**Appendix S2.** Testing the effects of linkage: single-site-per-locus analyses and locus-level bootstrapping with Ex*D*FOIL

**Introduction**

A potential flaw of *D*FOIL is that the Chi-square method that it uses for determining significance assumes that sites are independent. If this assumption was violated (e.g., by linkage between sites used for count data), it could theoretically increase type-I error. The authors of *D*FOIL demonstrated through simulation that when sufficient sites are sampled, the *D*FOIL components should approach a Chi-square distribution, despite non-independence between sites (Pease & Hahn, 2015). However, their formula for determining the number of sites required depends on estimates for effective population size and recombination rate, which we lack for our species. Instead, we used two different approaches to address the potential effects of linkage on our inferences of introgression. First, we sub-sampled our data to include only one informative site per locus and re-analyzed our data, to address the impact of linkage within loci (we refer to these as single-site analyses hereafter). Second, we performed bootstrapping analyses, resampling entire loci with replacement, to address the impact of linkage between loci. For both sets of analyses, we used the targeted dataset with reduced-individual sampling. This was the primary dataset used for introgression detection in the main text, and the dataset used for preparation of major figures (Figures 3,4).

**Methods**

*Single-site-per-locus sub-sampling*

In our primary dataset, we often used SNPs that occur on the same ddRAD locus. This may be problematic because sites located near one another on the genome are likely to be linked because of limited recombination. To eliminate linked sites within loci, we retained only a single SNP per ddRAD locus for any given *D*FOIL test. We then ran *D*FOIL on this reduced dataset and compared the number of inferred introgression events to those obtained from the full dataset (using multiple SNPs per ddRAD locus). If linkage between SNPs within loci was a salient issue we predicted that we would observe either (1) significantly fewer introgression events or (2) significantly different patterns of introgression across test cases. A drawback of this approach is that it greatly reduces the overall number of sites considered by *D*FOIL, presumably reducing statistical power. To combat this, we took steps to ensure that the retained sites were informative for a given *D*FOIL test.

In order to sub-sample our data while maintaining statistical power, we selected SNPs on a test-by-test basis and ensured that retained SNPs could be used by *D*FOIL. This approach can be broken down into two major steps: first, filtering the SNPs across all loci to ensure that only SNPs that can be used by *D*FOIL are retained; and second, retaining a maximum of one informative SNP per ddRAD locus, to remove presumably linked SNPs.

To filter SNPs, we first removed all individuals from the vcf file except those 5 individuals (P1–P4, and the outgroup) being considered for the test at hand, and then removed any sites with missing data for any of these taxa, as sites with any missing data are not considered by *D*FOIL. We then removed sites with any heterozygous individuals (also ignored by *D*FOIL). We also removed sites for which only one out of five taxa exhibited a base that was different from those in the other taxa. Thus, we excluded singleton counts, given that their inclusion may be problematic (discussed in main text). These steps ensured that all retained sites could be used in *D*FOIL calculations. Finally, we thinned the vcf files so that only a single SNP per ddRAD locus was retained and converted to fasta format. We used vcftools v0.1.15 (Danacek et al., 2011), custom shell scripts, and publicly available perl scripts to accomplish these filtering steps. We then executed a modified version of the Ex*D*FOIL pipeline that paired each test with its respective fasta file. All required scripts for filtering and execution of Ex*D*FOIL are available at <https://www.github.com/SheaML/ExDFOIL/AppendixS2_Materials>.

*Locus-level bootstrapping*

Another potential source of linkage in our dataset would be linkage between ddRAD loci. To address the potential effects of this linkage, we conducted bootstrapping of entire ddRAD loci. As with the single-site analyses, if linkage between loci was a salient issue, we expected to observe significantly fewer introgression events or significantly different patterns of introgression when assessing significance using the locus-level bootstrapping.

A potential issue when conducting bootstrapping for Ex*D*FOIL is that each replicate will require the same number of tests as the original dataset (>32,000 for the targeted dataset with reduced individual sampling). This makes conducting a large number of bootstrap replicates (e.g., 1000) computationally quite expensive, and in our case, led to conflicts with our available resources at the University of Arizona HPC. We took two steps to reduce the computational load of our bootstrap analyses.

First, we restricted the potential pool of tests to those with two individuals of *S. ornatu*s for P1 and P2, and two individuals of S. *oberon* for P3 and P4, with at least one representative of *oberon*-black (as in Tables 1–3 and Figure 3 of the main text). Our main analyses suggest that nearly all introgression signatures involve this set of taxa (see section **3.4** of the main text). This restricted set retains 12,376 tests in total.

Second, to examine the impact of using different numbers of bootstrap replicates in a computationally feasible manner, we first conducted 1000 bootstrap replicates using a randomly sampled set of 100 tests that returned ancestral introgression using the Chi-square method of *D*FOIL, and 100 tests that returned no introgression. We conducted random sampling using the sample() function of R v3.4.1 (R Core Team, 2017), sampling from the restricted set of 12,376 tests. We then compared the test results using the full set of 1000 bootstrap replicates to test results using 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, and 900 replicates. We randomly sampled the respective number of replicates from the full set 100 times to provide distributions of the potential results produced by each level of replication. To summarize the results, we used the number of tests that remained positive (out of 100 tests that returned ancestral introgression using the Chi-square method) and the number of tests that remained negative (out of 100 tests that returned no introgression using the Chi-square method). To compare results across the different levels of replication, we used the boxplot() function in R.

Figure S6 provides the results of our pilot bootstrapping analyses, using 1000 replicates on 100 tests that recovered ancestral introgression and 100 tests that recovered no introgression using the full data. As the number of bootstrap replicates increases, the number of tests successfully recovered as either positive or negative tends to increase. Using the full set of 1000 bootstrap replicates, we recovered 100/100 tests as negative, and 94/100 as positive. Based on these results, we chose to conduct 100 bootstrap replicates for the full set of 12,376 tests. Although a larger number of replicates might still be preferable, it appears that 100 bootstrap replicates should sufficiently mimic the results using 1000 replicates in most cases. Furthermore, our largest concern – increasing type-I error by using a smaller number of bootstrap replicates – does not appear to be an issue. If anything, we may experience slightly higher type-II error rates relative to a larger number of bootstrap replicates.

For the locus-level bootstrapping analyses, we re-sampled entire ddRAD loci with replacement (following Eaton & Ree 2013) such that each replicate had the same number of loci as in the original dataset. We used the command-line utility rl v0.2.7 (http://ch.tudelft.nl/~arthur/rl/) to accomplish this random re-sampling using plain-text lists of loci. Scripts for the generation of bootstrap replicates are available at <https://www.github.com/SheaML/ExDFOIL/AppendixS2_Materials>. We then ran the Ex*D*FOIL pipeline on each of the bootstrapped datasets, using the dfoilalt option to specify that singleton counts be ignored.

We used the results of our bootstrap analyses to measure the standard deviation of each *D*FOIL component for each test, using the std() function in R. We then used the number of standard deviations away from 0 (the expectation for *D*-statistics) as *Z*-scores for each *D*FOIL component. We conducted two-tailed tests of significance for each *Z*-score using a *z*-distribution and a significance threshold of 0.01 (matching the threshold used for the Chi-square distribution by *D*FOIL). Finally, we compiled the resulting *D*FOIL signatures for comparison to the results using the Chi-square method and full dataset.

**Results**

Complete results for the single-site and bootstrapping analyses are found at <https://www.github.com/SheaML/ExDFOIL/AppendixS2_Materials>. The results of these analyses are summarized and compared to the targeted dataset results from the main text in Table S17 and Figure S7, considering only tests with two representatives of *S. ornatus* and two representatives of *S. oberon* (with at least one *oberon*-black), as in Tables 1–3 of the main text. For Table S17, we summarize the results based on the species involved and provide the raw number of positive tests for each category. In Figure S7, we visually compare the results obtained using the full dataset, locus bootstrapping, and single-site analyses; using bar plots to indicate the number of tests that recovered each of the five most common results.

Overall, we found ~85% of the number of total introgression signatures using the single-site datasets (83.1%) and bootstrapped datasets (85.7%), relative to the targeted dataset from the main text (Table S17). For ancestral introgression signatures, we recovered 87.9% (single site per locus) and 86.7% (locus-level bootstrapping) of the total number of introgression results as compared to our main analyses using all loci and sites.

**Discussion**

We cannot say with confidence if any of our results are truly examples of type-I or type-II error, because we used an empirical dataset. Nevertheless, the patterns uncovered by our single-site and bootstrapping analyses provide an overall assessment of how influential linkage might be in our dataset. We focus on ancestral introgression events, given that directional events are potentially problematic (see main text). For the single-site and bootstrapping analyses, we found that fewer ancestral introgression events (~12–13% fewer; see Table S17) were detected. One interpretation of these results is that the Chi-square method occasionally infers ancestral introgression erroneously, owing to the effects of linkage between sites and/or loci. As such, we recommend that users combining RADseq data and *D*FOIL use bootstrapping and/or sub-sampling to provide additional support for any inferred cases of introgression.

At the same time, our overall results from the single-site and bootstrapping analyses were quite similar to the results using the full data (Figure S7). This similarity suggests that potential instances of linkage-related type-I error are not so misleading that general patterns of introgression are at risk of being misdiagnosed. Furthermore, we might expect that our single-site analyses suffer from reduced statistical power relative to the full data, as many fewer sites are considered. Also, our pilot bootstrapping analyses indicate that a slightly larger number of positive results may have been recovered if using a larger number of bootstrap replicates (e.g., 1000). In summary, by using single-site sampling and bootstrapping, we demonstrate that the Chi-square method of *D*FOIL was robust to linkage issues in ~85% of the cases considered. We encourage users to conduct similar sensitivity analyses when combining RADseq data and *D*FOIL.

**References**

Danecek, P., Auton, A., Abecasis, G., Albers, C., Banks, E., DePristo, M. A., Handsaker, R., Lunter, G., Marth, G., Sherry, S.T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27, 2156–2158. doi: 10.1093/bioinformatics/Fbtr330

Eaton, D. A. R., & Ree, R. H. (2013). Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology*, 62, 689–706. doi: 10.1093/sysbio/syt032

Pease, J. B., & Hahn, M. W. (2015). Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology*, 64, 651–662. doi: 10.1093/sysbio/syv023

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.