Genome-wide positive and negative selection in *Capsella grandiflora*

Robert J Williamson[1]*, Adrian E Platts[2], Khaled M Hazzouri[1], Stephen I Wright[1]

1. University of Toronto, Department of Ecology and Evolutionary Biology
2. McGill University, Department of Biology

Abstract

A central goal in evolutionary genetics is determining what parts, and how much of the genome is constrained and subject to adaptive evolution. Studies in *Drosophila* have found that between 10 % and 50% of between-species differences may have been driven to fixation by non-neutral processes, and that a large fraction of adaptive sites are in noncoding DNA. We are investigating genome wide positive and negative selection in this plants, using *Capsella grandiflora* as a model system. Using Illumina sequence data we determined allele frequency spectra for polymorphic sites in 26 Capsella grandiflora chromosomes. These spectra, and divergence data from *Neslia paniculata,* were then used to determine the distribution of fitness effects and proportion of adaptive substitutions using a recently-developed maximum likelihood approach (Keightley and Eyre-Walker 2007).

Why *Capsella*?

The *Capsella* system is an emerging model in population genetics. Previous studies have shown that *Capsella grandiflora* has low population structure and long term stable population size.

Capsella grandiflora also been shown to have high levels of polymorphism, and efficient positive selection compared to *Arabodopsis* (Slotte et al. 2010).

Data

We have whole genome sequence from 13 *Capsella grandiflora* individuals. After filtering for depth and quality our data set comprises 33,888,429 sites, of which 1,877,815 are SNPs.
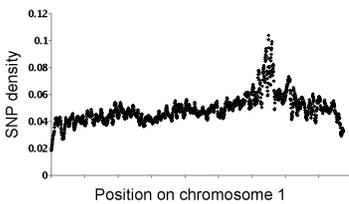

Figure 1. SNP density in 20,000 bp windows along chromosome 1. The peak in diversity is around the centromere.

Analysis pipeline

Illumina sequencing of 13 *Capsella grandiflora* individuals and one *Neslia paniculata* outgroup.

Alignment of samples to *Capsella rubella* genome using **bwa**.

Identification of 0-fold (nonsynonymous), 4-fold (synonymous), intron, flanking, and intergenic regions from *Capsella rubella* annotation.

Estimation of allele frequency spectrum (AFS) using **SAMtools**.

Maximum likelihood estimate of selection in different classes of sites using the Keightley Eyre-Walker DFE-alpha method.
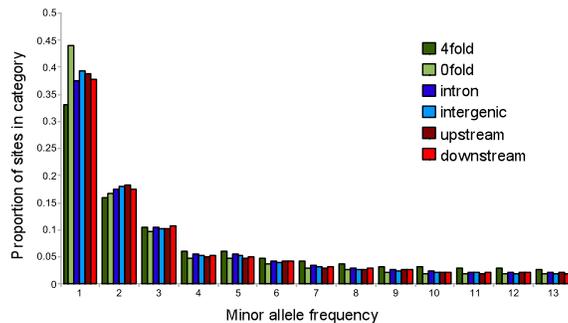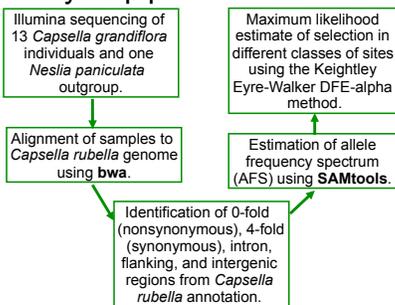

Figure 2. Allele frequency spectra for six classes of sites in the *Capsella grandiflora* data set. An excess of low frequency polymorphisms are seen in all classes compared to 4-fold degenerate sites, indicative of negative selection.
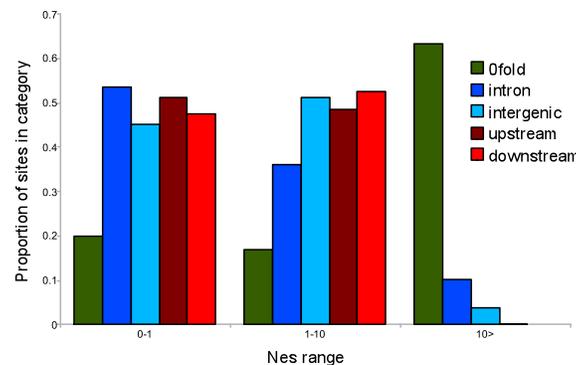

Figure 3. Selection strength on different categories of sites. Zero fold sites have many strongly selected sites compared to other categories, which have moderate levels of selection.

References

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.

Keightley, P. D. and 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.

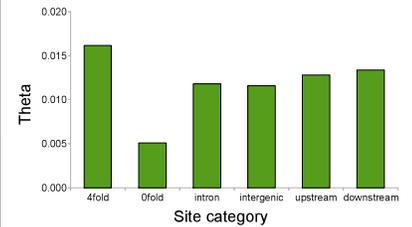*Capsella grandiflora* photo courtesy of Emily Josephs.

Figure 4. Levels of diversity in the different site categories, measured by theta.
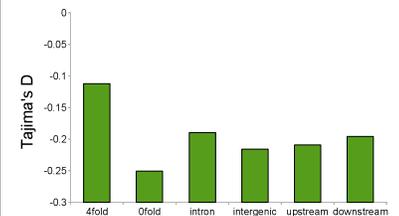

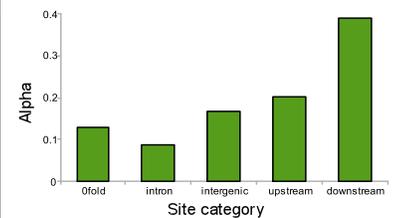Figure 5. Tajima's D for the six site categories. More negative values are indicative of negative selection.

Figure 6. Proportion of adaptive substitutions in each site category.

Conclusions

We have shown that noncoding DNA in *Capsella grandiflora* is constrained compared to 4-fold synonymous sites. There is a large proportion (~65%) of 0-fold sites under strong selection, while most other sites have moderate to weak selection acting on them. Additionally, there is a evidence of adaptive substitutions in noncoding regions of *Capsella grandiflora*, especially downstream of genes.

We will continue to refine these parameter estimates, generate confidence intervals, and investigate other site categories.