Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*

Robert J. Williamson*[1], Emily B. Josephs*[1,&], Adrian E. Platts[2,3], Khaled M. Hazzouri[1,4], Annabelle Haudry[1,5], Mathieu Blanchette[2,3], Stephen I. Wright[1,6]

[1] Department of Ecology and Evolutionary Biology, University of Toronto
25 Willcocks St, Toronto ON M5S 3B2

[2] Centre for Bioinformatics, McGill University, Montreal, Quebec, Canada

[3] School for Computer Science, McGill University, Montreal, Quebec, Canada

[4] Center for Genomics and Systems Biology, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

[5] Université Lyon 1, Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche (UMR) 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France.

[6] Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario, Canada

* These authors contributed equally to this work
& Author for correspondence: em.josephs@utoronto.ca

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 in 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.

Author Summary
Selection will have consequences for genomic variation, but it is unclear how much the genomic effects of selection vary across plant genomes, particularly in noncoding regions. To determine the strength and extent of selective signatures across the genome we sequenced and analyzed genomes from 13 *Capsella grandiflora* individuals. *C. grandiflora* is well-suited for this task because it has experienced a large, stable effective population size, so we expect that selection signatures will not be overly distorted by demographic effects. Our analysis shows that positive and negative selection for new mutations have broadly shaped patterns of genomic diversity in coding regions but not in most noncoding regions. However, when we focus only on noncoding regions that show evidence of constraint across species, we see evidence for strong positive and negative selection. In addition, we find that genes with high expression experience stronger negative selection than genes with low expression, but the extent of positive selection appears to be equivalent across expression categories.

Introduction

Determining the proportion of the genome subject to positive and negative selection has wide-ranging implications for understanding genome function, the maintenance of genetic variation, and the evolution of sex and mating systems [1]. Despite this question's importance, the rate, strength, and direction of natural selection across the genome remain largely unknown in most taxa. While early data on molecular evolution appeared to support the neutral theory [2] genome-scale data have suggested that, at least in some species with large effective population sizes (N_e), positive and negative selection are major determinants of genome-wide polymorphism and divergence [3-5]. Current evidence suggests contrasting patterns across taxa: *Drosophila melanogaster* estimates indicate high levels of constraint and high rates of positive selection genome-wide [6], while in humans data show much less evidence of positive selection, and noncoding constraint is primarily limited to a relatively small fraction of the genome [7-9].

While evidence to date suggests that the global strength and extent of selection varies considerably between species, many questions about the extent and nature of heterogeneity in selection across the genome remain unanswered. For example, gene expression level is one of the major determinants of rates of protein evolution in coding regions [10-12]; this could reflect relaxed purifying selection in lowly expressed genes, increased positive selection, or both. Similarly, noncoding regions as a whole are generally less conserved than coding regions [6,13] despite hypotheses that adaptive evolution will be particularly predominant in regulatory regions [14,15 but see 16]. Finally, the extent to which negative and positive selection predominate in species-specific sequences versus broadly conserved regions remains unclear.

In plant genomes, quantification of the extent of selection has been limited by a lack of genome-wide comparative and population genomics data. From the evidence to date, the amount of coding sequence under constraint tends to correlate positively with N_e, but most species show little evidence of positive selection and relatively little is known about the strength of selection on noncoding regions [17,18]. It has been suggested that plant noncoding regions are subject to less selective constraint than animal genomes, potentially because regulatory complexity is generated more via gene duplication than through complex *cis*-regulatory sequences [19]. If true, this could indicate very different functional genomic properties of plant and animal genomes.

In a recent study, Haudry and colleagues [20] identified between-species conserved regions in *Arabidopsis lyrata* and its close relatives by comparing the genomes of nine Brassicaceae species. They identified over 90,000 conserved noncoding regions genome-wide, and showed that approximately one quarter of the conserved sites in the *A. thaliana* genome were in noncoding regions. Here, we use population genomics to quantify the strength of selection inside and outside of an analogously identified set of conserved regions, and to investigate the extent of heterogeneity in positive and negative selection in a species-wide sample of 13 outbred *Capsella grandiflora* individuals. *C. grandiflora* is an obligately outcrossing member of the Brassicaceae family and conforms relatively closely to the assumptions of the standard neutral model because of its large and stable effective population size (N_e ~ 600,000) and relatively low population structure [17,21]. We estimate the strength of negative selection by fitting polymorphism data to a model of the distribution of negative fitness effects of mutations and compare between coding and noncoding regions and high and low expression genes. We quantify the contribution of positive selection to divergence in *C. grandiflora* using two complimentary approaches: an extension of the McDonald-Kreitman test [22] and an analysis of neutral variation linked to lineage-specific fixed substitutions [4]. 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.

Results

Genome wide patterns of polymorphism

We sequenced 13 outbred *C. grandiflora* individuals (26 chromosomes; ~140 Mb genome) sampled from across the species' range in northern Greece and called variants using the Genome Analysis Toolkit's Unified Genotyper [23]. After filtering for quality and depth (see Methods) we were left with 24,228,247 sites, 1,375,512 of which were single nucleotide polymorphisms (SNPs) (Table S1). 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 [24]. To avoid comparing sites that would not have equivalent mutation profiles, we excluded sites in coding regions that were neither 4-fold nor 0-fold degenerate. Consistent with previous estimates [25], average nucleotide diversity at 4-fold degenerate sites was 0.023 and there was evidence for an excess of rare variants genome-wide at 4-fold sites compared with the standard neutral model (Tajima's D = -0.506). Introns (Watterson's θ_w = 0.020) and intergenic regions (θ_w = 0.021) 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 synonymous sites. 5' and 3' UTRs showed a much stronger diversity reduction (θ_w = 0.015 and 0.014 respectively), while 0-fold degenerate nonsynonymous sites showed the strongest reduction (θ_w = [illegible]

Neutral diversity at 4-fold degenerate sites near centromeric regions was elevated on most chromosomes, similar to observations made in *Arabidopsis thaliana* [26], (Figure S1). As with *Arabidopsis*, this is not obviously caused by higher mutation rates in these regions, since divergence between *Capsella* and *Neslia* is not clearly elevated in these regions (Figure S1). Although elevated error rates in repetitive regions may be a contributor, our observation of high diversity in these regions is apparent after our extensive filtering (see Methods). Furthermore, diversity generally declines towards the ends of the chromosomes, potentially reflecting the 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. Indeed, we see a decrease in diversity in regions of low coding density (Figure S2).

We also examined individual heterozygosity in sliding windows along each chromosome. A number of individuals showed stretches of homozygosity indicative of biparental inbreeding (Figure S3). Consistent with these regions reflecting local biparental inbreeding, no such regions are found in our sample that is derived from a between-population cross, AXE (Figure S3). 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 genomic stretches across the genome. To avoid biased estimation of 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.

Genome-wide measures of purifying selection

Using the methods of Eyre-Walker & Keightley [1] we compared the allele frequency spectrum (AFS) and divergence for putatively neutral sites (4-fold degenerate) to these measures in every

other category of sites in order to quantify the amount negative and positive selection on each category (Figure 1A). Consistent with the patterns of diversity described above, negative selection is generally much stronger in coding regions than noncoding regions (Figure 1A). This pattern is most clearly seen in 0-fold degenerate sites, the only site category with a sizable fraction of sites in the strongest category of negative selection (41%). Of the noncoding categories UTRs show much stronger negative selection than other regions. In *C. grandiflora* 54% of both 5' and 3' UTRs are under moderate levels of purifying selection ($N_e s > 1$), but a considerably larger fraction of UTR sites are effectively neutral (46%) than 0-fold degenerate sites (14%). Additionally 5' UTRs are the only type of noncoding sites that show any signal of strong purifying selection ($N_e s$>10), other than conserved noncoding sequences (CNSs; see below).

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 as found in [20] coupled with typically weak to no selection outside of splice junctions. In order to test for this 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 partially be explained by variance around sampling sites in these different regions. These generally low estimates of $N_e s$ in (non-junction) intronic and intergenic sites imply either a general lack of purifying selection in most noncoding regions, a lack of sensitivity to detect small proportions of selected sites, and/or that these sites are under nearly equivalent purifying selection to synonymous sites, which are used as our neutral standard.

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. To assess this possibility, we identified conserved noncoding sequences (CNSs) across the 9 Brassicaceae genomes, following the implementation in Haudry et al. [20]. For this study, we used the *Capsella* genome as a reference for alignment, but excluded *Capsella* when identifying CNSs in order to avoid circularity when quantifying selection from diversity [27]. This allows our analysis of selection on noncoding sites using polymorphism to be more independent of the comparative analysis. When we look at only these conserved regions we see a small proportion of effectively neutral sites (29%), suggesting that the majority of CNS sequences are subject to purifying selection. However, estimates suggest that CNSs are generally under weaker purifying selection than nonsynonymous (0-fold) sites and experience primarily weak and intermediate purifying selection (Figure 1).

We also quantified selection at the different types of CNSs separately (Figure S4). In most categories about 25% of sites are nearly neutral, a slightly stronger signal of purifying selection than when we pool all CNSs. Intronic and CNSs have a larger proportion of effectively neutral sites than other categories, in agreement with the general neutrality of intronic sites (Figure 1). In contrast intergenic CNSs, small noncoding RNAs (sncCNSs), downstream CNSs, 5’ UTR CNSs, and upstream CNSs all show evidence of stronger selection than the pooled CNS sites. Small noncoding RNAs (snc) have the smallest fraction of effectively neutral sites (10%). The number of sites used to make the AFS for each of these categories varies substantially (Supp

Table S1), and our sample of sncCNSs have very little polymorphism (257 segregating sites). Nevertheless, despite the wide confidence intervals, sncCNSs still show significantly ($p < 0.001$) more moderate ($1<N_e s <10$) purifying selection than the pooled CNSs, which could be due to strong selection for secondary structure important for RNA activity [28]. This effect is consistent with sncCNSs showing a higher degree of conservation across the Brassicaceae [20] and having traceable orthologs in other plants.

Genome-wide estimates of positive selection

We used the approach of Eyre-Walker and Keightley [22] to estimate the proportion of fixations driven by positive selection (α) and the rate of positive selection (ω)while taking into account the effect of slightly deleterious mutations, which can bias estimates of positive selection downwards. To do this, we estimated divergence using whole genome alignments of *Capsella* to the close outgroup *Neslia paniculata*. 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 in noncoding DNA on CNS sites and UTRs. We found that 0-fold degenerate sites show a very high signal of divergence driven by positive selection (Figure 1C; $\alpha = 40.1\%$, $\omega = 0.08$). Similarly UTRs and CNS sites show evidence for positive selection (Figure 1C). These results generally suggest widespread positive selection on both nonsynonymous and functional noncoding genomic regions.

If many of the amino acid changes between *C. grandiflora* and its nearest relatives are due to positive selection from new mutations, we expect to see the signature of selective sweeps: reduced neutral diversity surrounding amino acid fixations [29,30]. We tested for this signature by measuring neutral diversity (the proportion of 4-fold degeneratesites in each window that were polymorphic) in non-overlapping 1kb windows surrounding fixed replacement (n = 60,378), and silent (n = 83,812) substitutions in *C. grandiflora*. Neutral diversity surrounding fixed replacement substitutions was lower than neutral diversity surrounding fixed silent substitutions in the 4kb window surrounding substitutions (Figure 2A). This result was robust to various window sizes (Figure S6) and a one-tailed test for reduced 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 [29], which would increase neutral diversity and divergence in *C. grandiflora*. Consistent with this prediction, divergence at 4-fold sites is elevated around synonymous and replacement substitutions (Figure 2B). To control for elevated mutation rate, we divided diversity by divergence at 4-fold degenerate sites. We observed a reduction in 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 (Figure 2C, $p < 0.01$ for 1 kb on either side of the substitution).

An analogous test for selective sweeps around fixations in noncoding regions is challenging because the test depends on accurately identifying interspersed functional and neutral sites, a difficult task in noncoding regions [27]. Instead, we compared 4-fold diversity around fixed substitutions in CNS regions (n = 12,578) with 4-fold diversity 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 noncoding regions (Figure S5). 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 4-fold degenerate divergence near CNSs. However, there is still a reduction in neutral diversity/divergence around fixed substitutions in CNSs compared to those in non-conserved intergenic regions, consistent with the action of recurrent selective sweeps (Figure 3A).

The observed reduction in diversity/divergence around CNS substitutions could also reflect the action of background purifying selection; sites closer to CNSs may experience a reduction of neutral diversity due to greater purifying selection on mutations in CNSs. To assess this further, we compared diversity and divergence surrounding CNSs containing at least one fixed substitution to diversity and divergence around those that do not. There is a reduction in 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 (Figure 3B).

Effects of expression and selection

We sequenced RNA extracted from leaf tissue of 10 of the 13 individuals (see also [24]), and measured expression levels of all expressed genes. 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, *dN*/*dS* is considerably lower in high expression genes (0.13) than low expression genes (0.21).

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 expressed genes, consistent with the hypothesis that negative selection acts more strongly in highly expressed genes (Figure S7). The distribution of fitness effects, estimated using the same methods described earlier, show that high expression genes have a much smaller proportion of effectively neutral sites (6.8%) than low expression genes (16%, $p < 0.001$) and a greater proportion of sites (86%) under strong purifying selection ($N_e s > 10$) compared with low expression genes (72%, $p < 0.001$) (Figure 4A).

Increased divergence in low expression genes relative to high expression genes could also be caused by increased positive selection in low expressed genes compared to highly expressed genes. To test this possibility, we calculated α and ω as described above. Highly expressed genes have a significantly higher value of α (0.66) than lowly expressed genes (0.42, $p < 0.01$) but the ω value for both classes is similar (high: 0.11, low: 0.10, $p = 0.38$), suggesting that the rate of positive selection does not differ between high and low expression genes (Figure 4B, 4C). The difference in α between the two categories likely reflects the reduction in the number of weakly deleterious and effectively neutral mutations that are able to fix due to stronger purifying selection in high expression genes compared to low expression genes, causing a higher proportion of those amino acids that do reach fixation to be positively selected.

Discussion

In this population genomic survey of *C. grandiflora*, we demonstrated that positive and negative selection are prevalent 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 on high expression genes, not increased positive selection on 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 long time since whole genome duplication [18].

In contrast, selection appears to be very rare in intergenic and intronic regions that are not conserved across *Brassicaceae* species. In particular, we cannot detect significant evidence of purifying selection on 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 proportion of the genome. This implies that this approach is likely to also be missing lineage-specific selection when it comprises a relatively small fraction of sites, and highlights the importance of integrating additional evidence of function (comparative and experimental), for improved quantification of selection.

However, the general neutrality of noncoding regions based on population genomic analysis is consistent with that of Haudry and colleagues [20] 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 [18]. Halligan et. al. [27] recently used information from the DFE to infer the number of adaptive substitutions in mice both in coding and noncoding regions. They show that many more of the adaptive substitutions are in noncoding regions, and suggest that they may have regulatory function. In contrast our data show that *C. grandiflora* has roughly equal numbers of adaptive substitutions in 0-fold sites (53.2 Mb) compared to sites that are noncoding (26.0 Mb, 3' UTR; 10.0 Mb, 5' UTR; 33.0 Mb, CNS). Our results are thus consistent with previous suggestions that, unlike in animals, plant noncoding regulatory sequences may be less complex in their control of gene expression in plants, possibly due to more frequent gene duplication and functional divergence [19].

One important consideration 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 [31], the strength of selection inferred is very weak, and close to effective neutrality. Furthermore, synonymous site selection is expected to be stronger in more highly expressed genes [32,33], which would cause us to underestimate, rather than overestimate, the difference in the strength of purifying selection compared with lowly expressed 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 [34,35].

Unlike other classes of noncoding sequence, UTRs show relatively high levels of purifying selection, likely reflective of their function in post-transcriptional regulation [36]. 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 87% of 3'UTR and 75% of 5' UTR selected sites are outside of conserved regions. This suggests that there may be many species-specific functional regions in UTRs.

Measuring positive selection

In this study we took advantage of the two detectable signatures expected to remain after a classic selective sweep from new mutation: 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 noncoding 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 [27,37]. Our ability to detect the signature of recurrent sweeps may be caused by the fact that *C. grandiflora*, like *Drosophila*, and unlike humans and mouse, 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 [38-40]. We expect that soft sweeps will also be common in *C. grandiflora* because of its large N_e [40,41]. In addition, adaptation in genes that contribute to polygenic traits is often expected to occur without fixation of new mutations [42], and 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.

Expression level and selection

Highly expressed genes diverge less than genes with low expression in many species [10-12,43-47]. 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 on 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 [33,44,48].

Methods

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 S2). 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. Quality control and single-end genomic sequencing were conducted at the Genome Quebec Innovation Centre at McGill University. RNA from 10 of the 13 individuals was extracted from leaf tissue and sequenced at the Genome Quebec Innovation Centre, as reported in [24]. Prior to sequencing, leaf samples were flash frozen in liquid nitrogen and RNA extractions were performed using Qiagen's RNAeasy plant RNA extraction kit.

Genotyping

Genomic reads were aligned to the *Capsella rubella* reference genome [24] using the Stampy aligner 1.0.13 with default settings [49]. Sites around indels were realigned using the Genome Analysis Toolkit (GATK) v1.05777 indel realigner [23]. Genotype and SNP calls were conducted using the GATK UnifiedGenotyper with default parameters [50]. AFSs were generated from counts of sites in the generated 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 and if (4) the most likely genotype of any individual did not have a phred scaled likelihood score of 0, (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 (Figure S3). We identified these regions by splitting the genome into 200kb windows, then calculating F_{IS}. If F_{IS} 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.

Divergence

We calculated lineage-specific divergence in two ways. First, we aligned the *C. rubella* reference sequence with sequence data from *Arabidopsis thaliana* and *Neslia paniculata* using lastZ [51] with chaining, as previously described [20]. 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.

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 [52] and MultiParanoid [53]. The peptide sequences of these orthologs were aligned using DialignTX [54] and reverse-translated into coding sequence. Whole-gene divergence at synonymous and nonsynonymous sites was calculated using PAML [55], under a model where ω was allowed to vary in the *Capsella* lineage compared to other branches.

We conducted comparisons of estimates of the distribution of fitness effects using the two methods above with identical gene sets, and found a very strong concordance of results (data not shown). 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 (data not shown).

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 [20,56] using *C. rubella* as the reference genome. The Capsella genome sequence was then neutralized (bases replaced with N) and the PhastCons tool used to quantify family-wide levels of conservation. CNSs were then identified based on extended (>12nt) near-continuous regions of high conservation as previously described [20].

Estimates of the distribution of fitness effects and α

Site categories were determined based on the Joint Genome Institute's gene annotation of the *C. rubella* reference genome [24]. The allele frequency spectra (AFS) and divergence values were calculated for each category of sites, and DFE-alpha [22,57] was used to estimate the fraction of sites under negative selection and α using 4-fold degenerate sites as the neutral reference. The genome was broken up into 10 kb regions and these regions were bootstrapped 200 times to generate 95% CIs for selection on each category of sites. We tested for a significant difference in selection between the pooled set of CNSs and each individual category of CNSs using a randomization test, as in Keightley & Eyre-Walker (2009), 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.

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. For each of the above measures, we bootstrapped by substitution (n=1000) and removed the top and bottom 25 bootstraps to construct 95% confidence intervals. Following Hernandez and colleagues [37], 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 linked selection in noncoding DNA, we compared diversity around fixed substitutions within CNSs to diversity around fixed substitutions in non-conserved intergenic regions. To find these substitutions we compared 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.

Gene expression

RNA sequence was mapped to the *C. rubella* reference genome using Tophat 1.2.0 [58] and expression level was quantified from these mapped reads using Cufflinks 1.3.0 [59]. Cufflinks standardizes expression levels by gene length and library size, returning values in units of 'fragments per kilobase of exon per million fragments mapped' (FPKM). We calculated the mean expression level for each gene across our 10 samples and removed those genes with <1 FKPM to eliminate genes that may have been mis-annotated. The remaining 11,564 genes were divided into four, roughly equally sized categories based on expression level: low (1-6.8 FPKM), mid-low (6.8 – 17.5 FPKM), mid-high (17.5-44.7 FPKM), and high (44.7 – 17,092 FPKM). The distribution of fitness effects, α, and ω 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 generated 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.

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.

Author contributions

Conceived and designed the experiments: RJW EBJ AEP MB SIW. Performed the experiments and analysis: RJW EBJ AEP AH KMH MB. Wrote the paper: RJW EBJ SIW.

References

1. Keightley PD, Eyre-Walker A (2010) What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philos Trans R Soc Lond B Biol Sci 365: 1187–1193. doi:10.1098/rstb.2009.0266.

2. Kimura M (1991) Recent Development of the Neutral Theory Viewed from the Wrightian Tradition of Theoretical Population Genetics. Proc Natl Acad Sci USA 88: 5969–5973.

3. Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, et al. (2007) Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol 5: e310. doi:10.1371/journal.pbio.0050310.

4. Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the Drosophila genome? PLoS Genet 5: e1000495. doi:10.1371/journal.pgen.1000495.

5. Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD (2010) Evidence for Pervasive Adaptive Protein Evolution in Wild Mice. PLoS Genet 6: e1000825. doi:10.1371/journal.pgen.1000825.

6. Andolfatto P (2005) Adaptive evolution of non-coding DNA in Drosophila. Nature 437: 1149–1152. doi:10.1038/nature04107.

7. Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, et al. (2009) Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet 5: e1000592. doi:10.1371/journal.pgen.1000592.

8. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482. doi:10.1038/nature10530.

9. Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, et al. (2013) Genome-wide inference of natural selection on human transcription factor binding sites. Nat Genet 45: 723–729. doi:10.1038/ng.2658.

10. Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158: 927–931.

11. Subramanian S, Kumar S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168: 373–381. doi:10.1534/genetics.104.028944.

12. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23: 327–337. doi:10.1093/molbev/msj038.

13. Keightley PD, Gaffney DJ (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc Natl Acad Sci USA 100: 13402–13406. doi:10.1073/pnas.2233252100.

14. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8: 206–216. doi:10.1038/nrg2063.

15. Carroll SB (2005) Evolution at Two Levels: On Genes and Form. PLoS Biol 3: e245. doi:10.1371/journal.pbio.0030245.

16. Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. Evolution 61: 995–1016. doi:10.1111/j.1558-5646.2007.00105.x.

17. Gossmann TI, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, et al. (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol 27: 1822–1832. doi:10.1093/molbev/msq079.

18. Hough J, Williamson RJ, Wright SI (2013) Patterns of selection in plant genomes. Annu rev Ecol Evol Syst 44:3.1–3.19. doi:10.1146/annurev-ecolsys-110512-135851.

19. Lockton S, Gaut BS (2005) Plant conserved non-coding sequences and paralogue evolution. Trends Genet 21: 60–65. doi:10.1016/j.tig.2004.11.013.

20. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Gen. doi:doi:10.1038/ng.2684.

21. St Onge KR, Källman T, Slotte T, Lascoux M, Palmé AE (2011) Contrasting demographic history and population structure in Capsella rubella and Capsella grandiflora, two closely related species with different mating systems. Mol Ecol 20: 3306–3320. doi:10.1111/j.1365-294X.2011.05189.x.

22. Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26: 2097–2108. doi:10.1093/molbev/msp119.

23. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303. doi:10.1101/gr.107524.110.

24. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, et al. (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Gen 45: 831–835. doi:10.1038/ng.2669.

25. Slotte T, Foxe JP, Hazzouri KM, Wright SI (2010) Genome-Wide Evidence for Efficient Positive and Purifying Selection in Capsella grandiflora, a Plant Species with a Large Effective Population Size. Mol Biol Evol 27: 1813–1821. doi:10.1093/molbev/msq062.

26. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, et al. (2007) Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana. Science 317: 338–342. doi:10.1126/science.1138632.

27. Halligan DL, Kousathanas A, Ness RW, Harr B, Eőry L, et al. (2013) Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents. PLoS Genet 9: e1003995. doi:10.1371/journal.pgen.1003995.

28. Ehrenreich IM, Purugganan MD (2008) Sequence variation of MicroRNAs and their binding sites in Arabidopsis. Plant Physiol 146: 1974–1982. doi:10.1104/pp.108.116582.

29. Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G (2011) Pervasive Adaptive Protein Evolution Apparent in Diversity Patterns around Amino Acid Substitutions in Drosophila simulans. PLoS Genet 7: e1001302. doi:10.1371/journal.pgen.1001302.

30. Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 23–35. doi:10.1017/S0016672308009579.

31. Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D (2011) Reduced Efficacy of Natural Selection on Codon Usage Bias in Selfing Arabidopsis and Capsella Species. Genome Biology and Evolution 3: 868–880. doi:10.1093/gbe/evr085.

32. Wright SI, Yau CBK, Looseley M, Meyers BC (2004) Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Mol Biol Evol 21: 1719–1726. doi:10.1093/molbev/msh191.

33. Park C, Chen X, Yang J-R, Zhang J (2013) Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA 110: E678–E686. doi:10.1073/pnas.1218066110.

34. Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci USA 96: 4482–4487.

35. Marais G, Mouchiroud D, Duret L (2001) Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci USA 98: 5688–5692.

36. Kim Y, Lee G, Jeon E, Sohn EJ, Lee Y, et al. (2013) The immediate upstream region of the 5'-UTR from the AUG start codon has a pronounced effect on the translational efficiency in Arabidopsis thaliana. Nucleic Acids Res. doi:10.1093/nar/gkt864.

37. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, et al. (2011) Classic Selective Sweeps Were Rare in Recent Human Evolution. Science 331: 920–924. doi:10.1126/science.1198878.

38. Pennings PS, Hermisson J (2006) Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol 23: 1076–1084. doi:10.1093/molbev/msj117.

39. Pennings PS, Hermisson J (2006) Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet 2: e186. doi:10.1371/journal.pgen.0020186.

40. Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169: 2335–2352. doi:10.1534/genetics.104.036947.

41. Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol 28: 659–669. doi:10.1016/j.tree.2013.08.003.

42. Pavlidis P, Metzler D, Stephan W (2012) Selective sweeps in multilocus models of quantitative traits. Genetics 192: 225–239. doi:10.1534/genetics.112.142547.

43. Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL (2005) Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol 22: 1345–1354. doi:10.1093/molbev/msi122.

44. Yang J-R, Liao B-Y, Zhuang S-M, Zhang J (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci USA 114:E831-40. 10.1073/pnas.1117408109.

45. Slotte T, Bataillon T, Hansen TT, St Onge K, Wright SI, et al. (2011) Genomic Determinants of Protein Evolution and Polymorphism in Arabidopsis. Genome Biol Evol 3: 1210–1219. doi:10.1093/gbe/evr094.

46. Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203–218. doi:10.1038/nature06341.

47. Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, et al. (2012) Evidence for Widespread Positive and Purifying Selection Across the European Rabbit (Oryctolagus cuniculus) Genome. Mol Biol Evol 29: 1837–1849. doi:10.1093/molbev/mss025.

48. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352. doi:10.1016/j.cell.2008.05.042.

49. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21: 936–939. doi:10.1101/gr.111120.110.

50. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genet 43: 491–498. doi:10.1038/ng.806.

51. Harris RS (2007) Improved Pairwise Alignment of Genomic DNA. Ph. D. Thesis, The Pennsylvania State University.

52. Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, et al. (2009) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38: D196–D203. doi:10.1093/nar/gkp931.

53. Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22: e9–e15. doi:10.1093/bioinformatics/btl213.

54. Subramanian AR, Kaufmann M, Morgenstern B (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol 3: 6. doi:10.1186/1748-7188-3-6.

55. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591. doi:10.1093/molbev/msm088.

56. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14: 708–715. doi:10.1101/gr.1933104.

57. Keightley PD, Eyre-Walker A (2007) Joint Inference of the Distribution of Fitness Effects of Deleterious Mutations and Population Demography Based on Nucleotide Polymorphism Frequencies. Genetics 177: 2251–2261. doi:10.1534/genetics.107.080663.

58. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111. doi:10.1093/bioinformatics/btp120.

59. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515. doi:10.1038/nbt.1621.

Figure Legends

Figure 1.
Estimates of negative and positive selection in coding and noncoding regions across the *C. grandiflora* genome.
A) The proportion of sites found in each bin of purifying selection strength, separated by site type. B) α the proportion of divergent sites fixed by positive selection and C) ω the rate of adaptive substitution relative to neutral divergence. Error bars represent 95% bootstrap confidence intervals.

Figure 2
Linked neutral diversity and divergence as a function of distance from fixed substitutions across the *C. grandiflora* genome.
A) 4-fold diversity, B) 4-fold divergence, C) 4-fold diversity/divergence. In all figures, black lines represent measures surrounding fixed replacement substitutions and gray shading represents 95% confidence intervals, from bootstrapping, around silent substitutions.

Figure 3
Linked neutral diversity/divergence surrounding conserved noncoding sequences (CNSs).
A) 4-fold diversity/divergence as a function of distance from CNSs containing fixed substitutions (black line) and CNSs without fixed substitutions (gray shading, 95% confidence interval). B) 4-fold diversity/divergence as a function of distance from fixed substitutions in CNSs (black lines) and non-conserved intergenic sequence (gray shading, 95% confidence interval).

Figure 4
Estimates of negative and positive selection on 0-fold sites in genes of varying expression level.
A) The proportion of sites found in each bin of purifying selection strength, separated by site type. B) α the proportion of divergent sites fixed by positive selection and C) ω the rate of adaptive substitution relative to neutral divergence. Error bars represent 95% bootstrap confidence intervals.

Figure S1
Pairwise diversity and divergence at 4-fold degenerate sites across the entire genome.
The x-axis represents position along the 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 S2
Coding density versus 4-fold diversity (π across the genome.
Each point represents one 10 kb window. Black points are windows that do not overlap centromeres while grey points are for windows that do overlap centromeres. Both with and without centromeric windows there is a slight negative correlation.

Figure S3
Regions of identity by descent in each sample.
Within 200 kb windows the ratio of homozygous to heterozygous calls at sites that are polymorphic across individuals is plotted against position across the genome. Each sample is plotted separately. Individual lines alternating between grey and blue represent chromosomes. Regions of IBD were defined as windows where F_{IS} 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, near the end of chromosome 1, therefore other filtering steps downsampled sites to 23 chromosomes.

Figure S4
Estimates of positive and negative selection in different categories of CNSs.
A) Distribution of fitness effects. Stars indicate categories in which a type of CNS is significantly different from the pooled set of CNSs. B) $\dot{a}$and C) w for each category. Error bars indicate 95% CIs from 200 bootstraps.

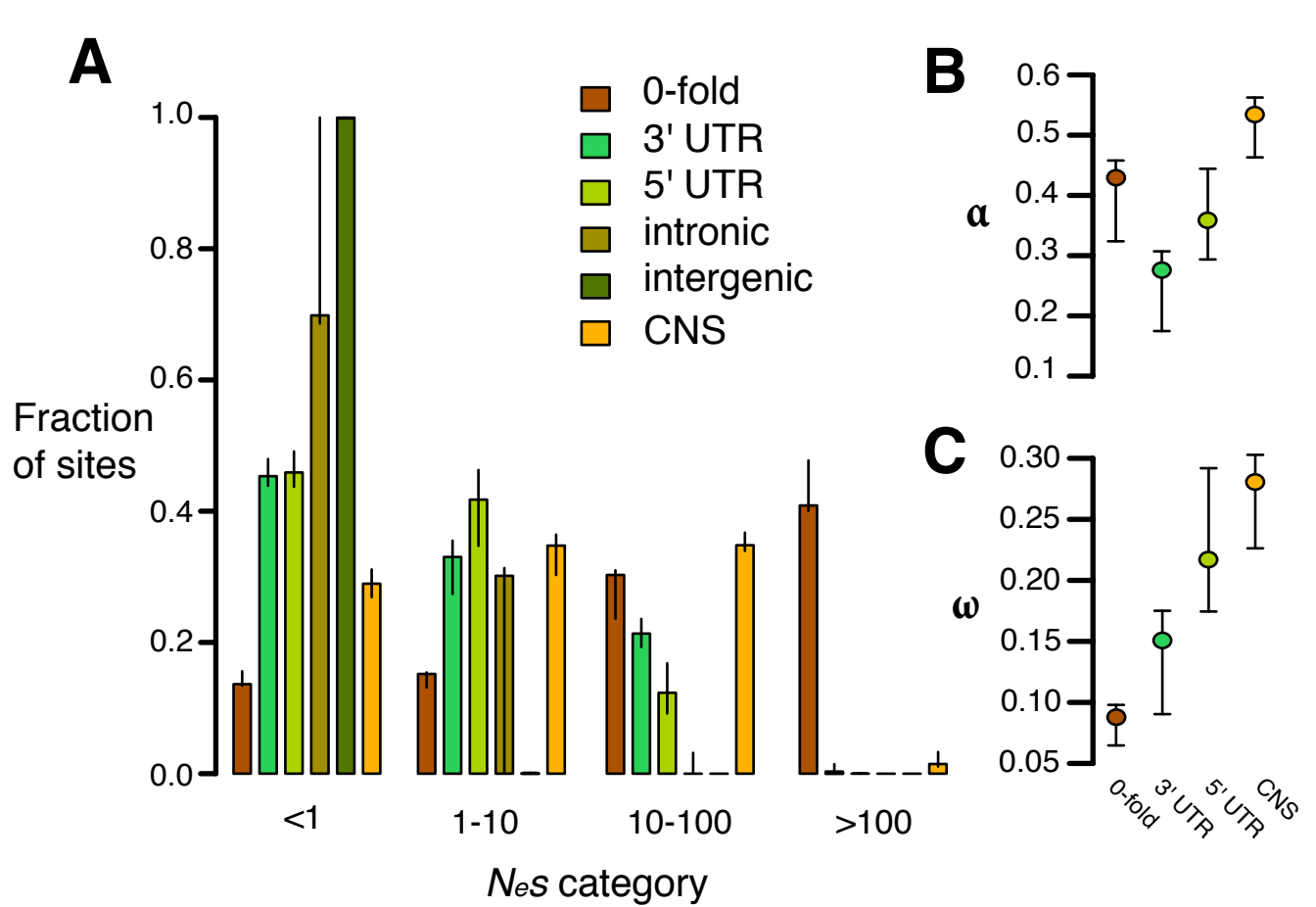
Figure S5
Robustness of sweep analysis to different window sizes.
This panel shows the results of our scans for recurrent selective sweeps using alternative window sizes: 500bp on left and 2kb on right. Otherwise, the methods are the same as described previously.

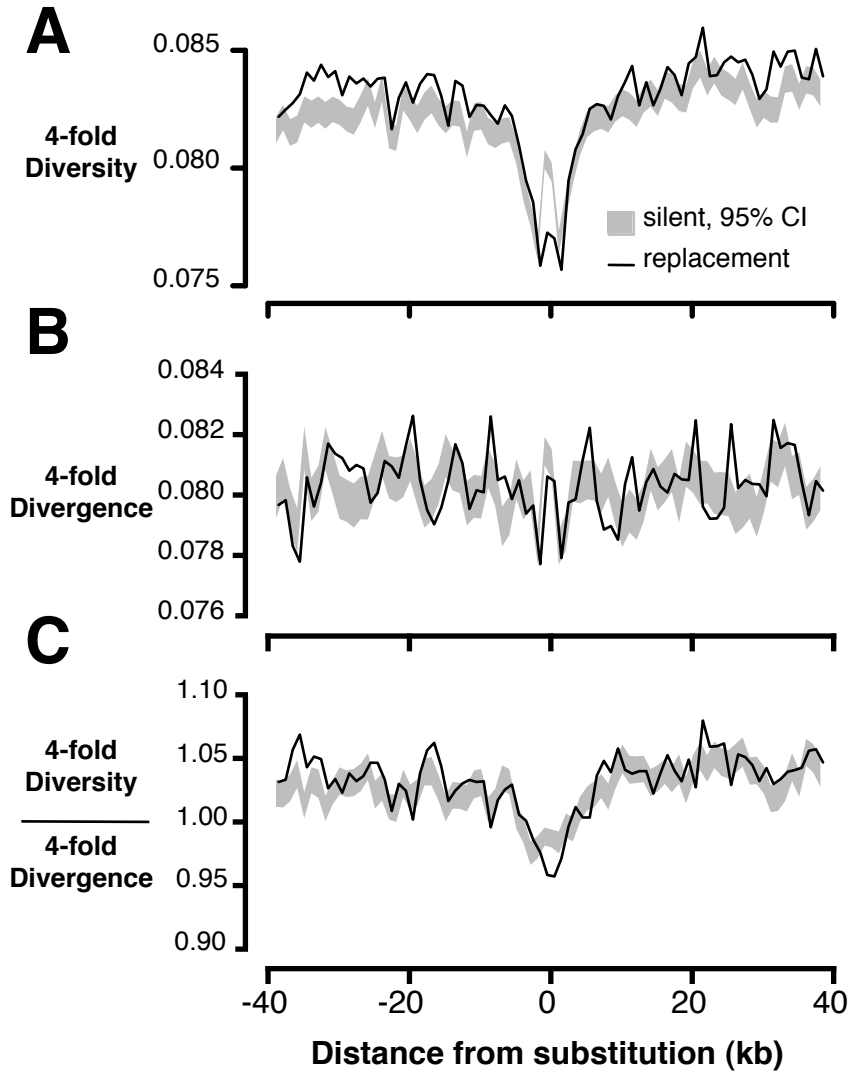
Figure S6
Additional diversity and divergence data for sweeps around substitutions in conserved noncoding regions.
The left panels show 4-fold diversity and 4-fold divergence around substitutions in conserved non-coding sequences (black lines) and non-conserved intergenic sequence (gray shading represents 95% confidence intervals). The right panels show the same information for 4-fold diversity around conserved noncoding sequences containing fixed substitutions (black lines) and conserved noncoding sequences without fixed substitutions (gray shading represents 95% confidence intervals).

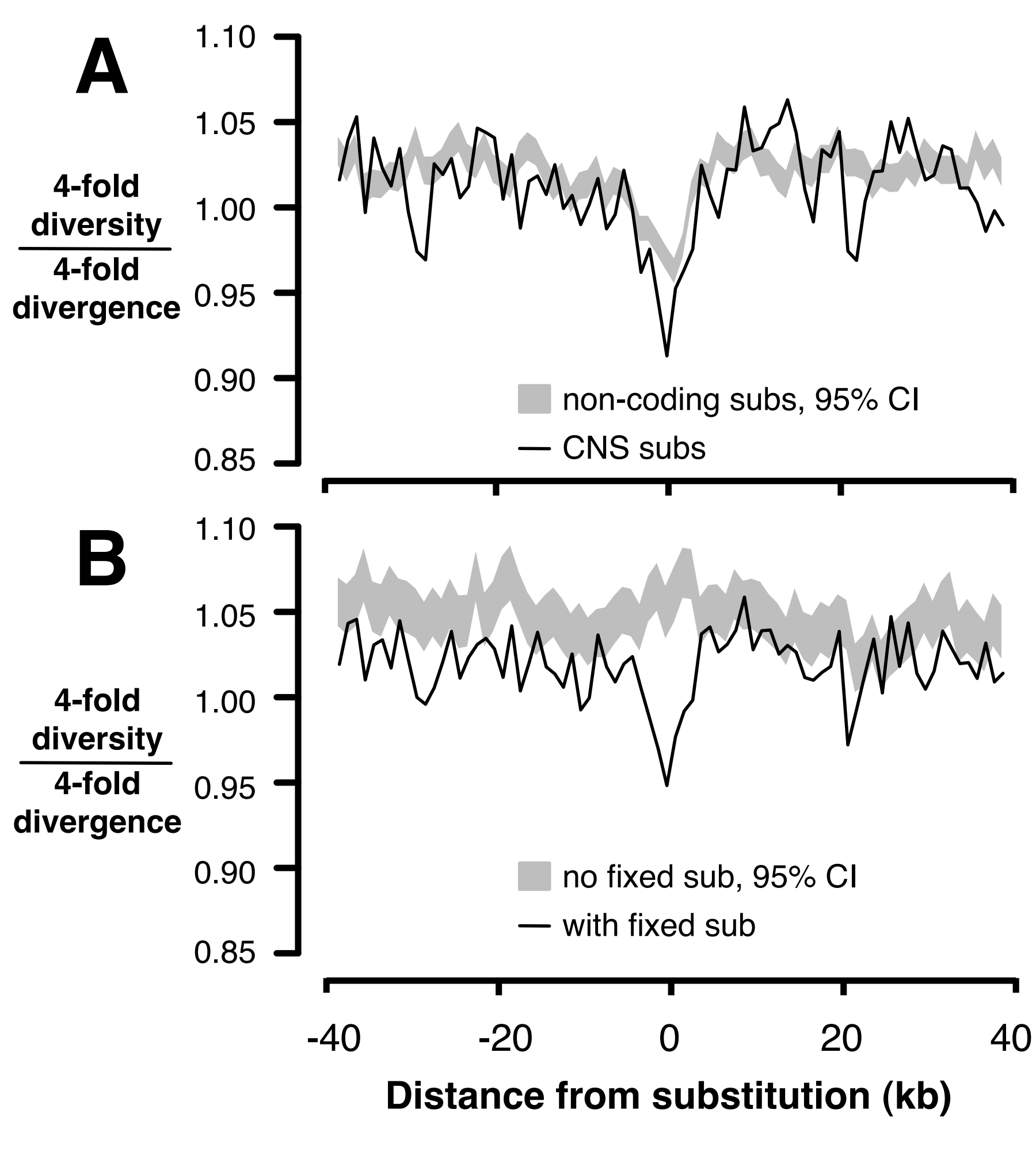
Figure S7
Allele frequency spectra of high and low expression genes.

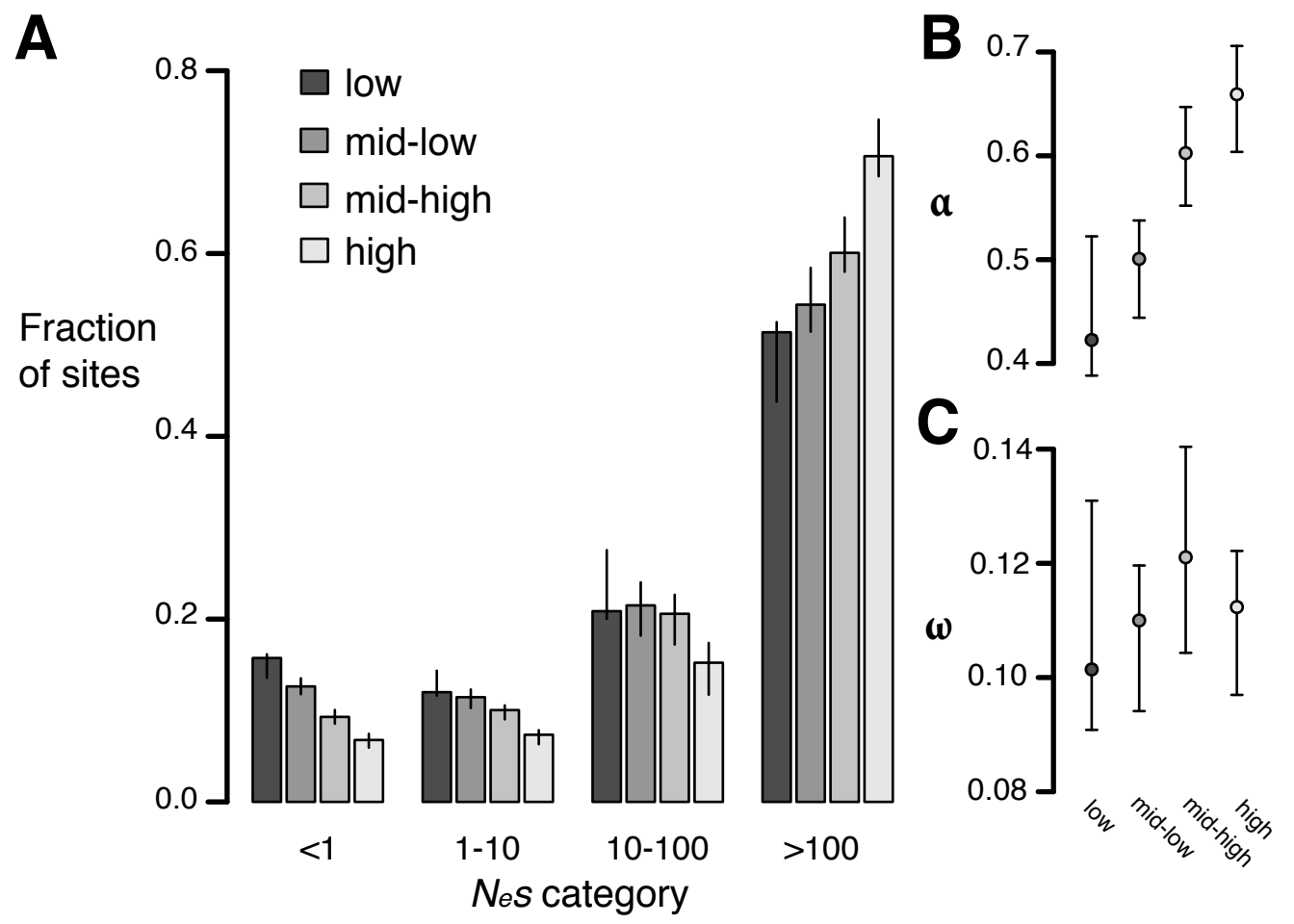
Table S1
Summary of diversity statistics of each site category across all samples.

Table S2
Sampling locations of each sample.


A
B
C
0-fold
3' UTR
5' UTR
intronic
intergenic
CNS
Fraction of sites
1.0
0.8
0.6
0.4
0.2
0.0
<1
1-10
10-100
>100
Nes category
α
0.6
0.5
0.4
0.3
0.2
0.1
ω
0.30
0.25
0.20
0.15
0.10
0.05
0-fold
3' UTR
5' UTR
CNS


A
4-fold Diversity
0.085
0.080
0.075
silent, 95% CI
replacement
B
4-fold Divergence
0.084
0.082
0.080
0.078
0.076
C
4-fold Diversity
4-fold Divergence
1.10
1.05
1.00
0.95
0.90
-40
-20
0
20
40
Distance from substitution (kb)


A
B
4-fold diversity / 4-fold divergence
1.10
1.05
1.00
0.95
0.90
0.85
non-coding subs, 95% CI
CNS subs
no fixed sub, 95% CI
with fixed sub
-40
-20
0
20
40
Distance from substitution (kb)


A
0.8
0.6
0.4
0.2
0.0
Fraction of sites
low
mid-low
mid-high
high
<1
1-10
10-100
>100
Nes category
B
0.7
0.6
0.5
0.4
α
C
0.14
0.12
0.10
0.08
ω
low
mid-low
mid-high
high