Association mapping reveals the role of mutation-selection balance in the maintenance of genomic variation for gene expression.

Emily B. Josephs[1], Young Wha Lee, John R. Stinchcombe*, Stephen I. Wright*

Dept. of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks st., Toronto ON, M6R 1M3, Canada.

* these authors contributed equally to the work
1 Corresponding author, em.josephs@utoronto.ca

Abstract:
The evolutionary forces that maintain genetic variation for quantitative traits within populations remain unknown. 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 effect size and frequency. Overall, our results show that mutation-selection balance is the dominant contributor to genomic variation for expression within a single, outcrossing population.

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 coefficient[1]. While studies of allele frequency spectra show that purifying selection is often prevalent in genomic sequence[2–4], little is known about how genetic variants under selection relate to phenotype, and ultimately, how phenotypic variation is maintained within populations. Association mapping can identify specific loci influencing traits, providing candidates for further analysis of selection[5]. 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[6].

Early eQTL studies mapped expression divergence between two lines, finding that many genes show local, polygenic expression divergence[7,8]. These studies have provided insight into selection on eQTLs; for example, a correlation between recombination rate and eQTL density implies that background selection is a dominant force acting on expression variation in *Caenorhabditus elegans*[9]. However, eQTL studies of

population-level genetic variation have thus far been limited to a few study systems[10–13] and only one study, in humans, has identified a negative correlation between effect size and frequency[13]. To date, eQTL studies in plants have used genetic crosses[14–16] or species-wide samples[17–19], 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.

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*[3,20], making it an ideal system for investigating selection on QTLs.

Results

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 at called SNPs by individual of 45. 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 (Supplementary Table 1).

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 (Figure 1). We omitted rare homozygotes from the analysis because most local regulation acts additively in *cis*[11] and low sample sizes for rare variants reduces power. Out of 5,507,316 SNPs tested, 39,628 SNPS are significantly associated with expression of 6,624 nearby genes (FDR = 0.1, $p < 0.00082$, Supplementary Fig. 1). These SNPs often clustered locally (Supplementary Figure 2), 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[12], 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).

If eQTLs act in *cis*, heterozygous eQTLs will cause allele-specific expression, providing an additional signature of regulatory variation. To detect and map QTLs for allele-specific expression (aseQTLs), we performed Mann-Whitney U tests comparing the average difference in expression standardized for sequencing depth (ASE), between alleles of the same gene in individuals that were homozygous at a local SNP and those that were heterozygous at that SNP (Figure 1). We excluded coding SNPs from this analysis because their genotype might confound ASE measurement. Out of 3,966,423 SNPS tested, 26,957 SNPs were significantly associated with ASE of 5,882 nearby genes (FDR = 0.1, $p < 0.00054$, Supplementary Fig. 1). We did not require a directional effect of SNP genotype on ASE, but 22,436 (83%)

of 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 SNP per gene for further analysis and we also required that ASE had to be higher in heterozygotes than homozygotes, leaving 4,580 aseQTLs.

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 (Figure 2), consistent with data from humans and Drosophila[10,11,13]. In addition, CNSs near the TSS were enriched for eQTLs and aseQTLs relative to non-conserved sites (Figure 2), suggesting that genetic variation within CNSs represents a major source of standing variation in gene expression. In contrast, CNSs in 5'UTRs and introns were not enriched for eQTLs or aseQTLs, consistent with observations that selection strength is relatively similar in conserved and non-conserved sites in these regions[3]. However, the detection of a large number of eQTLs outside of conserved regions suggests that regulatory element turnover is common in Brassicaceae (Supplementary Table 2). There were 2,236 genes that had both eQTLs and aseQTLs, significantly more than expected by chance ($X^2 = 471$, $p < 2.2x10^{-16}$). Of these 2,236 genes, 411 had the same SNP most significantly associated both with expression and ASE.

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

To generate an appropriate null distribution for QTL allele frequency, we permuted assignments between expression level and genotype for every gene 1000 times and ran eQTL analyses using permuted data. On average, 3,258 SNPs were associated with total expression in our permutations, consistent with an FDR of 0.1, since 39,628 SNPs were associated with 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 (Figure 4). This observation is conservative, because we have not accounted for reduced power to detect associations on rare alleles. We also investigated permuted aseQTLs, and found on average 3,194 SNPs associated with ASE in each permutation, which is slightly more than expected given our FDR of 10% (26,597 SNPs were associated with ASE in un-permuted data). As with eQTLs, aseQTLs were significantly rarer than those found in permuted data (Figure 4). 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[1], suggesting that QTLs under purifying selection should show a negative correlation between minor allele frequency and effect size, assuming that effect size correlates with the

strength of selection. However, this correlation is also expected if QTLs evolve neutrally because of winner's curse[21]. Essentially, effect size estimation error is greater for rare alleles, and when effect size is over-estimated, an association is more likely, leading to a negative correlation between effect size and minor allele frequency. To avoid the double-testing issue responsible for winner's curse, we decoupled the identification of associations from the estimation of effect size by comparing allele frequencies of SNPs identified as eQTLs with these SNP's effects on ASE. Consistent with mutation-selection balance, ASE effect size was negatively correlated with eQTL allele frequency (Figure 4a, R^2 = -0.128, p=3x10^-15, n = 2,832). There was also a negative correlation between aseQTL frequency and effect size for total expression (Figure 4b, correlation = -0.035, p = 0.02 , N = 4,580).

Our mapping of QTLs for expression and allele-specific expression genome-wide in a single population of *C. grandiflora* demonstrates that the frequencies and 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 CNS's potential role as regulatory elements; however, the large number of eQTLs discovered outside of conserved regions suggests significant turnover in regulatory elements between species. Taken together, our results, indicate that much of the expression variation observed at the population level is deleterious and support mutation-selection balance's role in maintaining genetic variation in populations.

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 Asher Cutter, Alan Moses, Tanja Slotte, Adrian Platts, and Graham Coop for helpful comments. We also thank Thomas Bureau, Mathieu Blanchette, Daniel Schoen, Paul Harrison, Alan Moses, Adrian Platts, and Eef Harmsen for their contributions to the Value-directed Evolutionary Genomics Initiative (VEGI) grant (Genome Quebec/Genome Canada). This work was supported by the previously mentioned VEGI grant from Genome Canada and Genome Quebec, an NSF Graduate Research Fellowship to EBJ (DGE-1048376), and NSERC Canada and CFI grants to JRS and SIW.

Materials and Methods
Study system and plant material

Capsella grandiflora is an obligately outcrossing member of the Brassicaceae family with a large effective population size (*Ne*~600,000), relatively low population structure and a range that spans northern Greece and southern Albania[20,22]. In June 2010, we collected seeds from approximately 400 plants growing in a roadside population of *C. grandiflora* near Monodendri, Greece (Population Cg-9[22]). We germinated and grew one individual from each parent in the University of Toronto greenhouses and performed crosses between independent random pairs of plants to generate the seeds used in this study. By growing the parents in a common environment and then assaying their progeny in a common environment, we reduced the influence of maternal effects and unknown micro-environmental effects on gene expression.

Approximately 10 seeds from each cross were sterilized in 10% bleach followed by 70% ethanol, placed on sterile plates filled with 0.8% agar with Mursashige-Skoog salts (2.15 g/L), stratified in the dark at 4ºC for one week, and then allowed to germinate in a growth chamber at 22ºC 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.

Genomic data

We mapped DNA sequence data to the *C. rubella* reference genome[23] 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*[3] as well as suspect realignments (transposable elements, centromeres, 600bp intervals containing extreme Hardy-Weinberg deviations, 1kb intervals with evidence of 3 or more snps in reference-to-reference mapping).

A relatedness analysis revealed that six individuals were more related to eachother than expected in an outcrossing population, perhaps because of introgression from *C. rubella*, so we removed these individuals from the analysis. In addition, RNAseq readmapping for two individuals was very poor quality (<10% reads mapped and paired correctly), so these individuals were removed as well. Our final sample size was 99 individuals.

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[23]. 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 (Supplementary Figure 3).

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[11], but we did not detect this (Supplementary Figure 4). 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.

Mapping local eQTL

We selected SNPs for our eQTL analysis by finding all SNPs within the window spanning 5 kb upstream of the gene's transcription start site and 5kb downstream from the gene's 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[17]. 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[23]. In addition, we identified SNPs located in non-coding sequence conserved across the Brassicaceae family[3]. We only included SNPs with at least 10 heterozygous individuals and 10 individuals that were homozygous for the common allele in our sample.

We wrote set of Python scripts to test for associations between expression level and genotype at nearby SNP 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 and we compared common homozygotes to heterozygotes because we expect most local eQTLs to act in *cis* and thus be additive[11], and because not being limited by the sample size of rare homozygotes allowed us to map eQTL at relatively rare alleles. We controlled for multiple testing by using a false discovery rate approach[24] and only considering eQTLs to be associated with expression if that association has 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. 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.

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 human studies[13]). We quantified allele-specific expression at all heterozygous sites inferred from the genomic data. We used the

count of reads mapped to each allele, taken from the 'AD' values in a VCF file constructed from the RNAseq data using GATK Unified Genotyper to calculate an allele-specific-expression measure ('ASE) for each gene in each individual. Specifically, we calculated the mean of the the differences in allelic expression values at all heterozygous sites across a gene and divided this mean by median expression level of all genes in the individual to control for sequencing depth. While we expected that our measure of gene-wide ASE would be more accurate when we required multiple heterozygous sites per gene, doing so did not significantly alter the number of aseQTLs we found or their allele frequency distribution, so we only required one heterozygous site per gene to measure ASE (Supplementary Figure 5).

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. 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 calculated ASE effect size for aseQTLs and eQTLs by taking the difference between mean ASE in homozygotes and mean ASE in heterozygotes.

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[12]. 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.

References

1. Haldane, J. B. S. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Math. Proc. Cambridge Philos. Soc.* **23,** 838–844 (1927).
2. Kousathanas, A., Oliver, F., Halligan, D. L. & Keightley, P. D. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Mol. Biol. Evol.* **28,** 1183–1191 (2011).
3. Williamson, R. J. *et al.* Evidence for widespread positive and negative selection in coding and conserved noncoding regions of Capsella grandiflora. *PLoS Genet.* **10,** e1004622 (2014).
4. Zhu, Q. *et al.* A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *Am. J. Hum. Genet.* **88,** 458–468 (2011).
5. Lee, Y. W., Gould, B. A. & Stinchcombe, J. R. Identifying the genes underlying quantitative traits: a rationale for the QTN programme. *AoB Plants* **6,** (2014).
6. Haudry, A. *et al.* An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* **45,** 891–898 (2013).
7. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752–755 (2002).
8. Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 1572–1577 (2005).
9. Rockman, M. V., Skrovanek, S. S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in C. elegans. *Science* **330,** 372–376 (2010).
10. Massouras, A. *et al.* Genomic variation and its impact on gene expression in Drosophila melanogaster. *PLoS Genet.* **8,** e1003055 (2012).
11. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
12. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
13. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24,** 14–24 (2014).
14. Potokina, E. *et al.* Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant J.* **53,** 90–101 (2008).
15. West, M. A. L. *et al.* Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* **175,** 1441–1450 (2007).
16. Bolon, Y.-T., Hyten, D. L., Orf, J. H., Vance, C. P. & Muehlbauer, G. J. eQTL Networks Reveal Complex Genetic Architecture in the Immature Soybean Seed. *Plant Genome* **7,** 0 (2014).
17. Gan, X. *et al.* Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477,** 419–423 (2011).
18. Zhang, X., Cal, A. J. & Borevitz, J. O. Genetic architecture of regulatory variation in Arabidopsis thaliana. *Genome Res.* **21,** 725–733 (2011).
19. Fu, J. *et al.* RNA sequencing reveals the complex regulatory network in the maize kernel. *Nat. Commun.* **4,** 2832 (2013).
20. Slotte, T., Foxe, J. P., Hazzouri, K. M. & Wright, S. I. 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 (2010).
21. Capen, E. C., Clapp, R. V. & Campbell, W. M. Competitive bidding in high-risk situations. *Journal of petroleum* (1971). at <https://www.onepetro.org/journal-paper/SPE-2993-PA>
22. ST. ONGE, K. R., Källman, T., Slotte, T., Lascoux, M. & Palmé, A. E. 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 (2011).
23. Slotte, T. *et al.* The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **45,** 831–835 (2013).
24. Dabney, A. & Storey, J. D. *qvalue: Qvalue estimation for false discovery rate control.* R *package version 1.140.0*

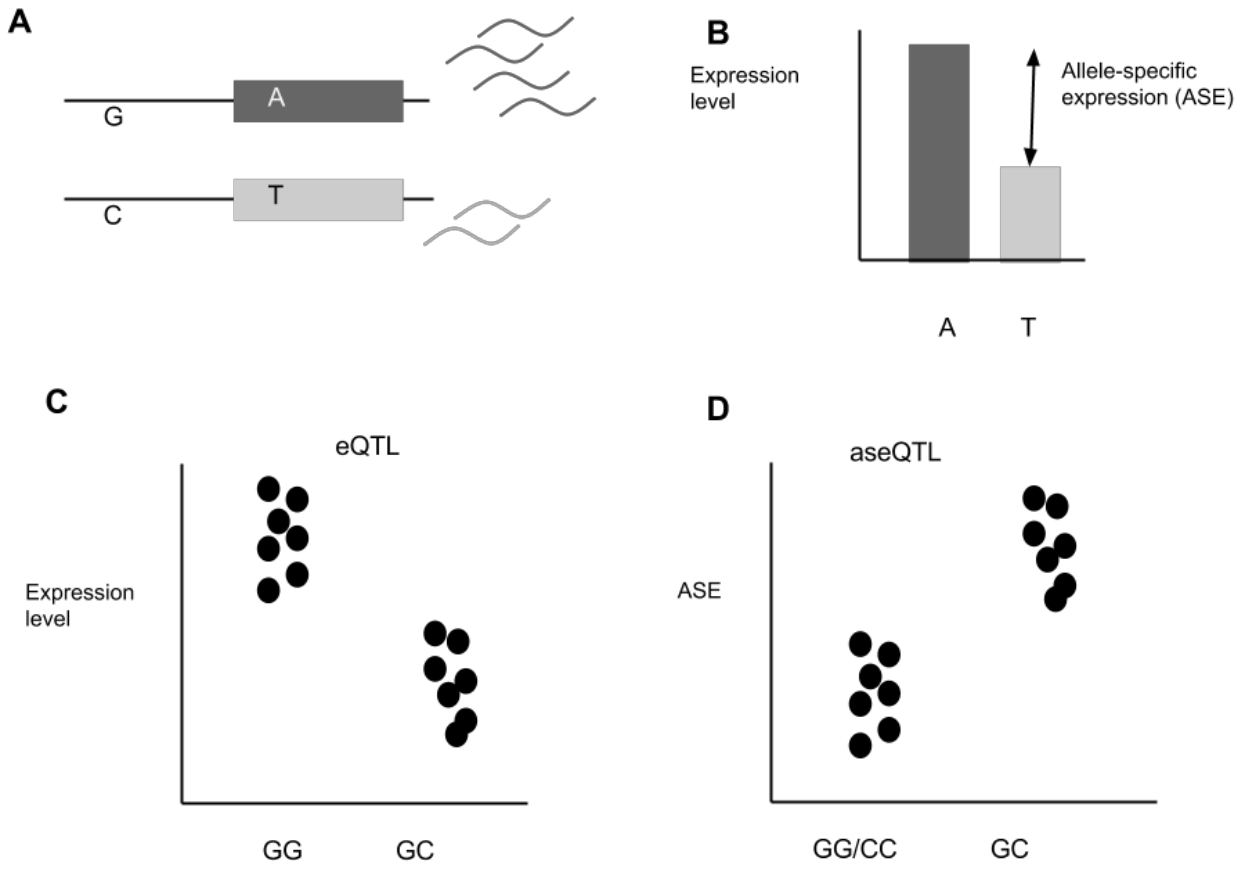

Figure 1: Detecting eQTLs and aseQTLs. (a) A gene model for an individual that is heterozygous at a regulatory locus (G/T) and at an informative coding site (A/T). The G allele increases expression relative to the C allele, (b) causing increased allelic expression of the reads carrying the A allele at 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 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.

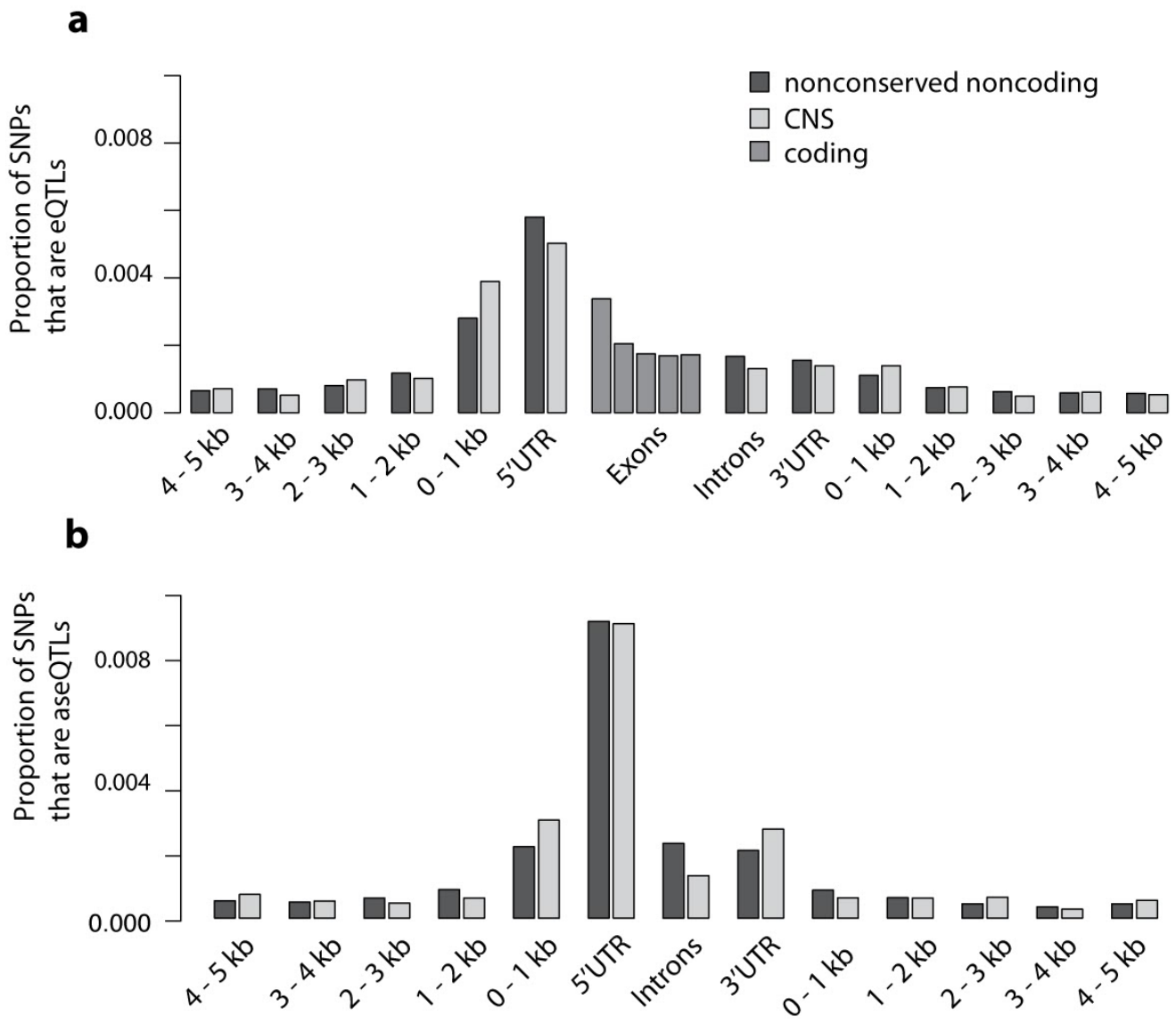

Figure 2: eQTL and aseQTL enrichments by site type. The proportion of SNPs tested in each category that were found to be eQTLs is plotted on the y axis for (a) eQTLs and (b) aseQTLs. Note that there were no exonic SNPs included in the aseQTL analysis.

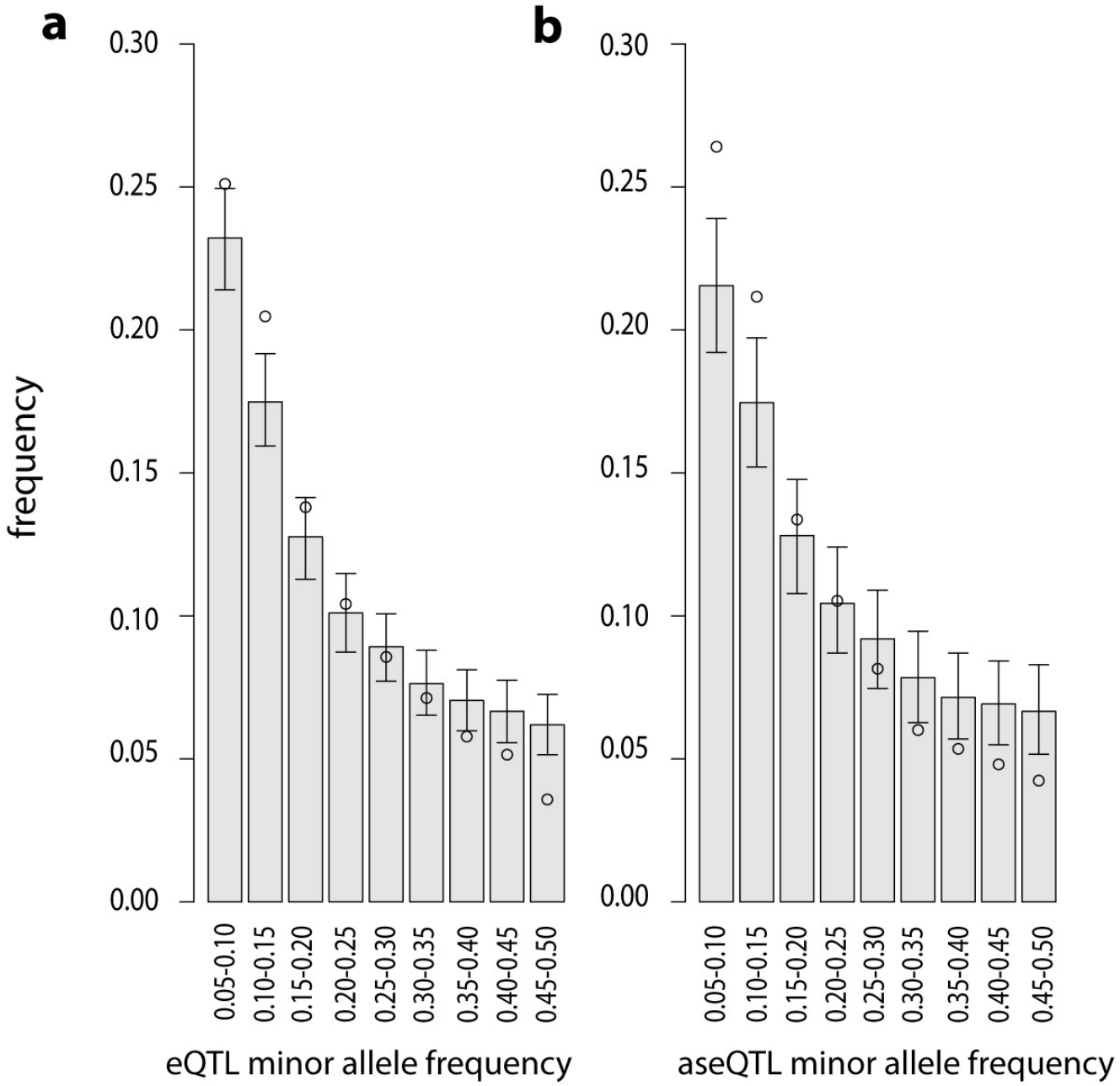

Figure 3: The site frequency spectra of eQTLs and aseQTLs. A histogram of minor allele frequencies of (a) eQTLs and (b) aseQTLs for observed data (black circles) and permuted data (gray bars with 95% confidence intervals).

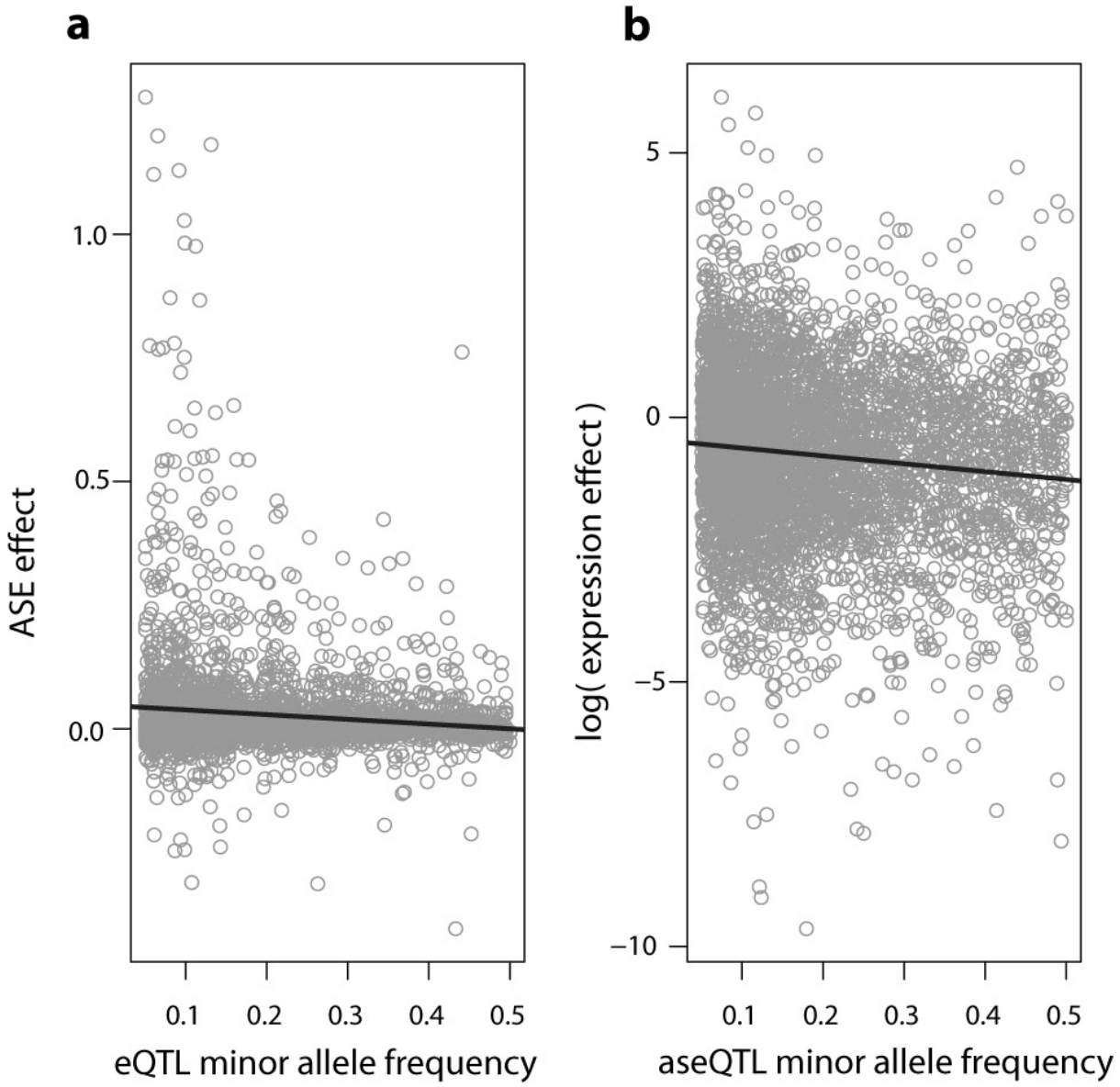

Figure 4: The relationship between minor allele frequency and effect size. (a) eQTL minor allele frequency is plotted against the effect of that SNP on ASE, calculated as the mean difference in ASE between individuals heterozygous at the eQTL and individuals homozygous at the eQTL. Negative values occur when the the homozygote for the eQTL has greater ASE than the heterozygote. The red 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.