The Evolutionary Genetics of Gene Expression in *Capsella grandiflora*

by

Emily Beth Josephs

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Ecology and Evolutionary Biology
University of Toronto

Abstract

The Evolutionary Genetics of Gene Expression in *Capsella grandiflora*

Emily Beth Josephs
Doctor of Philosophy
Graduate Department of Ecology and Evolutionary Biology
University of Toronto
2015

This is my abstract

To my grandmothers, Myra Josephs and Mary Barnard.

Acknowledgements

Here are my Acknowledgements. Thanks everybody!

Contents

1 General introduction **1**

2 Population genomics of *Capsella grandiflora* **2**

2.1 Abstract . . . 2
2.2 Introduction . . . 2
2.3 Results . . . 4
2.3.1 Genome-wide patterns of polymorphism . . . 4
2.3.2 Genome-wide measures of purifying selection . . . 5
2.3.3 Genome-wide estimates of positive selection . . . 7
2.3.4 Effects of expression and selection . . . 10
2.4 Discussion . . . 11
2.4.1 Measuring positive selection . . . 13
2.4.2 Expression level and selection . . . 13
2.5 Methods . . . 14
2.5.1 Sampling and sequencing . . . 14
2.5.2 Genotyping . . . 14
2.5.3 Divergence . . . 15
2.5.4 Identifying conserved noncoding sequences . . . 16
2.5.5 Estimates of the distribution of fitness effects and α . . . 16
2.5.6 Test for signatures of recurrent selective sweeps . . . 16
2.5.7 Gene expression . . . 17
2.6 Acknowledgements . . . 17
2.7 Appendix: Supplementary figures and tables . . . 17

3 Mutation-selection balance maintains gene expression variation **30**

3.1 Abstract . . . 30
3.2 Introduction . . . 30
3.3 Results and Discussion . . . 31
3.4 Materials and Methods . . . 37
3.4.1 Study system and plant material . . . 37
3.4.2 Genomic data . . . 37
3.4.3 Mapping local eQTL . . . 38
3.4.4 Mapping aseQTL . . . 39
3.4.5 Permutation analysis . . . 40

3.5 Acknowledgements 41
3.6 Appendix: Supplementary figures and tables 41

Bibliography **53**

List of Tables

2.1 **Sampling locations of each individual.** Note that individual AxE is a cross between 918/8 and Cg2e 18
2.2 **DFE-alpha model outputs for each site category** 18

List of Figures

2.1 **Estimates of negative and positive selection on coding and noncoding sites in C. grandiflora.**A) The proportion of sites found in each bin of purifying selection strength, separated by site type, B) The proportion of divergent sites fixed by positive selection, and C) the rate of adaptive substitution relative to neutral divergence. Error bars represent 95% bootstrap confidence intervals. 6

2.2 **Linked neutral diversity and divergence as a function of distance from fixed substitutions across the *C. grandiflora* genome.** A) Diversity at 4-fold degenerate sites, B) Divergence at 4-fold degenerate sites, and C) Diversity/divergence at 4-fold degenerate sites. In all figures, black lines represent measures surrounding fixed replacement substitutions and gray shading represents 95% confidence intervals, from bootstrapping, surrounding silent substitutions.). 8

2.3 **Linked neutral diversity/divergence surrounding conserved noncoding sequences (CNSs).** Diversity/divergence at 4-fold degenerate sites as a function of distance from fixed substitutions in CNSs (black lines) and fixed substitutions in non-conserved intergenic sequence (gray shading, 95% confidence interval). B) Diversity/divergence at 4-fold degenerate sites as a function of distance from CNSs containing fixed substitutions (black line) and CNSs without any fixed substitutions (gray shading, 95% confidence interval). . 9

2.4 **Estimates of negative and positive selection on nonsynonymous sites in genes of varying expression level.**A) The proportion of sites found in each bin of purifying selection strength, separated by expression level. B) The proportion of divergent sites fixed by positive selection and C) The rate of adaptive substitution relative to neutral divergence. Error bars represent 95% bootstrap confidence intervals. 10

2.5 **Coverage after filtering, across the genome.** A) The number of annotated sites in each category across the genome (light grey), and the number of sites that pass our filters and were used in analysis (dark grey). B) Proportion of sites that pass filters, calculated in 200kb windows, as a function of genomic position. 19

2.6 **Pairwise diversity and divergence at 4-fold degenerate sites across the entire genome.** Statistics were calculated in windows of 5,000 SNPs. Individual lines alternating between grey and blue represent chromosomes. The location of the centromere on each chromosome is indicated by the grey box along the x-axis. 20

2.7 **Coding density versus 4-fold degenerate diversity across the genome.** Each point represents one 10 kb window. Black points represent windows that do not overlap centromeres while grey points represent windows that do overlap centromeres. There is a slight negative correlation between diversity and coding density both with and without centromeric windows 21
2.8 **Regions of identity by descent in each sample.** The ratio of heterozygous to homozygous calls at sites that are polymorphic across individuals (in 200kb windows) plotted against position across the genome. Each sample is plotted separately and identified by sampled IDs. Individual lines alternating between grey and blue represent chromosomes. Regions of IBD were defined as windows where FIS was greater than 0.5 and are indicated by black lines along the x-axis. At most 3 regions of IBD overlap across all individuals. This occurs near the end of chromosome 1. 22
2.9 **FIS in windows across the genome in each sample.** FIS in 200kb windows is plotted across the genome. Each sample is plotted separately and identified by sample IDs. Individual lines alternating between grey and blue represent chromosomes. Regions of IBD were defined as windows where FIS was greater than 0.5 and are indicated by black lines along the 0 line of the y-axis. 23
2.10 **DFE-alpha results using all alleles, including IBD regions.** The distribution of fitness effects for 0-fold degenerate, 3 and 5 UTR, intronic, and intergenic sites are shown. For this analysis the genotyping calls were filtered as described in the methods, but the data was not downsampled in regions of IBD identified in Fig. 2.8. 24
2.11 **Estimates of positive and negative selection on different categories of CNSs.** A) Distribution of fitness effects. Stars indicate categories in which the fraction of nearly neutral sites was significantly different from the pooled sets of CNSs by a randomization test. B) α and C) ω for each category. Error bars indicate 95% CIs from 200 bootstraps. . 25
2.12 **Robustness of sweep analysis to different window sizes.** This panel shows the results of our scans for recurrent selective sweeps using alternative window sizes: 500bp on left and 2kb on right. Otherwise, the methods are the same as described previously. . . 26
2.13 **Additional diversity and divergence data for sweeps around substitutions in conserved noncoding regions.** The left panels show diversity at 4-fold degenerate sites and divergence at 4-fold degenerate sites around substitutions in conserved non-coding sequence (black lines) and non-conserved intergenic sequence (gray shading represents 95% confidence intervals). The right panels show the same information for diversity and divergence at 4-fold degenerate sites around conserved noncoding sequences containing fixed substitutions (black lines) and conserved noncoding sequences without fixed substitutions (gray shading represents 95% confidence intervals). 27
2.14 **Allele frequency spectra of replacement sites in genes with different expression levels.** 28

2.15 **Estimates of negative and positive selection on 0-fold sites in genes of varying expression level.** Data from this figure was generated using the divergence estimates from the whole genome alignments (as in Fig. 2.1) rather than divergence from PAML estimates (as in Fig. 2.4). Here AFS from 0-fold sites were compared to 4-fold sites, rather than non-synonymous to synonymous sites as in Fig. 2.4. A) The proportion of sites found in each bin of purifying selection strength, separated by expression level. B) The proportion of divergent sites fixed by positive selection and C) The rate of adaptive substitution relative to neutral divergence. Error bars represent 95% bootstrap confidence intervals. 29

3.1 **Detecting eQTLs and aseQTLs** (a) A gene model for an individual that is heterozygous at a regulatory locus (G/T) and at an informative coding site (A/T). The G allele increases expression relative to the C allele, (b) causing increased allelic expression of the reads carrying the A allele at the informative heterozygous site. We refer to this difference in allelic expression as ASE. (c) eQTLs are detected when there is a significant difference in total gene expression between individuals (represented by black circles) that are homozygous for the common allele of a SNP and individuals that are heterozygous at that SNP. (d) aseQTLs are detected when there is a significant difference in ASE between individuals that are heterozygous at a SNP and homozygous for either allele at that SNP. 32

3.2 **eQTL and aseQTL enrichments by site type.** The proportion of SNPs tested in each category that were found to be eQTLs is plotted on the y axis for (a) eQTLs and (b) aseQTLs. The exonic classes were determined by splitting the coding sequence of each gene into 5 equally sized pieces. Note that there were no exonic SNPs included in the aseQTL analysis. Error bars show the 95% confidence intervals from bootstrapping. . . . 34

3.3 **The site frequency spectra of eQTLs and aseQTLs** Minor allele frequencies of (a) eQTLs and (b) aseQTLs for observed data (red circles) and permuted data (gray circles, black lines are 95% confidence intervals). 35

3.4 **The relationship between minor allele frequency and effect size.** (a) eQTL minor allele frequency is plotted against the effect of that SNP on ASE, calculated as the mean difference in ASE between individuals heterozygous at the eQTL and individuals homozygous at the eQTL. Negative values occur when the the homozygote for the eQTL has greater ASE than the heterozygote. The trend line is calculated by linear regression (b) aseQTL minor allele frequency plotted against the effect of the aseQTL on total gene expression, calculated by taking the log of the absolute value of the mean difference in expression between individuals heterozygous at the aseQTL and individuals homozygous for the common allele at the aseQTL. The trend line was calculated by regression between minor allele frequency and the log of the expression effect. 42

3.5 **The site frequency spectra of QTLs detected in the frequency-controlled subsample.** 43

3.6 **Linkage disequilibrium in *C. grandiflora*.** Linkage disequilibrium was calculated for all SNPs within 1 kb of each other on scaffold 2. 1% of these pairs were randomly sampled for the above figure, which shows mean R2 between pairs in 10bp bins. 44

3.7 **The distribution of p values for all SNPs tested in eQTL analyses (a) and aseQTL analyses (b)** 45

3.8 **Example eQTL and aseQTL genes.** Manhattan plots for associations between SNPs and total expression (a and b) and ASE (c and d) for two genes, PAC:20895445 (a and c) and PAC:20904926 (b and d). Each black dot represents a SNP and is plotted by genomic position on the x axis and the negative log of the p value for association on the y axis. The gray line denotes the p value threshold corresponding to an FDR of 0.01. The gray boxes represent the exons of the gene. Note that PAC:20895445 is an ortholog of AT4G16250.1, PHYTOCHROME D and PAC:20904926 is an ortholog of ATG68185.1, a ubiquitin-like superfamily protein. 46

3.9 **The effect of designating a random associated SNP per gene as eQTL/aseQTL instead of the most associated SNP per gene.** The site frequency spectrum of eQTLs (a) and aseQTLs (b) for observed data (red circles) and permuted data (gray circles, black lines are 95% confidence intervals) when a random SNP is chosen per gene to be an eQTL or aseQTL. The same eQTLs and aseQTLs are plotted in (c) and (d). In (c), eQTL minor allele frequency is plotted against the effect of that SNP on ASE, calculated as the mean difference in ASE between individuals heterozygous at the eQTL and individuals homozygous at the eQTL. Negative values occur when the the homozygote for the eQTL has greater ASE than the heterozygote. The black line is calculated by linear regression. In (d), aseQTL minor allele frequency plotted against the effect of the aseQTL on total gene expression, calculated by taking the log of the absolute value of the mean difference in expression between individuals heterozygous at the aseQTL and individuals homozygous for the common allele at the aseQTL. The trend line was calculated by regression between minor allele frequency and the log of the expression effect. 47

3.10 **Tajimas D of eQTLs and aseQTLs within site type, recombination rate, and substitution type.**(a) shows Tajimas D for eQTLs of various categories. Red circles are the real data, gray circles show Tajimas D for permuted eQTLs, and black lines show 95% confidence intervals. Tajimas D was used to summarize the site frequency spectra and make plots more readable than they would be if raw frequencies were plotted. The total number of eQTLs in each category is shown with the red numbers and the mean number of permuted eQTLs in each category is shown with the black numbers (b) shows the same data as (a) but for aseQTLs. (c) shows Tajimas D for eQTLs and aseQTLs (red dots) and permuted eQTLs and aseQTLs (gray dots, black bars are 95% confidence intervals) for sites in low recombination regions (<3.45 cM/mB) and high recombination regions (>3.45 cM/mB). (d) shows Tajimas D for A/T to G/C substitutions that could be favored by gene conversion (conv) and other substitutions (notconv). Note that all Tajimas D values are significantly increased because only SNPS above a certain allele frequency were testable, so that even for 4fold degenerate sites in the analysis, Tajimas D is 2.403. 48

3.11 **A comparison of mapping programs in highly polymorphic regions.**RNAseq coverage for an example gene using mapping from Tophat (top) and Stampy (bottom). Colored lines indicate polymorphic sites compared to the reference. The arrows indicate regions where coverage was reduced in Tophat because of multiple polymorphisms. Note that Tophat reads have splice junctions while Stampy reads do not because we mapped to an exon-only reference. 49

3.12 **GC composition and expression by lane.** All genes included in the study were split into 20 equally sized bins by GC content. Expression in these bins was combined for each lane and plotted in box plots. 50
3.13 **The effect of increasing the number of SNPs required to measure ASE effects aseQTL detection.** The minor allele frequency of aseQTLs detected when ASE measurement required 1 heterozygous coding SNP (circles), 2 SNPs (triangles), and 3 SNPs (squares). While increasing the numbers of SNPs required to measure ASE reduced the number of aseQTLs detected, it did not qualitatively change our conclusions about the rareness of aseQTLs. 51
3.14 **The site frequency spectrum of aseQTLs when genes with SNPs showing ASE bias are removed from the analysis.** Red dots are frequencies of aseQTLs, gray dots are frequencies for permuted aseQTLs and black lines show 95% confidence intervals. . . 52

Chapter 1

General introduction

Chapter 2

Population genomics of *Capsella grandiflora*

2.1 Abstract

The extent that both positive and negative selection vary across different portions of plant genomes remains poorly understood. Here, we sequence whole genomes of 13 *Capsella grandiflora* individuals and quantify the amount of selection across the genome. Using an estimate of the distribution of fitness effects, we show that selection is strong in coding regions, but weak in most noncoding regions, with the exception of 5 and 3 untranslated regions (UTRs). However, estimates of selection on noncoding regions conserved across the Brassicaceae family show strong signals of selection. Additionally, we see reductions in neutral diversity around functional substitutions in both coding and conserved noncoding regions, indicating recent selective sweeps at these sites. Finally, using expression data from leaf tissue we show that genes that are more highly expressed experience stronger negative selection but comparable levels of positive selection to lowly expressed genes. Overall, we observe widespread positive and negative selection in coding and regulatory regions, but our results also suggest that both positive and negative selection on plant noncoding sequence are considerably rarer than in animal genomes.

2.2 Introduction

Determining the amount of positive and negative selection and how it varies across the genome has wide-ranging implications for understanding genome function and the maintenance of genetic variation [1]. Current evidence suggests that both positive and negative selection are common in coding and some noncoding sequences in several model systems [2-8]. However our understanding of genome-wide selection in plants remains relatively limited [6], particularly in noncoding regions.

One key question concerns the extent to which both positive and negative selection act in noncoding

regions of the genome compared with coding regions [2,5-8]. For example, it has been suggested that the majority of adaptive evolution may occur in noncoding regulatory regions, where new mutations may have fewer deleterious pleiotropic effects [9,10 but see 11]. Halligan and colleagues [8] showed that there have been many more adaptive substitutions in noncoding DNA than in coding regions in house mice, although adaptive substitutions in coding regions may experience stronger positive selection. Moreover, studies in Drosophila species and vertebrates have found that, although noncoding regions as a whole are generally less conserved than coding regions, there is more functional noncoding sequence than constrained coding sequence by a considerable margin [2,12].

Comparing these results to noncoding selection across plant genomes is of particular interest because it has been hypothesized that in plants, regulatory evolution may occur more often through gene duplication than cis-regulatory change [13], possibly leading to lower levels of functional constraint and positive selection on plant noncoding DNA. Consistent with this prediction, Haudry and colleagues [14] recently compared the genomes of nine Brassicaceae species, and showed that approximately one quarter of the conserved sites in the Arabidopsis thaliana genome were in noncoding regions, a much smaller fraction than found to date in studies of vertebrates and Drosophila. However, the strength of selection on these noncoding sites, the extent of species-specific selection in noncoding regions, and the extent of positive selection in noncoding regions compared with coding regions have not been quantified. While the strength of selection is expected to vary between coding and noncoding sequence, it also varies between genes. Gene expression level is one of the major determinants of rates of nonsynonymous evolution in coding regions in many species [15-17], including plants [18-21]. Variation in the strength of selection on genes could reflect differences in the relative importance of gene products for organism fitness, or it may simply relate to inherent properties of expression [22]. For example, deleterious mutations that cause misfolding or mis-interaction have more opportunity to interfere with cellular function when they occur in high expression genes [23-25]. Regardless of the underlying selective mechanisms, the negative correlation between expression level and nonsynonymous divergence could reflect relaxed purifying selection in lowly expressed genes, increased positive selection in lowly expressed genes, or both.

Here, we use population genomics to quantify the strength of both positive and negative selection inside and outside of coding regions and within highly and lowly expressed genes in a species-wide sample of 13 outbred *Capsella grandiflora* individuals. *C. grandiflora* is an obligately outcrossing member of the Brassicaceae family with a large effective population size (Ne $\sim$ 600,000) and relatively low population structure [26,27]. We estimate the strength of negative selection by fitting polymorphism data to a model of the distribution of negative fitness effects of mutations. We then quantify the contribution of positive selection to divergence in *C. grandiflora* using two complementary approaches: an extension of the McDonald-Kreitman test [28] and an analysis of neutral variation linked to lineage-specific fixed substitutions [29]. Our results demonstrate that both positive and negative selection are pervasive in coding regions, 5 and 3 untranslated regions (UTRs), and constrained noncoding regions of the *C. grandiflora* genome, but also that a large proportion of noncoding DNA may evolve neutrally. In addition, we find stronger negative selection in high expression genes compared to low expression genes, suggesting that differences in negative selection drive differences in rates of molecular evolution.

2.3 Results

2.3.1 Genome-wide patterns of polymorphism

We sequenced 13 outbred *C. grandiflora* individuals (26 sampled haploid chromosomes; ∼140 Mb genome assembly) sampled from across the species' range in northern Greece using single-end Illumina GAII sequencing (Table S1). The resulting 108 bp reads were mapped to the *Capsella rubella* reference genome [30] using the Stampy aligner resulting in a median coverage of 34 reads per sample per site. Genotypes were called using the Genome Analysis Toolkits Unified Genotyper [31]. After filtering for quality and depth (see Methods), we were left with ∼27 million sites, ∼1.5 million of which were single nucleotide polymorphisms (SNPs) (Table S2). Sites from across the genome were identified as 0-fold degenerate, 4-fold degenerate, intronic, 5 UTR, 3 UTR, or intergenic, based on the annotation of the *C. rubella* reference genome [30]. To avoid comparing sites that do not have equivalent mutation profiles, we excluded sites in coding regions that were neither 4-fold nor 0-fold degenerate. After filtering, our analysis includes 30-40% of coding and noncoding sites, except in intergenic regions where only approximately 10% of sites are retained due to the higher repeat content in these regions and the removal of highly repetitive pericentromeric DNA (Fig. 2.5).

Consistent with previous estimates made using a much smaller set of loci (257 Sanger-sequenced loci) and a different range-wide sample [32], average nucleotide diversity at 4-fold degenerate sites (Wattersons θ_w) was 0.022 and there was evidence for an excess of rare variants genome-wide at 4-fold degenerate sites compared with the standard neutral model (Tajimas D = -0.512). Introns (θ_w= 0.020) and intergenic regions (θ_w = 0.019) showed only slightly lower levels of nucleotide diversity than 4-fold degenerate sites, suggesting that the large majority of sites in these regions are effectively neutral, or subject to comparable levels of purifying selection as 4-fold degenerate sites. 5 and 3 UTRs showed a much stronger diversity reduction (θ_w= 0.015 and 0.014 respectively), while 0-fold degenerate nonsynonymous sites showed the strongest reduction (θ_w = 0.005).

Neutral diversity at 4-fold degenerate sites near centromeric regions was elevated on most chromosomes, similar to observations made in *A. thaliana* [33], *Arabidopsis lyrata* [34,35] and *Medicago truncatula* [36], (Fig. **??**. As with these other species, this effect is not obviously caused by higher mutation rates, since divergence between Capsella and Neslia is not clearly elevated in these regions (Fig. 2.6). Although elevated error rates in repetitive regions may contribute to high diversity, our observation of high diversity in these regions is still apparent after extensive filtering (see Methods). This increase in neutral diversity in pericentromeric regions may reflect a weakening of background selection in regions of low gene density, as recently shown in models of background selection applied to Arabidopsis [37]. Furthermore, diversity generally declines towards the ends of the chromosomes, potentially reflecting the stronger effects of background selection and/or selective sweeps in regions of relatively low recombination but high gene density, where the effects of linked selection are expected to be strongest. Consistent with these interpretations, we see an increase in diversity in regions of low coding density (Fig. 2.7).

We also examined individual heterozygosity in sliding windows along each chromosome. A number of individuals showed large stretches of homozygosity indicative of biparental inbreeding (Fig 2.8 and Fig. 2.9). Consistent with these regions reflecting local biparental inbreeding, no such regions are found

in our sample that is derived from a between-population cross, called AXE. These regions of identity-by-descent (IBD) in our data highlight that, despite being self-incompatible and obligately outcrossing, local biparental inbreeding can still generate excess homozygosity in stretches across the genome. To avoid biased estimation of species-wide allele frequencies in these regions, we subsampled the data to treat all IBD regions as haploid rather than diploid sequence for the purposes of allele frequency estimation, although treating these regions as diploid does not qualitatively change our conclusions (Fig. 2.10).

2.3.2 Genome-wide measures of purifying selection

In order to quantify the amount of negative selection acting on different categories of sites, we used the methods of Eyre-Walker and Keightley [1] to compare the allele frequency spectrum (AFS) and divergence of various site categories to those for 4-fold degenerate sites, which are putatively neutral (Fig. 2.1A). Consistent with the patterns of diversity described above, negative selection is generally much stronger in coding regions than noncoding regions (Fig. 2.1A). This pattern is most clearly seen in 0-fold degenerate sites, the only site category with a sizable fraction of sites in the strongest category of negative selection (41%). Of the noncoding categories, UTRs show much stronger negative selection than other regions. In C. grandiflora ∼55% of both 5' and 3' UTRs are under moderate levels of purifying selection (Nes ¿ 1), but a considerably larger fraction of UTR sites are effectively neutral (45%) than 0-fold degenerate sites (14%). Additionally while the UTRs and CNSs (see below) show a signal of strong purifying selection (Nes¿10), they experience less strong selection than 0-fold degenerate sites.

Genome-wide, we estimate that the proportion of intergenic sites that are nearly neutral approaches 100% and that approximately 70% of intronic sites are effectively neutral. Furthermore, bootstrapping results suggest that there is not significant support for less than 100% of intronic sites being effectively neutral. The large confidence intervals around estimates of selection on intronic sites may be due to strong selection at splice site junctions [14] coupled with typically weak to no selection outside of splice junctions. To test for selection near splice junctions, we quantified selection acting on the first and last 30 bp of each intron separately from sites in the middle of introns. While 100% of sites in the middle of introns are estimated to be effectively neutral, only 68% of sites in junctions are, suggesting that our wide confidence intervals around intronic sites can be partially explained by variance caused by sampling sites in these different regions between bootstraps. These generally low estimates of Nes in (non-junction) intronic and intergenic sites imply a general lack of purifying selection in most noncoding regions, a lack of sensitivity to detect small proportions of selected sites, and/or nearly equivalent purifying selection to synonymous sites.

Although our analysis suggests very low levels of purifying selection in noncoding regions other than UTRs and splice junctions, these global analyses may miss signatures of purifying selection on a small proportion of noncoding sites. One candidate set of sites that may have different signatures of selection are conserved noncoding sequences (CNSs); these are regions that show evidence of cross-species conservation, and are therefore prime candidates for functional noncoding sequences subject to selection. We identified CNSs across nine Brassicaceae genomes, following the implementation in Haudry et al. [14]. For this study, we used the Capsella genome as a reference for alignment, but excluded Capsella when identifying CNSs in order to avoid circularity when quantifying selection from diversity [8]. This

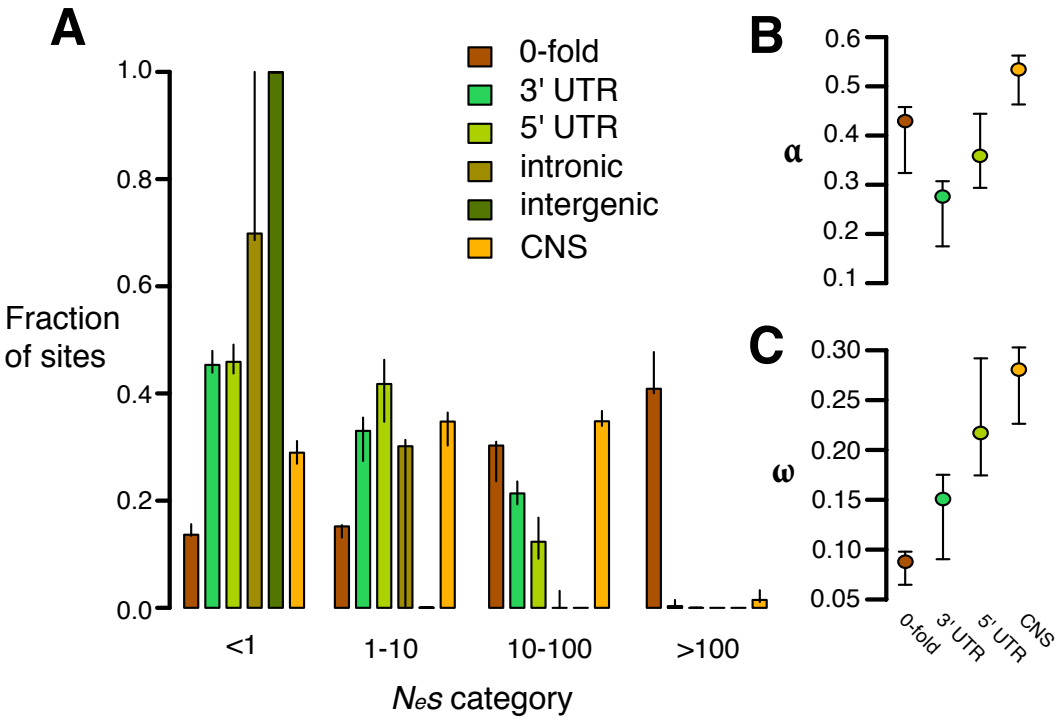

Figure 2.1: **Estimates of negative and positive selection on coding and noncoding sites in C. grandiflora.**A) The proportion of sites found in each bin of purifying selection strength, separated by site type, B) The proportion of divergent sites fixed by positive selection, and C) the rate of adaptive substitution relative to neutral divergence. Error bars represent 95% bootstrap confidence intervals.

method allows our analysis of selection on noncoding sites using polymorphism to be more independent of the comparative analysis. When we look at only these conserved regions in our C. grandiflora sample we see a small proportion of effectively neutral sites (28%) compared to the noncoding regions as whole, suggesting that the majority of CNS sequences are subject to purifying selection (Fig. 2.1A). However, estimates suggest that CNSs are generally under weaker purifying selection than nonsynonymous (0-fold) sites and experience primarily weak and intermediate purifying selection (Fig. 2.1A).

Although CNSs as a whole retain a considerable proportion of effectively neutral sites, it is of interest to examine whether particular classes of CNS show stronger selection. To examine differences between categories we quantified selection on the different types of CNSs separately (Fig. 2.11). In most categories, about 25% of sites are nearly neutral, a slightly stronger signal of purifying selection than when we pool all CNSs. Intronic CNSs have a larger proportion of effectively neutral sites than other categories, in agreement with the general neutrality of intronic sites (Fig. 2.1). In contrast, small noncoding RNAs (sncCNSs) have a stronger signal of selection than the other CNS categories. However, the number of sites used to make the AFS for each of these categories varies substantially (Table S2), and our sample of sncCNSs has very little polymorphism (155 segregating sites). Nevertheless, despite the wide confidence intervals, sncCNSs still show a significantly ($p < 0.001$) smaller fraction of sites that are nearly neutral ($Nes < 1$) than the pooled CNSs, which could be due to strong selection for sequence specificity to obtain the proper secondary structure important for RNA activity [38]. This effect is consistent with sncCNSs

showing a higher degree of conservation across the Brassicaceae [14] and having traceable orthologs in other plants.

2.3.3 Genome-wide estimates of positive selection

We used the approach of Eyre-Walker and Keightley [28] to estimate the proportion of fixations driven by positive selection (α) and the rate of positive selection (ω) while taking into account the effect of slightly deleterious mutations, which can bias estimates of positive selection downwards. To do this, we estimated divergence using whole genome alignments of *C. rubella*, *A. thaliana*, and *Neslia paniculata* (estimate of 4-fold synonymous divergence K_s between *C. rubella* and *N. paniculata* is K_s=0.14). Because the large majority of noncoding sites are estimated to be effectively neutral, and because of alignment concerns between species in unconstrained noncoding regions, we focus our estimates of positive selection on 0-fold degenerate sites, CNS sites, and UTRs. We found that 0-fold degenerate sites show a very high proportion of divergence driven by positive selection (Fig. 2.1B; $\alpha = 0.417$) and estimates of the rate of adaptive substitution relative to synonymous substitution (Fig. 2.1C ; $\omega = 0.08$). Similarly, UTRs and CNS sites show evidence for positive selection (Fig. 2.1B,C). These results generally suggest widespread positive selection in both nonsynonymous and functional noncoding genomic regions.

If many of the amino acid changes between textitC. grandiflora and its nearest relatives are due to recent, strong positive selection from new mutations, we expect to see the signature of selective sweeps: reduced neutral diversity surrounding amino acid fixations [39,40]. We tested for this signature by measuring the proportion of 4-fold degenerate sites in each window that were polymorphic (referred to hereafter as '4-fold diversity') in non-overlapping 1kb windows surrounding fixed replacement (n = 60,378), and silent (n = 83,812) substitutions in *C. grandiflora*. We found that 4-fold diversity surrounding fixed replacement substitutions was lower than 4-fold diversity surrounding fixed silent substitutions in the 4kb window surrounding substitutions (Fig. 2.2A). This result was robust to various window sizes from 500kb to 2kb (Fig. 2.12) and a one-tailed test for reduced 4-fold diversity around replacement sites was significant ($p < 0.01$ for 2 kb on either side of the substitution).

Patterns of diversity may be distorted by elevated mutation rates surrounding substitutions [39], which would increase diversity and divergence in *C. grandiflora*. Consistent with this prediction, divergence at 4-fold degenerate sites ('4-fold divergence') is elevated around synonymous and replacement substitutions (Fig. 2.2B). To control for elevated mutation rate, we divided diversity by divergence at 4-fold degenerate sites (subsequently referred to as '4-fold diversity/divergence'). We observed a reduction in 4-fold diversity/divergence around replacement substitutions compared to silent substitutions, demonstrating that the signature of recurrent sweeps is not an artifact caused by variation in mutation rate (Fig. 2.2C, $p < 0.01$ for 1 kb on either side of the substitution).

An analogous test for selective sweeps around fixations in noncoding regions is challenging because the test depends on accurately identifying interspersed functional and neutral sites, a difficult task in noncoding regions [8]. Instead, we compared 4-fold diversity and divergence around fixed substitutions in CNS regions (n = 12,578) with 4-fold diversity and divergence around fixed substitutions in non-conserved intergenic, intronic, and UTR regions (n = 117,178). Interestingly, there is a reduction in both 4-fold diversity and divergence surrounding fixed substitutions in CNSs compared to non-conserved

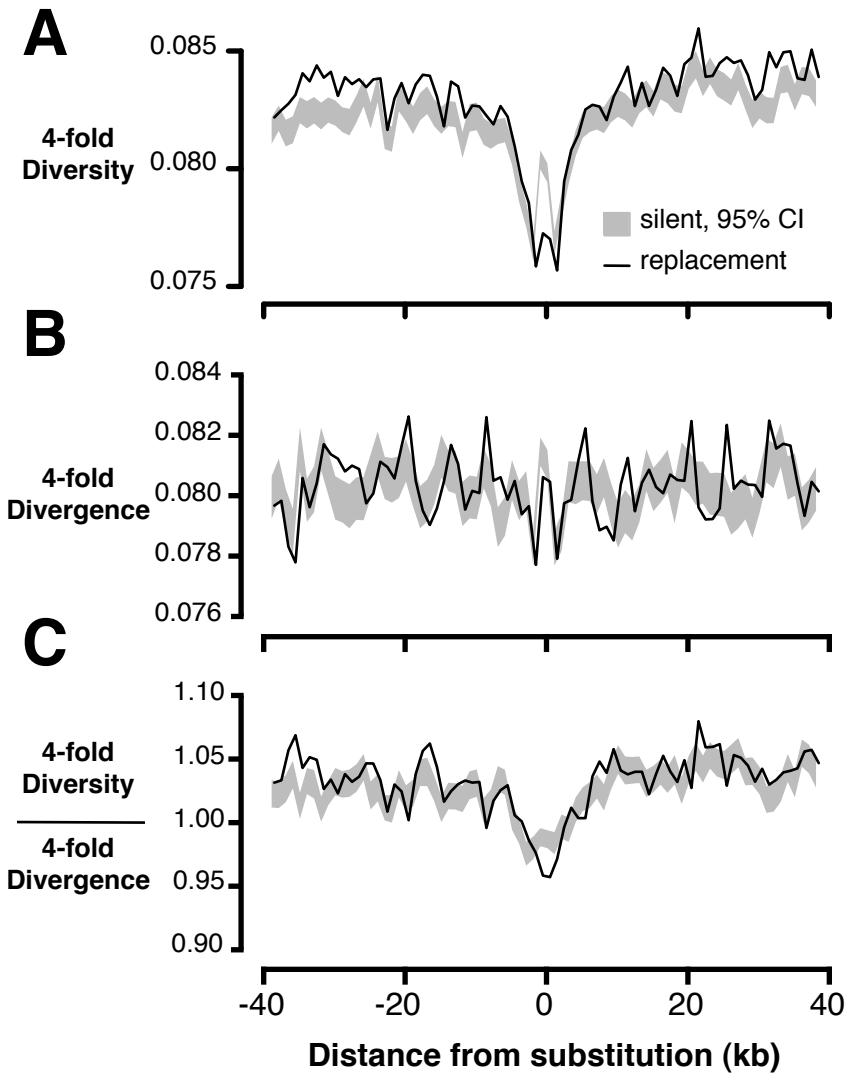

Figure 2.2: **Linked neutral diversity and divergence as a function of distance from fixed substitutions across the *C. grandiflora* genome.** A) Diversity at 4-fold degenerate sites, B) Divergence at 4-fold degenerate sites, and C) Diversity/divergence at 4-fold degenerate sites. In all figures, black lines represent measures surrounding fixed replacement substitutions and gray shading represents 95% confidence intervals, from bootstrapping, surrounding silent substitutions.).

noncoding regions (Fig. 2.13). It is not clear why 4-fold divergence decreases around CNS substitutions; it is possible that in genomic scans for conserved regions, large-scale constraint might span both coding and noncoding sequence, causing non-independence and reducing divergence at 4-fold degenerate sites near CNSs. However, there is still a reduction in 4-fold diversity/divergence around fixed substitutions in CNSs compared to those in non-conserved intergenic regions, consistent with the action of recurrent selective sweeps (Fig. 2.3A).

The observed reduction in diversity/divergence around CNS substitutions could also reflect the action of background purifying selection; sites closer to CNSs may experience a reduction of neutral diversity

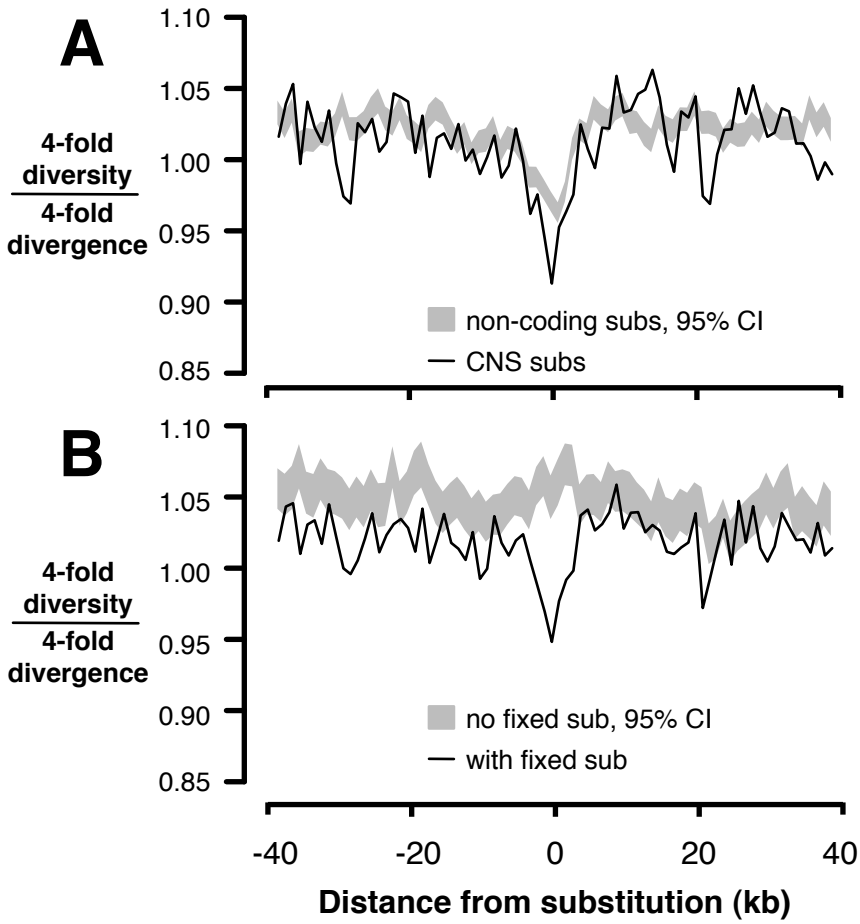

Figure 2.3: **Linked neutral diversity/divergence surrounding conserved noncoding sequences (CNSs).** Diversity/divergence at 4-fold degenerate sites as a function of distance from fixed substitutions in CNSs (black lines) and fixed substitutions in non-conserved intergenic sequence (gray shading, 95% confidence interval). B) Diversity/divergence at 4-fold degenerate sites as a function of distance from CNSs containing fixed substitutions (black line) and CNSs without any fixed substitutions (gray shading, 95% confidence interval).

due to greater purifying selection on mutations in CNSs. This effect is not a problem for comparisons between replacement and silent substitutions because they are interspersed within the same exons, so diversity and divergence around these sites experience the same background selection. To ensure that the reduction in diversity/divergence surrounding CNS substitutions compared to non-conserved noncoding substitutions is not due to differences in background selection between CNS and intergenic sites, we compared neutral diversity and divergence surrounding CNSs that contain at least one fixed substitution to neutral diversity and divergence around those that do not. There is a reduction in neutral diversity/divergence surrounding CNSs containing a fixed substitution (n = 12,884) compared to CNSs without fixed substitutions (n = 41,212), suggesting that this signature of recurrent sweeps is not driven by background selection specific to CNSs (Fig. 2.3B).

2.3.4 Effects of expression and selection

We measured expression levels of all expressed genes using RNA extracted from leaf tissue of 10 of the 13 *C. grandiflora* individuals. Genes were sorted by mean expression level and split into four equally sized groups, which will be referred to as high, mid-high, mid-low, and low expression genes. We calculated polymorphism within *C. grandiflora* and lineage-specific divergence from *N. paniculata* and *A. thaliana* for sites within these genes. As expected from previous studies, d_N/d_S is considerably lower in high expression genes (0.15) than low expression genes (0.22). In addition, d_N/d_S is negatively correlated with expression level across all genes (correlation coefficient = -0.051, p <0.001).

To test whether the strength of negative selection differs between expression categories we compared the allele frequency spectra of sites in different expression categories. Replacement polymorphisms in high expression genes show a stronger skew towards rare alleles than those in low expression genes (FIg. 2.14). In addition, a larger proportion of replacement sites are invariant in high expression genes (98.9%), than in low expression genes (97.8%), consistent with stronger negative selection. Comparisons of the distribution of fitness effects show that high expression genes have a much smaller proportion of effectively neutral sites (6.8%) than low expression genes (16%, randomization test [28], p <0.001) (Fig. 2.4A).

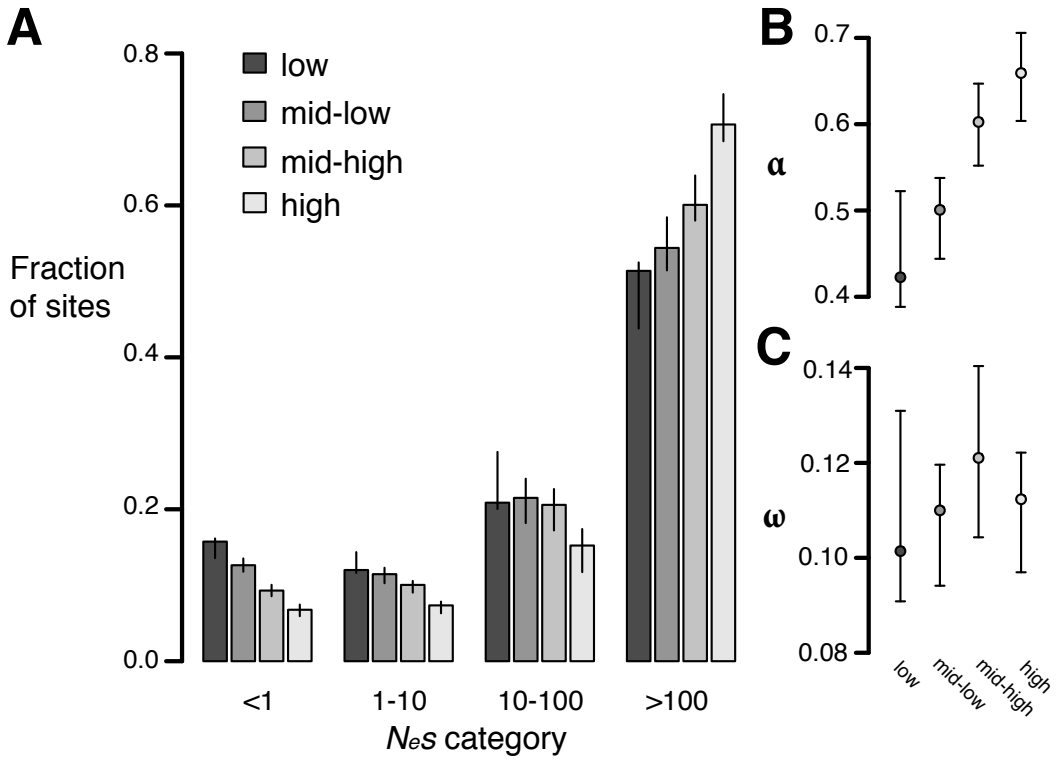

Figure 2.4: **Estimates of negative and positive selection on nonsynonymous sites in genes of varying expression level.** A) The proportion of sites found in each bin of purifying selection strength, separated by expression level. B) The proportion of divergent sites fixed by positive selection and C) The rate of adaptive substitution relative to neutral divergence. Error bars represent 95% bootstrap confidence intervals.

Increased divergence in low expression genes relative to high expression genes could also be caused by increased positive selection in low expressed genes compared to highly expressed genes. To test this possibility, we calculated α and ω as described above. High expression genes have a significantly higher value of α (0.66) than low expression genes (0.42, p <0.01) but the ω value for both classes is similar (high: 0.11, low: 0.10, p = 0.38), suggesting that the rate of positive selection does not differ between high and low expression genes (Fig. 2.4B,C). The difference in α between the two categories likely reflects the reduction in the number of weakly deleterious and effectively neutral mutations that are able to fix due to stronger purifying selection in high expression genes compared to low expression genes, causing a higher proportion of those amino acids that do reach fixation to be positively selected.

2.4 Discussion

In this population genomic survey of *C. grandiflora*, we demonstrated that positive and negative selection contribute to DNA sequence variation in protein-coding regions, UTRs, and CNSs. We also showed that differences in divergence between high and low expression genes are due to increased negative selection in high expression genes, not increased positive selection in low expression genes. In addition, we found a clear signature of recurrent selective sweeps contributing to divergence in coding regions as well as CNSs. Overall, our evidence for widespread positive and negative selection in *C. grandiflora* is in line with expectations, given its outcrossing mating system, large N_e, limited population structure, and lack of a recent whole genome duplication [6].

In contrast, selection appears to be very rare in intergenic and (non-junction) intronic regions that are not conserved across Brassicaceae species. In particular, we cannot detect significant evidence of purifying selection in intergenic or intronic regions as a whole, suggesting that selected sites within these regions must be rare or absent. However, when we only examine CNSs, we do see evidence of selection, indicating that at least 5% of sites in intergenic regions are selected, but the DFE approach is not sensitive enough to detect selection on such a small subset of intergenic sites. This result implies that this approach is likely to also be missing lineage-specific selection when it comprises a relatively small fraction of sites, and it highlights the importance of integrating additional evidence of function (comparative and experimental) for improved quantification of selection. The general neutrality of noncoding regions, based on population genomic analysis, is consistent with the conclusions of Haudry and colleagues [14], who used comparative genomics approaches to estimate that only 5% of noncoding bases are under selection in the Arabidopsis genome. This result contrasts with Drosophila and humans, where a relatively large fraction of selected sites are found in noncoding regions [6]. For example, in Drosophila, only 30%-70% of intronic and intergenic regions are nearly neutral [2,28,29]. Similarly, Halligan et. al. [8] recently used information from the DFE to infer the number of adaptive substitutions in mice both in coding and noncoding regions. They show that the majority (approximately 80%) of the adaptive substitutions in the mouse genome are in noncoding regions and suggest that they may have regulatory function. In contrast, our data show that C. grandiflora has similar numbers of adaptive substitutions in 0-fold sites (50.6 kb) and noncoding sites (21.6 kb, 3 UTR excluding CNSs; 10.2 kb, 5 UTR excluding CNSs; 32.7 kb, CNS; 64.4 kb total). Additionally, the width of diversity reductions surrounding replacement substitutions and substitutions in CNS regions appear comparable, suggesting

that there is little evidence for a difference in the strength of positive selection on substitutions in coding regions compared to conserved noncoding regions. Our results are consistent with previous suggestions that, unlike in animals, plant genomes may contain fewer noncoding regulatory sequences subject to positive and negative selection, possibly because gene expression can be modified through frequent gene duplication and functional divergence rather than through the evolution of novel regulatory elements [13]. In future work, it would be interesting to quantify the extent to which adaptive changes in gene expression in plants occur following gene duplication relative to between-species divergence at orthologous genes.

Unlike other classes of noncoding sequence, UTRs show relatively high levels of purifying selection, likely reflective of their function in post-transcriptional regulation [41]. UTRs are also under stronger negative selection than other noncoding regions in Drosophila [2], and this result is also in line with the previous study using comparative genomics in the Brassicaceae [14]. Interestingly, we infer that a large fraction of selected sites in UTRs may be outside of CNS regions identified in between-species comparisons. In particular, using estimates of the proportion of sites under selection, we estimate that 88% of 3 UTR and 77% of 5 UTR selected sites are outside of conserved regions. This result suggests that there may be many species-specific (i.e., non-CNS) functional regions in UTRs and they may therefore play an important role in recent or local adaptation.

One important consideration is the extent to which our analyses are truly reflective of genome-wide patterns of selection. Despite whole genome sequencing, our analyses are restricted to approximately 20% of the genome, and only 10% of intergenic sites, largely due to the fact that a large fraction of the genome is pericentromeric, repetitive and/or surrounds insertion/deletion events. It is important to recognize that our estimates of selection apply strictly to this accessible genome and that the extent of purifying and positive selection on the repetitive regions remains difficult to assess. Nevertheless, we would expect that our conclusions about low levels of purifying and positive selection across most noncoding regions are likely conservative with respect to these filters because a large proportion of repetitive DNA is likely to be neutral. On the other hand, rates of positive selection may be elevated in coding regions of duplicate genes filtered out of our analysis [42], suggesting that our estimates of positive selection in protein-coding regions may also be a lower bound. A second concern is the extent to which synonymous sites are neutrally evolving. Although analysis of codon usage bias from population genetic data does suggest the action of some purifying selection on synonymous sites in this species [43], the strength of selection inferred is close to effective neutrality. Furthermore, synonymous site selection is expected to be stronger in more high expression genes [23,44], causing us to underestimate, rather than overestimate, the difference in the strength of purifying selection compared with low expression genes. Thus, while selection on synonymous sites may bias our estimates of selection slightly downward, our general conclusions are likely to be robust to violations of neutrality. Nevertheless, more investigation of the action of selection on synonymous sites is important, particularly given growing evidence for synonymous site selection that may reflect gene regulation, in addition to codon usage [45,46].

At synonymous sites, we see an excess of rare variants, as indicated by a negative Tajimas D. The excess of rare variants is unlikely to be explained by a high Illumina error rate, as our observed value of -0.51 is nearly identical to a previous estimate (-0.52) from Sanger-sequenced loci and a comparable geographic sampling [27]. This previous study found that, while population subdivision was low compared to other herbaceous species studied, there were still three major geographic clusters (average between-population

Fst of 0.11). If we restrict our dataset to one of the three geographic regions based on these previous results, Tajimas D approaches zero (-0.16 at 4-fold degenerate sites), suggesting that the excess of rare variants at synonymous sites may be largely due to population structure.

2.4.1 Measuring positive selection

In this study, we took advantage of the two detectable signatures expected to remain after recurrent classic selective sweeps from new mutations: 1) an excess of replacement substitutions relative to expectations based on polymorphism, and 2) reduced neutral diversity near fixed differences. Our findings strongly suggest that positive selection has been common in coding regions, UTRs and conserved non-coding regions in *C. grandiflora* and that classic selective sweeps contribute significantly to divergence in these regions. To our knowledge, this is the first time that the signature of recurrent selective sweeps has been observed in a non-Drosophila species, despite being tested in other species [8,47]. Our ability to detect the signature of recurrent sweeps may be because *C. grandiflora* has relatively low linkage disequilibrium, increasing power.

However, many positively selected alleles may not follow the trajectory of a classic selective sweep. Soft sweeps adaptation from an allele previously maintained in the population by mutation-selection-drift balance or the simultaneous fixation of multiple independently derived mutations at the same allele may still increase the replacement to silent divergence ratio, but are expected to have a smaller effect on linked neutral diversity [48-50]. We expect that soft sweeps will also be common in *C. grandiflora* because of its large Ne [50,51]. In addition, adaptation in genes that contribute to polygenic traits is often expected to occur without fixation of new mutations [52], and this will be missed by both of our tests for positive selection. These considerations suggest that both measures of positive selection are conservative and may miss many instances of positive selection acting in the genome. Our conclusions about the prevalence of selective sweeps in *C. grandiflora* may seem to conflict with our observation that diversity and Tajima's D are slightly higher at 4-fold degenerate sites than intergenic sites, since frequent sweeps in coding regions should reduce diversity more strongly in sites near and within genes. There are two likely contributors to this discrepancy. First, recurrent sweeps may in fact reduce average diversity in 4-fold degenerate sites and, by using these sites to set neutral expectations, we are underestimating the strength of purifying selection in intergenic regions. Second, because recombination rates are relatively high, and intergenic regions near coding regions relatively small in Capsella, the average impact of linked selection may be similar at 4-fold degenerate sites and intergenic sequences.

2.4.2 Expression level and selection

Highly expressed genes diverge less than genes with low expression in many species [15-17,19,24,53-55]. This pattern could be due to stronger positive selection in low expression genes or stronger negative selection in high expression genes, or both. Our results suggest that variation in divergence rates between high and low expression genes is largely due to increased negative selection in high expression genes compared to low expression genes. This result is consistent with previous studies that have suggested that new nonsynonymous mutations that cause protein mis-folding or mis-interaction will have stronger

deleterious effects in high expression genes than low expression genes and that new mutations that cause mRNA mis-folding are under stronger negative selection in high expression genes than low expression genes [23-25]. In addition, our results agree with a similar study in Medicago truncatula that found stronger purifying selection in genes that were expressed than in genes that were not expressed [20].

2.5 Methods

2.5.1 Sampling and sequencing

Population samples for *C. grandiflora* represented a scattered sample of one individual per population for twelve populations from across the geographic range in Greece, plus a thirteenth sample that was the product of a cross of two additional populations (Table S1). Plants were grown for several months at the University of Toronto greenhouse, and genomic DNA was extracted from leaf tissue using a modified CTAB protocol. Library preparation and single-end genomic sequencing were conducted at the Genome Quebec Innovation Centre at McGill University on the Illumina GAII platform. Each sample was sequenced in 2 to 3 lanes and with a read length of 108 bp.

Leaves from 10 of the 13 individuals were collected and flash frozen for RNA extraction using Qiagen's RNAeasy plant extraction kit. This RNA was sequenced at the Genome Quebec Innovation Centre, on an Illumina GAII platform with one individual per lane, generating single-end 108 bp long reads. The RNA sequence from these 10 individuals was used for the annotation of the *C. rubella* reference genome, as reported in [30], but the raw sequence data was reanalyzed for this study (see below).

2.5.2 Genotyping

Genomic reads were aligned to the *C. rubella* reference genome [30] using the Stampy aligner 1.0.13 with default settings [56]. Sites around indels were realigned using the Genome Analysis Toolkit (GATK) v1.05777 indel realigner [31]. Genotype and SNP calls were conducted using the GATK UnifiedGenotyper with default parameters [57], after aligning and genotyping the median site quality was 89 and the median individual depth across all sites was 34.

To get a rough assessment of genotyping error rates, we conducted Sanger sequencing from nine coding regions in six of our individuals. From a total of 16,389 bp of Sanger sequence, we found 8 differences between Sanger and Illumina genotypes, giving an estimated error rate of 0.00049. Three of these disagreements were due to three segregating bases at a single site, which we excluded in our GATK genotyping protocol. As we suspect several of these disagreements may be due to Sanger sequencing errors due to variation in allelic representation of heterozygotes, this provides an upper bound estimate of error rate in coding regions, although higher indel rates and repetitive sequence in noncoding DNA may lead to a higher error rate in those regions.

AFSs were generated from counts of sites in the VCF. Invariant sites were excluded from the AFS if (1) the site quality score was below 90, (2) the fraction of reads containing spanning deletions was not

0 (i.e. the 'Dels' value was greater than zero), or (3) any individual's read depth was less than 20 or greater than 60. Additionally, polymorphic sites were excluded, based on filters 1-3, if (4) the most likely genotype of any individual did not have a phred scaled likelihood score of 0, and if (5) the second most likely genotype had a phred likelihood score less than 40. Additionally, entire regions of the genome were filtered out of the analysis if less than 30% of the sites in a 20kb window passed all other filters. This final filter primarily eliminated pericentromeric regions that were highly repetitive, where we were not confident in genotype calls and observed high heterozygosity.

Our data showed evidence of identity by descent (IBD) in some samples (Fig. 2.8). We identified these regions by splitting the genome into 200kb windows, then calculating FIS (Fig. 2.8). If FIS was greater than 0.5, the region was flagged as IBD. Across all samples no more than 3 of these regions overlapped. For further analyses we downsampled data in other regions down to 23 chromosomes treating any region of IBD as haploid to ensure that no IBD region was sampled twice from the same individual.

2.5.3 Divergence

We calculated lineage-specific divergence in two ways. First, we aligned the *C. rubella* reference sequence with sequence data from *A. thaliana* and *N. paniculata* using lastZ [58] with chaining, as previously described [14]. In order to get an estimate of divergence unique to the Capsella lineage, we called sites as diverged where *A. thaliana* and *N. paniculata* had the same nucleotide and this nucleotide differed in the *C. rubella* sequence. If any of the three species was missing data at a site, then that site, and sites 5 bp upstream and downstream of the site, were excluded from divergence analyses in order to avoid inflating divergence because of spurious alignments around indels.

We used a second method for calculating divergence for comparisons that included only coding sequences, particularly for the comparison of genes with different expression levels. We found orthologs between *C. rubella*, *A. thaliana* and *N. paniculata* genes using InParanoid [59] and MultiParanoid [60]. The peptide sequences of these orthologs were aligned using DialignTX [61], and reverse-translated into coding sequence. Whole-gene divergence at synonymous and nonsynonymous sites was calculated, using PAML [62], under a model where ω was allowed to vary in the Capsella lineage compared to other branches.

We conducted comparisons of estimates of the distribution of fitness effects using the two methods above with identical gene sets, and found a very strong concordance of results (see Fig. 2.4 compared to Fig. 2.15). Furthermore, while we don't predict a significant effect on results, it is important to note that the two methods also differed in how selected and nonselected classes were determined: the first distinguishes between 0-fold and 4-fold sites and discards other sites, while the second distinguishes between synonymous and nonsynonymous sites, including all data. However, both approaches gave comparable estimates of positive and negative selection.

2.5.4 Identifying conserved noncoding sequences

Conserved noncoding sequences (CNS) were identified in the *C. rubella* genome by first obtaining whole-genome multiple alignments, using a variant of the lastZ/Multiz pipeline previously described [14,63] and using *C. rubella* as the reference genome. The *C. rubella* genome sequence was then neutralized (bases replaced with N) and the PhastCons tool used to quantify family-wide levels of conservation. CNSs were then identified, based on extended (¿12nt) near-continuous regions of high conservation as previously described [14].

2.5.5 Estimates of the distribution of fitness effects and α

Site categories were determined based on the Joint Genome Institutes gene annotation of the *C. rubella* reference genome [30]. The allele frequency spectra (AFS) and divergence values were calculated for each category of sites, and DFE-alpha [28,64] was used to estimate the fraction of sites under negative selection and , using 4-fold degenerate sites as the neutral reference. The genome was broken up into 10 kb regions and these regions were bootstrapped 200 times to generate 95% CIs for selection on each category of sites. We tested for a significant difference in selection between the pooled set of CNSs and each individual category of CNSs using a randomization test, as in Keightley and Eyre-Walker [28], by calculating the proportion of bootstraps where selection was higher in the pooled set of CNS versus the category of interest. Because this is a two-tailed test, we report twice this proportion as the p value.

2.5.6 Test for signatures of recurrent selective sweeps

We used the multiple species alignments of orthologous genes, generated as described above, to identify silent and replacement single-nucleotide sites that were the same in A. thaliana and N. paniculata but differed in the *C. rubella* reference, suggesting that the substitution had most likely occurred in the Capsella lineage after divergence from *N. paniculata*. From these substitutions, we identified those that did not diverge between *C. rubella* and *C. grandiflora* and were fixed in *C. grandiflora*.

We calculated neutral diversity in sliding windows around fixed substitutions by calculating the proportion of 4-fold degenerate sites within these windows that were polymorphic in *C. grandiflora* (i.e., the proportion of segregating sites). Neutral divergence was measured by calculating the proportion of 4-fold sites within these windows that diverged in the Capsella lineage. Diversity/divergence was calculated by dividing diversity by divergence in each window. We conducted this analysis for windows of 500bp, 1kb, and 2kb, extending 40kb from each substitution. We chose this window size range to match analysis done in Sattath et al [39]. For each of the above measures, we bootstrapped by substitution (n=1000) and removed the top and bottom 25 bootstraps to construct 95% confidence intervals. Following Hernandez and colleagues [47], we tested the null hypothesis that diversity/divergence around replacement and silent substitutions does not differ by calculating a one-tailed p value for each window, equal to (i+1)/(n+2) where i is the number of bootstraps in which diversity/divergence around silent sites is lower or equal to the actual diversity/divergence around replacement sites, and n is the total number of bootstraps.

To detect the effects of linked selection on noncoding DNA, we compared diversity around fixed substitutions within CNSs to diversity around fixed substitutions in non-conserved intergenic regions. To find these substitutions, we compared the multiple sequence alignments of the CNSs between *C. grandiflora*, *N. paniculata*, and *A. thaliana* and chose sites that differed between *C. grandiflora* and the other species and were fixed within *C. grandiflora*. Additionally, we compared neutral diversity around CNSs with at least one fixed substitution to neutral diversity around CNSs without any fixed substitutions.

2.5.7 Gene expression

Illumina sequencing generated 331,629,531 reads for 10 individuals, ranging from 31,267,774 to 35,552,133 reads per individual. This RNA sequence was mapped to the *C. rubella* reference genome using Tophat 1.2.0 [65], and expression level was quantified from these mapped reads using Cufflinks 1.3.0 [66]. Cufflinks standardizes expression levels by gene length and library size, returning values in units of 'fragments per kilobase of exon per million fragments mapped' (FPKM). We calculated the mean expression level for each gene across our 10 samples and removed those genes with ¡1 FKPM to eliminate genes that may have been mis-annotated. The remaining 11,564 genes were divided into four, roughly equally sized categories based on expression level: low (1-6.8 FPKM), mid-low (6.8 17.5 FPKM), mid-high (17.5-44.7 FPKM), and high (44.7 17,092 FPKM). The distribution of fitness effects, , and were calculated for each gene set, using the same protocol described above. We bootstrapped each gene set by sampling genes with replacement 1000 times to generate 95% confidence intervals for selection strength. Using the same methods described for tests of differences within the CNSs categories above, we tested for a significant difference in selection strength between high and low expression genes.

2.6 Acknowledgements

We thank Peter Keightley and Dan Halligan for advice and custom scripts, Yunchen Gong and Emilio Vello for technical assistance, Tanja Slotte, Kate St. Onge, and John Paul Foxe for collecting seeds, and Detlef Weigel, Dan Koenig, Thomas Bureau, Alan Moses, Daniel Schoen, and John Stinchcombe for helpful discussion and/or comments on the manuscript. We would also like to thank Jeff Ross-Ibarra and two anonymous reviewers for helpful comments on the manuscript.

2.7 Appendix: Supplementary figures and tables

Table S2

Accession	Latitude	Longitude	# of BP sequenced
94.12	39.9597433333	20.7239333333	12360821508
83.17	38.4379	21.4243166667	7284286908
85.33	39.5552166667	20.9164166667	613454040
AxE	NA	NA	9941039172
918/8	39.75	19.8666666667	0
Cg2e	39.67175	19.7010166667	0
103.17	39.5183833333	21.5609166667	9999942588
5a	39.705025	19.757344	10451874852
91.23	39.86715	20.7070833333	12208550040
93.23	39.9644833333	20.71075	11152774332
95.15	39.1454166667	20.0581666667	7295444172
86.8	39.0172333333	20.1319333333	10473071796
88.56	39.0511833333	20.0666333333	6991120692
97.26	39.1164833333	21.1562166667	10100482488
98	38.0313166667	20.1539666667	6553236744

Table 2.1: **Sampling locations of each individual.** Note that individual AxE is a cross between 918/8 and Cg2e

	Positive selection		Distribution of fitness effects			
Site type	α	ω	0-1	1-10	10-100	100-Inf
0fold	0.417391	0.083841	0.136569	0.152498	0.303932	0.407
3utr	0.276083	0.150881	0.453332	0.33022	0.213531	0.002917
5utr	0.393226	0.251128	0.45836	0.417524	0.124102	0.000014
intergenic	NA	NA	0.999702	0.000298	0	0
intronic	NA	NA	0.698462	0.30153	0.000008	0
2/3 fold	0.173876	0.122638	0.611852	0.120999	0.139095	0.128055
CNS	0.545225	0.279022	0.275543	0.342483	0.363507	0.018467
3UTRcns 0.517609	0.255635	0.281223	0.332566	0.363655	0.022556	
DownstreamCNS	0.526032	0.225976	0.248254	0.434956	0.31525	0.001541
5UTRcns	0.424353	0.157474	0.253091	0.319996	0.39247	0.034442
UpstreamCNS	0.530125	0.224175	0.239224	0.366778	0.381506	0.012493
intronicCNS	0.50238	0.372934	0.41499	0.256174	0.28281	0.046026
IntergenicCNS	0.579499	0.255211	0.225545	0.401574	0.367618	0.005262
sncCNS	0.705816	0.174081	0.089996	0.38704	0.517114	0.00585
AmbiguousCNS	0.399807	0.158877	0.284093	0.368947	0.338238	0.008722
High expression	0.641085	0.107721	0.069669	0.066187	0.128722	0.735421
Mid-high expression	0.60684	0.122296	0.092334	0.099879	0.204814	0.602973
Mid-low expression	0.5052	0.11113	0.125454	0.11427	0.214865	0.545411
Low expression	0.504903	0.124287	0.14065	0.1352	0.256773	0.467376

Table 2.2: **DFE-alpha model outputs for each site category**

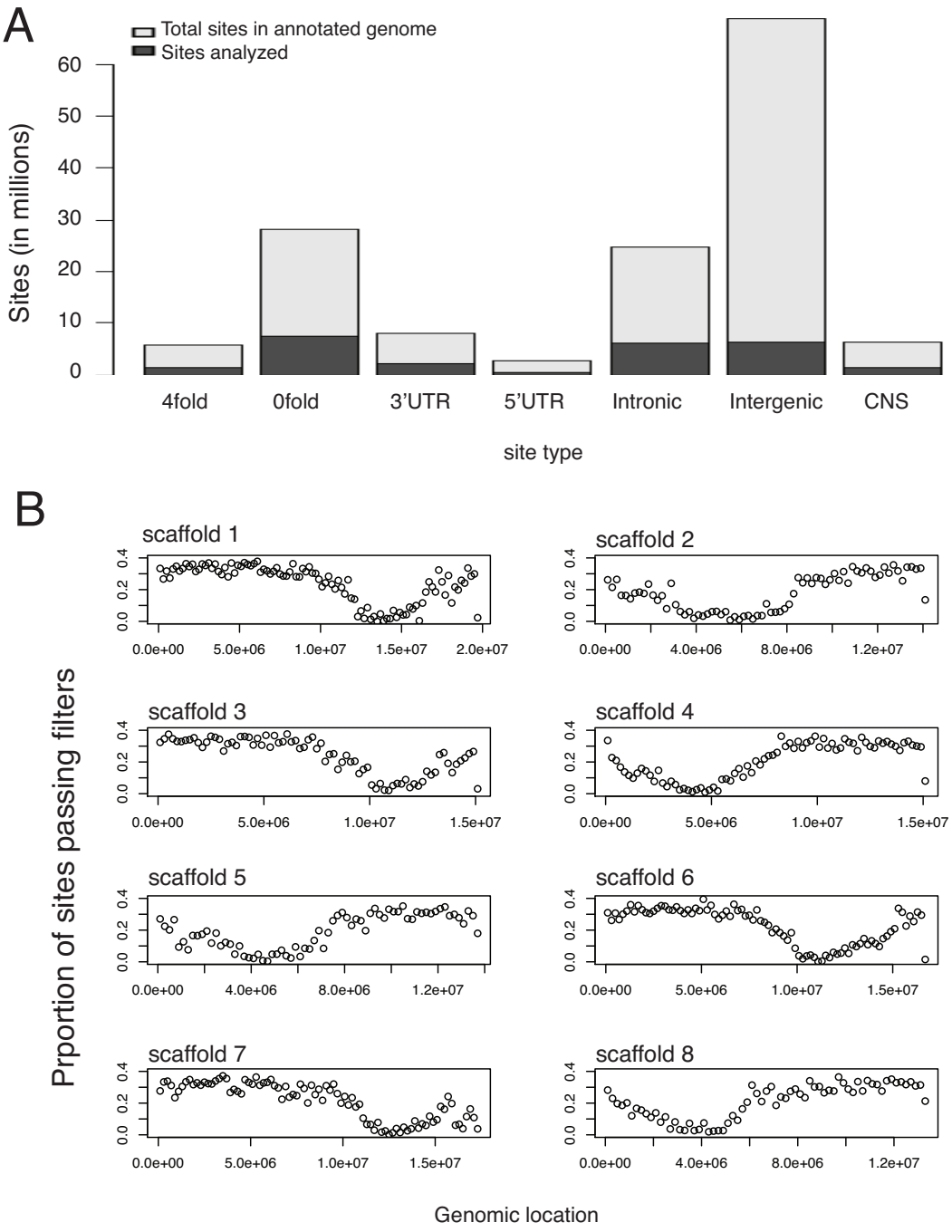

Figure 2.5: **Coverage after filtering, across the genome.** A) The number of annotated sites in each category across the genome (light grey), and the number of sites that pass our filters and were used in analysis (dark grey). B) Proportion of sites that pass filters, calculated in 200kb windows, as a function of genomic position.

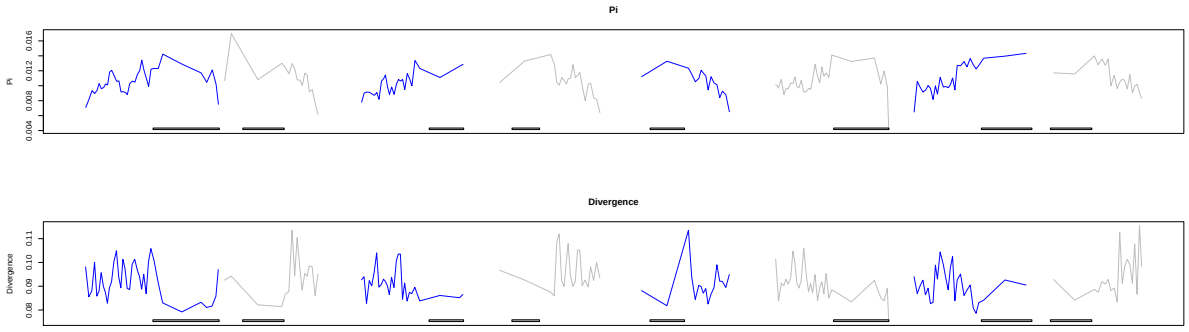

Figure 2.6: **Pairwise diversity and divergence at 4-fold degenerate sites across the entire genome.** Statistics were calculated in windows of 5,000 SNPs. Individual lines alternating between grey and blue represent chromosomes. The location of the centromere on each chromosome is indicated by the grey box along the x-axis.

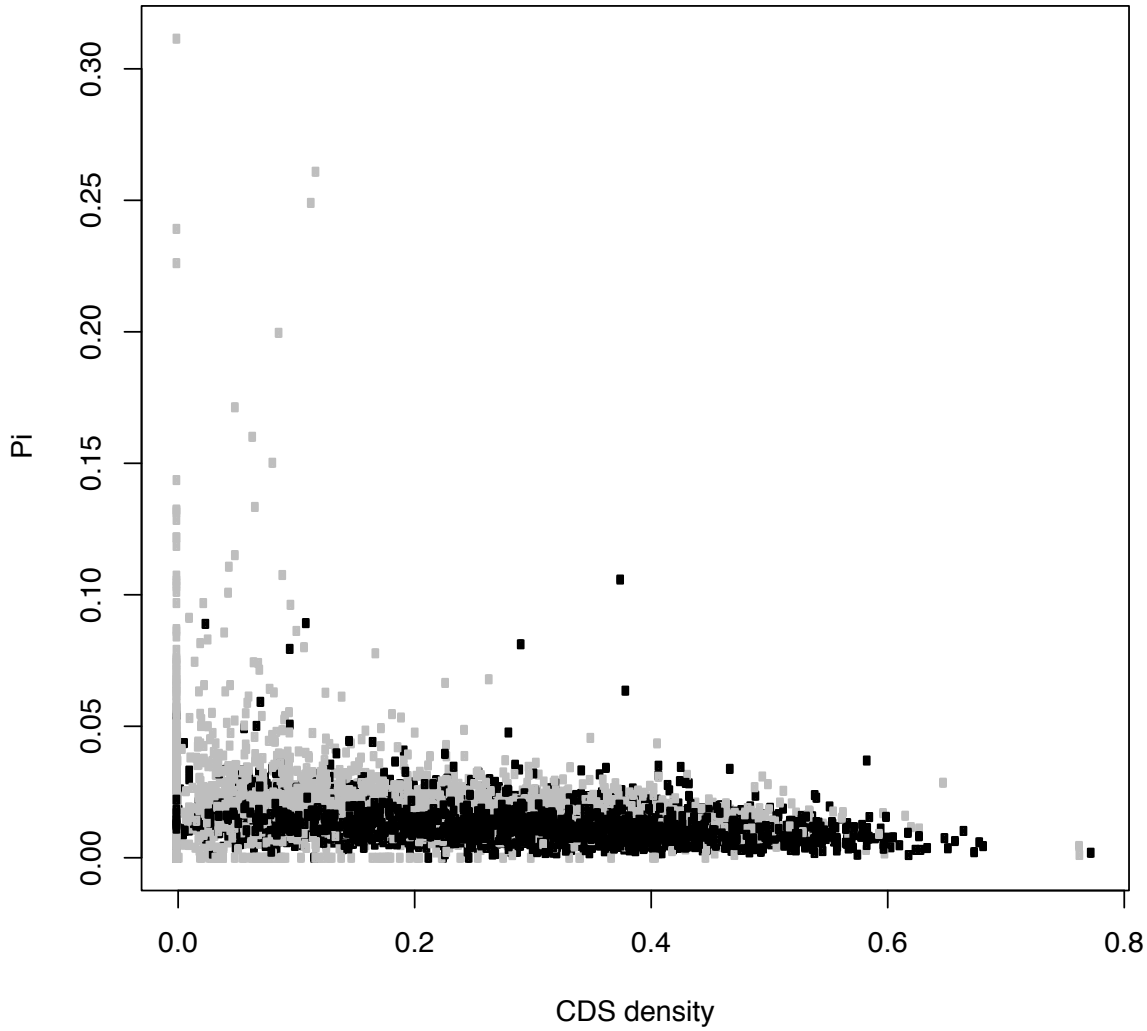

Figure 2.7: **Coding density versus 4-fold degenerate diversity across the genome.** Each point represents one 10 kb window. Black points represent windows that do not overlap centromeres while grey points represent windows that do overlap centromeres. There is a slight negative correlation between diversity and coding density both with and without centromeric windows

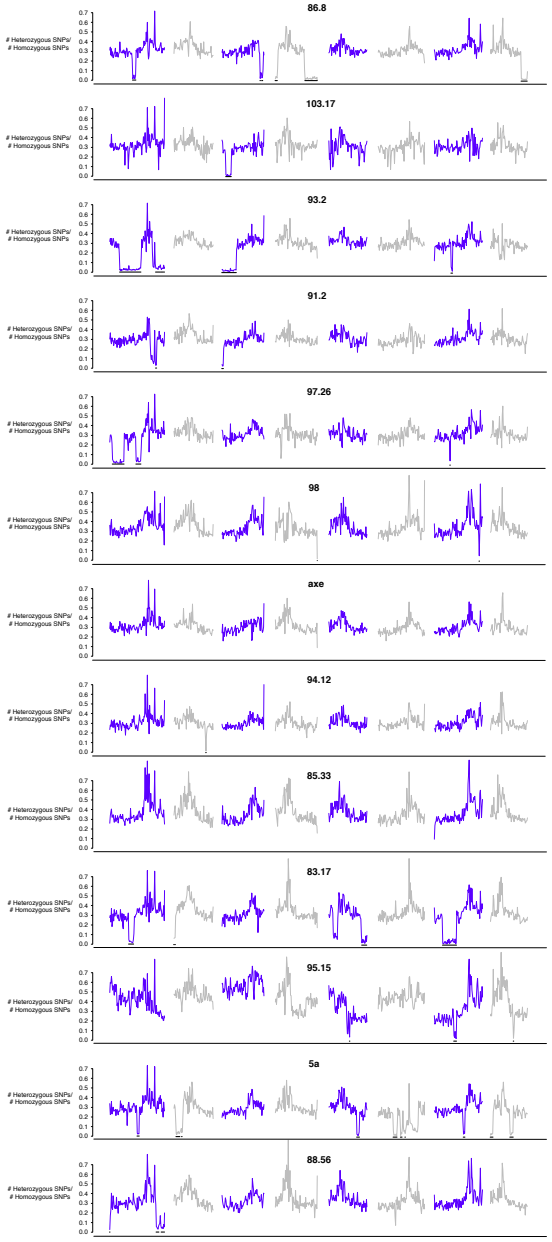

Figure 2.8: **Regions of identity by descent in each sample.** The ratio of heterozygous to homozygous calls at sites that are polymorphic across individuals (in 200kb windows) plotted against position across the genome. Each sample is plotted separately and identified by sampled IDs. Individual lines alternating between grey and blue represent chromosomes. Regions of IBD were defined as windows where FIS was greater than 0.5 and are indicated by black lines along the x-axis. At most 3 regions of IBD overlap across all individuals. This occurs near the end of chromosome 1.

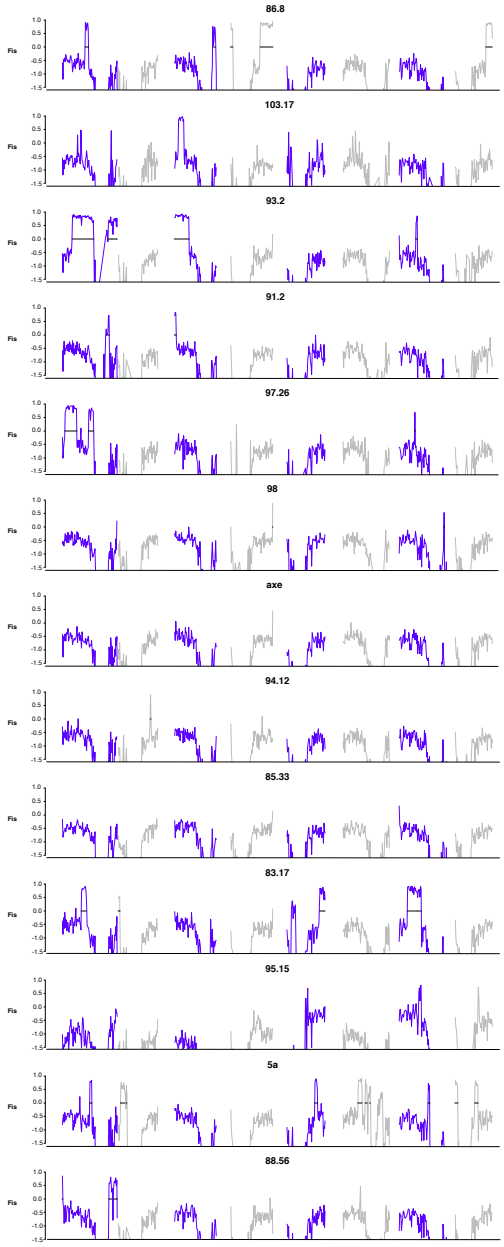

Figure 2.9: **FIS in windows across the genome in each sample.** FIS in 200kb windows is plotted across the genome. Each sample is plotted separately and identified by sample IDs. Individual lines alternating between grey and blue represent chromosomes. Regions of IBD were defined as windows where FIS was greater than 0.5 and are indicated by black lines along the 0 line of the y-axis.

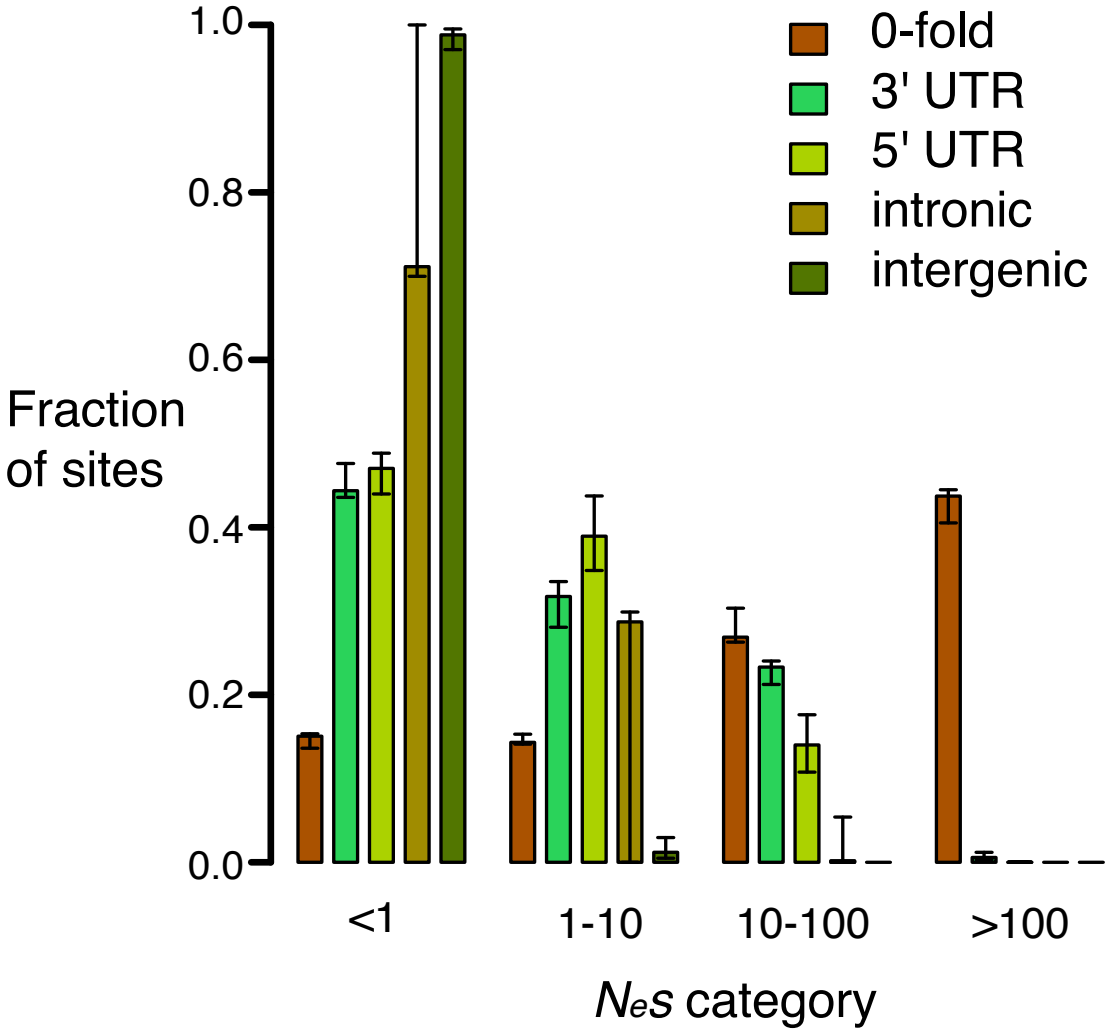

Figure 2.10: **DFE-alpha results using all alleles, including IBD regions.** The distribution of fitness effects for 0-fold degenerate, 3 and 5 UTR, intronic, and intergenic sites are shown. For this analysis the genotyping calls were filtered as described in the methods, but the data was not downsampled in regions of IBD identified in Fig. 2.8.

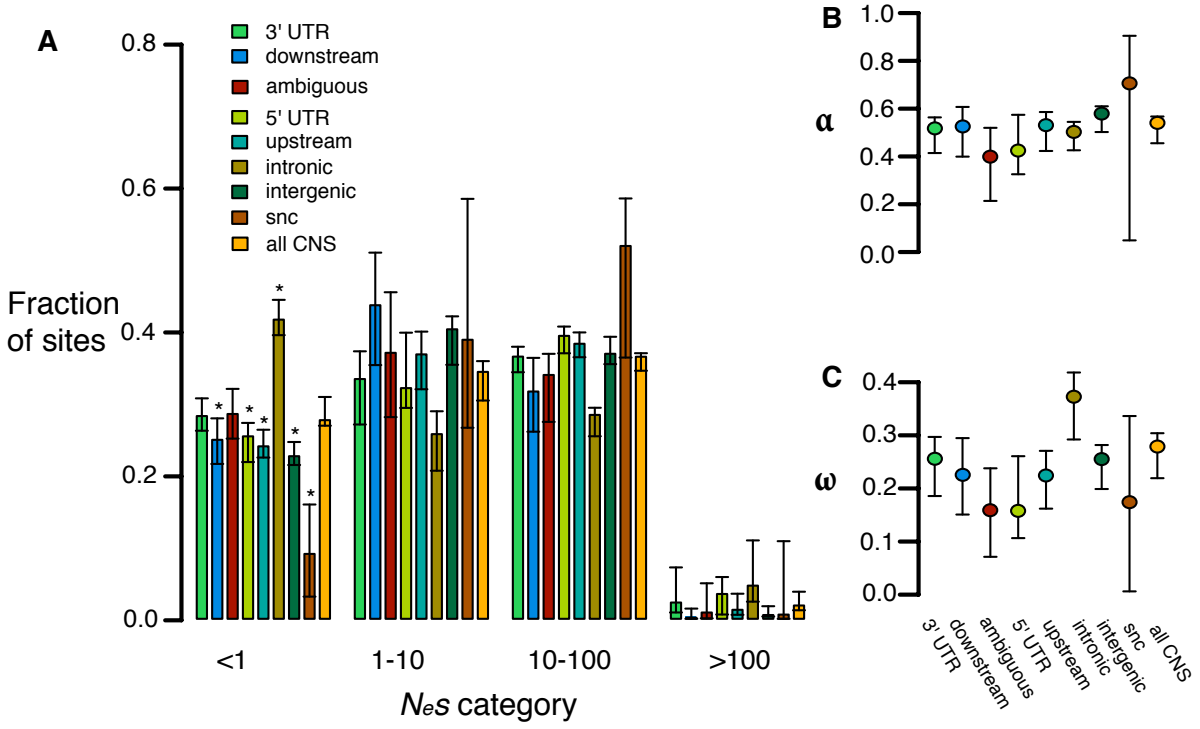

Figure 2.11: **Estimates of positive and negative selection on different categories of CNSs.** A) Distribution of fitness effects. Stars indicate categories in which the fraction of nearly neutral sites was significantly different from the pooled sets of CNSs by a randomization test. B) α and C) ω for each category. Error bars indicate 95% CIs from 200 bootstraps.

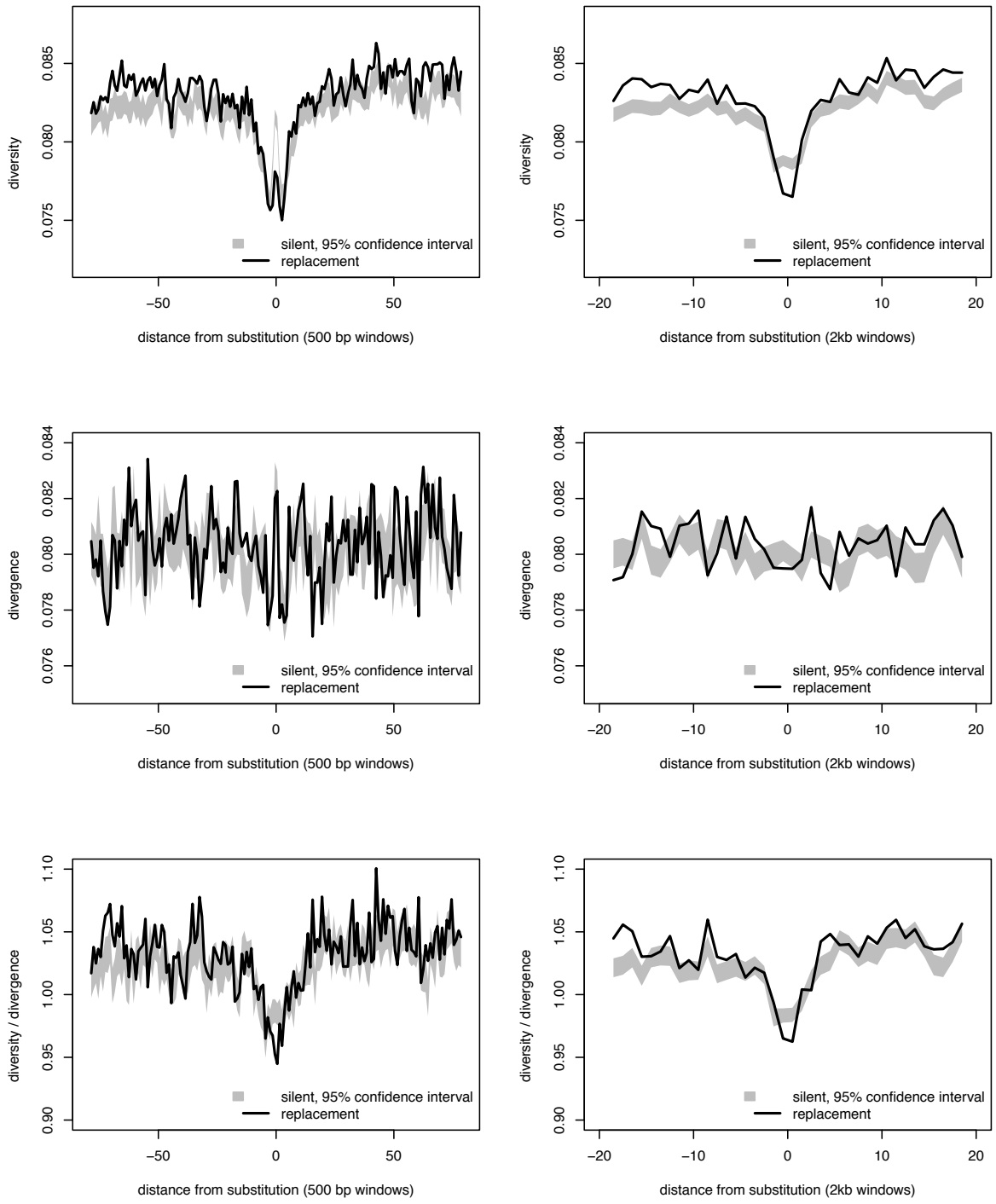

Figure 2.12: **Robustness of sweep analysis to different window sizes.** This panel shows the results of our scans for recurrent selective sweeps using alternative window sizes: 500bp on left and 2kb on right. Otherwise, the methods are the same as described previously.

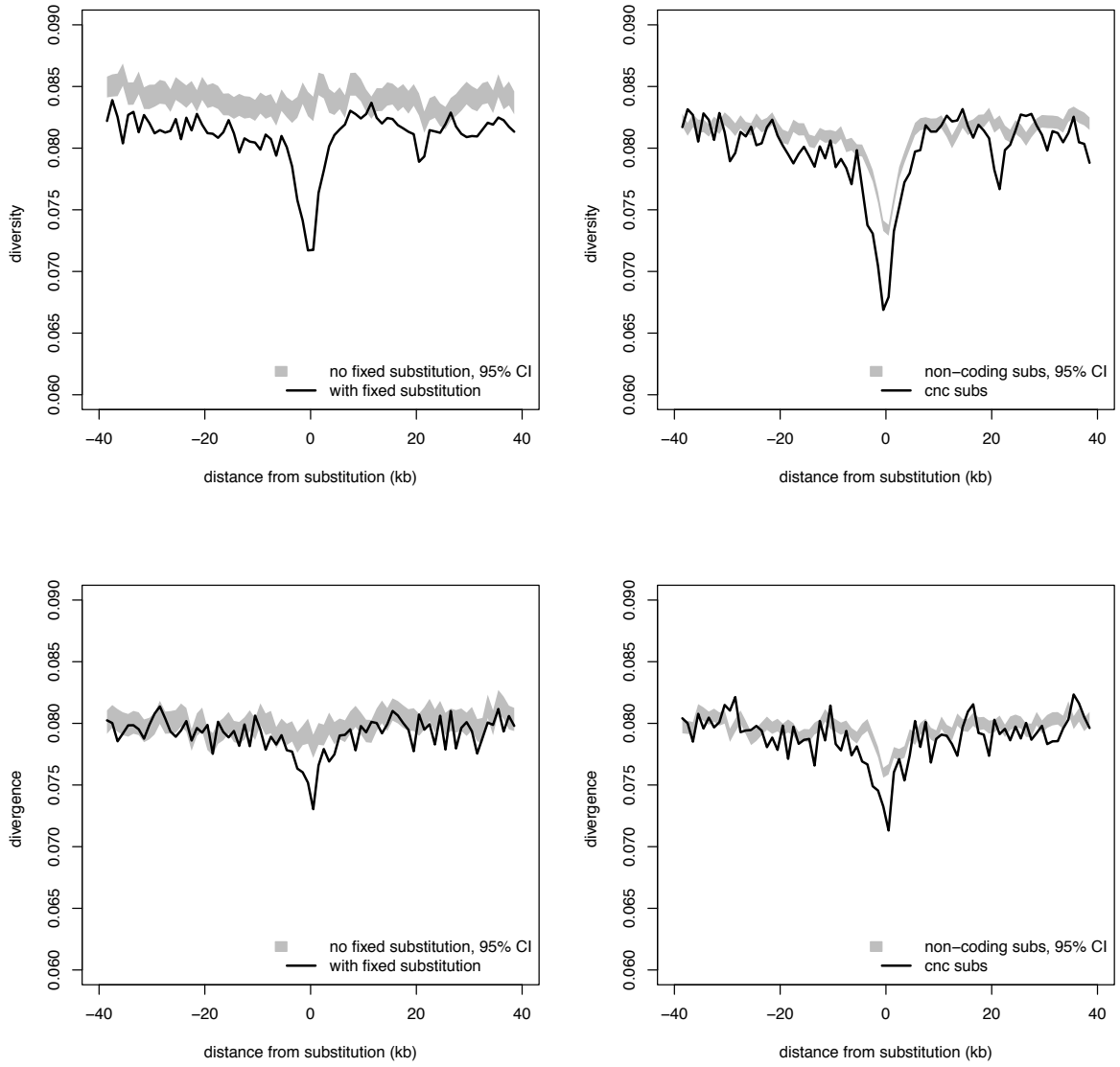

Figure 2.13: **Additional diversity and divergence data for sweeps around substitutions in conserved noncoding regions.** The left panels show diversity at 4-fold degenerate sites and divergence at 4-fold degenerate sites around substitutions in conserved non-coding sequence (black lines) and non-conserved intergenic sequence (gray shading represents 95% confidence intervals). The right panels show the same information for diversity and divergence at 4-fold degenerate sites around conserved noncoding sequences containing fixed substitutions (black lines) and conserved noncoding sequences without fixed substitutions (gray shading represents 95% confidence intervals).

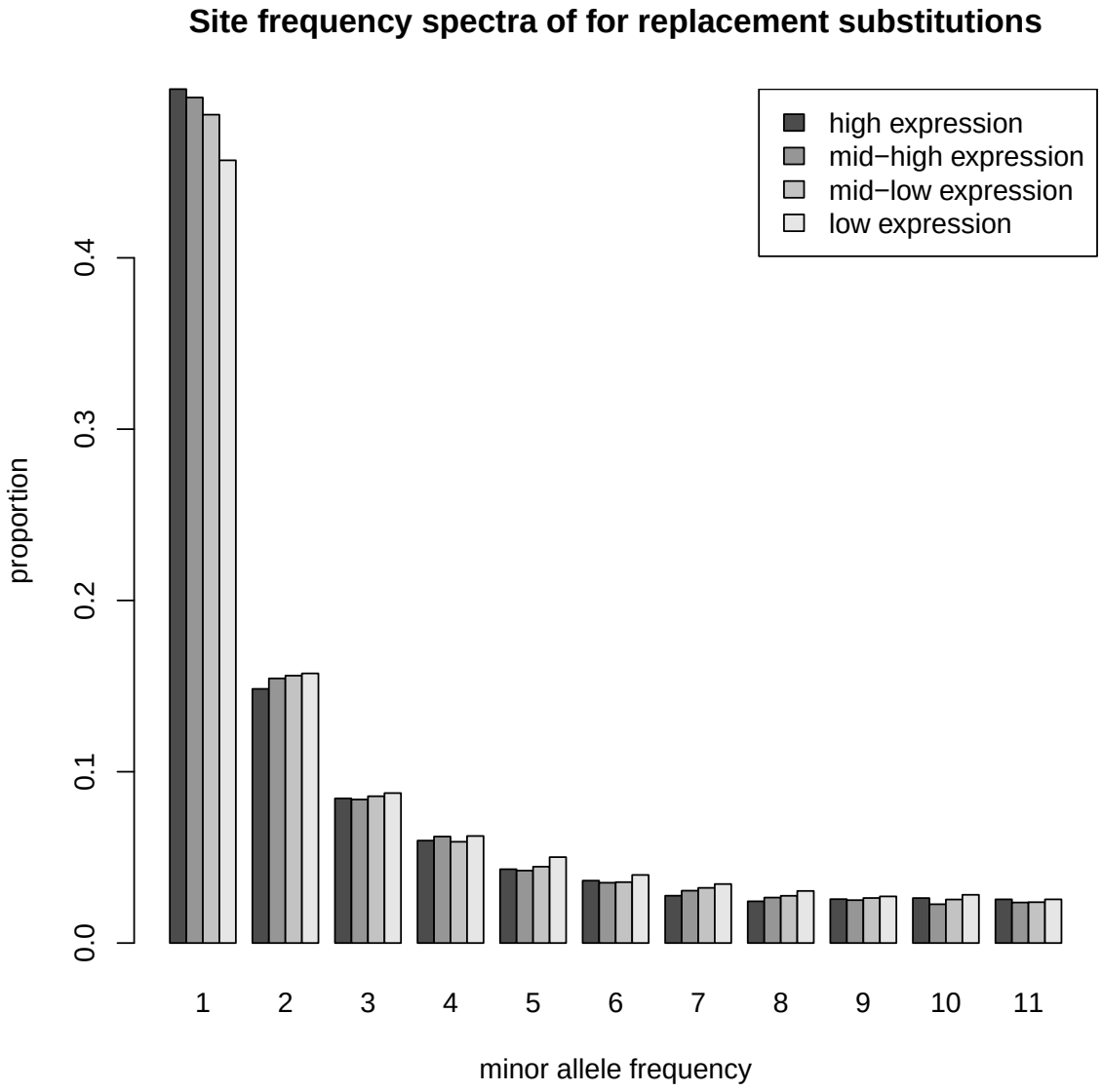

Figure 2.14: **Allele frequency spectra of replacement sites in genes with different expression levels.**

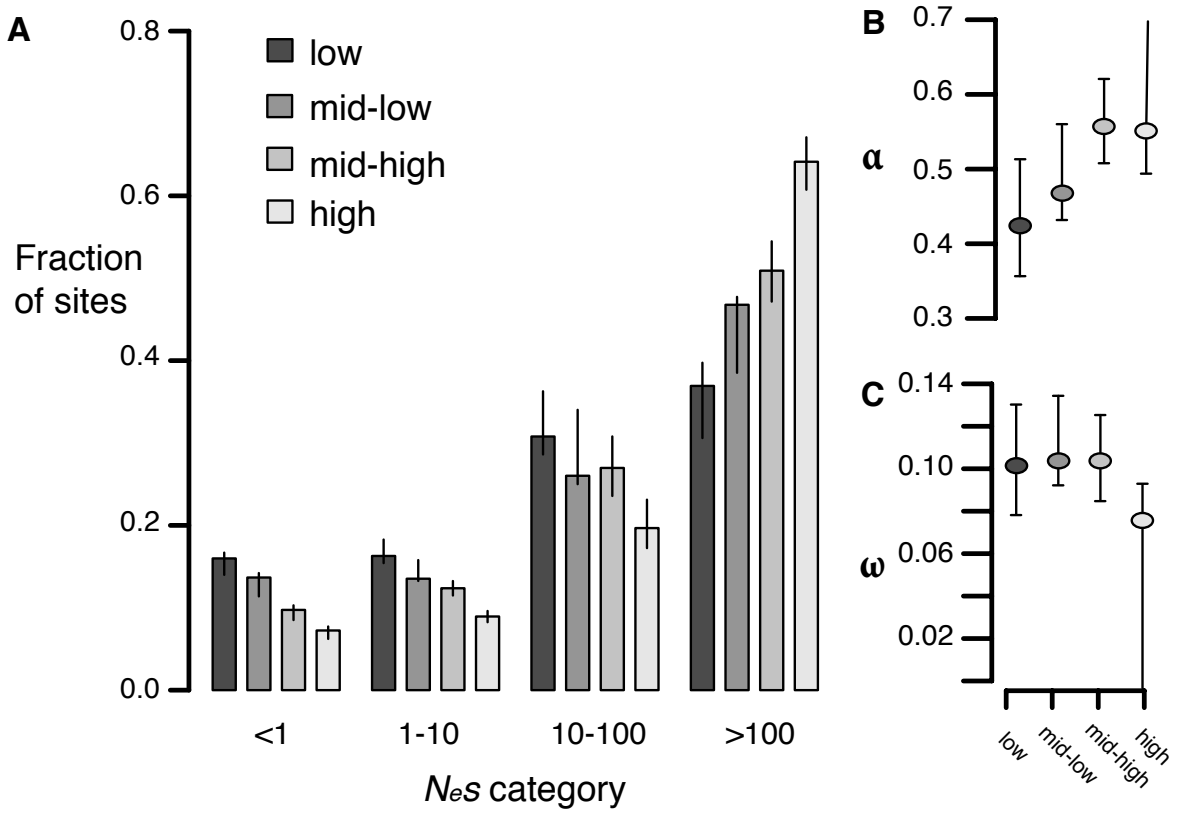

Figure 2.15: **Estimates of negative and positive selection on 0-fold sites in genes of varying expression level.** Data from this figure was generated using the divergence estimates from the whole genome alignments (as in Fig. 2.1) rather than divergence from PAML estimates (as in Fig. 2.4). Here AFS from 0-fold sites were compared to 4-fold sites, rather than non-synonymous to synonymous sites as in Fig. 2.4. A) The proportion of sites found in each bin of purifying selection strength, separated by expression level. B) The proportion of divergent sites fixed by positive selection and C) The rate of adaptive substitution relative to neutral divergence. Error bars represent 95% bootstrap confidence intervals.

Chapter 3

Mutation-selection balance maintains gene expression variation

3.1 Abstract

The evolutionary forces that maintain genetic variation for quantitative traits within populations remain poorly understood. One hypothesis suggests that variation is maintained by a balance between new mutations and their removal by selection and drift. Theory predicts that this mutation-selection balance will result in an excess of low-frequency variants and a negative correlation between minor allele frequency and selection coefficients. Here, we test these predictions using the genetic loci associated with total expression variation (eQTLs) and allele-specific expression variation (aseQTLs) mapped within a single population of the plant *Capsella grandiflora*. In addition to finding eQTLs and aseQTLs for a large fraction of genes, we show that alleles at these loci are rarer than expected and exhibit a negative correlation between phenotypic effect size and frequency. Overall, our results show that the distribution of frequencies and effect sizes of the loci responsible for local expression variation within a single, outcrossing population are consistent with mutation-selection balance.

3.2 Introduction

Genetic variation for quantitative traits persists within populations despite the expectation that prevalent stabilizing selection will reduce genetic variance. One hypothesis suggests that variation is maintained by a balance between new mutations and their removal by selection and drift, resulting in an excess of low-frequency variants and a negative correlation between minor allele frequency and selection coefficients (Haldane, 1927). While studies of allele frequency spectra show that purifying selection is often prevalent in genomic sequence (Kousathanas *et al.*, 2011; Zhu *et al.*, 2011; Williamson *et al.*, 2014), little is known about how the genetic variants under selection relate to phenotype, and ultimately, how phenotypic variation is maintained within populations. Association mapping can identify specific loci influencing

phenotype providing candidates for further analysis of selection (Lee *et al.*, 2014). In particular, mapping the local regulatory variants that affect gene expression can identify a large number of genetic loci that affect phenotype. Additionally, mapping the genetic basis of gene expression will answer questions about the basic biology of gene regulation, for example, by testing predictions that conserved non-coding sequences (CNSs) are constrained because they have regulatory function (**?**).

Early eQTL studies mapped expression divergence between two lines, finding that many genes have local expression QTL (7, 8). These studies have provided insight into selection on eQTLs; for example, a correlation between recombination rate and eQTL density implies that background selection is a dominant force acting on expression variation in Caenorhabditis elegans (9) and a skew towards rare allele frequencies in promoters of genes with eQTLs suggests that purifying selection may act on expression variation (10). However, eQTL studies of population-level genetic variation have thus far been limited to a few study systems (1115) and only one study, in humans, has identified a negative correlation between phenotypic effect size and frequency(14). In addition, human eQTL studies have shown that loci expected to be involved in selective sweeps are more likely to be eQTLs than other loci(16), allele frequencies of eQTLs that increase expression of a potentially deleterious coding SNP are under stronger purifying selection than those that do not (17), and eQTL allele frequencies within populations are linked to local adaptation(18, 19). To date, eQTL studies in plants have used genetic crosses (2022) or species-wide samples (2325), making it difficult to distinguish evolutionary forces acting within and between populations. In sum, we currently lack comprehensive tests of selection on within-population eQTLs in any system, especially in plants.

Here, we map local regulatory loci affecting expression in 99 members of a single large population of *Capsella grandiflora* (Brassicaceae), an obligate outcrosser. As might be expected from its large Ne and relative lack of population structure, purifying and positive selection are strong in *C. grandiflora*(3, 26), making it an ideal system for investigating the maintenance of genetic variation in the face of selection

3.3 Results and Discussion

We sequenced 22,895,738,517 100bp paired-end reads of DNA from 188 individuals, with a median of 119,321,591 reads per individual. Of these reads, a median of 93% mapped per individual (range: 51%-93%, the two individuals with <80% were not sampled for RNAseq). We called 9,526,786 SNPs with a mean depth of 45 reads per individual. Linkage disequilibrium between SNPs decays rapidly: mean R2 between SNPs less than 10bp apart is 0.25, and this decays to 0.12 within 100bp (Fig. 3.6). An analysis of population structure (27) found that the maximum likelihood number of populations was K=1, suggesting no widespread structure. We measured genome-wide gene expression in 99 of these individuals using RNAseq from young leaf tissue, generating 4,988,540,400 100bp paired-end RNAseq reads with a median of 49,549,336 reads per individual (range: 42,627,096-106,283,910). Of these, a median of 94% (range: 89-95%) mapped to genes (Table S1).

We mapped eQTLs by performing Mann-Whitney U tests comparing expression between individuals homozygous for the most common allele at a given SNP and those heterozygous at that SNP, for all SNPs within 5kb of the transcription start and end sites (Fig. 3.1). We omitted rare homozygotes

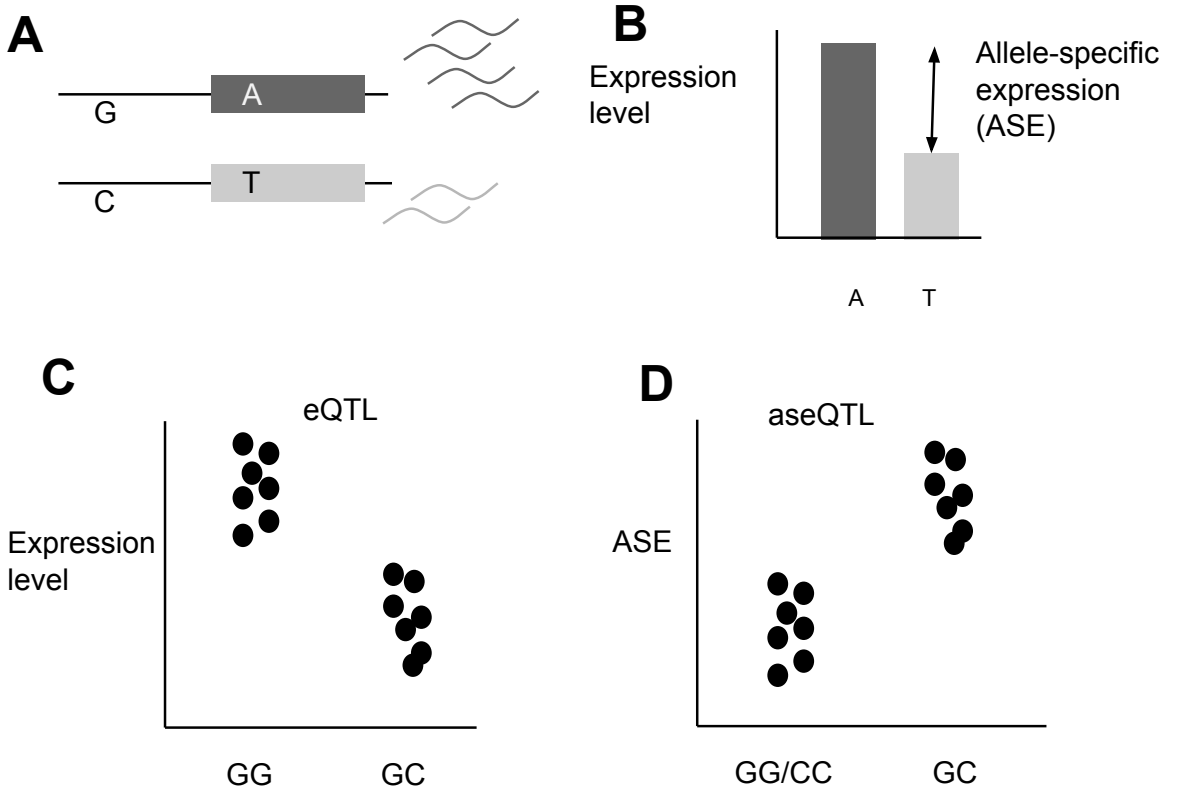

Figure 3.1: **Detecting eQTLs and aseQTLs** (a) A gene model for an individual that is heterozygous at a regulatory locus (G/T) and at an informative coding site (A/T). The G allele increases expression relative to the C allele, (b) causing increased allelic expression of the reads carrying the A allele at the informative heterozygous site. We refer to this difference in allelic expression as ASE. (c) eQTLs are detected when there is a significant difference in total gene expression between individuals (represented by black circles) that are homozygous for the common allele of a SNP and individuals that are heterozygous at that SNP. (d) aseQTLs are detected when there is a significant difference in ASE between individuals that are heterozygous at a SNP and homozygous for either allele at that SNP.

from the analysis because most local regulation acts additively in *cis* (12) and low sample sizes for rare variants reduce power. Out of 5,507,316 SNPs tested against the expression of 18,692 genes, 39,628 SNPS are significantly associated with expression of 6,624 nearby genes (FDR = 0.1, p $<8.2 \times 10^{-4}$, Fig. 3.7 A). These SNPs often clustered locally (Fig. 3.8 A,B), as would be expected if non-causal SNPs are in linkage disequilibrium with causal SNPs. Patterns of functional enrichment in human eQTLS suggest that SNPs most strongly associated with expression are more likely causal than those showing weaker associations(13), so to prevent variation in linkage disequilibrium from affecting subsequent analyses while increasing the likelihood of retaining causal SNPs, we chose the most significantly associated SNP for each gene for further analysis (N = 6,624). While there are likely multiple causal eQTLs for many genes, choosing one significant SNP per gene allows us to generate a large independent sample of eQTLs for further analysis.

If eQTLs act in *cis*, heterozygous eQTLs will cause allele-specific expression (ASE), providing an addi-

tional signature of regulatory variation. We measured ASE within individuals by calculating the mean expression difference between alleles, standardized for sequencing depth. We then mapped QTLs for ASE (aseQTLs) by performing Mann-Whitney U tests comparing ASE in individuals that were homozygous at a local SNP and those that were heterozygous at that SNP (Fig. 3.1). We excluded coding SNPs from this analysis because their genotype might confound ASE measurement. Out of 3,966,423 SNPS tested, 26,957 SNPs were significantly associated with ASE of 5,882 nearby genes (FDR = 0.1, p <5.4 x 10^{-4}, Fig. 3.6 B). Our analysis did not require a directional effect of SNP genotype on ASE, but 22,436 (83%) of the noncoding SNPs associated with ASE have higher ASE in heterozygotes, as would be expected if these SNPs control expression in *cis*. We selected the most strongly associated noncoding SNP per gene for further analysis and we also required that ASE had to be higher in heterozygotes at that SNP than homozygotes, leaving 4,580 aseQTLs (Fig. 3.7 B).

SNPs located near the transcription start site (TSS) and in 5 UTRs were more likely to be eQTLs and aseQTLs than SNPs further away from the gene (Fig. **??**A), consistent with data from humans and Drosophila (11, 12, 14). In addition, CNSs near the TSS were enriched for eQTLs and aseQTLs relative to non-conserved sites (Fig. **??**A), suggesting that genetic variation within CNSs represents a major source of standing variation in gene expression, although bootstrapped confidence limits for these overlap slightly in aseQTLs. In contrast, CNSs in 5UTRs were not enriched for eQTLs or aseQTLs, consistent with observations that selection strength is relatively similar in conserved and non-conserved sites in these regions(3). However, the detection of a large number of eQTLs outside of conserved regions suggests that regulatory element turnover is common in Brassicaceae (Table S2). There were 2,236 genes that had both eQTLs and aseQTLs, significantly more than expected by chance (X^2 = 471, p <2.2x10-16). Of these 2,236 genes, 411 had the same SNP most significantly associated both with expression and ASE.

Next, we tested eQTLs and aseQTLs for signatures of selection. Purifying selection will reduce the frequency of causal alleles at QTLs, but allele frequency also controls sample size in association studies, affecting QTL detection. Rare alleles have an increased likelihood of false negatives, because of lower power, and false positives, since expression is not normally distributed and an outlier in a small sample is more likely to lead to a positive association than an outlier in a large sample. The increased likelihood of false positives in rare alleles makes evolutionary inferences especially challenging because it mimics the signal of purifying selection.

To generate an appropriate null distribution for QTL allele frequency, we permuted assignments between expression level and genotype for every gene 1000 times and ran eQTL analyses using permuted data. On average, 3,258 SNPs were associated with total expression in our permutations, consistent with an FDR of 0.1, since 39,628 SNPs were associated with the observed data. However, observed eQTLs from un-permuted data were significantly rarer than those found in permuted data (mean N=2,047), consistent with the action of purifying selection (Fig. 3.3A). This observation is conservative, because we have not accounted for reduced power to detect associations on rare alleles. We also investigated permuted aseQTLs, and found on average 3,194 SNPs associated with ASE in each permutation, which is slightly more than expected given our FDR of 10% (26,597 SNPs were associated with ASE in un-permuted data). As with eQTLs, aseQTLs were significantly rarer than those found in permuted data (Fig. 3.3B). These results hold when we designate a random significantly-associated SNP per gene as the eQTL or

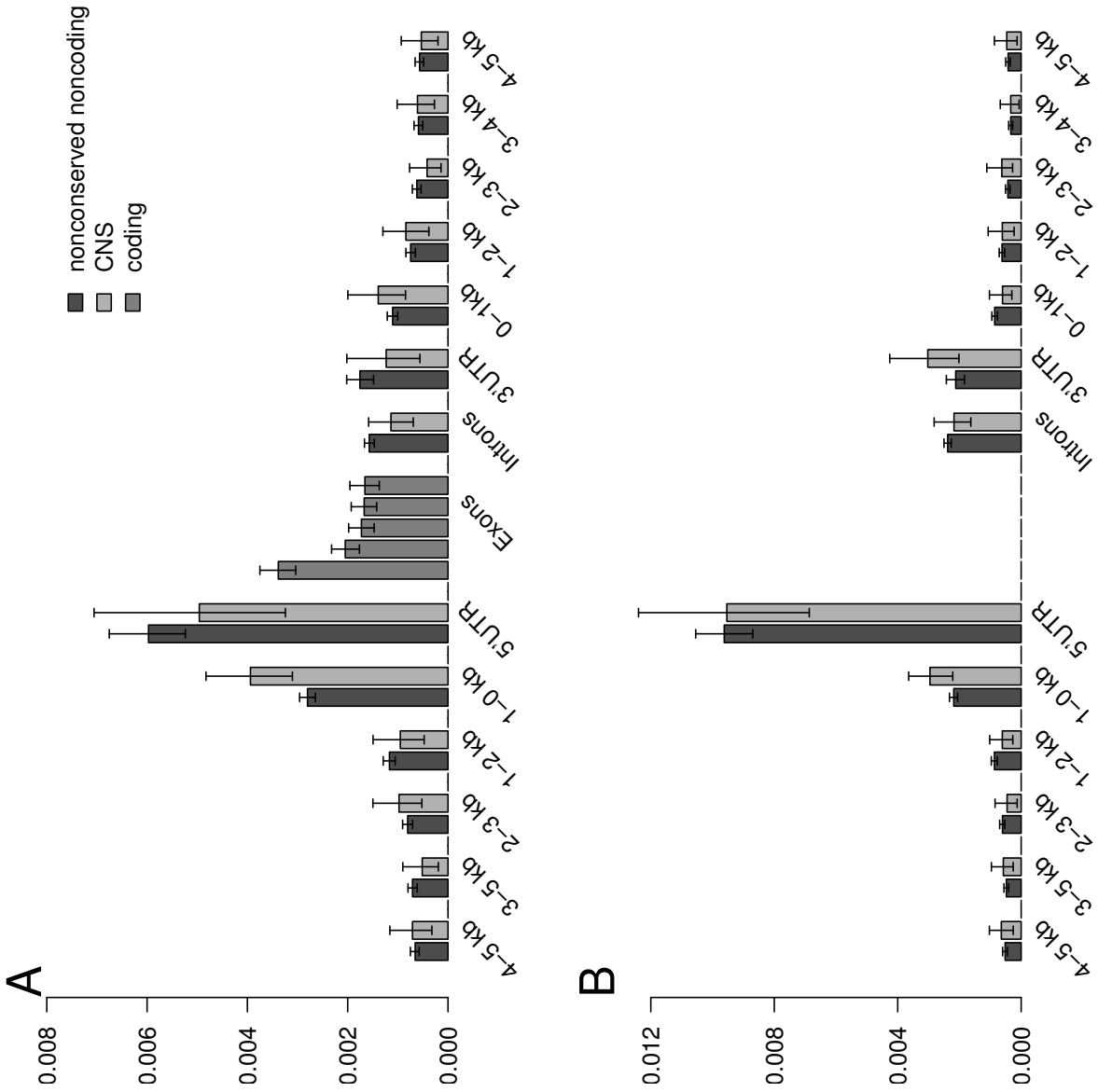

Figure 3.2: **eQTL and aseQTL enrichments by site type.** The proportion of SNPs tested in each category that were found to be eQTLs is plotted on the y axis for (a) eQTLs and (b) aseQTLs. The exonic classes were determined by splitting the coding sequence of each gene into 5 equally sized pieces. Note that there were no exonic SNPs included in the aseQTL analysis. Error bars show the 95% confidence intervals from bootstrapping.

aseQTL (Fig. **??** A,B). In addition, eQTLs and aseQTLs are significantly rarer than permuted eQTLs and aseQTLs when only SNPs 1-5 kb upstream or downstream of genes are considered (Fig. **??** A,B), and when sites are separated into high and low recombination sets or by substitution type (Fig. **??** C,D). Thus, the frequency distribution of both eQTLs and aseQTLs is consistent with the predominance of mutation-selection balance.

We incorporated effect sizes to test for an additional signature of selection. Theory predicts that

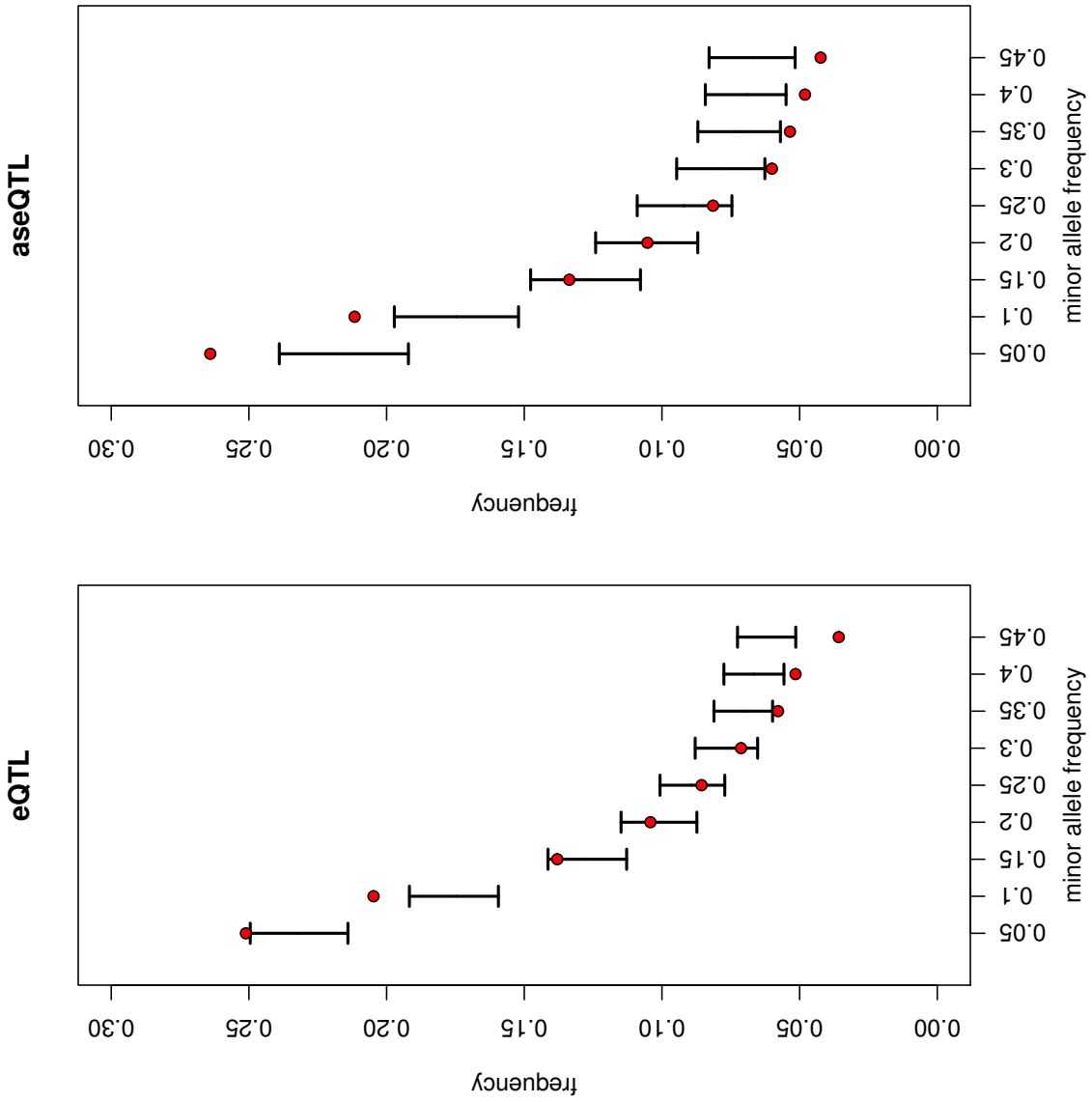

Figure 3.3: **The site frequency spectra of eQTLs and aseQTLs** Minor allele frequencies of (a) eQTLs and (b) aseQTLs for observed data (red circles) and permuted data (gray circles, black lines are 95% confidence intervals).

mutation-selection balance will maintain mutations at frequencies inversely proportional to the strength of selection acting against them (1), suggesting that QTLs under purifying selection should show a negative correlation between minor allele frequency and phenotypic effect size, assuming that phenotypic effect size correlates with the strength of selection. However, this correlation is also expected if QTLs evolve neutrally for two reasons. First, we have low power to detect rare small-effect QTLs. Second, effect size estimation error is greater for rare alleles, and when effect size is over-estimated, an association is more likely due to winners curse, leading to a negative correlation between effect size and minor allele frequency (28).

To avoid variation in power across allele frequency, we repeated the eQTL and aseQTL analysis, downsampling our population to 50 individuals in each test, such that 40 individuals were drawn from the more common genotype and 10 individuals were drawn from the less common type for each SNP tested. As a consequence, for every SNP we test, sample sizes of major and minor genotype classes are equalized regardless of allele frequency in the population. We also measured effect sizes in this subsample to avoid any relationship between allele frequency and effect size estimation error. Despite reducing our sample size by half, we still detected 594 eQTLs and 670 aseQTLs, when using the most significantly associated SNP per gene (above a p-value threshold corresponding to FDR = 0.1; p $<2.6\text{x}10^{-5}$ for eQTLs, p $<8.2\text{x}10^{-5}$ for aseQTLs). In addition, we decoupled the identification of associations from the estimation of effect size by comparing allele frequencies of SNPs identified as eQTLs with these SNPs effects on ASE, avoiding the double-testing issue responsible for winners curse.

Consistent with mutation-selection balance, an eQTLs effect on ASE was negatively correlated with eQTL allele frequency (p <0.05, correlation coefficient = -0.154, n=251) and total expression effect size was negatively correlated with aseQTL allele frequency (p <0.01, correlation coefficient = -0.104, n=670) (Fig. 3.4). eQTL and aseQTL allele frequency were also negatively correlated with the corresponding effect size when we designated a random significantly-associated SNP per gene as the focal eQTL or aseQTL (Fig. 3.9 C,D). One possible explanation for the stronger association between eQTL allele frequency and ASE effect than between aseQTL allele frequency and total expression effect may be that ASE variation results from cis regulatory variation while total expression variation is determined by both cis and trans regulatory variation. Since we only map local QTLs that mainly act in cis, extra noise from trans regulatory variation likely contributes to total expression variation, weakening the association between allele frequency and total expression effect.

We also investigated the allele frequency spectra of the eQTLs and aseQTLs detected using the downsampling approach. In this case it was appropriate to use the frequencies of all SNPs tested as a neutral hypothesis because false positive and false negative rates are independent of minor allele frequency. The minor allele frequencies of the eQTLs and aseQTLs detected with downsampling were rarer than the frequencies of all SNPs tested (Fig. 3.5). The skew towards rare alleles was stronger here than in the QTLs detected with the whole data set, perhaps because the reduced sample size of the downsampling approach allows us only to detect large effect QTLs, which are likely to be under stronger negative selection.

It is important to note that some of our QTLs may not be causal alleles, but are instead in linkage disequilibrium with a causal allele. However, this is unlikely to strongly affect the allele frequencies of the QTLs we detect because the extent of linkage disequilibrium is constrained by similarities in allele frequency since the coefficient of linkage disequilibrium (D) is highest when the frequencies of both loci are similar. Consistent with this inference, power analyses have shown that a causal SNP and a tagging SNP in incomplete LD must have similar allele frequencies for a GWAS to successfully identify an association with the tagging SNP (29). Therefore, our conclusions about the allele frequencies of QTLs should be robust to the inclusion of non-causal linked alleles.

Our mapping of QTLs for expression and allele-specific expression genome-wide in a single population of *C. grandiflora* demonstrates that the frequencies and phenotypic effect sizes of these QTLs are consistent with mutation-selection balance. In addition, the enrichment of eQTLs in CNSs directly upstream of

genes further supports CNSs potential role as regulatory elements; however, the large number of QTLs discovered outside of conserved regions suggests significant turnover in regulatory elements between species. Alternatively, QTLs may create new deleterious regulatory interactions, instead of disrupting conserved functional sites. Taken together, our results, indicate that much of local expression variation observed at the population level is deleterious and support the role of mutation-selection balance in maintaining genetic variation within populations.

3.4 Materials and Methods

3.4.1 Study system and plant material

Capsella grandiflora is an obligately outcrossing member of the Brassicaceae family with a large effective population size (Ne∼600,000), relatively low population structure and a range that spans northern Greece and southern Albania(26, 30). In June 2010, we collected seeds from approximately 400 plants growing in a roadside population of *C. grandiflora* near Monodendri, Greece (Population Cg-9(30)). We germinated and grew one individual from each parent in the University of Toronto greenhouses and performed crosses between independent random pairs of plants to generate the seeds used in this study. By growing the parents in a common environment and then assaying their progeny in a common environment, we reduced the influence of maternal effects and unknown micro-environmental effects on gene expression.

Approximately 10 seeds from each cross were sterilized in 10% bleach followed by 70% ethanol, placed on sterile plates filled with 0.8% agar with Mursashige-Skoog salts (2.15 g/L), stratified in the dark at 4^oC for one week, and then allowed to germinate in a growth chamber at 22^oC and 16 hour photoperiod. After one week, we transplanted two of the seedlings from each cross into 4 inch pots filled with ProMix BX soil and returned the pots to the growth chamber. After another week, pots were thinned down to one seed per cross. Throughout the experiment, pots were randomized once every week to minimize location effects.

Leaf tissue from young leaves was collected for RNA extraction four weeks after transplanting and immediately flash frozen in liquid nitrogen. RNA was extracted using plant RNA extraction kits (Sigma) from 2 or 3 samples from each plant. The extracted RNA was quantified with a Qubit spectrophotometer and the samples from each plant were pooled such that each pool contained the same amount of RNA from each sample. RNA was sequenced at the Genome Quebec Innovation Centre on two flow cells with 8 samples per lane. Reads were 100bp long and paired end. We extracted DNA from leaf tissue using a CTAB based protocol. Whole genome sequence from each individual was obtained through 100 cycles of paired-end sequencing in a Hiseq 2000 with Truseq libraries (Illumina), with three individuals sequenced per lane.

3.4.2 Genomic data

We mapped DNA sequence data to the C. rubella reference genome(31) with Stampy v1.0.19. After bioinformatic processing with Picard tools, we realigned reads around putative indels with GATK Re-

alignerTargetCreator and IndelRealigner and compressed the resulting bams with GATK ReduceReads. Raw SNP calls were generated by joint calling of all samples in GATK v2.81 UnifiedGenotyper. We subsequently followed GATK Best Practices for Variant Quality Recalibration using a high confidence subset of the raw calls generated by filtering snps for concordance with common variants (minor allele frequency >0.11) in a species-wide sample of *C. grandiflora* (3) as well as suspect realignments (transposable elements, centromeres, 600bp intervals containing extreme Hardy-Weinberg deviations, 1kb intervals with evidence of 3 or more snps in reference-to-reference mapping). A relatedness analysis revealed that six individuals were more related to eachother than expected in an outcrossing population, perhaps because of introgression from *C. rubella*, so we removed these individuals from the analysis. We measured population structure using fastStructure on a set of 56,011 biallelic snps distributed genome wide that had been pruned for LD following the recommended analysis stream (27).

To map RNA reads, we constructed our own codon-only reference sequence by stitching together the exons and UTRs of each gene into a scaffold using reference gene annotations (31). We mapped to this codon-only reference using Stampy 1.0.21 with default settings. We chose to use Stampy over other RNA-specific aligners, like Tophat, because visual examination of alignments showed that Stampy was better at mapping reads containing multiple polymorphisms, reducing the potential for false associations between expression level and the genotypic variants that affect mapping (Fig. 3.11). RNAseq readmapping for two individuals was very poor quality (<10% reads mapped and paired correctly), so these individuals were removed. Our final sample size was 99 individuals.

Expression level was measured with the HTSeq.scripts.count feature of HTSeq, which counts the number of read pairs that map to each gene. We normalized the read counts of each sample for library size by dividing read counts by the median read count of the entire sample. Previous studies on human gene expression have found interactions between GC content, lane, and expression level (12), but we did not detect this (Fig. 3.12). Genes with a median expression level below five reads per individual before normalization were removed from the analysis, leaving a total of 18,692 genes.

3.4.3 Mapping local eQTL

We selected SNPs for our eQTL analysis by finding all SNPs within the window spanning 5 kb upstream of the genes transcription start site and 5kb downstream from the genes transcription end site. We chose the 5kb range because a previous study in Arabidopsis thaliana mapping associations between expression and SNPs within 30kb of the gene found that 87% of local eQTLs were located within 5kb of the gene (23). SNPs were categorized as occurring in 0-fold degenerate sites, 4-fold degenerate sites, 2 or 3-fold degenerate sites, 5'UTRs, 3'UTRs, introns, stop codons, or intergenic regions based reference annotations(31). In addition, we identified SNPs located in non-coding sequence conserved across the Brassicaceae family(3). We only included SNPs with at least 10 heterozygous individuals and 10 individuals that were homozygous for the common allele in our sample.

We wrote set of Python scripts to test for associations between expression level and genotype at a nearby SNP. These scripts are available at https://github.com/emjosephs/eQTL. We mapped eQTLs by conducting a Mann-Whitney U test on the null hypothesis that gene expression does not differ between individuals that were homozygous for the common allele and individuals that were heterozygous. We

used non-parametric statistics because expression data is not normally distributed. We used the Mann-Whitney U test function in SciPy (scipy.stats.mannwhitneyu), which uses a continuity correction and corrects for ties. 8,302 of our genes had ties in expression level between individuals and these ties on average involved 4.5 individuals (32). In addition,we compared common homozygotes to heterozygotes because we expect most local eQTLs to act in cis and thus be additive (12), and because not being limited by the sample size of rare homozygotes allowed us to map eQTL at rarer alleles.

To avoid a relationship between allele frequency and sample size, we conducted a second eQTL analysis where we subsampled 50 individuals for each SNP tested so that 40 individuals had the most common genotypic category (usually the homozygote) and 10 had the less common genotypic category (usually heterozygote) (14). We chose these thresholds because they retained most individuals while allowing us to still test 3,972,771 of the 4,098,832 SNPs originally tested for eQTLs (96.9%).

For both eQTL analyses, we controlled for multiple testing by using a false discovery rate approach (33) and only considered eQTLs to be associated with expression if that association had a p value corresponding to a false discovery rate of ¡0.1. To avoid being biased by detecting multiple SNPs linked to only one causal site, we only selected one eQTL per gene, picking the SNP with the lowest p value for association. However, to investigate whether choosing the most associated SNP biased our results, we also performed all analyses with eQTLs that were randomly chosen from the pool of SNPs significantly associated with expression (FDR = 0.1). We calculated the expression effect size of eQTLs by taking the absolute value of the difference between mean expression in the common homozygote and mean expression in the heterozygote.

3.4.4 Mapping aseQTL

If local eQTLs act in cis, they should have allele-specific effects and individuals heterozygous for an eQTL will show a larger difference in expression between alleles than individuals homozygous for an eQTL. To take advantage of this second signature of expression variation, we developed a method to test for allele-specific expression QTL, or aseQTL (similar approaches have been used in humans (14)). We quantified allele-specific expression at all heterozygous sites inferred from the genomic data. We used the count of reads mapped to each allele, taken from the AD values in a VCF file constructed from the RNAseq data using GATK Unified Genotyper to calculate an allele-specific-expression measure (ASE) for each gene in each individual. Specifically, we calculated the mean of the the differences in allelic expression values at all heterozygous sites across a gene and divided this mean by median expression level of all genes in the individual to control for sequencing depth. While we expected that our measure of gene-wide ASE would be more accurate when we required multiple heterozygous sites per gene, doing so did not strongly alter the number of aseQTLs we found or their allele frequency distribution, so we only required one heterozygous site per gene to measure ASE (Fig. **??**).

ASE measures were not normally distributed, so we used a Mann-Whitney U test to test the null hypothesis that ASE did not differ between individuals that were heterozygous at a given SNP and individuals that were homozygous for either allele at that SNP. As before, we used the mannwhitneyu function in the SciPy package. 8,334 genes had at least one tie between individuals for ASE value and an average of 4 individuals were involved in ties within these genes. We only tested SNPs where we

had 10 individuals that were both heterozygous at the SNP and had a heterozygous marker site in the gene and 10 individuals that were homozygous at the SNP and had a heterozygous marker site in the gene, allowing us to test for associations at 17,880 genes. We designated aseQTLs as the most associated SNP per gene that had higher ASE in heterozygotes for that SNP than in homozygotes for that SNP. However, we also performed all analyses designating aseQTLS as a SNP that was randomly sampled from the set of SNPs that were significantly associated with expression (FDR = 0.1) and had greater ASE in homozygotes for that SNP than heterozygotes.

As in the eQTL analysis, we conducted a second aseQTL analysis where we subsampled 50 individuals for each SNP tested such that 40 had the most common genotypic category (usually the homozygote) and 10 had the less common genotypic category (usually heterozygote). This sample size allowed us to test 3,841,452 of the 3,966,364 SNPs originally tested for aseQTL (96.8%). For both sets of analyses, we conducted a false discovery rate analysis as described in the eQTL section and, selected all SNPS with a p value below the FDR threshold of 0.1, we chose the most significantly associated SNP per gene for further analysis, with the additional requirement that heterozygous individuals have higher ASE than homozygous individuals. We calculated ASE effect size for aseQTLs and eQTLs by taking the difference between mean ASE in homozygotes and mean ASE in heterozygotes. We only report ASE effects for eQTLs located outside the coding sequence of these genes they regulate.

Preferential mapping of reference alleles compared to alternative alleles could lead to spurious ASE. To evaluate the importance of this effect, we simulated all of the possible reads spanning each heterozygous site, containing either the reference allele or an alternate allele using scripts from Degner et al(34). There were up to 200 reads possible for each site, although reads near the start and end of genes had fewer reads covering them since we discarded all reads that were less than 100bp long. We mapped these reads with the same program and settings we used for the real data, with the exception that these reads were single-ended. Out of 2,365,590 SNPs in coding regions, 19,017 SNPs (<1%) had unequal numbers of reads mapping from each allele. 11,339 (60%) of these sites had more reads that mapped with the reference allele than with the alternative allele, suggesting that there is some reference bias. The 19,017 SNPs with evidence of mapping bias occurred in 3,059 genes. Removing these genes from the analysis did not qualitatively affect the minor allele frequency of aseQTLs (Fig. **??**)

3.4.5 Permutation analysis

Conducting millions of tests for genotype-expression associations with a relatively small (n=99) sample size exposes us to two potential sources of bias that correlate with the allele frequency of the SNPs we are testing. First, smaller sample sizes at low frequencies reduce power to detect associations. Second, smaller sample sizes at low frequencies increase our risk of false positives because expression data is non-normally distributed and outliers in a small sample will have a disproportionate effect on the mean(13). We found this second possibility especially concerning because it is not conservative with respect to our hypothesis that purifying selection will maintain eQTLs and aseQTLs at lower allele frequencies.

To ensure that our conclusions about allele frequencies were not due to false positives being more common at low allele frequencies, we compared the eQTLs and aseQTLs we found with those discovered using permuted data. We constructed permutations by randomly shuffling the assignments between genotype

and expression values or allele-specific expression values for each gene. This strategy allows us to retain the allele frequencies and spatial distributions of the SNPs we are testing along with the distribution of expression and allele-specific expression values of each gene. Each permuted set was analyzed using the same methods as the real data, with one exception: instead of calculating a FDR for each permuted data set, we used the p-value cut offs from the real data to identify false-positive eQTLs and aseQTLs in the permuted data. The frequency distributions of these false-positive QTLs were used as a null distribution for the expected frequency of QTLs.

The permutation analyses do not directly control for site type, recombination rate, or other factors that could both bias a SNP towards being an eQTL/aseQTL and reduce allele frequency. To ensure that these effects did not drive our observations, we divided our eQTLs and aseQTLs from the real data and from permuted data into subsets. First, for site type, we selected the most strongly associated eQTL and aseQTL per gene that came from a the site-type of interest. Our site types were 5UTRs, 3UTRs, introns, intronic CNSs, exons (divided into 5 regions based on distance from start and end of the gene), and upstream and downstream CNS and nonconserved regions. For upstream and downstream regions, we divided sites into those within 1 kb of the TSS/TES and those that were 1 to 5 kb from the TSS/TES.

To control for recombination rate, we divided SNPs into those coming from high recombination regions (¿ 3.45 cM/mB) and low recombination regions (¡ 3.45 cM/mB) using recombination rate data calculated by using a genetic map made from a cross between *C. rubella* and *C. grandiflora* (35). To test for confounding effects due to gene conversion, we divided SNPs into those whose mutations could be due to gene conversion (A or T and C or G) and others (A and T or G and C).

3.5 Acknowledgements

We thank Niroshini Epitawalage, Amanda Gorton, and Khaled Hazzouri for lab assistance, J. Paul Foxe for collection assistance, Wei Wang for computer assistance, and Aneil Agrawal, Graham Coop, Asher Cutter, Alan Moses, Adrian Platts, Tanja Slotte, and Robert Williamson for helpful suggestions. Thomas Bureau, Mathieu Blanchette, Daniel Schoen, Paul Harrison, Alan Moses, Adrian Platts, and Eef Harmsen contributed to the Value-directed Evolutionary Genomics Initiative (VEGI) grant (Genome Quebec/Genome Canada) which supported this work, along with an NSF Graduate Research Fellowship to EBJ (DGE-1048376), and NSERC Canada and CFI grants to JRS and SIW.

3.6 Appendix: Supplementary figures and tables

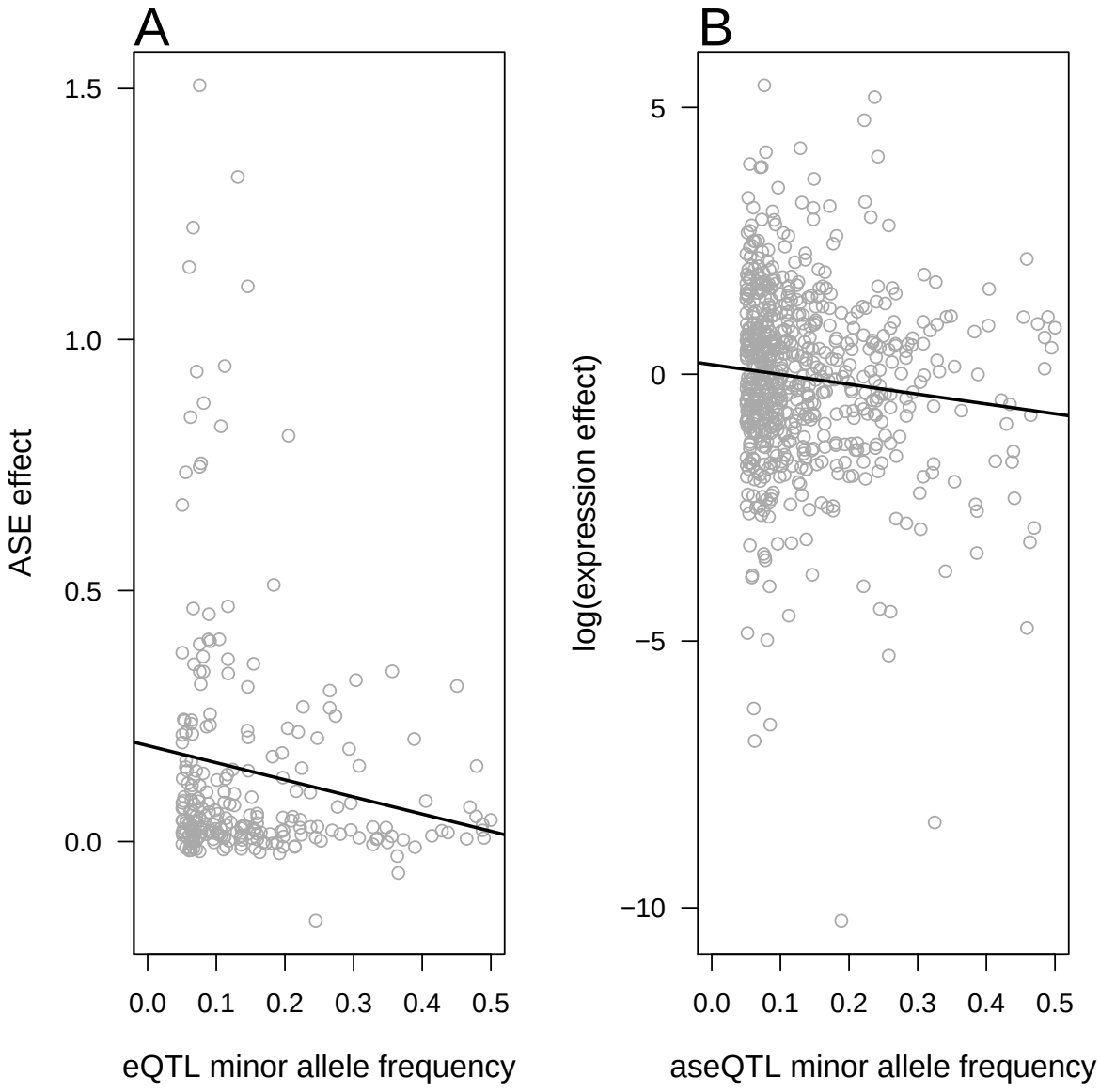

Figure 3.4: **The relationship between minor allele frequency and effect size.** (a) eQTL minor allele frequency is plotted against the effect of that SNP on ASE, calculated as the mean difference in ASE between individuals heterozygous at the eQTL and individuals homozygous at the eQTL. Negative values occur when the the homozygote for the eQTL has greater ASE than the heterozygote. The trend line is calculated by linear regression (b) aseQTL minor allele frequency plotted against the effect of the aseQTL on total gene expression, calculated by taking the log of the absolute value of the mean difference in expression between individuals heterozygous at the aseQTL and individuals homozygous for the common allele at the aseQTL. The trend line was calculated by regression between minor allele frequency and the log of the expression effect.

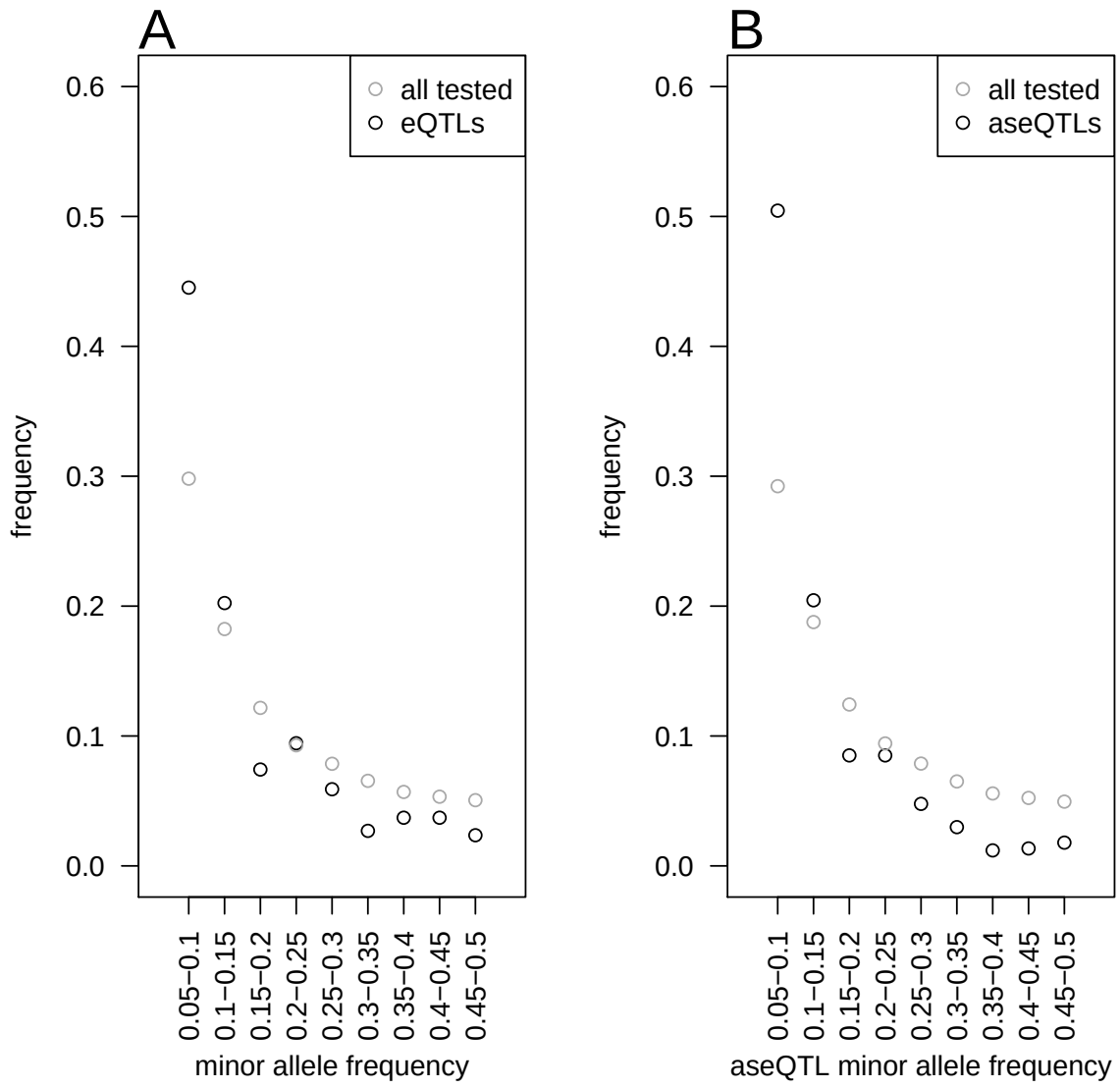

Figure 3.5: **The site frequency spectra of QTLs detected in the frequency-controlled subsample.**

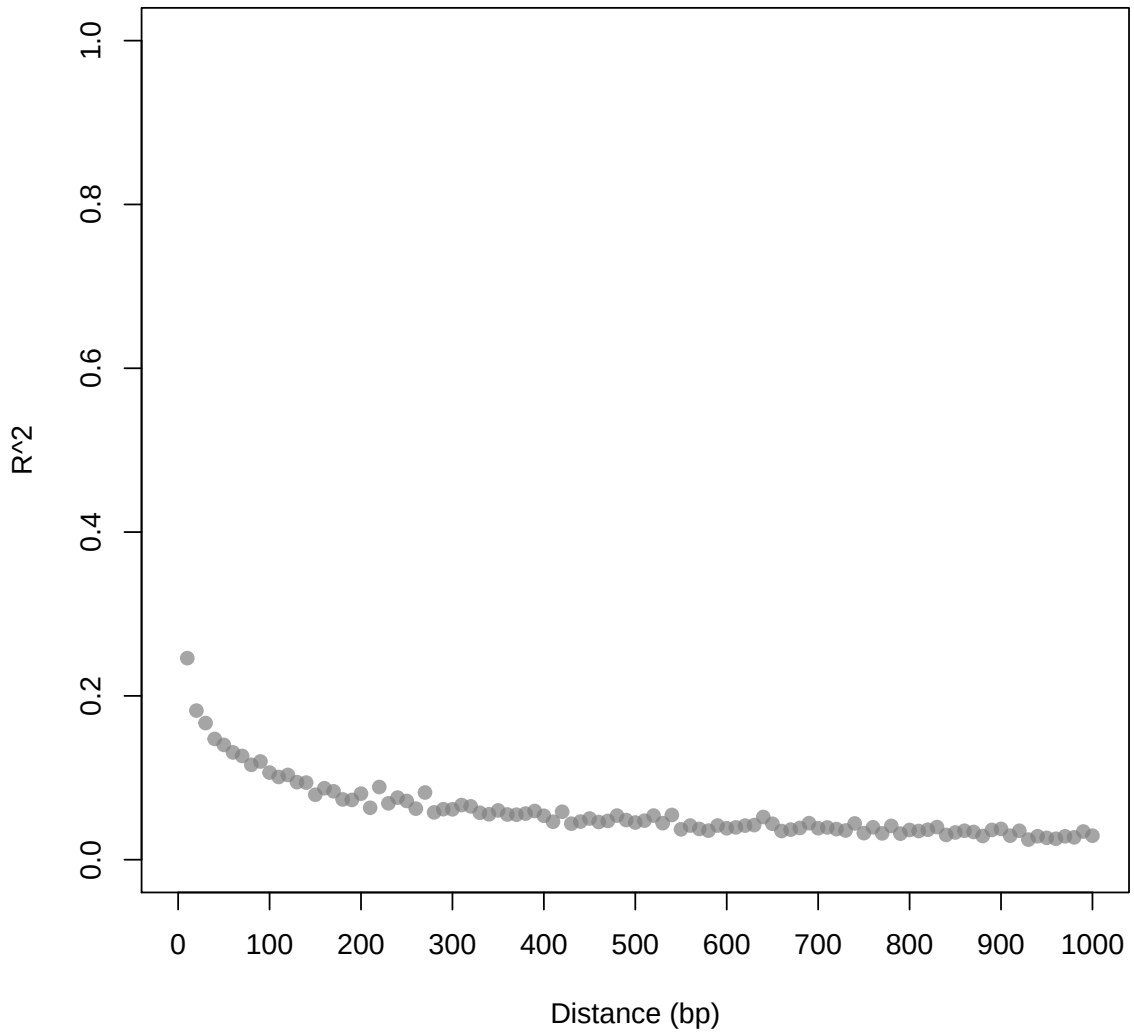

Figure 3.6: **Linkage disequilibrium in *C. grandiflora*.** Linkage disequilibrium was calculated for all SNPs within 1 kb of each other on scaffold 2. 1% of these pairs were randomly sampled for the above figure, which shows mean R2 between pairs in 10bp bins.

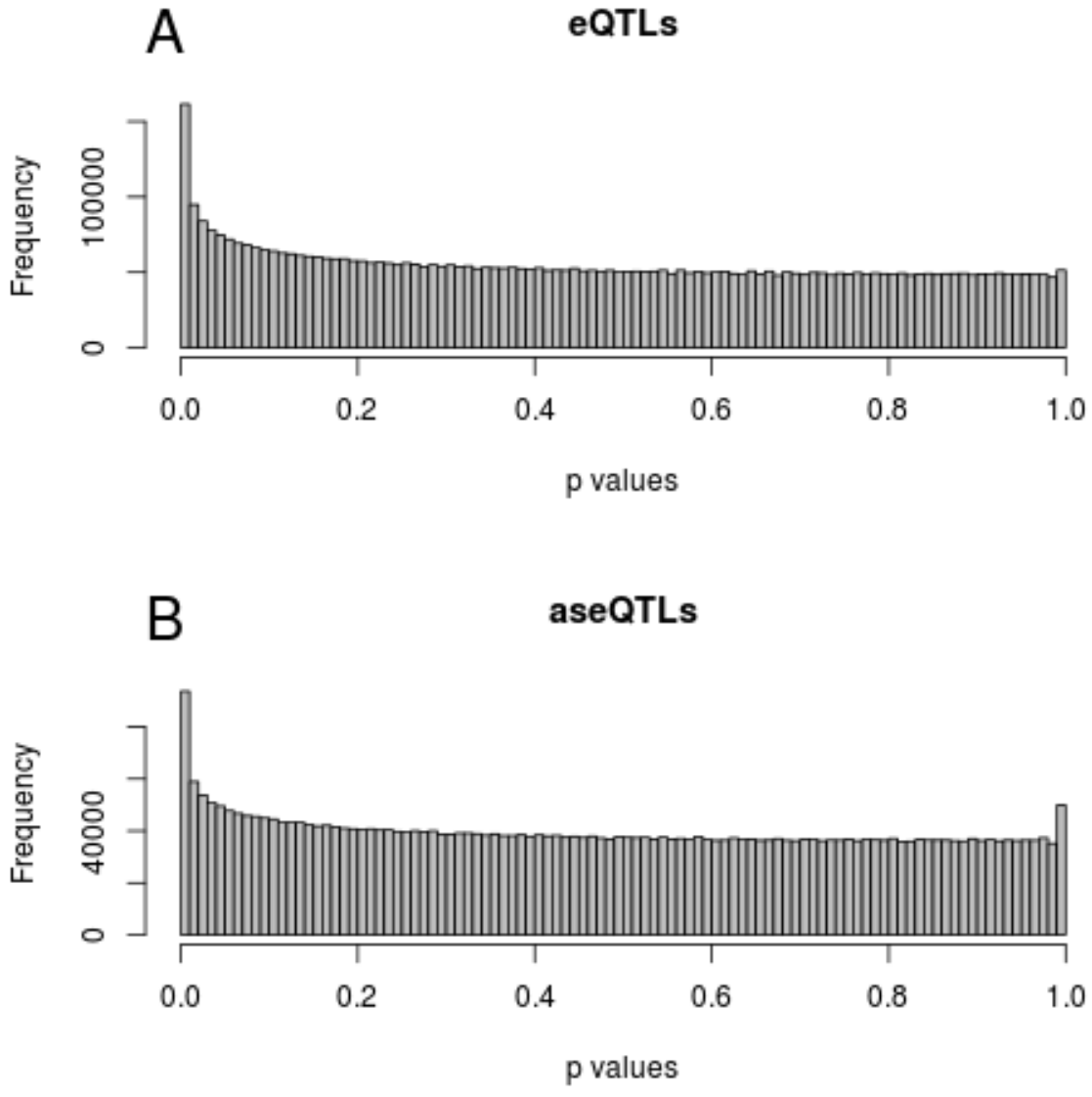

Figure 3.7: **The distribution of p values for all SNPs tested in eQTL analyses (a) and aseQTL analyses (b)**

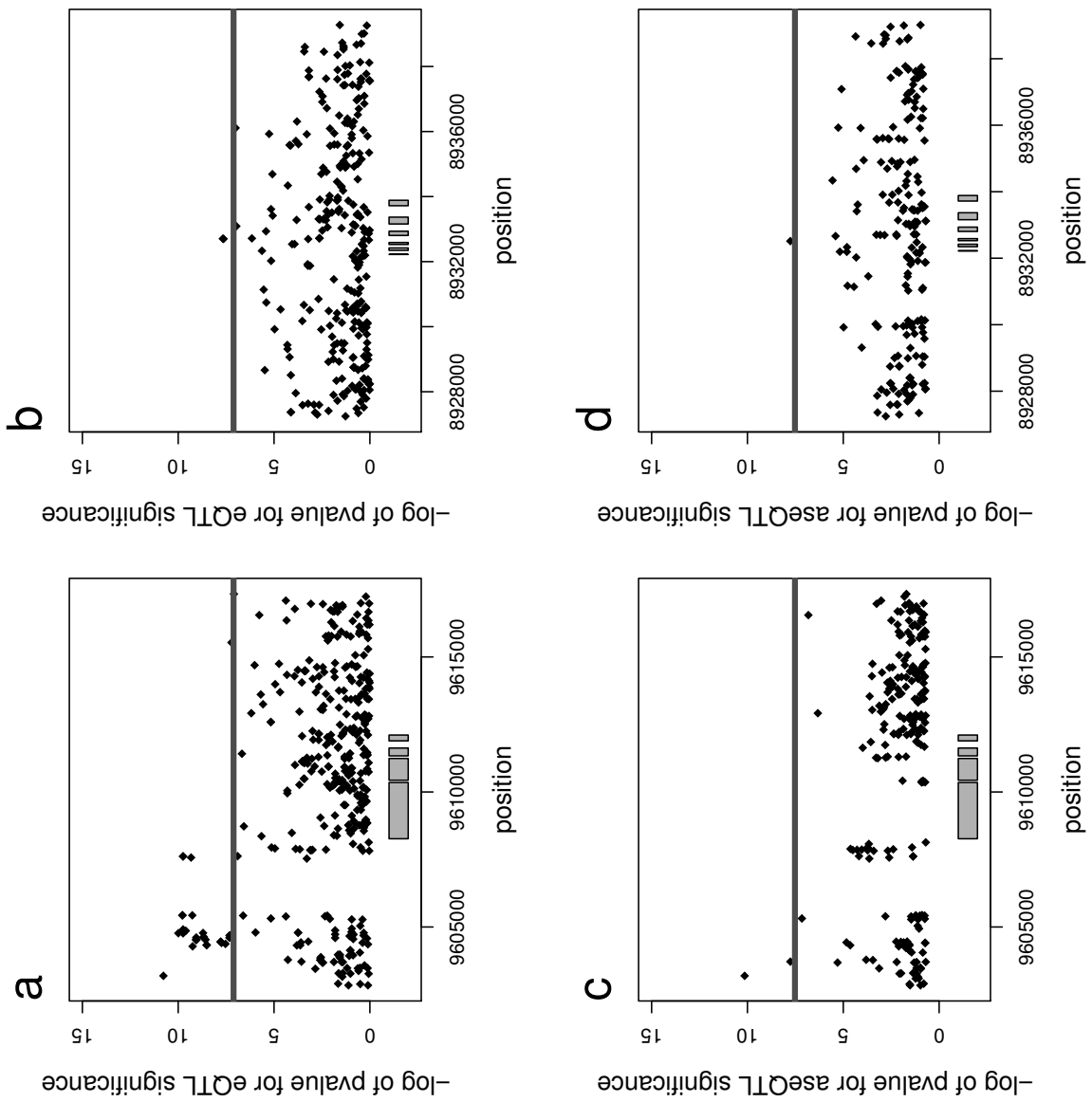

Figure 3.8: **Example eQTL and aseQTL genes.** Manhattan plots for associations between SNPs and total expression (a and b) and ASE (c and d) for two genes, PAC:20895445 (a and c) and PAC:20904926 (b and d). Each black dot represents a SNP and is plotted by genomic position on the x axis and the negative log of the p value for association on the y axis. The gray line denotes the p value threshold corresponding to an FDR of 0.01. The gray boxes represent the exons of the gene. Note that PAC:20895445 is an ortholog of AT4G16250.1, PHYTOCHROME D and PAC:20904926 is an ortholog of ATG68185.1, a ubiquitin-like superfamily protein.

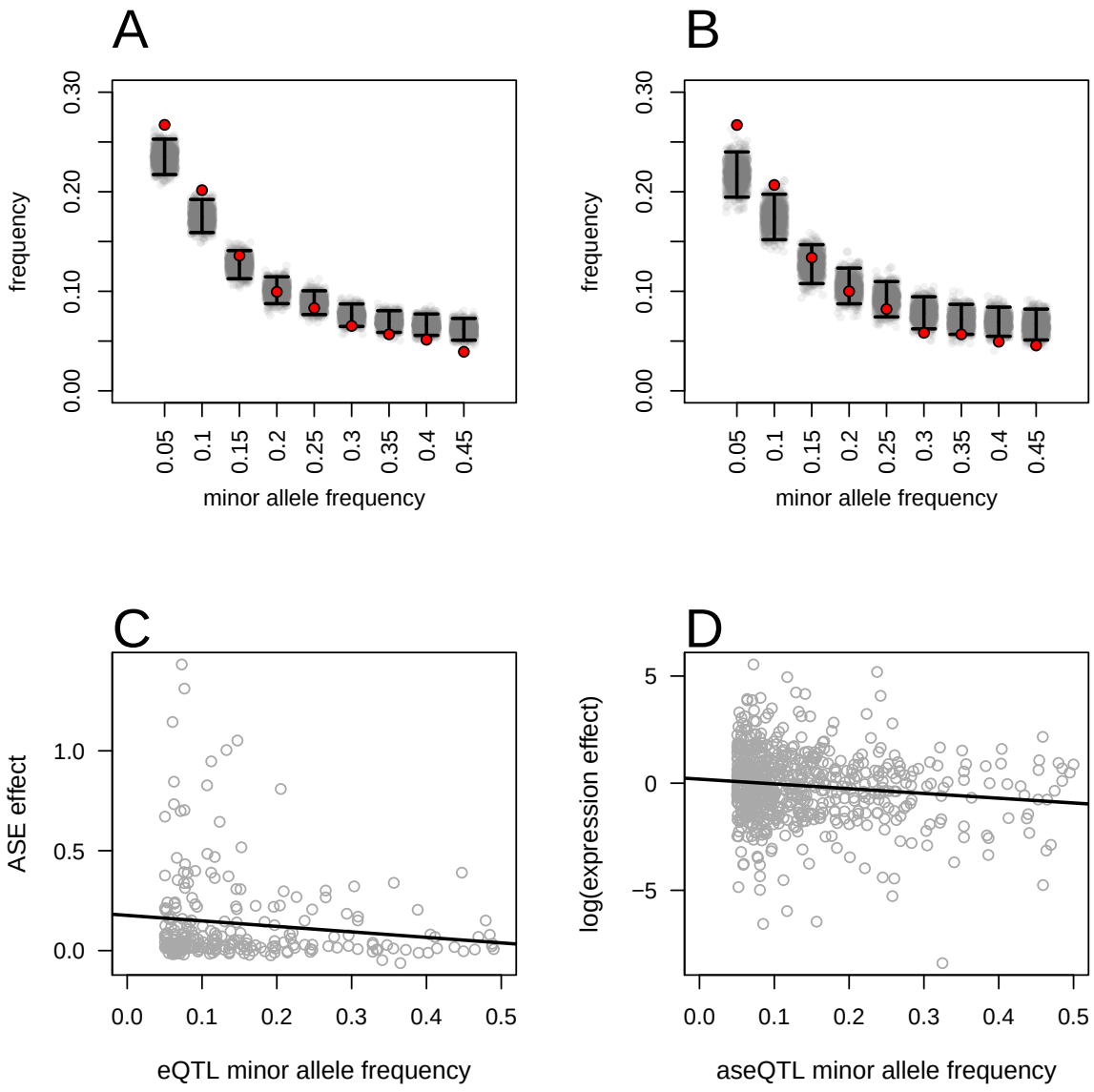

Figure 3.9: **The effect of designating a random associated SNP per gene as eQTL/aseQTL instead of the most associated SNP per gene.** The site frequency spectrum of eQTLs (a) and aseQTLs (b) for observed data (red circles) and permuted data (gray circles, black lines are 95% confidence intervals) when a random SNP is chosen per gene to be an eQTL or aseQTL. The same eQTLs and aseQTLs are plotted in (c) and (d). In (c), eQTL minor allele frequency is plotted against the effect of that SNP on ASE, calculated as the mean difference in ASE between individuals heterozygous at the eQTL and individuals homozygous at the eQTL. Negative values occur when the the homozygote for the eQTL has greater ASE than the heterozygote. The black line is calculated by linear regression. In (d), aseQTL minor allele frequency plotted against the effect of the aseQTL on total gene expression, calculated by taking the log of the absolute value of the mean difference in expression between individuals heterozygous at the aseQTL and individuals homozygous for the common allele at the aseQTL. The trend line was calculated by regression between minor allele frequency and the log of the expression effect.

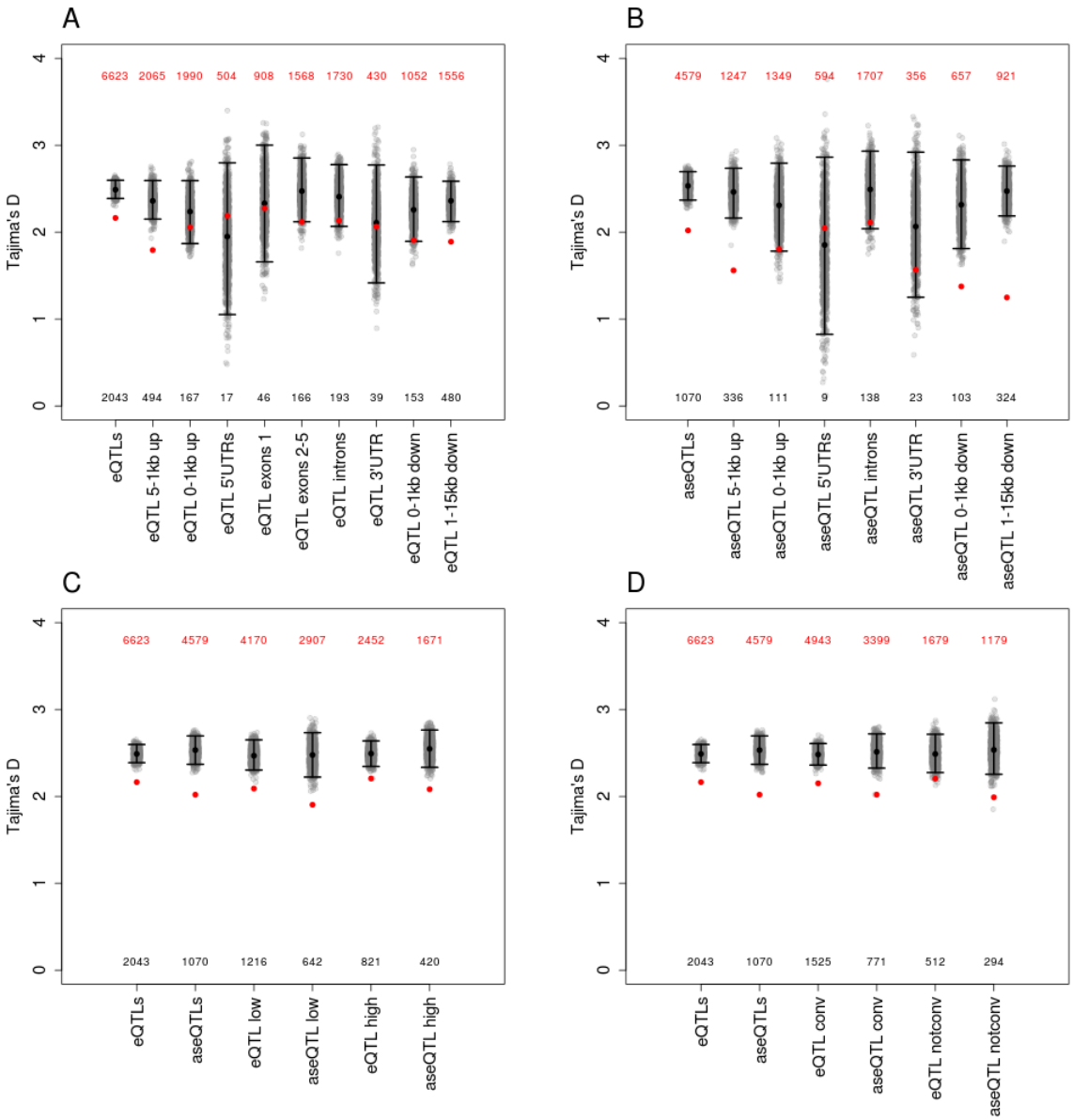

Figure 3.10: **Tajimas D of eQTLs and aseQTLs within site type, recombination rate, and substitution type.**(a) shows Tajimas D for eQTLs of various categories. Red circles are the real data, gray circles show Tajimas D for permuted eQTLs, and black lines show 95% confidence intervals. Tajimas D was used to summarize the site frequency spectra and make plots more readable than they would be if raw frequencies were plotted. The total number of eQTLs in each category is shown with the red numbers and the mean number of permuted eQTLs in each category is shown with the black numbers (b) shows the same data as (a) but for aseQTLs. (c) shows Tajimas D for eQTLs and aseQTLs (red dots) and permuted eQTLs and aseQTLs (gray dots, black bars are 95% confidence intervals) for sites in low recombination regions (<3.45 cM/mB) and high recombination regions (>3.45 cM/mB). (d) shows Tajimas D for A/T to G/C substitutions that could be favored by gene conversion (conv) and other substitutions (notconv). Note that all Tajimas D values are significantly increased because only SNPS above a certain allele frequency were testable, so that even for 4fold degenerate sites in the analysis, Tajimas D is 2.403.

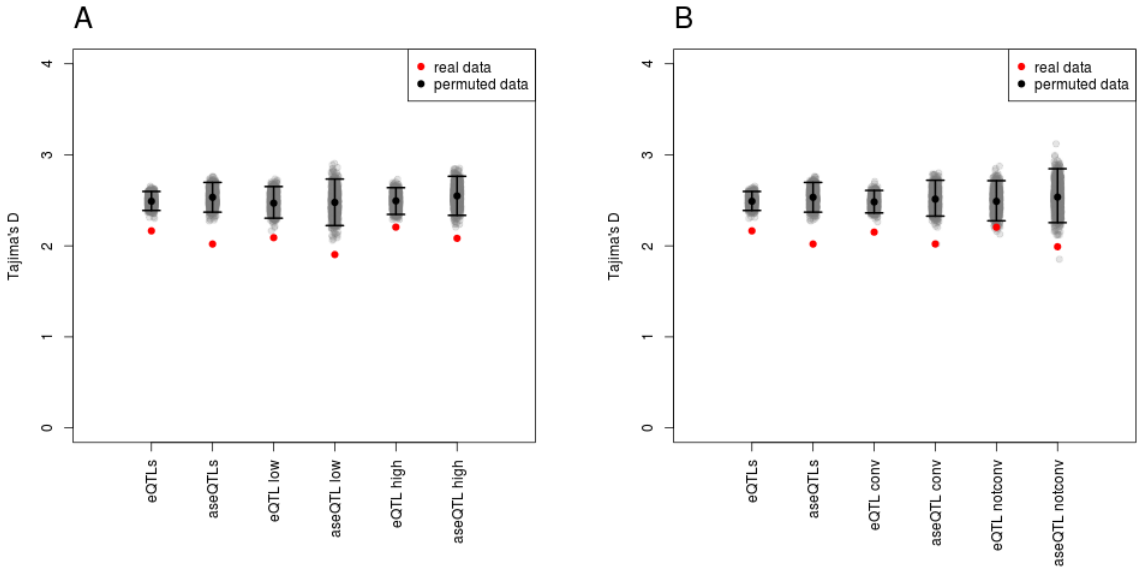

Figure 3.11: **A comparison of mapping programs in highly polymorphic regions.**RNAseq coverage for an example gene using mapping from Tophat (top) and Stampy (bottom). Colored lines indicate polymorphic sites compared to the reference. The arrows indicate regions where coverage was reduced in Tophat because of multiple polymorphisms. Note that Tophat reads have splice junctions while Stampy reads do not because we mapped to an exon-only reference.

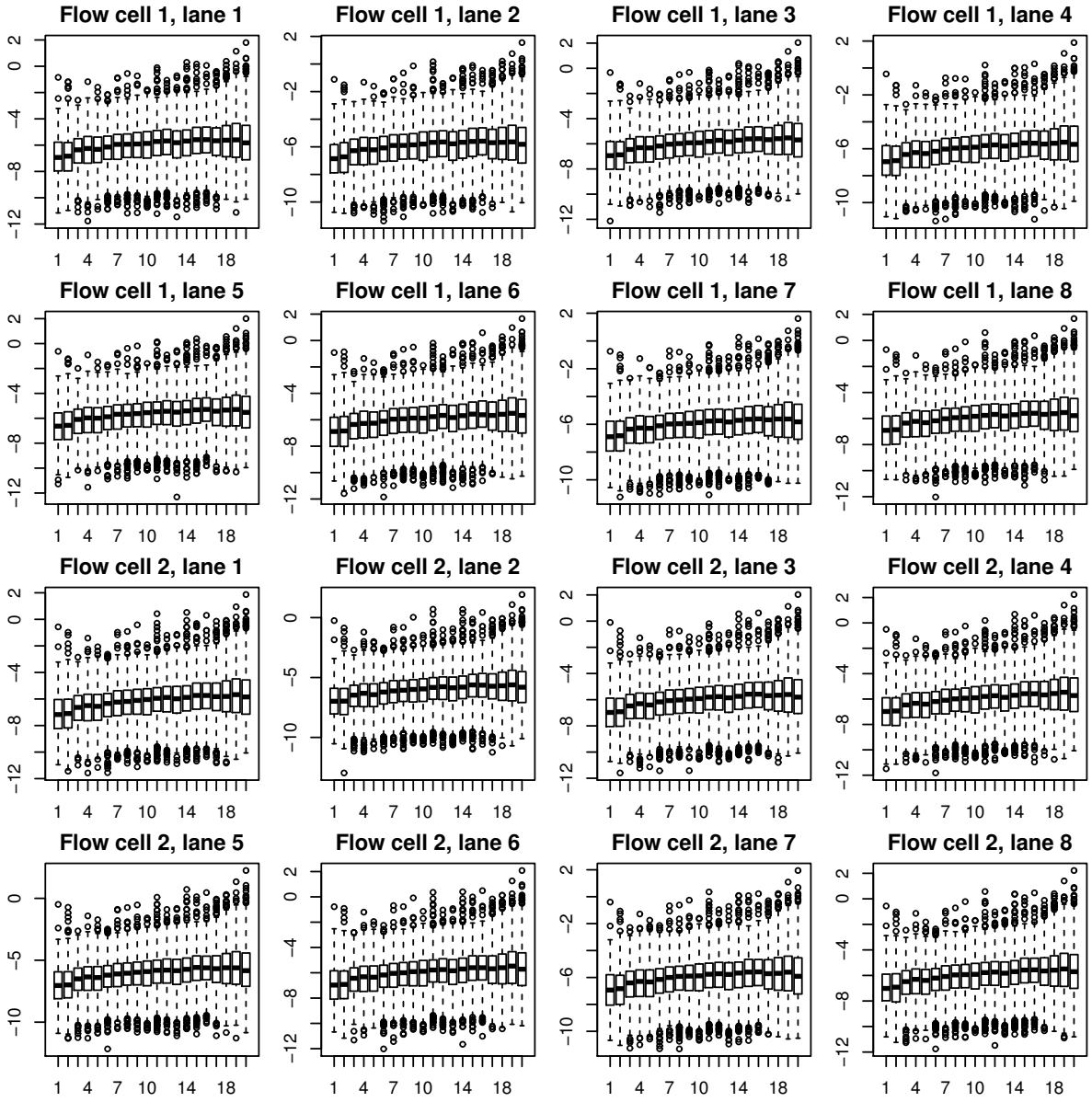

Figure 3.12: **GC composition and expression by lane.** All genes included in the study were split into 20 equally sized bins by GC content. Expression in these bins was combined for each lane and plotted in box plots.

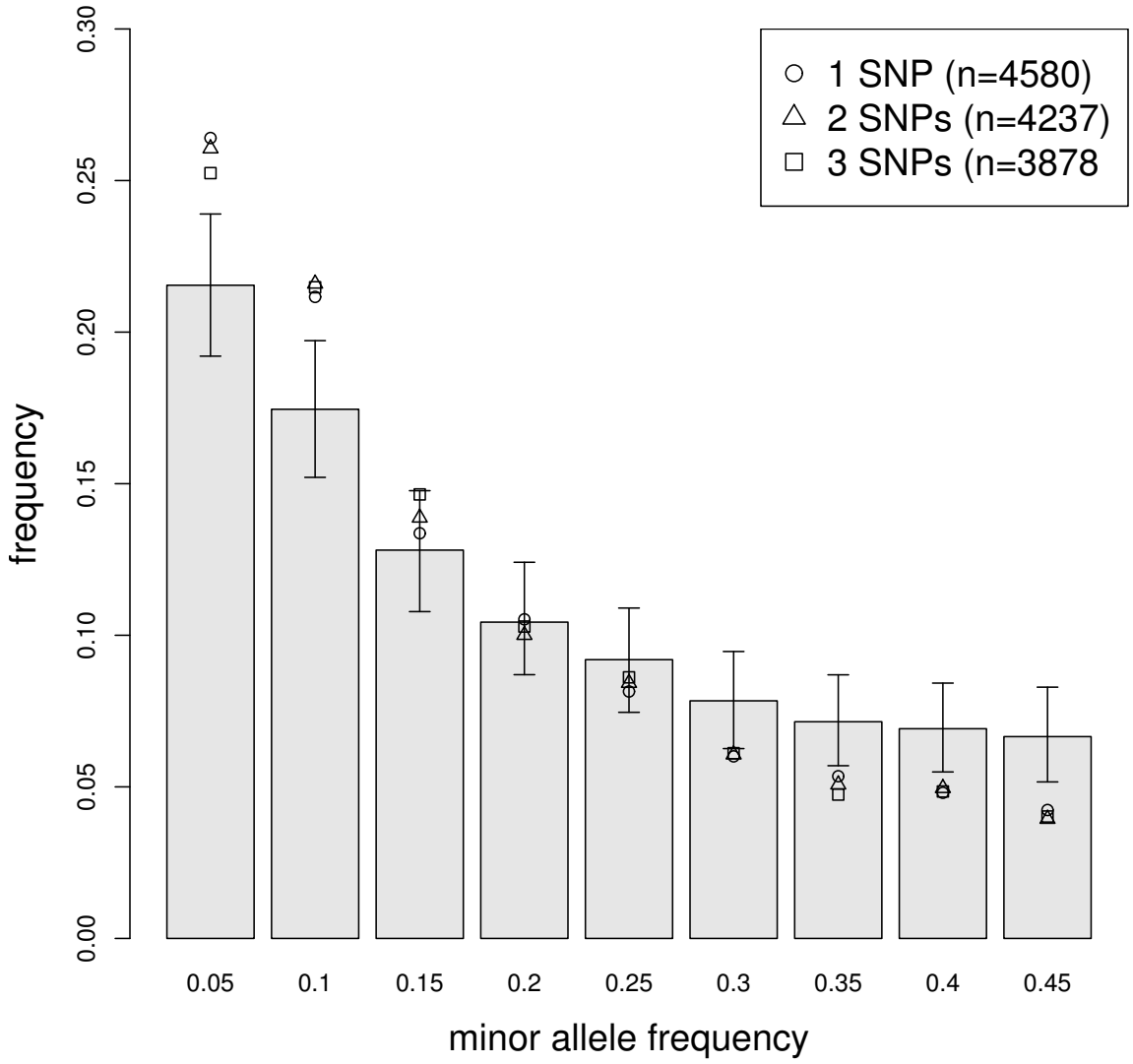

Figure 3.13: **The effect of increasing the number of SNPs required to measure ASE effects aseQTL detection.** The minor allele frequency of aseQTLs detected when ASE measurement required 1 heterozygous coding SNP (circles), 2 SNPs (triangles), and 3 SNPs (squares). While increasing the numbers of SNPs required to measure ASE reduced the number of aseQTLs detected, it did not qualitatively change our conclusions about the rareness of aseQTLs.

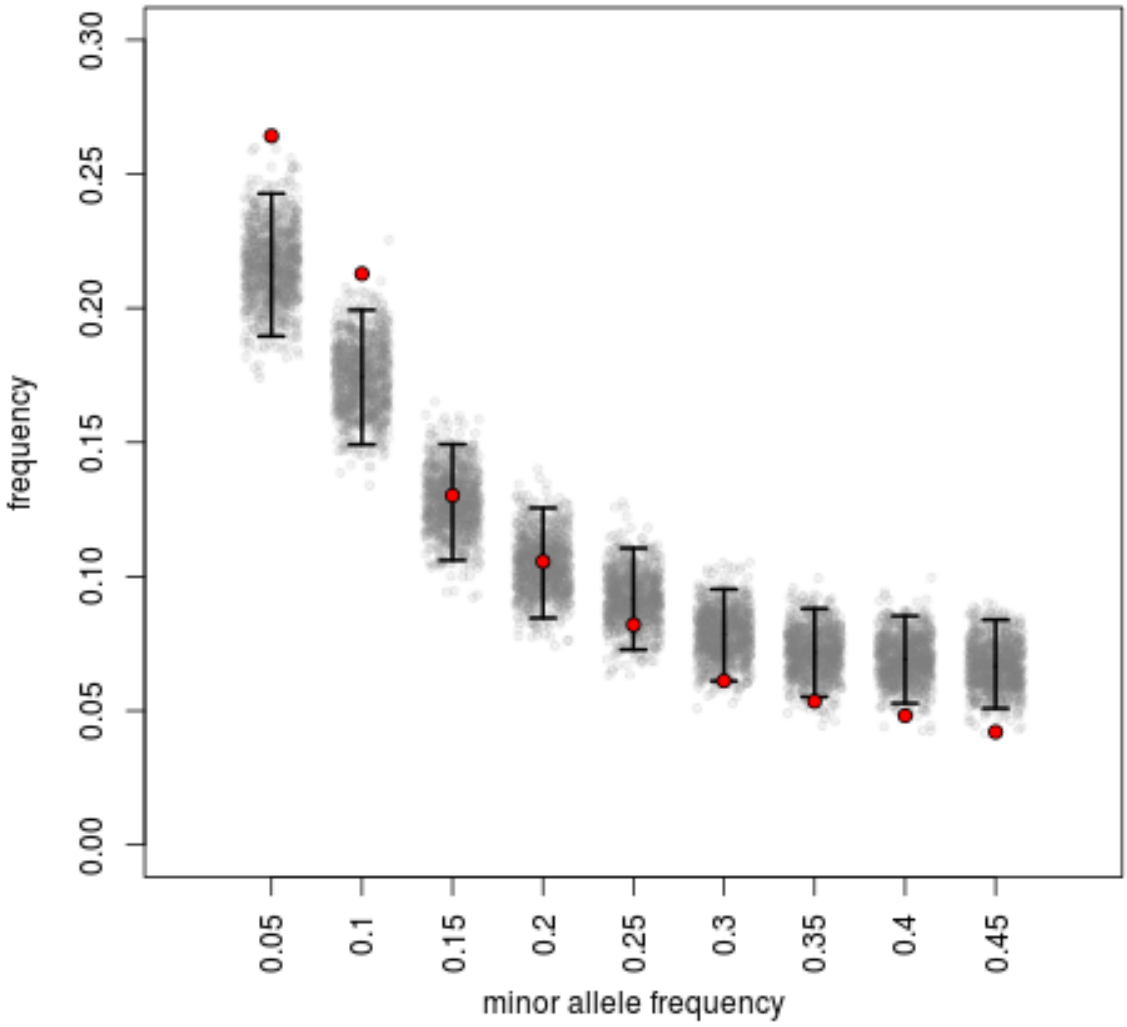

Figure 3.14: **The site frequency spectrum of aseQTLs when genes with SNPs showing ASE bias are removed from the analysis.** Red dots are frequencies of aseQTLs, gray dots are frequencies for permuted aseQTLs and black lines show 95% confidence intervals.

Bibliography

Haldane, J. B. S. (1927). A mathematical theory of natural and artificial selection, part v: Selection and mutation. *Math. Proc. Cambridge Philos. Soc.*, 23, 838–844.

Kousathanas, A., Oliver, F., Halligan, D. L. & Keightley, P. D. (2011). Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Mol. Biol. Evol.*, 28, 1183–1191.

Lee, Y. W., Gould, B. A. & Stinchcombe, J. R. (2014). Identifying the genes underlying quantitative traits: a rationale for the QTN programme. *AoB Plants*, 6.

Williamson, R. J., Josephs, E. B., Platts, A. E., Hazzouri, K. M., Haudry, A., Blanchette, M. & Wright, S. I. (2014). Evidence for widespread positive and negative selection in coding and conserved noncoding regions of capsella grandiflora. *PLoS Genet.*, 10, e1004622.

Zhu, Q., Ge, D., Maia, J. M., Zhu, M., Petrovski, S., Dickson, S. P., Heinzen, E. L., Shianna, K. V. & Goldstein, D. B. (2011). A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *Am. J. Hum. Genet.*, 88, 458–468.