Submitted to Genome Biology and Evolution on 23-Oct-2015

Deleterious mutation accumulation in *Arabidopsis thaliana* pollen genes: a role for a recent relaxation of selection

MC Harrison, EB Mallon, D Twell, RL Hammond

Dept. Genetics

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 or rh225@le.ac.uk

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

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

Abstract

As with reproductive genes generally, we expect a faster evolution of plant pollen genes compared to sporophytic genes. Haploid expression in pollen leads to advantageous and deleterious alleles not being masked by a dominant homologue. A combination of haploid expression and pollen competition have been suggested as being responsible for stronger purifying and positive selection on pollen genes. However, it is unlikely that this mechanism is so straightforward in the model plant species *Arabidopsis thaliana*. Since becoming self-compatible roughly 1 MYA high selfing rates have caused high homozygosity levels which in turn can be expected to reduce pollen competition and the effect of masking in diploid expressed, sporophytic genes. In this study we investigated the relative strength of selection on pollen genes compared to sporophytic genes in *A. thaliana*. We were especially interested in comparing ancestral selection patterns when *A. thaliana* was an obligate outcrosser with more recent selection patterns since the plant's loss of self-incompatibility. We present two major findings: 1) before becoming self-compatible both positive and purifying selection was stronger on pollen genes than sporophytic genes for *A. thaliana*; 2) an apparent shift in selection efficacy has taken place since *A. thaliana* became self-compatible. We present evidence which indicates that selection on pollen genes has become more relaxed and has led to higher polymorphism levels and a higher build-up of deleterious mutations in pollen genes compared to sporophytic genes.

Introduction

A faster evolution of reproductive genes compared to somatic genes has been documented for a wide range of taxa, including primates, rodents, mollusks, insects and fungi (Turner and Hoekstra, 2008; Swanson and Vacquier, 2002). The faster evolution is often observable in a higher number of non-synonymous nucleotide substitutions (base changes which alter the amino acid sequence of a protein) within the coding regions of homologues. In most cases stronger positive selection is described as the mechanism driving the divergence of these genes, generally due to some form of sexual selection like cryptic female choice or sperm competition.

Two recent studies on the strength of selection on reproductive and non-reproductive genes in *Arabidopsis thaliana* also presented somewhat conflicting findings (Szövényi *et al.*, 2013; Gossmann *et al.*, 2013). Szövényi *et al.* (2013) showed that the rate of protein evolution, measured in terms of dN/dS (ratio of nonsynonymous to synonymous per site substitution rates), between *Arabidopsis thaliana* and *A. lyrata* of pollen-specific genes was significantly higher than for sporophyte-specific genes (Szövényi *et al.*, 2013). The detection of higher polymorphism levels within pollen genes was compatible with relaxed purifying selection on pollen genes, as stronger positive selection, that could have caused the higher divergence rates, would have reduced intra-specific polymorphism levels. High tissue specificity and higher expression noise compared to sporophytic genes were considered the likely causes of relaxed selection on pollen genes. As pointed out in a further study, which focused on the comparison of genes with male biased or female biased expression (Gossmann *et al.*, 2013), inter-specific divergence and currently existing intra-specific polymorphisms likely arose under different selection regimes for *A. thaliana*. The divergence of *A. thaliana* from its closest relative *A. lyrata* happened largely during a period of outcrossing, since speciation occurred approximately 13 million years ago (Beilstein *et al.*, 2010), whereas *A. thaliana* became self-compatible only roughly one million years ago (Tang *et al.*, 2007). Divergence patterns for *A. thaliana* should therefore be similar to outcrossing species and reveal stronger selection on pollen genes. Existing, intra-specific polymorphisms, on the other hand, are expected to be influenced by high selfing rates in *A. thaliana* populations that have led to high levels of homozygosity across the whole genome (Nordborg, 2000; Wright *et al.*, 2008; Platt *et al.*, 2010). The outcome is a reduction in the masking of deleterious alleles in diploid sporophyte stages (because of high homozygosity) compared to the haploid gametophyte stage. Furthermore, selfing will result in fewer genotypes competing for fertilization so lowering the magnitude of pollen competition and reducing the strength of selection acting on pollen (Charlesworth and Charlesworth, 1992).

Gossmann *et al.* (2013) found protein divergence (dN/dS) to be higher for female biased genes compared to both male genes and 476 random, non-reproductive genes sampled from the *A. thaliana* genome.

However, pollen genes did not differ from the non-reproductive genes in terms of dN/dS. Despite using a larger number of accessions to measure polymorphism than in the Szövényi *et al.* study (80 compared to 19), Gossmann *et al.* did not detect any difference in nucleotide diversity between the non-reproductive genes and pollen-specific genes in general, although nucleotide diversity was significantly lower for sperm cell-specific genes (Gossmann *et al.*, 2013). When comparing polymorphism to divergence data with a modified version of the McDonald-Kreitman test (McDonald and Kreitman 1991, Distribution of Fitness Effects Software, DoFE; Eyre-Walker and Keightley 2009) a higher proportion of non-synonymous sites were found to be under purifying and adaptive selection for pollen genes compared to both female biased and non-reproductive genes.

The aim of our study was to attempt to resolve these apparently conflicting results for *A. thaliana* and to address the following questions. Are pollen proteins really more divergent than sporophyte proteins? If so, is this due to more relaxed purifying selection or increased positive selection on pollen genes? Have patterns of selection changed for *A. thaliana* since it became self-compatible? In a first step we estimated the protein divergence of 1,552 pollen and 5,494 sporophytic genes to both *A. lyrata* and *Capsella rubella* in terms of interspecific dN/dS. This larger gene set, combined with a larger number of accessions than both previous studies (269 compared to 80 and 19), increased the power to detect sites under positive and negative selection within the two groups of genes when conducting a DoFE analysis. As the polymorphism and divergence data likely reflect periods of differing selection regimes (divergence under self incompatibility, polymorphism under self compatibility) we additionally detected sites under positive selection using a site model of the Phylogenetic Analysis by Maximum Likelihood software (PAML 4.6; Yang 2007), which does not require polymorphism data and detects sites under positive selection by allowing dN/dS to vary within genes. In a second step, to investigate more recent selection patterns, we analyzed intra-specific polymorphism levels within each group of genes. Lower diversity, measured here via non-synonymous Watterson's θ and nucleotide diversity (π), would be expected for pollen genes compared to sporophyte genes in the case of stronger selection (Nielsen, 2005). In a further test we also compared existing levels of putative deleterious alleles (premature stop codons and frameshift mutations) between pollen genes and sporophyte genes. In each of these analyses we controlled for differences in genomic factors (expression level, GC content, codon bias, gene density, gene length and average intron length) between the pollen and sporophyte-specific genes which were correlated with the divergence, polymorphism and deleterious allele measurements.

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 269 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 7), were obtained from Borg *et al.* (2011). The expression data consisted of 7 pollen and 10 sporophyte data sets (table 7). 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 ($\rho = 0.976$; $p < 2.2\text{x}10^{-16}$; 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 7).

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 less conservative cut-off values 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 between *A. thaliana*, *A. lyrata* and *Capsella rubella* using PAML (Yang, 2007) and following the method described in Szövényi *et al.* (2013). In order to detect genes that contain codon sites under positive selection, we performed a likelihood-ratio test (LRT) between models 7 (null hypothesis; dN/dS limited between 0 and 1) and 8 (alternative hypothesis; additional parameter allows dN/dS $>$ 1). An LRT statistic (twice the difference in log-likelihood between the two models) greater than 9.210 indicated a highly significant difference ($p < 0.01$; LRT > 5.991: $p < 0.05$) between the two models suggesting the existence of sites under positive selection within the tested gene (Anisimova *et al.*, 2003; Yang, 2007).

Levels of purifying and positive selection were estimated with the Distribution of Fitness Effects Software (DoFE 3.0) using the Eyre-Walker and Keightley (2009) method. Input files were constructed using custom made Perl scripts. Synonymous and non-synonymous site spectra were obtained using the Pegas package (Paradis, 2010) in R (version 3.2.0; Team 2012), first and second codon positions were considered synonymous and third codon positions non-synonymous. The total number of synonymous and non-synonymous substitutions were extracted from the PAML output files used to previously calculate dN/dS. Ten random samples of 20 alleles and 50 genes were generated for each of the two groups of genes in R using the sample() function.

Intra-specific polymorphism

Nucleotide diversity (π) and Watterson's θ were calculated for non-synonymous sites using the R package PopGenome (version 2.1.6; Pfeifer *et al.* 2014). The diversity.stats() command was implemented and the subsites option was set to "nonsyn".

Putatively deleterious alleles

To quantify the frequency of deleterious mutations for each gene, the occurrence of premature stop codons and frameshifts was calculated for each gene locus across all 268 strains compared to the reference genome. Stop codons were recorded as the number of unique alternative alleles occurring within the 269 strains as a result of a premature stop codon. Frameshifts were calculated as a proportion of the strains containing a frameshift mutation for a particular gene. All analyses of coding regions were based on the representative splice models of the *A. thaliana* genes (TAIR10 genome release, www.arabidopsis.org).

Statistical analyses

All analyses were performed in R (version 3.2.0; Team 2012). To measure statistical difference between groups we utilized the non-parametric Mann Whitney U test (wilcox.test() function), and for correlations either the Spearman rank test (rcorr() function of Hmisc package; version 3.16-0; Jr and others 2015) or Spearman rank partial correlation (pcor.test() function; ppcor package; version 1.0; Kim 2012) was carried out.

Six genomic parameters were investigated as possible predictors of dN/dS, polymorphism levels and frequency of deleterious mutations. These were expression level, GC-content, codon bias variance, gene density, average gene length and average intron length. Expression level is described above in the section "Expression data". GC content, average gene length and average intron length were calculated using custom made scripts which extracted information either from the gene sequences or the genomic gff file. RSCU (relative synonymous codon usage) was used to measure codon bias. It was calculated for each codon of each locus with the R package 'seqinr' (uco() function; version 3.1-3; Perriere 2014). As the mean value per gene varied very little between loci but varied by site within genes, we used RSCU variance as a measure for codon bias. Gene density was calculated with custom Perl and R scripts by counting the number of genes within each block of 100kb along each chromosome. Gene densities were then attributed to each gene depending on the 100kb window, in which they were situated.

As most of the genomic parameters investigated here (gene expression, GC-content, codon bias variance, gene density, average gene length and average intron length) generally differed between groups of genes (see Results), it was important to control for their possible influence on divergence, polymorphism and frequencies of deleterious mutations. The 6 parameters were also inter-correlated, so we decided to implement principle component regression analyses (pcr() command, pls package, version 2.4-3; Mevik and Wehrens 2007) in order to combine these parameters into independent predictors of the variation in the investigated dependent variable (e.g. dN/dS). All variables, including the dependent variable, were log transformed (0.0001 was added to gene length and average intron length due to zero values). A jack knife test (jack.test()) was subsequently performed on each set of principal component regression results to test if the contribution of each predictor was significant. Non-significant predictors were then removed and the analyses were repeated. The principle component (PC), which explained the highest amount of variation in the dependent variable, was then used to represent the genomic predictors in an ANCOVA (e.g. lm(log(dN/dS) $\sim$ PC1 * ploidy)) with life-stage as the binary co-variate.

Results

Life-stage limited genes

Within the total data set, containing 20,839 genes, 4,304 (20.7%) had no reliably detectable expression (score < 0.9; see methods) in any of the analysed tissues and were removed from the analysis. Of the remaining 16,535 genes, 1,552 genes (9.4%) were expressed only in pollen and a further 5,494 (33.2%) were limited to sporophytic tissues (referred to as pollen-specific genes and sporophyte-specific genes in this study). The pollen-specific and sporophyte-specific genes were randomly distributed among the five chromosomes (table 1), and their distributions within the chromosomes did not differ significantly from each other (table 2).

Expression level was roughly twice as high within pollen-specific genes (median: 1,236.1) compared to sporophyte-specific genes (median: 654.7; W = 5.5 x 10^6; p = 1.2 x 10^{-63}; Mann Whitney U test; table 3). GC-content was significantly higher within sporophyte-specific genes (median: 44.6%) than in pollen-specific genes (median: 43.8%; W = 3.4 x 10^6; p = 1.0 x 10^{-19}; Mann Whitney U test; table 3). Sporophyte-specific genes were significantly longer and contained significantly longer introns than pollen-specific genes (table 3). Gene density was slightly, albeit significantly, higher in pollen-specific genes; codon bias variance did not differ significantly (table 3).

Pollen-specific proteins evolve at a higher rate than sporophyte-specific proteins

The rate of evolution of *Arabidopsis thaliana* proteins from *Arabidopsis lyrata* homologues was estimated using interspecific dN/dS. Protein divergence was significantly higher for pollen-specific genes than sporophyte-specific genes (p = 4.3 x 10^{-24}; Mann Whitney U test; table 5, fig. 1(c)). This was mainly due to a significant difference in the non-synonymous substitution rate (dN; p = 2.4 x 10^{-27}; Mann Whitney U test; fig. 1(a)) since the synonymous substitution rate (dS) differed less strongly - median dN was 30.8% higher while median dS was only 3.7% higher in pollen-specific genes - and less significantly between groups (p = 1.6 x 10^{-4}; Mann Whitney U test; fig. 1(b)).

Both expression level (ρ = -0.232; p = 5.6 x 10^{-169}; Spearman's rank partial correlation) and GC-content (ρ = -0.145; p = 4.3 x 10^{-64}; Spearman's rank partial correlation) were significantly negatively correlated with dN/dS while controlling for other factors (codon bias variance, gene length, average intron length and gene density; table 4). Codon bias variance, gene length and average intron length also each correlated significantly and negatively with dN/dS, while gene density was weakly, positively correlated (table 4).

In order to determine how the life-stage to which the expression of a gene is limited may be contributing

to the measured difference in dN/dS, it was important to control for the six previously mentioned genomic variables (expression level, GC-content, codon bias variance, gene length, average intron length and gene density). This was important since five of the six genomic variables differed significantly between pollen and sporophyte-specific genes (table 3) and all six were significantly correlated to dN/dS (table 4). A principal component regression was conducted to allow us to condense these several predictors of dN/dS into independent variables. We first included all 6 predictors in the principal component regression model, and they explained 9.10% of dN/dS variation. Principal component (PC) 2 explained the largest amount of variation at 6.15%. A jack knife test on this PC revealed significant p-values (< 0.05) only for expression, GC content and codon bias variance. After removal of the non-significant predictors (gene length, average intron length and gene density) codon bias variation was also no longer significant. The first PC of a model containing expression and GC content as the predictors of dN/dS had an explanation value of 7.15% (total 7.24%). This first PC was used as the continuous variable in an ANCOVA with dN/dS as the dependent variable and life-stage as the binary co-variable. Life-stage specificity had a significant influence on dN/dS: Pollen-specific genes had significantly higher dN/dS values when controlling for expression level and GC content with PC1 (fig. 2).

Pollen-specific genes contain a higher number of sites under positive and purifying selection

We investigated whether the higher divergence of pollen-specific proteins compared to sporophyte-specific proteins was restricted to *Arabidopis*, and possibly fueled by selection in either *A. thaliana* or *A. lyrata*, by investigating the protein divergence of both from *Capsella rubella*. Divergence was significantly higher for pollen-specific proteins in all three comparisons (table 5). Between branches only one comparison of divergence values differed significantly for sporophyte-specific proteins: *A. thaliana*-*A. lyrata* dN/dS > *A. lyrata*-*C. rubella* dN/dS (Bonferroni corrected p-value: 0.046); all other differences between branches were non-significant.

A higher dN/dS value, which is still lower than 1, generally indicates weaker purifying selection (Yang and Bielawski, 2000). Only 41 out of 13,518 genes had a dN/dS value greater than 1 and 65.1% of genes had a dN/dS less than 0.2. However, gene-wide estimates of dN/dS can be inflated by a few codon sites under positive selection (> 1) even if purifying selection is otherwise prevalent. In order to test this possibility we quantified levels of positive selection in both pollen- and sporophyte-specific genes.

Based on protein divergence (dN/dS between *A. thaliana* and *A. lyrata*) and polymorphism data (synonymous and non-synonymous substitutions within the 269 strains) we analysed the distribution of fitness effects within pollen and sporophyte-specific genes using the Distribution of Fitness Effects

Software (DoFE 3.0; Eyre-Walker and Keightley 2009). The analyses were repeated 10 times on random samples of 20 alleles and 50 genes from each group. The distribution of new deleterious mutations showed that a significantly larger fraction of non-synonymous mutations were deleterious ($N_e s > 1$; N_e: effective population size, s: selection coefficient) within pollen-specific genes (mean 0.725 ±0.001; sporophyte-specific genes: 0.654 ±0.001; $p < 2.2 \times 10^{-16}$; Mann Whitney U test). This indicates that a higher proportion of sites have evolved under purifying selection within the group of pollen-specific genes between *A. thaliana* and *A. lyrata* (Eyre-Walker and Keightley, 2009). Also, a significantly positive alpha (proportion of sites under positive selection) was found for 6 out of 10 samples from the pollen-specific genes (mean of six samples: 0.186 ±0.002) but for none of the sporophyte specific gene samples.

For the calculation of dN/dS on the multi-sequence alignment (*A. thaliana*, *A. lyrata* and *C. rubella*) described above we allowed dN/dS to vary among sites in order to detect sites under positive selection using PAML (Yang, 2007). This analysis confirmed the DoFE results: 15.6% of pollen-specific (p ¡ 0.05; 6.7 % at $p < 0.01$) compared to only 9.5% of sporophyte-specific genes (p ¡ 0.05; 5.0% at $p < 0.01$) contained codons under positive selection.

Pollen-specific genes are more polymorphic than sporophyte-specific genes

Pollen-specific genes were more polymorphic than sporophyte-specific genes with both non-synonymous nucleotide diversity (π_n) and non-synonymous Watterson's theta (θ_n) significantly higher in pollen-specific genes (fig. 3). Each of the six correlates of dN/dS listed above also correlated significantly with π_n and θ_n (all negatively except gene length; table 4). The six variables explained 25.98% of variation in π_n in a principal component regression. The first PC contributed most (21.13%). The same six factors explained a total of 40.53% of the variation in θ_n (first PC: 31.09%). The contribution of all six predictors was highly significant for both models. For each model the first PC was implemented in an ANCOVA testing the influence of life-stage as a co-variate. π_n and θ_n remained significantly higher for pollen-specific genes (fig. 4).

Higher frequency of deleterious mutations in pollen-specific genes

Higher polymorphism levels may indicate a recent relaxation of purifying selection on pollen-specific genes. To test this hypothesis further we investigated the frequency of putatively deleterious mutations - premature stop codons and frameshift mutations - within the 269 *A. thaliana* strains. Stop codon frequency, defined here as the relative number of unique alternative alleles due to premature stop codons occurring within the 269 strains, was significantly higher within pollen-specific genes (mean: 0.063 ±0.004; sporophyte mean: 0.049 ±0.002; $W = 2.7 \times 10^6$; $p = 4.1 \times 10^{-15}$; Mann Whitney U test; fig. 5(a)). The

frequency of strains containing at least one frameshift mutation was also significantly higher for pollen-specific genes (mean: 0.021 ±0.002) compared to sporophyte-specific genes (mean: 0.014 ±0.001; W = 2.7 x 10^6; p = 6.6 x 10^{-22}; Mann Whitney U test; fig. 5(b)). Significant correlations existed between these measures of deleterious mutations and the six correlates of dN/dS (table 4).

In a principal component regression analysis all six predictors (expression level, codon bias variance, GC content, gene length, average intron length and gene density) were significantly correlated with stop codon frequency. The six predictors explained a total of 20.04% of the variation in stop codon frequency, 17.42% explained by the first PC. An ANCOVA with life-stage as the binary co-variant showed that premature stop codons remained significantly more frequent within pollen-specific genes even when controlling for the six predictors via PC1 (fig. 6(a)). Four of the predictors (expression level, GC content, gene length and gene density) were also significantly correlated with the frequency of frameshift mutations. However, the four variables only explained a total of 5.49% of variation (first PC 5.08%). Again, in an ANCOVA analysis frameshift mutations remained significantly more frequent within pollen-specific genes when controlling for the predictors via the first PC (fig. 6(b)).

Tissue specific genes

Tissue specificity has been shown to be negatively correlated with selection efficiency (Duret and Mouchiroud, 2000; Liao *et al.*, 2006; Slotte *et al.*, 2011). The on average greater tissue specificity in pollen-specific genes compared to sporophyte specific genes could therefore explain the higher polymorphism levels and higher frequency of deleterious mutations found in pollen-specific genes. In order to control for this potential bias we compared polymorphism levels and the frequency of deleterious alleles in pollen-specific genes with a group of 340 genes whose expression is limited to a single sporophyte cell type (guard cell, xylem or root hair: referred to as 'tissue-specific sporophyte genes').

In this tissue-specificity controlled comparison, π_n and θ_n did not differ between pollen-specific genes and the tissue-specific sporophyte gene set. However, both π_n and θ_n were significantly higher in pollen-specific genes (p = 1.0 x 10^{-16} & 6.1 x 10^{-30}) and tissue specific sporophyte genes (p = 1.0 x 10^{-7} & 4.9 x 10^{-9}; Mann Whitney U test; fig. 7) than broadly expressed sporophyte-specific genes.

In terms of putatively deleterious changes, controlling the bias in tissue specificity reduced, in comparison to uncontrolled comparisons, the difference between pollen-specific and tissue-specificity controlled sporophyte genes for premature stop codons and frameshift mutations. Premature stop codons remained significantly more frequent in pollen-specific genes than in sporophytic, tissue specific genes (p = 0.018; fig. 8), although with a much reduced significance, whereas there was no significant difference in the frequency of frameshift mutations (fig. 8). The reduced significance (premature stop codons) or non-significance (frameshift mutations) compared to uncontrolled comparisons suggests that tissue specificity

is important. This was confirmed by there being significant differences between tissue-specific sporophyte genes and broadly expressed sporophyte genes for both estimates of deleteriousness (frameshift mutations: p = 4.7 x 10^{-15}; premature stop codons: p = 0.018; Mann Whitney U test; fig. 8).

Expression, GC content and codon bias variance differed significantly between pollen and tissue specific sporophyte genes (table 6) so we carried out a principal component regression and subsequent ANCOVA. In a first principal component regression all 6 correlates had a significant effect on the variation of π_n and θ_n, for which 25.8% and 29.6% of variation, respectively, was explained by the first PC. Pollen-specific genes were significantly more polymorphic (θ_n) than tissue-specific sporophyte genes when the first PC was controlled for (fig. 9(b)). The frequency of deleterious mutations was also significantly higher in pollen-specific genes than tissue specific, sporophytic genes, when controlling for the main predictors of stop codon mutations and frameshift mutations (fig. 9(c) & 9(d)).

Discussion

Our analysis showed that expression level, GC content, codon bias variance, gene length, average intron length and gene density correlated significantly with protein divergence, polymorphism levels and the frequency of deleterious mutations. All of these factors except codon bias variance differed significantly between pollen-specific and sporophyte-specific genes. Protein divergence, polymorphism levels and the frequency of deleterious mutations were significantly higher within pollen-specific genes compared to sporophyte-specific genes even when controlling for the six variables.

Divergence data suggest stronger positive and purifying selection on pollen-specific genes

Pollen-specific genes had significantly higher dN/dS between *A. thaliana* and *A. lyrata* than sporophyte-specific genes, a pattern that remained significant when differences in expression level and GC content were controlled for. Similar patterns in dN/dS were also detected between both *Arabidopsis* species and *Capsella rubella*. These higher dN/dS values were most likely explained by stronger positive selection acting on pollen-specific genes than on sporophyte specific genes. An analysis of the distribution of fitness effects (DoFE 3.0; Eyre-Walker and Keightley 2009) detected a significant α (proportion of advantageous non-synonymous substitutions) in 6 out of 10 random samples of the pollen-specific genes but in none of the sporophyte-specific samples. This was corroborated by a further analysis using PAML (Yang, 2007), in which a higher proportion of pollen-specific genes were found to contain codon sites under positive selection. Purifying selection, which dominated across genes (over 65% of genes with dN/dS < 0.2), was also higher on pollen-specific genes.

Overall, selection (positive and purifying) appears to have been stronger on pollen-specific genes than

sporophyte specific genes for *A. thaliana* during its divergence from *A. lyrata*. This confirms results found for *Capsella grandiflora*, in which pollen-specific genes also showed higher positive and purifying selection rates (Arunkumar *et al.*, 2013). It also agrees with findings for several taxa, in which reproductive genes evolve more quickly due to increased positive selection (Turner and Hoekstra, 2008; Swanson and Vacquier, 2002). Furthermore, our findings may help resolve recent, competing findings for *A. thaliana*. Higher dN/dS values for pollen-specific genes have also been found in *A. thaliana* but were interpreted as being caused by relaxed purifying selection (Szövényi *et al.*, 2013). However, signs of positive selection were not specifically investigated in that study and their interpretation of relaxed selection was based on polymorphism patterns (see below). In contrast, another study found no difference in dN/dS between pollen and sporophyte-specific genes, but, similar to our findings, higher levels of purifying and positive selection were detected for pollen-specific genes (Gossmann *et al.*, 2013).

Polymorphism levels suggest relaxed selection on pollen-specific genes

Stronger purifying and positive selection are expected to lead to lower intra-specific polymorphism levels. However, we found Watterson's θ and π of non-synonymous sites were significantly higher within pollen-specific genes, even when controlling for expression and five further genomic differences (GC content, codon bias variance, gene length, average intron length and gene density). In one of two recent studies higher polymorphism rates were also found in pollen-specific genes for *A. thaliana* (Szövényi *et al.*, 2013). In the second study, however, no difference was found between pollen-specific genes in general and random, non-reproductive genes in terms of nucleotide diversity; however, within genes specific to sperm cells nucleotide diversity was actually reduced (Gossmann *et al.*, 2013). It is possible that our larger sporophytic gene set (5,494 genes compared to 476) within a larger number of accessions (269 rather than 80) offered more power to detect differences in nucleotide diversity.

We also found significantly higher levels of putatively deleterious alleles (premature stop codons and frameshift mutations) within pollen-specific genes. This supports the conclusions of Szövényi *et al.* (2013) that the raised polymorphism levels indicate relaxed purifying selection on pollen-specific genes. In other words, comparatively weaker selective constraints are allowing deleterious alleles to accumulate at a greater rate within pollen-specific genes compared to those whose expression is restricted to the sporophyte. This is in stark contrast to the selection patterns we detected in the inter-specific dN/dS and indicates a possible recent relaxation of selection on pollen-specific genes for *A. thaliana*.

Has there been a recent shift in selection strength?

The patterns in our data are compatible with a change in selection efficacy that are likely to have taken place since the speciation of *A. thaliana* and *A. lyrata*. The relatively recent switch from self-incompatibility to self-compatability in *A. thaliana* (ca. 1MYA; Tang *et al.* 2007) explains why we have observed evidence for relaxed selection in polymorphism levels but stronger selection in divergence data for pollen-specific genes. The divergence data used to calculate dN/dS mainly represent a prolonged period of outcrossing ($\sim$ 12 MYA), since the speciation of *A. thaliana* from *A. lyrata* occurred roughly 13 million years ago (Beilstein *et al.*, 2010). In contrast, the polymorphism data and frequencies of putative deleterious alleles reflect the recent selective effects of high selfing rates.

The evidence we have found for a more recent weaker selection on pollen-specific genes contrasts with findings for the outcrossing *Capsella grandiflora* (Arunkumar *et al.*, 2013). In that study 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 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 likelihood that deleterious alleles are masked in heterozygous state when expressed in a diploid tissue (Platt *et al.*, 2010). A reduction in pollen competition can also 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 on pollen-specific genes than sporophyte-specific genes rather than similar?

We have shown here that tissue specificity partly explains why selection is more relaxed on pollen genes. The full set of sporophyte-specific genes contains genes expressed across several tissues, and broadly expressed genes have been known to be under more efficient selection than tissue-specific genes due to their exposure to a higher number of selective constraints (Duret and Mouchiroud, 2000; Liao *et al.*, 2006; Slotte *et al.*, 2011). Both pollen-specific genes and genes limited to one of three sporophytic tissues (xylem, guard cell or root hair) showed raised levels of polymorphism and frequency of deleterious mutations compared to broadly expressed sporophyte-specific genes (expressed in at least 5 tissues). Tissue specificity appeared to explain, to a certain extent, the reduced selection efficacy in

pollen-specific genes as there was no longer a significant difference in polymorphism levels (θ_n and π_n) or the frequency of frameshift mutations in pollen-specific genes compared to the tissue specific, sporophytic genes (the frequency of stop codon mutations remained significantly higher). However, tissue specificity alone only partly explains the apparent, current more relaxed selection on pollen-specific genes. Once further genomic features (expression level, GC content, codon bias variance, gene length, average intron length, gene density) were controlled for, all measures remained higher in pollen-specific genes even when compared to genes restricted to only one sporophytic tissue (significant for all except π_n).

Recent similar findings indicating relaxed purifying selection in pollen specific genes in *A. thaliana* (Szövényi *et al.*, 2013) were explained as possibly resulting from 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 to isolate the effect of tissue specificity. We have shown here that tissue specificity does appear to play a role but does not alone explain the difference in selection strength between both groups of genes. Higher expression noise could then 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 and high tissue specificity become dominant factors for pollen-specific genes of selfing plants. The loss of self-incompatibility in *A. thaliana* may therefore have led to a reduction in selection efficacy and the accumulation of deleterious alleles in pollen-specific genes.

Conclusion

We have shown that, as in many other taxa, genes expressed in male reproductive tissues evolve at a quicker rate than somatic genes in *A. thaliana*. The greater divergence of pollen proteins to both *A. lyrata* and *C. rubella* compared to sporophytic genes can be attributed to stronger positive and purifying selection. However, intra-specific polymorphism data indicate a strong shift in this selection pattern may have occurred. Since the more recent loss of incompatibility in *A. thaliana* selection appears to have become more relaxed in pollen-specific genes. This is likely due to a reduction in pollen competition and the masking of diploid, sporophytic genes as a result of high homozygosity levels. In outcrossing plants, haploid expression and pollen competition 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 has likely reduced the counteracting effects of pollen competition and haploid expression, leading to lower selection efficacy and an increased accumulation of deleterious mutations in pollen-specific compared to sporophyte-specific genes.

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

Anisimova, M., Nielsen, R., and Yang, Z. 2003. Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. *Genetics*, 164(3): 1229–1236.

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.

Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R., and Mathews, S. 2010. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, 107(43): 18724–18728.

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.

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.

Eyre-Walker, A. and Keightley, P. D. 2009. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution*, 26(9): 2097–2108.

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.

Jr, F. E. H. and others, w. c. f. C. D. a. m. 2015. Hmisc: Harrell Miscellaneous.

Kim, S. 2012. ppcor: Partial and Semi-partial (Part) correlation.

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.

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.

McDonald, J. H. and Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328): 652–654.

458 Mevik, B.-h. and Wehrens, R. 2007. The pls Package: Principal Component and Partial Least Squares
459 Regression in R. *Journal of Statistical Software*, pages 1–24.

460 Nielsen, R. 2005. Molecular Signatures of Natural Selection. *Annual Review of Genetics*, 39(1): 197–218.

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

463 Paradis, E. 2010. pegas: an R package for population genetics with an integrated–modular approach.
464 *Bioinformatics*, 26(3): 419–420.

465 Perriere, D. C. a. J. R. L. a. A. N. a. L. P. a. S. P. a. G. 2014. seqinr: Biological Sequences Retrieval and
466 Analysis.

467 Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., and Lercher, M. J. 2014. PopGenome: An Efficient
468 Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution*, 31(7):
469 1929–1936.

470 Platt, A., Horton, M., Huang, Y. S., Li, Y., Anastasio, A. E., Mulyati, N. W., Ågren, J., Bossdorf, O.,
471 Byers, D., Donohue, K., Dunning, M., Holub, E. B., Hudson, A., Le Corre, V., Loudet, O., Roux, F.,
472 Warthmann, N., Weigel, D., Rivero, L., Scholl, R., Nordborg, M., Bergelson, J., and Borevitz, J. O.
473 2010. The Scale of Population Structure in Arabidopsis thaliana. *PLoS Genet*, 6(2): e1000843.

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

477 Slotte, T., Bataillon, T., Hansen, T. T., St. Onge, K., Wright, S. I., and Schierup, M. H. 2011. Genomic
478 Determinants of Protein Evolution and Polymorphism in Arabidopsis. *Genome Biology and Evolution*,
479 3: 1210–1219.

480 Swanson, W. J. and Vacquier, V. D. 2002. Reproductive Protein Evolution. *Annual Review of Ecology*
481 *and Systematics*, 33: 161–179. ArticleType: research-article / Full publication date: 2002 / Copyright
482 © 2002 Annual Reviews.

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

486 Tang, C., Toomajian, C., Sherman-Broyles, S., Plagnol, V., Guo, Y.-L., Hu, T. T., Clark, R. M., Nas-
487 rallah, J. B., Weigel, D., and Nordborg, M. 2007. The Evolution of Selfing in Arabidopsis thaliana.
488 *Science*, 317(5841): 1070–1072.

489 Team, R. C. 2012. R: A language and environment for statistical computing.

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

492 Wang, Z. and Zhang, J. 2011. Impact of gene expression noise on organismal fitness and the efficacy of
493 natural selection. *Proceedings of the National Academy of Sciences of the United States of America*,
494 108(16): E67–E76.

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

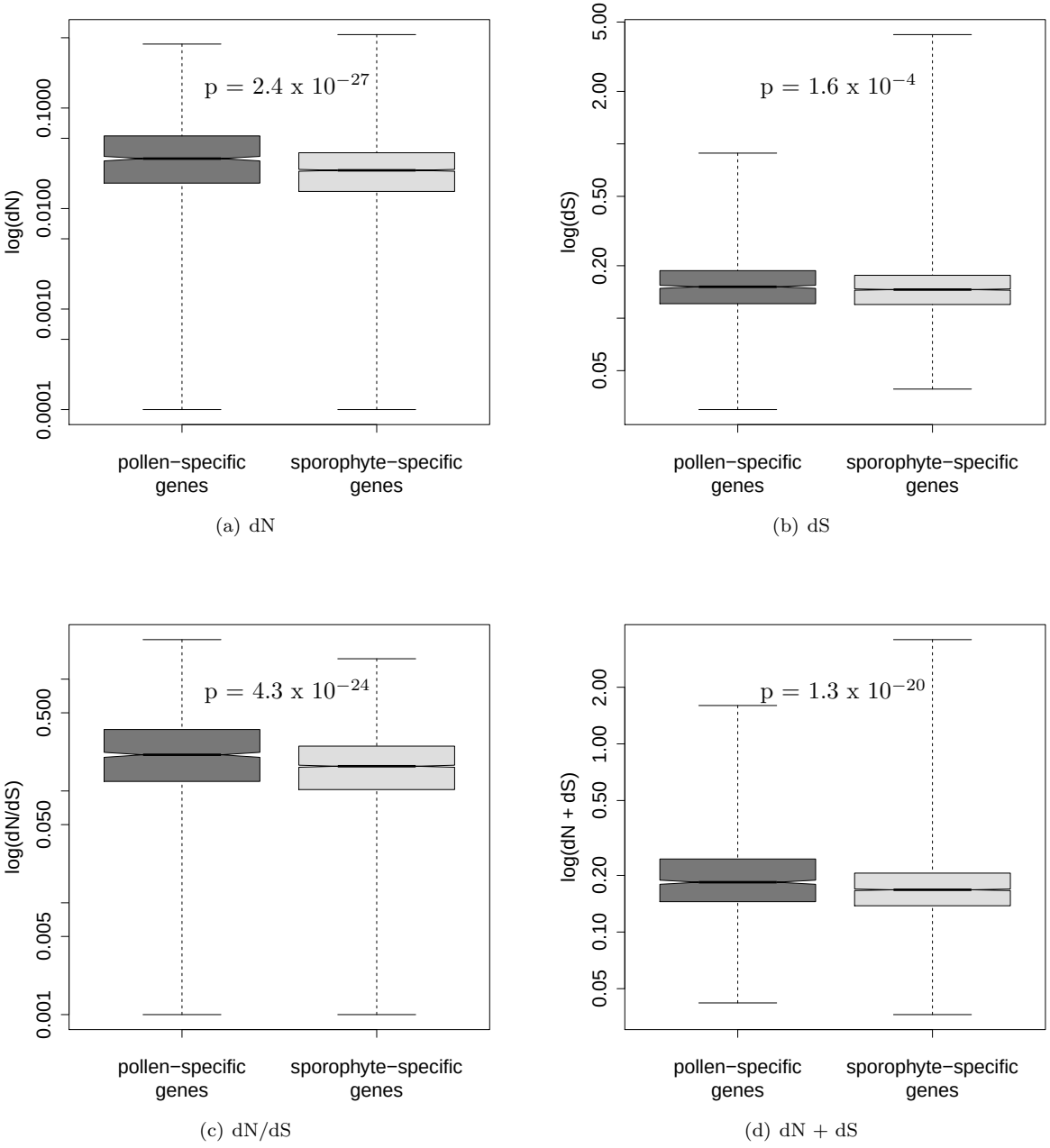
497 Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolu-*
498 *tion*, 24(8): 1586–1591.

499 Yang, Z. and Bielawski, J. P. 2000. Statistical methods for detecting molecular adaptation. *Trends in*
500 *Ecology & Evolution*, 15(12): 496–503.

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

504 *Figures*



(a) dN (b) dS

(c) dN/dS (d) dN + dS

Figure 1: Non-synonymous (dN; a), synonymous (dS; b), dN/dS (c) and total nucleotide substitution rate (dN + dS; d) within pollen-specific and sporophyte-specific genes. Significance tested with Mann Whitney U test.

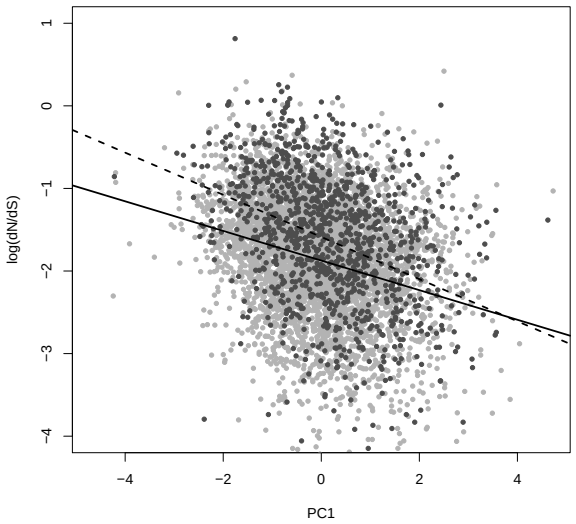


Figure 2: ANCOVA analysis with PC1 (expression and GC content) as continuous variable reveals higher dN/dS among pollen-specific (dark grey points and dashed line) than sporophyte-specific genes (light grey points and solid line).

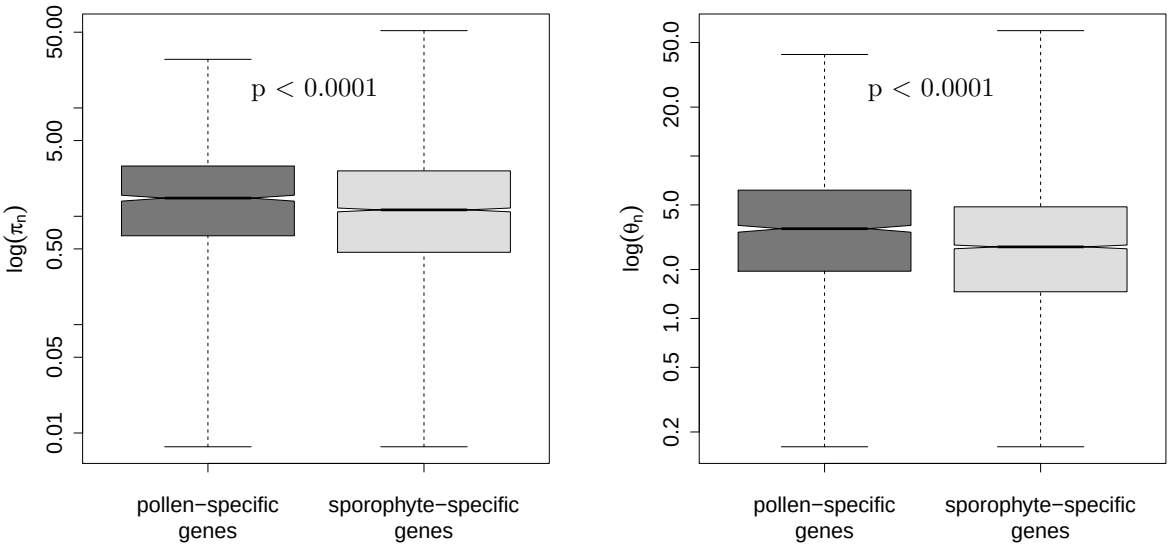
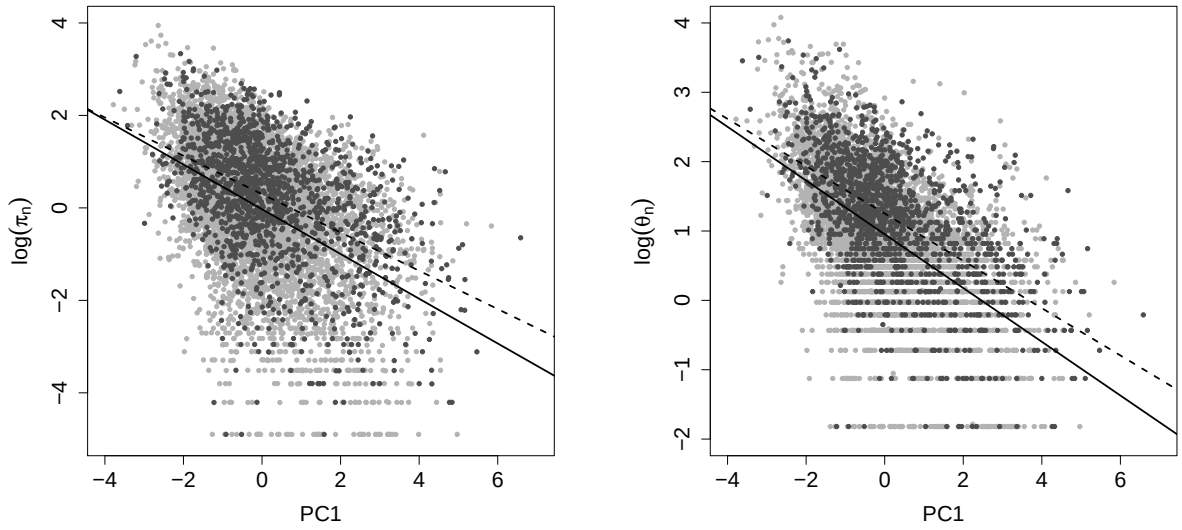
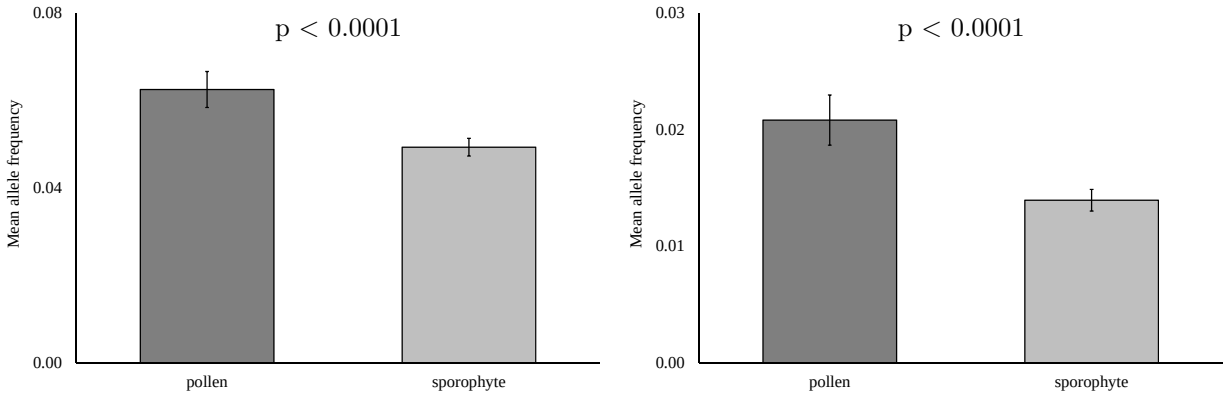


Figure 3: Non-synonymous nucleotide diversity (a) and non-synonymous Watterson's theta (b) within pollen-specific and sporophyte-specific genes. Significance tested with Mann Whitney U test.



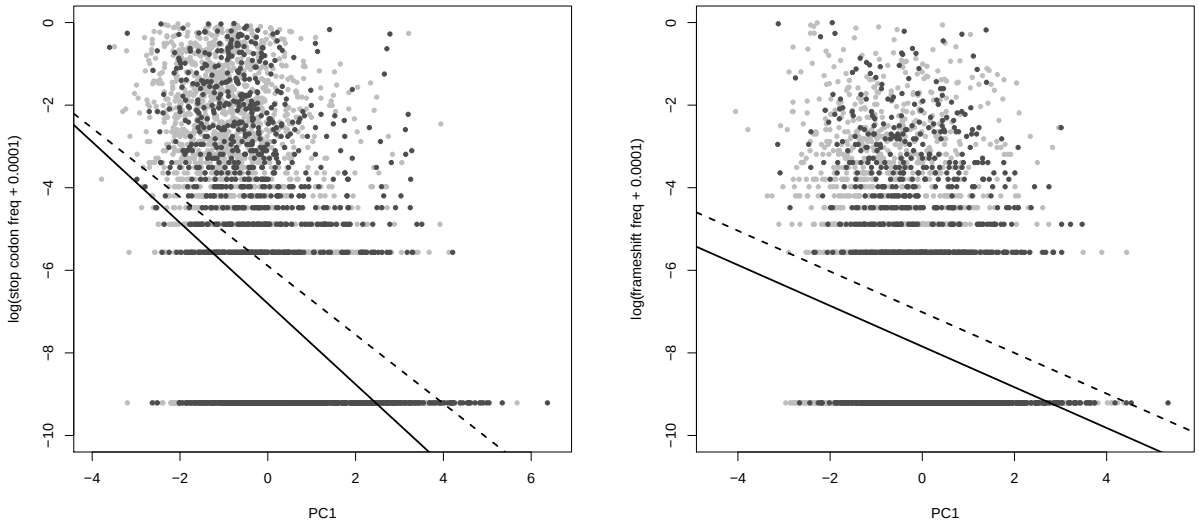
(a) (b)

Figure 4: ANCOVA analysis with PC1 (6 genomic variables) as continuous variable reveals both higher π_n (a) and higher θ_n (b) among pollen-specific (dark grey points and dashed line) than sporophyte-specific genes (light grey points and solid line).



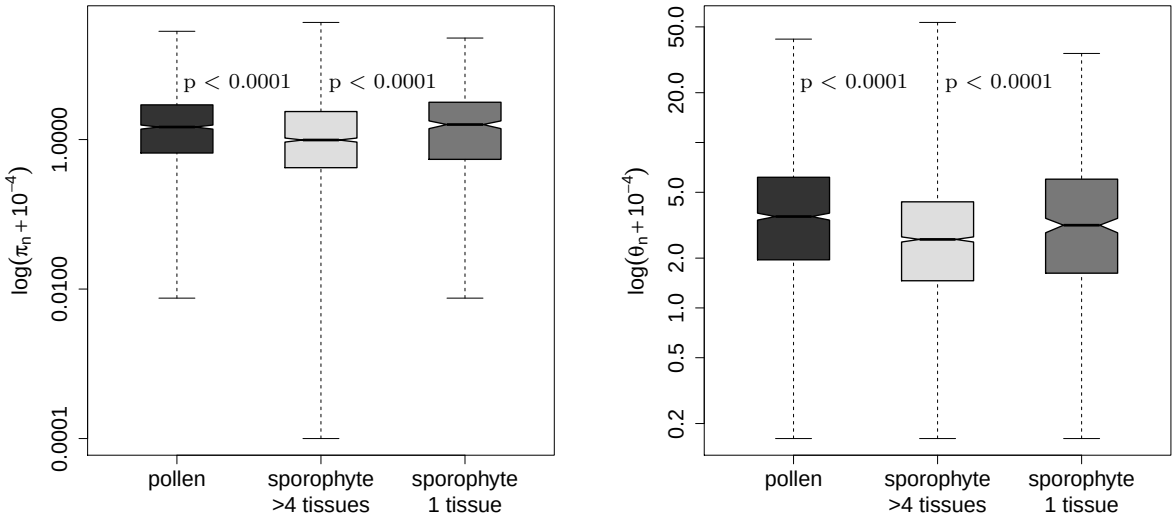
(a) (b)

Figure 5: Frequency of alleles containing premature stop codon mutations (a) and frameshift mutations (b) in pollen-specific and sporophyte-specific genes. Significance tested with Mann Whitney U test.



(a) (b)

Figure 6: ANCOVA analysis with PC1 (6 genomic variables) as continuous variable reveals significantly higher frequency of stop codon mutations (a) and frameshift mutations (b) among pollen-specific (dark grey points and dashed line) than sporophyte-specific genes (light grey points and solid line).



(a) (b)

Figure 7: Non-synonymous nucleotide diversity (a) and non-synonymous Watterson's theta (b) within genes specific to guard cells, xylem or root hair, sporophyte-specific and pollen-specific genes.

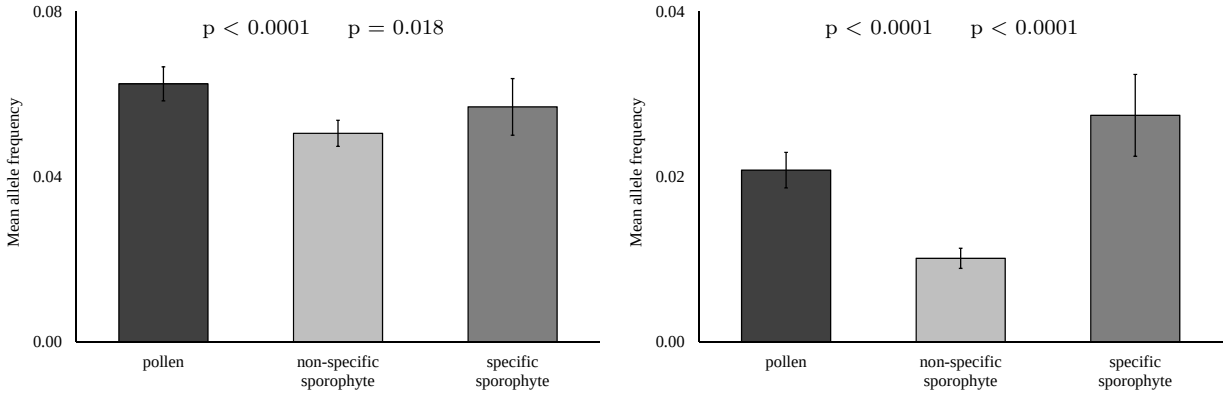


Figure 8: Mean frequency of alleles containing a premature stop codon (a) or frameshift mutation (b) among pollen-specific genes, broadly expressed sporophytic genes (at least 5 tissues) and tissue specific genes (expression restricted to guard cell, xylem or root hair tissues).

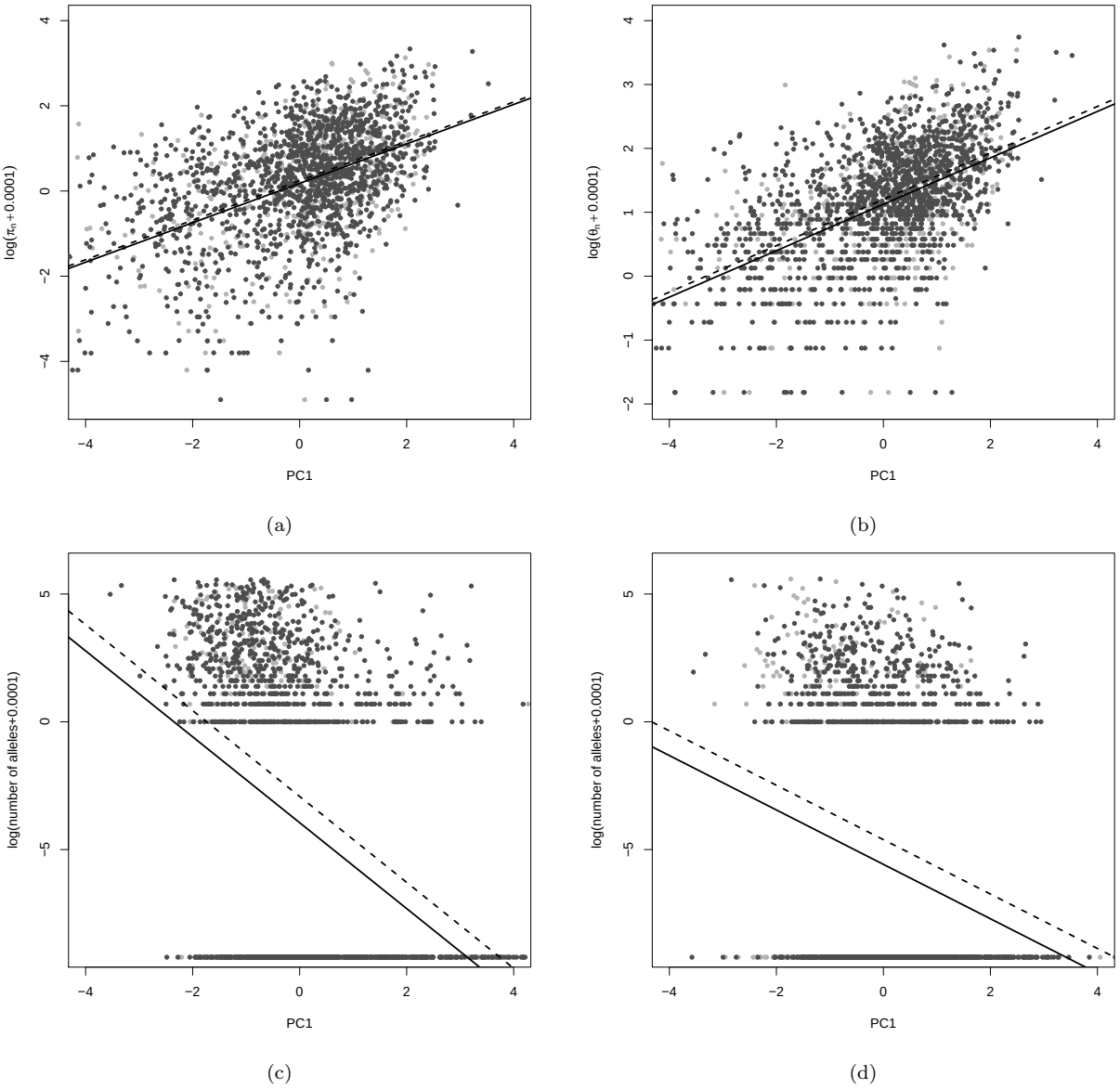


Figure 9: ANCOVAS comparing pollen-limited genes to tissue-specific, sporophytic genes while controlling for the first PC of a PCR. Significantly higher values were measured among pollen-limited genes for (b) θ_n, (c) frequency of premature stop codon mutations and (c) frequency of frameshift mutations; and higher but non-significant for (a) π_n .

505 *Tables*

Table 1: Chi squared test of the distribution of pollen and sporophyte limited genes among the five nuclear *A. thaliana* chromosomes. Degrees of freedom: 4.

Chromosome	All genes	Pollen	Sporophyte
1	4,348	392	1,495
2	2,522	251	862
3	3,326	340	1,049
4	2,451	214	839
5	3,888	355	1,249
Σ	16,535	1,552	5,494
	χ^2	5.367	7.456
	p	0.252	0.136

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

Chromosome	W	p
1	2.79×10^5	0.137
2	1.00×10^5	0.071
3	1.72×10^5	0.315
4	8.54×10^4	0.267
5	2.31×10^5	0.241

Table 3: Differences in 6 genomic variables between pollen-specific and sporophyte-specific genes. Values are means ± standard error of the mean; significance was tested with Mann Whitney U test; p-values are Bonferroni corrected for multiple testing.

Genomic variable	Pollen-specific genes			Sporophyte-specific genes		p
Expression level	2,562.30	± 86.49	>	1,256.21	± 23.80	1.2×10^{-63}
GC content (%)	44.20	± 0.08	<	45.08	± 0.04	1.0×10^{-19}
Codon bias variance	0.46	± 0.01	=	0.43	± 0.00	not significant
gene length	1,570.30	± 24.41	<	1,634.39	± 11.62	2.3×10^{-4}
average intron length	124.44	± 3.23	<	160.08	± 2.49	8.6×10^{-10}
gene density (per 100kb)	29.99	± 0.12	>	29.57	± 0.07	1.5×10^{-3}

Table 4: Partial correlations of 6 genomic variables with dN/dS, θ_n, π_n, frequency of premature stop codons and frameshift mutations. Spearman rank correlations controlling for remaining 5 variables; p-values are Bonferroni corrected for multiple testing.

	dN/dS		θ_n		π_n		stop codons		frameshifts	
Expression level	-0.232	***	-0.131	***	-0.086	***	not significant		-0.090	***
GC content (%)	-0.145	***	-0.192	***	-0.166	***	-0.180	***	-0.143	***
Codon bias variance	-0.104	***	-0.210	***	-0.161	***	-0.124	***	-0.088	***
gene length	-0.108	***	0.325	***	0.181	***	0.136	***	-0.037	*
average intron length	-0.061	***	-0.191	***	-0.123	***	0.084	***	-0.109	***
gene density (per 100kb)	0.039	*	-0.137	***	-0.116	***	-0.054	***	-0.029	*

*p<0.01; **p<10^{-6}; ***p<10^{-9}

Table 5: dN/dS between *A. thaliana*, *A. lyrata* and *C. rubella*. Values are means (and medians); significance was tested with Mann Whitney U test; p-values are Bonferroni corrected for multiple testing.

	Pollen	Sporophyte	p value
A. thaliana vs. *A. lyrata*	0.2689 (0.2106)	0.1963 (0.1664)	4.3 x 10^{-24}
A. thaliana vs. *C. rubella*	0.2409 (0.2036)	0.1801 (0.1567)	8.8 x 10^{-22}
A. lyrata vs. *C. rubella*	0.2370 (0.1945)	0.1818 (0.1568)	1.3 x 10^{-15}

Table 6: Differences in 6 genomic variables between pollen-specific genes and genes limited to one of three sporophytic tissues. Values are means ± standard error of the mean; significance was tested with Mann Whitney U test; p-values are Bonferroni corrected for multiple testing.

	Pollen-specific genes			guard cell, xylem or root hair		p
Expression level	2,562.30	± 86.49	>	446.24	± 26.82	1.0 x 10^{-77}
GC content (%)	44.20	± 0.08	<	44.80	± 0.17	4.5 x 10^{-3}
Codon bias variance	0.46	± 0.01	>	0.39	± 0.01	2.2 x 10^{-6}
gene length	1,570.30	± 24.41	=	1,561.71	± 36.20	not significant
average intron length	124.44	± 3.23	=	152.49	± 9.14	not significant
gene density (per 100kb)	29.99	± 0.12	=	29.48	± 0.30	not significant

Table 7: 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