Genomic analysis reveals major determinants of *cis*-regulatory variation in *Capsella grandiflora*

Kim A. Steige[1,2,§], 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 to the work.

*Authors for correspondence, email: Benjamin Laenen; Benjamin.Laenen@su.se; Tanja Slotte; Tanja.Slotte@su.se

Keywords
allele-specific expression; distribution of fitness effects; purifying selection; positive selection; transposable elements; gene-body methylation

Understanding the causes of *cis*-regulatory variation is a long-standing aim in evolutionary biology. Although *cis*-regulatory variation has long been considered important for adaptation, we still have a limited understanding of the selective importance and genomic determinants of standing *cis*-regulatory variation. To address these questions, we studied the prevalence, genomic determinants and selective forces shaping *cis*-regulatory variation in the outcrossing plant *Capsella grandiflora*. We first identified a set of 1,010 genes with common *cis*-regulatory variation using analyses of allele-specific expression (ASE). Population genomic analyses of whole-genome sequences from 32 individuals showed that genes with common *cis*-regulatory variation are 1) under weaker purifying selection and 2) undergo less frequent positive selection than other genes. We further identified genomic determinants of *cis*-regulatory variation. Gene-body methylation (gbM) was a major factor constraining *cis*-regulatory variation, whereas presence of nearby TEs and tissue specificity of expression increased the odds of ASE. Our results suggest that most common *cis*-regulatory variation in *C. grandiflora* is under weak purifying selection, and that gene-specific functional constraints are more important for the maintenance of *cis*-regulatory variation than genome-scale variation in the intensity of selection. Our results agree with previous findings that suggest TE silencing affects nearby gene expression, and provide novel evidence for a link between gbM and *cis*-regulatory constraint, possibly reflecting greater dosage-sensitivity of body-methylated genes. Given the extensive conservation of gene-body methylation in flowering plants, this suggests that gene-body methylation could be an important predictor of *cis*-regulatory variation in a wide range of plant species.

52 **Significance**

53 Despite long-standing interest in the contribution of *cis*-regulatory changes to
54 adaptation, we still have a limited understanding of the selective importance and
55 genomic determinants of *cis*-regulatory variation in natural populations. Using a
56 combination of analyses of allele-specific expression and population genomic
57 analyses, we investigate the selective forces and genomic determinants of *cis*-
58 regulatory variation in the outcrossing plant species *Capsella grandiflora*. We
59 conclude that gene-specific functional constraints shape *cis*-regulatory variation and
60 that genes with *cis*-regulatory variation are under relaxed purifying selection
61 compared to other genes. Finally, we identify a novel link between gene-body
62 methylation and the extent of *cis*-regulatory constraint in natural populations.

63

64 **Introduction**

65 Understanding the causes of regulatory variation is of major importance for many
66 areas of biology and medicine (1). Much interest has centered on *cis*-regulatory
67 variation, which has long been thought to be particularly important for adaptation (2-
68 5). Like other quantitative traits, *cis*-regulatory variation is expected to be shaped by
69 the interplay of mutation, selection, and drift. However, the relative importance of
70 these forces remains unclear in most species.

71 Recently, prospects for quantifying *cis*-regulatory variation have greatly
72 improved, and as a result, ample heritable cis-regulatory variation has been identified
73 in many species (reviewed in (6)). This is resulting in a growing consensus that a
74 large amount of standing *cis*-regulatory variation is under weak purifying selection (7-
75 9). Clarifying why the impact of purifying selection varies across the genome is
76 therefore important to understand the maintenance of *cis*-regulatory variation.

77 Variation in the intensity of purifying selection across the genome can result
78 from differences in selective constraint that are due to the specific functions of the
79 genes involved. For example, according to the dosage balance hypothesis, genes that
80 encode interacting proteins are expected to experience stronger constraint than other
81 genes (10). In yeast, there is empirical evidence that purifying selection on expression
82 noise constrains regulatory evolution of dosage-sensitive genes (11-13) and in plants,
83 dosage-sensitivity affects the retention of duplicate genes following whole-genome
84 duplication (14). However, many other genomic features, including expression level,

85 tissue specificity and gene-body methylation (gbM), are also known to be associated
86 with constraint (15-18) and could affect *cis*-regulatory variation.
87 Variation in purifying selection can also result from broad, genome-scale
88 forces that affect genes mainly as a result of their genomic environment, and not due
89 to their specific function. For instance, in the self-fertilizing species *Caenorhabditis*
90 *elegans*, variation in the impact of background selection across the genome had a
91 major effect on the distribution of *cis*-regulatory variation across the genome (8). If
92 background selection is important, then one might generally expect levels of *cis*-
93 regulatory variation to be associated with recombination rate and/or gene density (19).
94 At present however, the relative importance of gene-level constraint vs. genome-scale
95 evolutionary forces for the distribution of *cis*-regulatory variation remains unclear in
96 most species.
97 In this study, we have investigated the selective importance and genomic
98 correlates of common *cis*-regulatory variation in the outcrossing crucifer species
99 *Capsella grandiflora*. This species is particularly well suited for studying differences
100 in the impact of selection across the genome, as it has relatively low population
101 structure (20) and a large, stable effective population size (21, 22). Indeed, selection
102 on both protein-coding (23) and regulatory regions (18) is highly efficient in *C.*
103 *grandiflora,* and high levels of polymorphism enhance the power to detect *cis*-
104 regulatory variation and quantify selection. Genomic studies are facilitated by the
105 close relationship between *C. grandiflora* and the selfing species *Capsella rubella*, for
106 which a genome sequence is available (22).
107 Here, we identified genes with common *cis*-regulatory variation in *C.*
108 *grandiflora* based on analyses of allele-specific expression (ASE) in deep
109 transcriptome sequencing data. To quantify the impact of positive and purifying
110 selection on genes with *cis*-regulatory variation, we conducted population genomic
111 analyses of high-coverage whole genome resequencing data from 32 *C. grandiflora*
112 individuals. Finally, we identified genomic predictors of *cis*-regulatory variation. Our
113 results show that there is pervasive *cis*-regulatory variation in *C. grandiflora*, and
114 genes that harbor *cis*-regulatory variation are under weaker purifying selection and
115 undergo less frequent positive selection than other genes. We find no evidence for a
116 role of recombination rate or gene density in shaping *cis*-regulatory variation,
117 suggesting that gene-specific variation in functional constraint is more important in
118 this species. We further identify gbM as a major factor constraining *cis*-regulatory

119 variation, whereas presence of nearby TEs and tissue specificity of expression
120 increase the odds of ASE. Our results provide novel evidence for a link between gbM
121 and *cis*-regulatory constraint, possibly reflecting greater dosage-sensitivity of body-
122 methylated genes.
123
124 **Results**
125
126 *Widespread cis-regulatory variation in C. grandiflora*
127 To identify genes with *cis*-regulatory variation, we quantified allele-specific
128 expression based on deep whole transcriptome sequencing data (total 93.5 Gbp with
129 Q≥30) from flower buds and leaves of three *C. grandiflora* F1s (Supplementary Table
130 S1). Each F1 harbored an average of about 235,700 high-confidence heterozygous
131 coding SNPs, which were phased prior to analyses of ASE. After filtering, on average
132 approximately 13,400 genes per F1 were amenable to ASE analyses (Table 1).
133 We assessed ASE using a Bayesian method (24), accounting for technical
134 variation in allelic counts using high-coverage whole genome resequencing data for
135 each F1 (mean coverage of 40x, total 26.6 Gbp with Q≥30; Supplementary Table S2).
136 We estimated that a mean of 35% (range 33-39%) of analyzed genes show ASE in
137 individual *C. grandiflora* F1s (Table 1). Similar proportions of genes had ASE in both
138 leaves and flower buds (Table 1) and allelic expression biases were moderate for most
139 genes with ASE, with strong allelic expression biases (0.2 ≤ ASE ratio ≥ 0.8) shown
140 by an average of 5.1% of genes (Figure 1, Supplementary Figures S1 and S2).
141 Out of a total of 11,532 genes that were amenable to analysis of ASE in all
142 F1s, there were 1,010 genes that showed ASE in either leaves or flower buds, 313
143 genes showed ASE in flower buds but not leaves, 404 genes showed ASE in leaf
144 samples but not flower buds, and 293 genes had ASE in both flower buds and leaves
145 of all F1s (Supplementary Figure S3). Among the 1,010 genes with ASE leaves or
146 flower buds of all F1s, one GO category, GO:0006952, "defense response" was
147 significantly enriched at FDR ≤ 0.01. This was likely driven by genes with ASE in
148 leaves, as there was no significant enrichment of Gene Ontology (GO) terms among
149 genes with ASE in flower buds, whereas six biological process GO terms associated
150 with photosynthesis and defense responses were significantly enriched (FDR ≤0.01)
151 among genes with ASE in leaves (Supplementary Table S3). Among control genes,

152 there was a nominally significant enrichment of genes in only two GO terms, protein
153 binding (GO:0005515) and zinc ion binding (GO:0008270) (Weighted Fisher $P \leq$
154 0.01), but this was not significant at FDR $\leq$ 0.01.
155
156 *Lower intensity of purifying selection on genes with cis-regulatory variation*
157 To assess the impact of selection on genes showing *cis*-regulatory variation in *C.*
158 *grandiflora*, we sequenced the genomes of 21 individuals from one population in the
159 Zagory region of Greece (the 'population sample') as well as 12 individuals from
160 separate populations across the species range (the 'range-wide sample') using 233.2
161 Gbp of high-quality (Q≥30) paired-end 100 bp Illumina reads and a mean coverage of
162 25x per individual (Supplementary Table S2). We called variants using GATK best
163 practices and filtered genomic regions as previously described (25) to identify a total
164 of 6,492,075 high-quality SNPs, most of which (5,240,485) were also segregating in
165 the population sample.
166 We compared levels of polymorphism at genes that show ASE in all of our
167 F1s (1,010 genes; 'ASE genes'), using as a control set the 10,552 genes that were
168 amenable to ASE analyses in all F1s but did not show significant ASE (termed
169 'control genes') (Supplementary Figure S3). To reduce bias resulting from the
170 requirement of expressed polymorphisms for analyses of ASE, all population genetic
171 analyses were conducted only on these paired gene sets, and genes that were not
172 amenable to analysis of ASE were not included. ASE genes had elevated
173 polymorphism levels compared to the control at all investigated site classes, as well as
174 an elevated ratio of nonsynonymous to synonymous polymorphism (Table 2;
175 Supplementary Table S4), suggesting that the impact of purifying selection might
176 differ between ASE and control gene sets (Table 2; Supplementary Table S4).
177 To quantify the impact of purifying selection on ASE genes and control genes,
178 we used the DFE-alpha method (26, 27), which allows estimation of a gamma-
179 distribution of negative fitness effects based on site frequency spectra (SFS) at
180 putatively neutral and selected sites. We found that ASE genes have a significantly
181 higher proportion of nearly neutral nonsynonymous mutations than control genes, as
182 well as a significantly reduced proportion of nonsynonymous mutations under strong
183 purifying selection (strength of purifying selection $N_e s$>10) (Figure 2). This result
184 applies broadly, both for the population and the range-wide samples, and when

185 assuming a constant population size as well as after correcting for population size
186 change (Supplementary Figure S4). The result also holds after controlling for
187 differences in the expression level among genes with and without ASE
188 (Supplementary Figure S5), and when classifying genes based on a single F1
189 individual (Supplementary Figure S6), suggesting that the results hold broadly for
190 common *cis*-regulatory variation. Our results further remain unchanged after
191 removing defense-response genes (GO:0006952) with ASE (Supplementary Figure
192 S7) prior to DFE-alpha analyses, and thus strong balancing selection on these genes
193 does not drive the patterns we observe.
194 In contrast to the clear evidence for weaker purifying selection at
195 nonsynonymous sites for genes with ASE, there were no significant differences in the
196 DFE depending on ASE status at 5'-UTRs (Supplementary Figure S8). For introns,
197 results were inconsistent, with some but not all analyses pointing to weaker purifying
198 selection on control genes (Figure 2, Supplementary Figure S9, Supplementary Table
199 S5). This could suggest that patterns of selection differ among coding and noncoding
200 regions. However, at other noncoding regions than introns, such as promoter regions
201 500 bp upstream of the TSS and at 3'-UTRs, there was some evidence for relaxed
202 purifying selection at ASE genes (Figure 2; Supplementary Figures S10-S11,
203 Supplementary Table S5). These results held only under the 1-epoch model, which
204 could in part be due to a lack of power, as regulatory motifs are expected to make up a
205 small fraction of the analyzed sites. Consistent with this, we infer weaker purifying
206 selection on upstream regions and UTRs than on nonsynonymous mutations
207 (Supplementary Figures S8-S11; Supplementary Table S5).
208
209 *Genes with cis-regulatory variation undergo less frequent adaptive evolution*
210 To investigate the impact of positive selection on genes with and without ASE we
211 obtained estimates of ω_a, the rate of adaptive substitutions relative to neutral
212 divergence (28) in DFE-alpha. For this purpose, we relied on genome-wide
213 divergence between *Capsella* and *Arabidopsis*, with 4-fold synonymous sites
214 considered to be evolving mainly neutrally (see Methods for details). Using this
215 method, we find that ASE genes show a significantly lower proportion of adaptive
216 nonsynonymous substitutions than control genes (Figure 3). In contrast, we found no
217 significant differences in ω_a among ASE genes or control genes for UTRs or regions
218 500 bp upstream of the TSS (Supplementary Table S5). Second, we estimated α, the

219 proportion of adaptive fixations in the selected site class, based on the approximate
220 method of (29), designed to yield accurate estimates in the presence of linked
221 selection. Results generated with this method were consistent with DFE-alpha, with a
222 significantly lower estimate of the proportion of adaptive nonsynonymous
223 substitutions at genes with *cis*-regulatory variation than at control genes in *C.*
224 *grandiflora* (Figure 3).
225
226 *Determinants of cis-regulatory variation in C. grandiflora*
227 To identify genomic factors and potential drivers of *cis*-regulatory variation, we
228 conducted logistic regression analyses with presence/absence of ASE as the response
229 variable. We included a total of 12 predictor variables, chosen to include proxies for
230 variation in mutation rate, recombination rate, gene density, expression level and
231 degree of constraint, which could be expected to affect levels of *cis*-regulatory
232 variation (see Methods for details). The best-fit model retained eight of these
233 predictor variables (Table 3). In this model, gbM had the greatest effect on *cis*-
234 regulatory variation, resulting in a reduction of 49% in the odds of observing ASE
235 (Table 3), whereas the presence of polymorphic TEs within 1 kb of the gene also had
236 a substantial effect, increasing the odds of ASE by 38%, followed in turn by tissue
237 specificity of expression, promoter diversity, expression level, gene length and
238 nonsynonymous/synonymous polymorphism, all of which increased the odds of ASE
239 (Table 3). Including network connectivity improved model fit, although the effect was
240 not individually significant (Table 3). Notably, gene density and recombination rate,
241 which affect the intensity of linked selection, were not included in the best-fit model.
242 Similar results were obtained in an analysis that followed the approach of (30) to
243 ensure orthogonality of predictors by using principal components of all continuous
244 predictors in logistic regression analyses (Supplementary Tables S6 and S7). These
245 analyses suggest that variation in gene-specific constraint are important for shaping
246 the distribution of *cis*-regulatory variation across the *C. grandiflora* genome, and that
247 gbM and presence of nearby TEs are strong predictors of *cis*-regulatory constraint.
248
249 **Discussion**
250 Our results show that genes that harbor common *cis*-regulatory variation in *C.*
251 *grandiflora* are under weaker purifying selection, and experience less frequent
252 positive selection than other genes. We further find that gene-specific features that are

253 likely to reflect the degree of functional constraint and mutational input are better
254 predictors of *cis*-regulatory variation than those that are expected to shape the broad
255 impact of linked selection across the genome. These functional constraints do not
256 appear to limit the potential for adaptation at coding sequences, as positive selection
257 had a greater impact on coding divergence at genes that did not exhibit common *cis*-
258 regulatory variation in *C. grandiflora*.

259 Our findings support the view that most standing *cis*-regulatory variation in
260 natural populations is weakly deleterious (7), and our robust inference of relaxed
261 purifying selection on genes with common *cis*-regulatory variation agrees well with
262 those of a recent eQTL mapping study in *C. grandiflora* (9). Our inference of relaxed
263 purifying selection on genes with common *cis*-regulatory variation do not appear to be
264 driven by balancing selection or conditional neutrality affecting a subset of defense-
265 related genes that show ASE, as our results remain unchanged after removing such
266 genes.

267 The major association between gbM and *cis*-regulatory constraint that we
268 detected is particularly interesting, because the function of gbM is currently unclear
269 (31, 32). The conservation of gbM of orthologs in very distantly related plant species
270 suggests that gbM has functional importance, but intriguingly, some plants lack gbM
271 (31-33). Body-methylated genes tend to be longer than other genes, expressed at
272 intermediate levels, evolve slowly at the sequence level (17, 34, 35), and are stably
273 expressed under different conditions (36). A recent study found that *A. thaliana* from
274 northern Sweden show elevated gene body methylation, mainly due to *trans*-acting
275 loci (36), but as far as we are aware, no study has directly linked gbM to *cis*-
276 regulatory variation in natural plant populations.

277 It is possible that the associations we detected between genomic features and
278 *cis*-regulatory variation are caused by underlying drivers that were not directly
279 measured. One natural candidate is gene essentiality. However, while gbM is
280 significantly associated with predicted gene essentiality (37) (Fisher exact test
281 $P<0.001$), our results do not appear to be driven by essentiality, which was not
282 retained in our best-fit logistic regression model for *cis*-regulatory variation. Instead,
283 we hypothesize that selection for increased stability of expression of dosage-sensitive
284 genes could underlie several of the associations we observe. Dosage-sensitive genes
285 exhibit less expression noise (12, 38), show less variation in expression among tissues,
286 and are expected to be part of larger regulatory network modules(10, 12). In our study,

287 reduced tissue-specificity of expression and increased network connectivity were
288 associated with a reduced likelihood of ASE. Furthermore, expression variation
289 among three biological replicates of a *C. rubella* genotype (25) that likely represents
290 mainly noise, is significantly lower for genes with no ASE than for those with ASE
291 (median CV of FPKM=0.28 for genes with ASE, 0.18 for control genes, Wilcoxon
292 rank-sum test, P-value$<10^{-5}$). Finally, defense-related genes, which are thought to be
293 dosage-insensitive in plants (39), were significantly enriched among genes with *cis*-
294 regulatory variation in our study, whereas protein-binding genes were nominally
295 enriched among control genes without ASE. Both promoter polymorphism and TE
296 insertions, which can impact expression in several ways (40), might be expected to be
297 more likely to be tolerated near dosage-insensitive genes. Our results are therefore
298 consistent with dosage sensitivity causing strong constraint on *cis*-regulatory variation
299 and shaping the impact of positive and purifying selection on coding variation. Thus,
300 similar functional constraints that shape duplicate gene retention after whole genome
301 duplication (14) may also be key for the genomic distribution of *cis*-regulatory
302 variation in natural plant populations. Future studies should explore the connection
303 between dosage-sensitivity, gbM, and *cis*-regulatory variation in greater detail across
304 a wider range of plant species.
305
306 **Materials and Methods**
307
308 *Plant Material*
309 For analyses of ASE, we generated three intraspecific *C. grandiflora* F1s by crossing
310 six individuals sampled across the range of *C. grandiflora* (Supplementary Table S8).
311 For population genomic analyses of *C. grandiflora*, we grew a single offspring from
312 field-collected seeds of each of 32 plants ('the population genomic sample';
313 Supplementary Table S9), representing 21 plants from one population from Greece
314 (the ‘population sample’), and 11 additional plants from 11 separate Greek
315 populations covering the species’ range. Together with an individual from the
316 population sample, these represent a 12-plant ‘range-wide sample’. We grew plants at
317 standard long-day conditions and collected leaf and mixed stage flower bud samples
318 for RNA sequencing, and leaf samples for whole genome sequencing as previously
319 described (25).
320

321 *Sample preparation and sequencing*
322 We extracted total RNA from the intraspecific F1s using a Qiagen RNEasy Plant Mini
323 Kit (Qiagen, Hilden, Germany). RNAseq libraries were constructed using the TruSeq
324 RNA v2 kit. For genomic resequencing, we extracted genomic DNA using a modified
325 CTAB extraction method. Whole genome sequencing libraries with an insert size of
326 300-400 bp were prepared using the TruSeq DNA v2 protocol. Sequencing of 100bp
327 paired-end reads was performed on an Illumina HiSeq 2000 instrument (Illumina, San
328 Diego, CA, USA). All sequence data has been submitted to the European
329 Bioinformatics Institute (www.ebi.ac.uk), with study accession numbers:
330 PRJEB12070 and PRJEB12072.
331
332 *Sequence quality and trimming*
333 RNA and DNA reads from the F1s were trimmed as previously described (25).
334 Adapters and low quality sequence were trimmed using CutAdapt 1.3. We analyzed
335 genome coverage using BEDTools v.2.17.0 (41) and removed potential PCR
336 duplicates using Picard v.1.92 (http://picard.sourceforge.net).
337
338 *Read mapping, variant calling and filtering*
339 We mapped RNAseq reads from the F1s to the v1.0 reference *C. rubella* assembly
340 (22) using STAR v.2.3.0.1 (42) with default parameters. For genomic reads from F1s,
341 we mapped reads with STAR as in (25). Genomic reads from the population genomic
342 sample were mapped using BWA-MEM v.0.7.12 (43) using default parameters and
343 the –M flag.
344 Variant calling was done using GATK v. 2.5-2 UnifiedGenotyper (44)
345 according to GATK best practices (45, 46). We conducted duplicate marking, local
346 realignment around indels and recalibrated base quality scores using a set of
347 1,538,085 SNPs identified in *C. grandiflora* (18) as known variants, and retained only
348 SNPs considered high quality by GATK.
349 We removed centromeric and pericentromeric regions where we have low
350 confidence in our variant calls, and prior to ASE analysis, we conducted additional
351 filtering of SNPs as in (25). Using this procedure, we identified an average of 235,719
352 heterozygous coding SNPs in 17,973 genes in each F1. For population genomic
353 analyses, we further filtered all genomic regions annotated as repeats using
354 RepeatMasker 4.0.1, and removed sites with extreme coverage (DP < 15 or DP > 200)

355 and too many missing individuals (≥20%) using VCFtools (47). Indels and non-
356 biallelic SNP were also pruned prior to analysis.
357
358 *Phasing*
359 Prior to ASE analysis, we conducted read-backed phasing of genomic variants in F1s
360 using GATK v. 2.5-2 ReadBackPhasing (-phaseQualityThresh 10). RNAseq data
361 from all F1s were subsequently phased by reference to the phased genomic variants.
362 Read counts for all phased fragments were obtained using Samtools mpileup. This
363 resulted in a mean number of 31,313 contiguous phased fragments per F1 (Table 1).
364 To validate our phasing procedure, we compared the phased fragments, based
365 on reads, with the phased chromosomes, based on heritage, in three interspecific *C.*
366 *grandiflora* x *C. rubella* F1s from (25). For most genes, over 95% of SNPs were
367 correctly phased in the interspecific F1s, demonstrating that our phasing procedure is
368 reliable (Supplementary Figure S13; Supplementary Figure S14).
369
370 *Analyses of allele-specific expression*
371 We analyzed ASE using a hierarchical Bayesian method which requires phased data,
372 in the form of read counts at heterozygous SNPs for both genomic and transcriptomic
373 data (24). Genomic read counts are used to obtain an empirical estimate of technical
374 variation which is then used in analyses of the RNAseq data. We used this method to
375 obtain estimates of the posterior probability and degree of ASE, for the longest phased
376 fragment per gene with at least three transcribed SNPs. We analyzed ~14,000 reliably
377 expressed genes for ASE in flower buds, and ~13,400 genes in leaves (Table 1). All
378 analyses were run in triplicate, and we checked MCMC convergence by comparing
379 parameter estimates from independent runs with different starting points, and by
380 assessing mixing. Runs were completed using the pqR version of R (http://www.pqr-
381 project.org) for 200,000 generations or a maximum runtime of 10 days, with the first
382 10% of each run discarded as burn-in.
383
384 *Population genomic analyses*
385 To assess whether patterns of polymorphism differ among ASE and control genes, we
386 tested for a difference in median levels of polymorphism and Tajima's *D* in the *C.*
387 *grandiflora* population sample, using Mann-Whitney U-tests, with Benjamini-
388 Hochberg correction for multiple comparisons. Estimates of nucleotide diversity (π),

389 Watterson's theta (θ_W) and Tajima's D (D_T) were obtained using custom R scripts by
390 BL. Separate estimates were obtained for 6 classes of sites: 4-fold degenerate sites, 0-
391 fold degenerate sites, 3'- and 5'-untranslated regions (UTRs), introns, and intergenic
392 regions 500 bp upstream of the transcription start site (TSS).
393
394 *Selection on genes with ASE*
395 To test whether there was evidence for a difference in the strength and direction of
396 natural selection on ASE and control genes, we first estimated the distribution of
397 fitness effects (DFE) as in (26), and the proportion of adaptive substitutions relative to
398 the total number of synonymous substitutions (ω_a) using the method of (28). The DFE
399 was estimated under a constant population size model and under a model with
400 stepwise population size change. We obtained confidence intervals for our estimates
401 of three bins of the DFE ($0<N_es<1$; $1<N_es<10$; $10< N_es$) and for α and ω_a by
402 resampling genes in 200 bootstrap replicates and tested for a difference in the DFE,
403 and ω_a among sets of genes with ASE and control genes, as in (27). Separate
404 estimates were obtained for 0-fold degenerate sites, 3'- and 5'-untranslated regions
405 (UTRs), introns, and promoter regions 500 bp upstream of the TSS likely enriched for
406 regulatory elements, using 4-fold degenerate sites as neutral standard. For estimates of
407 α and ω_a, we relied on divergence to *Arabidopsis*; specifically, we generated a whole
408 genome alignment using lastz v. 1.03.54, with chaining of *C. rubella*, *A. thaliana* and
409 *A. lyrata* as described in (48), and counted divergence differences and sites as in (18).
410 DFE-alpha analyses were run using Method I (27).
411 To assess the effect of expression level on our DFE-alpha inference, we
412 selected genes among the control set of genes to match the distribution of expression
413 level of ASE gene by resampling the control genes to match the distribution of
414 expression levels in the ASE gene set. Purifying selection and positive selection were
415 then re-evaluated in DFE-alpha, using the resampled control gene set and the ASE set.
416 To assess whether our results were robust to the sampling strategy for ASE analyses,
417 we based our classification of ASE and control genes based on a single F1 individual
418 and repeated the DFE analyses. To investigate whether our results could be driven by
419 the inclusion of defense-related genes, we removed genes annotated as defense-
420 response genes (GO:0006952), and repeated the DFE-alpha analyses.
421
422

423 *Genomic determinants of cis-regulatory variation*
424 We assessed the relative importance of a number of genomic features for
425 presence/absence of ASE using logistic regression, on a set of genes that was
426 restricted to those for which we could assess ASE. We included the following
427 genomic features that may affect linked selection: recombination rate and gene
428 density (in 50 kb windows). Gene density were based on the annotation of *C. rubella*
429 v1.0 reference genome (22). We obtained recombination rates per 50kb windows
430 based on 878 markers from (49) by fitting a smooth spline. We further included gene
431 length, tissue specificity (*t*;(16)), expression level (log FPKM values), and as a proxy
432 for mutation rate variation, we included 4-fold synonymous divergence to
433 Arabidopsis (d_S). Because promoter polymorphism may cause *cis*-regulatory variation,
434 we included nucleotide diversity (π) for the region 500bp upstream of the TSS. We
435 included nonsynonymous/synonymous nucleotide diversity (π_N/π_S) to reflect the level
436 of constraint at the coding sequence level. According to the dosage balance
437 hypothesis, genes in smaller co-expression modules may be under reduced regulatory
438 constraint. We therefore included information on *A. thaliana* co-expression module
439 size (37) in our analyses. We further included information on the presence of retained
440 paralogs from the Brassicaceae α whole genome duplication or the β and γ whole
441 genome duplication (37). We identified a set of genes with gbM in both *C. rubella*
442 (33) and *A. thaliana* (17), which are highly likely to also harbor gbM in *C.*
443 *grandiflora*. Finally, we included information on polymorphic TEs within 1 kb of
444 genes in the range-wide sample. We identified TE insertions in our range-wide
445 sample as in (25), except that we required a minimum of 5 reads to call a TE insertion.
446 All continuous variables were centered and scaled prior to logistic regression, with
447 model selection using a stepwise AIC procedure with backward and forward selection
448 of variables to find the best-fit model (Table 3). We repeated the analysis using an
449 analysis strategy which is superior to partial correlation analysis and robust in the
450 presence of noisy genomic data and multicollinearity of predictor variables (30). We
451 used a set of orthogonal predictor variables obtained by identifying principal
452 components for a data set including all the continuous variables using the "pls"
453 package in R, well as gene body methylation and presence of heterozygous TEs as
454 binary factors, and conducted model selection as described above.
455
456 **Acknowledgements**

457 We thank Lauren McIntyre and Stephen Wright for valuable discussion, Daniel
458 Halligan for sharing scripts for DFE-alpha analyses and Daniel Skelly for advice on
459 ASE analyses. We thank Veronika Scholz and Michael Nowak for bioinformatic
460 assistance, and Julia Dankanich and Cindy Canton for assistance with experiments
461 and lab work. Sequencing was performed by the SNP&SEQ Technology Platform in
462 Uppsala. The computations were performed on resources provided by SNIC through
463 Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX)
464 under projects b2012122 and b2012190. This study was funded by grants from the
465 Swedish Research Council, the Nilsson-Ehle foundation, the Magnus Bergvall
466 foundation, and the Erik Philip-Sörensen foundation to T.S.
467

468 **References**

469 1. Albert FW, Kruglyak L (2015) The role of regulatory variation in complex
470 traits and disease. *Nat Rev Genet* 16(4):197–212.
471 2. King MC, Wilson AC (1975) Evolution at two levels in humans and
472 chimpanzees. *Science* 188(4184):107-116
473 3. Wray GA (2007) The evolutionary significance of cis-regulatory mutations.
474 *Nat Rev Genet* 8(3):206–216.
475 4. Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a
476 genetic theory of morphological evolution. *Cell* 134(1):25–36.
477 5. Wittkopp PJ, Kalay G (2012) Cis-regulatory elements: molecular mechanisms
478 and evolutionary processes underlying divergence. *Nat Rev Genet* 13(1):59–69.
479 6. Fraser HB (2011) Genome-wide approaches to the study of adaptive gene
480 expression evolution. *Bioessays* 33(6):469–477.
481 7. Fay JC, Wittkopp PJ (2008) Evaluating the role of natural selection in the
482 evolution of gene regulation. *Heredity* 100(2):191–199.
483 8. Rockman MV, Skrovanek SS, Kruglyak L (2010) Selection at linked sites
484 shapes heritable phenotypic variation in C. elegans. *Science* 330(6002):372–
485 376.
486 9. Josephs EB, Lee YW, Stinchcombe JR, Wright SI (2015) Association mapping
487 reveals the role of purifying selection in the maintenance of genomic variation
488 in gene expression. *Proc Natl Acad Sci USA* 112(50):15390-15395.
489 10. Birchler JA, Veitia RA (2012) Gene balance hypothesis: connecting issues of
490 dosage sensitivity across biological disciplines. *Proc Natl Acad Sci USA*

491 109(37):14746–14753.
492 11. Lemos B, Meiklejohn CD, Hartl DL (2004) Regulatory evolution across the
493 protein interaction network. *Nat Genet* 36(10):1059–1060.
494 12. Lehner B (2008) Selection to minimise noise in living systems and its
495 implications for the evolution of gene expression. *Mol Syst Biol* 4(1):170.
496 13. Metzger BPH, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ (2015) Selection
497 on noise constrains variation in a eukaryotic promoter. *Nature* 521(7552):344-
498 347.
499 14. Li Z, et al. (2016) Gene duplicability of core genes is highly consistent across
500 all angiosperms. *Plant Cell* 28(2):326–344.
501 15. Rocha EPC (2006) The quest for the universals of protein evolution. *Trends*
502 *Genet* 22(8):412–416.
503 16. Slotte T, et al. (2011) Genomic determinants of protein evolution and
504 polymorphism in Arabidopsis. *Genome Biol Evol* 3:1210–1219.
505 17. Takuno S, Gaut BS (2012) Body-methylated genes in *Arabidopsis thaliana* are
506 functionally important and evolve slowly. *Mol Biol Evol* 29(1):219–227.
507 18. Williamson RJ, et al. (2014) Evidence for widespread positive and negative
508 selection in coding and conserved noncoding regions of *Capsella grandiflora.*
509 *PLoS Genet* 10(9):e1004622.
510 19. Slotte T (2014) The impact of linked selection on plant genomic variation.
511 *Brief Funct Genomics* 13(4):268–275.
512 20. St Onge KR, Källman T, Slotte T, Lascoux M, Palmé AE (2011) Contrasting
513 demographic history and population structure in *Capsella rubella* and *Capsella*
514 *grandiflora*, two closely related species with different mating systems. *Mol*
515 *Ecol* 20(16):3306–3320.
516 21. Foxe JP, et al. (2009) Recent speciation associated with the evolution of selfing
517 in *Capsella*. *Proc Natl Acad Sci USA* 106(13):5241–5245.
518 22. Slotte T, et al. (2013) The *Capsella rubella* genome and the genomic
519 consequences of rapid mating system evolution. *Nat Genet*. 45:831-835.
520 23. Slotte T, Foxe JP, Hazzouri KM, Wright SI (2010) Genome-wide evidence for
521 efficient positive and purifying selection in *Capsella grandiflora*, a plant
522 species with a large effective population size. *Mol Biol Evol* 27(8):1813–1821.
523 24. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM (2011) A powerful
524 and flexible statistical framework for testing hypotheses of allele-specific gene

525 expression from RNA-seq data. *Genome Res* 21(10):1728–1737.
526 25. Steige KA, Reimegård J, Koenig D, Scofield DG, Slotte T (2015) *Cis*-
527 regulatory changes associated with a recent mating system shift and floral
528 adaptation in *Capsella*. *Mol Biol Evol* 32(10):2501–2514.
529 26. Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of
530 fitness effects of deleterious mutations and population demography based on
531 nucleotide polymorphism frequencies. *Genetics* 177(4):2251–2261.
532 27. Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular
533 evolution in the presence of slightly deleterious mutations and population size
534 change. *Mol Biol Evol* 26(9):2097–2108.
535 28. Gossmann TI, et al. (2010) Genome wide analyses reveal little evidence for
536 adaptive evolution in many plant species. *Mol Biol Evol* 27(8):1822–1832.
537 29. Messer PW, Petrov DA (2013) Frequent adaptation and the McDonald-
538 Kreitman test. *Proc Natl Acad Sci USA* 110(21):8615-8620.
539 30. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates
540 the rate of yeast protein evolution. *Mol Biol Evol* 23(2):327–337.
541 31. Takuno S, Ran J-H, Gaut BS (2016) Evolutionary patterns of genic DNA
542 methylation vary across land plants. *Nature Plants* 2(2):15222.
543 32. Bewick AJ, et al. (2016) On the origin and evolutionary consequences of gene
544 body DNA methylation *Proc Natl Acad Sci USA* doi:
545 10.1073/pnas.1604666113
546 33. Niederhuth CE, et al. (2016) Widespread natural variation of DNA methylation
547 within angiosperms *bioRxiv* doi:10.1101/045880.
548 34. Takuno S, Gaut BS (2013) Gene body methylation is conserved between plant
549 orthologs and is of evolutionary consequence. *Proc Natl Acad Sci USA*
550 110(5):1797–1802.
551 35. Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D (2014) Evolution of
552 DNA methylation patterns in the Brassicaceae is driven by differences in
553 genome organization. *PLoS Genet* 10(11):e1004785.
554 36. Dubin MJ, et al. (2015) DNA methylation in Arabidopsis has a genetic basis
555 and shows evidence of local adaptation. *Elife* 4:e05255.
556 37. Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H (2015)
557 Characteristics of plant essential genes allow for within- and between-species
558 prediction of lethal mutant phenotypes. *Plant Cell* 27(8):2133–2147.

559 38. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise
560 minimization in eukaryotic gene expression. *Plos Biol* 2(6):e137.
561 39. Coate JE, Song MJ, Bombarely A, Doyle JJ (2016) Expression-level support
562 for gene dosage sensitivity in three *Glycine* subgenus *Glycine* polyploids and
563 their diploid progenitors. *New Phytol*. doi:10.1111/nph.14090.
564 40. Feschotte C (2008) Transposable elements and the evolution of regulatory
565 networks. *Nat Rev Genet* 9(5):397–405.
566 41. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for
567 comparing genomic features. *Bioinformatics* 26(6):841–842.
568 42. Dobin A, et al. (2013) STAR: ultrafast universal RNA-seq aligner.
569 *Bioinformatics* 29(1):15–21.
570 43. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs
571 with BWA-MEM. *arXiv* 1303.3997v2.
572 44. McKenna A, et al. (2010) The Genome Analysis Toolkit: a MapReduce
573 framework for analyzing next-generation DNA sequencing data. *Genome Res*
574 20(9):1297–1303.
575 45. DePristo MA, et al. (2011) A framework for variation discovery and
576 genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–
577 498.
578 46. Van der Auwera GA, et al. (2013) From FastQ data to high confidence variant
579 calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc*
580 *Bioinformatics* 11(1110):11.10.1–11.10.33.
581 47. Danecek P, et al. (2011) The variant call format and VCFtools. *Bioinformatics*
582 27(15):2156–2158.
583 48. Haudry A, et al. (2013) An atlas of over 90,000 conserved noncoding
584 sequences provides insight into crucifer regulatory regions. *Nat Genet*
585 45(8):891–898.
586 49. Slotte T, Hazzouri KM, Stern D, Andolfatto P, Wright SI (2012) Genetic
587 architecture and adaptive significance of the selfing syndrome in *Capsella*.
588 *Evolution* 66(5):1360–1374.
589

590 **Figure Legends**

591

592 Figure 1. The extent of ASE in leaves (A, B) and flower buds (C, D) of one
593 representative *C. grandiflora* F1. Panels A and C show the deviation from equal
594 expression for all assayed genes. Genes with strong evidence for ASE (posterior
595 probability of ASE ≥ 0.95) show stronger deviations from equal expression (B and D).

596

597 Figure 2. The impact of purifying selection differs between genes with and without
598 ASE in *C. grandiflora.* The estimated proportion of mutations in each $N_e s$ bin of the
599 distribution of negative fitness effects (DFE) is shown, with whiskers corresponding
600 to 95% confidence intervals. The panels show the DFE for nonsynonymous sites (0-
601 fold degenerate sites) (A), for introns (B), for promoter regions 500 bp upstream of
602 the transcription start site (C) and for 3'-UTRs (D). Significance levels for
603 comparisons of ASE and control genes, that were also amenable to analysis of ASE,
604 are indicated by asterisks (*: P-value ≤ 0.05, ** P-value ≤ 0.01). These results are
605 based on the population sample and the one-epoch model.

606

607 Figure 3. A lower proportion of adaptive nonsynonymous fixations at genes with ASE.
608 (A) Estimation of α using the asymptotic method of Messer and Petrov (2013), which
609 fits an exponential function to estimates of α based on polymorphisms at different
610 frequencies. Orange dots show values for control genes, and green dots show values
611 for genes with ASE. The grey shaded area indicates 95% confidence intervals. The
612 point estimate for genes with and without ASE is 0.06 vs 0.28, respectively. (B) The
613 estimated proportion of adaptive fixations relative to 4-fold synonymous substitutions
614 (ω_a) for genes with and without ASE. Whiskers correspond to 95% confidence
615 intervals, and significance levels for comparisons of ASE and control genes are
616 indicated by asterisks (*: P-value ≤ 0.05, ** P-value ≤ 0.01).

617

618 **Tables**

619

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

Sample type	F1	Analyzed[1]	ASE genes[2]	ASE prop.[3]	FDR[4]
Flower buds	6.3	13521	3065	0.33	0.00134
	7.2	14390	3829	0.36	0.00240
	8.2	14232	3601	0.35	0.00198
Leaves	6.3	12390	3425	0.34	0.00182
	7.2	13074	3749	0.39	0.00242
	8.2	12796	3550	0.34	0.00195

622

623 [1]Genes with expression data for at least one replicate, and with a phased fragment
624 containing at least three transcribed SNPs after filtering.
625 [2]Genes with posterior probability of ASE ≥ 0.95.
626 [3]Estimated proportion of genes with ASE
627 [4]False Discovery Rate

628 Table 2. Population genetic summary statistics and divergence estimates for the
629 different site classes, separately for ASE and control genes.

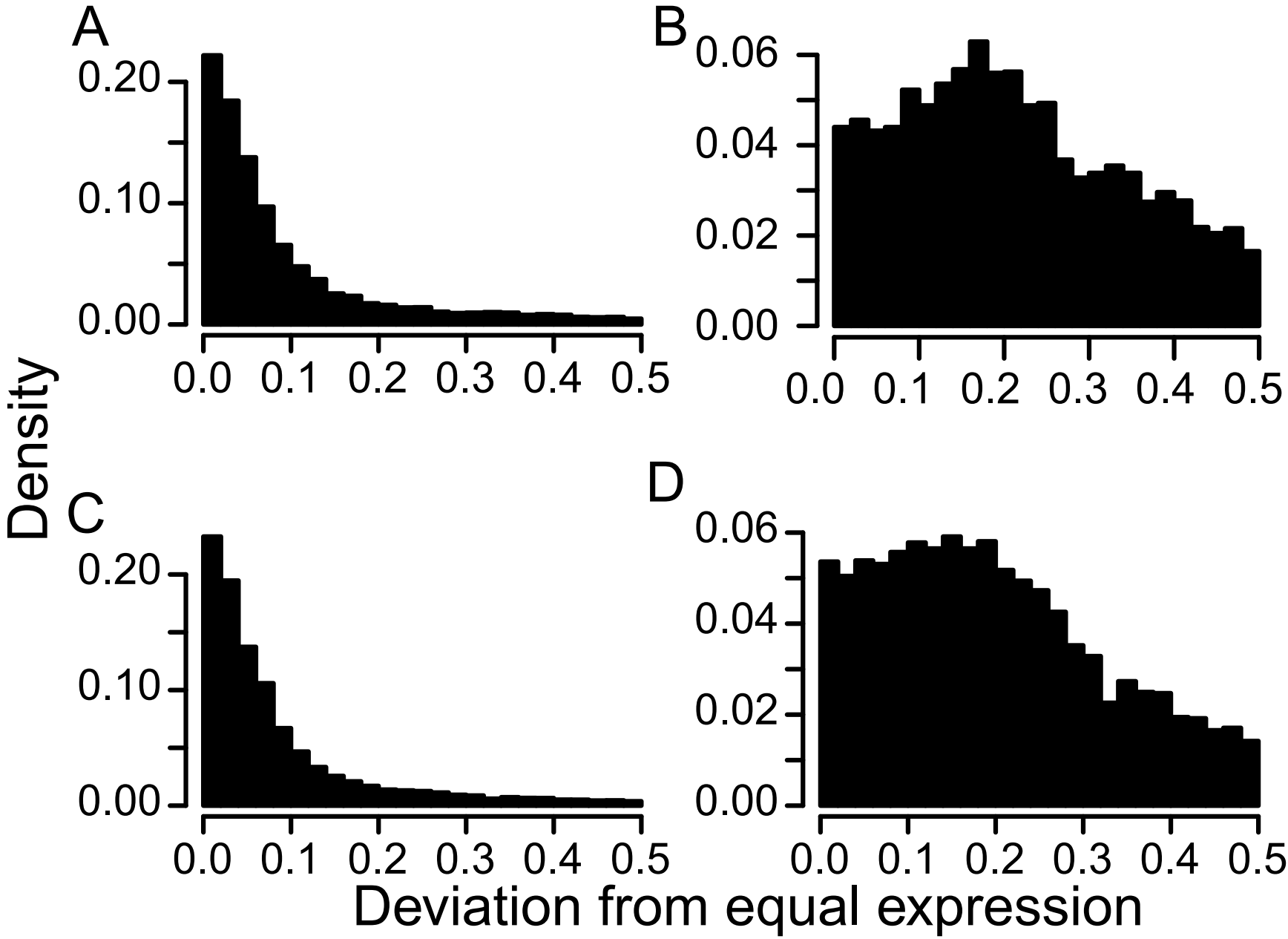
Site class[1]	Gene set	Mean π	Mean θ_W	$\pi_{siteclass}/\pi_{4\text{-fold}}$[2]	d
4-fold	ASE	0.029	0.029	NA	0.16
	control	0.023	0.024	NA	0.15
0-fold	ASE	0.009	0.011	0.32	0.04
	control	0.005	0.007	0.23	0.03
3'UTR	ASE	0.018	0.021	0.62	0.13
	control	0.014	0.018	0.62	0.12
5'UTR	ASE	0.016	0.016	0.55	0.12
	control	0.012	0.012	0.54	0.12
500 bp upstream	ASE	0.017	0.020	0.6	0.16
	control	0.015	0.019	0.68	0.15
intron	ASE	0.020	0.022	0.69	0.15
	control	0.018	0.020	0.79	0.14

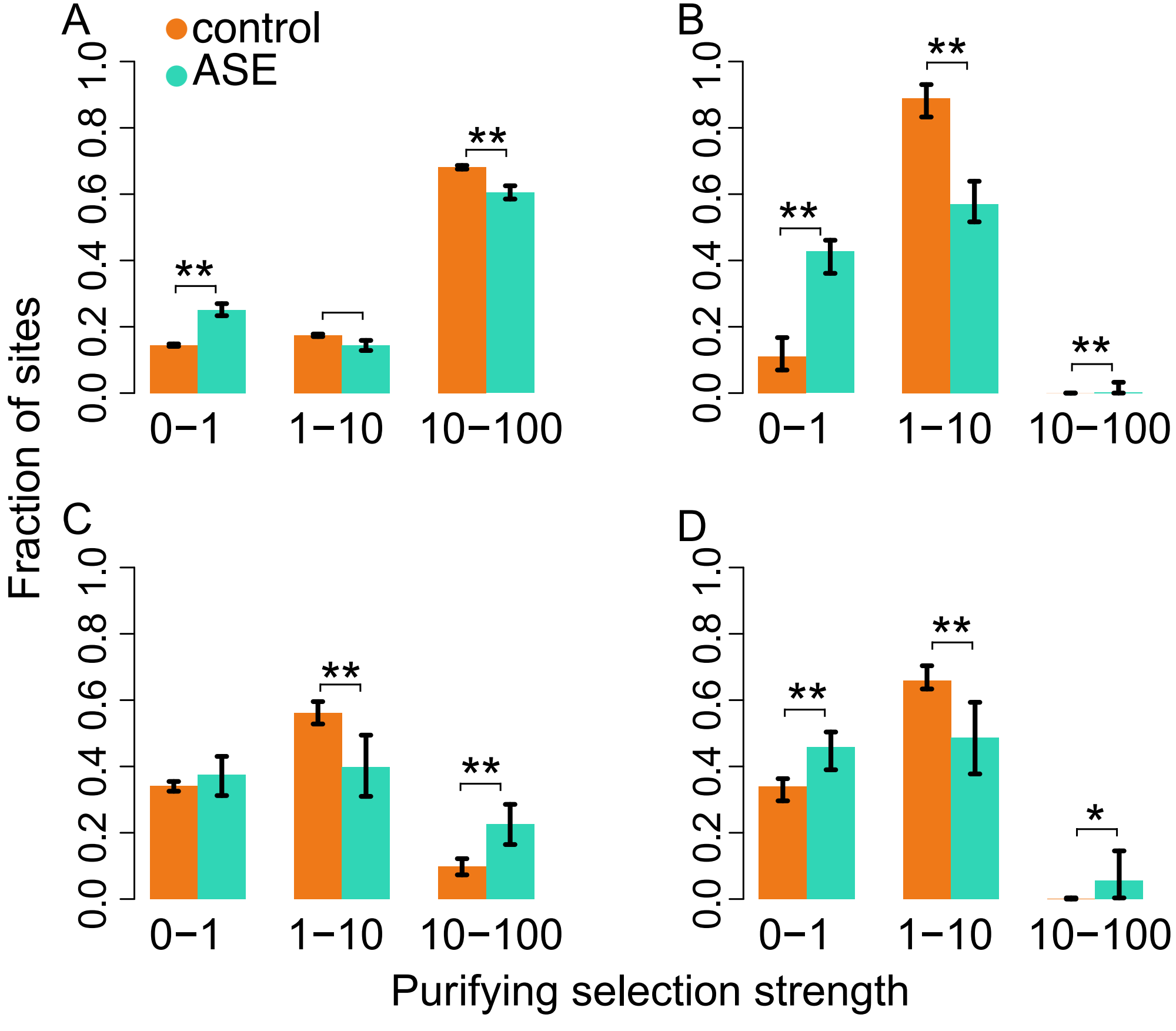
630 [1]Class of sites investigated, including 4-fold degenerate sites (4-fold), 0-fold
631 degenerate sites (0-fold), 5'UTRs, 3'UTRs, 500 bp upstream of the TSS (500 bp
632 upstream) and introns.
633 [2] Ratio of nucleotide diversity at focal site class to nucleotide diversity at 4-fold
634 synonymous sites.
635

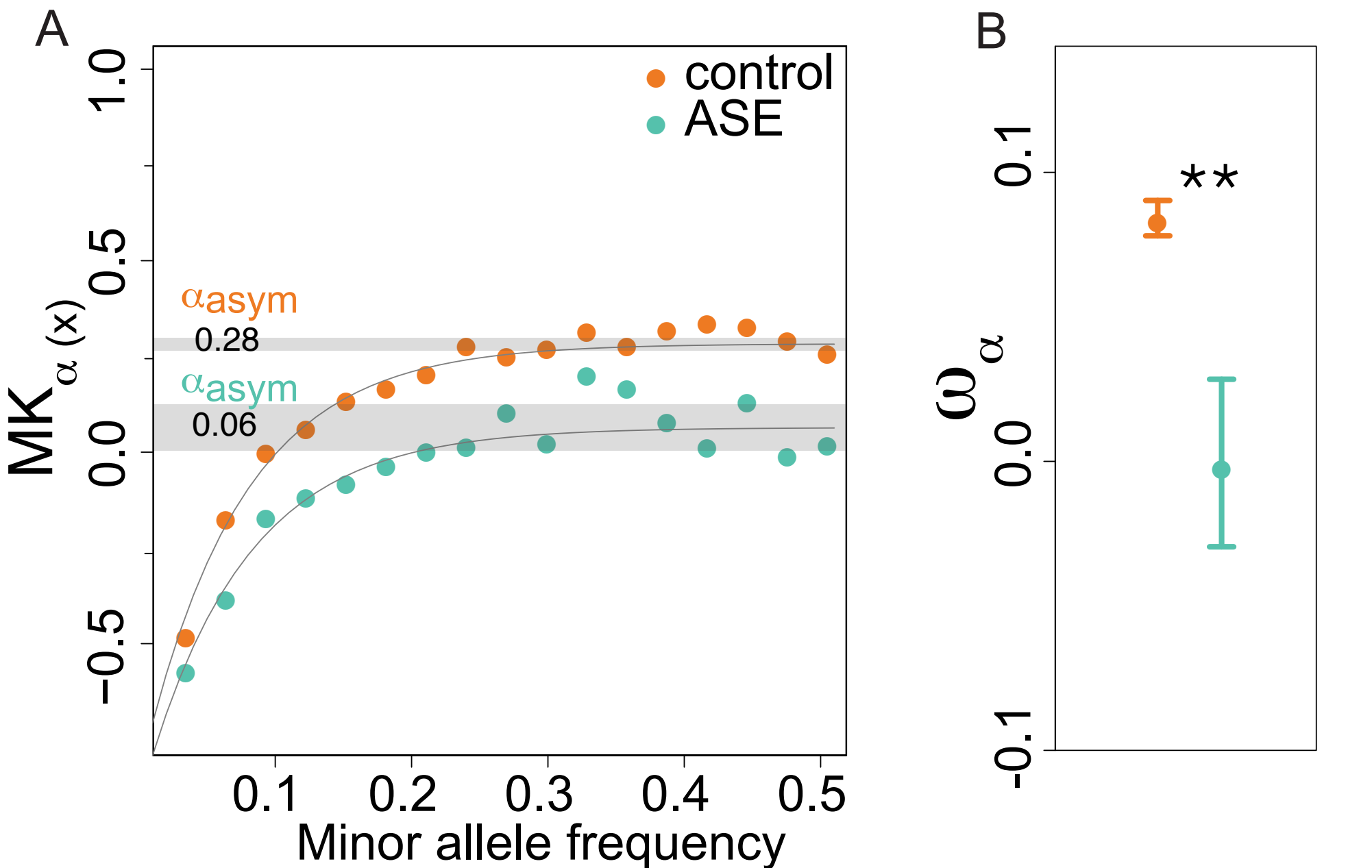
636 Table 3. The best-fit logistic regression model (AIC=3086.9) predicting ASE from
637 genomic features. Regression coefficients and their standard error, z-statistics and
638 associated P-values, and odds ratios (OR) are shown. Coexpression module size and
639 gene length were included in the best-fit model, but did not have individually
640 significant effects.

Model parameter	Coeff. (SE)	z value	P-value	OR
Gene-body methylation	-0.67 (0.20)	-3.41	$<10^{-3}$	0.51
π_N/π_S	0.08 (0.04)	2.25	0.024	1.09
Expression level	0.20 (0.06)	3.31	<0.001	1.22
Promoter polymorphism	0.21 (0.05)	4.45	$<10^{-3}$	1.23
Tissue specificity	0.30 (0.06)	5.03	$<10^{-3}$	1.35
TE within 1 kb	0.32 (0.13)	2.50	0.013	1.38
Coexpression module size	-0.08 (0.05)	-1.59	NS	0.92
Gene length	0.08 (0.06)	1.49	NS	1.09
Intercept	-2.60 (0.06)	-42.91	$<10^{-3}$	0.07

641


A
B
C
D
Density
0.20
0.10
0.00
0.06
0.04
0.02
0.00
0.0 0.1 0.2 0.3 0.4 0.5
Deviation from equal expression


A
control
ASE
B
C
D
Fraction of sites
0.0
0.2
0.4
0.6
0.8
1.0
**
*
0−1
1−10
10−100
Purifying selection strength


A
control
ASE
αasym
0.28
αasym
0.06
MKα (x)
1.0
0.5
0.0
−0.5
0.1
0.2
0.3
0.4
0.5
Minor allele frequency
B
ωα
0.1
0.0
-0.1
**