Pollen-Specific, but Not Sperm-Specific, Genes Show Stronger Purifying Selection and Higher Rates of Positive Selection Than Sporophytic Genes in Capsella grandiflora

Ramesh Arunkumar,*1 Emily B. Josephs,1 Robert J. Williamson,1 and Stephen I. Wright1
1Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada
*Corresponding author: E-mail: ram.arunkumar@utoronto.ca.
Associate editor: John Parsch

Abstract

Selection on the gametophyte can be a major force shaping plant genomes as 7–11% of genes are expressed only in that phase and 60% of genes are expressed in both the gametophytic and sporophytic phases. The efficacy of selection on gametophytic tissues is likely to be influenced by sexual selection acting on male and female functions of hermaphroditic plants. Moreover, the haploid nature of the gametophytic phase allows selection to be efficient in removing recessive deleterious mutations and fixing recessive beneficial mutations. To assess the importance of gametophytic selection, we compared the strength of purifying selection and extent of positive selection on gametophyte- and sporophyte-specific genes in the highly outcrossing plant Capsella grandiflora. We found that pollen-exclusive genes had a larger fraction of sites under strong purifying selection, a greater proportion of adaptive substitutions, and faster protein evolution compared with seedling-exclusive genes. In contrast, sperm cell-exclusive genes had a smaller fraction of sites under strong purifying selection, a lower proportion of adaptive substitutions, and slower protein evolution compared with seedling-exclusive genes. Observations of strong selection acting on pollen-expressed genes are likely explained by sexual selection resulting from pollen competition aided by the haploid nature of that tissue. The relaxation of selection in sperm might be due to the reduced influence of intrasexual competition, but reduced gene expression may also be playing an important role.

Key words: adaptive evolution, gametophytic selection, distribution of fitness effects, haploid, pollen competition, Capsella.

Introduction

Sex-biased gene expression is widespread across animals and often such genes evolve rapidly due to various forces including those involved in sexual selection (Swanson and Vacquier 2002; Ellegren and Parsch 2007; Parsch and Ellegren 2013). Plants serve as an effective comparison to the animal models as although sexual selection also occurs in plants, the targets are often different (Willson 1979; Moore and Pannell 2011). The angiosperm life cycle is characterized by the alternation of generations with gametophytic and sporophytic phases (Bower 1908). Fertilization is initiated as the haploid pollen from the male gametophyte lands on a receptive stigma and germinates (Heslop-Harrison 1987). The pollen tubes compete with each other to extend down the style toward the ovary (Delph and Havens 1998). This pollen competition effect can be a major selective force. Furthermore, selection is likely to be more efficient in genes expressed in haploid phase, as recessive and partially recessive mutations are not masked (Orr and Otto 1994; Gerstein and Otto 2009). The exposure of recessive mutations in the haploid phase could intensify the selective pressures on the male gametophyte, increasing the fixation probability of beneficial mutations and the purging of deleterious mutations. As around 7–11% of genes are expressed only in the gametophyte and 60% of genes are expressed in both the gametophytic and sporophytic phases (Honys and Twell 2004; Pina et al. 2005; Borges et al. 2008; Borg et al. 2009), selection on gametophytic genes can be a considerable force that affects the overall fitness of the plant. Thus, the potential for intense pollen competition and the expression of mutations in the haploid phase imply that both positive and negative selection may be more efficient in pollen-expressed genes.

Currently, empirical support for increased selective pressures on gametophytic genes in plants is limited, with multiple alternate explanations that could account for the observed patterns. Previous studies have found evidence for rapid evolution in pollen-specific oleosin-like proteins (Schein et al. 2004) and glycine-rich pollen surface proteins (Fiebig et al. 2004). Although these results are encouraging, the entire collection of proteins expressed in a given tissue needs to be analyzed if one is interested in detecting tissue-wide patterns of selection. Seoighe et al. (2005) found that genes expressed exclusively in pollen have shorter introns compared with genes expressed in the sporophytic tissues of Arabidopsis thaliana. The authors argue that shorter introns might be selectively favoured to reduce transcriptional costs. Longer introns might be involved in complex regulatory processes.
(Louro et al. 2009); therefore, the pattern seen in pollen-expressed genes might be due to less complex regulation occurring in such tissues. Whittle et al. (2007) and Camiolo et al. (2012) found that male tissues experience greater usage of preferential codons compared with sporophytic ones in *Brassica napus* and *A. thaliana*, respectively. Camiolo et al. (2012) found similar patterns even when the GC content of intergenic regions and expression levels was included as covariates. Hence, they argue that factors other than mutational biases and levels of gene expression were responsible for the observed patterns. As indicated in these studies, one consideration is that the patterns of codon usage biases might have resulted either from stronger selection on gametophytic genes or are due to differential tRNA composition in various tissues. Finally, Szővényi et al. (2013) found faster evolution of haploid-specific compared with diploid-specific genes in *A. thaliana* and the moss Funaria hygrometrica by comparing patterns of polymorphism and divergence. Their observations were not altered even after they accounted for differences in levels of gene expression, gene length, GC content, average intron length, and distribution of molecular functions. They postulate that the observed differences are due to relaxed selection acting on haploid-specific genes, because such genes are expressed across fewer tissues and at reduced consistency due to greater expression noise in the haploid phase. However, it remains unclear from this study whether genes with similar expression breadth (i.e., tissue-specific genes) would show evidence for more efficient selection if expressed in the gametophytic phase. Moreover, explicit tests for differences in the extent of positive selection were not conducted.

Another potential limitation of previous studies on gametophytic selection is that many have focused on the selfing species *A. thaliana*. Because selfers have highly homozygous genomes (Nordborg 2000; Charlesworth and Wright 2001; Glémin et al. 2006; Wright et al. 2008), there is little difference in the ability to mask mutations between the gametophytic and sporophytic phases, thereby making haploid-level selection a negligible force. Moreover, the effects of pollen competition will likely be minimal in such species, as recurrent selfing would have reduced the number of distinct pollen genotypes competing on a stigma (Mazer et al. 2010). Thus, obligate outcrossing species might represent better candidates for studying the efficacy of selection on gametophytic genes.

Here, we examined selection on pollen-specific and sporophyte-specific genes in *Capsella grandiflora*, a self-incompatible obligate outcrossing close relative of *A. thaliana*. Previous work has suggested that *C. grandiflora* is characterized by a large effective population size, low population structure, and efficient positive and purifying selection (Slotte et al. 2010), making it a good model system for testing hypotheses about the extent of positive and negative selection. We analyzed whole genome resequencing data from 13 outbred *C. grandiflora* individuals (Williamson RJ, Josephs EB, Wright SI, unpublished data) and identified male gametophytic (pollen and sperm) and sporophytic (seedling, leaf, silique, and sporophytic nucellus) genes by comparing the *Capsella* genome to the tissue-specific *A. thaliana* transcriptomes characterized by Borges et al. (2008), Pina et al. (2005), and Schmidt et al. (2011). For all genes exclusively expressed in each tissue, we estimated levels of purifying selection and extent of positive selection by using the method of Eyre-Walker and Keightley (2009), which jointly infers selective and demographic parameters using polymorphism and divergence data. We find evidence suggesting that pollen-specific genes indeed show stronger purifying and positive selection than sporophyte-specific genes, but this was not true for sperm cell-specific genes. These results suggest the combined importance of haploid selection and pollen competition in determining the efficacy of positive and negative selection.

**Results**

**Tissue-Specific Estimates of Selection**

Using the expression profile of the *A. thaliana* transcriptome in Borges et al. (2008), we identified the *C. grandiflora* orthologs for the genes expressed in the pollen (vegetative cell), sperm, and seedling tissues (fig. 1). To test for differences in selection across phases, we estimated selection on genes expressed exclusively in each tissue and genes expressed in all tissues. We first used divergence-based tests to estimate the rates of evolution for the *C. grandiflora* orthologs in each tissue. Under a model where a single rate of evolution was calculated across all sites in a given gene, both pollen-exclusive genes and sperm-exclusive genes had a larger ratio of the number of nonsynonymous substitutions per nonsynonymous site (d_{ns}) to the number of synonymous substitutions per synonymous site (d_{s}) compared with seedling-exclusive genes (fig. 2 and table 1). As higher d_{ns}/d_{s} values could be due to higher rates of adaptive evolution or a relaxation of selective constraints, we performed further tests that would quantify the contribution of the two forces.

We estimated parameters to infer both the efficacy of purifying selection and extent of positive selection using both polymorphism and divergence. For the *C. grandiflora* orthologs, the estimated distribution of fitness effects of new nonsynonymous mutations (DFE) (fig. 3a) indicated that genes expressed in the different tissues had 70–85% of mutations falling in the N_{5} > 10 category. Close to 82% of mutations that occurred in genes that were expressed in all

![Figure 1](https://example.com/fig1.png)

**Fig. 1.** Venn diagram depicting numbers of *Capsella grandiflora* orthologs of *Arabidopsis thaliana* genes expressed in the pollen, sperm, and seedling tissues as profiled in Borges et al. (2008).
three tissues fell in the $N_{s} > 10$ category, suggesting widespread purifying selection. Consistent with our predictions, pollen-exclusive genes had approximately 1–2% fewer mutations in the $N_{s} < 1$ and $N_{s} = 1–10$ categories and ~6% more mutations in the $N_{s} > 10$ category compared with seedling-exclusive genes (fig. 3a). In contrast, sperm-exclusive genes had ~7% more mutations in the $N_{s} < 1$ and ~8% fewer mutations in the $N_{s} > 10$ category compared with seedling-exclusive genes. All these differences were assessed to be statistically significant (table 1) although note that the same tissue-specific comparisons made using each of the three $N_{s}$ categories will not be independent, as a shift in the DFE will change all or multiple categories within it. To validate our estimates of purifying selection, we compared the nonsynonymous and synonymous derived allele frequency spectra for the previous tissue comparisons. For nonsynonymous sites, we found that pollen-exclusive genes

![Fig. 2](image-url) The ratio of the number of nonsynonymous substitutions per nonsynonymous site ($d_{n}$) to the number of synonymous substitutions per synonymous site ($d_{s}$) for genes expressed in pollen, sperm, and seedling tissues for genes identified using the Borges et al. (2008) study. Error bars are 95% confidence intervals from 200 bootstrap replicates generated by resampling over genes.

![Fig. 3](image-url) (a) The distribution of fitness effects of new nonsynonymous mutations (DFE), (b) the proportion of adaptive substitutions ($\alpha$), and (c) the rate of adaptive nonsynonymous to synonymous substitutions ($\omega_{a}$) for genes expressed in pollen, sperm, and seedling tissues for genes identified using the Borges et al. (2008) study. Error bars are 95% confidence intervals from 200 bootstrap replicates generated by resampling over genes.
had a larger proportion of singletons relative to all other frequency classes compared with seedling-exclusive genes (supplementary fig. S1a and table S1, Supplementary Material online) as expected if purifying selection was stronger in pollen. In contrast, sperm-exclusive genes had a smaller proportion of singletons relative to all other frequency classes compared with seedling-exclusive genes as expected if purifying selection was relaxed in sperm cells. No such patterns were observed for the synonymous sites for the pollen versus seedling and the sperm versus seedling comparisons (supplementary fig. S1b and table S1, Supplementary Material online), indicating that there were no discernible differences in patterns of neutral processes between the tissues.

We also used the method of Eyre-Walker and Keightley (2009) to estimate the proportion of adaptive substitutions (\(\alpha\)), as well as the rate of adaptive substitutions relative to the rate of synonymous substitution (\(\omega_a\) Gossmann et al. 2010). Because differences across genes in synonymous substitution rate and extent of purifying selection can cause heterogeneity in \(\alpha\), \(\omega_a\) comparisons provide the most explicit test of differences in the rate of positive selection (Gossmann et al. 2010). Moreover, the \(\omega_a\) parameter differs from the \(d_{ns}/d_s\) estimates that we derived previously as it quantifies the rate and extent of purifying selection relative to adaptive substitutions. We used alignments of our C. grandiflora genes to two outgroups, Neslia paniculata and A. thaliana, in order to estimate lineage-specific rates of substitution for C. grandiflora. For the C. grandiflora orthologs identified using Borges et al. (2008), genes expressed in all tissues had \(\alpha\) and \(\omega_a\) values of 0.58 and 0.11, respectively (fig. 3b and c). \(\alpha\) and \(\omega_a\) were significantly higher for pollen-exclusive genes when compared with seedling-exclusive genes by 0.17 and 0.05, respectively (table 1). In contrast, there were no significant differences between sperm-exclusive genes and seedling-exclusive genes for \(\alpha\) and \(\omega_a\) (fig. 3b and c, table 1). We also validated our \(\alpha\) comparisons using other methods of Fay et al. (2001), Smith and Eyre-Walker (2002), and the maximum likelihood approach of Bierne and Eyre-Walker (2004) to estimate that parameter (supplementary table S2, Supplementary Material online). Although point estimates of \(\alpha\) for a given tissue varied depending on the method, the trend remained the same where pollen-exclusive genes had larger and sperm-exclusive genes had smaller \(\alpha\) values compared with seedling-exclusive genes respectively. Finally, we also tested the congruency of our results with divergence-based tests of positive selection using PAML. We first fit PAML models where there were two classes of sites, including neutral evolution and purifying selection. Then, we fit models that had additional parameters to allow for sites under positive selection. The model incorporating positive selection was a better fit for 17%, 20%, and 12.5% of pollen-, sperm-, and seedling-exclusive genes, respectively, compared with models that accounted for purifying selection and neutral evolution alone.

### Finer Scales of Tissue Specificity

One possible limitation with the above analysis is that the tissue specificity of the seedling-specific genes is not clear, as only a single representative sporophytic tissue was examined in this study. To address this, we also compared selection on pollen- and sperm-specific genes with selection on tissue-specific genes from multiple sporophytic tissues from other studies (Pina et al. 2005; Schmidt et al. 2011). Furthermore, repeating the analyses on the multiple expression profiles allowed us to control for the methodological variation associated with the generation of those profiles as such differences led to the discovery of novel transcripts in each study (supplementary fig. S2, Supplementary Material online). For tests of differences in purifying selection, similar patterns were found when examining additional sporophyte-specific gene sets. In particular, for the C. grandiflora orthologs identified using Pina et al. (2005) (supplementary fig. S3, Supplementary Material online), pollen-exclusive genes had significantly fewer mutations in the \(N_{\omega} < 1\) and significantly more mutations in the \(N_{\omega} > 10\) category compared with leaf-, seedling-, and silique-exclusive genes (table 2 and supplementary fig. S4a, Supplementary Material online).

**Table 2.** P Values for Comparisons between Tissues for the \(N_{\omega}\) Categories (<1, 1–10, >10) in the DFE, \(\alpha\), and \(\omega_a\) for Genes Identified Using the Pina et al. (2005) Study.

<table>
<thead>
<tr>
<th>Test</th>
<th>(N_{\omega} &lt; 1)</th>
<th>(N_{\omega} 1–10)</th>
<th>(N_{\omega} &gt; 10)</th>
<th>(\alpha)</th>
<th>(\omega_a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pollen vs. leaf*</td>
<td>0.01*</td>
<td>0.06</td>
<td>0.01*</td>
<td>0.07</td>
<td>0.24</td>
</tr>
<tr>
<td>Pollen vs. seedling*</td>
<td>0.01*</td>
<td>0.04*</td>
<td>0.01*</td>
<td>0.01*</td>
<td>0.01*</td>
</tr>
<tr>
<td>Pollen vs. silique*</td>
<td>0.01*</td>
<td>0.01*</td>
<td>0.01*</td>
<td>0.11</td>
<td>0.79</td>
</tr>
<tr>
<td>Pollen vs. silique transport*</td>
<td>0.47</td>
<td>0.01*</td>
<td>0.01*</td>
<td>0.65</td>
<td>0.95</td>
</tr>
<tr>
<td>Pollen vs. silique all transport*</td>
<td>0.01*</td>
<td>0.04*</td>
<td>0.01*</td>
<td>0.06</td>
<td>0.92</td>
</tr>
</tbody>
</table>

*Note.—Values for Comparisons between Tissues for the \(N_{\omega}\) Categories (<1, 1–10, >10) in the DFE, \(\alpha\), and \(\omega_a\) for Genes Identified Using the Schmidt et al. (2011) Study.

<table>
<thead>
<tr>
<th>Test</th>
<th>(N_{\omega} &lt; 1)</th>
<th>(N_{\omega} 1–10)</th>
<th>(N_{\omega} &gt; 10)</th>
<th>(\alpha)</th>
<th>(\omega_a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pollen vs. sporophytic silique</td>
<td>0.01*</td>
<td>0.01*</td>
<td>0.01*</td>
<td>0.01*</td>
<td>0.01*</td>
</tr>
<tr>
<td>Sperm vs. sporophytic silique</td>
<td>0.32</td>
<td>0.4</td>
<td>0.02*</td>
<td>0.92</td>
<td>0.63</td>
</tr>
</tbody>
</table>

*Note.—As both tests are part of the same statistical inference, their P value threshold for significance was adjusted by applying the FDR correction (Benjamini and Hochberg 1995) with the FDR threshold at 5%. Significant P values are indicated with an asterisk beside them.
Sperm-exclusive genes had significantly fewer mutations in the $N_s > 10$ category compared with sporophytic nucellus-exclusive genes. However, tests of the extent of positive selection showed similar trends but were not always significant among the various tissue-specific expression profile data sets. For the C. grandiflora orthologs identified using Pina et al. (2005), both $\alpha$ and $\omega_s$ were significantly higher for pollen-exclusive genes when compared with seedling-exclusive genes. Although point estimates were higher for pollen-exclusive genes when compared with leaf- or silique-exclusive genes, these differences were not significant (table 2 and supplementary fig. S4, and Supplementary Material online). In contrast, for the C. grandiflora orthologs identified using Schmidt et al. (2011), $\alpha$ and $\omega_s$ were significantly higher for pollen-exclusive genes when compared with sporophytic nucellus-exclusive genes (table 3 and supplementary fig. S5b and c, Supplementary Material online). One important assumption in our analyses is that expression profiles remain conserved from Arabidopsis to Capsella. To assess the robustness of our analyses, we further filtered our set of pollen-exclusive genes and only retained genes that were identified as being pollen-specific in a third more distantly related member of the Brassicaceae, Brassica napus (Whittle et al. 2010). Sixty-seven of the 375 genes in our original pollen-specific gene set were designated as being exclusive to pollen tissue in both Arabidopsis and B. napus. We repeated the DFE-$\alpha$ analyses on this more highly conserved pollen-exclusive gene set. We found that our point estimates for $N_s > 10$, $\alpha$, and $\omega_s$ were higher and the estimates for $N_s < 1$ were lower for the highly conserved set compared with the pollen-exclusive gene set generated using the Arabidopsis expression profile alone (supplementary fig. S6, Supplementary Material online). Moreover, these more highly conserved pollen-exclusive genes had significantly fewer mutations in the $N_s < 1$ and $N_s = 1–10$ categories and significantly more mutations in the $N_s > 10$ category of the DFE and had significantly higher $\alpha$ and $\omega_s$ compared with seedling-exclusive genes (table 1).

**Effective strength of selection ($N_s$)**

![Effective strength of selection ($N_s$)](https://academic.oup.com/mbe/article-abstract/30/11/2475/1263063)

**Tissue-Specific Differences in Molecular Function**

Another factor that might potentially confound assessments of differences in the efficacy of selection between the gametophytic and sporophytic phases would be if the collection of gene classes present in each phase varied considerably. In such a case, it is possible that the observed patterns are solely due to differences in how selection operates on some gene classes compared with others. To address this issue, we attempted to validate the observed patterns by repeating the DFE-$\alpha$ analyses on genes with similar functions. Gene ontology (GO) analyses revealed that 47 of the 375 pollen-exclusive genes and 294 of the 4,245 seedling-exclusive genes characterized using Borges et al. (2008) were enriched for the transport function (GO:0006810) (supplementary fig. S7a, Supplementary Material online). Genes annotated for the transport function in pollen had significantly fewer mutations in the $N_s < 1$ category and significantly more mutations in the $N_s > 10$ category of the DFE (fig. 4a), and had significantly higher $\alpha$ (fig. 4b) and $\omega_s$ (fig. 4c) compared with genes annotated for the same function in seedlings (table 1). These patterns remained unchanged when we repeated the DFE-$\alpha$ analyses on all genes except those annotated for the transport function (table 1 and supplementary fig. S8, Supplementary Material online). Similarly, 26 of the 238 pollen-exclusive genes and 12 of the 384 silique-exclusive genes characterized using Pina et al. (2005) were annotated for the transport function (supplementary fig. S7b, Supplementary Material online). Genes annotated for the transport function in pollen had significantly fewer mutations in the $N_s = 1–10$ category and significantly more mutations in $N_s > 10$ category of the DFE compared with genes enriched for the same function in silique (table 2 and supplementary fig. S9a, Supplementary Material online) but did not show any significant differences for $\alpha$ and $\omega_s$ (supplementary fig. S9b and c, Supplementary Material online), and similarly the patterns remained unchanged when we repeated the DFE-$\alpha$ analyses on all genes except those annotated for the transport function (table 2 and supplementary fig. S10, Supplementary Material online). No other gametophyte and sporophyte comparisons yielded greater than 10 genes in each tissue with similar functions.

**Species-Specific Expression Validation**

A further factor that might have influenced our results is the presence of differences in the expression intensities among the genes expressed in various tissues, because expression level is known to be a major predictor of protein evolution (Pál et al. 2001). Furthermore, as mentioned above, a key assumption of our analyses is that expression specificities
remain conserved between *A. thaliana* and *C. grandiflora*. Although tissue-specific gametophytic and sporophytic expression profiles are not available for *C. grandiflora*, we can make use of flower bud and leaf transcriptomes (Slotte et al. 2013) as proxies, to gain species-specific information on expression levels. Although not a perfect proxy, leaf transcriptomes serve as a good substitute for estimating expression intensities of seedling genes as they are expected to express sporophytic genes such from seedling tissues but not gametophytic genes from pollen and sperm cells. Moreover, as the list of seedling genes generated by Borges et al. (2008) was done without comparisons to multiple sporophytic tissues, it is likely that genes expressed in this list will overlap with genes expressed in other sporophytic tissues (supplementary fig. S2c, Supplementary Material online). Furthermore, flower bud transcriptomes will include gametophytic-expressed pollen and sperm genes and, therefore, they should be highly enriched relative to leaf transcriptomes for gametophytic gene expression.

We estimated levels of expression for pollen-, sperm-, and seedling-exclusive genes from *C. grandiflora* flower bud and leaf transcriptomes and calculated the mean difference in expression between the two transcriptomes for each gene. As expected, expression levels for pollen-exclusive genes were much higher in the flower bud compared with leaf transcriptomes; however, the magnitude of the difference was smaller for sperm-exclusive genes (supplementary fig. S11, Supplementary Material online), largely due to this set of genes being lowly expressed. On the other hand, expression levels for seedling-exclusive genes were higher in leaf compared with flower bud transcriptomes.

To control for differences among genes in expression level, we repeated the DFE-α analyses on genes that had expression levels between 5 and 30 fragments per million mapped reads (FPKM), using flower bud expression estimates for pollen- and sperm-specific genes and leaf expression estimates for sporophytic seedling genes. We excluded pollen- and sperm-specific genes that had FPKM values >5 in the leaf transcriptomes. Once we restricted the set of genes to those only with FPKM values from 5 to 30, a two-sample t-test assuming unequal variances revealed that there were no significant differences in levels of expression between pollen-exclusive genes and sporophytic seedling-exclusive genes (t-stat = 0.54, df = 169, P = 0.58). Consistent with earlier results, genes with 5–30 FPKM in pollen had significantly fewer mutations in the \( N_e < 1 \) category and significantly more mutations in the \( N_e > 10 \) category of the DFE (fig. 5a), and had significantly higher α (fig. 5b) compared with sporophytic genes with similar levels of expression in leaves (table 1). Although \( \omega_a \) values (fig. 5c) were also higher for the pollen-exclusive genes compared with sporophytic genes, the wider confidence intervals due to a reduction in the number of genes sampled made the tests insignificant at the 5% level after correcting for multiple comparisons (table 1). Genes with FPKM values from 5 to 30 in sperm had significantly more mutations in the \( N_e < 1 \) category and significantly fewer mutations in the \( N_e > 10 \) category of the DFE (fig. 5a) compared with sporophytic genes with similar levels of expression in leaves (table 1). However, a two-sample t-test assuming unequal variances revealed that there were significant differences in levels of expression between sperm-exclusive genes and sporophytic seedling-exclusive genes even after restricting the set of genes to those only those with FPKM values from 5 to 30 (t-stat = 6.18, df = 123, P = 8.31e−09).

**Fig. 5.** (a) The distribution of fitness effects of new nonsynonymous mutations (DFE), (b) the proportion of adaptive substitutions (α), and (c) the rate of adaptive nonsynonymous to synonymous substitutions (\( \omega_a \)) for genes with mean values for fragments per kilobase of transcript per million mapped reads (FPKM) between 5 and 30 and that are either exclusively expressed in pollen (\( n = 120 \)), sperm (\( n = 55 \)), or seedling (\( n = 1133 \)) tissues from the list of tissue-specific genes identified using the Borges et al. (2008) study. Error bars are 95% confidence intervals from 200 bootstrap replicates generated by resampling over genes.
Discussion

Our results indicate that in the obligately outcrossing *C. grandiflora*, purifying and positive selection are stronger in pollen-exclusive genes compared with genes that are exclusively expressed in seedlings and the sporophytic nucellus, possibly due to the joint effects of haploid-level selection and pollen competition. Interestingly, the efficacy of purifying selection on genes expressed in another male gametophytic tissue, the sperm, is opposite to that observed when pollen-expressed genes were used for the gametophyte and sporophyte comparisons. One possible reason for the relaxation of selection observed in sperm could be due to them being released from competition (Dresselhaus and Sprunck 2012). In the angiosperm life cycle, it is the pollen tubes that compete to gain access to ovules where they release two sperm cells. As both sperm cells fuse with cells in the female gametophyte and contribute to embryo development (Berger et al. 2008), mutations occurring in sperm-specific genes may be generally subject to less selective pressures. The exception would be very strongly deleterious mutations that prevent sperm cells from fusing. The contrasting patterns observed for sperm- versus pollen-expressed genes could also be due to imprinted genes. Such genes are known to be rapidly evolving (Wolff et al. 2011) with evidence for positive selection acting on some of them (Spillane et al. 2007; Miyake et al. 2009; O’Connell et al. 2010). As imprinted genes are silenced in sperm cells but expressed in the vegetative cell (Calarco et al. 2012), their effects in our study may be mostly seen from assessments of selection on genes exclusively expressed in the pollen tube.

Although most of our conclusions remained consistent across methods, some discrepancies were observed that likely reflect the power and assumptions of the tests. Divergence-based (PAML) methods applying a single rate of evolution for all sites in a given gene indicated greater rates of evolution in pollen and sperm tissues compared with seedlings. One limitation of this approach was the inability to disentangle the relative contributions of faster adaptive evolution and relaxed selective constraints leading to increased accumulation of substitutions in the gametophytic tissues. On the other hand, the DFE-α approach used both polymorphism and divergence to estimate the relative contributions of positive and purifying selection. The DFE-α approach might overestimate α if it does not have sufficient power to estimate the true DFE especially in cases when it is complex and multimodal (Keightley and Eyre-Walker 2010; Kousathanas and Keightley 2013). As we have sampled more than 10 alleles for each gene, the method should have had sufficient power to estimate the DFE (Keightley and Eyre-Walker 2010). Another consideration is that the DFE-α approach is better able to estimate α and the shape parameter governing the DFE under a model incorporating population size changes as we had utilized. At the same time, in such cases, it seems less robust at making demographic inferences and can overestimate the strength of purifying selection in the face of linked selection (Messer and Petrov 2013). Moreover, α can be influenced by many factors such as the rate of environmental change and the complexity of the organism (Lourenço et al. 2013). However, these concerns have limited influence in our study as we are comparing α and ωₐ values across genes within the same species. Although point estimates of α varied depending on the approach we used, the pattern remained the same where pollen-exclusive genes had larger values compared with seedling-exclusive genes, indicating that comparisons between α values are a more useful indicator for assessing the impact of positive selection rather than simply relying on their respective magnitudes. Finally, we do see some validation of our results from additional analyses based on divergence-based (PAML) methods incorporating different site classes, as models accounting for positive selection seem to be a better fit for a greater proportion of pollen-exclusive genes compared with seedling-exclusive genes. However, we also found that models accounting for positive selection seem to be a better fit for a greater proportion of sperm-exclusive genes compared with seedling-exclusive genes, in contrast with the results from the DFE-α analyses where we did not find evidence for more efficient positive selection. This result is left as an inconsistency in approaches in our study. As the DFE-α approach explicitly accounts for a distribution of deleterious selection coefficients and species-specific selection pressure, we suggest that it provides conclusions that are more reliable, indicating only less efficient purifying selection in sperm-specific genes.

One important limitation of our study is that it relies on the assumption that gene expression profiles are conserved from *Arabidopsis*. Some of our results support this assumption, as the lists of pollen- and sperm-exclusive genes we generated have greater expression intensities in *C. grandiflora* flower buds, which have both gametophytic and sporophytic tissues, when compared with sporophytic leaf transcriptomes. The magnitude of the difference in expression levels for genes in the sperm cells was not large but this is likely because the expression intensities for such genes are marginal. This pattern is consistent with the notion that expression intensity also plays a role in the efficacy of selection in sperm cells. In contrast, our lists of seedling-exclusive genes generated from Borges et al. (2008) have greater expression intensities in the sporophytic transcriptomes generated from leaf tissues when compared with those generated from the flower buds. Nevertheless, if some genes have changed their expression profiles since the species diverged, this can complicate inferences. However, our assumption is conservative with respect to testing for differences in the strength of selection, since genes that have not maintained sporophytic versus gametophytic expression would erode the differences between our gene sets. This is indicated by the much greater difference we see between pollen and seedling tissues for the NS categories (<1, 1–10, >10) in the DFE, α, and ωₐ when we use a more highly conserved list of pollen-exclusive genes for performing the comparisons. Therefore, our conclusions appear to be robust to the assumptions of conservation of expression profiles, with the caveat that low expression levels likely contribute to relaxed selection in sperm-expressed genes.
Although stronger purifying selection was also observed when pollen-exclusive genes were compared with sporophytic tissues other than seedling, including leaf- and silique-exclusive genes, those comparisons showed less consistent patterns when testing differences in positive selection. Leaf-exclusive genes had large confidence intervals for $\omega_a$ and $\omega_{a'}$, possibly reflecting the low precision of these estimates made using a small number of genes. The pattern observed for silique-exclusive genes is harder to explain. One possibility is that the high $\alpha$ and $\omega_a$ values are due to functional enrichment for genes involved in reproduction (GO:0000003) and developmental processes (GO:0032502) in silique tissues (Supplementary fig. S12a, Supplementary Material online). We did not observe enrichment for these functional categories in both the leaf (Supplementary fig. S12b, Supplementary Material online) and seedling (Supplementary fig. S12c, Supplementary Material online) tissues. Other empirical evidence also indicates that silique tissues are enriched for proteins such as transcription factors that function in reproduction (de Folter et al. 2004; Hennig et al. 2004). If so, the extent of positive selection observed in silique tissues could be due to rapid evolution acting on proteins involved in reproduction (Swanson and Vacquier 2002). Under this expectation, we should have observed significant differences in the extent of positive selection when we constrained our analysis to pollen- and silique-exclusive genes functioning in transport. As only 12 silique-exclusive transport genes were used to generate the confidence intervals for $\alpha$ and $\omega_a$, the resulting power to detect a significant difference from this comparison was very low. Although our preliminary results are encouraging, more gametophyte and sporophyte comparisons with more genes in each category are needed to be confident about the differences in extent of positive selection between the two phases. One way to achieve that might be to use sequencing approaches to uncover novel, species-specific transcripts in each tissue.

The observed patterns of the efficacy of purifying selection and level of positive selection remain consistent when we control for the effects of functionality and expression and repeated the tissue-specific comparisons using genes that are only involved in transport function or have levels of expression varying between 5 and 30. Nevertheless, many transport genes are upregulated during pollen tube growth (Bock et al. 2006; Wang et al. 2008) and, therefore, their expression level might influence the efficacy of selection. Highly expressed genes are known to encode slowly evolving proteins partly due to stronger selection against mRNA misfolding (Park et al. 2013), protein misfolding (Drummond and Wilke 2008), and protein misinteractions (Yang et al. 2012). It would be difficult to account for the observation of an increased proportion of substitutions fixed by positive selection for pollen-exclusive transport genes if level of expression was the sole factor responsible for the differential selection efficacy among tissues. Moreover, we still observed more efficient positive and purifying selection acting on pollen-exclusive genes when we controlled for the effect of differences in expression intensity. However, we were unable gain support for the possibility that factors other than expression intensity could also explain the less efficient purifying selection observed for sperm-exclusive genes as we were unable to generate a large enough number of genes that fell within a narrow breath of expression. More precise characterization of the sperm transcriptome might lead to the generation of such a list of genes.

Our study suggesting stronger positive and purifying selection in the outcrossing C. grandiflora contrasts with a study on A. thaliana (Szövényi et al. 2013), where the authors concluded that purifying selection was not stronger in gametophytic genes. Although this contrast is consistent with the expectation of stronger gametophytic selection in outcrossers, a recent study by Gossmann et al. (Gossmann TI, Schmid MW, Grossniklaus U, Schmid KJ, personal communication) applied DFE-$\alpha$ approaches to whole genome resequencing data from A. thaliana and found similar patterns to our study of C. grandiflora, showing more efficient purifying and positive selection in pollen-expressed genes. Although this pattern is unexpected, one factor highlighted by the authors to explain it is that some populations experience sufficient outcrossing (15%) in natural populations (Bomblies et al. 2010) to drive pollen competition. However, they also note that the increased outcrossing effect was likely recent and, therefore, unlikely to have led to much segregating variation within the species. The observation of stronger purifying selection on pollen genes in A. thaliana could suggest that this set of genes may also experience generally stronger selective pressures due to reproduction that may be unrelated to sexual selection. Nevertheless, their observations of a greater extent of adaptive evolution in pollen-expressed genes could be explained by the use of the outcrosser, A. lyrata, to make divergence estimates between the two species; observations of elevated positive selection may reflect the outcrossing history of the genus. The observed patterns of greater positive selection on pollen-expressed genes in outcrossers including A. lyrata and C. grandiflora is expected, as haploid-level selection and pollen competition can be substantial evolutionary forces shaping their genomes. Overall, the scope for pollen competition and sexual selection in mixed mating species merits more empirical and theoretical attention.

Genes expressed in multiple tissues seem to have a large fraction of mutations under strong purification selection (Duret and Mouchiroud 2000; Zhang and Li 2004; Liao and Zhang 2006; Slotte et al. 2011; Yang and Gaut 2011). About $58\%$ of their substitutions have been fixed by positive selection. Although this value is higher than the previous estimate for C. grandiflora of $40\%$ as indicated by Slotte et al. (2010), it is likely that their estimate was closer to the lower bound as it was based on 257 coding loci sequenced using conserved primers. The levels of strongly deleterious mutations are similar between genes expressed in multiple tissues and genes exclusively expressed in pollen tissues. One possibility is that this pattern indicates that gametophytic selection is a major force influencing genes co-expressed in both the gametophytic and sporophytic phases. Alternatively, genes expressed in multiple tissues are likely to be constitutive genes where most new nonsynonymous mutations are likely to be deleterious as they might alter critical protein function.
(Zuckerkandl 1976; Wilson et al. 1977). Such an effect is consistent with the observations that $\alpha$ and $\omega$ values are not comparable to those for pollen-expressed genes. If fewer substitutions are likely to be beneficial in such genes, the extent positive selection will likely be small.

Our results indicate that there are differences in the efficacy of selection and extent of positive selection within the gametophyte and between gametophytic and sporophytic tissues in *C. grandiflora*. Particularly, pollen-expressed genes in the male gametophyte seem to be under stronger purifying and positive selection compared with genes in vegetative sporophytic tissues. Strong selection acting on pollen-expressed genes provides molecular evidence for intrasexual competition occurring in plants where the rapid evolution can be aided by the haploid nature of that phase. Further support for this hypothesis can be gathered from the contrasting patterns observed in sperm where the genes are under weaker selection and seem to have undergone slower evolution compared with the sporophytic-expressed genes. As sperm cells are released from the effects of intrasexual competition in plants, they are less likely to be targets of sexual selection. Moreover, the patterns observed for genes co-expressed in the gametophytic and sporophytic stages are similar to those observed for pollen-expressed genes indicating that gametophytic selection might be shaping plant genomes.

**Materials and Methods**

**Genome Resequencing and SNP Calling**

We used whole genome resequencing data from 13 outbred *C. grandiflora* individuals (26 chromosomes), as described in our companion article (Williamson RJ, Josephs EB, Wright SI, unpublished data). Briefly, this study conducted Illumina 100 bp PE sequencing of a “scattered” sample (Wakeley 1998, 1999; Wakeley and Lessard 2003) of individuals from 12 populations, plus a thirteenth individual derived from a cross between two additional populations. Illumina reads were mapped to the very closely related *C. rubella* genome (Slotte et al. 2013) using the Stampy aligner I.0.13 (Lunter and Goodson 2011). These two species diverged in the past 100,000 years (Slotte et al. 2013), so cross-species mapping efficiency is high. After initial alignments, indels were realigned using the Genome Analysis Toolkit (GATK) v1.05777 indel realigner and genotype and SNP calls were conducted using the GATK Unified Genotyper under default parameters (DePristo et al. 2011). Sites with quality score $<90$ were removed. Analysis was conducted only on sites where all the individual’s read depth was between 20 and 60 and the genotype quality scores of all individuals was greater than 40. Orthologs for each gene were identified using whole genome alignments of *A. thaliana* (Haudry et al. 2013) and the close outgroup *N. paniculata* (Slotte et al. 2013). Notably, the high evolutionary conservation (~90%) between *A. thaliana* and *Capsella* aided the identification of orthologs (Boivin et al. 2004). Only loci with orthologs from all three species were retained for the analysis. The use of whole-genome alignment for ortholog identification ensured that both synteny and sequence similarity information were combined for robust ortholog identification. The coding sequences of the identified orthologs were translated and then aligned using Dialign-TX (Subramanian et al. 2008) to inform the nucleotide alignments. At the end of these analyses, there were 12,520 genes from 13 diploid *C. grandiflora* samples, *A. thaliana*, and *N. paniculata* for further analyses. We also obtained the expression levels for each *C. grandiflora* gene from 5 flower bud transcriptomes (Slotte et al. 2013) and 10 leaf transcriptomes (Williamson RJ, Josephs EB, Wright SI, unpublished data). In both cases, the pipeline laid out by Trapnell et al. (2010) for transcript assembly and quantification was used. Using *C. rubella* to provide the reference annotation, the expression level was for each sample was estimated by the fragments per kilobase of transcript per million mapped reads (FPKM) metric. We calculated the mean expression level for each gene across all samples from a given tissue and used it for our analyses.

**Tissue-Specific Expression Characterization**

We compared the *C. grandiflora* genome with the closely related related *A. thaliana* transcriptome to identify genes expressed in gametophytic and sporophytic phases. First, we used the study by Borges et al. (2008) that had identified genes exclusively expressed in the pollen, sperm, and seedling tissues and genes that were expressed in all three tissues in *A. thaliana*. Borges et al. (2008) obtained expression data for pollen and seedling tissues from Pina et al. (2005). They additionally used the expression data they generated for sperm cells to refine the tissue-specific expression profile for *A. thaliana*. Such profiling led to the identification of genes that were exclusively expressed in the vegetative tissues of pollen grains versus those that are exclusively expressed in the reproductive sperm cells. Moreover, we validated our results by repeating the tissue-specific comparisons using sporophytic tissues from two other expression profile data sets for *A. thaliana*. Pina et al. (2005) identified genes exclusively expressed in the pollen, leaf, seedling, and silique tissues. Because they did not characterize the sperm tissue, we used the data generated by Borges et al. (2008) to remove genes that overlapped between the sperm tissue and the pollen, leaf, seedling, and silique tissues from our analyses. We also used the study by Schmidt et al. (2011) that identified genes exclusively expressed in pollen, sperm, and sporophytic nucellus tissues. We also filtered and generated a more highly conserved list of pollen genes using the expression profile generated for *B. napus*. This more highly conserved list only contained *C. grandiflora* orthologs for genes that were profiled as being exclusive to the pollen tissue in both *A. thaliana* by Borges et al. (2008) and *B. napus*. The lists of genes we used in our study are given in supplementary table S3, Supplementary Material online.

**Functional Categorization**

We performed a GO analysis of the genes expressed in sporophytic tissues from Pina et al. (2005) using the agriGO toolkit and database (Du et al. 2010). Specifically, we
performed a singular enrichment analysis (SEA) using the genes expressed in each tissue as the query and Arabidopsis gene model (TAIR9) as the reference. Significant GO terms were found after performing hypergeometric tests with the Yekutieli multiple test correction at a significance level of 0.05.

We performed a SEA using the pollen-exclusive genes as the query list, the seedling-exclusive genes as the reference list, and using the previous approaches to detect significant GO terms and identified genes with similar functions in both categories. We used the results to classify genes in pollen and seedling tissues according to their biological function. Only functional categories with more than 10 genes in both pollen and seedling tissues were used for further analyses. We also repeated the previous SEA using the pollen-exclusive genes as the query list and the silique-exclusive genes as the reference list from the list of genes identified from Pina et al. (2005).

Selection Estimation

We used a modified version of Polymorphorama (Bachtrog and Andolfatto 2006; Andolfatto 2007; Haddrill et al. 2008) to generate nonsynonymous and synonymous allele frequency spectra (AFS) for each gene. As preliminary test for purifying selection, we performed \( \chi^2 \) tests of homogeneity comparing singletons to all other frequency classes as implemented in Sawyer et al. (1987). We did this test for both nonsynonymous and synonymous sites and obtained nondirectional probability estimates after applying a Yates’ correction for continuity. With \( N. \) paniculata and \( A. \) thaliana as outgroups, we additionally calculated the counts of nonsynonymous and synonymous sites, lineage-specific divergence, and \( d_{ns}/d_{s} \) for each \( C. \) grandiflora gene using PAML (Yang 2007). We calculated \( d_{ns}/d_{s} \) for \( C. \) grandiflora genes under models where 1) a single parameter was applied for all sites in a given gene, 2) individual parameters were given for sites under neutral evolution and purifying selection, and 3) individual parameters were given for sites under neutral evolution, purifying selection, and positive selection. For a subset of genes, we did a likelihood ratio test comparing the latter two models to assess whether the model incorporating positive selection was a better fit than one without. Tissue-specific nonsynonymous and synonymous AFS, numbers of sites, and divergence summed across all genes used in the study are given in supplementary table S4, Supplementary Material online.

We used the DFE-\( \alpha \) software to infer the DFE for deleterious mutations, \( \alpha_c \) and \( \omega_c \), which implements the methods of Eyre-Walker and Keightley (2009) and Gossmann et al. (2010). For data summaries and statistical tests, we combined the \( N_s \) 10–100 and > 100 categories into a single : : > 10 category as with a sample size of 26 chromosomes, there is unlikely to be high power to distinguish mutations in these two categories. As the DFE-\( \alpha \) is a maximum likelihood model, individual runs can stagnate at local optima. Using the source code version of the DFE-\( \alpha \) program available on Peter Keightley’s lab website (http://lanner.cap.ed.ac.uk/~eang33/est-dfe-files.tar.gz, last accessed September 17, 2013), we performed 10 runs for each analysis with random starting values for the mean s for deleterious mutations, \( t_2 \), and \( \beta \) parameters and picked the parameterization that resulted in the highest maximum likelihood.

For the \( N_s \) categories, \( \alpha_c \), \( \omega_c \), and \( d_{ns}/d_{s} \) parameters, we generated 200 bootstrap replicates by resampling over genes using R (R Development Core Team 2011). For each parameter, we used the bootstrap results to estimate 95% confidence intervals. We tested whether the \( N_s \) categories, \( \alpha_c \), \( \omega_c \), and \( d_{ns}/d_{s} \) differed among various tissue categories using the randomization test as applied in Keightley and Eyre-Walker (2007). When necessary, we adjusted the \( P \) values for multiple tests by applying the false discovery rate (FDR) correction (Benjamini and Hochberg 1995) with the FDR threshold at 5%. We also used the approaches of Fay et al. (2001), Smith and Eyre-Walker (2002) and the maximum likelihood approach of Bienne and Eyre-Walker (2004) to validate our point estimates of \( \alpha_c \) as implemented in the DoFE program available on AdamEyre-Walker’s lab website (http://www.lifeisci.sussex.ac.uk/home/Adam_Eyre-Walker/Website/Software_files/DoFE%203.0%20-%20windows.zip, last accessed September 17, 2013).

Supplementary Material

Supplementary figures S1—S12 and table S1–S4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Toni I. Gossmann and colleagues for providing an early draft of their manuscript on tissue-specific patterns of selection in \( A. \) thaliana, which is currently in revision. This work was supported by the Natural Science and Engineering Research Council of Canada (NSERC) (Discovery Grant to S.I.W. and Canada Graduate Scholarships support to R.A.), Ministry of Training, Colleges and Universities Ontario Graduate Scholarship to R.A., National Science Foundation (NSF) Graduate Research Fellowship support (DGE-1048376 to E.B.J.), and by the Genome Canada/Genome Quebec grant to T. Bureau, S.I.W., M. Blanchette, D. Schoen, P. Harrison, and J. Stinchcombe. \( C. \) rubella genome sequence data were produced by the U.S. Department of Energy Joint Genome Institute, supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References


Selection on Pollen- and Sperm-Specific Genes · doi:10.1093/molbev/mst149


