

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

© Copyright 2015 by Emily Beth Josephs

# **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

Understanding the evolutionary forces that maintain variation at the sequence and phenotypic level is crucial for developing a full understanding of how evolution works in nature. In my thesis, I investigate the evolutionary forces that maintain genetic variation at the sequence level and at the phenotype level using genomic and transcriptomic data from the plant *Capsella grandiflora*. In Chapter 2, I show that negative selection and positive selection shape sequence variation in coding and conserved noncoding sequence and that a gene's expression level influences the strength of negative selection. In Chapter 3, I map the loci that affect expression level genome-wide and show that negative selection is the dominant force acting on these loci, consistent with the hypothesis of mutation-selection balance. In Chapter 4, I show that genes with higher coexpression network connectivity experience stronger positive and negative selection, suggesting that pleiotropy, as measured by network connectivity, increases constraint but does not limit adaptation. Overall, my results suggest that mutation-selection balance explains much of the genetic variation observed within populations at both the sequence and phenotypic level.

To my grandmothers, Myra Josephs and Mary Barnard.

## Acknowledgements

This thesis would not have been possible, nor would my experience as a Ph.D. student been as pleasant, without support from a number of people.

My supervisors John Stinchcombe\* and Stephen Wright\* have been wonderful mentors and friends. I am very grateful for their guidance, encouragement, respect, and patience. (The \*'s denote equal contributions).

This thesis benefited greatly from the advice from my committee: Asher Cutter, Alan Moses, and Art Weis, along with my appraisal committee members Nicholas Provart and Marie-Josée Fortin. I am preemptively grateful to Helen Rodd and Matthew Rockman for commenting on my thesis and sitting on my defense committee.

Robert Williamson, Arvid Ågren, and Young Wha Lee joined the department at the same time that I did and I feel incredibly lucky to have them as my academic twins and friends. I am especially grateful to Robert for teaching me most of what I know about bioinformatics, figure design, pedagogy, and internet memes, to Young Wha, for becoming one of my closest friends and someone I know I can always turn to for help, and Arvid, for his boundless enthusiasm and positive energy that never fail to brighten up my day.

I am very grateful for the friendship and support of a number of graduate students and postdocs in the department, who have been, among other things, friends, lunchmates, gym buddies, and drinking companions. I would especially like to thank: Alison Wardlaw, Susie Wadgymar, Nathaniel Sharp, Anna Simonsen, Amanda Stock, Adriana Salcedo, Brechann McGoe, Yuheng Huang, Khaled Hazzouri, Matty Hartfield, Tia Harrison, Megan Greischar, Amanda Gorton, Stephen DeLisle, Brandon Campitelli, Joanna Bundus, Felix Beaudry, Maggie Bartkowska, Ramesh Arunkumar, Emily Austen, and Michelle Afkhami. I would also like to make very special thanks to Jenn Coughlan, for being the best.

A number of faculty who were not formal mentors have still provided valuable guidance, including Aneil Agrawal, Locke Rowe, and Helen Rodd. My thesis work would not have been possible without the assistance Adrian Platts and Wei Wang, who always had helpful answers for my bioinformatics questions, Bruce Hall and Andrew Petrie, who assisted in the greenhouse, and countless EEB department staff members, especially Stephanie Mello, Courtney Key, Pam Pecoskie, and Kitty Lam.

I would not have made it to grad school in the first place without guidance and encouragement from my early mentors: Annie Schmitt and Leonie Moyle, along with many members of the Schmitt and Moyle labs (especially Yaniv Brandvain).

I am incredibly grateful to my parents, Bill and Zina Josephs, for their unconditional love, support, and also proofreading. I would also like to thank my husband and partner Kevin Lucas, for always accepting me in my endeavours but also persistently making sure that I have some fun outside of school. I am very lucky to have him in my life.

Finally, I am grateful to funding sources. The first three years of my Ph.D. were supported by a United States National Science Foundation Graduate Research Fellowship and the sequencing work that formed the basis of my thesis was paid for by a Genome Canada grant.

– Emily

# Contents

# List of Tables

# List of Figures

# Chapter 1

## General introduction

No one supposes that all the individuals of the same species are cast in the very same mould. These individual differences are highly important for us, as they afford materials for natural selection to accumulate.

— ?

The individuals belonging to a species differ to a greater or less extent.

— ?

Attempts to understand the causes and significance of organic diversity have been made ever since antiquity; the problem seems to possess an irresistible aesthetic appeal.

— ?

There is a considerable amount of genic variation segregating in all of the populations studied and .. the real variation in these populations must be greater than we are able to demonstrate.

— ?

As many great evolutionary biologists have noted, natural variation is ubiquitous. However, the relative importance of various evolutionary forces in maintaining this variation, especially at the population level, is still unknown.

### 1.1 Population genetic approaches to the maintenance of variation.

The question of what forces maintain genetic variation has been investigated extensively in population genetics. Two competing hypotheses formed early on for explaining within-population variation (reviewed in ?). The classical school argued that little variation segregates in populations, most new mutations are deleterious, and most observed variation is transient, on the way to either adaptive fixation or removal. In contrast, the balance school argued that large amounts of genetic variation are maintained within populations due to heterosis or other forms of balancing selection.

As the first protein polymorphism data emerged, it appeared that neither hypothesis could explain the data (??). Levels of heterozygosity in proteins were too high to be explained by balancing selection alone since at every polymorphic site, homozygotes will still make up at least 50% of the population at Hardy-Weinberg equilibrium, leading to extreme variation in fitness between individuals and inbreeding depression (?). However, high levels of sequence variation did not fit with the original classical theory's prediction that there is low variation within populations. Kimura and Ohta reconciled high within-population variation with the classical school by proposing that most of the sequence variation observed in nature, both within and between populations, is either selectively neutral or effectively selectively neutral (????).

However, genome-wide sequence data has shown that selection does shape observed genetic variation and molecular divergence between species, both directly in coding and noncoding sequence, and indirectly through linked selection. McDonald-Kreitman based approaches have found that >30% of replacement fixations in *Drosophila* and *Escherichia coli* are due to positive selection (????). Within population estimates of the distribution of fitness effects of segregating mutations show that weakly deleterious mutations contribute significantly to standing genetic variation in humans (??).

There is also evidence that selection shapes genetic variation outside genic regions. ? found constraint across intergenic, intronic, and untranslated regions of the *D. simulans* genome and there is constraint in the noncoding sequences that flank genes of other species (?). However, quantifying selection in noncoding regions is difficult because these regions do not break down into obvious selected and neutral sites. Methods that have used information about specific function have been able to detect selection on functional sites such as transcription factor binding sites (?) and phylogenetically-conserved noncoding sequences (?).

There is mounting evidence that selection shapes diversity at neutral sites in linkage disequilibrium with directly selected sites. Linked neutral diversity will be removed by both fixations due to positive selection (?) and the removal of polymorphisms by negative selection (?). The combined effects of selective sweeps and background selection are widespread, as demonstrated by positive correlations between recombination rate and polymorphism (?????), negative correlations between divergence and polymorphism (???), and reduced average neutral diversity around replacement fixed substitutions relative to silent fixed substitutions (?) but see (?). Recent estimates suggest that linked selection dramatically reduces genome-wide diversity in the *Drosophila simulans* genome (?).

Most of the discussion above focussed on negative and positive selection, which both remove genetic variation from within populations, but selection can also directly increase genetic variation through balancing selection. However, the impact of balancing selection on sequence variation has generally been difficult to quantify. Balancing selection can result from a number of different processes, including heterosis, frequency-dependent selection, spatially-variable selection, and temporally-variable selection (??) and the temporal scale at which balancing selection occurs will affect the types of genetic signatures it leaves behind (?). For example, long-term balancing selection appears to maintain genetic variation between humans and chimpanzees at a number of loci (?), but the genetic signature of short-term balancing selection is difficult to disentangle from other demographic events (?).

Overall, it is clear that some genetic variation is neutral, some is deleterious, and some is beneficial.

However, the relative importance of these processes for shaping genome-wide variation is still unclear.

## 1.2 Quantitative genetic approaches to the maintenance of variation

The types of selection that shape sequence variation may differ from the types of selection that shape phenotypic variation because much of the variation at sequence level may not translate into phenotype (this is a central tenet of the nearly neutral theory). Since selection generally acts on the phenotype, understanding the selective forces that maintain genetic variation for phenotype is crucial. The main hypotheses for the maintenance of quantitative genetic variation for phenotypic traits mirror those in population genetics: mutation-selection balance and balancing selection (??).

An exemplar of the kinds of approaches that quantitative geneticists have used to address the maintenance of genetic variation comes from Kelly, who proposed that the contribution of deleterious alleles to genetic variation could be quantified by investigating the directional dominance of variation in a trait (?). Applications of this approach have shown that intermediate-frequency alleles, not rare recessive alleles as expected under mutation-selection balance, can explain variation for flower size in *Mimulus guttatus* and female fecundity in *Drosophila* (??). However, this approach requires multiple generations of breeding in order to measure directional dominance and has so far only been applied to a few traits that may not be representative of all traits.

## 1.3 Linking genotype and phenotype

Developing an understanding of the selective forces that maintain genetic variation for phenotype on a systematic level will require linking genetic variation to phenotypic variation. An enormous effort has been made over the last few decades to do exactly this, using techniques like linkage mapping, quantitative trait loci mapping, and genome-wide association studies (GWAS). I will refer to all of these techniques under the umbrella term QTN program. following ?.

Often the QTN program has not been motivated by evolutionary questions, but instead by an attempt to better understand the molecular basis of a trait and then use this information to develop treatments for a disease, find genetic regions that could be introgressed into crop plants, or other practical uses. However, evolutionary biologists have also made use of QTN techniques to describe the genetic architecture of traits (?). It seems especially appealing to use the sample of known QTNs (sometimes distressingly referred to as the QTLome (?)) to make general conclusions about the types of variation maintained within populations (?).

Unfortunately, inferring the selective forces acting on variation from QTNs is not straightforward (??). The act of finding QTNs biases us towards finding large-effect alleles but other lines of evidence suggest that small-effect QTNs are also important for genetic variation. For example, the inability of genome-wide association mapping studies to identify large-effect common alleles for many heritable human

diseases suggests that either small-effect and/or very rare large-effect alleles make up genetic variation for these traits (reviewed in Rockman 2012).

As emphasized by (?), identifying QTNs within a population is still a useful approach to understanding how selection acts on traits. Specifically, QTNs maintained by mutation-selection balance are expected to be found at low allele frequencies while QTNs maintained by balancing selection are expected to be found at intermediate allele frequencies (?). However, efforts to infer selective forces acting on QTNs must be careful of the ascertainment biases that filter the sample of known QTNs. One successful approach has been a set of studies that estimate the proportion of variation in a trait explained by all genotyped loci to show that small-effect alleles do contribute to genetic variation, even though the specific loci involved are unknown (??).

A potential strategy for generating a sample of QTNs with relatively small fitness effects is to investigate the loci that affect gene expression. While expression is often thought of as an intermediate phenotype, it is more useful in this case to think of it as a model phenotype that integrates across phenotypic space, is straightforward to measure on a large scale, and is measurable without prior, potentially biasing information about how a certain expression level may vary in nature or influence fitness (?).

## 1.4 The importance of standing variation for adaptation

While understanding the maintenance of variation is fundamental for a full understanding of evolutionary biology, it is also relevant because the types of variation maintained within populations will control adaptation. The classical theory predicts that adaptation occurs through new mutations, leading to the signatures of classic selective sweeps and an adaptive walk that occurs in short mutation-limited bursts (???). However, the balance theory suggests that ecologically relevant variation waits within populations, ready to contribute to adaptation if conditions change. Adaptive mutations that rise to fixation from standing variation will leave a subtler sweep signal in linked neutral variation since more than one haplotype may fix in the population (???). If adaptation occurs at many loci present in standing variation, the population can reach a new trait optimum without any fixations at all (??).

The origin of adaptive alleles will not only affect our ability to detect that adaptation using population genetic data, but will also affect the kinds of mutations that actually fix during adaptation. For example, adaptation from standing variation may involve more small-effect alleles than adaptation from new mutation (?). In addition, local adaptation between divergent populations from standing variation may involve more instances of conditional neutrality, where one allele is selected in one population and neutral in another population, than trade-offs, where alleles are alternately favored and disfavored in populations, since the alleles involved in trade-offs are less likely to be maintained within populations than neutral alleles (?).

## 1.5 Research Objectives

Throughout the course of my Ph.D., I have investigated the selective forces that maintain genetic variation within populations and that shape sequence variation at the population, within-species, and between-species level. Below I describe my study system and outline the chapters of my thesis.

### 1.5.1 Study System

I used the species *Capsella grandiflora* throughout my thesis. *C. grandiflora* has been a useful system for the investigation of selective forces for a number of reasons. First, *C. grandiflora* has had a large, relatively stable effective population size making both positive and negative selection strong relative to drift (?). Second, *C. grandiflora* has relatively low population structure (?), making it possible to do population genomics on a species-wide sample and minimizing the impact that migration from nearby populations will have on a focal population. Third, unlike most plant model systems, *C. grandiflora* is obligately outcrossing, making it appropriate for answering questions about selective forces that act within populations. Finally, *C. grandiflora* is a highly tractable genomic system. It has a small genome for a plant (219 MB), making it feasible to generate and map sequence data from a population sample and is closely related to *Arabidopsis thaliana* ( $\sim$ 10 million years ago), allowing me to use well-curated annotation information on gene function and structure.

### 1.5.2 Chapter 2

In this chapter I tested for the signature of recurrent selective sweeps in coding and conserved non-coding regions of the *C. grandiflora* genome and described the effects of expression level on the strength of selection in genic regions. This work was part of a collaborative effort with Robert Williamson to describe the strength of positive and negative selection across the *C. grandiflora* genome.

Robert used the site frequency spectrum to estimate the distribution of fitness effects of new mutations and, ultimately, quantify the relative strength of selection across the *C. grandiflora* genome. He found that both positive and negative selection appeared to be strong in coding and conserved noncoding regions. I detected reduced neutral diversity surrounding fixed replacement substitutions in *C. grandiflora* relative to fixed synonymous substitutions, consistent with the action of recurrent selective sweeps. I also detected the signature of recurrent selective sweeps around conserved noncoding sequences.

In addition, I investigated the importance of expression level in shaping between-gene variation in selection. Previous work has shown that highly expressed genes diverge more slowly than lowly expressed genes. However, this could result from either stronger constraint on highly expressed genes or stronger positive selection on low expressed genes. I showed that stronger negative selection in highly expressed genes explains the observed correlation between expression and divergence.

This work was published in Plos Genetics in 2014: Williamson RJ\*, Josephs EB\*, Platts AE, Hazzouri KM, Haudry A, Blanchette M, et al. (2014) Evidence for Widespread Positive and Negative Selection in Coding and Conserved Noncoding Regions of *Capsella grandiflora*. PLoS Genet 10(9): e1004622.

doi:10.1371/journal.pgen.1004622 (\*The two first authors contributed equally)

### 1.5.3 Chapter 3

In this chapter I set out to map the genetic variants that affect expression in a population of *C. glandiflora* and used the allele frequencies and effect sizes of these variants to show that negative selection acts on these variants, and ultimately, on expression variation. This project was done in collaboration with Young Wha Lee, who mapped the genomic data and called variants. A preprint is posted on Biorxiv: Josephs, EB, Lee, YW, Stinchcombe, HR, & Wright, SI. Association mapping reveals the role of mutation-selection balance in the maintenance of genomic variation for gene expression <http://biorxiv.org/content/early/2015/09/21/015743>

### 1.5.4 Chapter 4

In this chapter I investigated the relationship between expression network structure and selection. I collaborated with Dan Schoen at McGill University, who used the expression data generated in Chapter 3 to construct coexpression networks. I then used these networks to test for the effects of network connectivity and module membership on selection. I found that connectivity increased both positive and negative selection, suggesting that pleiotropy, as measured through connectivity, does not constrain adaptation.

# 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 (?). Current evidence suggests that both positive and negative selection are common in coding and some noncoding sequences in several model systems (??????). However our understanding of genome-wide selection in plants remains relatively limited (?), 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 (????). 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 (?? but see ?). Halligan and colleagues (2013) 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 (??).

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 (?), possibly leading to lower levels of functional constraint and positive selection on plant noncoding DNA. Consistent with this prediction, Haudry and colleagues (2013) 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 (????), including plants (????). 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 (?). 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 (???). 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 ( $N_e \sim 600,000$ ) and relatively low population structure (??). 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 (?) and an analysis of neutral variation linked to lineage-specific fixed substitutions (?). 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;  $\sim$ 140 Mb genome assembly) sampled from across the species' range in northern Greece using single-end Illumina GAII sequencing (Table ??). The resulting 108 bp reads were mapped to the *Capsella rubella* reference genome (?) 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 (?). After filtering for quality and depth (see Methods), we were left with  $\sim$ 27 million sites,  $\sim$ 1.5 million of which were single nucleotide polymorphisms (SNPs). 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 (?). 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. ??).

Consistent with previous estimates made using a much smaller set of loci (257 Sanger-sequenced loci) and a different range-wide sample (?), average nucleotide diversity at 4-fold degenerate sites (Wattersons  $\theta_w$ ) 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 ( $\theta_w$  = 0.020) and intergenic regions ( $\theta_w$  = 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 ( $\theta_w$  = 0.015 and 0.014 respectively), while 0-fold degenerate nonsynonymous sites showed the strongest reduction ( $\theta_w$  = 0.005).

Neutral diversity at 4-fold degenerate sites near centromeric regions was elevated on most chromosomes, similar to observations made in *A. thaliana* (?), *Arabidopsis lyrata* (??) and *Medicago truncatula* (?), (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. ??). 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* (?). 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. ??).

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 ?? and Fig. ??). 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. ??).

### 2.3.2 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 ? 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. ??A). Consistent with the patterns of diversity described above, negative selection is generally much stronger in coding regions than noncoding regions (Fig. ??A). 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 ( $N_{es} > 1$ ), 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 ( $N_{es} > 10$ ), 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 (?) 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  $N_{es}$  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 ?. For this study, we used the *C. rubella* genome as a reference for alignment, but excluded *Capsella* when identifying CNSs in order to avoid circularity when quantifying selection from diversity (?). This method allows our

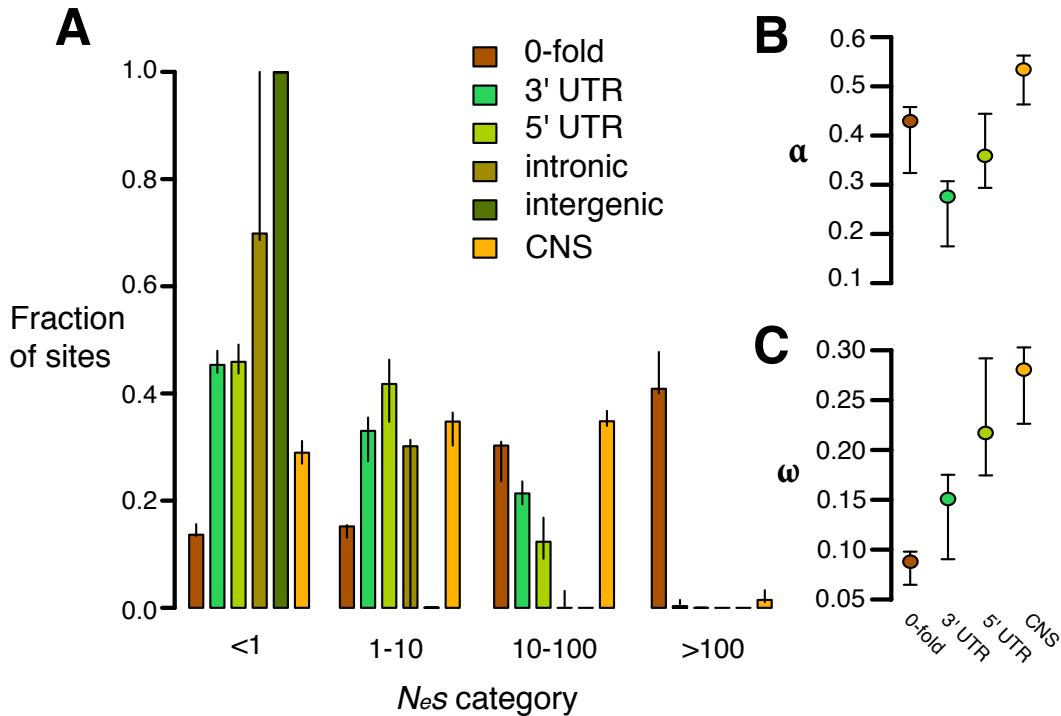


Figure 2.1: Estimates of negative and positive selection on coding and noncoding sites in *C. grandiflora*. A) The proportion of sites found in each bin of purifying selection strength, separated by site type, B) the proportion of divergent sites fixed by positive selection, and C) the rate of adaptive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

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 CNSs are subject to purifying selection (Fig. ??A). 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. ??A).

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. ??). 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. ??). 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 ??), 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 ( $N_{eS} < 1$ ) than the pooled CNSs, which could be due to strong selection for sequence specificity to obtain the proper secondary structure important for RNA activity (?). This effect is consistent with sncCNSs

showing a higher degree of conservation across the Brassicaceae (?) and having traceable orthologs in other plants.

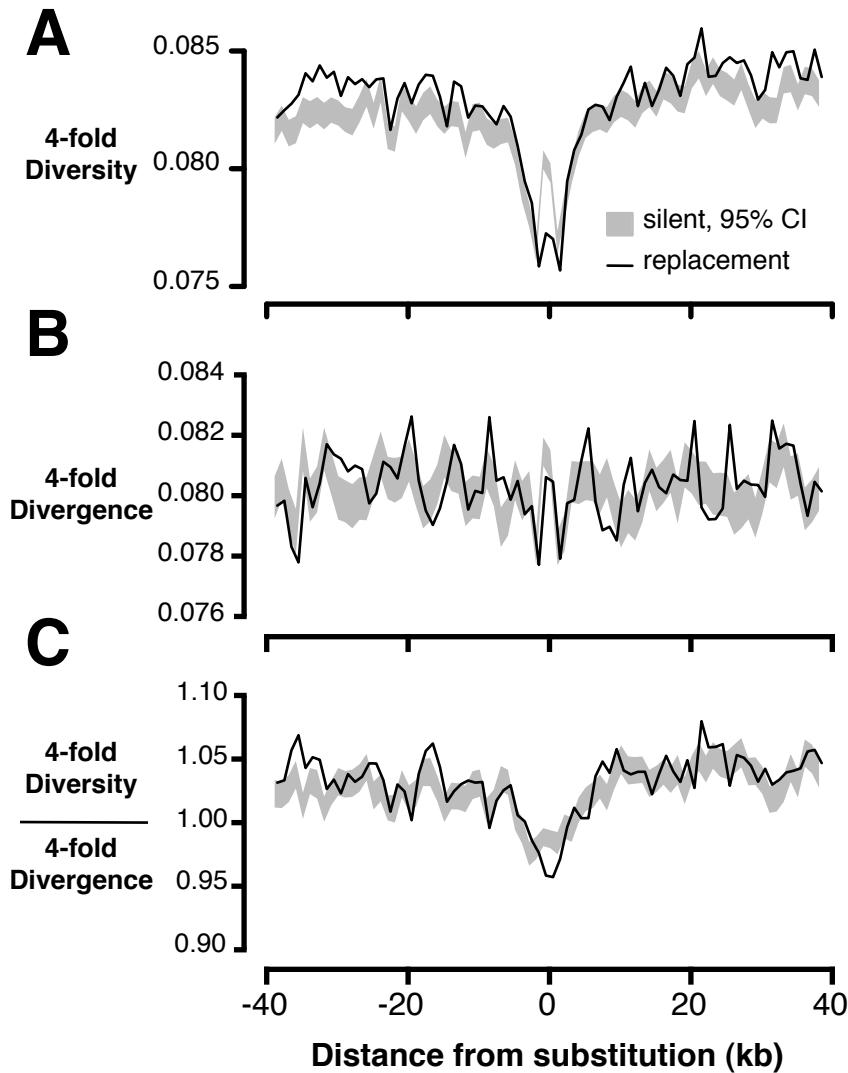
### 2.3.3 Genome-wide estimates of positive selection

We used the approach of ? to estimate the proportion of fixations driven by positive selection ( $\alpha$ ) and the rate of positive selection ( $\omega$ ) 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. ??B;  $\alpha = 0.417$ ) and estimates of the rate of adaptive substitution relative to synonymous substitution (Fig. ??C ;  $\omega = 0.08$ ). Similarly, UTRs and CNS sites show evidence for positive selection (Fig. ??B,C). These results generally suggest widespread positive selection in both nonsynonymous and functional noncoding genomic regions.

If many of the amino acid changes between *C. 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 (?). 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. ??A). This result was robust to various window sizes from 500kb to 2kb (Fig. ??) 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 (?), 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. ??B). 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. ??C,  $p < 0.01$  for 1 kb on either side of the substitution).

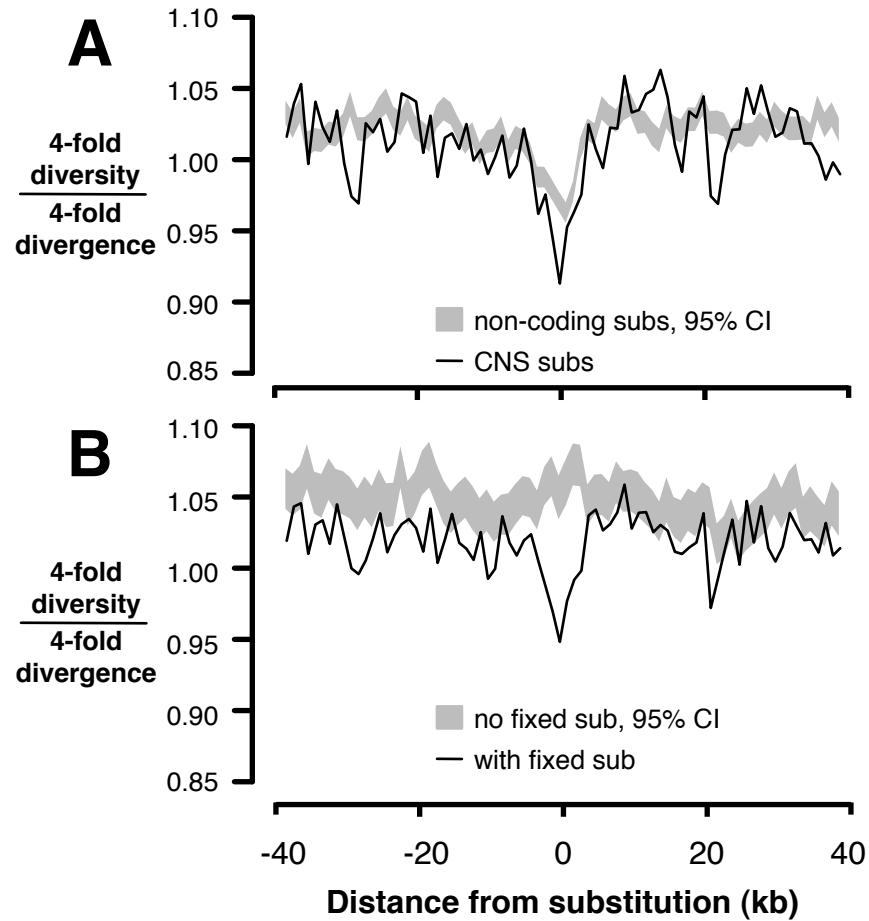
An analogous test for selective sweeps around fixations in noncoding regions is challenging because the test depends on accurately identifying interspersed functional and neutral sites, a difficult task in noncoding regions (?). Instead, we compared 4-fold diversity and divergence around fixed substitutions 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. ??). 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. ??A).

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).** A) 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. ??B).

### 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. ??). 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 from ?,  $p < 0.001$ ) (Fig. ??A).

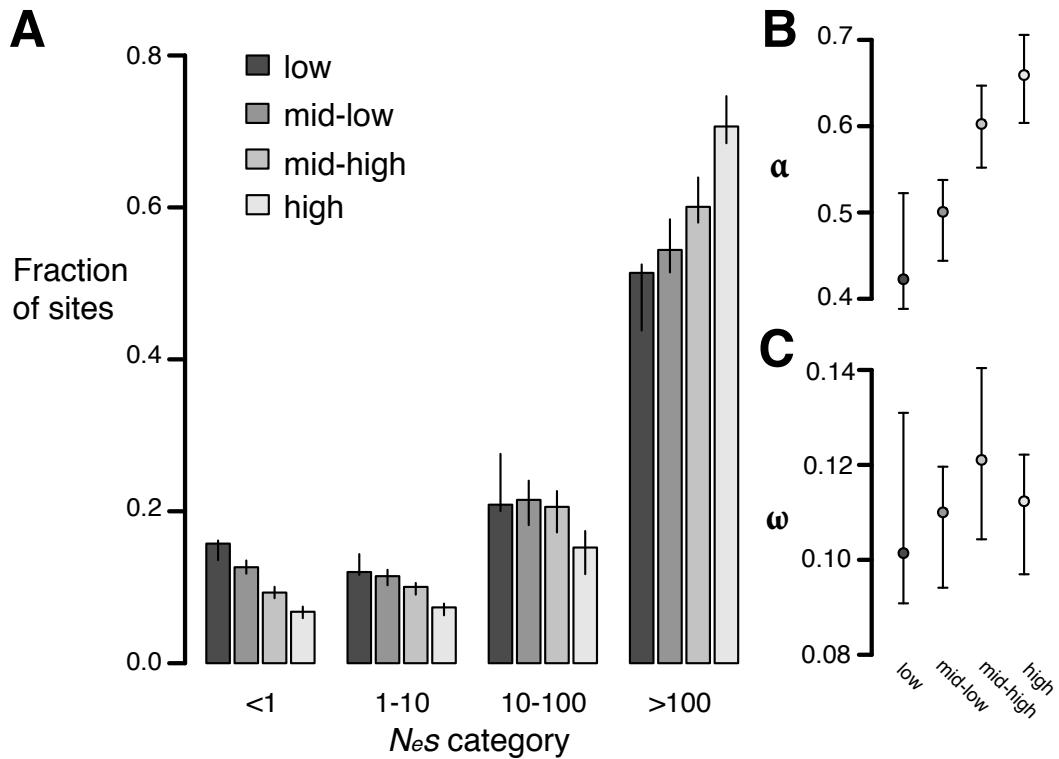


Figure 2.4: Estimates of negative and positive selection on nonsynonymous sites in genes of varying expression level. A) The proportion of sites found in each bin of purifying selection strength, separated by expression level, B) the proportion of divergent sites fixed by positive selection and C) the 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  $\alpha$  and  $\omega$  as described above. High expression genes have a significantly higher value of  $\alpha$  (0.66) than low expression genes (0.42,  $p < 0.01$ ) but the  $\omega$  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. ??B,C). The difference in  $\alpha$  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 (?).

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 ?, 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 (?). For example, in *Drosophila*, only 30%-70% of intronic and intergenic regions are nearly neutral (??). Similarly, ? 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 (?). 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 (?). UTRs are also under stronger negative selection than other noncoding regions in *Drosophila* (?), and this result is also in line with the previous study using comparative genomics in the Brassicaceae (?). 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 (?), 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 (?), the strength of selection inferred is close to effective neutrality. Furthermore, synonymous site selection is expected to be stronger in more high expression genes (??), 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 (??).

At synonymous sites, we see an excess of rare variants, as indicated by a negative Tajima's 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 (?). 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, Tajima's 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 (??). 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 (??). We expect that soft sweeps will also be common in *C. grandiflora* because of its large  $N_e$  (??). In addition, adaptation in genes that contribute to polygenic traits is often expected to occur without fixation of new mutations (?), 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 (????????). 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 (???). 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 (?).

## 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 ??). 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 (?), 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 (?) using the Stampy aligner 1.0.13 with default settings (?). Sites around indels were realigned using the Genome Analysis Toolkit (GATK) v1.05777 indel realigner (?). Genotype and SNP calls were conducted using the GATK UnifiedGenotyper with default parameters, 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. ??). We identified these regions by splitting the genome into 200kb windows, then calculating FIS (Fig. ??). 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 (?) with chaining, as previously described (?). 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 (?) and MultiParanoid (?). The peptide sequences of these orthologs were aligned using DialignTX (?), and reverse-translated into coding sequence. Whole-gene divergence at synonymous and nonsynonymous sites was calculated, using PAML (?), under a model where  $\omega$  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. ?? compared to Fig. ??). 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 (??) 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 ( $\geq 12$ nt) near-continuous regions of high conservation as previously described (?).

### 2.5.5 Estimates of the distribution of fitness effects and $\alpha$

Site categories were determined based on the Joint Genome Institutes gene annotation of the *C. rubella* reference genome (?). The allele frequency spectra (AFS) and divergence values were calculated for each category of sites, and DFE-alpha (??) 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 ?, 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 ?. 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 ?, 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 substi-

tutions 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 (?), and expression level was quantified from these mapped reads using Cufflinks 1.3.0 (?). 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 (6.817.5 FPKM), mid-high (17.5-44.7 FPKM), and high (44.7-17,092 FPKM). The distribution of fitness effects,  $\alpha$ , and  $\omega$  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

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

Site type	Positive selection		Distribution of fitness effects			
	$\alpha$	$\omega$	0-1	1-10	10-100	100- $\infty$
0fold	0.417	0.084	0.137	0.152	0.304	0.407
3utr	0.276	0.151	0.453	0.330	0.214	0.003
5utr	0.393	0.251	0.458	0.418	0.124	0.000
intergenic			1.000	0.000	0.000	0.000
intronic			0.698	0.302	0.000	0.000
2/3 fold	0.174	0.123	0.612	0.121	0.139	0.128
CNS	0.545	0.279	0.276	0.342	0.364	0.018
3UTRcns	0.518	0.256	0.281	0.333	0.364	0.023
DownstreamCNS	0.526	0.226	0.248	0.435	0.315	0.002
5UTRcns	0.424	0.157	0.253	0.320	0.392	0.034
UpstreamCNS	0.530	0.224	0.239	0.367	0.382	0.012
intronicCNS	0.502	0.373	0.415	0.256	0.283	0.046
IntergenicCNS	0.579	0.255	0.226	0.402	0.368	0.005
sncCNS	0.706	0.174	0.090	0.387	0.517	0.006
AmbiguousCNS	0.400	0.159	0.284	0.369	0.338	0.009
High expression	0.641	0.108	0.070	0.066	0.129	0.735
Mid-high expression	0.607	0.122	0.092	0.100	0.205	0.603
Mid-low expression	0.505	0.111	0.125	0.114	0.215	0.545
Low expression	0.505	0.124	0.141	0.135	0.257	0.467

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

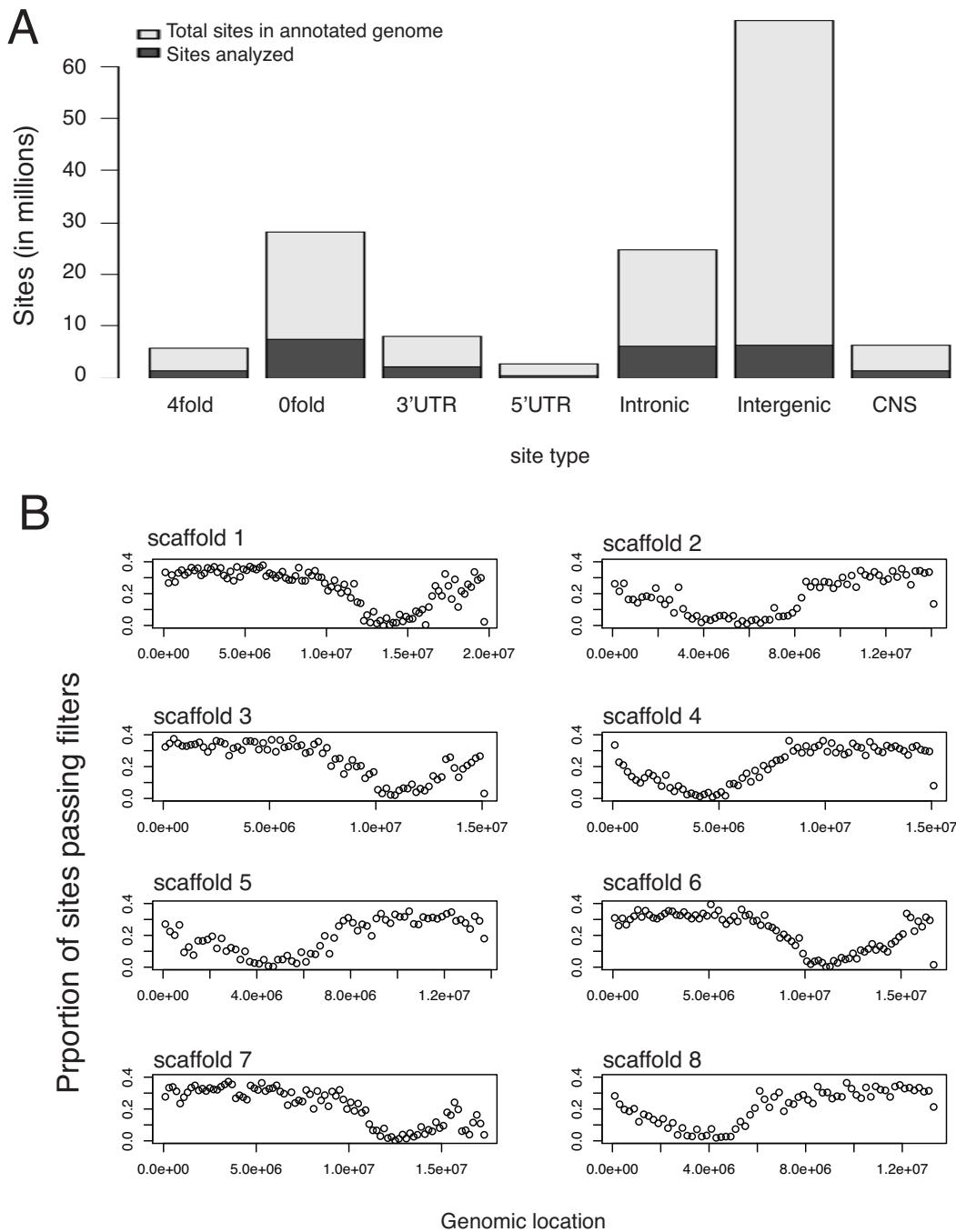


Figure 2.5: **Coverage after filtering, across the genome.** A) The number of annotated sites in each category across the genome (light grey), and the number of sites that pass our filters and were used in analysis (dark grey). B) The proportion of sites that pass filters, calculated in 200kb windows,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

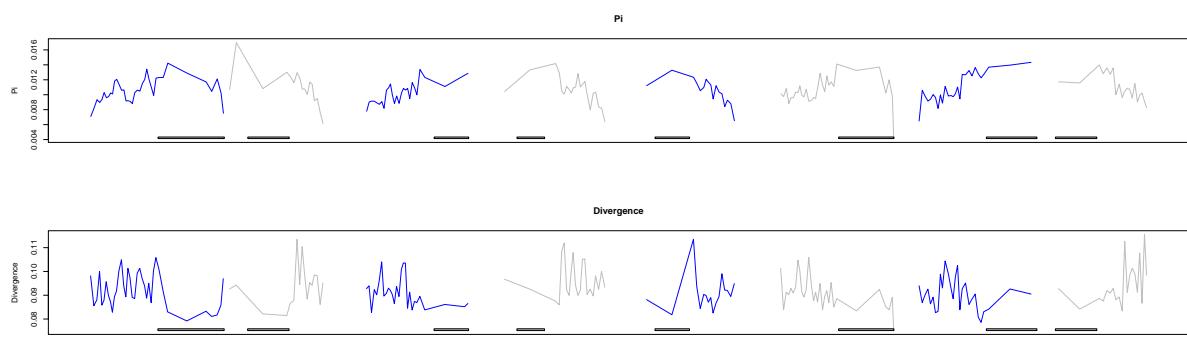


Figure 2.6: **Pairwise diversity and divergence at 4-fold degenerate sites across the entire genome.** Statistics were calculated in windows of 5,000 SNPs. Individual lines alternating between grey and blue represent chromosomes. The location of the centromere on each chromosome is indicated by the grey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

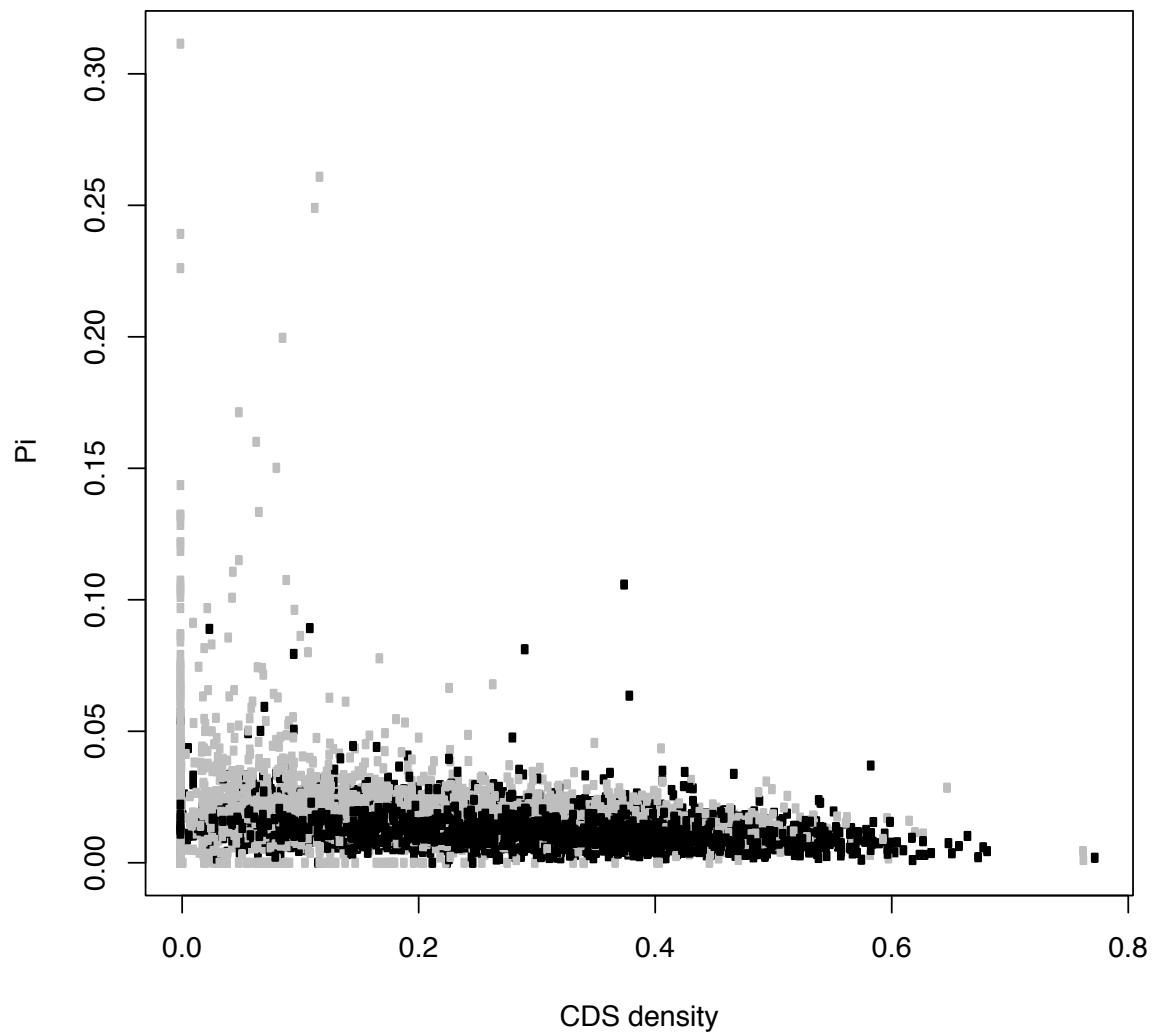
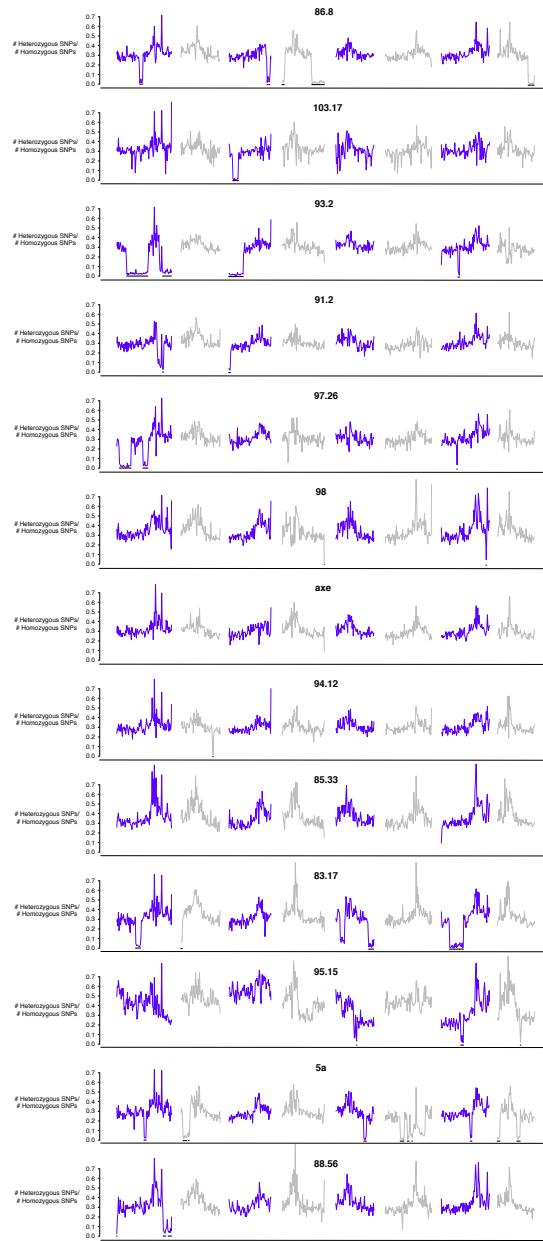


Figure 2.7: **Coding density versus 4-fold degenerate diversity across the genome.** Each point represents one 10 kb window. Black points represent windows that do not overlap centromeres while grey points represent windows that do overlap centromeres. There is a slight negative correlation between diversity 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.

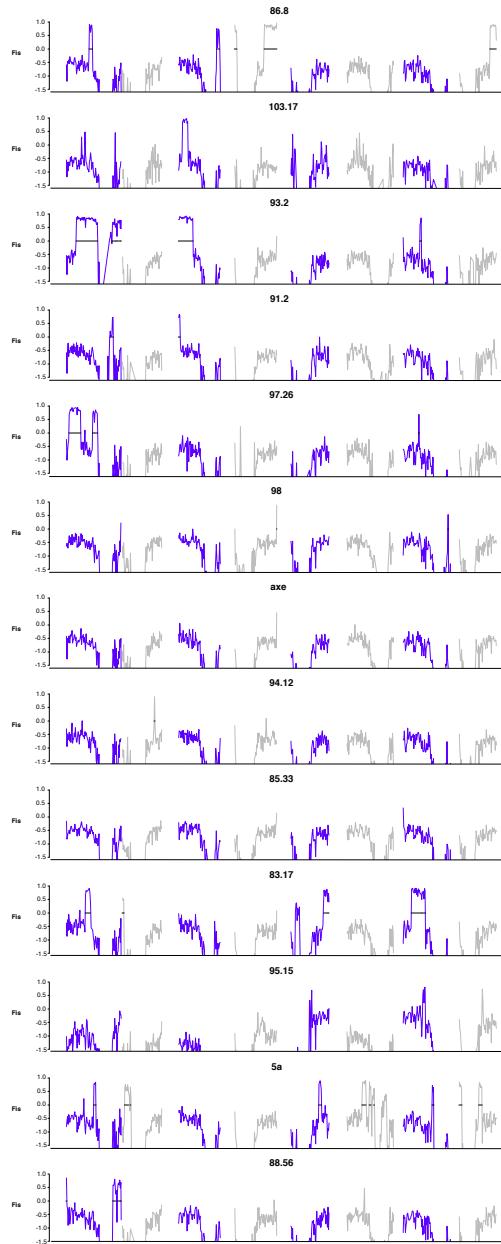


Figure 2.9: **FIS in windows across the genome in each sample.** FIS in 200kb windows is plotted across the genome. Each sample is plotted separately and identified by sample IDs. Individual lines alternating between grey and blue represent chromosomes. Regions of IBD were defined as windows where FIS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

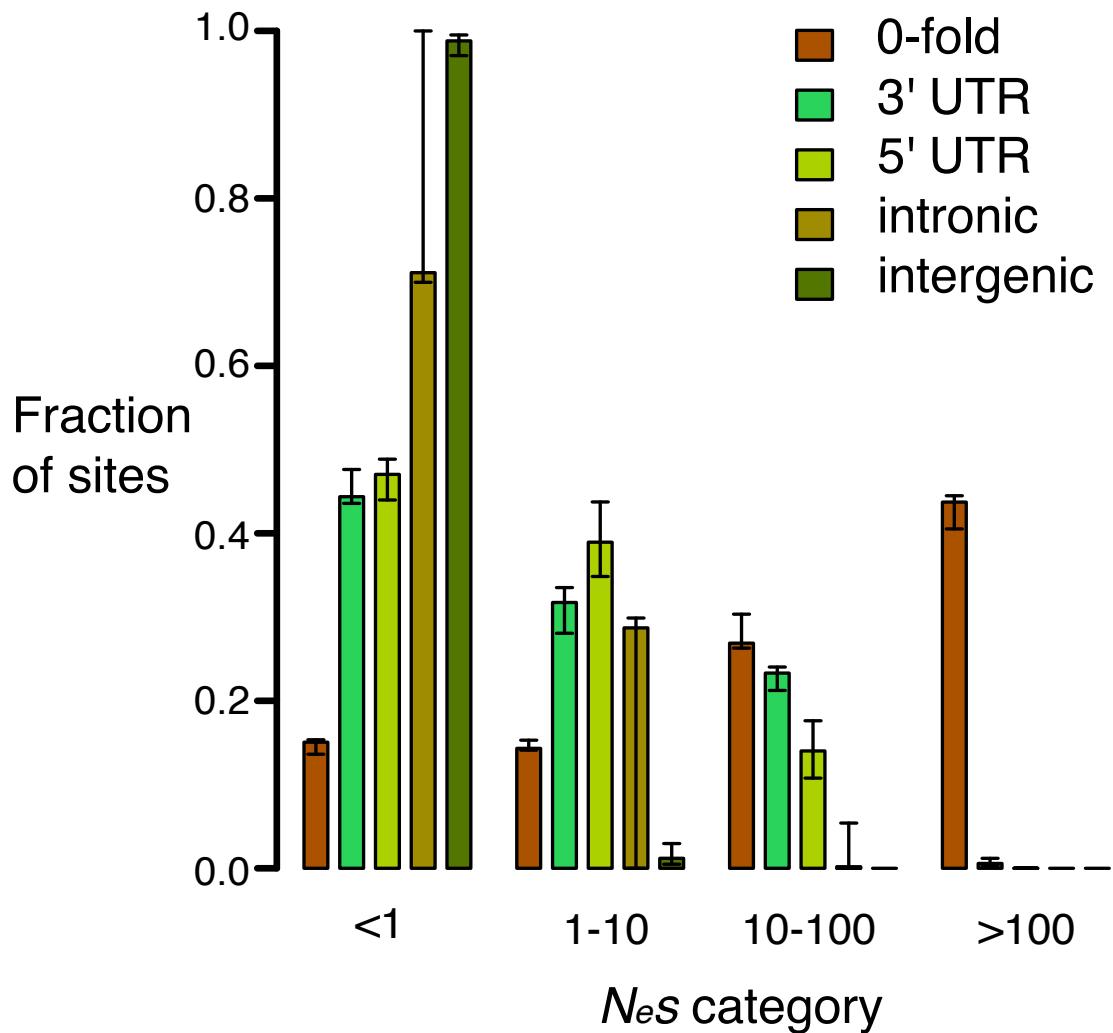


Figure 2.10: **DFE-alpha results using all alleles, including IBD regions.** The distribution of fitness effects for 0-fold degenerate, 3' and 5' UTR, intronic, and intergenic sites are shown. For this analysis the genotyping calls were filtered as described in the methods, but the data was not downsampled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

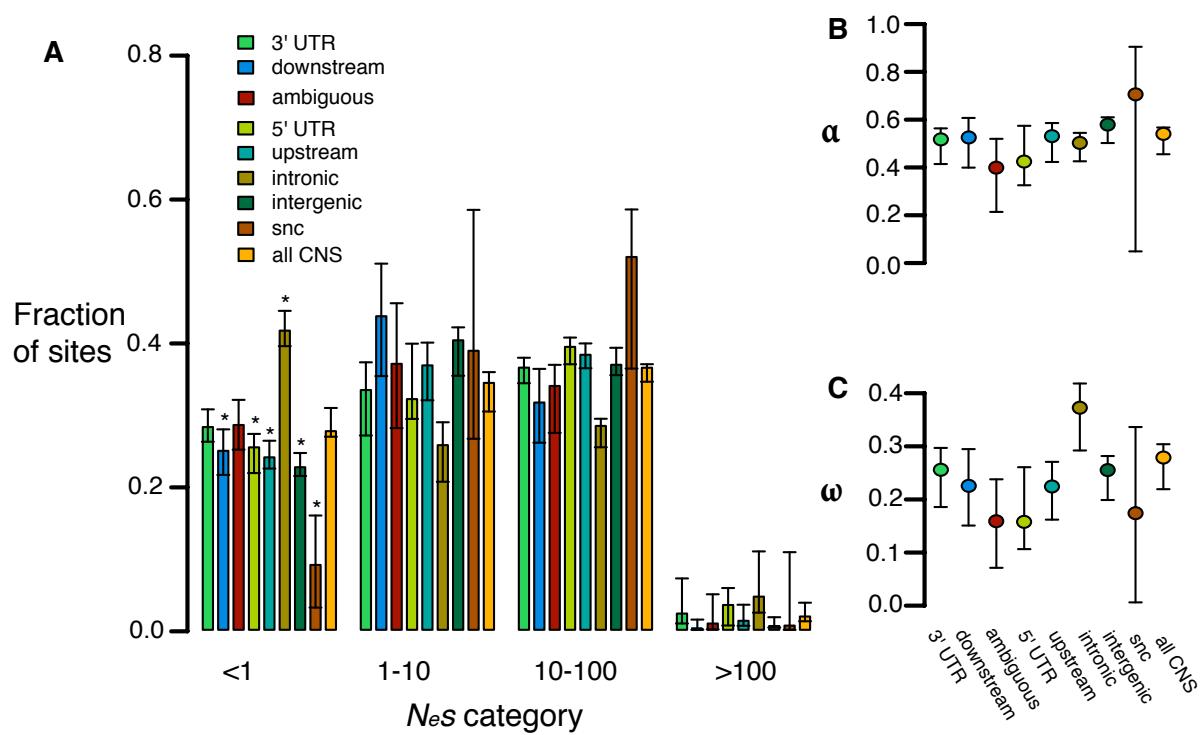


Figure 2.11: **Estimates of positive and negative selection on different categories of CNSs.** A) Distribution of fitness effects. Stars indicate categories in which the fraction of nearly neutral sites was significantly different from the pooled sets of CNSs by a randomization test. B)  $\alpha$  and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

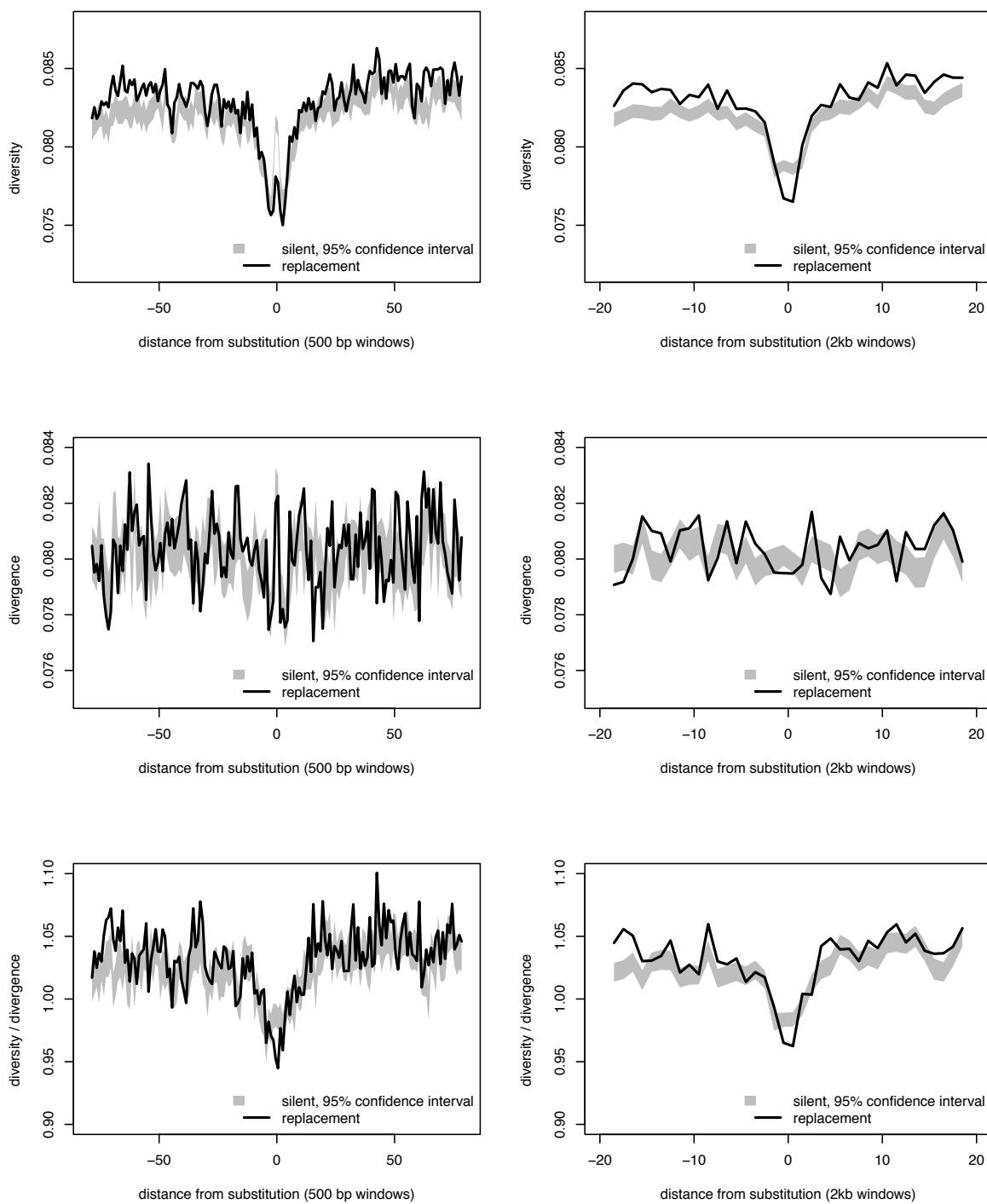
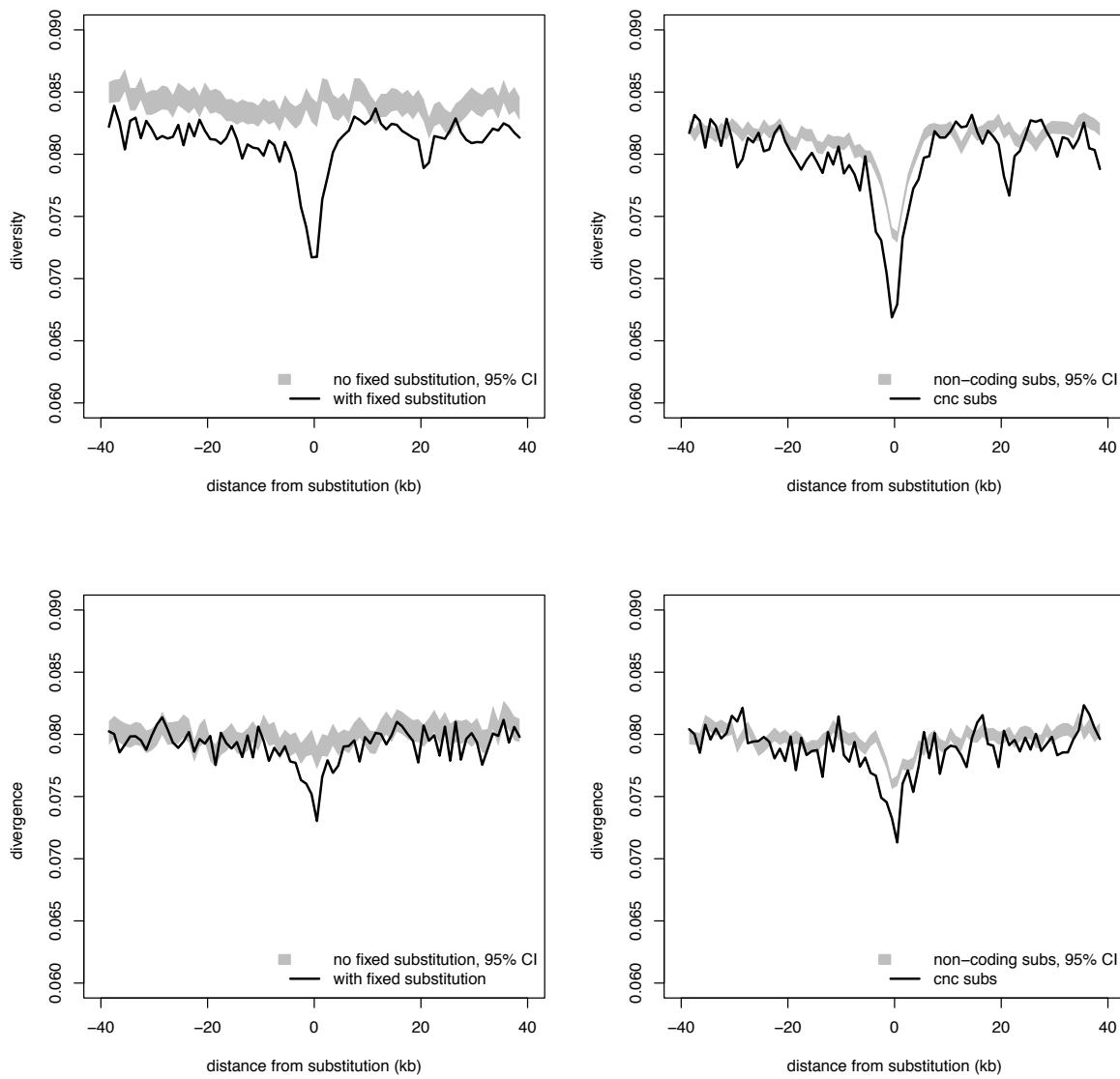


Figure 2.12: **Robustness of sweep analysis to different window sizes.** This panel shows the results of our scans for recurrent selective sweeps using alternative window sizes: 500bp on left and 2kb on right. Otherwise, the methods are the same as described previously.



**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).

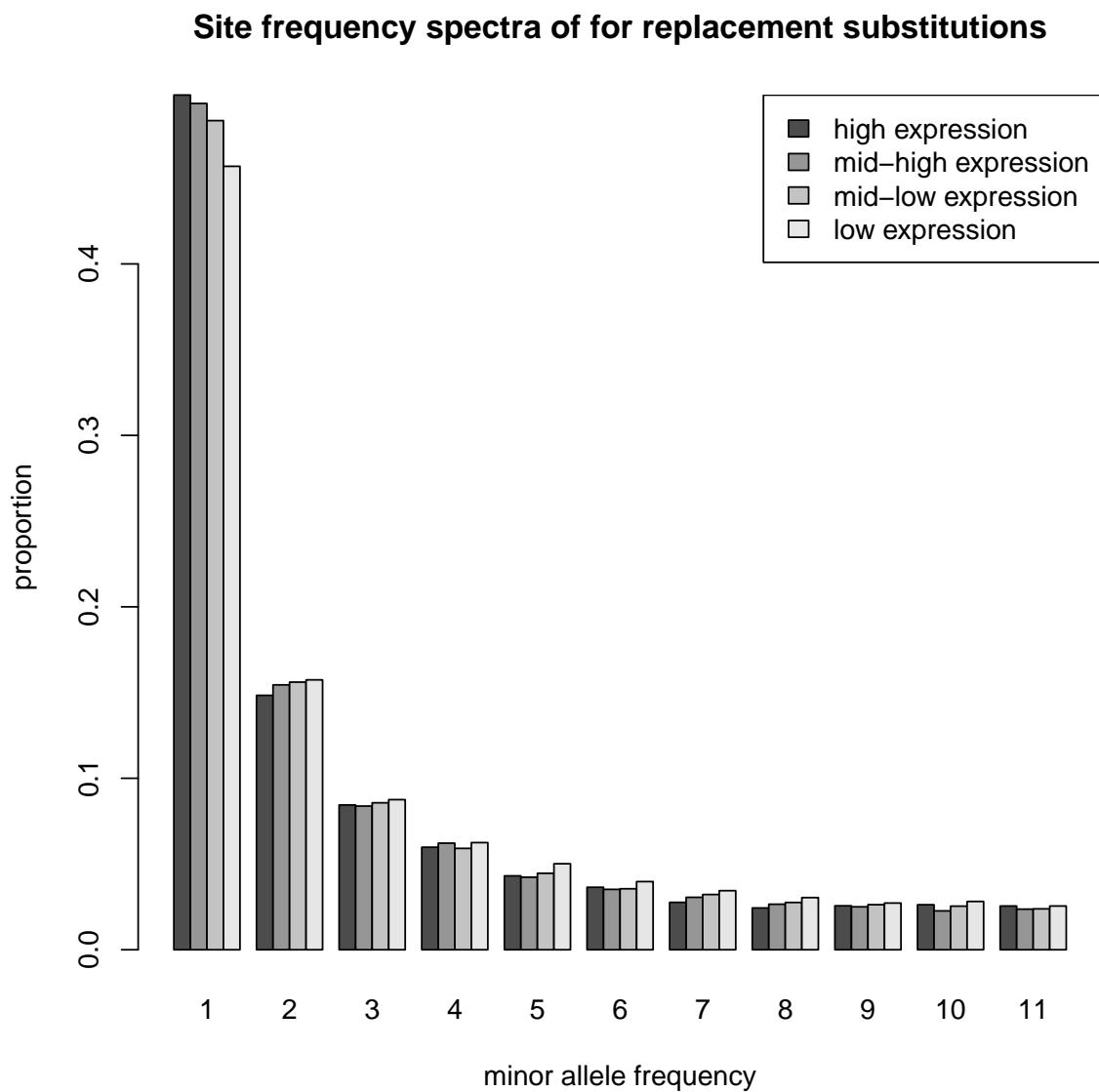


Figure 2.14: Allele frequency spectra of replacement sites in genes with different expression levels.

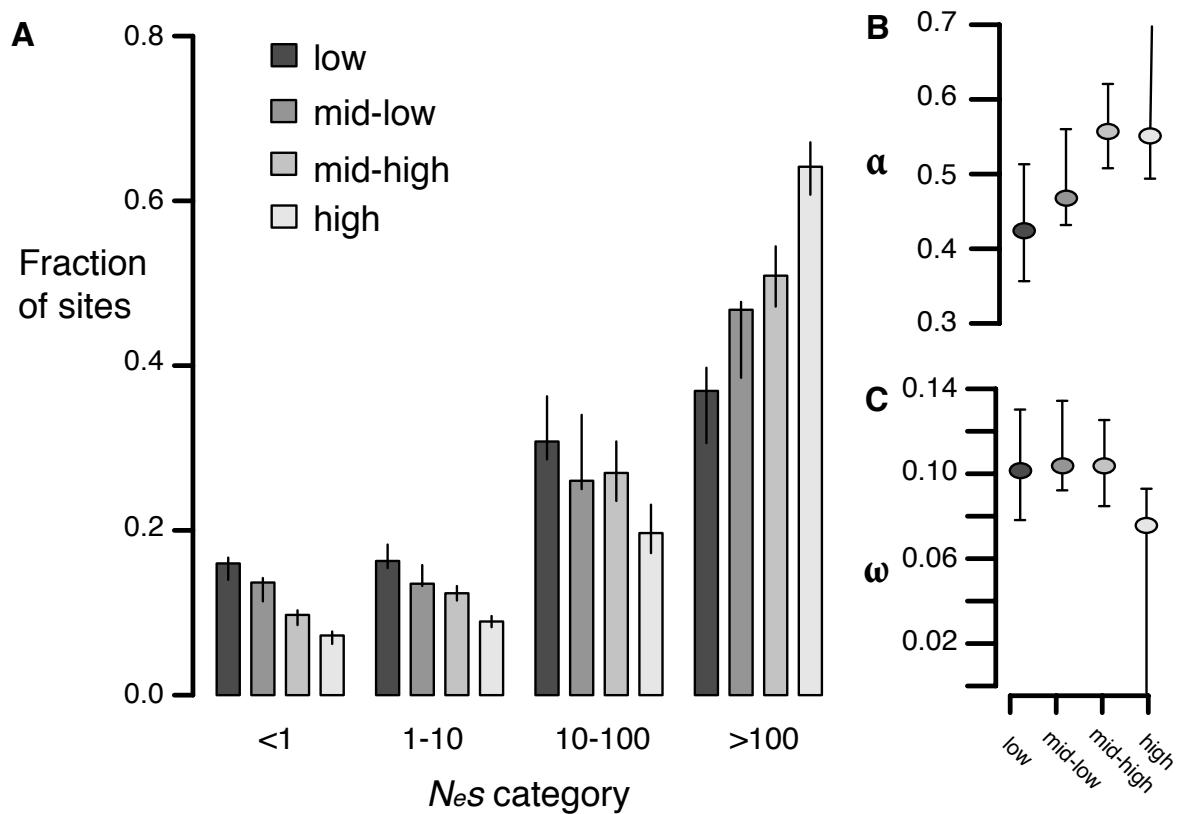


Figure 2.15: **Estimates of negative and positive selection on 0-fold sites in genes of varying expression level.** Data from this figure was generated using the divergence estimates from the whole genome alignments (as in Fig. ??) rather than divergence from PAML estimates (as in Fig. ??). Here AFS from 0-fold sites were compared to 4-fold sites, rather than non-synonymous to synonymous sites as in Fig. ???. 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 (?). While studies of allele frequency spectra show that purifying selection is often prevalent in genomic sequence (??), 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 (?). 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 (?).

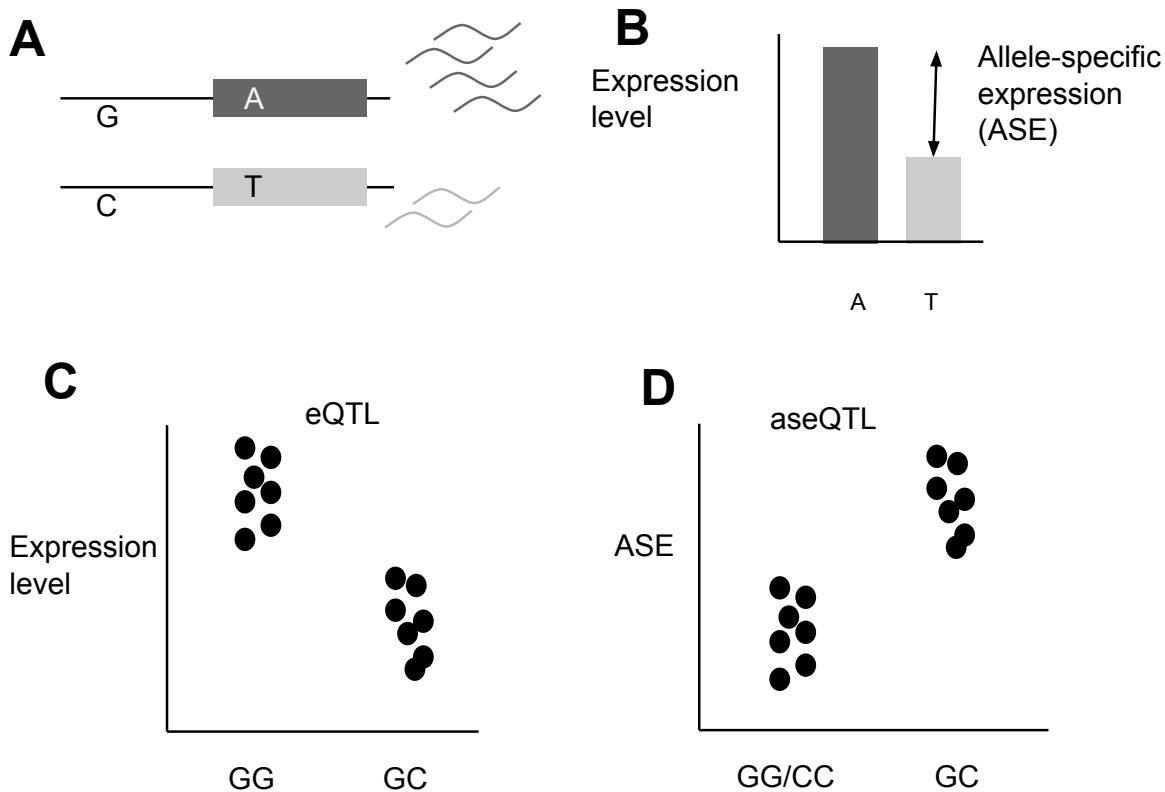
Early eQTL studies mapped expression divergence between two lines, finding that many genes have local expression QTL (??). These studies have provided insight into selection on eQTLs; for example, a correlation between recombination rate and eQTL density implies that background selection is a dominant force acting on expression variation in *Caenorhabditis elegans* (?) and a skew towards rare allele frequencies in promoters of genes with eQTLs suggests that purifying selection may act on expression variation (?). However, eQTL studies of population-level genetic variation have thus far been limited to a few study systems (?????) and only one study, in humans, has identified a negative correlation between phenotypic effect size and frequency (?). 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 (?), alleles eQTLs that increase expression of a potentially deleterious coding SNP are under stronger purifying selection than those that do not (?), and eQTL allele frequencies within populations are linked to local adaptation (??). To date, eQTL studies in plants have used genetic crosses (???) or species-wide samples (??), 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  $N_e$  and relative lack of population structure, purifying and positive selection are strong in *C. grandiflora* (??), 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  $r^2$  between SNPs less than 10bp apart is 0.25, and this decays to 0.12 within 100bp (Fig. ??). An analysis of population structure (?) 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.

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. ??). We omitted rare homozygotes from the analysis because most local regulation acts additively in *cis* (12) and low sample sizes for rare variants



**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.

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 ( $\text{FDR} = 0.1$ ,  $p < 8.2 \times 10^{-4}$ , Fig. ?? A). These SNPs often clustered locally (Fig. ?? 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 (?), 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.

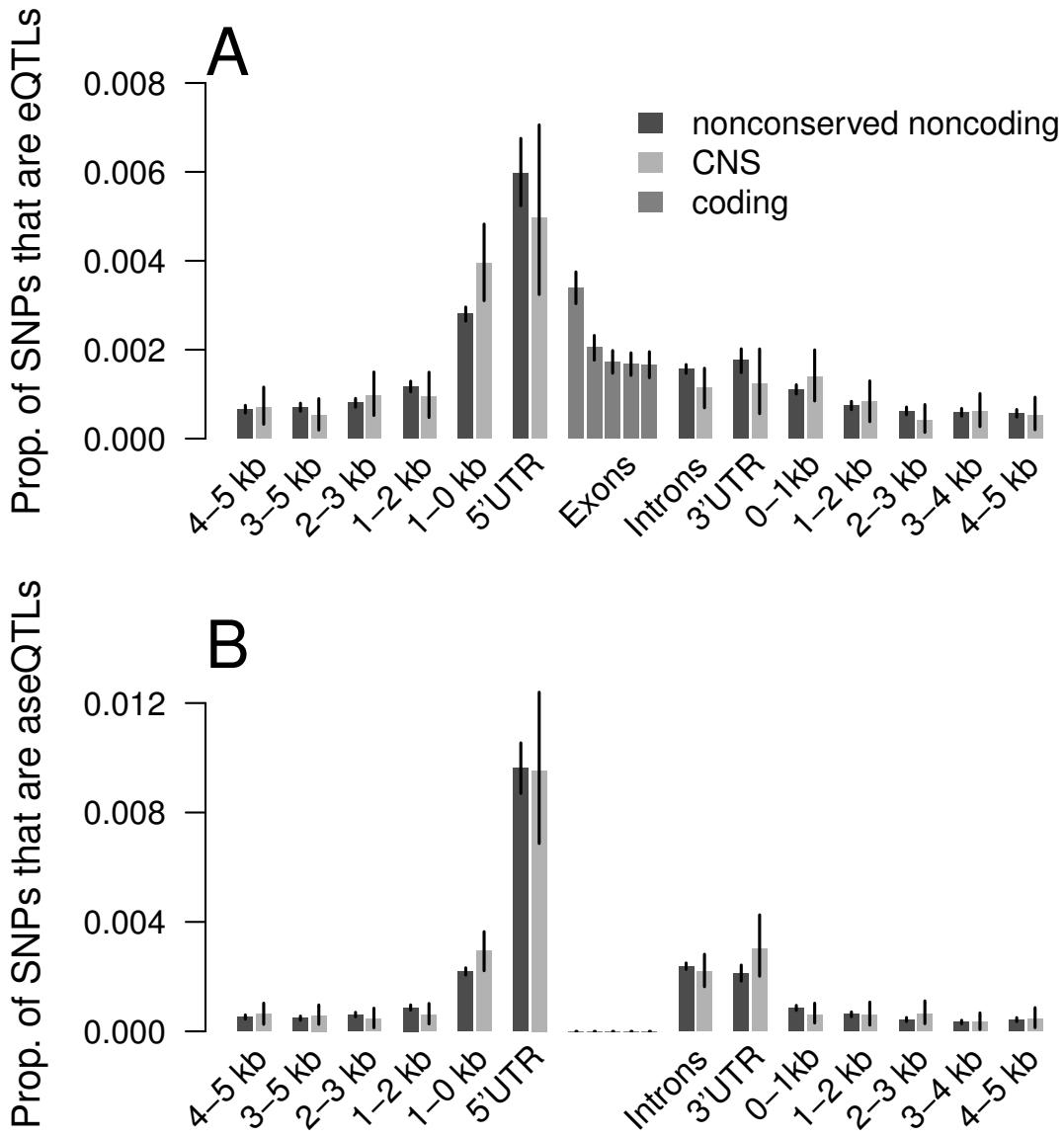
If eQTLs act in *cis*, heterozygous eQTLs will cause allele-specific expression (ASE), providing an additional signature of regulatory variation. We measured ASE within individuals by calculating the mean expression difference between alleles, standardized for sequencing depth. We then mapped QTLs for 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. ??). 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,  $p < 5.4 \times 10^{-4}$ , Fig. ?? 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. ?? 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* (??). 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 (?). However, the detection of a large number of eQTLs outside of conserved regions suggests that regulatory element turnover is common in Brassicaceae. There were 2,236 genes that had both eQTLs and aseQTLs, significantly more than expected by chance ( $X^2 = 471$ ,  $p < 2.2 \times 10^{-16}$ ). 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. ??A). 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 unpermuted data). As with eQTLs, aseQTLs were significantly rarer than those found in permuted data (Fig. ??B). These results hold when we designate a random significantly-associated SNP per gene as the eQTL or 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. ??



**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 C) 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.

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 mutation-selection balance will maintain mutations at frequencies inversely proportional to the strength of selection acting against them (?), suggesting that QTLs under purifying selection should show a negative correlation between minor allele frequency and phenotypic effect size, assuming that phenotypic

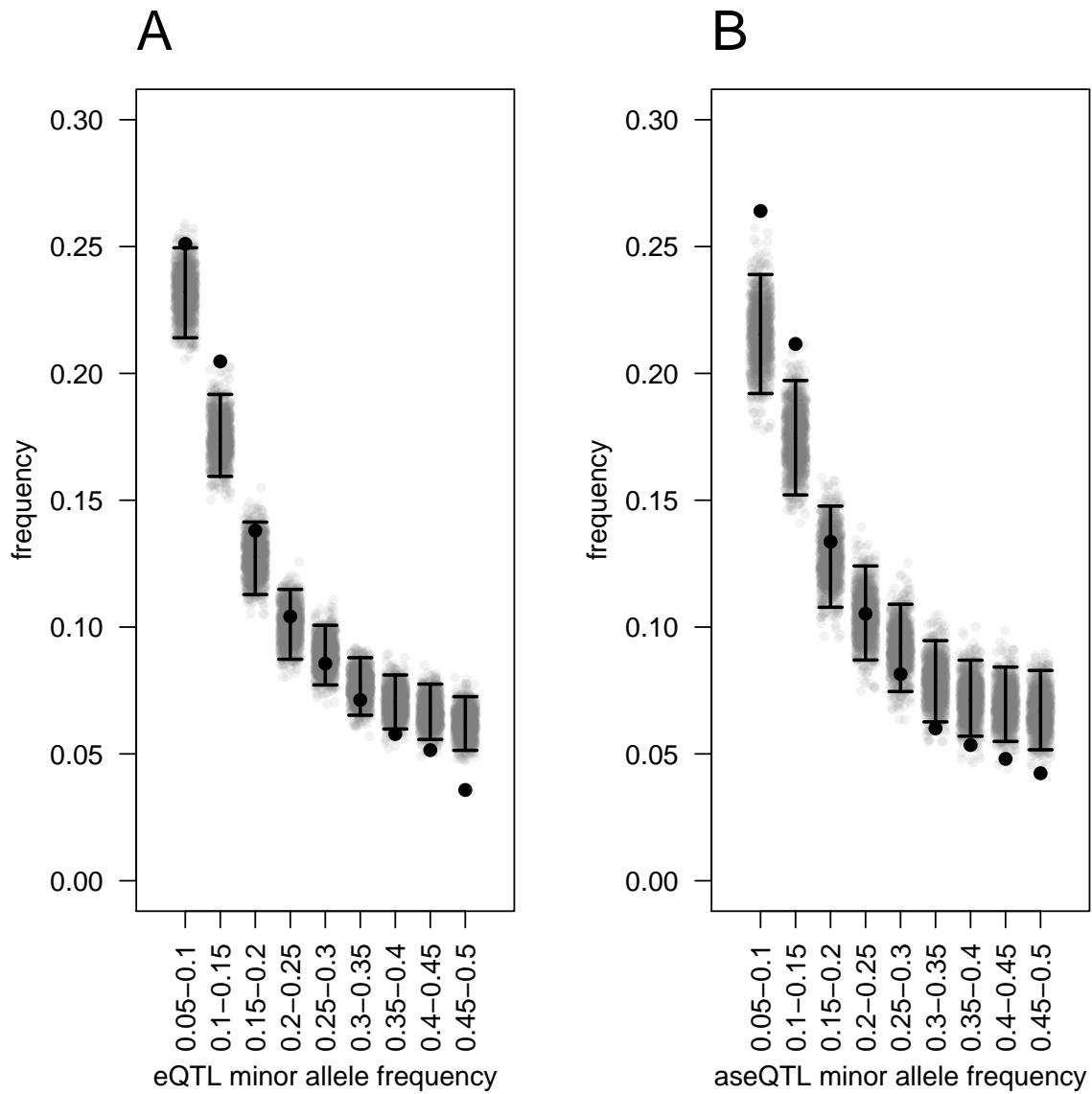


Figure 3.3: The site frequency spectra of eQTLs and aseQTLs. Minor allele frequencies of A) eQTLs and B) aseQTLs for observed data (black circles) and permuted data (gray circles, black lines are 95% confidence intervals).

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 (?).

To avoid variation in power across allele frequency, we repeated the eQTL and aseQTL analysis, down-sampling 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.

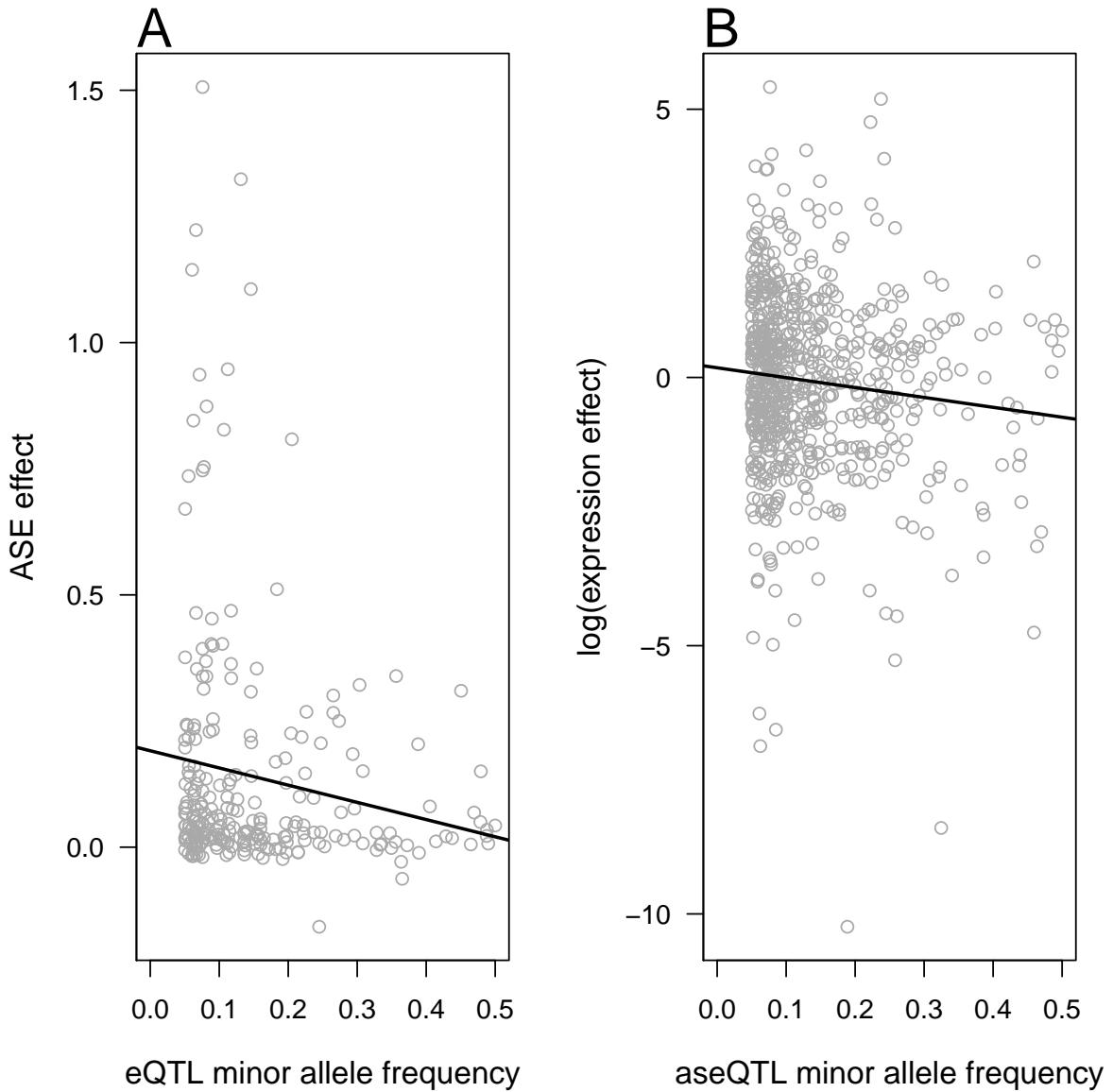
As a consequence, for every SNP we test, sample sizes of major and minor genotype classes are equalized regardless of allele frequency in the population. We also measured effect sizes in this subsample to avoid any relationship between allele frequency and effect size estimation error. Despite reducing 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;  $p < 2.6 \times 10^{-5}$  for eQTLs,  $p < 8.2 \times 10^{-5}$  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. ??). 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. ?? 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 down-sampling 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. ??). 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 (?). 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



**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.

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.

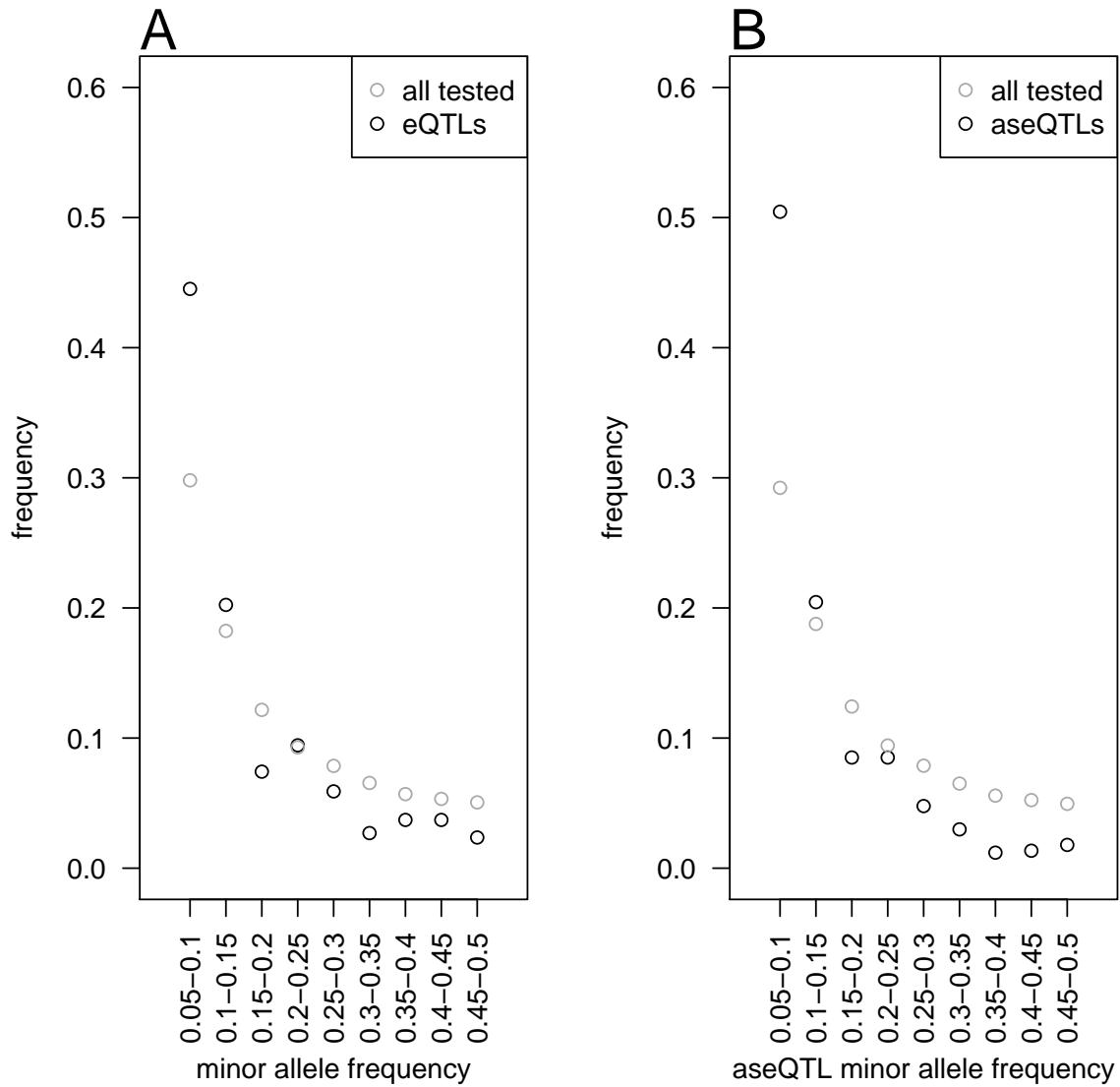


Figure 3.5: The site frequency spectra of QTLs detected in the frequency-controlled sub-sample.

### 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 ( $N_e \sim 600,000$ ), relatively low population structure and a range that spans northern Greece and southern Albania (??). 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 Murashige-Skoog salts (2.15 g/L), stratified in the dark at 4°C for one week, and then allowed to germinate in a growth chamber at 22°C 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 (?) with Stampy v1.0.19 (?). After bioinformatic processing with Picard tools, we realigned reads around putative indels with GATK RealignerTargetCreator 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* (?) 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 each other 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 (?).

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 (?). 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. ??). 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. ??). 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 (?). 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(?). In addition, we identified SNPs located in non-coding sequence conserved across the Brassicaceae family (?). 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. In addition, we compared common homozygotes to heterozygotes because we expect most local eQTLs to act in cis and thus be additive (?), 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 randomly 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) (?). 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 (?) 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 in ?). 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 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 (?). 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. 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 ( $> 3.45 \text{ cM/mB}$ ) and low recombination regions ( $\leq 3.45 \text{ cM/mB}$ ) using recombination rate data calculated by using a genetic map made from a cross between *C. rubella* and *C. grandiflora* (?). 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

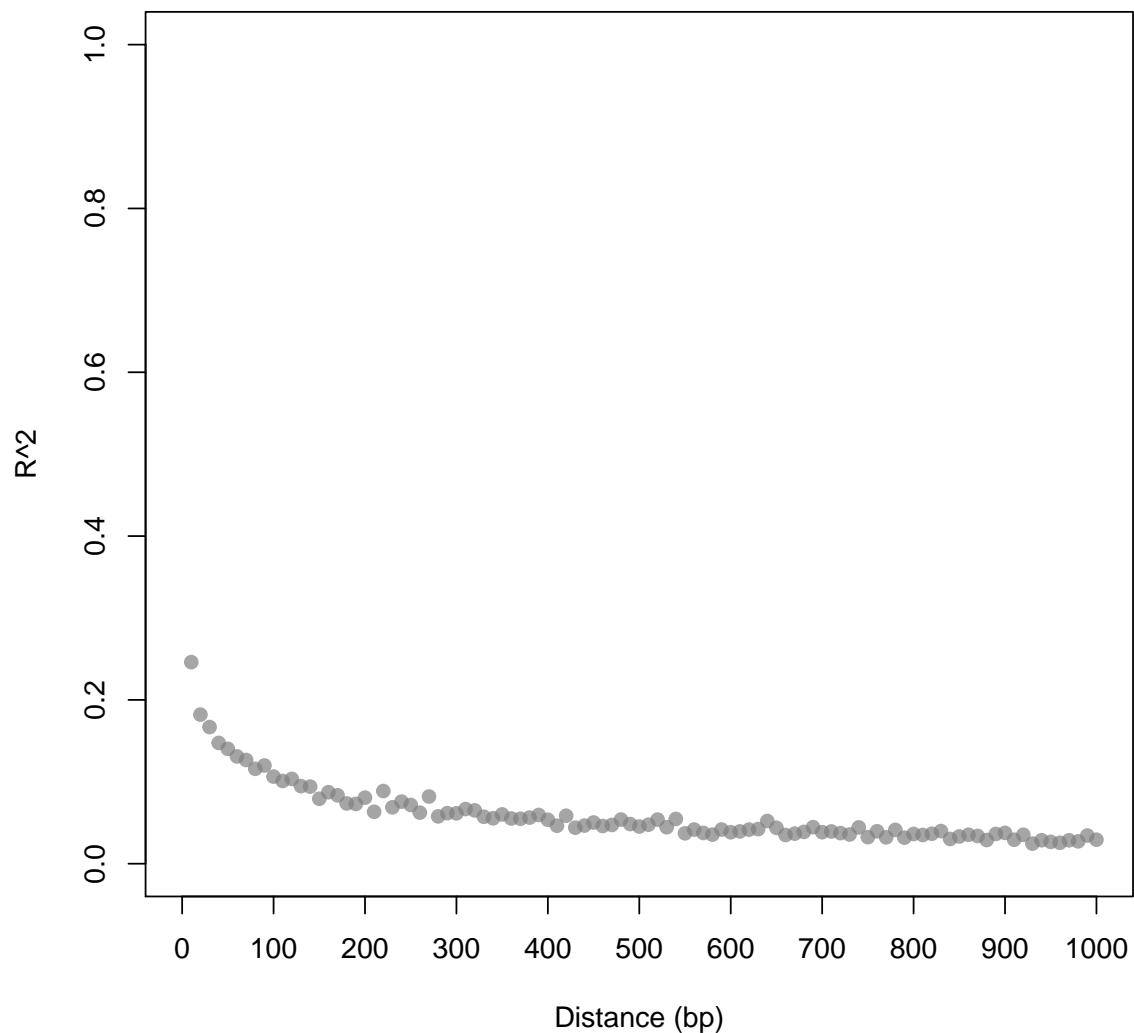


Figure 3.6: **Linkage disequilibrium in *C. grandiflora*.** Linkage disequilibrium was calculated for all SNPs within 1kb of each other on scaffold 2. 1% of these pairs were randomly sampled for the above figure, which shows mean  $r^2$  between pairs in 10bp bins.

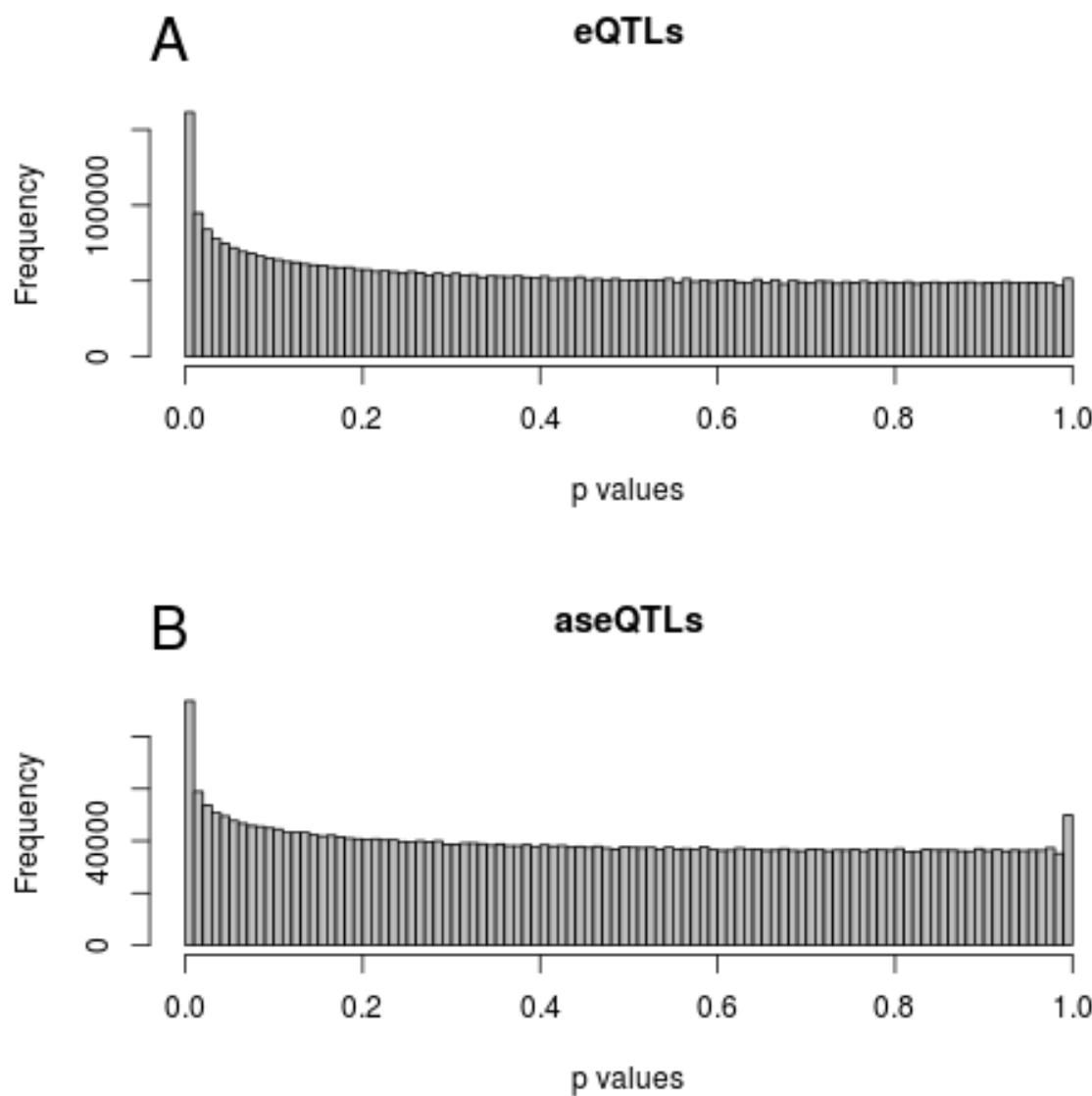
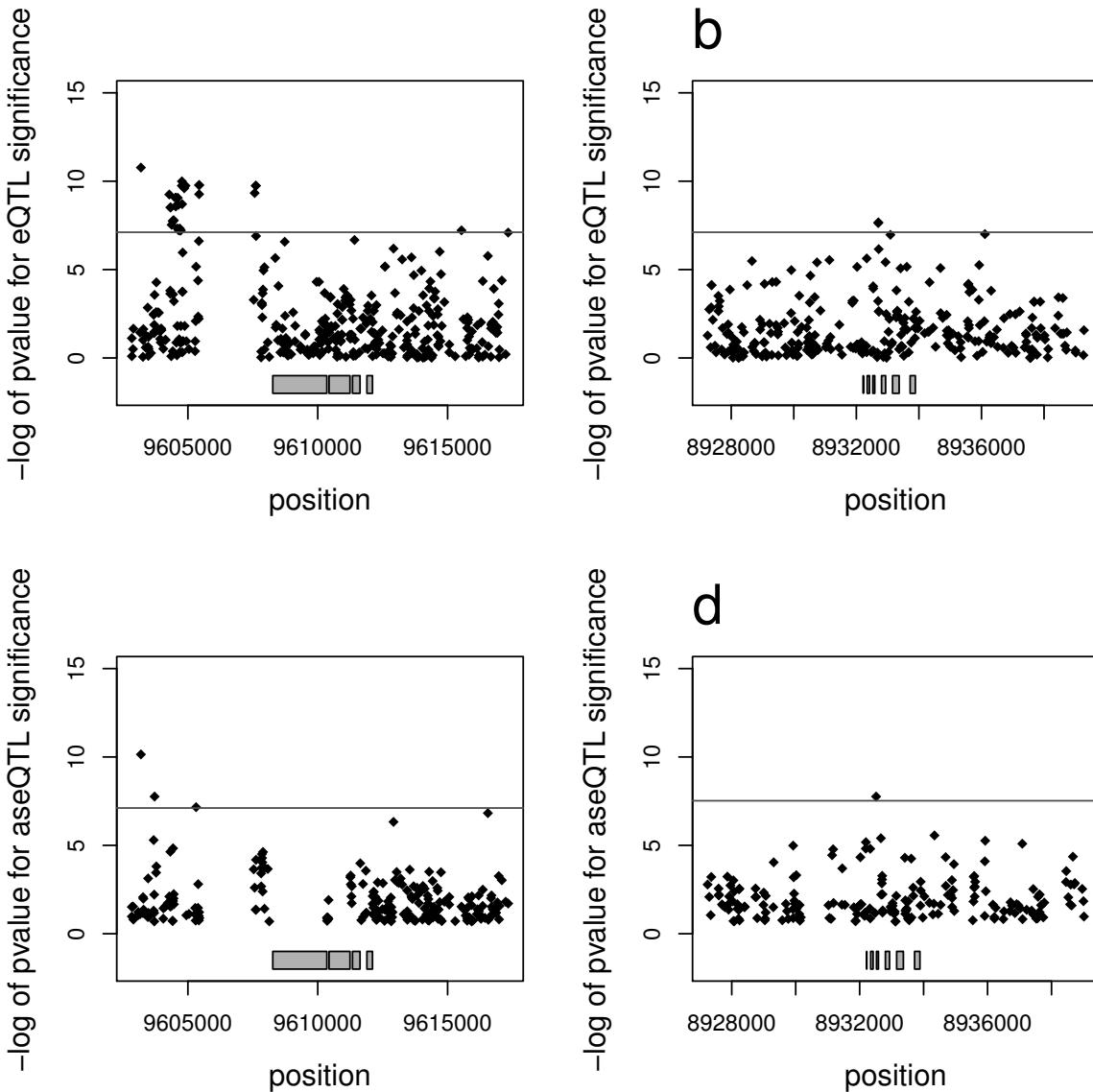
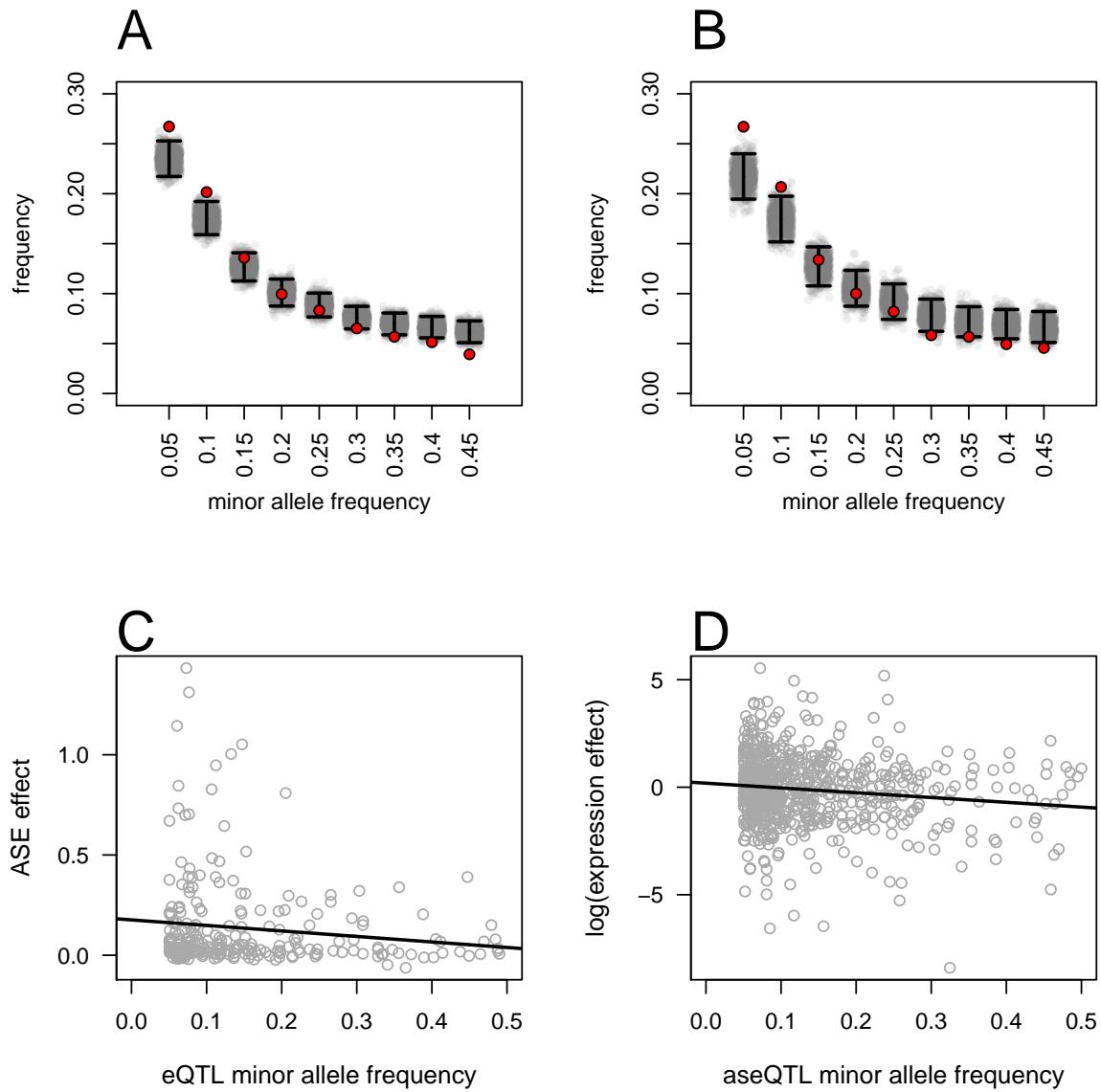


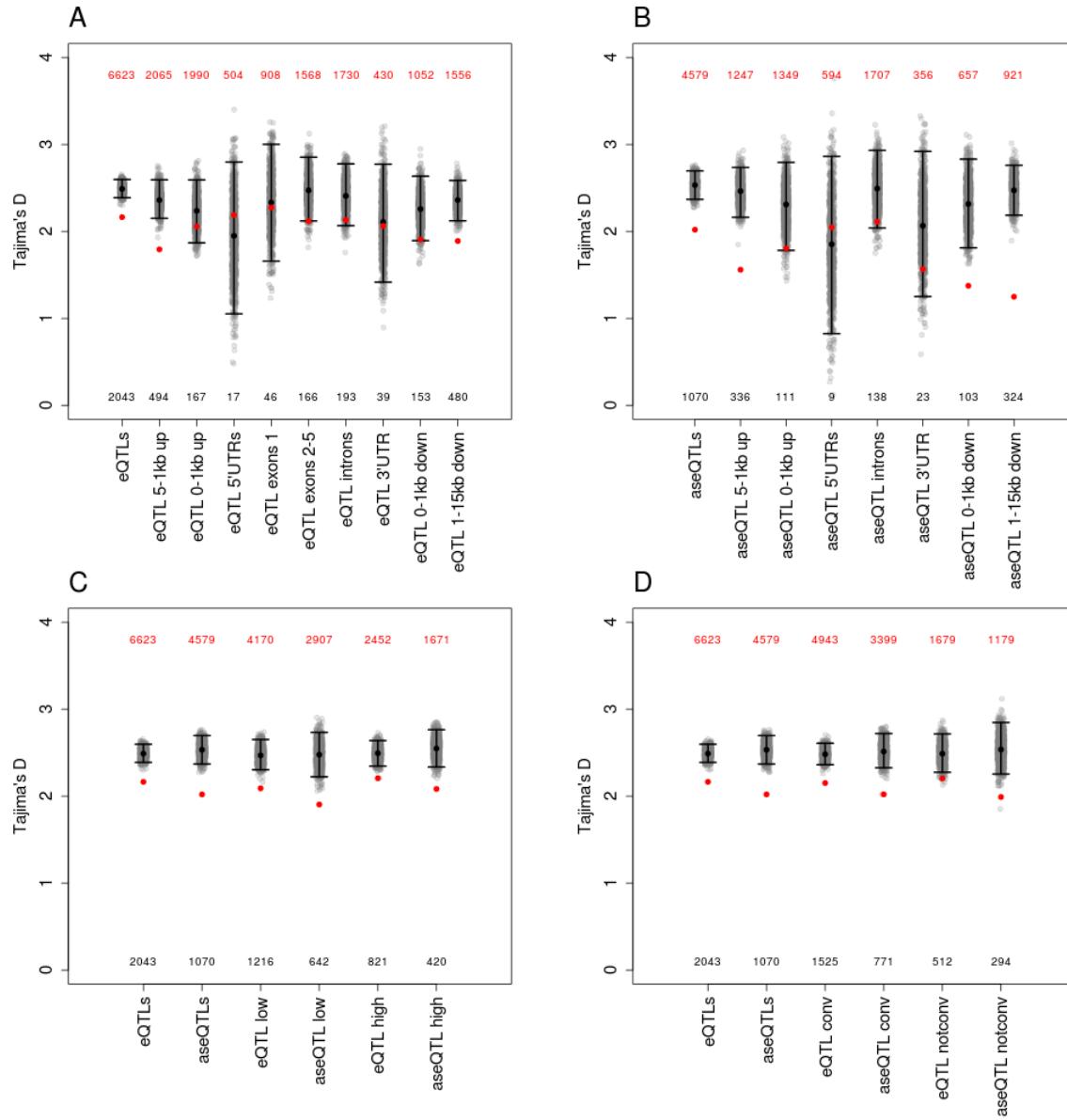
Figure 3.7: The distribution of p values for all SNPs tested in A) eQTL analyses and B) aseQTL analyses



**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 A) eQTLs and B) aseQTLs 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: Tajima's D of eQTLs and aseQTLs within site type, recombination rate, and substitution type.** A) Tajima's D for eQTLs of various categories. Red circles are the real data, gray circles show Tajima's D for permuted eQTLs, and black lines show 95% confidence intervals. Tajima's 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) The same data as A) but for aseQTLs. C) Tajima's 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) Tajima's D for A/T to G/C substitutions that could be favored by gene conversion (conv) and other substitutions (notconv). Note that all Tajima's 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, Tajima's 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.

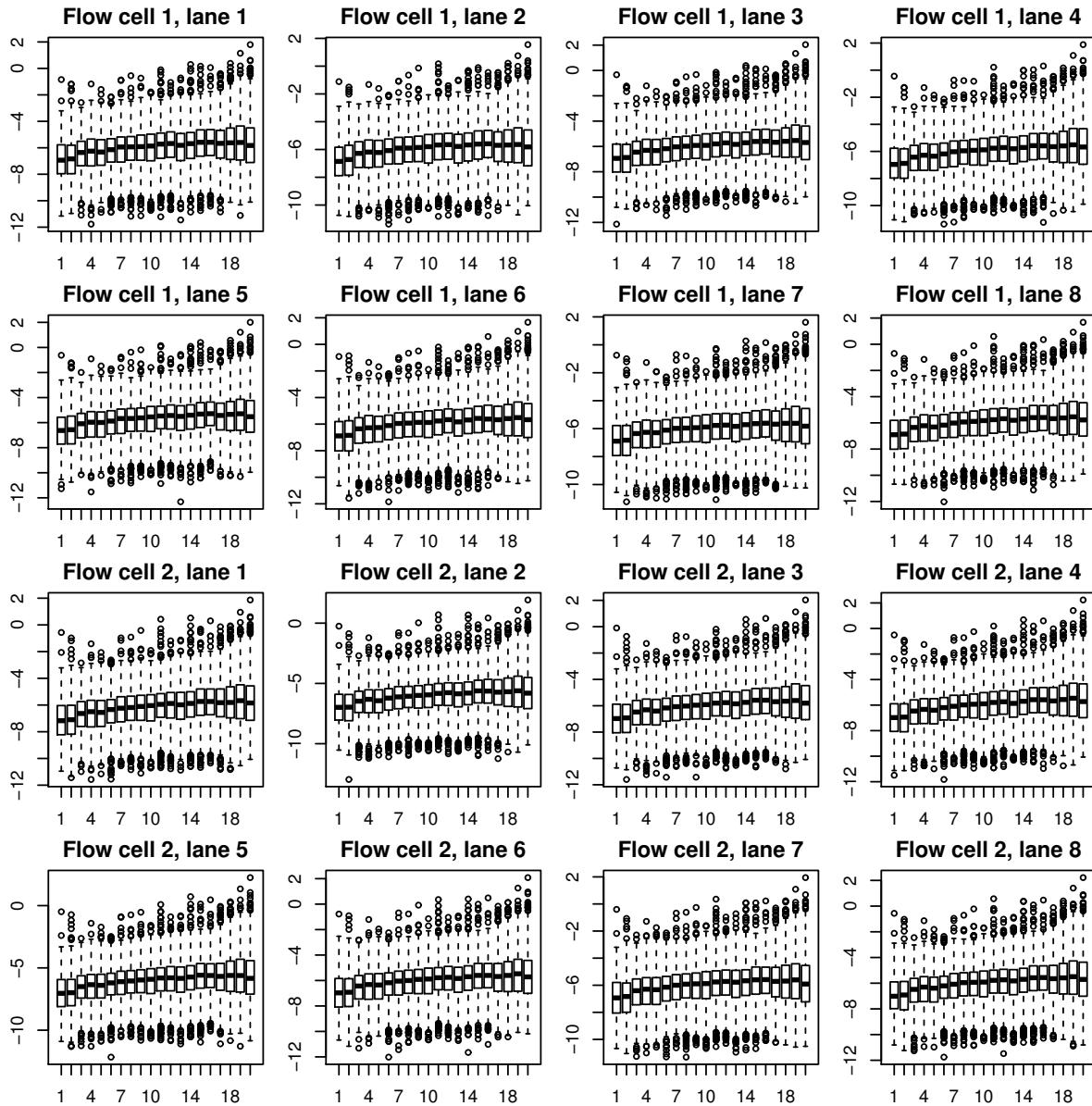
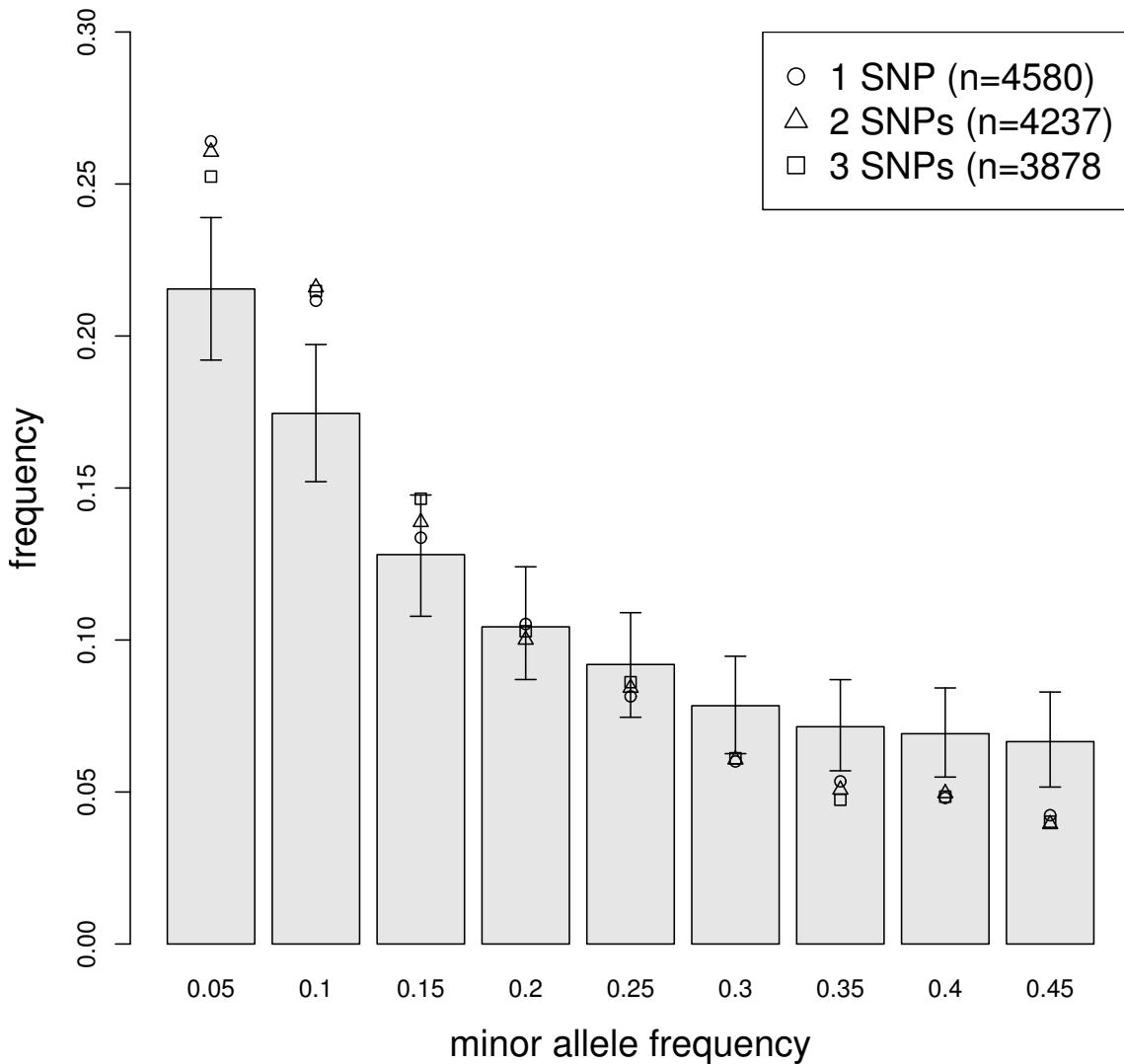


Figure 3.12: **GC composition and expression by lane.** All genes included in the study were split into 20 equally sized bins by GC content. Expression in these bins was combined for each lane and plotted in box plots.



**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.

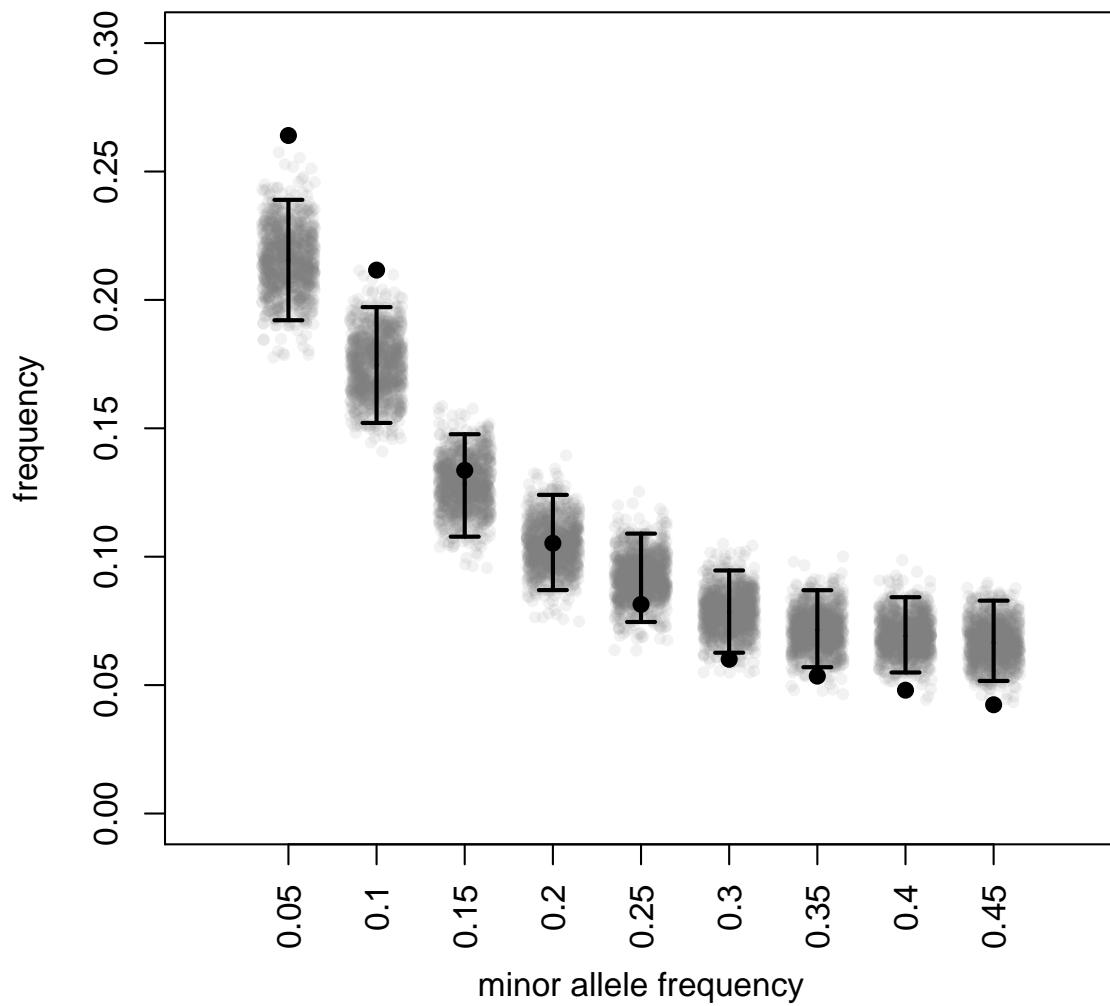


Figure 3.14: The site frequency spectrum of aseQTLs when genes with SNPs showing ASE bias are removed from the analysis. Red dots are frequencies of aseQTLs, gray dots are frequencies for permuted aseQTLs and black lines show 95% confidence intervals.

# Chapter 4

## Coexpression network connectivity affects selection

### 4.1 Abstract

The number and strength of interactions between genes will shape the selective effects of mutations in these genes. Specifically, genes with more interactions are likely to be more pleiotropic and may therefore be under stronger constraint and will be less free to adapt. We used coexpression network connectivity in the plant species *Capsella grandiflora* to investigate the relationship between connectivity and selection. We found that genes with high connectivity have lower amino acid divergence, even when controlling for expression. The relationship between connectivity and divergence appears to be driven by stronger negative selection in highly connected genes. However, positive selection is also stronger in more highly connected genes, implying that positive and negative selection have opposing effects on divergence. Overall, our results show that pleiotropy, as measured through network connectivity, increases selective constraint but does not reduce adaptive potential.

### 4.2 Introduction

Variation in quantitative traits results from the combined effects of genetic variation at many genes (?), and the ways in which these genes interact can influence how selection shapes this genetic variation. Specifically, genes that have large effects on a specific trait or with pleiotropic effects are likely to be under stronger purifying selection than genes with smaller effects or less pleiotropy (??). Similarly, genes with smaller effects on traits or reduced pleiotropy may be more subject to weaker evolutionary constraint and thus be more free to adapt (??). While these hypotheses are intuitively appealing, we lack still have lack data on how important pleiotropy genome-wide scale. One way to estimate the effect size and/or pleiotropy of genes on a genome-wide is to use network connectivity information: genes with higher connectivity are more likely to be pleiotropic or large-effect than genes with low connectivity (?).

Here, we investigate the relationship between positive and negative selection and network connectivity in the plant *Capsella grandiflora*.

A large body of work has shown that amino acid divergence is affected by network properties, including the number of protein-protein interactions (???, but see ??), protein-network centrality (?), pathway position (???), metabolic network connectivity (?), regulatory network centrality (?), and coexpression network connectivity (??). However, reduced amino acid divergence in highly connected genes has been interpreted to be due to either stronger constraint on well-connected genes (??) or more frequent positive selection in genes on the network periphery (??), both of which are possible, but are difficult to distinguish using just divergence-based approaches. In addition, network properties often correlate with expression level, which itself correlates with amino acid divergence, suggesting that expression variation may sometimes underlie observed relationships between amino acid divergence and connectivity (?). However, previous studies have not consistently controlled for the effects of expression on the relationship between connectivity and selection (??).

Here, we evaluate the relationship between coexpression network connectivity and both positive and negative selection by combining expression data with information on polymorphism and divergence in the plant *Capsella grandiflora*, which has previously been shown to have efficient positive and negative selection (?). We find that connectivity is negatively correlated with amino acid divergence and use polymorphism information to show that this correlation is driven by stronger negative selection on highly connected genes. In addition, we observe stronger positive selection in highly connected genes, suggesting that pleiotropy is not the main factor constraining adaptation across genes of different connectivities. Finally, we observe differences in the strength of positive and negative selection between modules. Collectively, our results suggest that connectivity influences selection on a genome-wide scale.

## 4.3 Results

We analyzed genome-wide gene expression data from the leaves of 146 *C. grandiflora* individuals from one large population grown in common garden. The expression data from 99 of these individuals has been previously reported (?). For the previously unreported 47 individuals, we generated approximately 1.3 billion single-end RNAseq reads, with a median of 26.7 million reads per individual (range: 19.4 to 44.3 million). Of these, a median of 94.2% reads mapped to genes (range: 93.1% to 94.9%).

### 4.3.1 Coexpression modules

Twenty-four gene coexpression modules were detected in the *C. grandiflora* expression data using WGCNA (?). These modules ranged in size from 43 to 4,094 genes. Clustering of the module eigengenes revealed that they fell into several broad groups of roughly equal size (Fig. ??). Total connectivity values for each gene were calculated as the sum of the connection strengths with all other genes and intramodular connectivity values were calculated as the sum of the connection strengths with the other genes in the same module. These connectivity values show the expected scale-free distribution: there is a linear relationship between the connectivity to a gene ( $k$ ) and the proportion of nodes with the

connectivity ( $p(k)$ ) when plotted on a log scale (slope = -1.61,  $R^2 = 0.96$ ). (Fig. ??). This scale free distribution means that most genes have relatively few connections, but some genes have a large number of connections. Total connectivity correlates strongly with intramodular connectivity (correlation coefficient = 0.807), so we chose to focus on intramodular connectivity for the rest of the paper.

Of the 24 *C. grandiflora* coexpression modules detected, eight modules showed a high degree of preservation in *Arabidopsis thaliana*, and another 12 showed a moderate degree of preservation (Table 1). All but one of the modules showed enrichment for GO-process terms that suggest specific biological roles for the different modules (Table 1).

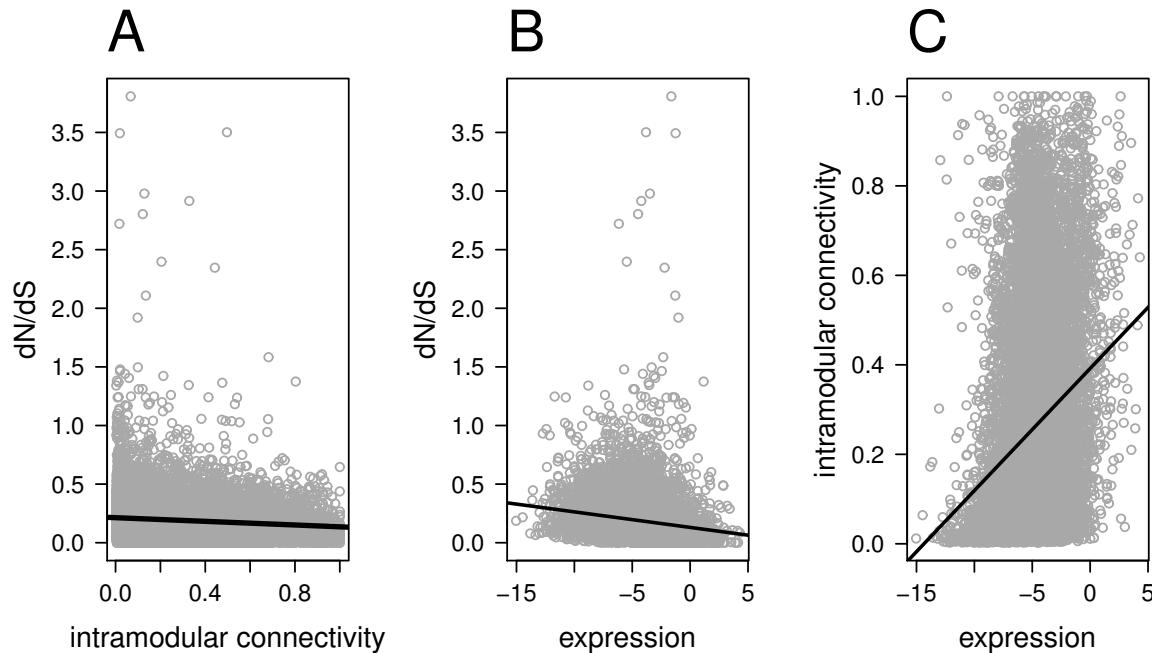
### 4.3.2 Connectivity affects divergence

Connectivity correlated negatively with  $d_N$  ( $\rho = -0.167$ ,  $p < 0.001$ , Fig. ??A),  $d_S$  ( $\rho = -0.104$ ,  $p < 0.001$ , Fig. ??B), and  $d_N/d_S$  ( $\rho = -0.127$ ,  $p < 0.001$ ) (Fig. ??A). The correlation between  $d_N/d_S$  and connectivity was significant within 7 out of 24 modules ( $p < 0.002$ ), but these were also the largest modules, where there was the most power to detect a correlation (Fig. ??). Expression also correlated negatively with  $d_S$  ( $\rho = -0.037$ ,  $p < 0.001$  Fig. ??C),  $d_N$  ( $\rho = -0.276$ ,  $p < 0.001$ , Fig. ??D), and  $d_N/d_s$  ( $\rho = -0.261$ ,  $p < 0.001$ , Fig. ??B), although the strength of the correlation with  $d_S$  was quite weak. (Fig. ??). Expression correlates positively with connectivity ( $\rho = 0.289$ ,  $p < 0.001$ , Fig. ??C) suggesting that expression variation might explain observed divergence patterns. However, partial correlations that account for expression variation still show a significant although weaker relationship between connectivity and  $d_N$  ( $\rho = -0.09$ ,  $p < 0.001$ ),  $d_S$  ( $\rho = -0.09$ ,  $p < 0.01$ ) and  $d_N/d_S$  ( $\rho = -0.05$ ,  $p < 0.001$ ).

However, experimental noise in measuring expression can make it difficult to fully remove expression effects in a partial correlation, potentially generating spurious correlations (?). To avoid this problem, we followed Drummond and colleagues recommendation and conducted a principal component analysis to identify independent sources of variation in our explanatory variables:  $d_S$ , expression, intramodular connectivity, and the coefficient of variation in expression (expression variation). We found that  $d_S$  and expression variation loaded positively on PC1, while traits expression level and intramodular connectivity loaded negatively; PC1 explained 32% of the variation across genes, while PC2, where all explanatory variables loaded negatively, explained 25% of the variation (Table ??). A linear model based on the four principal components and all possible interactions explained  $\sim 22\%$  of the variation in  $d_N$  (Table ??). The first principal component explained 14% of the variation in  $d_N$  and was determined by all four explanatory variables, suggesting that  $d_S$  and expression variation contribute positively to  $d_N$  while intramodular connectivity and expression level contribute negatively to  $d_N$ .

### 4.3.3 Connectivity affects positive and negative selection

Reduced divergence in highly connected genes could be caused either by stronger negative selection or weaker positive selection. To investigate these alternatives, we divided all genes that were expressed in *C. grandiflora* leaf tissue into four equally-sized categories based on intramodular connectivity and estimated the proportion of 0-fold degenerate sites in these categories under various strengths of purifying selection (?).



**Figure 4.1: Correlations between protein divergence, intramodular connectivity, and expression.** A)  $d_N/d_S$  correlates negatively with intramodular connectivity ( $\rho = -0.127$ ,  $p < 0.001$ ). B)  $d_N/d_S$  correlates negatively with expression ( $\rho = -0.261$ ,  $p < 0.001$ ). C) Expression correlates positively with intramodular connectivity ( $\rho = 0.289$ ,  $p < 0.001$ )

	PC1	PC2	PC3	PC4	All
% variation in $d_N$ explained	13.75	2.23	0.98	1.34	21.67
P value	0	$1.5 \times 10^{-82}$	$2.0 \times 10^{-37}$	$1.2 \times 10^{-50}$	
Direction of effect	+	-	+	-	
% of variance explained	32.9	25.5	21.5	20.6	100
Contributions to PCs					
$d_S$	0.512	-0.481	0.420	0.566	
Expression level	-0.539	-0.332	0.702	-0.328	
Expression variation	0.573	-0.328	-0.065	-0.748	
Intramodular connectivity	-0.344	-0.742	-0.564	0.110	

Table 4.1: Parameters and results from principal components analysis

The strength of negative selection across all expressed genes was similar to those observed in a previous study of a species-wide sample of *C. grandiflora* (Fig. ??A). However,  $\alpha$  (the proportion of all 0-fold degenerate fixations due to positive selection) and  $\omega$  (the rate of 0-fold degenerate fixations due to positive selection) were somewhat lower than observed in the species-wide sample:  $\alpha$  in the population sample was 0.32 compared to 0.41 in the species-wide sample and  $\omega$  was 0.06 compared to 0.08 in the species-wide sample (Fig. ??B, C) (?).

In general, negative selection acted more strongly on genes with higher connectivity (Fig. ??). For example, 0.193 (95% CIs: 0.186 - 0.197) of 0-fold degenerate sites in genes in the lowest connectivity category are effectively neutral ( $N_e S < 1$ ) while 0.115 (95% CIs: 0.112 - 0.127) of 0-fold degenerate sites in genes in the highest connectivity category have  $N_e S < 1$  ( $p < 0.01$ ) (Fig. ??A). Positive selection

also acted more strongly in highly connected genes.  $\alpha$ , was 0.209 (95% CIs: 0.295 - 0.476) in the lowest connectivity category and 0.455 (95% CIs: 0.395-0.476) in the highest connectivity category ( $p < 0.01$ )(Fig. ??B).  $\omega$  was also higher in highly connected genes (0.0834, 95% CIs: 0.0722-0.0895) than genes with low connectivity (0.0449, 95% CIs: 0.00740-0.0594) ( $p < 0.005$ )(Fig. ??C).

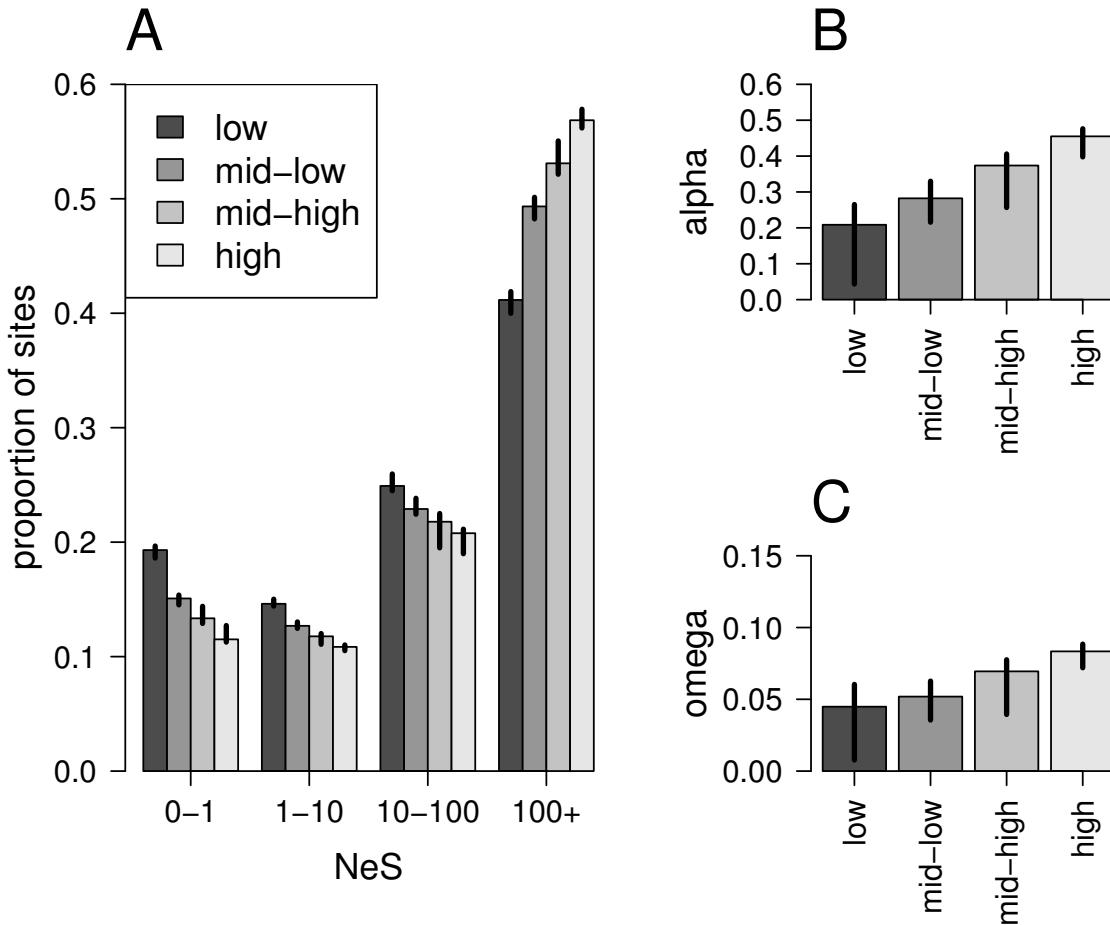
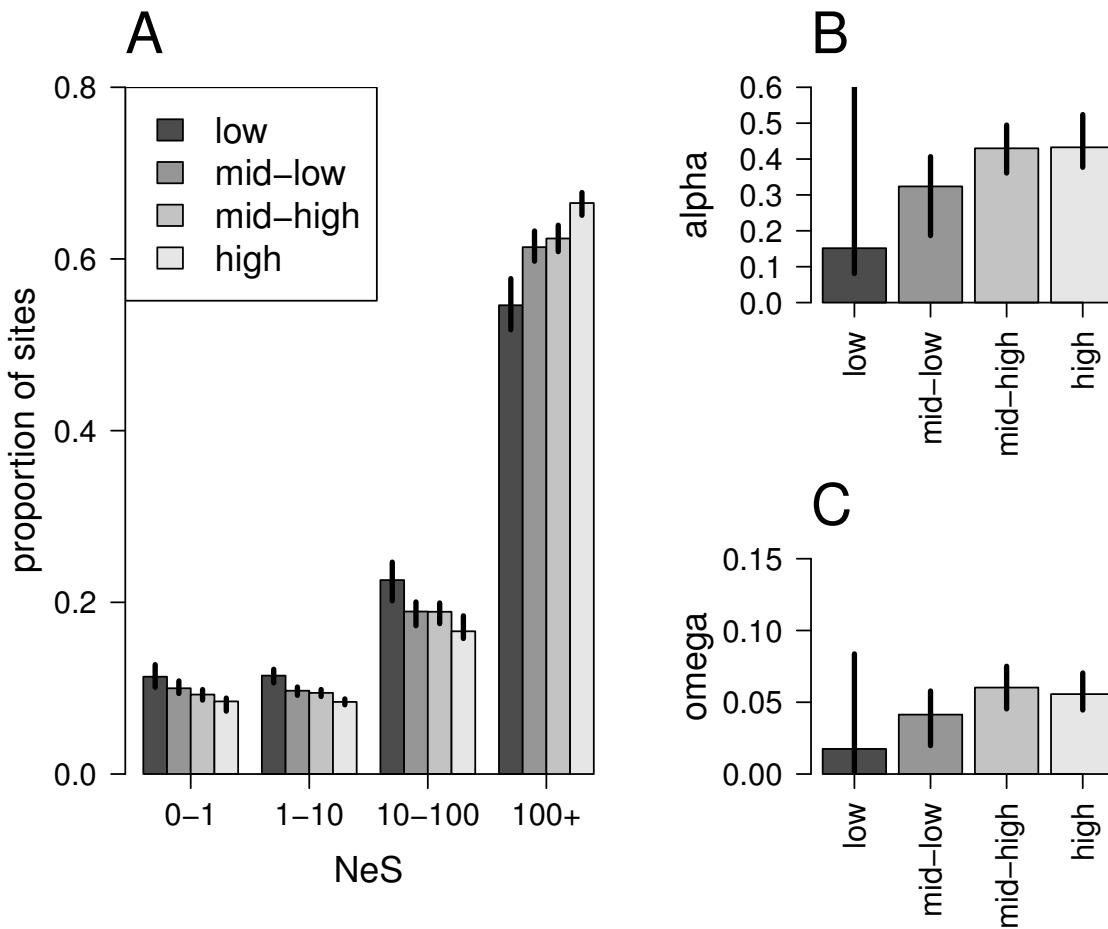


Figure 4.2: **Estimates of negative and positive selection in genes of different connectivity categories.** A) The proportion of sites found in each bin of purifying selection strength, separated by connectivity category. B) The proportion of sites fixed by positive selection, and C) the rate of adaptive substitution. Error bars represent 95% confidence intervals.

However, expression level could also explain stronger negative selection in highly connected genes, since expression level is correlated with connectivity and genes with higher expression in leaf tissue in *C. grandiflora* experience stronger negative selection (?). To control for expression level in this analysis, we found a subset of our dataset within a narrow range of expression levels where expression level and connectivity were not correlated and repeated our analysis of selection on this subset of genes ( $N = 2,714$ ). Within this subset, negative selection was still stronger on genes with higher connectivity values, although this effect was somewhat weaker: 0.113 (95% CIs: 0.100 - 0.129) of sites in the lowest connectivity category were effectively neutral ( $N_e S < 1$ ) while only 0.08 (95% CIs: 0.0741-0.0882) of 0-fold degenerate sites in the highest connectivity category were neutral ( $p < 0.01$ ) (Fig. ??A).

Since no relationship between the rate of positive fixations ( $\omega$ ) and expression has been observed (?), it

seems likely that the differences in  $\omega$  observed between connectivity classes in the entire dataset are free of the confounding effects of expression. However, we still investigated positive selection in our subset and found that positive selection was stronger on highly connected genes, although reduced sample size made it difficult to calculate measures of positive selection for the lowest connectivity category, which had only 251 genes. Still, the trend suggests that positive selection is stronger in highly connected genes.  $\alpha$  is 0.433 (95% CIs: 0.371 - 0.510) in genes in the highest connectivity class and 0.324 (95% CIs: 0.182 - 0.411) in the mid-low connectivity category ( $p < 0.05$ ) (Fig. ??B). Similarly,  $\omega$  was higher in the highest connectivity class (0.0557, 95% CIs: 0.0443-0.0682) than in the mid-low connectivity class (0.0413, 95% CIs: 0.0191-0.0585,  $p = 0.19$ ) (Fig. ??C).



**Figure 4.3: Estimates of negative and positive selection in genes of different connectivity categories within the subset where expression does not correlate with connectivity.** A) The proportion of sites found in each bin of purifying selection strength, separated by connectivity category. B) The proportion of sites fixed by positive selection, and C) the rate of adaptive substitution. Error bars represent 95% confidence intervals.

#### 4.3.4 Selection across coexpression modules

The strength of positive and negative selection varies across modules (Fig. ??). The proportion of 0-fold degenerate sites evolving neutrally ( $N_e S < 1$ ) ranged from 0.102 (grey60) to 0.216 (skyblue3) (Fig. ??A). The proportion of 0-fold degenerate fixations due to positive selection within modules ranged from 0.412 (skyblue3) to 0.531 (yellowgreen) and the rate of 0-fold degenerate fixation due to positive selection ranged from 0.025 (magenta) to 0.145 (yellowgreen) (Fig. ??B). There was no correlation between the proportion of 0-fold degenerate sites in a module that evolve neutrally ( $N_e S < 1$ ) and the rate of fixation due to positive selection ( $\omega$ ). (Pearson correlation,  $p = 0.34$ ).

We investigated the GO processes enriched in modules that experience the strongest positive selection to get a better picture of what biological processes may be under selection in our population. The module that experiences the strongest positive selection (yellowgreen,  $\alpha = 0.531$ ,  $\omega = 0.147$ ) had a number of enrichments for GO processes, including response to chitin, response to organonitrogen compound, response to nitrogen compound, and a number of processes involved in defense response and immune systems. (Table 1). The module with the second highest  $\alpha$  and third highest  $\omega$  (lightsteelblue1,  $\alpha=0.517$ ,  $\omega = 0.103$ ) also had enrichments in defense related genes. (Table 1). The module with the second highest  $\omega$  (0.1046) and seventh highest  $\alpha$  (0.403), saddlebrown, had enrichments in processes involved in DNA replication and epigenetic regulation of gene expression.

For the most part, modules under weak positive selection had enrichments in different processes than those under strong positive selection. For example, the magenta module ( $\alpha = 0.166$ ,  $\omega = 0.025$ ) had enrichments in phosphatidylinositol biosynthetic and metabolic processes, but the darkgrey module ( $\alpha=0.182$ ,  $\omega = 0.027$ ) did not have any significant enrichments. The midnightblue module, with the third lowest measures of positive selection ( $\alpha = 0.189$ ,  $\omega=0.030$  ) had enrichments in response to chitin, organonitrogen compounds, and nitrogen compounds. The surprising result that both a module under the strong positive selection (yellowgreen) and a module under weak positive selection (midnightblue) share similar GO processes suggests that not all the genes in this GO process are under the same selective pressures and that our coexpression analysis was able to separate genes by selective pressure.

## 4.4 Discussion

In this study we investigated the relationship between connectivity and other coexpression network properties and selection in the plant *C. grandiflora*. Intramodular connectivity was negatively associated with  $d_N/d_S$ , suggesting that constraint is stronger on highly connected genes. However, when we analyzed positive and negative selection separately, we found that connectivity increases the strength of both negative and positive selection. Finally, there were significant differences in the strength of selection between coexpression modules.

Most previous work on the relationship between connectivity and selection has focused on amino acid divergence, making it difficult to separate the effects of positive and negative selection. However, by using polymorphism to investigate the distribution of fitness effects and, then, an extension of the McDonald-Kreitman test to estimate positive selection, we were able to detect positive selection and

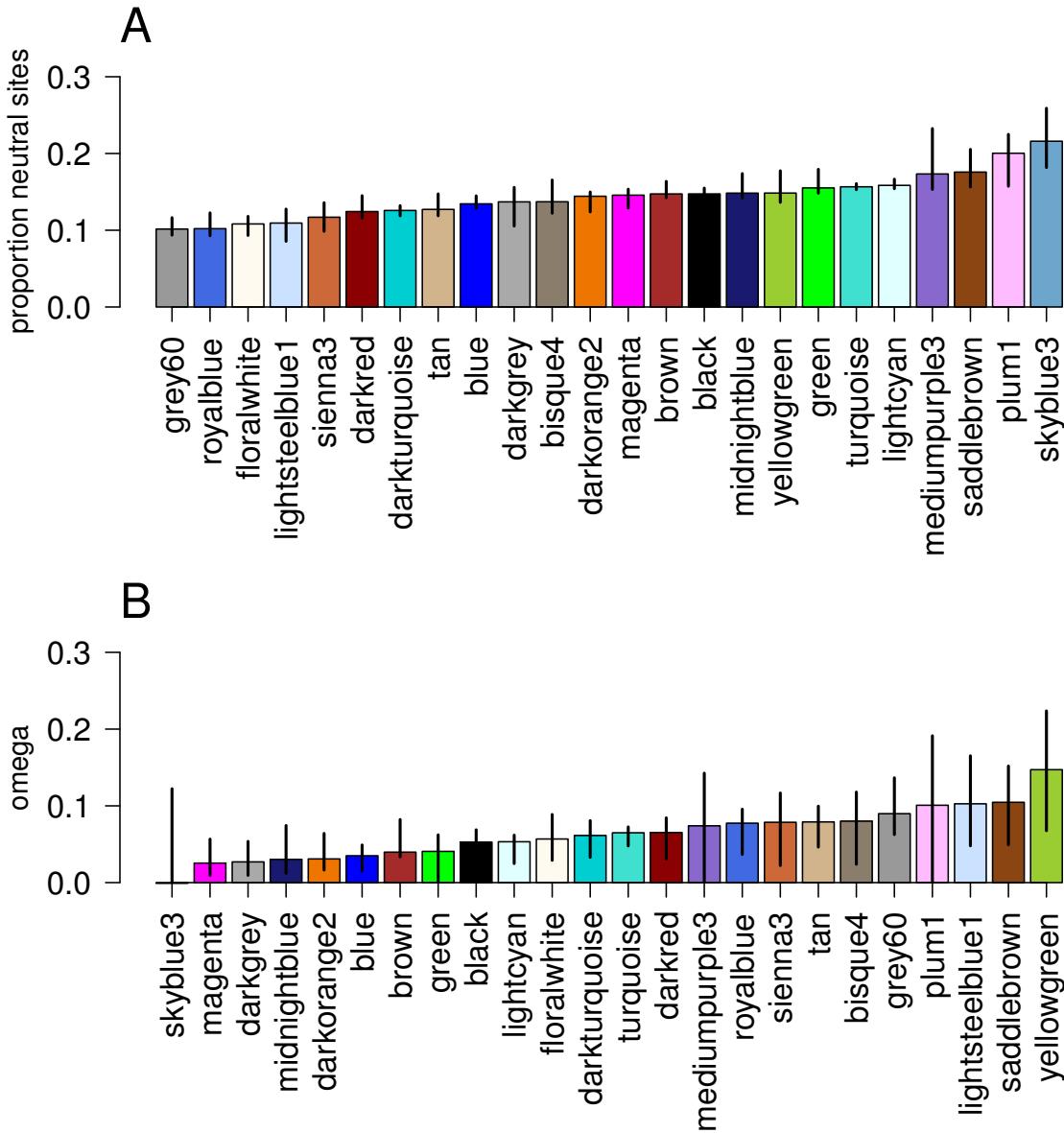


Figure 4.4: **Positive selection across modules.** A) shows the proportion of sites where  $\text{NeS} > 1$  across modules. Black bars represent 95% confidence intervals. B) shows  $\omega$  across modules. Black bars also represent 95% confidence intervals.

negative selection that would have opposite effects on divergence. One other study that investigated positive selections relationship to network centrality also ran into difficulties separating the effects of negative and positive selection on divergence. ? used human data to detect a positive relationship between network degree and signatures of linked selection on neutral polymorphisms but they found a negative relationship between network centrality and estimates of positive selection based on divergence. Our results suggest that previous studies that detected an effect of connectivity on protein divergence may be underestimating the effect of connectivity on negative selection and entirely missing the effect of connectivity on positive selection because of these forces conflicting effects on divergence.

We are not the first to note a relationship between connectivity and synonymous divergence (?). The cause of this correlation seems unclear. Jordan and colleagues (2004) suggested that highly connected genes may have more constraint on synonymous sites due to translational regulation. A full understanding of the causes of the relationship between connectivity and dS will likely depend on developing a better picture of the mechanism driving the relationship between connectivity and other selective parameters.

A relationship between positive and negative selection and coexpression network connectivity could result from a few different biological mechanisms. First, genes with higher connectivity could have more pleiotropic effects (?), which would constrain sequence evolution. However, while it has been suggested that highly pleiotropic genes should experience reduced positive selection (??), we do not observe this in our data. Alternatively, mutations in genes with higher connectivity may have larger phenotypic effect sizes, suggesting that both positive and negative selection may be stronger in highly connected genes (?). This latter hypothesis better fits our own findings that positive selection is stronger in well-connected genes.

Linking the relationship between connectivity and selection to a specific biological mechanism will require a closer understanding of how coexpression connectivity relates to pathway and network structure. A number of studies have investigated pathway position and selection in specific pathways. For example, genes in the human innate immune system are more constrained when they have many connections with other genes while genes at the top and the bottom of the network have the strongest evidence of positive selection (?). However, the insulin and target of rapamycin pathways show opposite patterns (reviewed in (?)). However, there is no clear way to link coexpression network connectivity with directionality and pathway structure, making it difficult to evaluate specific predictions about network structure with coexpression networks.

Observations that the evolutionary rate varies between genes were initially hypothesized to be due to variation in the proportion of gene sites under constraint: genes with a larger portion of constrained sites evolved slowly. (?). This view implies that variation in evolutionary rates relates directly to the function of the proteins. However, later investigations suggested that most or even all of the variation in evolutionary rates between proteins was not due to function at all, but instead a by-product of expression level variation that makes deleterious mutations more deleterious in highly expressed genes. (??). While it is undeniable that expression level does influence evolutionary rates, by demonstrating a relationship between connectivity and both negative and positive selection, we have shown that protein function does contribute to the variation in selection between genes, suggesting that gene function matters on a genome-wide scale for determining genetic variation.

## 4.5 Materials and Methods

### 4.5.1 Measuring expression

We used leaf expression data from a population sample of 99 individuals of *C. grandiflora* reported in ? and an additional 48 individuals, first reported here. All 147 individuals descend from the same

population collected near Monodendri in Greece, were grown in the same growth chamber, and underwent the same tissue collection and RNA extraction procedure, described in ?. The RNA from the previously unreported 48 individuals was sequenced at the Genome Quebec Innovation Centre in an Illumina HiSeq in one flow cell with 7 or 8 samples per lane (additional samples were sequenced and but were not included in this analysis). All reads from these 48 individuals were 100bp long and single ended. The first 99 individuals were sequenced paired-end in 100bp long reads, so to avoid the confounding effects of mapping accuracy, we randomly selected one read per pair for further analysis.

All sequence data was mapped using Stampy 1.0.21 (?) with default settings to an exon-only reference (?). We measured expression level for each gene by counting the number of reads that mapped to each gene using the HTSeq.scripts.count feature from HTSeq and we normalized for sequencing depth by dividing by the median expression level for each individual (?). We did not detect interactions between GC content, expression level, and lane (?). Genes with a median expression of 0 were removed from the analysis, leaving a total of 20,570 genes. For comparisons of expression level and divergence and selection we divided read count by gene length, measured as the number of nucleotides in the coding sequence, to make expression level comparable between genes.

#### 4.5.2 Construction of *Capsella grandiflora* gene co-expression network

The R package, WGCNA version 1.34, running under R version 2.15.1 (?) was employed to construct gene coexpression networks for the *C. grandiflora* expression data (?). This analysis groups together genes with similar patterns of pairwise correlation of expression, patterns that in theory are expected to reflect a shared role within similar biologically meaningful functional pathways and processes. We were interested in retaining the information embodied in the sign of the gene expression correlations, so we conducted a signed network analysis using the following adjacency function:  $a_{ij} = -0.5 + 0.5 \times \text{cor}(x_i, x_j)$ , where  $\text{cor}(x_i, x_j)$  is the correlation of gene expression of the  $i^{th}$  and  $j^{th}$  gene, and is the soft thresholding value. Soft thresholding is used in WGCNA, as it typically leads to co-expression networks that exhibit a scale-free degree distribution pattern of node connectivity, an inherent property of biological networks.

The *C. grandiflora* gene co-expression networks were obtained using a soft thresholding power of 12, as suggested by the authors of the WGCNA package for signed networks. Genes that exhibited similar patterns of connectivity (i.e., genes showing high topological overlap) were grouped together in the same coexpression modules, based on hierarchical clustering of topological overlap values, in which a dynamic branch-cutting algorithm was used to define initial gene co-expression modules (?). Module eigengenes (the first principal component of the gene expression values of modules) were calculated and modules whose eigengenes were highly correlated were merged to arrive at the final set of co-expression modules. The resulting modules were labeled with different colors for ease of referencing. A preliminary coexpression analysis showed that one individual was a strong outlier, so we removed it and reran the analysis on the remaining 146 individuals. Total connectivity for each gene was calculated as the sum of connection strengths between the focal gene and all other genes and intramodular connectivity was calculated as the sum of connections between the focal gene and all other genes in the same module.

#### 4.5.3 Co-expression module preservation analysis

Module preservation analysis aims to determine whether and to what extent co-expression modules detected in one data set (the reference data set) are conserved in a second (or test) data set. The reference gene co-expression modules and data set in this case were from the *C. grandiflora* gene expression data and analysis described above, while the test data were from RNAseq reads for 19,706 genes obtained from RNA collected from seedling tissue of 20 ecotypes of *A. thaliana*, obtained from published data (?). Module preservation analysis does not require network construction for the test data set. Instead, the module assignment based on analysis of the reference (*C. grandiflora*) set is applied to the test (*A. thaliana*) set, and network-based preservation statistics are calculated and combined into a single composite preservation measure that reflects preservation of both module density and module connectivity patterns (?). By randomly permuting the module labels assigned to the test data and re-calculating the network-based preservation statistics n times (n = 200 times for the analyses reported here), one can calculate a Z-statistic for the composite preservation measure associated with each reference module, referred to here as a Z-summary statistic. The Z-summary statistic asymptotically follows the normal distribution with mean 0 and SD = 1, and can be converted to a p-value under the standard normal distribution. Simulations conducted by Langfelder et al. (2011) show that Z-summary values <10 correspond with strong evidence of module preservation in the test data, Z-summary values between 2 and 10 correspond with moderate evidence of module preservation in the test data.

#### 4.5.4 GO enrichment analysis of genes in modules

We used the GOrilla software package to search for GO terms enriched in the *A. thaliana* orthologs (from ?) of each co-expression module (?). 16,672 of our expressed genes had *A. thaliana*, though GOrilla had GO terms for only 10,079 of these genes, and so the enrichment analysis is based on this subset.

#### 4.5.5 Measuring divergence

Divergence measurements from each gene were previously published in (?) and detailed methods are included there. Briefly, we found orthologs between the *Capsella rubella* reference genome and its two closest outgroups: *Neslia paniculata* and *A. thaliana* and aligned the protein sequences of these orthologs. These alignments were used to calculate dN and dS using codeml in PAML in a model where divergence was allowed to vary in the lineage leading to *C. rubella* (?). Nine genes with zero synonymous substitutions were removed from the analysis so that  $d_N/d_S$  could be calculated.

#### 4.5.6 Relating divergence to connectivity and other parameters

We conducted Spearman correlations between connectivity and divergence and connectivity and expression for each gene (N=13,211) using R's cor.test() function (?). Partial Spearman correlations between connectivity and divergence while accounting for expression were conducted with the pcor.test() function from the ppcor library in R. A principal component analysis for dS, mean expression level, the coefficient

of variation for expression level, and intramodular connectivity was conducted in R using the prcomp() function, with scaling. We also used the R lm() function to construct a linear model for how principal components predict variation in  $d_N$ .

#### 4.5.7 Measuring selection

The strength of positive and negative selection in the *C. grandiflora* genome was measured using single nucleotide polymorphism (SNP) data from 178 individuals collected from the same population described above and previously reported in (??). 146 of these 178 individuals overlap with the individuals used to generate coexpression networks.

DNA was extracted by either a CTAB based protocol or by DNeasy Plant Mini Kit (Qiagen). We obtained whole genome sequences from each individual through 100 cycles of paired end sequencing in a Hiseq 2000 with Truseq libraries (Illumina) and three individuals were sequenced per lane. We subsequently followed GATK Best Practices for Variant Quality Recalibration circa GATK 2.7 (?) using a high confidence truth set generated by filtering SNPs for concordance with common variants (<0.11) in a species-wide sample of *C. grandiflora* as well as suspect realignments (transposable elements, centromeres, 600 bp intervals containing extreme Hardy Weinberg deviations, 1 kb intervals that showed evidence of three or more SNPs in a reference-to-reference mapping of 150 bp paired end reads from the reference genome line). Inter-relatedness among lines was <5%. For population genetic analysis, we downsampled to 320 alleles by randomly classifying alleles as missing, which allowed retention of 94.2% of sites.

Divergence was measured from an outgroup, *N. paniculata*, aligned to *C. rubella* using LastZ with chaining, as described in (?). We used the Joint Genome Institutes gene annotation of the *C. rubella* reference genome to identify 0-fold degenerate and 4-fold degenerate sites within genes. We assigned these genes into categories based on intramodular connectivity values and coexpression module membership and calculated site frequency spectra (SFS) and divergence at 0-fold degenerate and 4-fold degenerate sites within these categories. We used SFS and divergence to estimate the fraction of 0-fold degenerate sites within each category under negative selection and  $\alpha$  using 4-fold sites as a neutral reference and the methods described in ???. 200 bootstraps were conducted by resampling the genes included within each category with replacement. We used these bootstraps to construct 95% confidence intervals and we generated two-tailed p-values by calculating the proportion of bootstraps where selection was higher in the less selected category than the more selected category and multiplying by two (?).

### 4.6 Acknowledgements

Robert Williamson, Adrian Platts, Wei Wang, Alan Moses, Asher Cutter.

### 4.7 Appendix: Supplementary figures and tables

Module	Z	N	$N_e S < 1$	$\alpha$	$\omega$	GO Term	Description	p
bisque4	3.827	191	0.137	0.402	0.080	GO:0010207	photosystem II assembly	1.9E-09
						GO:0019252	starch biosynthetic process	6.1E-08
						GO:0005982	starch metabolic process	1.2E-07
						GO:0000023	maltose metabolic process	1.6E-07
						GO:0005984	disaccharide metabolic process	6.4E-07
black	8.104	1106	0.147	0.290	0.053	GO:1902589	single-organism organelle organization	4.3E-05
						GO:0022402	cell cycle process	4.5E-05
						GO:0009630	gravitropism	8.9E-05
						GO:0009606	tropism	8.9E-05
						GO:0009629	response to gravity	1.4E-04
blue	12.664	2066	0.134	0.230	0.035	GO:0043043	peptide biosynthetic process	1.6E-90
						GO:0006412	translation	3.8E-90
						GO:0043604	amide biosynthetic process	6.0E-87
						GO:0006518	peptide metabolic process	4.2E-86
						GO:0043603	cellular amide metabolic process	7.8E-77
brown	-1.149	2757	0.147	0.236	0.040	GO:0006914	autophagy	1.2E-09
						GO:0044248	cellular catabolic process	6.9E-07
						GO:0046700	heterocycle catabolic process	1.5E-04
						GO:0044242	cellular lipid catabolic process	1.7E-04
						GO:0019439	aromatic compound catabolic process	2.2E-04
darkgrey	1.195	357	0.137	0.182	0.027	NA	NA	NA
darkorange2	4.598	982	0.144	0.196	0.031	GO:0010264	myo-inositol hexakisphosphate biosynthetic process	2.5E-05
						GO:0033517	myo-inositol hexakisphosphate metabolic process	2.5E-05
						GO:0032958	inositol phosphate biosynthetic process	4.0E-05
						GO:0046173	polyol biosynthetic process	6.3E-05
						GO:0043647	inositol phosphate metabolic process	3.4E-04

Module	Z	N	$N_e S < 1$	$\alpha$	$\omega$	GO Term	Description	p
darkred	3.277	456	0.124	0.375	0.065	GO:0048583	regulation of response to stimulus	2.0E-05
						GO:0031347	regulation of defense response	3.2E-05
						GO:0045087	innate immune response	4.5E-05
						GO:1900674	olefin biosynthetic process	5.6E-05
						GO:1900673	olefin metabolic process	5.6E-05
darkturquoise	6.152	1003	0.126	0.359	0.061	GO:1901700	response to oxygen-containing compound	5.4E-11
						GO:0042221	response to chemical	5.5E-09
						GO:0050896	response to stimulus	1.5E-08
						GO:0001101	response to acid chemical	7.1E-07
						GO:0048511	rhythmic process	1.6E-06
floralwhite	5.314	249	0.108	0.376	0.057	GO:0019682	glyceraldehyde-3-phosphate metabolic process	1.4E-22
						GO:0019637	organophosphate metabolic process	9.5E-22
						GO:0019288	isopentenyl diphosphate biosynthetic process	3.6E-21
						GO:0009240	methylerythritol 4-phosphate pathway	
						GO:0046490	isopentenyl diphosphate biosynthetic process	4.7E-21
						GO:0046490	isopentenyl diphosphate metabolic process	4.7E-21
green	3.104	1269	0.155	0.231	0.041	GO:0009853	photorespiration	1.2E-11
						GO:0043094	cellular metabolic compound salvage	7.8E-10
						GO:0080129	proteasome core complex assembly	1.2E-06
						GO:0043248	proteasome assembly	1.3E-06
						GO:0051788	response to misfolded protein	1.3E-06

Module	Z	N	$N_e S < 1$	$\alpha$	$\omega$	GO Term	Description	p
grey60	3.874	278	0.102	0.507	0.090	GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	4.1E-12
						GO:0007167	enzyme linked receptor protein signaling pathway	4.1E-12
						GO:0007166	cell surface receptor signaling pathway	2.1E-11
						GO:0006468	protein phosphorylation	2.6E-07
						GO:0071554	cell wall organization or biogenesis	3.3E-05
lightcyan	18.410	2886	0.158	0.277	0.053	GO:1901293	nucleoside phosphate biosynthetic process	7.6E-53
						GO:0009165	nucleotide biosynthetic process	4.0E-52
						GO:0006221	pyrimidine nucleotide biosynthetic process	9.8E-41
						GO:0006220	pyrimidine nucleotide metabolic process	9.8E-41
						GO:0046390	ribose phosphate biosynthetic process	2.6E-40
lightsteelblue1	-0.078	73	0.109	0.517	0.103	GO:0006897	endocytosis	6.2E-08
						GO:0048268	clathrin coat assembly	6.8E-08
						GO:0006901	vesicle coating	9.5E-08
						GO:0016050	vesicle organization	1.3E-07
						GO:0071822	protein complex subunit organization	3.6E-04
magenta	0.085	807	0.146	0.166	0.025	GO:0006661	phosphatidylinositol biosynthetic process	5.9E-06
						GO:0046488	phosphatidylinositol metabolic process	3.7E-05
						GO:0046474	glycerophospholipid biosynthetic process	1.0E-04
						GO:0006650	glycerophospholipid metabolic process	4.3E-04
						GO:0045017	glycerolipid biosynthetic process	5.2E-04

Module	Z	N	$N_e S < 1$	$\alpha$	$\omega$	GO Term	Description	p
mediumpurple3	13.539	74	0.173	0.326	0.074	GO:0007017	microtubule-based process	1.4E-30
						GO:0000911	cytokinesis by cell plate formation	7.0E-30
						GO:0032506	cytokinetic process	1.3E-29
						GO:1902410	mitotic cytokinetic process	1.3E-29
						GO:1903047	mitotic cell cycle process	8.6E-27
						GO:0010200	response to chitin	4.0E-50
midnightblue	24.002	771	0.148	0.189	0.030	GO:0010243	response to organonitrogen compound	4.0E-50
						GO:1901698	response to nitrogen compound	3.2E-47
						GO:0009719	response to endogenous stimulus	5.1E-40
						GO:0010033	response to organic substance	1.7E-33
						GO:0009620	response to fungus	3.3E-05
						GO:0009723	response to ethylene	7.6E-04
plum1	10.286	98	0.200	0.361	0.101	GO:0051707	response to other organism	7.6E-04
						GO:0051704	multi-organism process	8.9E-04
						GO:0055085	transmembrane transport	3.2E-04
						GO:0005984	disaccharide metabolic process	4.5E-04
royalblue	14.434	263	0.102	0.468	0.077	GO:0009311	oligosaccharide metabolic process	8.8E-04
						GO:0006260	DNA replication	1.4E-11
						GO:0006270	DNA replication initiation	1.9E-10
						GO:0040029	regulation of gene expression	3.4E-10
						GO:0061647	epigenetic histone H3-K9 modification	8.3E-10
						GO:0051567	histone H3-K9 methylation	8.3E-10
saddlebrown	10.662	147	0.176	0.403	0.105	GO:0044710	single-organism metabolic process	2.6E-07
						GO:0016109	tetraterpenoid biosynthetic process	1.2E-06
						GO:0016108	tetraterpenoid metabolic process	1.2E-06
						GO:0016116	carotenoid metabolic process	1.2E-06
						GO:0016117	carotenoid biosynthetic process	1.2E-06

Module	Z	N	$N_e S < 1$	$\alpha$	$\omega$	GO Term	Description	p
skyblue3	21.610	119	0.216	-0.005	-0.001	GO:0009697	salicylic acid biosynthetic process	9.3E-12
						GO:0009696	salicylic acid metabolic process	9.3E-12
						GO:0009627	systemic acquired resistance	1.0E-11
						GO:0046189	phenol-containing compound biosynthetic process	1.3E-11
						GO:0018958	phenol-containing compound metabolic process	1.9E-11
tan	2.090	591	0.127	0.419	0.079	GO:0045491	xylan metabolic process	9.0E-12
						GO:0010410	hemicellulose metabolic process	1.5E-11
						GO:0010413	glucuronoxyan metabolic process	5.7E-11
						GO:0045492	xylan biosynthetic process	7.3E-11
						GO:0070592	cell wall polysaccharide biosynthetic process	1.2E-10
turquoise	6.411	4005	0.157	0.321	0.065	GO:0006486	protein glycosylation	3.1E-14
						GO:0043413	macromolecule glycosylation	3.1E-14
						GO:0070085	glycosylation	4.2E-14
						GO:0009630	gravitropism	1.8E-09
						GO:0009606	tropism	1.8E-09
yellowgreen	10.232	120	0.148	0.531	0.147	GO:0010200	response to chitin	1.2E-09
						GO:0010243	response to organonitrogen compound	1.2E-09
						GO:1901698	response to nitrogen compound	1.4E-09
						GO:0006865	amino acid transport	3.4E-08
						GO:0002679	respiratory burst involved in defense response	4.0E-08

Table 4.2: Population genetics and top 5 enriched GO process terms for each module.

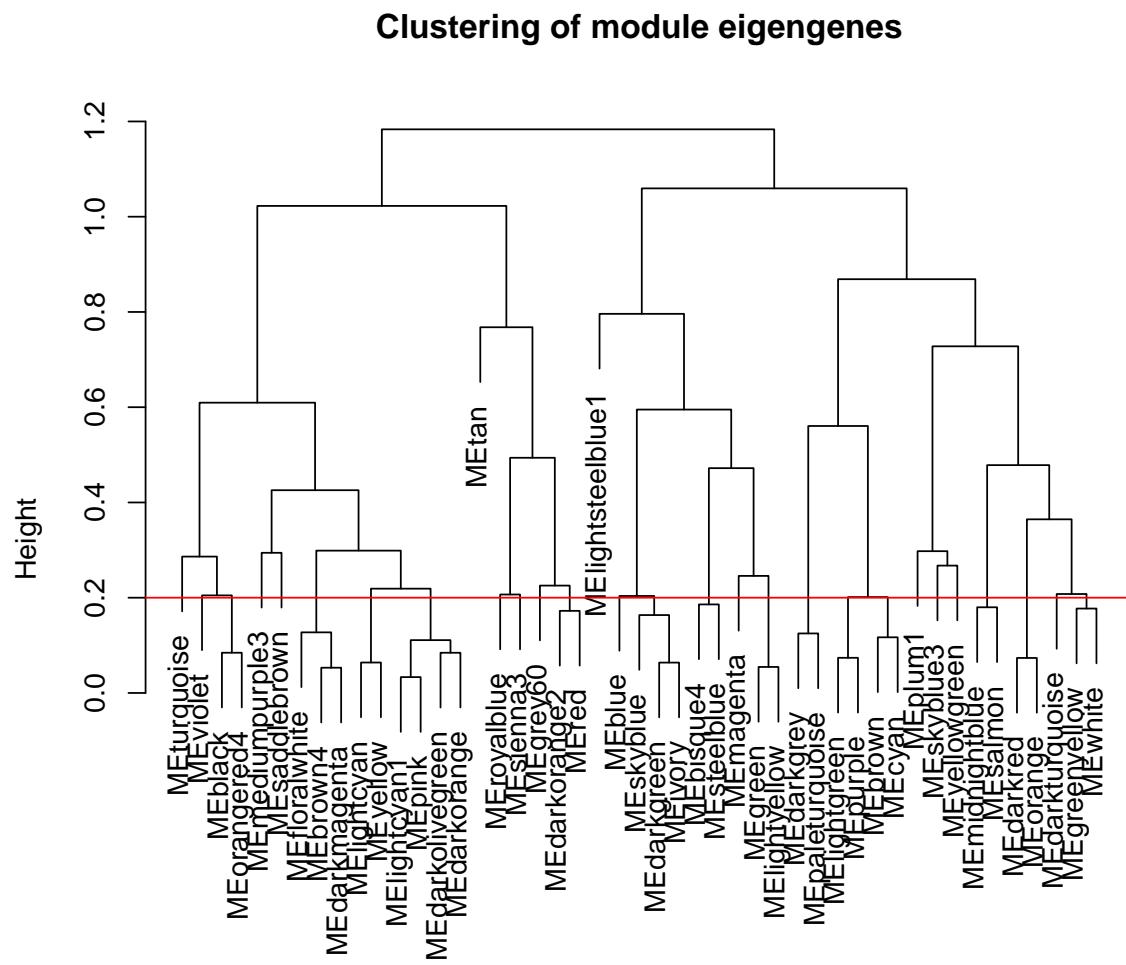


Figure 4.5: Clustering of module eigengenes

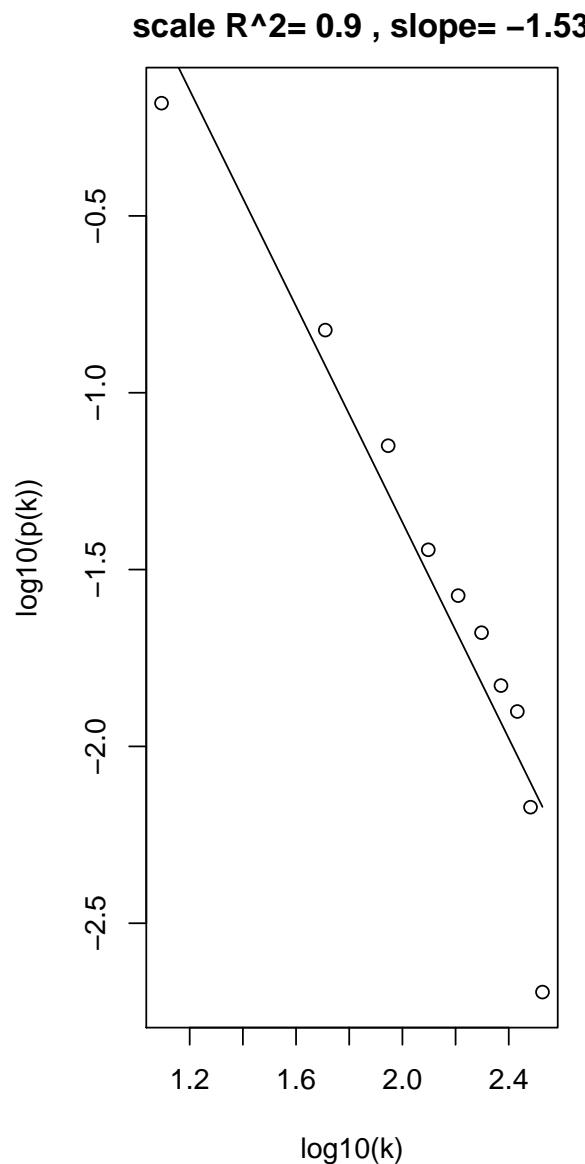


Figure 4.6: **Coexpression modules are scale-free.** Genes were binned by connectivity level,  $k$ , and are plotted against the proportion of genes in each connectivity bin,  $p(k)$  on a log scale.  $R^2 = 0.9$ , slope =  $-1.53$ .

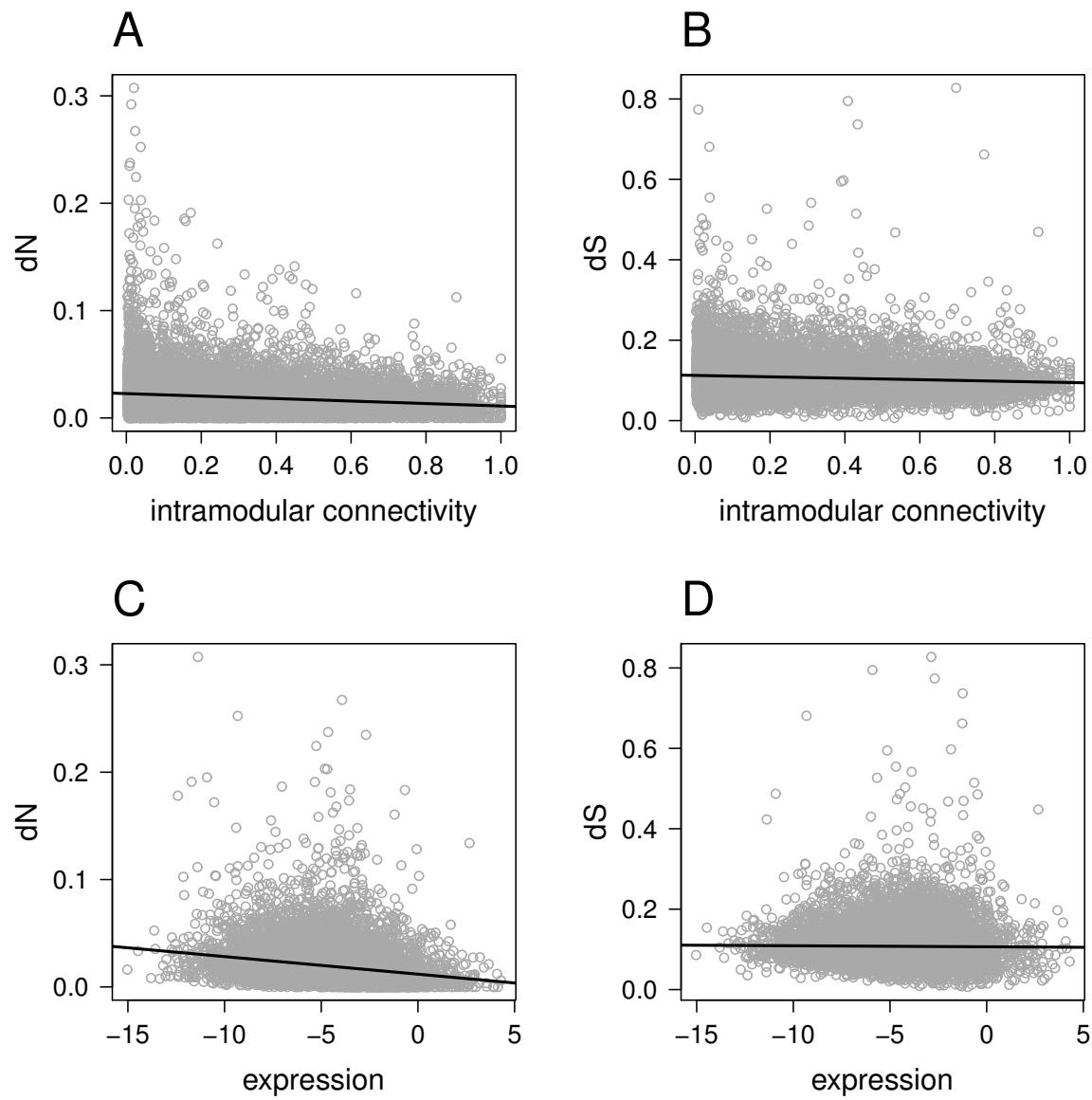


Figure 4.7: Correlations between expression and intramodular connectivity and  $d_N$  and  $d_S$ .

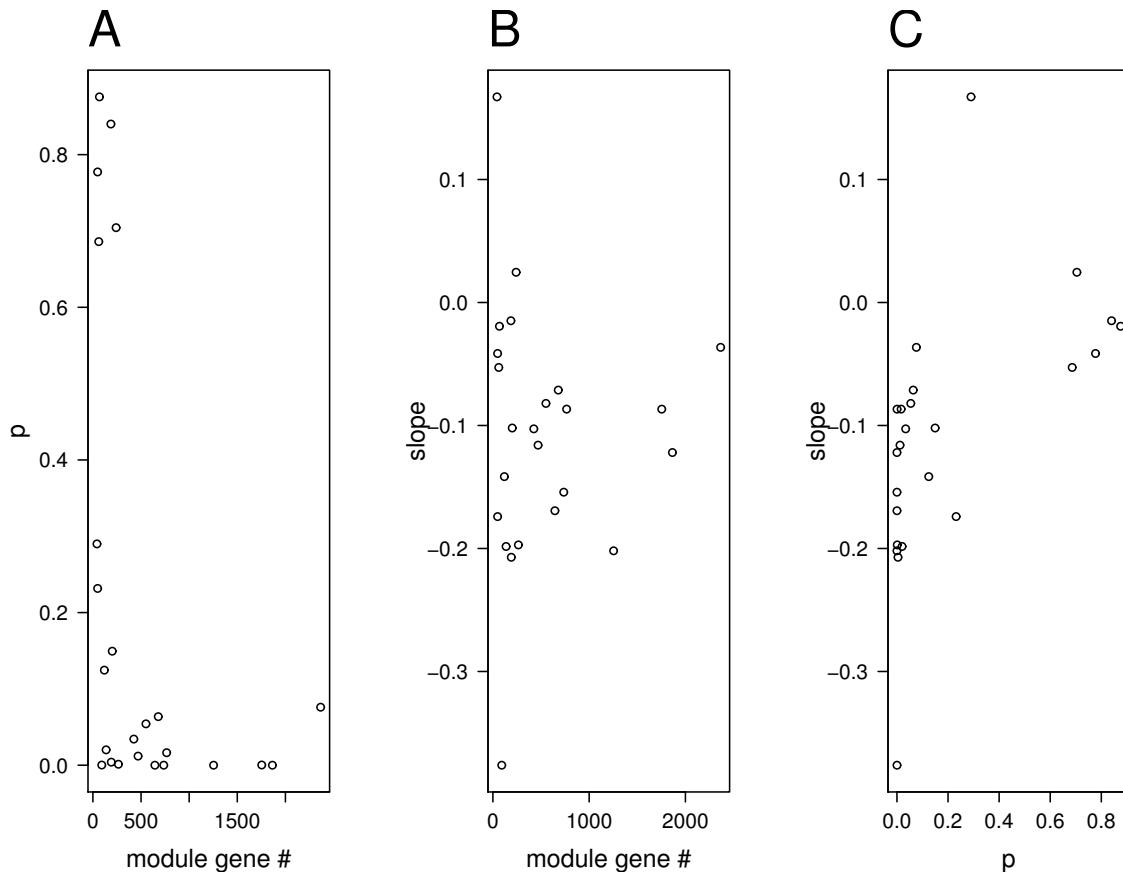
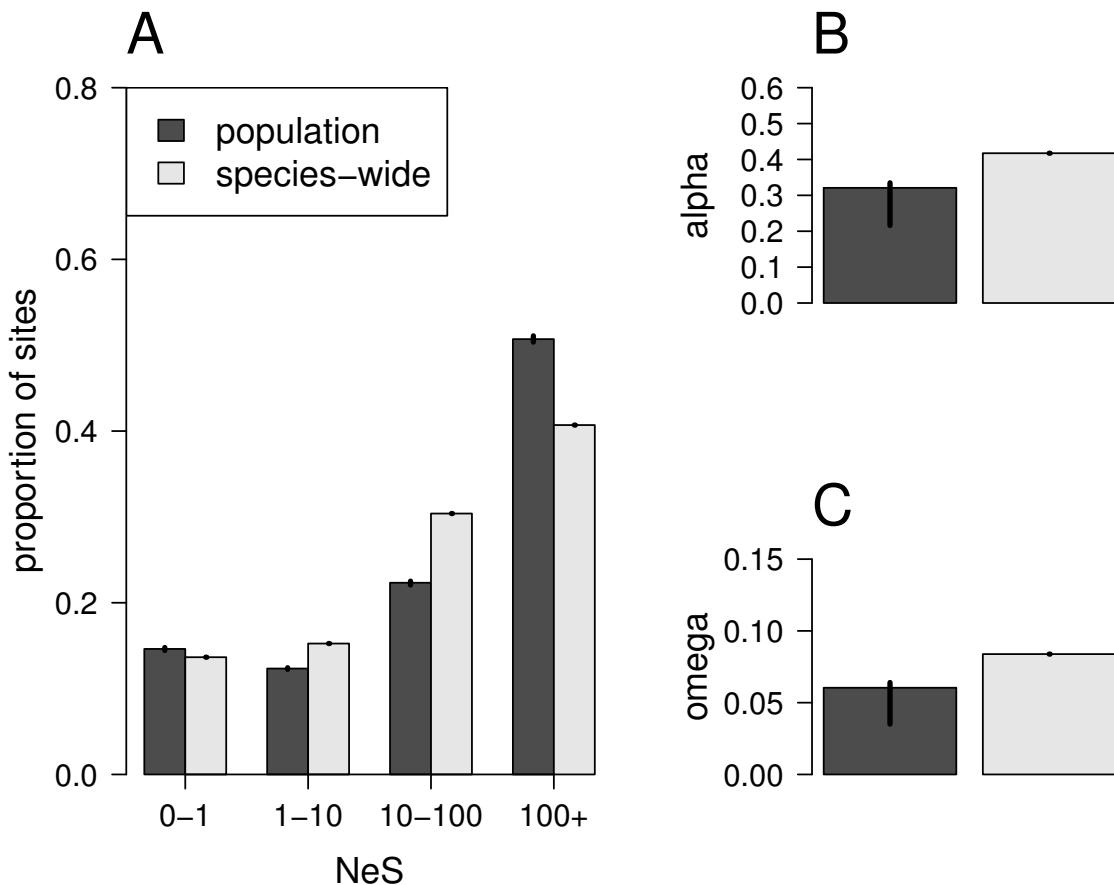


Figure 4.8: **Within-module estimates of the correlation between intramodular connectivity and  $d_N/d_S$  depend on module size** A) The number of genes in a module (x-axis) plotted against the p value for significance of the correlation between connectivity and  $d_N/d_S$  within that module (y-axis). B) The number of genes in a module (x-axis) plotted against the slope of the relationship between connectivity and  $d_N/d_S$  (y-axis). C) The relationship between (x-axis) and slope (y-axis) for within-module correlations of connectivity and  $d_N/d_S$



**Figure 4.9: Positive and negative selection at 0-fold degenerate sites in all genes included in the coexpression analysis compared to selection on all 0-fold sites in a previously-reported species-wide sample.** A) The proportion of sites found in each bin of purifying selection strength. B) The proportion of sites fixed by positive selection, and C) the rate of adaptive substitution. Error bars represent 95% confidence intervals and are only reported for the data from this study. Measurements of selection strength on the species-wide sample come from (?).