Pollen-specific genes accumulate more deleterious mutations than sporophytic genes under relaxed purifying selection in *Arabidopsis thaliana.*

MC Harrison, EB Mallon, D Twell, RL Hammond

Dept. Biology

University of Leicester

University Road

Leicester, LE1 7RH

Phone: ++ (0)116 252 3339

Fax: +44 (0)116 252 3330

E-mail: mch44@le.ac.uk

Keywords: Purifying selection, sporophyte, pollen, ploidy, deleterious, masking

Running title: "Relaxed selection on pollen-specific genes."

Abstract

The strength of purifying selection varies among loci and leads to differing frequencies of deleterious alleles within genomes. Selection is generally stronger for highly and broadly expressed genes but can be less efficient for diploid expressed, deleterious alleles if heterozygous. In plants expression level, tissue specificity and ploidy level differ between pollen specific and sporophyte specific genes. This may explain why the reported strength and direction of the relationship between selection and the specificity of a gene to either pollen or sporophytic tissues varies between studies and species. In this study, we investigate the individual effects of expression level and tissue specificity on selection efficacy within pollen genes and sporophytic genes of *Arabidopsis thaliana*. Due to high homozygosity levels caused by selfing, masking is expected to play a lesser role. We find that expression level and tissue specificity independently influence selection in *A. thaliana*. Furthermore, contrary to expectations, pollen genes are evolving faster due to relaxed purifying selection and have accumulated a higher frequency of deleterious alleles. This suggests that high homozygosity levels resulting from high selfing rates reduce the effects of pollen competition and masking in *A. thaliana*, so that the high tissue specificity and expression noise of pollen genes are leading to lower selection efficacy compared to sporophyte genes.

Introduction

Gene expression is arguably the most important component in how variation at the genetic level leads to variation at the phenotypic level, and therefore on how selection acts (Fay and Wu, 2003; Drummond *et al.*, 2005; Rocha, 2006). Likewise, variation in expression among genes will lead to varying levels of selection within the same genome. Indeed, a significant correlation between expression level and the evolutionary rate of proteins has been reported for a wide range of taxa including bacteria (Rocha and Danchin, 2004), yeast (Pál *et al.*, 2001; Drummond *et al.*, 2006), *Drosophila* (Marais *et al.*, 2004) and *Arabidopsis thaliana* (Wright *et al.*, 2004; Wright and Andolfatto, 2008; Slotte *et al.*, 2011; Yang and Gaut, 2011). Furthermore, selection on broadly expressed genes is generally stronger than for tissue specific genes (Duret and Mouchiroud, 2000; Liao *et al.*, 2006).

The restriction of a gene's expression to reproductive tissues also has an effect on selection strength. Across a broad range of taxa, including mammals, *Drosophila*, mollusks and fungi, genes involved in reproduction have been reported to evolve more rapidly than somatic genes due to increased positive selection (Swanson and Vacquier, 2002; Haerty *et al.*, 2007; Turner and Hoekstra, 2008). However, isolating the strength and direction of the relationship between the involvement of a gene in reproduction and the efficacy of selection acting on that gene is not straight forward for some plant species (Arunkumar *et al.*, 2013; Gossmann *et al.*, 2013; Szövényi *et al.*, 2013). This is because of the potentially confounding effects of differences between pollen genes and sporophytic genes in expression level and breadth, but also in ploidy level. Whether a gene is haploid or diploid can also effect its visibility to selection. The masking hypothesis describes the less efficient purging of deleterious alleles in diploids than in haploids due to masking by a dominant homologue when heterozygous (Kondrashov and Crow, 1991). For example, in the outcrossing crucifer, *Capsella grandiflora*, genes with expression restricted to the male gametophyte revealed evidence for more efficient purifying and adaptive selection than for sporophytic genes (Arunkumar *et al.*, 2013). The stronger selection on male gametophytic genes was interpreted as resulting from the combined effects of haploid expression and pollen competition, however, the relative contributions of these two factors were difficult to disentangle. In the moss *Funaria hygrometrica*, on the other hand, little or no difference was observed between the divergence rates of pollen and sporophyte-specific proteins (Szövényi *et al.*, 2013), but variation in tissue specificity, a potentially important confounding factor, was not considered.

Self-compatible *Arabidopsis thaliana* may offer the opportunity to isolate the contribution of the reproductive role of pollen-specific genes on selection efficacy from differences in ploidy. This is because high selfing rates lead to high homozygosity in *A. thaliana* populations (Nordborg, 2000; Wright *et al.*, 2008; Platt *et al.*, 2010), so the masking of deleterious alleles in diploid sporophyte stages compared to

the haploid gametophyte stage is *a priori* likely to be much reduced. Furthermore, selfing reduces the magnitude of pollen competition, and so the strength of selection acting on pollen, as fewer genotypes compete for fertilization (Charlesworth and Charlesworth, 1992). In a recent study pollen-specific genes were found to contain a higher number of non-synonymous sites under purifying and adaptive selection than random genes sampled from the *A. thaliana* genome (Gossmann *et al.*, 2013). Importantly though, differences in expression level and tissue specificity between gene groups were not controlled for in that study. In contrast, a further study found pollen-specific genes to be evolving faster than sporophytic genes due to relaxed purifying selection in *A. thaliana* (Szövényi *et al.*, 2013). This was believed to be caused by a combination of high tissue specificity and higher expression noise in pollen compared to sporophytic genes. However, the individual effect of tissue specificity was not isolated.

In this study we aimed to isolate the individual effects of expression level, tissue specificity and the reproductive role of a gene on selection in *A. thaliana*. To investigate efficacy of selection, we analyzed levels of polymorphism within 269 *A. thaliana* strains and sequence divergence from the sister taxon *A. lyrata*. We also compared the frequency of deleterious mutations (premature stop codons and frameshift mutations) among loci. Expression level was expected to correlate positively and tissue specificity negatively with selection pressure. We, therefore, controlled for expression level and tissue specificity when comparing between pollen and sporophyte genes.

Results

Expression level, tissue specificity and life-stage limited expression are inter-related

Within the total data set containing 19,970 genes, expression level per gene (see *Methods* for details) ranged from 0 (not reliably detectable) to 19,470 with a median of 794.5 (IQR: 1,454) and a mean of 1,449 ±14.6 (standard error of the mean, sem). Tissue specificity (τ), which ranged from 0 to 1.0 with a mean of 0.572 ±0.002 (sem) and a median of 0.566 (IQR: 0.510), was significantly negatively correlated with expression level (ρ = -0.41; p < 2.2x10^{-16}; Spearman's rank correlation). That is, broadly expressed genes were generally expressed at a higher level.

Of the 16,360 genes with reliably detectable expression (see *Methods*), 1,503 genes were expressed only in pollen and a further 5,398 were limited to sporophytic tissues (referred to as pollen-specific genes and sporophyte-specific genes in this study). Pollen-specific and sporophyte-specific genes were randomly distributed among the five chromosomes (table 1), and their distributions within the chromosomes also did not differ significantly from each other (table 2).

Pollen and sporophyte-limited genes differed significantly from each other in terms of expression level and tissue specificity. Naturally, tissue specificity was higher among pollen genes (median: 0.934, IQR:

0.160) than sporophyte genes (median: 0.812, IQR: 0.301), and the difference was highly significant (W = 5.7 x 10^6; p = 8.3 x 10^{-130}; Mann Whitney U test; fig. 1). Although broadly expressed genes were generally highly expressed, the sporophyte genes were expressed at a significantly lower level than the highly tissue specific pollen genes (pollen median: 1,293, IQR: 2,590; sporophyte median: 659, IQR: 1,022; W= 5.3x10^6, p=1.8x10^{-69}; Mann Whitney U test; fig. 1).

Expression level correlates with dN/dS, pN/pS and frequency of deleterious alleles

Sequence divergence, measured via interspecific dN/dS (rate of non-synonymous substitutions per non-synonymous site versus rate of synonymous substitutions per synonymous site between *A. thaliana* and *A. lyrata*), was significantly negatively correlated with expression level (ρ = -0.32; p < 2.2x10^{-16}; Spearman's rank correlation; table 3). This means that genes expressed at a low level have evolved more quickly than highly expressed genes between the two taxa. To determine whether stronger purifying selection among highly expressed genes is causing their slower evolution or lowly expressed genes are in fact evolving quickly as a consequence of elevated positive selection, intraspecific pN/pS (as with dN/dS but using within species substitution rates) was analyzed. pN/pS also significantly negatively correlated with expression (ρ = -0.17; p < 2.2x10^{-16}; Spearman's rank correlation; table 3, first row), meaning highly expressed genes not only diverge more slowly from the sister taxon *A. lyrata*, but are also less divergent between strains of *A. thaliana*. This is an indication of stronger purifying selection acting on highly expressed genes and relaxed selection among lowly expressed genes. This was corroborated by significant, negative correlations of expression level with the frequency of unique alleles resulting from premature stop codons (ρ = -0.12; p < 2.2x10^{-16}; Spearman's rank correlation; table 3) and the frequency of frameshift mutations (ρ = -0.25; p < 2.2x10^{-16}; Spearman's rank correlation; table 3). In order to control for τ the correlations were calculated within ten sub-groups of genes according to their τ values. All correlations remained negative and the majority significant (34 out of 40 significant correlations; table 3).

Tissue specificity correlates with dN/dS, pN/pS and frequency of deleterious alleles

A significant, positive correlation existed between tissue specificity and sequence divergence (ρ = 0.25; p < 2.2x10^{-16}; Spearman's rank correlation; table 4) suggesting more broadly expressed genes are subjected to stronger purifying selection. This was further supported by a positive correlation between τ and pN/pS (ρ = 0.17; p < 2.2x10^{-16}; Spearman's rank correlation; table 4). The frequency of deleterious alleles also correlated positively and significantly with τ, the highest frequency of stop codons and frameshifts occurring among the most tissue specific genes (stop codons: ρ = 0.07; p < 2.2x10^{-16}; frameshifts: ρ

= 0.20; p < 2.2x10^{-16}; Spearman's rank correlation; table 4). In order to control for the influence of expression level, the genes were allocated to four equally sized subgroups according to their expression level, and the correlations with τ were re-calculated within these subgroups. The correlations remained positive and significant for all four quartile groups for dN/dS, pN/pS, and frameshifts, and for the 3rd and 4th expression quartiles for stop codons (table 4).

Pollen genes under weaker selection

Pollen-specific genes seem to be evolving more quickly than sporophyte-specific genes in *A. thaliana* indicated by significantly higher dN/dS ratios (pollen median: 0.206, IQR: 0.217; sporophyte median: 0.164, IQR: 0.144; W= 1.3x10^{6}, p=3.2x10^{-14}, Mann Whitney U test; fig. 2). This appears to be due to more relaxed purifying selection acting on pollen-specific genes revealed by significantly higher pN/pS values (pollen median: 0.095, IQR: 0.208; sporophyte median: 0.072, IQR: 0.177; W= 4.0x10^{6}, p=8.4x10^{-6}, Mann Whitney U test; fig. 2) and significantly higher frequencies of stop codons (pollen mean: 1.200 ±0.041 sem; sporophyte mean: 0.873 ±0.019 sem; W= 4.6x10^{6}, p=1.1x10^{-18}, Mann Whitney U test; fig. 2) and frameshifts (pollen mean: 0.020 ±0.002; sporophyte mean: 0.014 ±0.001; W= 4.6x10^{6}, p=8.3x10^{-26}, Mann Whitney U test; fig. 2) among pollen-specific genes compared to sporophyte-specific genes.

To test whether the more relaxed selection pressure on pollen-specific genes was due to their higher tissue specificity, divergence, polymorphism and frequency of deleterious alleles were also calculated among tissue specific sporophyte-specific genes. Among the 1,690 sporophyte-specific genes (31.3%) and 790 pollen-specific genes (52.6%) with a τ value between 0.9 and 1.0, divergence, polymorphism and frequency of deleterious alleles remained significantly higher among the pollen-specific gene subset (fig. 3).

Within these gene sub-groups of high tissue specificity, expression was significantly higher within pollen-specific genes than sporophyte-specific genes. In order to control for expression level, we further analyzed those highly tissue-specific genes ($\tau \geq$0.9), which had an expression level over 1,000. Within this group neither expression nor τ differed significantly between pollen-specific and sporophyte-specific genes. However, dN/dS, pN/pS, stop codons and frameshifts were all significantly higher among pollen-specific than sporophyte-specific genes (fig. 4).

Discussion

We investigated the role of three factors on the efficacy of selection on genes in *Arabidopsis thaliana*: expression level, tissue specificity and the restriction of expression to pollen. Higher selection efficacy

was expected among highly and broadly expressed genes and even more so in pollen genes compared to sporophyte genes.

First, we found a significant negative correlation between expression level and rates of protein evolution dN/dS), polymorphism (pN/pS) and the frequency of deleterious alleles (stop codon and frameshift mutations). Second, there is a significant positive correlation between tissue specificity and dN/dS, pN/pS, stop codon frequency and frameshift frequency. Third, dN/dS, pN/pS and the frequency of deleterious mutations were all significantly higher among pollen genes than sporophyte genes, even when controlling for tissue specificity and expression level.

The importance of gene visibility to selection

The negative correlation of expression level with dN/dS and pN/pS indicates a positive relationship between expression level and purifying selection. This suggests that highly expressed genes are more likely to be constrained by purifying selection, whereas genes expressed at lower levels are less constrained. Indeed, more relaxed selection reducing the purging of deleterious alleles is confirmed by the significantly higher frequency of deleterious alleles among lowly expressed genes. Purifying selection was also stronger for broadly expressed genes, while the faster evolution of tissue-specific genes suggests relaxed selection. Importantly, however, although tissue specificity and expression level were significantly negatively correlated with each other, correlations with dN/dS and pN/pS remained significant when each was controlled for.

The effect of expression level (Rocha and Danchin, 2004; Pál *et al.*, 2001; Drummond *et al.*, 2006; Marais *et al.*, 2004; Wright *et al.*, 2004; Wright and Andolfatto, 2008; Slotte *et al.*, 2011; Yang and Gaut, 2011) and tissue specificity (Duret and Mouchiroud, 2000; Liao *et al.*, 2006) on selection has been confirmed in many previous studies for a broad range of taxa. Importantly, in this study we have confirmed that both factors independently have a significant effect on the efficacy of selection acting on genetic variation in *A. thaliana*.

Purifying selection is more relaxed for pollen-specific genes

Contrary to our expectations, we have discovered evidence for more relaxed purifying selection among genes exclusively expressed in pollen compared to sporophyte limited genes. This was true despite significantly higher expression levels among pollen genes and remained true when controlling for tissue specificity by comparing pollen genes only with the most tissue specific sporophyte genes. Therefore, the faster evolutionary rates of pollen-specific compared to sporophyte-specific genes due to relaxed purifying selection cannot be explained by differences in expression level or tissue specificity.

These results are in contrast to the findings of two recent studies, in which pollen genes were found to be under stronger purifying and adaptive selection than sporophyte genes in *Capsella grandiflora* (Arunkumar *et al.*, 2013) and *A. thaliana* (Gossmann *et al.*, 2013). The results of the *A. thaliana* study were based on a comparison between pollen-specific genes and a relatively small group of 476 random genes (excluding reproductive genes), presumably comprising mainly sporophytic genes (Gossmann *et al.*, 2013). In this comparison differences in expression level and tissue specificity between gene groups were not controlled for. However, we have shown here that pollen-specific genes are expressed at a significantly higher level than sporophytic genes as previously shown for *Arabidopsis* (Honys and Twell, 2003), making them more visible to selection. This was even more apparent in the Gossmann *et al.* study (2013) because they separated sperm-specific genes, which are generally expressed at a lower level, from pollen-specific genes.

In the case of the outcrossing *C. grandiflora*, the more efficient purifying and adaptive selection on pollen genes was linked to two possible factors: haploid expression and pollen competition. *A. thaliana* is a highly self-fertilizing species with selfing rates generally in the range of 95 - 99% (Platt *et al.*, 2010), so *a priori* haploid expression is unlikely to improve the efficacy of selection on pollen-specific genes relative to sporophyte genes. This is because most individuals found in natural populations are homozygous for the majority of loci, reducing the masking of deleterious alleles in heterozygous state when expressed in a diploid tissue (Platt *et al.*, 2010).

But even in the complete absence of masking, pollen competition may be expected to generate more effective selection on pollen genes than sporophyte genes. A reduction in pollen competition can be expected due to the probably limited number of pollen genotypes in highly selfing populations (Charlesworth and Charlesworth, 1992; Mazer *et al.*, 2010). However, outcrossing does occur in natural *A. thaliana* populations with one study reporting an effective outcrossing rate in one German population of 14.5% (Bomblies *et al.*, 2010). Nevertheless, it appears that these generally rare outcrossing events may not be sufficient to prevent a reduction in pollen competition for *A. thaliana*.

So if we assume both masking and pollen competition are negligible forces when comparing selection on pollen-specific genes to sporophyte-specific genes, why is selection more relaxed among pollen-specific genes than sporophyte-specific genes? In fact, our results confirm recent findings indicating relaxed purifying selection in pollen specific genes compared to sporophytic genes in *A. thaliana* (Szövényi *et al.*, 2013), a pattern explained by a combination of high tissue specificity and higher expression noise in pollen compared to sporophytic genes. However, the authors did not compare selection on pollen genes to tissue specific sporophyte genes suggesting tissue specificity as an alternative explanation. We have shown here that tissue specificity does not explain why selection is more relaxed among pollen genes, as divergence, polymorphism and the frequency of deleterious alleles were still significantly lower in

tissue specific sporophyte genes than pollen-specific genes. Higher expression noise could, however, be an important factor influencing the level of deleterious alleles which exist for pollen genes in *A. thaliana*.

Expression noise has been found to reduce the efficacy of selection substantially and is expected to be considerably higher for haploid expressed genes (Wang and Zhang, 2011). It is, therefore, likely that in the absence of pollen competition and the masking of deleterious sporophyte-specific genes, expression noise becomes a dominant factor for pollen-specific genes of selfing plants. This leads to a reduction in selection efficacy and the accumulation of deleterious alleles in pollen-specific genes.

Conclusion

Our results confirm the effect of both expression level and tissue specificity on selection efficacy. In outcrossing plants, haploid expression and pollen competition, combined with high expression levels outweigh the negative impact of high tissue specificity and expression noise on the selection efficacy of pollen-specific genes. In the self-compatible *A. thaliana* high homozygosity likely reduces the counteracting effects of pollen competition and haploid expression, leading to lower selection efficacy and increased accumulation of deleterious mutations in pollen-specific compared to sporophyte-specific genes.

Methods

Genomic data

Publicly available variation data were obtained for 269 inbred strains of *A. thaliana*. Beside the reference genome of the Columbia strain (Col-0), which was released in 2000 (*Arabidopsis*, Genome Initiative), 250 were obtained from the 1001 genomes data center (http://1001genomes.org/datacenter/; accessed September 2013), 170 of which were sequenced by the Salk Institute (Schmitz *et al.*, 2013) and 80 at the Max Planck Institute, Tübingen (Cao *et al.*, 2011). A further 18 were downloaded from the 19 genomes project (http://mus.well.ox.ac.uk/; accessed September 2013; Gan *et al.* (2011)). These 268 files contained information on SNPs and indels recorded for separate inbred strains compared to the reference genome. A quality filter was applied to all files, in order to retain only SNPs and indels with a phred score of at least 25.

Expression data

Normalized microarray data, covering 19,970 genes specific to different developmental stages and tissues of *A. thaliana* (table 5), were obtained from Borg *et al.* (2011). The expression data consisted of 7 pollen and 10 sporophyte data sets (table 5). Four of the pollen data sets represented expression patterns of the pollen developmental stages, uninucleate, bicellular, tricellular and mature pollen grain, one contained

expression data of sperm cells and the remaining two were pollen tube data sets. There was a strong, significant correlation between the two pollen tube data sets ($\rho = 0.976$; $p < 2.2x10^{-16}$; Spearman's rank correlation), so both were combined and the highest expression value of the two sets was used for each gene. Each of the 10 sporophyte data sets contained expression data for specific sporophytic tissues (table 5).

Each expression data point consisted of a normalized expression level (ranging from 0 to around 20,000, scalable and linear across all data points and data sets) and a presence score ranging from 0 to 1 based on its reliability of detection across repeats, as calculated by the MAS5.0 algorithm (Borg *et al.*, 2011). In our analyses expression levels were conservatively considered as present if they had a presence score of at least 0.9, while all other values were regarded as zero expression. All analyses were repeated using a less conservative cut-off value of 0.7 and 0.5 (data not shown). This did not change the tendency of results obtained with the 0.9 cut-off.

Genes were classed as either pollen or sporophyte-specific genes, if expression was reliably detectable in only pollen or only sporophyte tissues or developmental stages. The highest expression value across all tissues or developmental stages was used to define the expression level of a particular gene.

Detecting signatures of selection

Evolutionary Rates

To estimate evolutionary rates of genes, dN/dS ratios (ratio of non-synonymous to synonymous substitution rates relative to the number of corresponding non-synonymous and synonymous sites) were calculated for all orthologous genes (15,772) between *A. thaliana* and *A. lyrata* and based on the TAIR 9 genome release (Szövényi *et al.*, 2013).

Intra-specific polymorphism

pN/pS ratios were calculated with the yn00 programme within PAML (Phylogenetic Analysis by Maximum Likelihood, version 4.6, Yang 2007) for each pairwise comparison of strains. The individual pN/pS estimates achieved via the Nei-Gojobori method were extracted from the output files and averaged across all pairwise comparisons for each gene.

Putatively deleterious alleles

To quantify the frequency of deleterious mutations for each gene, the occurrence of premature stop codons and frame shifts was calculated for each gene locus among all 269 strains. Stop codons were recorded as the number of unique alternative alleles occurring within the 269 strains as a result of a premature stop codon. Frame shifts were calculated as a proportion of the strains containing a frame shift mutation for

a particular gene. All analyses of coding regions were based on the representative splice models of the 27,202 *A. thaliana* genes (TAIR10 genome release www.arabidopsis.org).

Acknowledgements

MCH was supported by a PhD research grant from the Natural Environment Research Council (NERC). DT would like to acknowledge financial support from the UK Biotechnology and Biological Science Research Council (BBSRC).

References

Arunkumar, R., Josephs, E. B., Williamson, R. J., and Wright, S. I. 2013. Pollen-Specific, but Not Sperm-Specific, Genes Show Stronger Purifying Selection and Higher Rates of Positive Selection Than Sporophytic Genes in Capsella grandiflora. *Molecular Biology and Evolution*, 30(11): 2475–2486.

Bomblies, K., Yant, L., Laitinen, R. A., Kim, S.-T., Hollister, J. D., Warthmann, N., Fitz, J., and Weigel, D. 2010. Local-Scale Patterns of Genetic Variability, Outcrossing, and Spatial Structure in Natural Stands of Arabidopsis thaliana. *PLoS Genet*, 6(3): e1000890.

Borg, M., Brownfield, L., Khatab, H., Sidorova, A., Lingaya, M., and Twell, D. 2011. The R2r3 MYB Transcription Factor DUO1 Activates a Male Germline-Specific Regulon Essential for Sperm Cell Differentiation in Arabidopsis. *The Plant Cell Online*, 23(2): 534–549.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43(10): 956–963.

Charlesworth, D. and Charlesworth, B. 1992. The Effects of Selection in the Gametophyte Stage on Mutational Load. *Evolution*, 46(3): 703–720.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40): 14338–14343.

Drummond, D. A., Raval, A., and Wilke, C. O. 2006. A Single Determinant Dominates the Rate of Yeast Protein Evolution. *Molecular Biology and Evolution*, 23(2): 327–337.

Duret, L. and Mouchiroud, D. 2000. Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Molecular Biology and Evolution*, 17(1): 68–070.

Fay, J. C. and Wu, C.-I. 2003. Sequence Divergence, Functional Constraint, and Selection in Protein Evolution. *Annual Review of Genomics and Human Genetics*, 4(1): 213–235.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., Kover, P. X., Clark, R. M., Rätsch, G., and Mott, R. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, 477(7365): 419–423.

Gossmann, T. I., Schmid, M. W., Grossniklaus, U., and Schmid, K. J. 2013. Selection-Driven Evolution of Sex-Biased Genes Is Consistent with Sexual Selection in Arabidopsis thaliana. *Molecular Biology and Evolution*, page mst226.

Haerty, W., Jagadeeshan, S., Kulathinal, R. J., Wong, A., Ravi Ram, K., Sirot, L. K., Levesque, L., Artieri, C. G., Wolfner, M. F., Civetta, A., and Singh, R. S. 2007. Evolution in the Fast Lane: Rapidly Evolving Sex-Related Genes in Drosophila. *Genetics*, 177(3): 1321–1335.

Honys, D. and Twell, D. 2003. Comparative Analysis of the Arabidopsis Pollen Transcriptome. *Plant Physiology*, 132(2): 640–652.

Kondrashov, A. S. and Crow, J. F. 1991. Haploidy or diploidy: which is better? *Nature*, 351(6324): 314–315.

Liao, B.-Y., Scott, N. M., and Zhang, J. 2006. Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian Proteins. *Molecular Biology and Evolution*, 23(11): 2072–2080.

Marais, G., Domazet-Lošo, T., Tautz, D., and Charlesworth, B. 2004. Correlated Evolution of Synonymous and Nonsynonymous Sites in Drosophila. *Journal of Molecular Evolution*, 59(6): 771–779.

Mazer, S. J., Hove, A. A., Miller, B. S., and Barbet-Massin, M. 2010. The joint evolution of mating system and pollen performance: Predictions regarding male gametophytic evolution in selfers vs. outcrossers. *Perspectives in Plant Ecology, Evolution and Systematics*, 12(1): 31–41.

Nordborg, M. 2000. Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization. *Genetics*, 154(2): 923–929.

Platt, A., Horton, M., Huang, Y. S., Li, Y., Anastasio, A. E., Mulyati, N. W., Ågren, J., Bossdorf, O., Byers, D., Donohue, K., Dunning, M., Holub, E. B., Hudson, A., Le Corre, V., Loudet, O., Roux, F., Warthmann, N., Weigel, D., Rivero, L., Scholl, R., Nordborg, M., Bergelson, J., and Borevitz, J. O. 2010. The Scale of Population Structure in Arabidopsis thaliana. *PLoS Genet*, 6(2): e1000843.

Pál, C., Papp, B., and Hurst, L. D. 2001. Highly Expressed Genes in Yeast Evolve Slowly. *Genetics*, 158(2): 927–931.

Rocha, E. P. C. 2006. The quest for the universals of protein evolution. *Trends in Genetics*, 22(8): 412–416.

Rocha, E. P. C. and Danchin, A. 2004. An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. *Molecular Biology and Evolution*, 21(1): 108–116.

Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R. B., Chen, H., Schork, N. J., and Ecker, J. R. 2013. Patterns of population epigenomic diversity. *Nature*, 495(7440): 193–198.

Slotte, T., Bataillon, T., Hansen, T. T., St. Onge, K., Wright, S. I., and Schierup, M. H. 2011. Genomic Determinants of Protein Evolution and Polymorphism in Arabidopsis. *Genome Biology and Evolution*, 3: 1210–1219.

Swanson, W. J. and Vacquier, V. D. 2002. Reproductive Protein Evolution. *Annual Review of Ecology and Systematics*, 33: 161–179. ArticleType: research-article / Full publication date: 2002 / Copyright © 2002 Annual Reviews.

Szövényi, P., Ricca, M., Hock, Z., Shaw, J. A., Shimizu, K. K., and Wagner, A. 2013. Selection is no more efficient in haploid than in diploid life stages of an angiosperm and a moss. *Molecular Biology and Evolution*, page mst095.

Turner, L. M. and Hoekstra, H. E. 2008. Causes and consequences of the evolution of reproductive proteins. *The International Journal of Developmental Biology*, 52(5-6): 769–780.

Wang, Z. and Zhang, J. 2011. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proceedings of the National Academy of Sciences of the United States of America*, 108(16): E67–E76.

Wright, S., Ness, R., Foxe, J., and Barrett, S. 2008. Genomic Consequences of Outcrossing and Selfing in Plants. *International Journal of Plant Sciences*, 169(1): 105–118.

Wright, S. I. and Andolfatto, P. 2008. The Impact of Natural Selection on the Genome: Emerging Patterns in *Drosophila* and *Arabidopsis*. *Annual Review of Ecology, Evolution, and Systematics*, 39(1): 193–213.

Wright, S. I., Yau, C. B. K., Looseley, M., and Meyers, B. C. 2004. Effects of Gene Expression on Molecular Evolution in Arabidopsis thaliana and Arabidopsis lyrata. *Molecular Biology and Evolution*, 21(9): 1719–1726.

Yang, L. and Gaut, B. S. 2011. Factors that Contribute to Variation in Evolutionary Rate among Arabidopsis Genes. *Molecular Biology and Evolution*, 28(8): 2359–2369.

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8): 1586–1591.

Author contributions

All four authors developed the project idea and were involved in the interpretation of data and finalization of the manuscript. MCH analyzed the data and drafted the manuscript

353 *Figures*

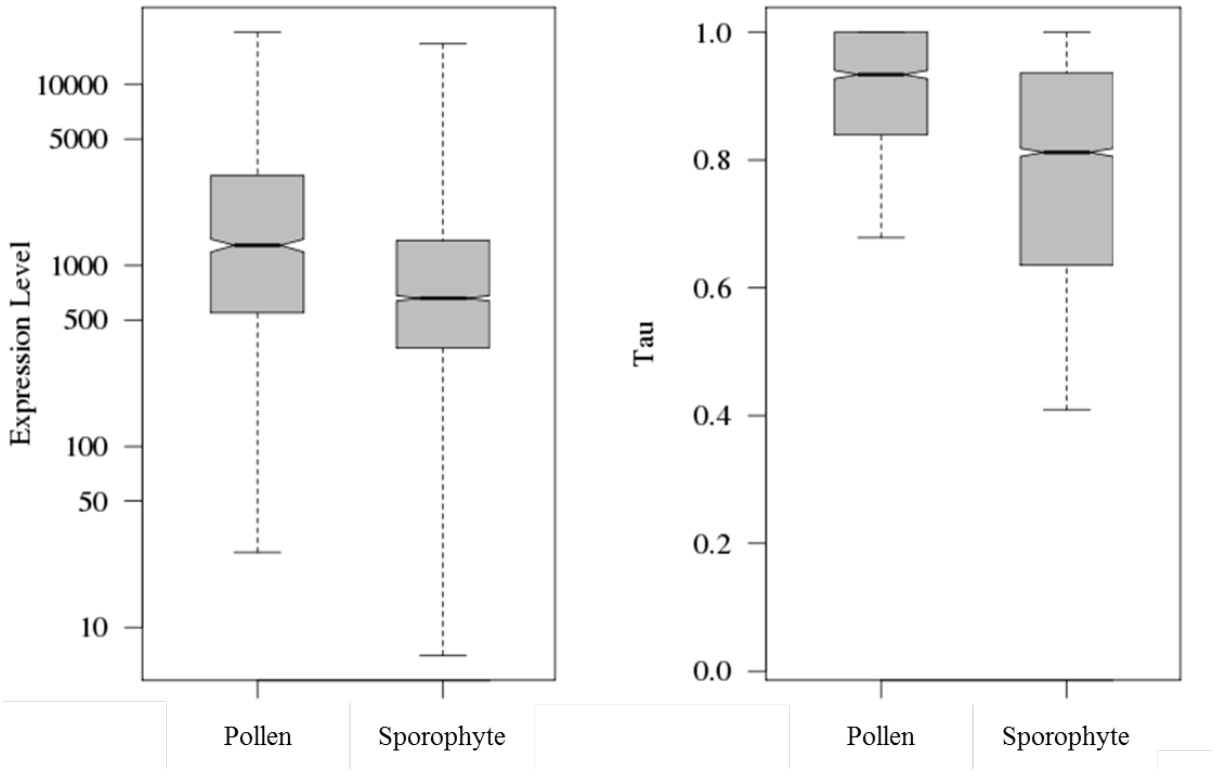

Figure 1: Expression level (left) and tissue specificity (τ, right) for pollen and sporophyte genes. Expression level (W = 5.3 x 10^6; p = 1.8 x 10^{-69}) and τ (W = 5.7 x 10^6; p = 8.3 x 10^{-130}) differ significantly between pollen and sporophyte genes; Mann Whitney U test.

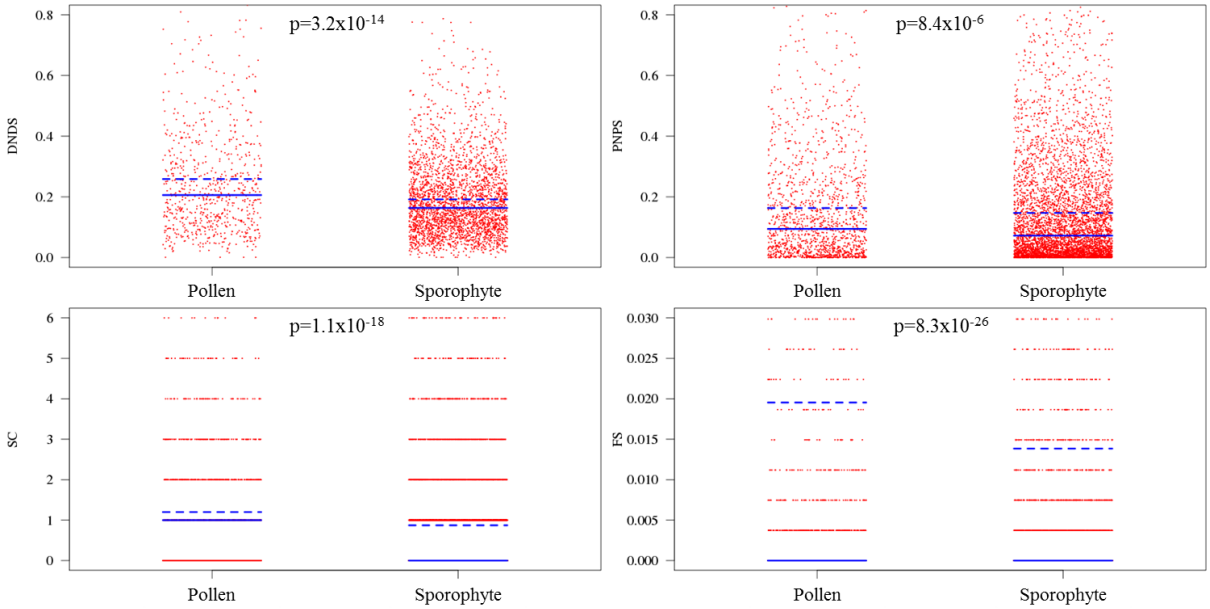

Figure 2: dN/dS (top left), pN/pS (top right), stop codon frequency (bottom left) and frameshift frequency for pollen-specific (n=1,503) and sporophyte-specific (n=5,398) genes. Each data point represents a gene; solid line is the mean and dashed line the median. Differences tested with Mann-Whitney U test.

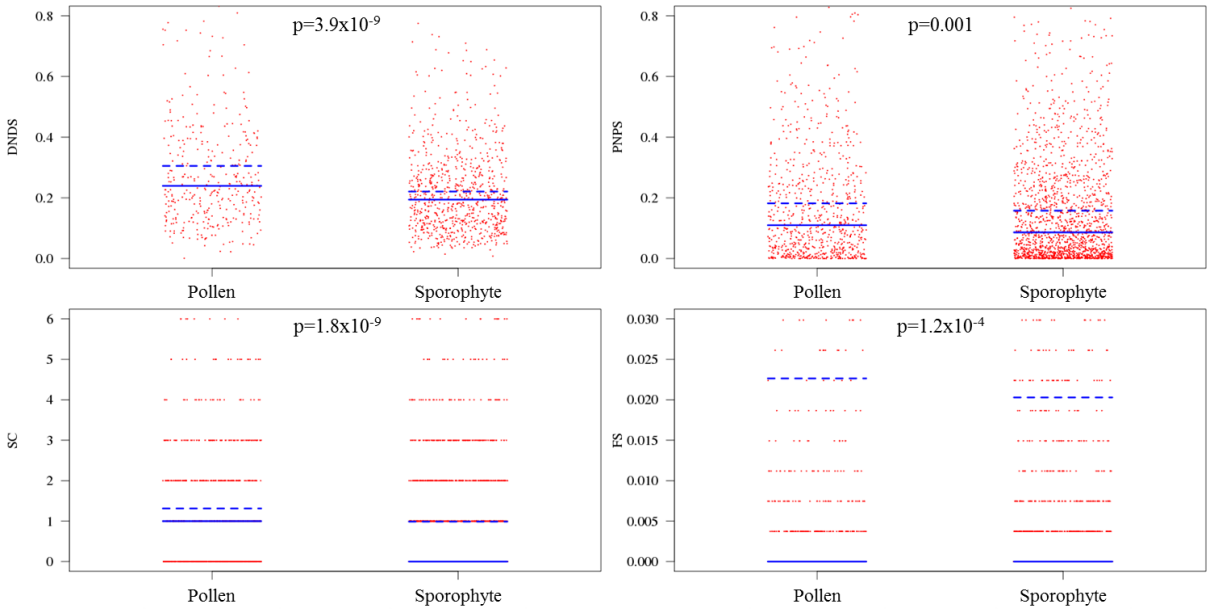

Figure 3: dN/dS (top left), pN/pS (top right), stop codon frequency (bottom left) and frameshift frequency for pollen-specific (n=790) and sporophyte-specific (n=1,690) genes with $\tau \geq 0.9$. Each data point represents a gene; solid line is the mean and dashed line the median. Differences tested with Mann-Whitney U test.

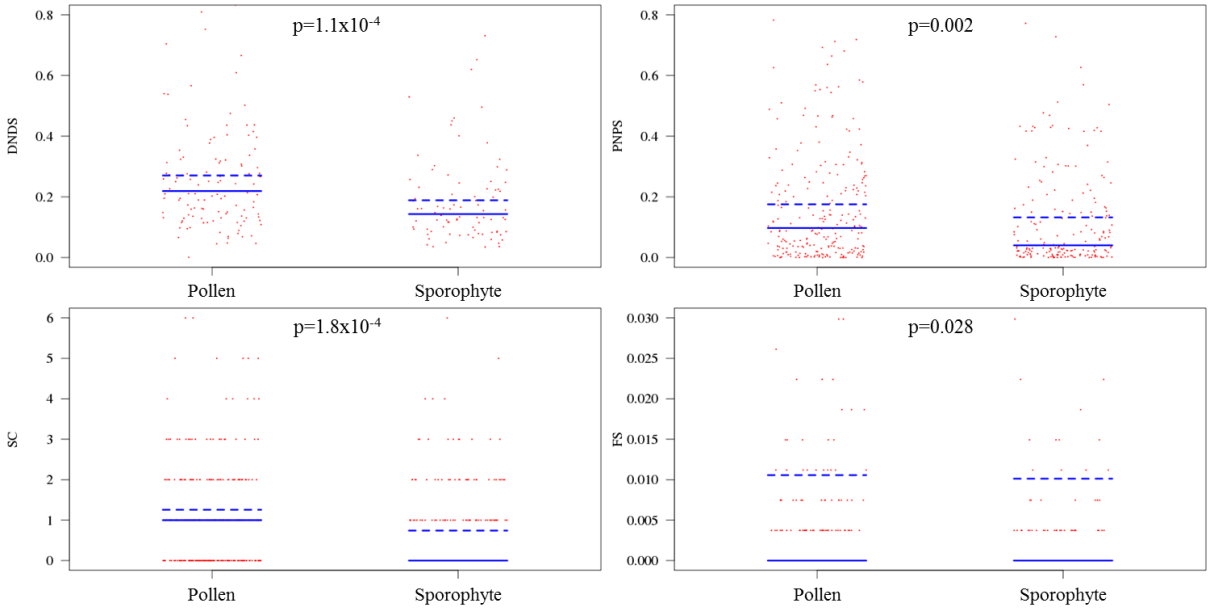

Figure 4: dN/dS (top left), pN/pS (top right), stop codon frequency (bottom left) and frameshift frequency for pollen-specific (n=253) and sporophyte-specific (n=209) genes with $\tau \geq 0.9$ and expression $>$ 1,000. Each data point represents a gene; solid line is the mean and dashed line the median. Differences tested with Mann-Whitney U test.

354 *Tables*

Table 1: Chi squared test of the distribution of pollen and sporophyte limited genes among the five nuclear *A. thaliana* chromosomes.

Chromosome	All genes	Pollen	Sporophyte
1	4,307	379	1,467
2	2,473	236	845
3	3,287	329	1,033
4	2,425	207	820
5	3,868	352	1,233
σ	16,360	1,503	5,398
	χ^2	4.613	6.925
	p	0.329	0.140

Table 2: Comparison of chromosomal positions of pollen and sporophyte genes. Mann Whitney U test.

Chromosome	W	p
1	2.63×10^5	0.108
2	9.42×10^4	0.192
3	1.63×10^4	0.272
4	8.22×10^4	0.484
5	2.26×10^5	0.244

Table 3: Correlation coefficients of expression with dN/dS, pN/pS, SC and FS within all genes (first row) and within each of the 10-quantiles of τ.

		n	dN/dS		pN/pS		SC		FS	
	Total	19,970	-0.322	***	-0.172	***	-0.119	***	-0.250	***
Tissue specificity (τ)	0 - 0.1	356	-0.277	**	-0.169	*	-0.183	**	-0.125	ns
	0.1 - 0.2	1,442	-0.252	***	-0.176	***	-0.153	***	-0.119	***
	0.2 - 0.3	1,873	-0.253	***	-0.137	***	-0.122	***	-0.074	*
	0.3 - 0.4	1,990	-0.186	***	-0.132	***	-0.095	***	-0.072	*
	0.4 - 0.5	1,602	-0.243	***	-0.075	*	-0.114	***	-0.095	**
	0.5 - 0.6	1,368	-0.174	***	-0.096	**	-0.052	ns	-0.135	***
	0.6 - 0.7	1,289	-0.188	***	-0.049	ns	-0.125	***	-0.052	ns
	0.7 - 0.8	1,549	-0.200	***	-0.101	***	-0.074	*	-0.105	***
	0.8 - 0.9	2,274	-0.159	***	-0.063	*	-0.083	***	-0.055	ns
	0.9 - 1.0	2,617	-0.119	***	-0.069	**	-0.031	ns	-0.118	***

Spearman's rank correlation. 0 *** 0.001 ** 0.01 * 0.05; ns = not significant (Bonferroni corrected)

Table 4: Correlation coefficients of τ with dN/dS, pN/pS, SC and FS within all genes (first row) and within each expression level quartile.

		n	dN/dS		pN/pS		SC		FS	
	Total	19,970	0.250	***	0.166	***	0.070	***	0.198	***
Expression Level	1st Q	4,091	0.102	***	0.085	***	-0.002	ns	0.148	***
	2nd Q	4,092	0.170	***	0.089	***	-0.010	ns	0.112	***
	3rd Q	4,088	0.206	***	0.151	***	0.060	***	0.182	***
	4th Q	4,089	0.214	***	0.176	***	0.081	***	0.167	***

Spearman's rank correlation. 0 *** 0.001 ** 0.01 * 0.05; ns = not significant (Bonferroni corrected)

Table 5: Expression data sets.

	Dataset	Description	Chips	Original source
Haploid	UNM	Uninucleate microspore	2	Honys & Twell, 2004
	BCP	Bicellular pollen	2	Honys & Twell, 2004
	TCP	Tricellular Pollen	2	Honys & Twell, 2004
	MPG	Mature Pollen	2	Honys & Twell, 2004
	GP*	Pollen Tube Grouped	6	Qin et al., 2009 ; Wang et al., 2008
	PT4*	Pollen Tube Grouped	6	Qin et al., 2009 ; Wang et al., 2008
	SPC	Sperm Cell	3	Borges et al., 2008
Diploid	SL	Silique	30	NASC
	LF	Leaves	36	NASC
	GC	Guard Cell	3	NASC
	PT**	Petiole	3	NASC
	ST	Stems	2	NASC
	HP	Hypocotyl	8	NASC
	XL	Xylem	3	NASC
	CR	Cork	3	NASC
	RT	Roots	11	NASC
	RH	Root hair elongation zone	3	NASC

NASC: Nottingham Arabidopsis Stock Centre.
* GP and PT4 were combined to one data set called PT, selecting the highest expression level of the two for each gene.
** Renamed PET