopsis lyrata* (Guo et al. 2009) and is well suited as a model for studying differences in the impact of selection across the genome, as it has relatively low population structure (St Onge et al. 2011) and a large, relatively stable effective population size (Foxe et al. 2009; Slotte et al. 2013). Indeed, selection on both protein-coding (Slotte et al. 2010) and regulatory regions (Williamson et al. 2014) is highly efficient in *C. grandiflora,* and high levels of polymorphism further enhance the power to detect *cis*-regulatory variation and to quantify the impact of selection. Genomic studies are facilitated by the close relationship (split time estimated to <200 kya; Slotte et al 2013) between *C. grandiflora* and the selfing species *Capsella rubella*, for which a genome sequence is available (Slotte et al. 2013).

To investigate the prevalence, genomic correlates and selective importance of *cis*-regulatory variation in *C. grandiflora*, we conducted deep transcriptome sequencing of mRNA from flower buds and leaves, and identified genes with *cis*-regulatory variation based on analyses of ASE. We further obtained high coverage whole genome resequencing data for both population and species-wide samples, to quantify the impact of both positive and purifying selection on genes that harbor *cis*-regulatory variation. Finally, we conduct linear modelling to identify genomic predictors of *cis*-regulatory variation and purifying selection. Our results show that in *C. grandiflora*, *cis*-regulatory variation is pervasive, and genes that harbor standing *cis*-regulatory variation are under weaker purifying selection and experience less frequent positive selection. Thus, variation in the impact of positive and purifying selection across the genome appears to be a major determinant of the presence of intraspecific *cis*-regulatory variation in the outcrosser *C. grandiflora*.

Results

Identification and phasing of SNPs for analysis of ASE

In order to identify genes with *cis*-regulatory variation within *C. grandiflora*, we generated deep whole transcriptome RNAseq data from flower buds and leaves of three *C. grandiflora* F1s resulting from crosses of outbred *C. grandiflora* individuals (total 93.5 Gbp having Q≥30, with 43.1 Gbp for flower buds and 50.4 Gbp for leaves, respectively; Supplementary Table S1). To account for read mapping biases and technical variation in analyses of ASE, we further conducted deep whole genome resequencing of all F1s (mean expected coverage per individual of 40x, total 26.6 Gbp with Q≥30; Supplementary Table S2).

We used a previously established bioinformatic pipeline to identify reliable SNPs for analyses of ASE (Steige et al. 2015). Briefly, we relied on best-practice procedures for variant calling in GATK, coupled with stringent filtering of genomic regions where we had low confidence in our SNP calls (mainly pericentromeric regions; see (Steige et al. 2015) and Methods for details). Using this procedure, we identified an average of 235,719 heterozygous coding SNPs in 17,973 genes in each F1. We then conducted read-backed phasing of genomic SNPs in GATK. This resulted in a mean number of 31,313 contiguous phased fragments per F1, with an average of 8 phased SNPs per fragment (Table 1). We empirically validated this procedure by assessing the proportion of correctly read-back phased SNPs in genomic data for three interspecific *C. grandiflora* x *C. rubella* F1s with known haplotypes genome-wide (inferred through phasing by transmission using genomic data of F1s and their highly homozygous *C. rubella* parents in Steige et al 2015) (see Methods for details). We found that for most genes, the vast majority of SNPs (over 95%) were correctly phased in the interspecific F1s (Figure 1). We therefore proceeded to use the longest contiguous read-phased fragment per gene harboring at least 3 heterozygous SNPs for all subsequent analyses of ASE in *C. grandiflora* F1s. After removing genes which were not detectably expressed, we retained ~14,000 genes for analyses of ASE in flower buds, and ~13,400 genes for analyses of ASE in leaves (Table 1).

ASE results show widespread *cis*-regulatory variation in *C. grandiflora*

We assessed ASE with a Bayesian method that uses genomic reads to account for technical variation in allelic counts and that has a reduced false positive rate compared to the standard binomial test (Skelly et al. 2011). The method requires phased data, and yields direct estimates of the proportion of genes with ASE independent of significance cutoffs, as well as gene-level estimates of the posterior probability of ASE, the magnitude of ASE, and the degree of variability in ASE along a gene.

Using the Skelly et al (2011) method, a mean of 35% (range 26%-39%) of analyzed genes showed ASE (posterior probability of ASE ≥ 0.95) in each of our *C. grandiflora* F1s (Table 1). Similar proportions of genes had evidence for ASE in both leaves and flower buds, and all posterior probability distributions for ASE showed a clear separation between genes with high vs. low posterior probability of ASE (Table 1; Figure 2; Figure 3). Allelic expression biases were moderate for most genes with ASE (Figure 2; Figure 3), with strong strong allelic expression biases (0.2 ≤ ASE ratio ≥ 0.8) shown by an average of just 5.1% of genes. There was little evidence for strong variability in ASE along genes (Figure 2; Figure 3).

While a relatively large proportion of genes showed ASE in individual F1s, most cases of ASE were unique to a particular genotype or sample. Indeed, out of a total of 11,532 genes that were amenable to analysis of ASE in all F1s, only 294 genes had ASE in both leaves and flower buds of all F1s. In total, 1,010 genes showed ASE in either leaves or flower buds, 312 genes showed ASE in flower buds but not leaves, and 404 genes showed ASE in leaf samples but not flower buds of all F1s.

Elevated polymorphism at genes with standing *cis*-regulatory variation

In order to assess the impact of selection on genes showing *cis*-regulatory variation in *C. grandiflora*, we sequenced the genomes of 21 individuals from one population in the Zagory region of Greece (hereafter called the 'population sample') as well as 12 individuals from separate populations across the species range (hereafter called the 'range-wide sample') using paired-end 100 bp Illumina reads and a mean coverage 25x per individual (Supplementary Table S2). We called variants using GATK best practices and filtered genomic regions as previously described (Steige et al. 2015) to identify a total of 6,492,075 high-quality SNPs, most of which (5,240,485) were also segregating in the population sample.

We compared levels of polymorphism at genes that show ASE in all of our F1s (1,010 genes; hereafter 'ASE genes'), using as a control set the 10,552 genes that were amenable to ASE analyses in all F1s but did not show significant ASE (hereafter termed "control"). To reduce bias resulting from the requirement of expressed polymorphisms from analyses of ASE, all population genetic analyses were conducted only on these paired gene sets, and genes that were not amenable to analysis of ASE were not included. ASE genes had elevated polymorphism levels compared to the control at all investigated site classes, as well as an elevated ratio of nonsynonymous to synonymous polymorphism (Table 2; Supplementary Table S3). Control genes without ASE had elevated levels of low frequency polymorphisms at nonsynonymous sites, 5'-UTRs, 3'-UTRs, introns and regions 500 bp upstream of the TSS than those with ASE, suggesting that the impact of purifying selection might differ between ASE and control gene sets (Table 2; Supplementary Table S3).

Reduced intensity of purifying selection on genes with *cis*-regulatory variation

To quantify the impact of purifying selection on ASE genes and control genes, we used the DFE-alpha method (Keightley and Eyre-Walker 2007). Briefly, this method allows estimation of a gamma-distribution of negative fitness effects based on site frequency spectra (SFS) at two classes of sites, one that is assumed to evolve neutrally, and one that is assumed to be subject to selection. Using this method, we found that ASE genes have a significantly higher proportion of nearly neutral nonsynonymous mutations than control genes, as well as a significantly reduced proportion of nonsynonymous mutations under strong purifying selection (strength of purifying selection N_es>10) (Figure 4). This result applies broadly, both for the population and the range-wide samples, and when assuming a constant population size as well as after correcting for population size change (Figure 4). The result also holds after controlling for differences in the expression level among genes with and without ASE (Figure 4). There were no significant differences in the DFE depending on ASE status at 5'-UTRs (Supplementary Figures S1-S4, Supplementary Table S4). Promoter regions 500 bp upstream of the TSS and and 3'-UTRs showed significantly relaxed purifying selection in ASE genes, but this result held only under the 1-epoch model (Supplementary Figures S1-S4, Supplementary Table S4) and could in part be due to a lack of power, as regulatory motifs are expected to make up a small fraction of the analyzed sites. Consistent with this, we infer weaker purifying selection on upstream regions and UTRs than on nonsynonymous mutations (Supplementary Figures S1-S4; Supplementary Table S4).

As many other factors than *cis*-regulatory variation could be associated with variation in positive and purifying selection, we sought to identify factors influencing purifying selection through a general linear model, using the ratio of nonsynonymous to synonymous nucleotide diversity (π_N/π_S) as a proxy for strength of purifying selection. Our model used π_N/π_S as the response variable and several predictors: the presence/absence of ASE as a binary variable; tissue specificity (τ); gene density; gene length; expression level; map-based recombination rate; and divergence at four-fold synonymous sites (d_S) as a proxy for mutation rate (see Methods for details). Model selection by stepwise AIC indicated that the best-fitting model (AIC: 17564) was the one that included all predictors. In this model, ASE had the strongest effect on π_N/π_S (Table 3). The presence of ASE was positively correlated with π_N/π_S, suggesting weaker purifying selection on genes with *cis*-regulatory variation. Tissue specificity showed the same trend, whereas d_S, recombination rate, expression level and gene length were negatively correlated with π_N/π_S (Table 3).

Reduced adaptive evolution at genes with *cis*-regulatory variation

To investigate the impact of positive selection on genes with and without ASE we obtained estimates of ω_α, the rate of adaptive substitutions relative to neutral divergence (Gossmann et al. 2010) in DFE-alpha (Eyre-Walker and Keightley 2009). For this purpose, we relied on genome-wide divergence between *Capsella* and *Arabidopsis*, with 4-fold synonymous sites considered to be evolving mainly neutrally (see Methods for details). Using this method, we find that ASE genes show a significantly lower proportion of adaptive nonsynonymous substitutions than control genes (Figure 5). In contrast, we found no significant differences in ω_α among ASE genes or control genes for UTRs or regions 500 bp upstream of the TSS (Supplementary Table S3). Second, we estimated α, the proportion of adaptive fixations in the selected site class, based on the approximate method of Messer and Petrov (2013) which was designed to yield accurate estimates in the presence of linked selection. Results generated with this method were consistent with DFA-alpha, with a significantly lower estimate of the proportion of adaptive nonsynonymous substitutions at genes with *cis*-regulatory variation than at control genes in *C. grandiflora* (Figure 5).

TE polymorphism is strongly associated with ASE

We have recently shown that TEs targeted by small RNAs are associated with ASE in interspecific *C. grandiflora* x *C. rubella* F1 hybrids (Steige et al. 2015). To assess whether there is also an enrichment of TEs near genes with *cis*-regulatory variation within *C. grandiflora*, we scored heterozygous TE insertions in our F1s as in (Ågren et al. 2014) and tested for an association between heterozygous TE insertions and ASE using Fisher exact tests. On average we detected 1,455 homozygous TE insertions and 1,181 heterozygous TE insertions per *C. grandiflora* F1; the majority of these were retroelements (Supplementary Table S5). There was a significant association between genes with ASE and the presence of heterozygous TE insertions within 1 kb of the gene (Figure 6; Supplementary Table S6). This was true for all F1s when considering flower bud samples, and for two out of three F1s when considering leaf samples.

To test whether polymorphic TEs still had an impact after correcting for other genomic factors, we conducted a logistic regression with presence/absence of ASE as the response variable and a number of predictor variables in addition to polymorphic TEs (see Methods for details). To ensure independence of the TE data from the ASE data, we used TE information gained from the range-wide population sample, which is independent from the specific samples we used to score ASE. We selected the best model using a stepwise AIC procedure.

The best-fitting model (AIC: 3202) included polymorphic TE status, expression level, π_N/π_S, tissue specificity (τ) and promoter polymorphism as predictor variables (Table 4). The presence of polymorphic TEs was the most influential predictor based on the odds ratio; it resulted in a ~40% increase in the odds of observing ASE. The other predictors resulted in an increase of 9%-36% in the odds of observing ASE, and the second most important predictor was τ (Table 4). The presence of polymorphic TEs is thus an important feature associated with *cis*-regulatory variation in *C. grandiflora*.

Discussion

It has long been hypothesized that *cis*-regulatory variation is an important contributor to adaptive evolution, yet the selective forces and genomic correlates of standing *cis*-regulatory variation remain poorly understood in most species. Here, we have shown that there is pervasive *cis*-regulatory variation (via its proxy, ASE) in the outcrossing plant species *Capsella grandiflora*, and that genes with *cis*-regulatory variation are under weaker purifying selection and have undergone a lower proportion of adaptive substitutions than control genes. We found that presence or absence of ASE is a strong predictor of the intensity of purifying selection as measured by the ratio of nonsynonymous to synonymous polymorphism, and ASE is indeed the best predictor when considered alongside several other widely used predictors of purifying selection (Table 3).

The impact of selection on standing *cis*-regulatory variation remains poorly characterized in most systems. Several recent studies have found evidence for a contribution of positive selection to *cis*-regulatory divergence between closely related species (Wittkopp et al. 2008; Fraser et al. 2010; Graze et al. 2012). Our results suggest that, at least for our outcrossing plant species, intraspecific *cis*-regulatory variation is under relaxed positive as well as purifying selection. This finding does not necessarily contradict important contributions of *cis*-regulatory variation to adaptive interspecific evolution. In contrast, it is possible that recurrent sweeps have removed variation specifically at genes without ASE. Supporting this scenario, recent work with the present plant species suggests a general role for recurrent hitchhiking in shaping the distribution of genomic variation (Williamson et al. 2014). In contrast with results for the selfer *C. elegans*, where background selection seems to shape *cis*-

regulatory variation (Rockman et al 2010), we find no clear evidence for clustering of genes with *cis*-regulatory variation in certain chromosomal regions (Supplementary Figures S6-S11).

If our results for *C. grandiflora* hold more generally, this has implications for theoretical modeling of adaptation from *cis*-regulatory variation. For instance, if most standing *cis*-regulatory variation in natural populations is weakly deleterious, models of adaptation from initially weakly deleterious standing variation (e.g. Glémin and Ronfort 2013) would be especially relevant for an improved understanding of the contribution of *cis*-regulatory variation to adaptation. One specific case in which this could be useful would be to aid in our understanding the contribution of *cis*-regulatory changes to the recent adaptive evolution of floral and reproductive traits accompanying the recent shift to selfing in the *C. rubella* (Steige et al. 2015).

Our robust finding of relaxed purifying selection on genes with *cis*-regulatory variation is in good agreement with the results of a recent eQTL mapping study which analyzed 99 individuals from a natural *C. grandiflora* population and found that SNPs associated with expression variation were skewed towards low frequencies, as expected under weak purifying selection (Josephs et al 2015). Our results hold after correction for expression level variation, under different demographic model assumptions, and regardless of whether analyses are conducted on a population sample or a range-wide sample of *C. grandiflora*. Furthermore, our results also hold if we classify genes as ASE or control genes based on a single F1 individual (Supplementary Figure S12). Although many factors have been shown to be correlated with patterns of selection in the genome, when we considered *cis*-regulatory variation (via its proxy, the presence/absence of ASE) alongside several of these factors, we found ASE was the predictor with the largest effect on π_N/π_S. This suggests that, even after accounting for other confounding factors, *cis*-regulatory variation is associated with relaxed purifying selection.

It has recently been suggested that in humans, deleterious nonsynonymous variants can accumulate on the same haplotypes as regulatory variants that result in lower expression, due to their lower penetrance in this regulatory context (Lappalainen et al. 2011). This model seems unlikely to explain our results, as regulatory and coding SNPs are not expected to remain in strong LD in *C. grandiflora* (r^2 decays to less than 0.1 within approximately 500 bp; Supplementary Figure S5). Instead, we suggest that variation in the impact of selection across the genome is more important, and that genes that are generally under weaker selection in *C. grandiflora* are more likely to harbor both *cis*-regulatory and nonsynonymous variation.

A number of recent studies have suggested that TEs may be important for *cis*-regulatory variation and divergence in plants (Hollister and Gaut 2009; Hollister et al. 2011; Wang et al. 2013; Steige et al. 2015). Our results provide tentative support for this conclusion, as we found an enrichment of polymorphic TE insertions in the vicinity of genes with *cis*-regulatory variation, and in our logistic model with presence/absence of ASE as the response, the presence of nearby polymorphic TEs was the strongest factor affecting ASE. These results suggest the importance of TEs in creating ASE under at least some conditions, for instance through effects of TEs silencing on the expression of nearby genes (e.g. Lippman et al. 2004; Hollister and Gaut 2009; Ahmed et al 2011). However, with the currently available data, we cannot rule out the alternative hypothesis that TE insertions have been able to accumulate specifically near genes that are under weaker purifying selection and are also more likely to tolerate nonsynonymous or *cis*-regulatory variation.

In sum, our results suggest that most common standing *cis*-regulatory variation in *C. grandiflora* is under weak purifying selection. Future empirical studies should investigate the impact of TE silencing on *cis*-regulatory variation in *C. grandiflora*, as well as how selection might jointly affect *cis*-regulatory variation and TE accumulation.

Material and Methods

Plant material

For analyses of ASE, we generated three intraspecific *C. grandiflora* F1s by crossing six individuals sampled across the range of *C. grandiflora* (Supplementary Table S7). For validation of our bioinformatic procedures, we also used data from three interspecific F1 individuals from *C. grandiflora* x *C. rubella* F1s that have previously been described (Steige et al. 2015).

For population genomic analyses of *C. grandiflora*, we grew a single offspring from field-collected seeds of each of 32 plants, representing 21 plants from one population near the village of Koukouli in the Zagory region, Greece (the 'population sample'), and 11 additional plants from throughout the species' range representing each of 11 additional Greek populations. Together with an individual from the Koukouli population, these represent a 12-plant 'range-wide sample'. Collectively the 32 plants are termed the 'population genomic sample'. Geographical origins of all samples are given in Supplementary Table S8.

Seeds were surface-sterilized, stratified at 4°C for a week, and germinated on 0.5 x Murashige-Skoog medium. One-week old seedlings were transplanted to pots in soil, which were placed in a growth chamber under long-day conditions (16 h light: 8 h dark; 20° C: 14° C). We collected leaf and mixed stage flower bud samples for RNA sequencing, and leaf samples for whole genome sequencing from all F1 plants, as previously described (Steige et al. 2015). For population genomic analyses, we collected leaf samples for whole genome sequencing from all 32 *C. grandiflora* plants.

Sample preparation and sequencing

We extracted total RNA from all flower bud and leaf samples of the intraspecific F1s using a Qiagen RNEasy Plant Mini Kit (Qiagen, Hilden, Germany). RNAseq libraries were constructed using the TruSeq RNA v2 kit. For genomic resequencing, we extracted predominantly nuclear DNA using a modified CTAB extraction method. Whole genome sequencing libraries with an insert size of 300-400 bp were prepared using the TruSeq DNA v2 protocol. Sequencing of 100bp paired-end reads was performed on an Illumina HiSeq 2000 instrument (Illumina, San Diego, CA, USA) at the Uppsala SNP&SEQ Technology Platform, Uppsala University. In total, we obtained 93.6 Gbp (Q≥30) of RNAseq data, with an average of 15.6 Gbp per sample from intraspecific F1s. In addition we obtained 26.6 Gbp (Q≥30) of DNAseq data, corresponding to a mean expected coverage per individual of 39x for the intraspecific F1s. For population genomic analyses of *C. grandiflora* samples, we obtained a total of 233.2 Gbp (Q≥30) with an average of 7.3 Gbp (Q≥30) per sample. All sequence data has been submitted to the European Bioinformatics Institute (www.ebi.ac.uk), with study accession numbers: PRJEB12070 and PRJEB12072.

Sequence quality and trimming

RNA and DNA reads from the F1s were trimmed as previously described (Steige et al. 2015). For the 32 *C. grandiflora* individuals sequenced for population genomic analyses, we used custom Perl scripts written by DGS to detect adapters and PCR primers present in the raw reads. Adapters and low quality sequence were trimmed using CutAdapt 1.3 (Martin 2011). We analyzed genome coverage using BEDTools v.2.17.0 (Quinlan and Hall 2010) and removed potential PCR duplicates using Picard v.1.92 (http://picard.sourceforge.net).

Read mapping, variant calling and filtering

We mapped RNAseq reads from the F1s to the v1.0 reference *C. rubella* assembly (Slotte et al. 2013) (http://www.phytozome.net/capsella) using STAR v.2.3.0.1 (Dobin et al. 2013) with default parameters. For genomic reads from F1s, we used STAR with settings modified to avoid splitting up reads (see Steige et al. 2015). Genomic reads from the population genomic sample were mapped using BWA-MEM v.0.7.12 (Li 2013) using default parameters and the –M flag.

Variant calling was done using GATK v. 2.5-2 UnifiedGenotyper (McKenna et al. 2010) according to GATK best practices (DePristo et al. 2011; Van der Auwera et al. 2013). We conducted duplicate marking, local realignment around indels and recalibrated base quality scores using a set of 1,538,085 SNPs identified in *C. grandiflora* (Williamson et al. 2014) as known variants and retained only SNPs considered high quality by GATK.

Prior to further analyses, we removed previously identified regions where we have low confidence in our variant calls due to the presence of large-scale copy number variation and repeats; these mainly consist of centromeric and pericentromeric regions (Steige et al. 2015). Before analyses of ASE, we additionally removed SNPs that were in the 1% tails of a beta-binomial distribution fit to all heterozygous SNPs in each F1, as such highly biased SNPs may result in false inference of variable ASE if retained (Skelly et al. 2011). We also removed overlapping parts of genes. For population genomic analyses, we further filtered all genomic regions annotated as repeats using RepeatMasker 4.0.1 (http://www.repeatmasker.org), and removed sites with extreme coverage (DP < 15 or DP > 200) and too many missing individuals (≥20%) using VCFtools (Danecek et al. 2011). Indels and non-biallelic SNP were also pruned prior to any analysis.

Phasing

To allow for ASE analysis based on multiple phased SNPs per gene (see section 'Analyses of allele-specific expression' below), we conducted read-backed phasing of previously annotated genomic variants in both the intraspecific and interspecific F1s using GATK v. 2.5-2 ReadBackPhasing (-phaseQualityThresh 10). RNAseq data from all F1s were subsequently phased by reference to the phased genomic variants. Read counts for all phased fragments were obtained using Samtools mpileup and a custom software written in javascript by JR.

To assess the quality of the read phasing we compared the phased fragments, based on reads, with the phased chromosomes, based on heritage, in three interspecific *C. grandiflora* x *C. rubella* F1s included in a previous study (Steige et al. 2015). For these interspecific F1s chromosome phasing has previously been inferred by reference to whole genome sequences of their highly inbred *C. rubella* parents (Steige et al. 2015). As intra- and interspecific F1s harbored similar numbers of phased SNPs per gene (median of 5 SNPs per gene in both types of F1s; Supplementary Figure S13), the success of the phasing procedure in the interspecific F1s is likely to reflect the phasing success in intraspecific *C. grandiflora* F1s.

Analyses of allele-specific expression

Analyses of allele-specific expression were conducted using a hierarchical Bayesian method developed by Skelly et al (2011). The method requires phased data, in the form of read counts at heterozygous SNPs for both genomic and transcriptomic data. Genomic read counts are used to obtain an empirical estimate of the distribution of technical variation in read counts, which is assumed to follow a beta-binomial distribution. This distribution is subsequently used in analyses of RNAseq data where genes are assigned posterior probabilities of having ASE. The method also results in estimates of the ASE proportion and variation in ASE along the gene.

We analyzed the longest phased fragment per gene with at least three transcribed SNPs. All analyses were run in triplicate, and we checked MCMC convergence by comparing parameter estimates from independent runs with different starting points, and by assessing the mixing of chains. Runs were completed on a high-performance computing cluster at Uppsala University (UPPMAX) using the pqR version of R (http://www.pqr-project.org) for 200,000 generations or a maximum runtime of 10 days. The first 10% of each run was discarded as burn-in and parameter estimates were then obtained as described in Skelly et al (2011).

Identification of TE insertions and association with ASE

To test whether heterozygous TE insertions are associated with ASE in *C. grandiflora,* we used PoPoolationTE (Kofler et al. 2012) and a custom library of TE sequences based on multiple Brassicaceae species (Maumus and Quesneville 2014) to identify TEs in the genomes of our range-wide sample and the intraspecific F1s. We required a minimum of 5 reads to call a TE insertion, and followed the procedure of Ågren et al. (2014) to determine homozygosity or heterozygosity of TE insertions.

Population genomic analyses

In order to assess whether patterns of polymorphism differ among genes with vs. without ASE, we tested for a difference in median levels of polymorphism and Tajima's *D* at all site classes specified above using Mann-Whitney U-tests, with Benjamini-Hochberg correction for multiple comparisons.

Estimates of nucleotide diversity (π), Watterson's theta (θ_W) and Tajima's *D* (D_{T}) were obtained using custom R scripts by BL. Separate estimates were obtained for 6 classes of sites: 4-fold degenerate sites, 0-fold degenerate sites, 3'- and 5'-untranslated regions (UTRs), introns, and intergenic regions 500 bp upstream of the transcription start site (TSS). In order to assess whether species-wide patterns of polymorphism differed from those observed at the population level, we conducted separate analyses on the 12 individuals from the range-wide sample, and the 21 individuals from the population sample.

Selection on genes with ASE

To test whether there was evidence for a difference in the strength and direction of natural selection on sets of genes with and without ASE, we first estimated the distribution of fitness effects (DFE) using the method of Keightley and Eyre-Walker (2007), and the proportion of adaptive selected substitutions relative to the total number of synonymous substitutions (ω_α) using the methods of Eyre-Walker and Keightley (2009) and Gossmann et al (2010). This method allows us to assess the distribution of negative fitness effects (DFE) using the site frequency spectrum (SFS) and corrects for weak purifying selection when estimating ω_α. The DFE was estimated under a constant population size demographic model and under a model with stepwise change in population size between two epochs. We obtained confidence intervals for our estimates of three bins of the DFE ($0<N_es<1$; $1<N_es<10$; $10< N_es$) and for α and ω_α by resampling genes in 200 bootstrap replicates. We tested for a difference in the DFE, and ω_α among sets of genes with ASE (as outlined above) and control genes as in Eyre-Walker and Keightley (2009). Separate estimates were obtained for 0-fold degenerate sites, 3'- and 5'-untranslated regions (UTRs), and regions 500 bp upstream of the TSS likely enriched for regulatory elements. We used both 4-fold degenerate sites as well as introns as the class of sites likely to harbor mainly neutrally evolving variants. For estimates of α and ω_α, we relied on divergence to *Arabidopsis*; specifically, we generated a whole genome alignment using lastz v. 1.03.54 (Harris 2007) with chaining of *C. rubella*, *Arabidopsis thaliana* and *Arabidopsis lyrata* as described in (Haudry et al. 2013), and counted divergence differences and sites as in Williamson et al (2014) for the site categories outlined above. DFE-alpha analyses were run using the method developed by Peter Keightley (Method I in Eyre-Walker and Keightley 2009).

Expression level is one of the most prominent genomic features correlated with purifying selection within plant species (Paape et al. 2013; Williamson et al. 2014) and rates of protein evolution across a broad range of species (e.g. Drummond and Wilke 2008; Larracuente et al. 2008; Slotte et al. 2011). In order to assess the effect of expression level on our DFE-alpha inference, we selected genes among the control set of genes to match the distribution of expression level of ASE genes as follows. For each gene, we obtained the maximum FPKM value among tissues in each F1 individual and then took the average over the three F1s. We divided the distribution in ten bins, excluded the first and last bin to avoid including outliers with very high or low expression level, and then resampled the control genes to match the distribution of expression levels in the ASE gene set. Purifying selection and positive selection were then re-evaluated in DFE-alpha, using the resampled control gene set and the ASE set without the first and last bin, as described above.

In order to test the impact of genomic features on purifying selection we conducted general linear modeling. We used π_N/π_S estimated for the population sample as a proxy for intensity of purifying selection as the response variable and included a suite of genomic predictors including recombination rates, tissue specificity in *A. thaliana* (τ; from Slotte et al. 2011), gene length, expression level (log FPKM value), gene density in 50kb windows, synonymous divergence (d_S) and presence/absence of ASE. Only genes that were amenable to ASE analysis were included in this analysis. Gene length and density were based on the annotation of *C. rubella* v1.0 reference genome (Slotte et al. 2013). We obtained recombination rates per 50kb windows based on 878 markers from (Slotte et al. 2012) by fitting a smooth spline. All the continuous predictors were centered and scaled prior to the regression. We first fit a full model in R and then used a stepwise AIC procedure with backward and forward selection of variables to find the best-fitting model (Table 3).

Linkage disequilibrium decay

To assess the expected LD between regulatory and coding SNPs, we assessed the decay of linkage disequilibrium for the population sample based on r^2 in 2kb windows along each scaffold. The mean r^2 was plotted against physical distance to assess the relative decay of linkage disequilibrium (Supplementary Figure S5). All the calculations were done in plink v. 1.90 (http://pngu.mgh.harvard.edu/purcell/plink/, Purcell et al. 2007).

Relative importance of genomic correlates for *cis*-regulatory variation

We assessed the relative importance of a number of genomic correlates for presence/absence of ASE using logistic regression. The set of analyzed genes was restricted to those for which we could assess

ASE, and we included the following genomic features in our analyses: recombination rate, gene density, tissue specificity (τ), gene length, expression level (log FPKM values), proportion of divergence at synonymous sites (d_S), π for the region 500bp upstream of the TSS and π_N/π_S. All of these variables were obtained as described in "Selection on genes associated with ASE" above. We conducted logistic regression using ‘glm’ in R with model selection using a stepwise AIC procedure with backward and forward selection of variables to find the best-fitting model (Table 4).

Availability of supporting data

All sequence data has been submitted to the European Bioinformatics Institute (www.ebi.ac.uk), with study accession numbers: PRJEB12070 and PRJEB12072.

Description of additional data files

Supplementary Information contains all supplementary Tables and Figures referred to in the text.

Acknowledgements

We thank Lauren McIntyre and Stephen Wright for valuable discussions, Daniel Halligan for sharing scripts for DFE-alpha analyses and Daniel Skelly for advice on ASE analyses. We thank Veronika Scholz and Michael Nowak for bioinformatic assistance, and Julia Dankanich and Cindy Canton for assistance with experiments and lab work. 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 were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under projects b2012122 and b2012190. This study was funded by grants from the Swedish Research Council, the Nilsson-Ehle foundation, the Magnus Bergvall foundation, and the Erik Philip-Sörensen foundation to T.S.

References

Ahmed I, Sarazin A, Bowler C, Colot V, Quesneville H. 2011. Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. Nucleic Acids Res 39:6919–6931.

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. Nat Rev Genet 16:197–212.

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–755.

Carroll SB. 2000. Endless forms: the evolution of gene regulation and morphological diversity. Cell

101:577–580.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134:25–36.

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365–1369.

Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, Kim Y, Wang JR, Morgan AP, Calaway JD, Aylor DL, et al. 2015. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat Genet 47:353-360.

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. 27:2156–2158.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.

Doss S, Schadt EE, Drake TA, Lusis AJ. 2005. *Cis*-acting expression quantitative trait loci in mice. Genome Res 15:681–691.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.

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

Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. 2009. Recent speciation associated with the evolution of selfing in *Capsella*. Proceedings of the National Academy of Sciences 106:5241–5245.

Fraser HB, Moses AM, Schadt EE. 2010. Evidence for widespread adaptive evolution of gene expression in budding yeast. Proceedings of the National Academy of Sciences 107:2977–2982.

Fraser HB. 2011. Genome-wide approaches to the study of adaptive gene expression evolution: systematic studies of evolutionary adaptations involving gene expression will allow many fundamental questions in evolutionary biology to be addressed. Bioessays 33:469–477.

Glémin S, Ronfort J. 2013. Adaptation and maladaptation in selfing and outcrossing species: new mutations versus standing variation. Evolution 67:225–240.

Gossmann TI, Song B-H, 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 27:1822–1832.

Graze RM, Novelo LL, Amin V, Fear JM, Casella G, Nuzhdin SV, McIntyre LM. 2012. Allelic imbalance in Drosophila hybrid heads: exons, isoforms, and evolution. Mol Biol Evol 29:1521–1532.

Guo Y-L, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D, Schierup MH. 2009. Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. Proceedings of the National Academy of Sciences 106:5246–5251.

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 45:891–898.

Harris, R.S. 2007. Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University

Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution 61:995–1016.

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 19:1419–1428.

Hollister JD, Smith LM, Guo Y-L, Ott F, Weigel D, Gaut BS. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proceedings of the National Academy of Sciences 108:2322–2327.

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. Proceedings of the National Academy of Sciences. In press.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107-116.

Kofler R, Betancourt AJ, Schlötterer C. 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet 8:e1002487.

Lappalainen T, Montgomery SB, Nica AC, Dermitzakis ET. 2011. Epistatic selection between coding and regulatory variation in human evolution and disease. Am J Hum Genet 89:459–463.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in *Drosophila*. Trends Genet 24:114–123.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1

Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. Nature 430:471–476.

Lowry DB, Logan TL, Santuari L, Hardtke CS, Richards JH, Derose-Wilson LJ, McKay JK, Sen S, Juenger TE. 2013. Expression Quantitative Trait Locus Mapping across Water Availability Environments Reveals Contrasting Associations with Genomic Features in *Arabidopsis*. Plant Cell 25:3266-3279.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10-12.

Massouras A, Waszak SM, Albarca-Aguilera M, Hens K, Holcombe W, Ayroles JF, Dermitzakis ET, Stone EA, Jensen JD, Mackay TFC, et al. 2012. Genomic variation and its impact on gene expression in *Drosophila melanogaster*. PLoS Genet 8:e1003055.

Maumus F, Quesneville H. 2014. Ancestral repeats have shaped epigenome and genome composition for millions of years in *Arabidopsis thaliana*. Nat Commun 5:4104.

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 20:1297–1303.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. Proceedings of the National Academy of Sciences 110:8615-8620.

Paape T, Bataillon T, Zhou P, J Y Kono T, Briskine R, Young ND, Tiffin P. 2013. Selection, genome-wide fitness effects and evolutionary rates in the model legume *Medicago truncatula*. Mol Ecol 22:3525–3538.

Pastinen T. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet 11:533–538.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464:768–772.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 81:559-575

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

Rockman MV, Skrovanek SS, Kruglyak L. 2010. Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. Science 330:372–376.

Ronald J, Brem RB, Whittle J, Kruglyak L. 2005. Local regulatory variation in *Saccharomyces cerevisiae*. PLoS Genet 1:e25.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297–302.

Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. 2011. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome Res 21:1728–1737.

Slotte T, Bataillon T, Hansen TT, St Onge K, Wright SI, Schierup MH. 2011. Genomic determinants of protein evolution and polymorphism in *Arabidopsis*. Genome Biol Evol 3:1210–1219.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol Biol Evol 27:1813–1821.

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, 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 45:831-835

Slotte T, Hazzouri KM, Stern D, Andolfatto P, Wright SI. 2012. Genetic architecture and adaptive significance of the selfing syndrome in *Capsella*. Evolution 66:1360–1374.

Slotte T. 2014. The impact of linked selection on plant genomic variation. Brief Funct Genomics 13:268–275.

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 20:3306–3320.

Steige KA, Reimegård J, Koenig D, Scofield DG, Slotte T. 2015. *Cis*-Regulatory Changes Associated with a Recent Mating System Shift and Floral Adaptation in *Capsella*. Mol Biol Evol 32:2501–2514.

Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, et al. 2012. Patterns of *cis* regulatory variation in diverse human populations. PLoS Genet 8:e1002639.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, et al. 2007. Population genomics of human gene expression. Nat Genet 39:1217–1224.

Stupar RM, Springer NM. 2006. *Cis*-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. Genetics 173:2199–2210.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 11:11.10.1–11.10.33.

Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4:e1000214.

Wang X, Weigel D, Smith LM. 2013. Transposon variants and their effects on gene expression in *Arabidopsis*. PLoS Genet 9:e1003255.

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:e1004622.

Wittkopp PJ, Haerum BK, Clark AG. 2008. Regulatory changes underlying expression differences within and between *Drosophila* species. Nat Genet 40:346–350.

Wittkopp PJ, Kalay G. 2012. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet 13:59–69.

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

Zhang X, Cal AJ, Borevitz JO. 2011. Genetic architecture of regulatory variation in *Arabidopsis thaliana*. Genome Res 21:725–733.

Ågren JA, Wang W, Koenig D, Neuffer B, Weigel D, Wright SI. 2014. Mating system shifts and transposable element evolution in the plant genus Capsella. BMC Genomics 15:602.

Tables

Table 1. Genes amenable to analysis of ASE in flower buds and leaves and ASE results.

Sample type	F1 designation	Analyzed genes[1]	Genes with ASE[2]	ASE proportion[3]	FDR[4]
Flower buds	Intra6.3	13521	3065	0.33	0.00134
	Intra7.2	14390	3829	0.36	0.00240
	Intra8.2	14232	3601	0.35	0.00198
Leaves	Intra6.3	12390	3425	0.34	0.00182
	Intra7.2	13074	3749	0.39	0.00242
	Intra8.2	12796	3550	0.34	0.00195

[1] Genes with expression data for at least one replicate, and with a phased fragment containing at least three transcribed SNPs after filtering.

[2] Genes with posterior probability of ASE ≥ 0.95.

[3] Estimated proportion of genes with ASE

[4] False Discovery Rate

Table 2. Population genetic summary statistics for the different site classes, separately for ASE and control genes. Estimates are based on the *C. grandiflora* population sample.

Site class[1]	Gene set	Mean π	Mean θ_W	Mean D_{Tajima}	$\pi_{siteclass}/\pi_{4\text{-}fold}$[2]	$N_{DivSites}$[3]	$n_{DivDiff}$[4]	d	$N_{PolSites}$[5]	*S*[6]
4-fold	ASE	0.029	0.029	-0.024	NA	107230	16797	0.16	146996	16244
	control	0.023	0.024	-0.170	NA	1492746	224365	0.15	1717139	157030
0-fold	ASE	0.009	0.011	-0.345	0.32	514975	21169	0.04	660669	24901
	control	0.005	0.007	-0.501	0.23	7137303	245981	0.03	7649524	183401
3'UTR	ASE	0.018	0.021	-0.344	0.62	80721	10296	0.13	133690	10282
	control	0.014	0.018	-0.501	0.62	1004228	119779	0.12	1378898	89674
5'UTR	ASE	0.016	0.016	-0.344	0.55	60109	7261	0.12	87810	5688
	control	0.012	0.012	-0.501	0.54	681489	80787	0.12	851102	48383
500 bp upstream	ASE	0.017	0.020	-0.395	0.6	162790	25389	0.16	314722	25374
	control	0.015	0.019	-0.543	0.68	2113681	314431	0.15	3457930	254809
intron	ASE	0.020	0.022	-0.345	0.69	294445	43827	0.15	456600	39330
	control	0.018	0.020	-0.502	0.79	4520964	632022	0.14	6170854	484589

[1] Class of sites, including 4-fold degenerate sites (4-fold), 0-fold degenerate sites (0-fold), 5'UTRs, 3'UTRs, 500 bp upstream of the TSS (500 bp upstream) and introns.

[2] Ratio of nucleotide diversity at focal site class to nucleotide diversity at 4-fold synonymous sites.

[3] Number of sites assayed for divergence differences.

[4] Number of divergence differences

[5] Number of sites assayed for polymorphisms.

[6] Segregating sites

Table 3. Results of the best-fitting general linear model predicting π_N/π_S from genomic features. Coefficient of the regression and their standard error (SE), z statistics and associated P-value are shown.

	Estimate (SE)	z	P-value
(Intercept)	-0.01 (0.01)	-0.671	0.5023
ASE	0.12 (0.05)	2.491	0.0127*
tissue specificity (τ)	0.07 (0.02)	4.200	0.0000271***
gene density	0.03 (0.01)	2.296	0.0217000*
gene length	-0.03 (0.01)	-2.434	0.0150000*
expression level	-0.11 (0.02)	-7.143	<0.000001***
recombination rate	-0.06 (0.01)	-5.053	0.0000004***
synonymous divergence (d_S)	-0.07 (0.01)	-5.402	0.0000002***

Table 4. Results of the best fit logistic regression model predicting ASE from genomic features. Coefficient of the regression and their standard error (SE), z statistics and associated p-values, and odds ratios (OR) are shown.

	Coeff. (SE)	z	P-value	OR
(Intercept)	-2.66 (0.06)	-47.4	<0.00001***	0.07
TE	0.34 (0.13)	2.68	0.00729**	1.40
tissue specificity (τ)	0.31 (0.06)	5.33	<0.00001***	1.36
expression level	0.21 (0.06)	3.50	0.000472***	1.23
$\pi_{\text{500bp upstream}}$	0.21 (0.05)	4.54	<0.00001***	1.23
π_N/π_S	0.08 (0.04)	2.28	0.022642*	1.09

Figure legends

Figure 1. Success of read-back phasing. The distribution of the proportion of correctly read-back phased SNPs for three interspecific F1s (inter3.1, inter4.1 and inter5.1) with known haplotypes.

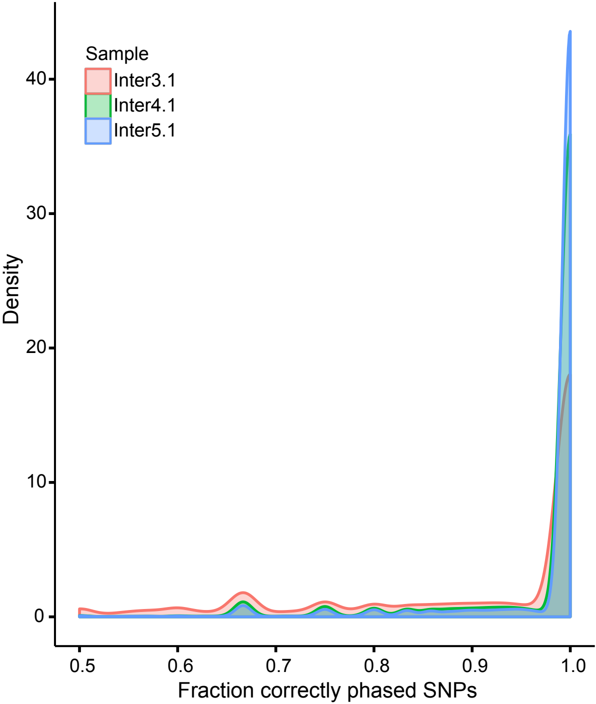
Figure 2. ASE in flower buds of three intraspecific *C. grandiflora* F1s. Distributions of the deviation from equal expression for all assayed genes (A-C) and for genes with at least 0.95 posterior probability of ASE (D-F), estimates of the dispersion parameter (G-I), and the posterior probability of ASE (J-L). All distributions are shown for each of the three intraspecific F1s intra6.3 (left), intra7.2 (middle) and intra8.2 (right).

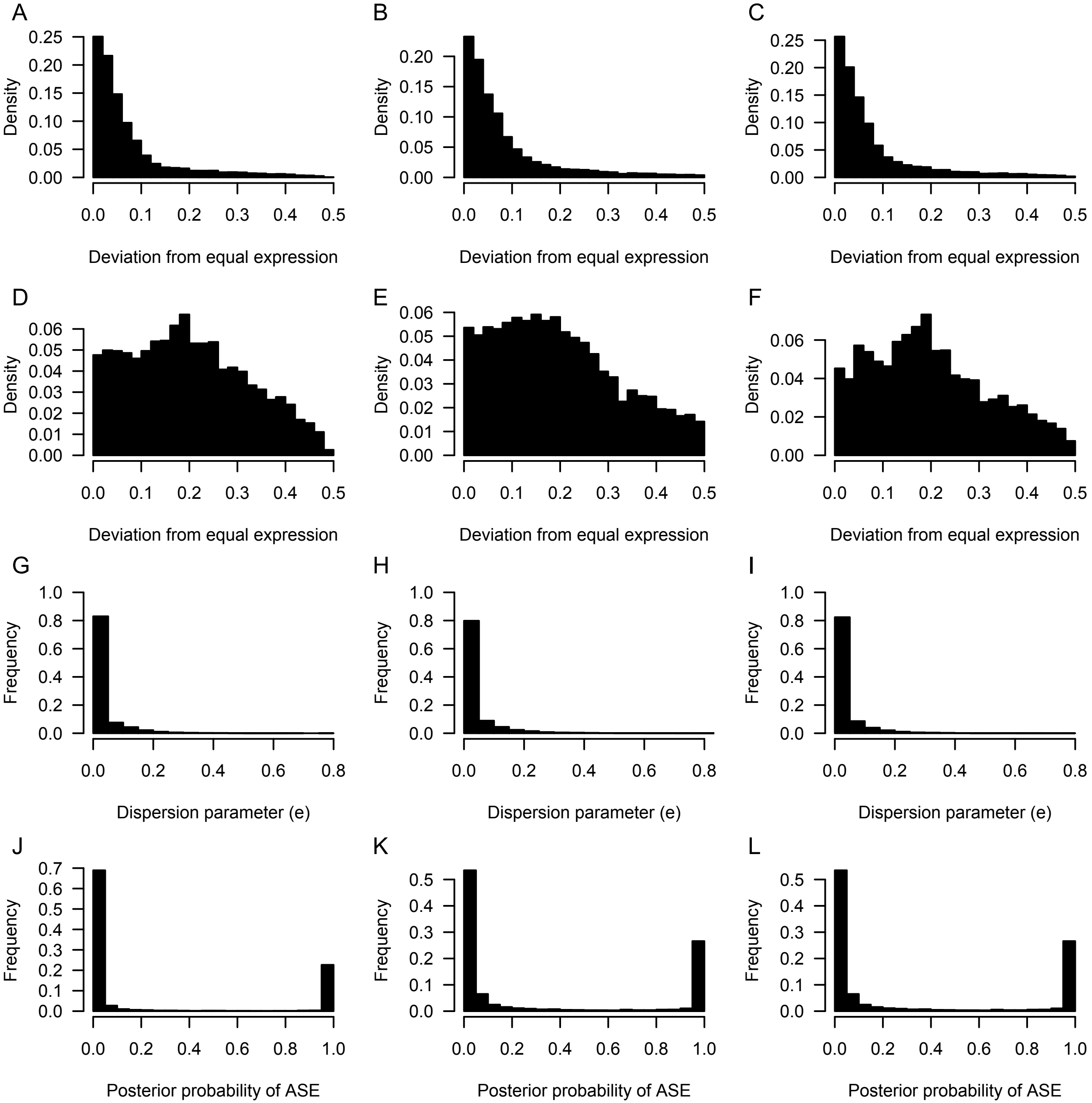
Figure 3. ASE in leaves of three intraspecific *C. grandiflora* F1s. Distributions of the deviation from equal expression for all assayed genes (A-C) and for genes with at least 0.95 posterior probability of ASE (D-F), estimates of the dispersion parameter (G-I), and the posterior probability of ASE (J-L). All distributions are shown for each of the three intraspecific F1s intra6.3 (left), intra7.2 (middle) and intra8.2 (right).

Figure 4. Relaxed purifying selection on genes with ASE in C. grandiflora. The estimated proportion of new nonsynonymous mutations in each bin of the distribution of negative fitness effects is shown, with whiskers corresponding to 95% confidence intervals based on 200 bootstrap replicates, separately for genes with ASE and control genes. Panels A, C, and D show estimates for the population sample, under the (A) two-epoch model, (C) one-epoch model, and (D) two-epoch model, after controlling for expression level differences among ASE and control genes, whereas (B) shows that estimates for the species-wide sample are similar to those for the population sample. Significance levels of the p-value: * ≤ 0.05; ** ≤ 0.01.

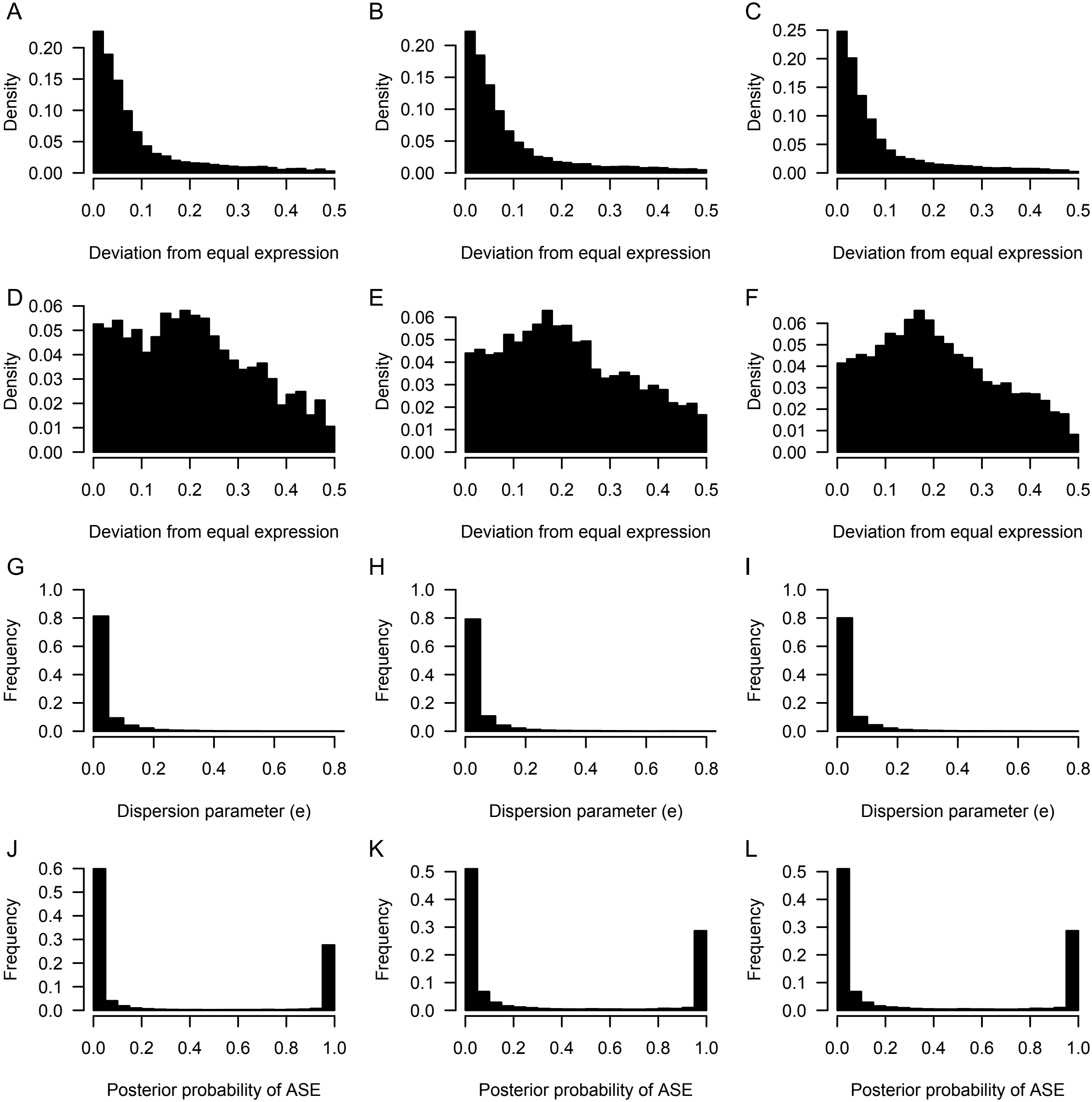
Figure 5. A lower proportion of adaptive nonsynonymous fixations at genes with ASE. (A) The estimated proportion of adaptive fixations relative to 4-fold synonymous substitutions (wa) for genes with and without ASE. Whiskers correspond to 95% confidence intervals based on 200 bootstrap replicates. (B) Estimation of a using the asymptotic method of Messer and Petrov (2013), which fits an exponential function to estimates of a based on polymorphisms at different frequencies. Orange dots show values for control genes, and green dots show values for genes with ASE. The grey shaded area indicates 95% confidence intervals based on 200 bootstrap replicates. The point estimate (aasym) for genes with and without ASE is 0.06 vs 0.28, respectively. Significance levels of the p-value: * ≤ 0.05; ** ≤ 0.01.

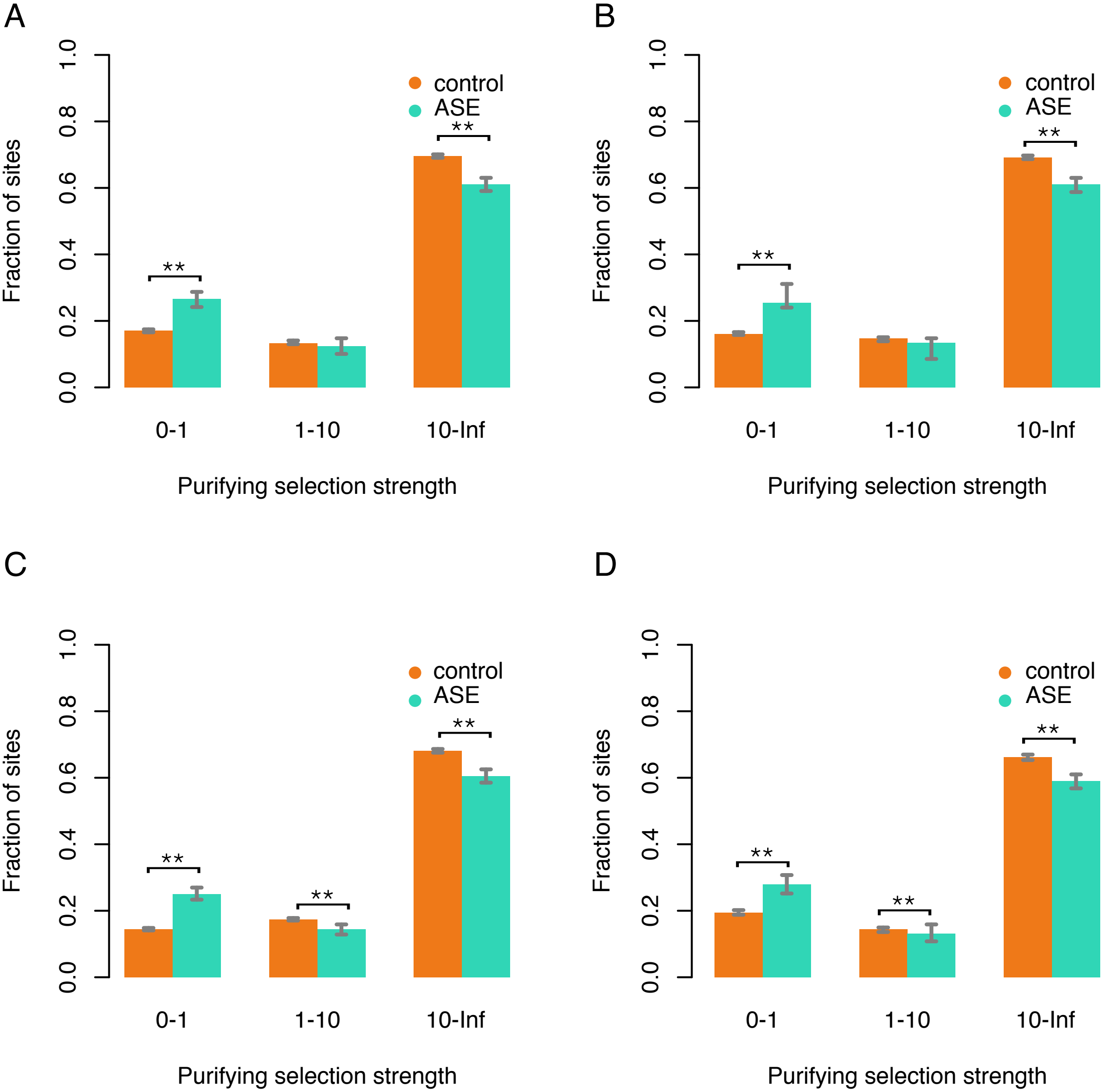
Figure 6. Enrichment of TEs near genes with ASE in *C. grandiflora* F1s. Odds ratios of the association between genes with ASE and TEs, with TE insertions scored in four different window sizes (within a distance of 0 bp, 1 kb, 2 kb, 5 kb and 10 kb of each gene).

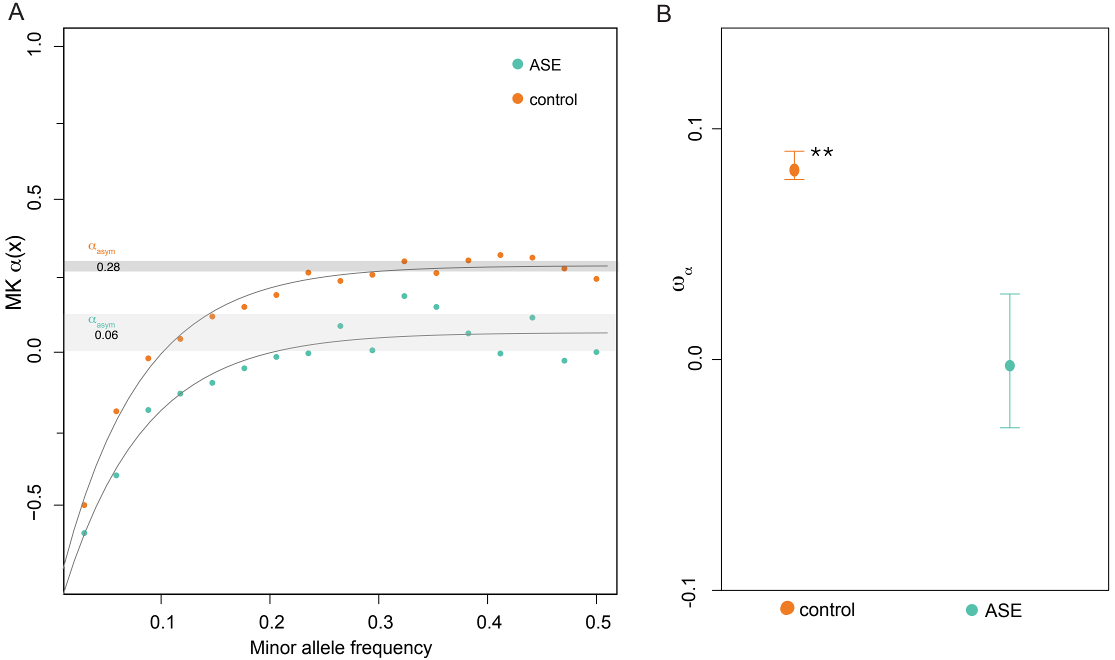

Sample
Inter3.1
Inter4.1
Inter5.1
Density
0
10
20
30
40
0.5
0.6
0.7
0.8
0.9
1.0
Fraction correctly phased SNPs

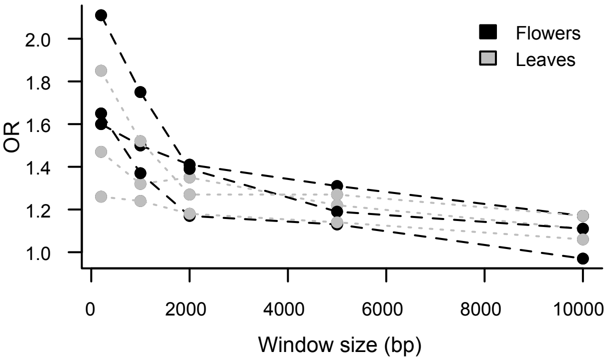


A
B
C
D
E
F
G
H
I
J
K
L
Density
Frequency
Deviation from equal expression
Dispersion parameter (e)
Posterior probability of ASE
0.00
0.05
0.10
0.15
0.20
0.25
0.01
0.02
0.03
0.04
0.06
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1.0


A
B
C
Density
Deviation from equal expression
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.1
0.2
0.3
0.4
0.5
D
E
F
0.01
0.02
0.03
0.04
0.05
0.06
G
H
I
Frequency
Dispersion parameter (e)
0.0
0.2
0.4
0.6
0.8
1.0
J
K
L
Posterior probability of ASE
0.1
0.3
0.5


A
B
C
D
Fraction of sites
0.0
0.2
0.4
0.6
0.8
1.0
control
ASE
**
0-1
1-10
10-Inf
Purifying selection strength


A
B
ASE
control
MK α(x)
1.0
0.5
0.0
−0.5
α_{asym}
0.28
α_{asym}
0.06
0.1
0.2
0.3
0.4
0.5
Minor allele frequency
ω_α
0.1
0.0
-0.1
**
control
ASE


Flowers
Leaves
OR
2.0
1.8
1.6
1.4
1.2
1.0
0
2000
4000
6000
8000
10000
Window size (bp)