For consideration as a Research Article in the Discoveries Section of MBE.

Cis-regulatory changes associated with a recent mating system shift and floral adaptation in *Capsella*

Kim A. Steige[1], Johan Reimegård[2], Daniel Koenig[3], Douglas G. Scofield[1], Tanja Slotte[1,4,*]

[1]Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden
[2]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden
[3]Max Planck Institute for Developmental Biology, Tübingen, Germany
[4]Department of Ecology, Environment and Plant Sciences, Science for Life Laboratory, Stockholm University, Stockholm, Sweden

*Corresponding author:
Tanja Slotte
Email: tanja.slotte@su.se (TS)

Abstract

The selfing syndrome constitutes a suite of floral and reproductive trait changes that have evolved repeatedly across many evolutionary lineages in response to the shift to selfing. Convergent evolution of the selfing syndrome suggests that these changes are adaptive, yet our understanding of the detailed molecular genetic basis of the selfing syndrome remains limited. Here, we investigate the role of *cis*-regulatory changes during the recent evolution of the selfing syndrome in *Capsella rubella*, which split from the outcrosser *Capsella grandiflora* less than 200 kya. We assess allele-specific expression (ASE) in leaves and flower buds at a total of 18,452 genes in three interspecific F1 *C. grandiflora* x *C. rubella* hybrids. Using a hierarchical Bayesian approach that accounts for technical variation using genomic reads, we find evidence for extensive *cis*-regulatory changes. On average, 44% of the assayed genes show evidence of ASE, however only 6% show strong allelic expression biases. Flower buds, but not leaves, show an enrichment of *cis*-regulatory changes in genomic regions responsible for floral and reproductive trait divergence between *C. rubella* and *C. grandiflora*. We further detected an excess of heterozygous transposable element (TE) insertions near genes with ASE, and TE insertions targeted by uniquely mapping 24-nt small RNAs were associated with reduced expression of nearby genes. Our results suggest that *cis*-regulatory changes have been important during the recent adaptive floral evolution in *Capsella* and that differences in TE dynamics between selfing and outcrossing species could be important for rapid regulatory divergence in association with mating system shifts.

Introduction

The transition from outcrossing to predominant self-fertilization has occurred repeatedly in flowering plants (Stebbins 1950). In association with this shift, marked changes in floral and reproductive traits have occurred independently in many different lineages (Barrett 2002). In general, selfers tend to show reduced allocation of resources to traits involved in pollinator attraction and reward (e.g. smaller petals, less nectar per flower, less scent), exhibit changes in floral morphology that may improve the efficacy of autonomous self-pollination (e.g. reduced separation between stigma and anthers), and show reduced allocation of resources to male function (reduced ratio of pollen to ovules) (reviewed in Sicard and Lenhard 2011). Together, this combination of floral and reproductive traits is termed "the selfing syndrome" (Ornduff 1969).

Despite the striking pattern of convergent floral evolution in association with the shift to selfing, we currently have a limited understanding of the molecular genetic basis of the selfing syndrome. Quantitative trait loci (QTL) for the selfing syndrome have been identified in a handful of systems (e.g. *Capsella*; Sicard et al 2011; Slotte et al 2012; *Leptosiphon*; Goodwillie et al 2006; *Mimulus*; Fishman et al 2002; Fishman et al 2015; *Oryza*; Grillo et al 2009; *Solanum*; Bernacchi and Tanksley 1997). In domesticated tomatoes, *cis*-regulatory changes at the *Style2.1* gene have been implicated in reduced stigma exsertion (Chen et al 2007), but in most other systems, the molecular basis of the selfing syndrome is not known. A major unresolved question thus concerns the general importance of *cis*-regulatory changes vs. other types of molecular changes for the evolution of the selfing syndrome.

Cis-regulatory changes have long been hypothesized to be important for organismal adaptation (Doebley and Lukens 1998; Carroll 2000; Wray 2007; Carroll 2008; Stern and Orgogozo 2008; but see Hoekstra and Coyne 2007), due to their potentially limited negative pleiotropic effects (Wray 2007). The prospects for identifying *cis*-regulatory changes on a transcriptome-wide scale have greatly improved due to the advent of massively parallel sequencing (Fraser 2011). In particular, methods for assessing allele-specific expression (ASE) that contrast the relative levels of expression of two alleles in an individual allow for transcriptome-scale assessment of *cis*-regulatory changes. ASE studies require the presence of transcribed polymorphisms as well as rigorous bioinformatic approaches, but have benefits over mapping approaches (e.g. eQTL mapping) in terms of cost and resolution, and can identify individual genes with *cis*-regulatory changes (Pastinen 2010).

As part of our broad goal to examine molecular genetic changes associated with the selfing syndrome, we examine the influence of *cis*-regulatory changes on the evolution of the selfing syndrome in *Capsella rubella*. We further test whether silencing of TEs through the RNA-directed methylation pathway is important for global *cis*-regulatory divergence in association with the shift to selfing. The crucifer genus *Capsella* is a promising system for assessing the role of *cis*-regulatory changes in association with plant mating system shifts and adaptation, because of the availability of a sequenced genome of *C. rubella* (Slotte et al 2013) and because it is possible to generate viable

offspring from crosses between *Capsella* species that differ in their mating system (e.g. Slotte et al 2012, Rebernig et al 2015).

In *C. rubella*, the transition to selfing occurred relatively recently (<200 kya), and was associated with speciation from an outcrossing progenitor similar to present-day *C. grandiflora* (Slotte et al. 2013, Foxe et al. 2009, Guo et al. 2009, St Onge et al. 2011, Brandvain et al. 2013). Despite the recent shift to selfing, *C. rubella* already exhibits a derived reduction in petal size and a reduced pollen-ovule ratio, as well as a reduction of the degree of flower opening (Sicard et al 2011, Slotte et al. 2012). *C. rubella* therefore exhibits floral and reproductive characters typical of a selfing syndrome. The selfing syndrome of *C. rubella* is associated with improved efficacy of autonomous self-pollination (Sicard et al. 2011), and regions with quantitative trait loci (QTL) for floral divergence between *C. rubella* and *C. grandiflora* exhibit an excess of fixed differences and reduced polymorphism in *C. rubella* (Slotte et al. 2012). Together, these observations suggest that the rapid evolution of the selfing syndrome in *C. rubella* was driven by positive selection.

While the molecular genetic basis of the selfing syndrome in *C. rubella* has not been identified, it has been suggested that *cis*-regulatory changes could be involved, and a previous study found many flower and pollen development genes to be differentially expressed in flower buds of *C. grandiflora* and *C. rubella* (Slotte et al. 2013). However, these results could be confounded by differences in floral organ sizes and pollen number between *C. rubella* and *C. grandiflora,* and Slotte et al. (2013) did not directly assess *cis*-regulatory changes or investigate possible causes of *cis*-regulatory divergence. There is reason to believe that *cis*-regulatory changes could be partly caused by differences in TE abundance between selfers and outcrossers, as TE silencing can affect nearby gene expression in plants (Hollister and Gaut 2009; Hollister et al. 2011). As *C. rubella* harbors fewer TEs close to genes than *C. grandiflora* (Ågren et al. 2014), this system offers an opportunity to investigate the role of TEs for *cis*-regulatory evolution and for the evolution of floral and reproductive traits in association with the shift to selfing.

In this study we directly assessed *cis*-regulatory divergence by analyzing allele-specific expression in F1 hybrids of *C. grandiflora* and *C. rubella*, and investigated the role of *cis*-regulatory changes for the selfing syndrome in *C. rubella*. We conducted deep sequencing of transcriptomes, small RNAs, as well as genomes of *C. grandiflora* x *C. rubella* hybrids to identify genes with *cis*-regulatory divergence in flower buds and leaves, and tested whether *cis*-regulatory changes in flowers were overrepresented in genomic regions responsible for adaptive phenotypic divergence. We further identifed TEs in *C. rubella* and *C. grandiflora* and tested whether TE insertions targeted by uniquely mapping 24-nt siRNAs were associated with *cis*-regulatory divergence. Our results provide insight into the role of *cis*-regulatory changes in association with the shift to selfing in a wild plant system.

Results

112 **Many genes exhibit allele-specific expression in interspecific F1 hybrids**

113 In order to quantify ASE between *C. grandiflora* and *C. rubella*, we generated deep whole
114 transcriptome RNAseq data from flower buds and leaves of three *C. grandiflora* x *C. rubella* F1
115 hybrids (total 52.1 vs 41.8 Gbp with Q≥30 for flower buds and leaves, respectively). We included
116 three technical replicates for one F1 in order to examine the reliability of our expression data. For all
117 F1s and their *C. rubella* parents, we also generated deep (38-68x) whole genome resequencing data in
118 order to reconstruct parental haplotypes and account for read mapping biases.

119 F1 RNAseq reads were mapped with high stringency to reconstructed parental haplotypes
120 specific for each F1, i.e. reconstructed reference genomes containing whole-genome haplotypes for
121 both the *C. grandiflora* and the *C. rubella* parent of each F1 (see Material and Methods). We
122 conducted stringent filtering of genomic regions where SNPs were deemed unreliable for ASE
123 analyses due to e.g. high repeat content, copy number variation, or a high proportion of heterozygous
124 genotypes in an inbred *C. rubella* line (for details, see Material and Methods and S1 text); this mainly
125 resulted in removal of pericentromeric regions (S2 Fig - S5 Fig). After filtering, we identified ~18,200
126 genes with ~274,000 transcribed heterozygous SNPs that were amenable to ASE analysis in each F1
127 (Table 1). The mean allelic ratio of genomic read counts at these SNPs was 0.5 (S6 Fig), suggesting
128 that our bioinformatic procedures efficiently minimized read mapping biases. Furthermore, technical
129 reliability of our RNAseq data was high, as indicated by a mean Spearman's ρ between replicates of
130 0.98 (range 0.94-0.99).

131 We assessed ASE using a Bayesian statistical method with a reduced false positive rate
132 compared to the standard binomial test (Skelly et al. 2011). The method uses genomic read counts to
133 model technical variation in ASE and estimates the global proportion of genes with ASE, independent
134 of specific significance cutoffs, and also yields gene-specific estimates of the ASE ratio and the
135 posterior probability of ASE. The model also allows for and estimates the degree of variability in ASE
136 along the gene, through the inclusion of a dispersion parameter.

137 Based on this method, we estimate that on average, the proportion of assayed genes with ASE
138 is 44.6% (Table 1; S8 Table). In general, most allelic expression biases were moderate, and only 5.9%
139 of assayed genes showed ASE ratios greater than 0.8 or less than 0.2 (Figs. 1 and 2). There was little
140 variation in ASE ratios along genes, as indicated by the distribution of the dispersion parameter
141 estimates having a mode close to zero and a narrow range (Figs. 1 and 2). This suggests that unequal
142 expression of differentially spliced transcripts is not a major contributor to regulatory divergence
143 between *C. rubella* and *C. grandiflora* (Figs. 1 and 2).

144 For genes with evidence for ASE (hereafter defined as posterior probability of ASE ≥ 0.95),
145 there was a moderate shift toward higher expression of the *C. rubella* allele (mean ratio *C.*
146 *rubella*/total=0.56; Figs. 1 and 2). This shift was present for all F1s, for both leaves and flowers (Figs.
147 1 and 2). No such shift was apparent for genomic reads, and ratios of genomic read counts for SNPs in

148 genes with ASE were very close to 0.5 (mean ratio *C. rubella*/total=0.51; Figs 1 and 2). Furthermore,
149 qPCR with allele-specific probes for five genes validated our ASE results empirically (S9 Table).
150 Thus, *C. rubella* alleles appear to be on average expressed at a higher level than *C. grandiflora* alleles
151 in our F1s.
152 The mean ASE proportion, as well as the absolute number of genes with ASE was greater for
153 leaves (49%; 6010 genes) than for flower buds (40%; 5216 genes), although this difference was
154 largely driven by leaf samples from one of our F1s (Table 1). Most instances of ASE were specific to
155 either leaves or flower buds, and on average, only 15% of genes expressed in both leaves and flower
156 buds showed consistent ASE in both organs (Fig. 3). Many cases of ASE were also specific to a
157 particular F1, and across all three F1s, there were 1305 genes that showed consistent ASE in flower
158 buds, and 1663 in leaves (Fig. 3).
159
160 **Enrichment of *cis*-regulatory changes in genomic regions responsible for phenotypic divergence**
161 We used permutation tests to check for an excess of genes showing ASE within five previously-
162 identified narrow (<2 Mb) QTL regions responsible for floral and reproductive trait divergence (Slotte
163 et al. 2012). These genomic regions harbor major QTL for petal size and flowering time, but also
164 encompass part of the confidence intervals for QTL for sepal size, stamen length and ovule number, as
165 QTL for different floral and reproductive traits are highly overlapping (Slotte et al 2012). As the
166 selfing syndrome has a shared genetic basis in independent *C. rubella* accessions (Sicard et al. 2011,
167 Slotte et al. 2012), we reasoned that genes with consistent ASE across all F1s would be most likely to
168 represent candidate *cis*-regulatory changes underlying QTL. Out of the 1305 genes with ASE in flower
169 buds of all F1s, 85 were found in narrow QTL regions, and this overlap was significantly greater than
170 expected by chance (permutation test, P=0.03; Fig. 4; see Material and Methods for details). In
171 contrast, for leaves, there was no significant excess of genes showing ASE in narrow QTL
172 (permutation test, P=1; Fig. 4). Thus, the association between QTL and ASE in flower buds is unlikely
173 to be an artifact of locally elevated heterozygosity facilitating both ASE and QTL detection, which
174 should affect analyses of both leaf and flower samples.
175
176 **List enrichment analyses reveal floral candidate genes with ASE**
177 We conducted list enrichment analyses to characterize the functions of genes showing ASE relative to
178 all genes amenable to analysis of ASE (i.e. harboring heterozygous transcribed SNPs and expressed at
179 detectable levels). There was an enrichment of Gene Ontology (GO) terms involved in defense and
180 stress responses for genes with ASE in flower buds and in leaves (S10 Table). GO terms related to
181 hormonal responses, including brassinosteroid and auxin biosynthetic processes, were specifically
182 enriched among genes with ASE in flower buds (S10 Table). Genes with nearby heterozygous TE
183 insertions were also enriched for a number of GO terms related to reproduction and defense (S11-S12

184 Table), suggesting that heterozygous TE insertions could be important for patterns of GO term
185 enrichment for genes with ASE
186 We further identified nineteen genes involved in floral and reproductive development in *A.*
187 *thaliana*, which are located in QTL regions (see above), and show ASE in flower buds (Table 2).
188 These genes are of special interest as candidate genes for detailed studies of the genetic basis of the
189 selfing syndrome in *C. rubella*.
190
191 **Intergenic divergence is elevated near genes with ASE**
192 To investigate the role of polymorphisms in regulatory regions for ASE, we assessed levels of
193 heterozygosity in intergenic regions 1 kb upstream of genes, and in previously identified conserved
194 noncoding regions (Williamson et al. 2014) within 5 kb and 10 kb of genes. Genes with ASE were not
195 significantly more likely to be associated with conserved noncoding regions with heterozygous SNPs
196 than genes without ASE. However, levels of intergenic heterozygosity 1 kb upstream of genes were
197 slightly but significantly higher for genes with ASE than for those without ASE (median
198 heterozygosity of 0.016 vs. 0.014, respectively in leaves (Wilcoxon rank sum test, W = 295692325, p-
199 value = $2.26*10^{-115}$), median heterozygosity of 0.017 vs. 0.014, respectively in flowers (Wilcoxon rank
200 sum test, W = 297625040, p-value = $6.16*10^{-142}$), S13 Table), suggesting that polymorphisms in
201 regulatory regions upstream of genes might contribute to *cis*-regulatory divergence.
202
203 **Enrichment of TEs near genes with ASE**
204 To test whether differences in TE content might contribute to *cis*-regulatory divergence between *C.*
205 *rubella* and *C. grandiflora*, we examined whether heterozygous TE insertions near genes were
206 associated with ASE. We identified TE insertions specific to the *C. grandiflora* or *C. rubella* parents
207 of our F1s using genomic read data, as in Ågren et al. (2014) (Table 3; see Material and Methods).
208 Overall, we found that *C. rubella* harbored fewer TE insertions close to genes than *C. grandiflora* (on
209 average, 482 vs 1154 insertions within 1 kb of genes in *C. rubella* and *C. grandiflora*, respectively).
210 Among heterozygous TE insertions, *Gypsy* insertions were the most frequent (Table 3); they were also
211 the most frequent genome-wide (Table 3). There was a significant association between heterozygous
212 TE insertions within 1 kb of genes and ASE, for both leaves and flower buds, and the strength of the
213 association was greater for TE insertions closer to genes (Table 4; Fig. 5). This was true for individual
214 F1s, as well as for all F1s collectively (Table 4; Fig. 5; S14 Table).
215
216
217 **TEs targeted by uniquely mapping 24-nt small RNAs are associated with reduced expression of**
218 **nearby genes**
219 To test whether siRNA-based silencing of TEs might be responsible for the association between TE
220 insertions and ASE in *Capsella*, we analyzed data for flower buds from one of our F1s, for which we

221 had matching small RNA data (see Material and Methods). We selected only those 24-nt siRNA reads
222 that mapped uniquely, without mismatch, to one site within each of our F1s, because uniquely
223 mapping siRNAs have been shown to have a more marked association with gene expression in
224 *Arabidopsis* (Hollister et al. 2009). For each gene, we then assessed the ASE ratio of the allele on the
225 same chromosome as a TE insertion (i.e. ASE ratios were polarized such that relative ASE was equal
226 to the ratio of the expression of the allele with a TE insertion on the same chromosome over the total
227 expression of both alleles), and then further examined the influence of nearby siRNAs.
228 Overall, the mean relative ASE was reduced for genes with nearby TE insertions (Fig. 6) with
229 a more pronounced effect for TE insertions within 1 kb (within the gene: Wilcoxon rank sum test, W =
230 1392103, p-value = $8.76*10^{-3}$; within 200 bp: Wilcoxon rank sum test, W = 1903047, p-value =
231 $7.17*10^{-3}$; within 1 kb: Wilcoxon rank sum test, W = 3687972, p-value = $8.19*10^{-3}$). The magnitude
232 of the effect on ASE was more pronounced for genes near TE insertions targeted by uniquely mapping
233 24-nt siRNAs (Fig. 6; for genes with a TE insertion within the gene: Wilcoxon rank sum test, W =
234 423369, p-value = $1.36*10^{-4}$; within 200 bp: W = 540926, p-value = $1.82*10^{-5}$; within 1 kb: W =
235 983938, p-value = $3.13*10^{-3}$). In contrast, no significant effect on ASE was apparent for genes near TE
236 insertions that were not targeted by uniquely mapping 24-nt siRNAs (Fig. 6). Thus, uniquely mapping
237 siRNAs targeting TE insertions appear to be responsible for the association we observe between ASE
238 and TE insertions. Globally, Gypsy and hAT insertions made up a greater proportion of the TE
239 insertions that were targeted by siRNA, compared to those that were not (Chi-squared test, χ=35.9468,
240 P=$1.796*10^{-5}$, Supplementary Figure S7). However, for heterozygous TE insertions within 1 kb of
241 genes there were no significant differences in the composition of TEs that were vs. were not targeted
242 by uniquely mapping siRNAs.
243

244 Discussion

245 In this study, we have quantified allele-specific expression in order to understand the role of *cis*-
246 regulatory changes in association with a recent plant mating system shift. Our results indicate that
247 many genes, on average over 40%, harbor *cis*-regulatory differences between *C. rubella* and *C.*
248 *grandiflora*. The proportion of genes with ASE may seem high given the recent divergence (~100 kya)
249 between *C. rubella* and *C. grandiflora* (Brandvain et al. 2013, Slotte et al. 2013). However, the
250 majority of genes with ASE showed relatively mild allelic expression biases, and while our estimates
251 are higher than those in a recent microarray-based study of interspecific *Arabidopsis* hybrids (<10%)
252 (He et al. 2012a), our results are consistent with recent analyses of RNAseq data from intraspecific F1
253 hybrids of *Arabidopsis* accessions (~30%) (Cubillos et al. 2014). Somewhat higher levels of ASE
254 were found in a recent study of maize and teosinte (~70% of genes showed ASE in at least one tissue
255 and F1 individual (Lemmon et al. 2014), and using RNAseq data and the same hierarchical Bayesian
256 analysis that we employed, Skelly et al. (2011) estimated that a substantially higher proportion, >70%

257 of assayed genes, showed ASE among two strains of *Saccharomyces cerevisiae.* Thus, our estimates
258 of the proportion of genes with ASE fall within the range commonly observed for recently diverged
259 accessions or lines based on RNAseq data.

260 Two lines of evidence suggest that *cis*-regulatory changes have contributed to floral and
261 reproductive adaptation to selfing in *C. rubella*. First, we find an excess of genes with ASE in flower
262 buds within previously identified narrow QTL regions for floral and reproductive traits that harbor a
263 signature of selection (Slotte et al. 2012). This suggests either that multiple *cis*-regulatory changes
264 were involved in the evolution of the selfing syndrome in *C. rubella*, or that these regions harbor an
265 excess of *cis*-regulatory changes for other reasons, for instance due to hitchhiking of *cis*-regulatory
266 variants with causal variants for the selfing syndrome. Distinguishing between these hypotheses will
267 require identification of causal genetic changes for the selfing syndrome in *C. rubella*. In contrast, no
268 such excess is present for genes with ASE in leaves, suggesting that this observation is not simply a
269 product of higher levels of divergence among *C. rubella* and *C. grandiflora* in certain genomic regions
270 facilitating both QTL delimitation and ASE analysis. Second, we find that genes involved in hormonal
271 responses, including brassinosteroid biosynthesis, are overrepresented among genes with ASE in
272 flower buds, but not in leaves. Based on a study of differential expression and functional information
273 from *Arabidopsis thaliana*, regulatory changes in this pathway were previously suggested to be
274 important for the selfing syndrome in *C. rubella* (Slotte et al. 2013). While we do not identify ASE at
275 the same genes as in Slotte et al. 2013, our work nonetheless provides support for *cis*-regulatory
276 changes at other genes in the brassinosteroid pathway contributing to the selfing syndrome of *C.*
277 *rubella*. Future studies should conduct fine-scale mapping and functional validation to fully explore
278 this hypothesis. To facilitate this work, we have identified a set of candidate genes with ASE that are
279 located in genomic regions harboring QTL for floral and reproductive trait divergence between *C.*
280 *rubella* and *C. grandiflora*. Of particular interest in this list is the gene *JAGGED* (*JAG*), which is
281 involved in determining petal growth and shape by promoting cell proliferation in *A. thaliana* (Sauret-
282 Güeto et al. 2013, Schiessl et al. 2014). As *C. rubella* has reduced petal size due to a shortened period
283 of proliferative growth (Sicard et al. 2011), and the *C. rubella* allele is expressed at a lower level than
284 the *C. grandiflora* allele, this gene is a very promising candidate gene for the selfing syndrome.

285 Our work also provides general insights into the nature of *cis*-regulatory divergence. Indeed,
286 many instances of ASE were specific to a particular individual or tissue, an observation also supported
287 by recent studies (e.g. Lemmon et al. 2014, He et al. 2012a). This suggests that there is substantial
288 variation in ASE depending on genotype and developmental stage, consistent with the reasoning that
289 *cis*-regulatory changes can have very specific effects, but expression noise is probably also a
290 contributing factor. It is also difficult to completely rule out the possibility that some cases of subtle
291 ASE may not represent biologically meaningful *cis*-regulatory variation. However, in our analyses, we
292 took several steps to model and account for technical variation in order to reduce the incidence of false
293 positives. We also cannot fully rule out imprinting effects as potential causes of ASE, because

294 generating reciprocal F1 hybrids was not possible due to seed abortion in *C. rubella* x *C. grandiflora*
295 crosses. However, we do not expect these effects to make a major contribution to the patterns we
296 observed; in *Arabidopsis*, imprinting effects are only prevalent in endosperm tissue, and are rare in
297 more advanced stage tissues such as those analyzed here (Scott et al. 1998, Wolff et al. 2011, Cubillos
298 et al. 2014), which suggests that imprinting is not likely to be responsible for the patterns we observe.
299 One somewhat unexpected finding was the global shift in expression levels toward higher
300 relative expression of the *C. rubella* allele in the F1 hybrids. No marked bias was present for the same
301 SNPs and genes in our genomic data, suggesting that if systematic bioinformatic biases are the cause,
302 the effect is specific to transcriptomic reads. This seems unlikely to completely explain the shift in
303 expression that we observe, as we made considerable effort to avoid reference mapping bias, including
304 high stringency mapping of transcriptomic reads to reconstructed parental haplotypes specific to each
305 F1. Similar global shifts toward higher expression of the alleles from one parent have also been
306 observed in F1s of maize and teosinte (Lemmon et al. 2014) and *Drosophila* (McManus et al. 2010).
307 An even stronger bias toward higher expression of the *A. lyrata* allele was recently observed in F1s of
308 *A. thaliana* and *A. lyrata* (He et al. 2012a), and was attributed to interspecific differences in gene
309 silencing. Our results mirror those seen in some allopolyploids, where homeologs from one parental
310 species can be expressed at a markedly higher level than those from the other parental species (e.g.
311 Chang et al 2010; Flagel & Wendel 2010; Schnable et al 2011; Yoo et al. 2013).
312 To investigate potential mechanisms for *cis*-regulatory divergence, we first examined
313 heterozygosity in regulatory regions and conserved noncoding regions close to genes. While genes
314 with ASE in general showed slightly elevated levels of heterozygosity upstream of genes, there was no
315 enrichment of conserved noncoding regions with heterozygous SNPs close to genes with ASE. It thus
316 seems likely that divergence in regulatory regions in the proximity of genes, but not specifically in
317 conserved noncoding regions, has contributed to global *cis*-regulatory divergence between *C. rubella*
318 and *C. grandiflora*.
319 To examine biological explanations for the shift toward a higher relative expression of *C.*
320 *rubella* alleles, we examined the relationship between TE insertions and ASE. As *C. rubella* harbors a
321 lower number of TE insertions near genes than *C. grandiflora,* we reasoned that TE silencing might
322 contribute to the global shift in expression toward higher relative expression of the *C. rubella* allele,
323 with *C. grandiflora* alleles being preferentially silenced due to targeted methylation of nearby TEs,
324 through transcriptional gene silencing mediated by 24-nt siRNAs. Our results are consistent with this
325 hypothesis. Not only is there is an association between genes with TEs and heterozygous TE insertions
326 in our F1s, there is also reduced expression of alleles that reside on the same haplotype as a nearby TE
327 insertion, and the reduction is particularly strong for TEs that are targeted by uniquely mapping
328 siRNAs. In contrast, no effect on ASE is apparent for TEs that are not targeted by uniquely mapping
329 siRNAs. Moreover, the relatively limited spatial scale over which siRNA-targeted TE insertions are
330 associated with reduced expression of nearby genes (<1 kb) is consistent with previous results from

331 *Arabidopsis* (Hollister et al. 2009, Hollister et al. 2011, Wang et al. 2013). Our findings therefore
332 suggest that silencing of TE insertions close to genes is important for global *cis*-regulatory divergence
333 between *C. rubella* and *C. grandiflora*.
334 Why then do *C. rubella* and *C. grandiflora* differ with respect to silenced TEs near genes? In
335 *Arabidopsis*, methylated TE insertions near genes appear to be predominantly deleterious, and exhibit
336 a signature of purifying selection (Hollister et al. 2009). The reduced prevalence of TE insertions near
337 genes in *C. rubella* could be caused by rapid purging of recessive deleterious alleles due to increased
338 homozygosity as a result of self-fertilization (Arunkumar et al. 2014). However, we prefer the
339 alternative interpretation that deleterious alleles that were rare in the outcrossing ancestor were
340 preferentially lost in *C. rubella*, mainly as a consequence of the reduced effective population size
341 associated with the shift to selfing. This is in line with analyses of polymorphism and divergence at
342 nonsynonymous sites, for which *C. rubella* exhibits patterns consistent with a general relaxation of
343 purifying selection (Slotte et al. 2013).
344 If TE dynamics are generally important for *cis*-regulatory divergence in association with plant
345 mating system shifts, we might expect different effects on *cis*-regulatory divergence depending not
346 only on the genome-wide distribution of TEs, but also on the efficacy of silencing mechanisms in the
347 host (Hollister et al. 2009, Hollister et al. 2011, Ågren et al. 2015). For instance, He et al. (2012a)
348 found a shift toward higher relative expression of alleles from the outcrosser *A. lyrata,* which harbors
349 a higher TE content, a fact which they attributed to differences in silencing efficacy between *A.*
350 *thaliana* and *A. lyrata*; indeed, TEs also showed upregulation of the *A. lyrata* allele (He et al. 2012b)
351 and *A. lyrata* TEs were targeted by a lower fraction of uniquely mapping siRNAs (Hollister et al.
352 2011). In contrast, we found no evidence for a difference in silencing efficacy between *C. rubella* and
353 *C. grandiflora*, which harbor similar fractions of uniquely mapping siRNAs (12% vs 10% uniquely
354 mapping/total 24-nt RNA reads for *C. rubella* and *C. grandiflora*, respectively). Thus, in the absence
355 of strong divergence in silencing efficacy, differences in the spatial distribution of TEs, such as those
356 we observe between *C. rubella* and *C. grandiflora*, might be more important for *cis*-regulatory
357 divergence. More studies of ASE in F1s of selfers of different ages and their outcrossing relatives are
358 needed to assess the general contribution of differences in silencing efficacy versus genomic
359 distribution of TE insertions for *cis*-regulatory divergence in association with mating system shifts.
360
361 **Conclusions**
362 We have shown that many genes exhibit *cis*-regulatory changes between *C. rubella* and *C. grandiflora*
363 and that there is an enrichment of genes with floral ASE in genomic regions responsible for
364 phenotypic divergence. In combination with analyses of the function of genes with floral ASE, this
365 suggests that *cis*-regulatory changes have contributed to the evolution of the selfing syndrome in *C.*
366 *rubella*. We further observe a general shift toward higher relative expression of the *C. rubella* allele,
367 an observation that can in part be explained by elevated TE content close to genes in *C. grandiflora*

368 and reduced expression of *C. grandiflora* alleles due to silencing of nearby TEs. These results support
369 the idea that TE dynamics and silencing are of general importance for *cis*-regulatory divergence in
370 association with plant mating system shifts.
371

372 ## Material and Methods

373

374 **Plant material**

375 We generated three interspecific *C. grandiflora* x *C. rubella* F1s by crossing two accessions of the
376 selfer *C. rubella* as pollen donor with three accessions of the outcrosser *C. grandiflora* as seed parent
377 (S16 Table). No viable seeds were obtained from reciprocal crosses. Seeds from F1s and their *C.*
378 *rubella* parental lines were surface-sterilized and germinated on 0.5 x Murashige-Skoog medium. We
379 transferred one-week old seedlings to soil in pots that were placed in randomized order in a growth
380 chamber (16 h light: 8 h dark; 20° C: 14° C). After four weeks, but prior to bolting, we sampled young
381 leaves for RNA sequencing. Mixed-stage flower buds were sampled 3 weeks later, when all F1s were
382 flowering. To assess data reliability, we collected three separate samples of leaves and flower buds
383 from one F1 individual, and three biological replicates of one *C. rubella* parental line. For genomic
384 DNA extraction, we sampled leaves from all three F1 individuals as well as from their *C. rubella*
385 parents. For small RNA sequencing, we germinated six F2 offspring from one of our F1 individuals
386 and sampled flower buds as described above.
387

388 **Sample preparation and sequencing**

389 We extracted total RNA for whole transcriptome sequencing with the RNEasy Plant Mini Kit (Qiagen,
390 Hilden, Germany). For small RNA sequencing, we extracted total RNA using the mirVana kit (Life
391 Technologies). For whole genome sequencing, we used a modified CTAB DNA extraction (Doyle and
392 Doyle 1987) to obtain predominantly nuclear DNA. RNA sequencing libraries were prepared using
393 the TruSeq RNA v2 protocol (Illumina, San Diego, CA, USA). DNA sequencing libraries were
394 prepared using the TruSeq DNA v2 protocol. Small RNA libraries were prepared from 1 µg of total
395 RNA using the TruSeq SmallRNA SamplePrep fom Illumina according to the manufacturer's protocol
396 (#15004197 rev E; Illumina, San Diego, CA, USA). Sequencing was performed on an Illumina HiSeq
397 2000 instrument (Illumina, San Diego, CA, USA) to gain 100bp paired end reads, except for small
398 RNA samples for which single end 50 bp reads were obtained. Sequencing was done at the Uppsala
399 SNP & SEQ Technology Platform, Uppsala University, except for accession *C. rubella* Cr39.1 where
400 genomic DNA sequencing was done at the Max Planck Institute of Developmental Biology, Tübingen.
401 In total, we obtained 93.9 Gbp (Q≥30) of RNAseq data, with an average of 9.3 Gbp per sample. In
402 addition we obtained 45.6 Gbp (Q≥30) of DNAseq data, corresponding to a mean expected coverage
403 per individual of 52x, and 106,110,000 high-quality (Q≥30) 50 bp small RNA reads. All sequence data

404 has been submitted to the European Bioinformatics Institute (www.ebi.ac.uk), with study accession
405 number: PRJEB9020.
406
407 **Sequence quality and trimming**
408 We merged read pairs from fragment spanning less than 185 nt (this also removes potential adapter
409 sequences) in SeqPrep (https://github.com/jstjohn/SeqPrep) and trimmed reads based on sequence
410 quality (phred cutoff of 30) in CutAdapt 1.3 (Martin 2011). For DNA and RNAseq reads, we removed
411 all read pairs where either of the reads was shorter than 50 nt. We then analyzed each sample
412 individually using fastQC v. 0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to
413 identify potential errors that could have occurred in the process of amplifying DNA and RNA. We
414 assessed RNA integrity by analyzing the overall depth of coverage over annotated coding genes, using
415 geneBody_coverage.py that is part of the RSeQC package v. 2.3.3 (Wang et al. 2012). For DNA reads
416 we analyzed the genome coverage using bedtools v.2.17.0 (Quinlan and Hall 2010) and removed all
417 potential PCR duplicates using Picard v.1.92 (http://picard.sourceforge.net). Small RNA reads were
418 trimmed using custom scripts and CutAdapt 1.3 and filtered to retain only reads of 24 nt length.
419
420 **Read mapping and variant calling**
421 We mapped both genomic reads and RNAseq reads to the v1.0 reference *C. rubella* assembly (Slotte
422 et al. 2013) (http://www.phytozome.net/capsella). For RNAseq reads we used STAR v.2.3.0.1 (Dobin
423 et al. 2013) with default parameters. For genomic reads we modified the default STAR settings to
424 avoid splitting up reads, and for mapping 24-nt small RNA we used STAR with settings modified to
425 require perfect matches to the parental haplotypes of the F1s as well as to a TE library based on
426 multiple Brassicaceae species and previously used in Slotte et al. (2013).
427 Variant calling was done in GATK v. 2.5-2 (McKenna et al. 2010) according to GATK best
428 practices (DePristo et al. 2011, Van der Auwera et al. 2013). Briefly, after duplicate marking, local
429 realignment around indels was undertaken, and base quality scores were recalibrated, using a set of
430 1,538,085 SNPs identified in *C. grandiflora* (Williamson et al. 2014) as known variants. Only SNPs
431 considered high quality by GATK were kept for further analysis. Variant discovery was done jointly
432 on all samples using the UnifiedGenotyper, and for each F1, genotypes were phased by transmission,
433 by reference to the genotype of its highly inbred *C. rubella* parental accession.
434 We validated our procedure for calling variants in genomic data by comparing our calls for the
435 inbred line *C. rubella* 1GR1 at 176,670 sites sequenced in a different individual from the same line by
436 Sanger sequencing (Slotte et al. 2010). Overall, we found 29 calls that differed among the two sets,
437 resulting in an error rate of 0.00016, considerably lower than the level of divergence among *C. rubella*
438 and *C. grandiflora* (0.02; Brandvain et al. 2013).
439
440 **Reconstruction of parental haplotypes of interspecific F1s**

441 We reconstructed genome-wide parental haplotype sequences for each interspecific F1 and used these
442 as a reference sequence for mapping genomic and transcriptomic reads for ASE analyses. This was
443 done to reduce effects of read mapping biases on our analyses of ASE by increasing the number of
444 mapped reads and reducing mismapping that can result when masking heterozygous SNPs in F1s
445 (Degner et al. 2009).
446 To reconstruct parental genomes for each F1, we first conducted genomic read mapping,
447 variant calling and phasing by reference to the inbred *C. rubella* parent as described in the section
448 "Read Mapping and Variant Calling" above. The resulting phased vcf files were used in conjunction
449 with the *C. rubella* reference genome sequence to create a new reference for each F1, containing both
450 of its parental genome-wide haplotypes. Read mapping of both genomic and RNA reads from each F1
451 was then redone to its specific parental haplotype reference genome, and read counts at all reliable
452 SNPs (see section "Filtering" below) were obtained using Samtools mpileup and a custom software
453 written in javascript by Johan Reimegård. The resulting files with allele counts for genomic and
454 transcriptomic data were used in all downstream analyses of allelic expression biases (see section
455 "Analysis of Allele-Specific Expression" below).
456
457 **Filtering**
458 We used two approaches to filter the genome assembly to identify regions where we have high
459 confidence in our SNP calls. Genomic regions with evidence for large-scale copy number variation
460 were identified using Control-FREEC (Boeva et al. 2011), and repeats and selfish genetic elements
461 were identified using RepeatMasker 4.0.1 (http://www.repeatmasker.org). Additionally, we identified
462 genomic regions with unusually high proportions of heterozygous genotype calls in a lab-inbred *C.*
463 *rubella* line, which is expected to be highly homozygous. Regions with evidence for high proportions
464 of repeats, copy number variation or high proportion of heterozygous calls in the inbred line mainly
465 corresponded to centromeric and pericentromeric regions, and these were removed from consideration
466 in further analyses of allele-specific expression (S2 Fig. - S5 Fig.).
467
468 **Analysis of allele-specific expression**
469 Analyses of allele-specific expression (ASE) were done using a hierarchical Bayesian method
470 developed by Skelly et al. (2011). The method requires read counts at heterozygous coding SNPs for
471 both genomic and transcriptomic data. Genomic read counts are used to fit the parameters of a beta-
472 binomial distribution, in order to obtain an empirical estimate of the distribution of variation in allelic
473 ratios due to technical variation (as there is no true ASE for genomic data on read counts for
474 heterozygous SNPs). This distribution is then used in analyses of RNAseq data where genes are
475 assigned posterior probabilities of exhibiting ASE.
476 We conducted ASE analyses using the method of Skelly et al. (2011) for each of our three F1
477 individuals. Prior to analyses, we filtered the genomic data to only retain read counts for heterozygous

478 SNPs in coding regions that did not overlap with neighboring genes, and following Skelly et al.
479 (2011), we also removed SNPs that were the most strongly biased in the genomic data (specifically, in
480 the 1% tails of a beta-binomial distribution fit to all heterozygous SNPs in each sample), as such
481 highly biased SNPs may result in false inference of variable ASE if retained. The resulting data set
482 showed very little evidence for read mapping bias affecting allelic ratios: the mean ratio of *C. rubella*
483 alleles to total was 0.507 (S6 Fig).
484 All analyses were run in triplicate and MCMC convergence was checked by comparing
485 parameter estimates across independent runs from different starting points, and by assessing the
486 degree of mixing of chains. For all analyses of RNA counts, we used median estimates of the
487 parameters of the beta-binomial distribution from analyses of genomic data for all three F1s (S8
488 Table). Runs were completed on a high-performance computing cluster at Uppsala University
489 (UPPMAX) using the pqR implementation of R (http://www.pqr-project.org), for 200,000 generations
490 or a maximum runtime of 10 days. We discarded the first 10% of each run as burn-in prior to
491 obtaining parameter estimates.
492
493 **ASE validation by qPCR**
494 We validated ASE results by performing qPCR with TaqMan® Reverse Transcription Reagents
495 (LifeTechnologies, Carlsbad, CA, USA) using oligo$(dT)_{16}$s to convert mRNA into cDNA using the
496 manufacturers protocol and performed qPCR with the Custom TaqMan® Gene Expression Assay
497 (LifeTechnologies, Carlsbad, CA, USA) with the colors FAM and VIC using manufacturers protocol.
498 The qPCR for both alleles was multiplexed in one well to directly compare the two alleles using a Bio-
499 Rad CFX96 Touch™ Real-Time PCR Detection System (Bio-Rad, Hercules, CA, USA). To exclude
500 color bias, we used reciprocal probes with VIC and FAM colorant (S15 Table). The expression
501 difference between the *C. rubella* and *C. grandiflora* allele was quantified using the difference in
502 relative expression between the two alleles, as well as the Quantification Cycle (Cq value). A lower
503 Cq value correlates with a higher amount of starting material in the sample. If the direction of allelic
504 imbalance inferred by qPCR was the same as for ASE inferred by the method by Skelly et al. (2011),
505 we considered that the qPCR supported the ASE results. For further details see S1 Text.
506
507 **Enrichment of genes with ASE in genomic regions responsible for phenotypic divergence**
508 We tested whether there was an excess of genes with evidence for ASE (posterior probability of ASE
509 ≥ 0.95 in all three F1 hybrids) in previously identified genomic regions harboring QTL for phenotypic
510 divergence between *C. rubella* and *C. grandiflora* (Slotte et al. 2012). For this purpose, we
511 concentrated on narrow QTL regions, defined as in a previous study (Slotte et al. 2012) (i.e. QTL
512 regions with 1.5-LOD confidence intervals <2 Mb). The five QTL regions that met our criteria for
513 inclusion as narrow QTL were non-overlapping and corresponded to previously identified QTL for
514 floral and reproductive traits (on scaffolds 2 and 7 for petal width, on scaffold 7 for petal length and

515 on scaffolds 1 and 3 for flowering time). As QTL for floral and reproductive traits are generally highly
516 overlapping these genomic regions also encompass part of the confidence intervals for other QTL,
517 including a major QTL for petal length on scaffold 2, and QTL for sepal length, stamen length and
518 ovule number on scaffold 7). Significance was based on a permutation test (1000 permutations) in R
519 3.1.2.
520
521 **List enrichment tests of GO terms**
522 We tested for enrichment of GO biological process terms using Fisher exact tests in the R package
523 TopGO (Alexa et al. 2006). GO terms were downloaded from TAIR (http://www.arabidopsis.org) on
524 September 3rd, 2013, for all *A. thaliana* genes that have orthologs in the *C. rubella* v1.0 annotation,
525 and we only considered GO terms with at least two annotated members in the background set.
526 We tested for enrichment of GO biological process terms among genes with ASE in all of our
527 F1s Separate tests were conducted for leaf and flower bud samples, and background sets consisted of
528 all genes where we could assess ASE in either leaves or flower buds.
529 We used the same approach to test for enrichment of GO biological process terms among
530 genes within 1 kb and 2 kb of heterozygous TE insertions in F1 Inter4.1, for which we had matching
531 small RNA data. For this purpose, separate tests were done for all heterozygous TE insertions,
532 heterozygous TE insertions targeted by uniquely mapping siRNAs, and heterozygous TE insertions
533 not targeted by siRNAs. For these tests, the background sets consisted of all annotated *C. rubella*
534 genes.
535
536 **Intergenic heterozygosity in regulatory and conserved noncoding regions**
537 We quantified intergenic heterozygosity 1 kb upstream of genes using VCFTools (Danecek et al.
538 2011), and compared levels of polymorphism among genes with and without ASE using a Wilcoxon
539 rank sum test. We further assessed whether there was an enrichment of conserved noncoding elements
540 (identified in Williamson et al. (2014)) with heterozygous SNPs within 5 kb of genes with ASE, using
541 Fisher exact tests. Separate tests were conducted for each F1.
542
543 **Identification of TE insertions and association with ASE**
544 We used PoPoolationTE (Kofler et al. 2012) to identify transposable elements in our F1s. While
545 intended for pooled datasets, this method can also be used on genomic reads from single individuals
546 (Ågren et al. 2014). For this purpose we used a library of TE sequences based on several Brassicaceae
547 species (Slotte et al. 2013). We used the default pipeline for PoPoolationTE, modified to require a
548 minimum of 5 reads to call a TE insertion, and the procedure in Ågren et al. (2014) to determine
549 heterozygosity or homozygosity of TE insertions. Parental origins of TE insertions were inferred by
550 combining information from runs on F1s and their *C. rubella* parents. We used chi-square tests to

assess tested whether the composition of heterozygous TE insertions targeted by uniquely mapping siRNAs differed from those not targeted by siRNAs.

We tested whether heterozygous TE insertions within a range of different window sizes close to genes (200 bp, 1 kbp, 2 kbp, 5 kbp, and 10 kbp) were associated with ASE by performing Fisher exact tests. We tested whether the expression of the allele on the same chromosome as a nearby (within 1 kbp) TE insertion was reduced compared to ASE at against genes without nearby TE insertions using a Wilcoxon rank sum test. Similar tests were conducted to test for an effect on relative ASE of TE insertions with uniquely mapping siRNAs.

Acknowledgements

The authors thank Daniel Skelly, Duke University, for helpful advice on ASE analyses, Emily Josephs, University of Toronto and Adrian Platts, McGill University for information on conserved noncoding regions in *Capsella*, and Michael Nowak, Stockholm University, for valuable comments on the manuscript. 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 Project b2012122. This work was supported by grants from the Swedish Research Council, the Erik Philip-Sörensen foundation, the Nilsson-Ehle foundation, the Magnus Bergvall foundation, and the Royal Swedish Academy of Sciences to T.S. D.K. acknowledges funding from the Human Frontier Science Program (LT000783) and the German Research Foundation Priority Program 1529 – 'Adaptomics' (WE 2897).

References

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 22:1600–1607.

Arunkumar R, Ness RW, Wright SI, Barrett SCH. 2014. The Evolution of Selfing Is Accompanied by Reduced Efficacy of Selection and Purging of Deleterious Mutations. Genetics. 199(3):817-829

Barrett SC. 2002. The evolution of plant sexual diversity. Nature Reviews Genetics. 3:274-284

Bernacchi D, Tanksley SD. 1997. An interspecific backcross of *Lycopersicon esculentum* x *L. hirsutum*: linkage analysis and a QTL study of sexual compatibility factors and floral traits. Genetics. 147:861-877

Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, Barillot E. 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics. 27:268–269.

587 Brandvain Y, Slotte T, Hazzouri KM, Wright SI, Coop G. 2013. Genomic Identification of Founding
588 Haplotypes Reveals the History of the Selfing Species *Capsella rubella*. PLoS Genet.
589 9:e1003754.
590 Carroll SB. 2000. Endless forms: the evolution of gene regulation and morphological diversity. Cell.
591 101:577–580.
592 Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological
593 evolution. Cell. 134:25–36.
594 Chang PL, Dilkes BP, McMahon M, Comai L, Nuzhdin SV. 2010. Homoeolog-specific retention and
595 use in allotetraploid *Arabidopsis suecica* depends on parent of origin and network partners.
596 Genome Biology. 11:R125
597 Chen K-Y, Cong B, Wing R, Vrebalov J, Tanksley SD. 2007. Changes in regulation of a transcription
598 factor lead to autogamy in cultivated tomatoes. Science. 318:643-645
599 Cubillos FA, Stegle O, Grondin C, Canut M, Tisné S, Gy I, Loudet O. 2014. Extensive *cis*-regulatory
600 variation robust to environmental perturbation in *Arabidopsis*. Plant Cell. 26:4298–4310.
601 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth
602 GT, Sherry ST et al. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158.
603 Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-
604 mapping biases on detecting allele-specific expression from RNA-sequencing data.
605 Bioinformatics. 25:3207–3212.
606 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G,
607 Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-
608 generation DNA sequencing data. Nat Genet. 43:491–498.
609 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR.
610 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29:15–21.
611 Doebley J, Lukens L. 1998. Transcriptional regulators and the evolution of plant form. Plant Cell.
612 10:1075–1082.
613 Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue.
614 Phytochem bull. 19: 11-15.
615 Fishman L, Kelly AJ, Willis JH. 2002. Minor quantitative trait loci underlie floral traits associated with
616 mating system divergence in Mimulus. Evolution. 56:2138–2155.
617 Fishman L, Beardsley PM, Stathos A, Williams CF, Hill JP. 2015. The genetic architecture of traits
618 associated with the evolution of self-pollination in Mimulus. New Phytologist. 205:907-917
619 Flagel LE, Wendel JF 2010. Evolutionary rate variation, genomic dominance and duplicate gene
620 expression evolution during allotetraploid cotton speciation. New Phytologist. 186:184-193
621 Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. 2009. Recent speciation associated with
622 the evolution of selfing in *Capsella*. Proceedings of the National Academy of Sciences.
623 106:5241–5245.

624 Fraser HB. 2011. Genome-wide approaches to the study of adaptive gene expression evolution:
625 systematic studies of evolutionary adaptations involving gene expression will allow many
626 fundamental questions in evolutionary biology to be addressed. Bioessays. 33:469–477.
627 Goodwillie C, Ritland C, Ritland K. 2006 The genetic basis of floral traits associated with mating
628 system evolution in Leptosiphon (Polemoniaceae): an analysis of quantitative trait loci. Evolution.
629 60:491-504
630 Grillo MA, Changbao L, Fowlkes AM, Briggeman TM, Zhou A. Schemske DW, Sang T. 2009. Genetic
631 architecture for the adaptive origin of wild rice, *Oryza nivara*. Evolution. 63:870-883
632 Guo Y-L, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D Schierup MH. 2009. Recent
633 speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-
634 incompatibility and an extreme bottleneck. Proceedings of the National Academy of Sciences.
635 106:5246–5251.
636 He F, Zhang X, Hu J, Turck F, Dong X, Goebel U, Borevitz J, de Meaux J. 2012a. Genome-wide
637 analysis of *cis*-regulatory divergence between species in the *Arabidopsis* genus. Mol Biol Evol.
638 29:3385–3395.
639 He F, Zhang X, Hu J-Y, Turck F, Dong X, Goebel U, Borevitz JO, de Meaux J. 2012b. Widespread
640 interspecific divergence in cis-regulation of transposable elements in the *Arabidopsis* genus. Mol
641 Biol Evol. 29:1081–1091.
642 Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation.
643 Evolution. 61:995–1016.
644 Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced
645 transposition and deleterious effects on neighboring gene expression. Genome Res. 19:1419–
646 1428.
647 Hollister JD, Smith LM, Guo Y-L, Ott F, Weigel D, Gaut BS. 2011. Transposable elements and small
648 RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis*
649 *lyrata*. Proceedings of the National Academy of Sciences. 108:2322–2327.
650 Kofler R, Betancourt AJ, Schlötterer C. 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers
651 complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet.
652 8:e1002487.
653 Lemmon ZH, Bukowski R, Sun Q, Doebley JF. 2014. The Role of cis Regulatory Evolution in Maize
654 Domestication. PLoS Genet. 10:e1004745.
655 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.
656 EMBnet.journal. 17:10–12.
657 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D,
658 Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for
659 analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.
660 McManus CJ, Coolon JD, O'Duff M, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory

661 divergence in *Drosophila* revealed by mRNA-seq. Genome Res. 20: 816-825
662 Ornduff R. 1969. Reproductive Biology in Relation to Systematics. Taxon. 18(2):121-133
663 Pastinen T. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev
664 Genet. 11:533–538.
665 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
666 Bioinformatics. 26:841–842.
667 Rebernig CA, Lafon-Placette C, Hatorangan MR, Slotte T, Köhler C. 2015. Non-reciprocal interspecies
668 hybridization barriers in the Capsella genus are established in the endosperm. PLoS Genetics. 11:
669 e1005295
670 Sauret-Güeto S, Schiessl K, Bangham A, Sablowski R, Coen E. 2013. JAGGED controls *Arabidopsis*
671 petal growth and shape by interacting with a divergent polarity field. Plos Biol. 11:e1001550.
672 Schiessl K, Muiño JM, Sablowski R. 2014. *Arabidopsis* JAGGED links floral organ patterning to tissue
673 growth by repressing Kip-related cell cycle inhibitors. Proceedings of the National Academy of
674 Sciences. 111:2830–2835.
675 Schnable JC, Springer NM, Freeling M. 2011 Differentiation of the maize subgenomes by genome
676 dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of
677 Sciences. 108:4069-4074
678 Scott RJ, Spielman M, Bailey J, Dickinson HG. 1998. Parent-of-origin effects on seed development in
679 *Arabidopsis thaliana*. Development. 125:3329–3341.
680 Sicard A, Lenhard M. 2011. The selfing syndrome: a model for studying the genetic and evolutionary
681 basis of morphological adaptation in plants. Annals of Botany. 107(9):1433-1443
682 Sicard A, Stacey N, Hermann K, Dessoly J, Neuffer B, Bäurle I, Lenhard M. 2011. Genetics, evolution,
683 and adaptive significance of the selfing syndrome in the genus *Capsella*. Plant Cell. 23:3156–
684 3171.
685 Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. 2011. A powerful and flexible statistical
686 framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome
687 Res. 21:1728–1737.
688 Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and
689 purifying selection in *Capsella grandiflora*, a plant species with a large effective population size.
690 Mol Biol Evol. 27:1813-1821.
691 Slotte T, Hazzouri KM, Stern D, Andolfatto P, Wright SI. 2012. Genetic architecture and adaptive
692 significance of the selfing syndrome in *Capsella*. Evolution. 66:1360–1374.
693 Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo YL, Steige K, Platts AE, Escobar JS,
694 Newman LK et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid
695 mating system evolution. Nat Genet. 2013; 45(7):831-835
696 St Onge KR, Källman T, Slotte T, Lascoux M, Palmé AE. 2011. Contrasting demographic history and
697 population structure in *Capsella rubella* and *Capsella grandiflora*, two closely related species

698 with different mating systems. Mol Ecol. 20:3306–3320.
699 Stebbins GL. 1950. Variation and Evolution in Plants. Columbia Univ. Press, New York.
700 Stern DL, Orgogozo V. 2008. The loci of evolution: how predictable is genetic evolution? Evolution.
701 62(9):2155–2177.
702 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T,
703 Shakir K, Roazen D, Thibault J et al. 2013. From FastQ data to high confidence variant calls: the
704 Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 11:11.10.1–
705 11.10.33.
706 Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments. Bioinformatics.
707 28:2184–2185.
708 Wang X, Weigel D, Smith LM. 2013. Transposon variants and their effects on gene expression in
709 *Arabidopsis*. PLoS Genet. 9:e1003255.
710 Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014.
711 Evidence for widespread positive and negative selection in coding and conserved noncoding
712 regions of *Capsella grandiflora*. PLoS Genet. 10:e1004622.
713 Wolff P, Weinhofer I, Seguin J, Roszak P, Beisel C, Donoghue MT, Spillane C, Nordborg M,
714 Rehmsmeier M, Köhler C. 2011. High-resolution analysis of parent-of-origin allelic expression in
715 the *Arabidopsis* Endosperm. PLoS Genet. 7:e1002126.
716 Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. Nat Rev Genet. 8:206–216
717 Yoo MJ, Szadkowski E., Wendel JF. 2013 Homoeolog expression bias and expression level dominance
718 in allopolyploid cotton. Heredity. 110:171-180
719 Ågren JA, Wang W, Koenig D, Neuffer B, Weigel D, Wright SI. 2014. Mating system shifts and
720 transposable element evolution in the plant genus *Capsella*. BMC Genomics. 15:602.
721 Ågren JA, Wright SI. 2015. Selfish genetic elements and plant genome size evolution. Trends Plant Sci.
722 doi: 10.1016/j.tplants.2015.03.007
723

724 ## Supporting Information

725 S1 Information: S1 Text containing detailed procedures for filtering genomic regions, qPCR details,
726 Supporting Figures (S2-S7) and Tables (S8-S16).

727 **Tables**

728

729 Table 1. Genes amenable to analysis of ASE in flower bud and leaf samples from the three *C.*
730 *grandiflora* x *C. rubella* F1s, counts of genes with evidence for ASE and the estimated false discovery
731 rate (FDR) and proportion of genes with ASE.

F1 designation	Sample	Genes amenable to ASE analysis[a]	Analyzed genes[b]	Heterozygous SNPs in analyzed genes	Genes with ASE PP ≥ 0.95[c]	FDR	ASE proportion[d]
Inter3.1	Flower buds	18299	16857	262120	4728	0.0013	0.38
Inter4.1		18270	17837	272126	5744	0.0022	0.42
Inter5.1		18144	17448	262696	5176	0.0020	0.40
Inter3.1	Leaves	18299	14877	238786	5105	0.0012	0.44
Inter4.1		18270	15784	249181	8129	0.0024	0.62
Inter5.1		18144	15478	240653	4795	0.0018	0.41

732 [a]Total number of genes with heterozygous SNPs in coding regions remaining after filtering.
733 [b]Number of genes amenable to ASE analyses with expression data in at least one of the replicates of
734 the sample.
735 [c]Number of genes with evidence for ASE (posterior probability ≥ 0.95).
736 [d]Direct estimate of the ASE proportion independent of significance cutoffs.

737 Table 2. Selfing syndrome candidate genes identified based on ASE, QTL information, and Arabidopsis annotation.

C. rubella ortholog	Arabidopsis ortholog	Arabidopsis annotation	GO biological process terms related to floral and reproductive development
Carubv10012851m[a,b]	AT3G24340	CHR40	regulation of flower development
Carubv10016094m[a,b]	AT3G24650	ATABI3, ABI3, SIS10	embryo development, cotyledon development
Carubv10007602m[a,b]	AT4G21600	ENDO5	brassinosteroid biosynthetic process
Carubv10000655m[b,d]	AT5G08130	BIM1	brassinosteroid mediated signaling pathway, primary shoot apical meristem specification
Carubv10006681m[b,d]	AT4G28720	YUC8	brassinosteroid mediated signaling pathway
Carubv10021883m[a,c]	AT1G68480	JAG	sepal formation, flower development, abaxial cell fate specification, anther development, carpel development, stamen development, petal formation, specification of floral organ identity
Carubv10021345m[a,c]	AT1G68640	PAN, TGA8	petal formation, sepal formation, regulation of flower development
Carubv10013321m[a,c]	AT3G22420	ATWNK2, WNK2, ZIK3	photoperiodism, flowering
Carubv10016406m[a,c]	AT3G23270	-	pollen tube growth
Carubv10014951m[a,c]	AT3G23440	EDA6, MEE37	megagametogenesis
Carubv10014152m[a,c]	AT3G23630	ATIPT7, IPT7	pollen tube growth, reciprocal meiotic recombination
Carubv10010238m[a,c]	AT3G62210	EDA32	polar nucleus fusion
Carubv10004312m[a,c]	AT4G16760	ATACX1, ACX1	pollen development
Carubv10005585m[a,c]	AT4G17030	AT-EXPR, EXPR,	sexual reproduction

		ATEXLB1, ATEXPR1, EXLB1	
Carubv10007441m[a,c]	AT4G20370	TSF	regulation of flower development, photoperiodism, flowering, positive regulation of flower development
Carubv10004229m[a,c]	AT4G20910	CRM2, HEN1	specification of floral organ identity, floral organ formation, petal formation, regulation of flower development, sepal formation, meristem initiation, meristem development, ovule development
Carubv10015623m[a,c]	AT4G21380	ARK3, RK3	recognition of pollen
Carubv10007227m[a,c]	AT4G21530	APC4	ovule development
Carubv10007633m[a,c]	AT4G21590	ENDO3	petal development, stamen development, pollen tube growth, ovule development

738 [a]located within narrow QTL regions

739 [b]ASE in all three F1s

740 [c]ASE in the F1 with data for three replicates, but not in all three F1s

741 [d]located within QTL regions, but not narrow QTL regions

742 Table 3. Mean number of TE insertions in three interspecific F1s. The table shows the overall number,
743 as well as heterozygous insertions with parent of origin information.

TE family	Mean copy number	Heterozygous insertions	Insertions specific to the *C. rubella* parental genome	Insertions specific to the *C. grandiflora* parental genome
CACTA	84	40	10	30
Copia	710	483	144	339
Gypsy	1124	602	153	449
Harbinger	176	109	26	83
hAT	83	55	16	40
Helitron	236	127	30	97
LINE	229	165	38	128
MuDR	203	109	28	81
SINE	113	92	9	83
Total	2958	1782	454	1330

744

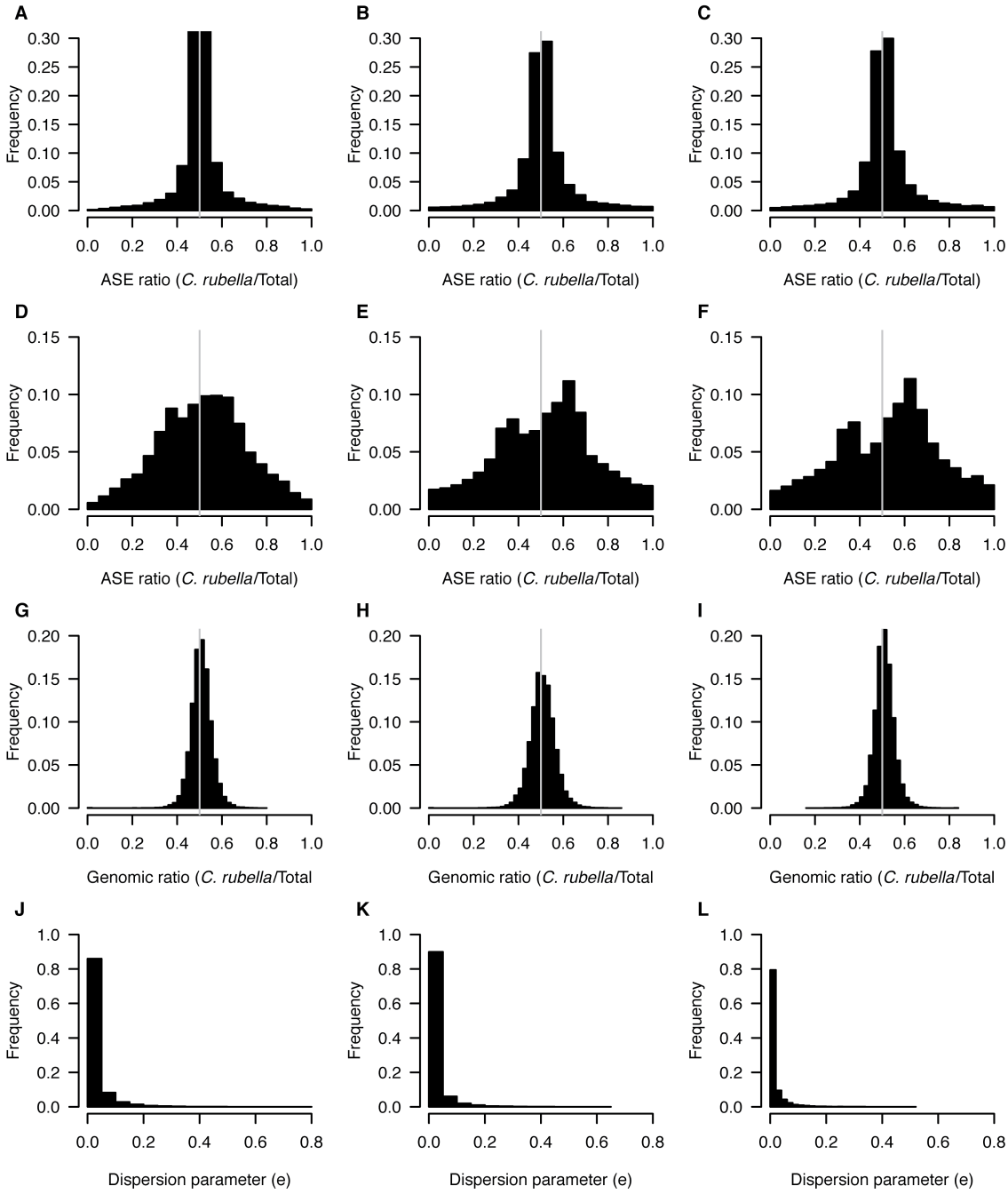
745 Table 4. Enrichment of heterozygous TEs near genes with ASE. The table shows mean counts over all
746 three F1s, and Fisher exact test *P*-values. The four categories of counts correspond to numbers of
747 genes with ASE (posterior probability of ASE ≥ 0.95) and TE insertions within a specific window size
748 near the gene (+ASE,+TE), with ASE but without TEs (+ASE,-TE), without ASE but with TE
749 insertions (-ASE,+TE), and with neither ASE nor TEs (-ASE,-TE). NS indicates not significant.

Sample	Window size (bp)	+ASE,+TE	+ASE-TE	-ASE, +TE	-ASE,-TE	*P*
Flower buds	200	113	5103	136	12029	$4.32*10^{-19}$
	1000	218	4998	339	11826	$5.07*10^{-16}$
	2000	307	4909	540	11624	$6.53*10^{-12}$
	5000	566	4650	1108	11057	$8.22*10^{-10}$
	10000	958	4258	2006	10159	$2.32*10^{-7}$
Leaves	200	108	5902	115	9255	$8.52*10^{-7}$
	1000	216	5793	277	9093	$1.49*10^{-4}$
	2000	317	5693	435	8935	$2.25*10^{-3}$
	5000	595	5415	877	8493	NS
	10000	1027	4983	1576	7795	NS

750

751 **Figures**

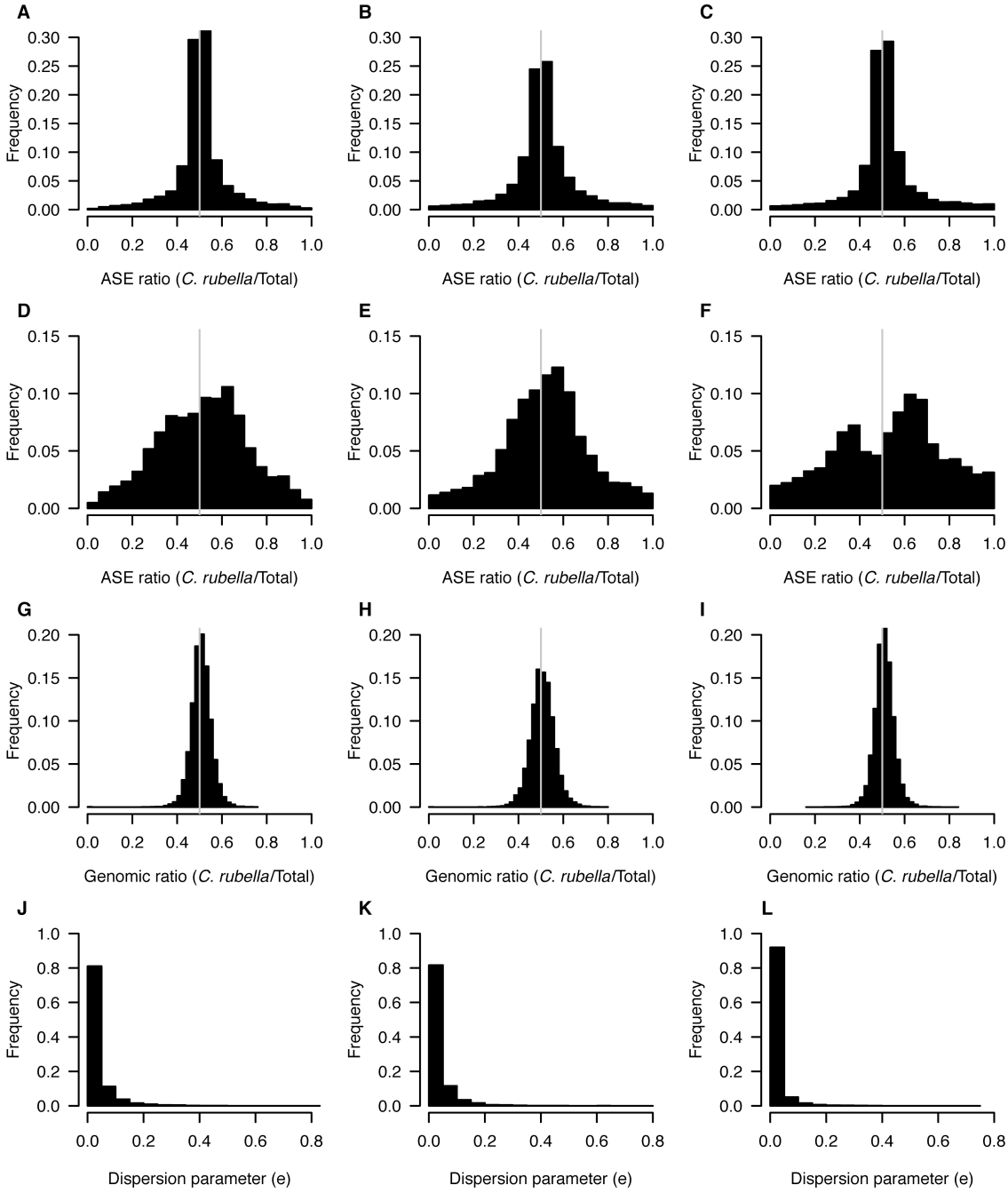
752 Figure 1



753

754 Fig. 1. ASE in flower buds. Distributions of ASE ratios (*C. rubella*/Total) for all assayed genes (A, B,
755 C), and for genes with at least 0.95 posterior probability of ASE (D, E, F). Ratio of *C. rubella* to total
756 for genomic reads, for genes with significant ASE (G, H, I), and the distribution of the dispersion
757 parameter that quantifies variability in ASE across genes (J, K, L). All distributions are shown for
758 each of the three interspecific F1s inter 3.1 (left), inter4.1 (middle) and inter5.1 (right).

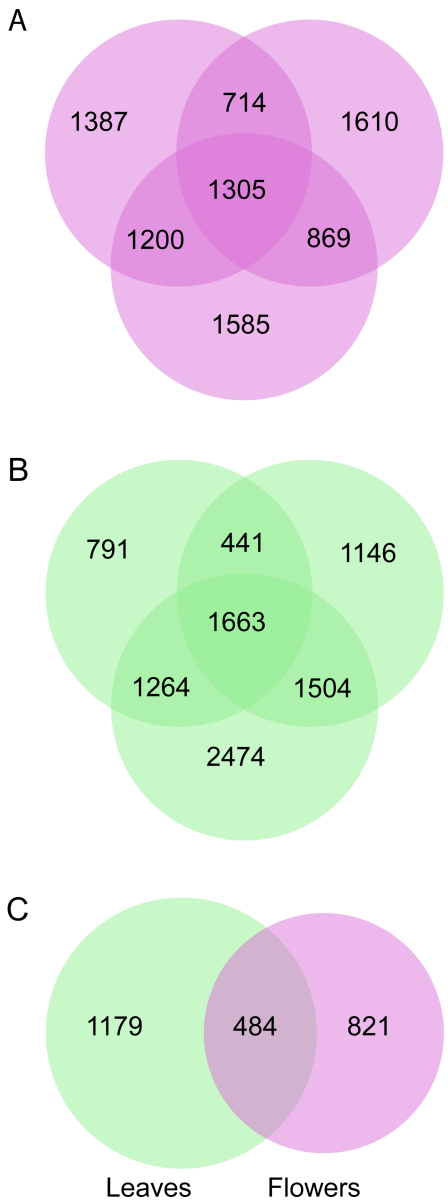
759 Figure 2



760

761 Fig. 2. ASE in leaves. Distributions of ASE ratios (*C. rubella*/Total) for all assayed genes (A, B, C),
762 and for genes with at least 0.95 posterior probability of ASE (D, E, F). Ratio of *C. rubella* to total for
763 genomic reads, for genes with significant ASE (G, H, I), and the distribution of the dispersion
764 parameter that quantifies variability in ASE across genes (J, K, L). All distributions are shown for
765 each of the three interspecific F1s inter 3.1 (left), inter4.1 (middle) and inter5.1 (right).

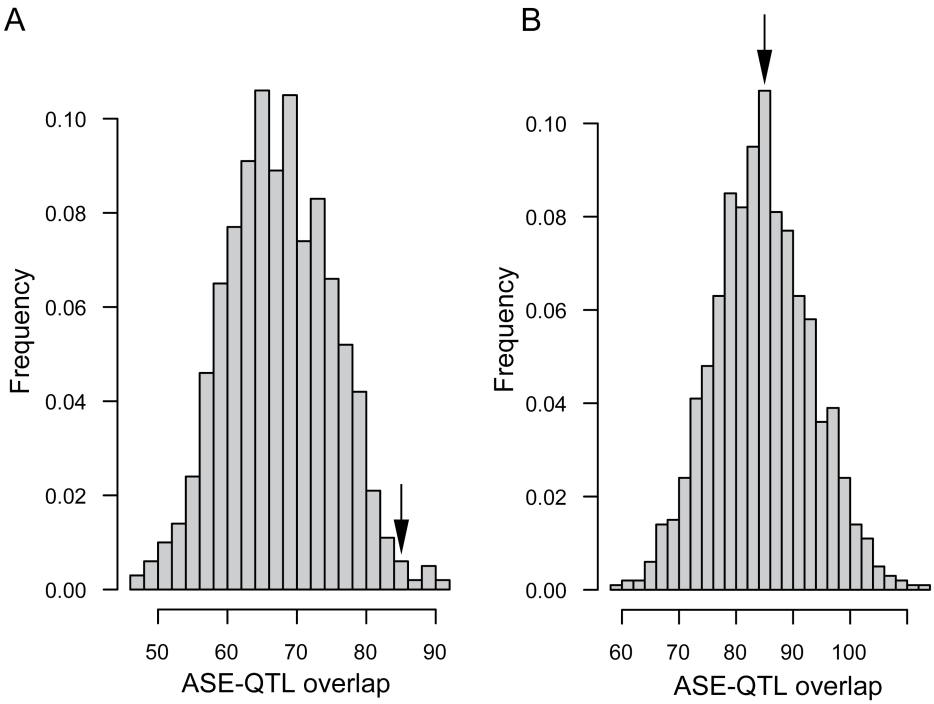
766 Figure 3



767

768 Fig. 3. Many cases of ASE are specific to individuals or samples. Venn diagrams showing
769 intersections of genes with ASE in flower buds (A) and leaves (B) of the three F1 individuals, and (C)
770 in all leaf and flower samples, for the set of genes assayed in all F1s.

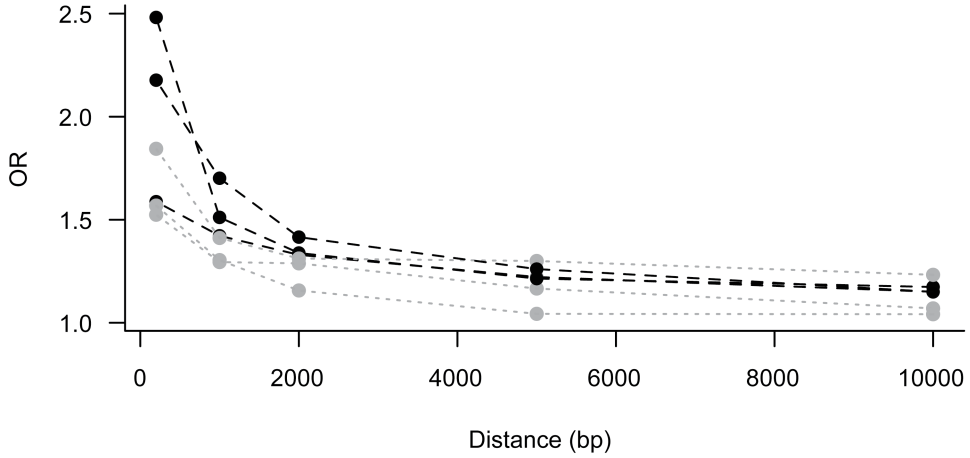
771 Figure 4



772

773 Fig. 4. Enrichment of genes with ASE in narrow QTL regions. There is an excess of genes with ASE
774 in narrow QTL regions for flower buds (A) but not for leaves (B). Histograms show the distribution of
775 numbers of genes with ASE that fall within narrow QTL regions, based on 1000 random permutations
776 of the observed number of genes with ASE among all genes where we could assess ASE. Arrows
777 indicate the observed number of genes with ASE that are located in narrow QTL regions.

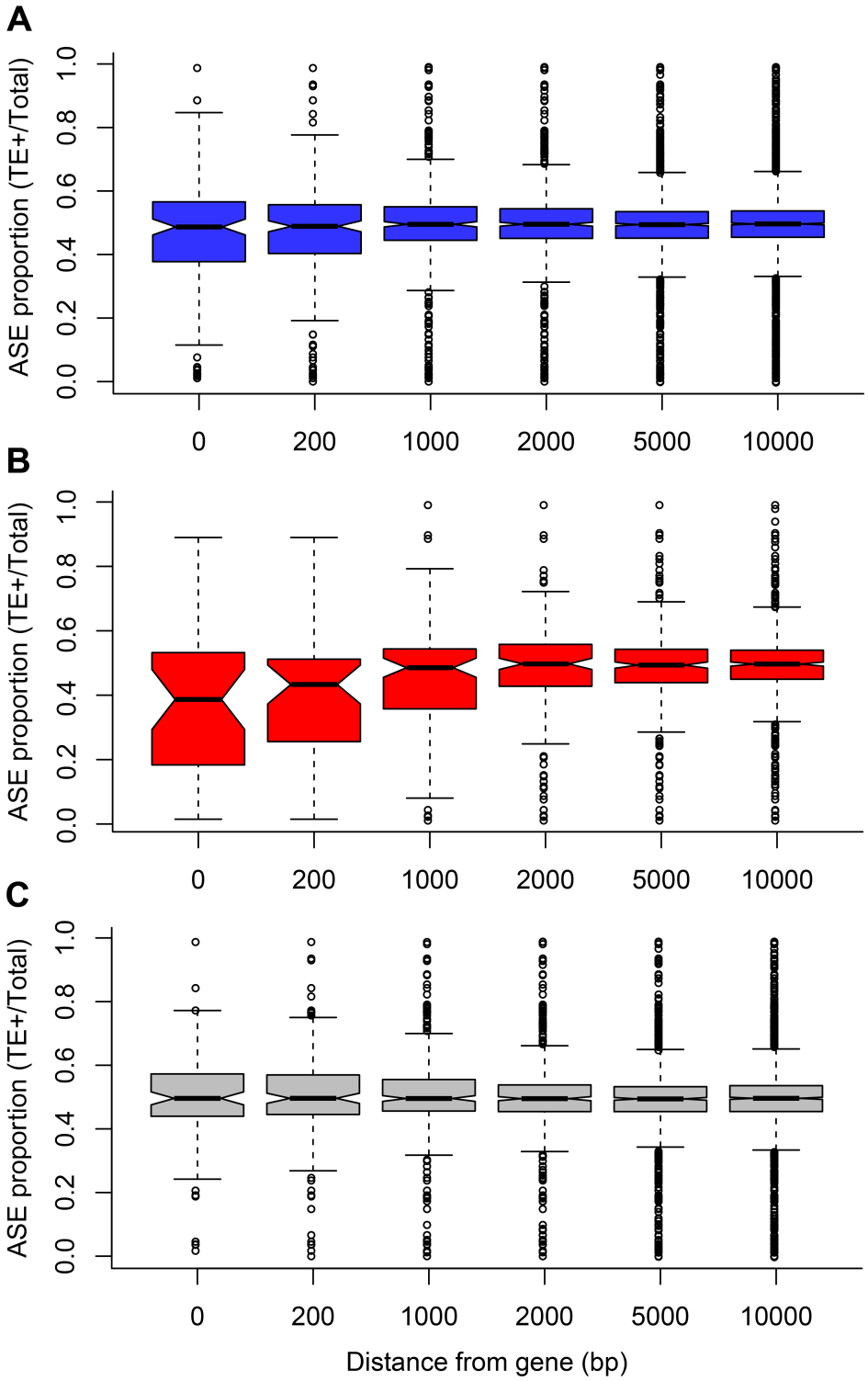
778 Figure 5



779

780 Fig. 5. Enrichment of TEs near genes with ASE. The Figure shows odds ratios (OR) of the association
781 between genes with ASE and TEs, with TE insertions scored in four different window sizes (within a
782 distance of 0 bp, 1 kbp, 2 kbp, 5 kbp, and 10 kbp of each gene) . Odds ratios for flower buds are
783 shown for all three F1s studies, with values for flower buds in black and leaves in grey.

784 Figure 6



785

786 Fig. 6. The effect of TE insertions on relative allelic expression. Boxplots show the relative allelic
787 expression (expression of the allele on same haplotype as TE insertion relative to expression of both
788 alleles) for genes near heterozygous TE insertions, scored in a range of window sizes ranging from 0
789 bp (within the gene) to 10 kbp from the gene. A. The relative allelic expression is reduced for genes
790 with nearby TE insertions. B. The degree of reduction of relative allelic expression is stronger for

791 genes near TE insertions targeted by uniquely mapping siRNA. C. There is no reduction of relative
792 allelic expression for genes near TE insertions that are not targeted by uniquely mapping siRNA.