Check F_{is} and LD of microsatellite loci

Jenna Melanson

2025-06-20

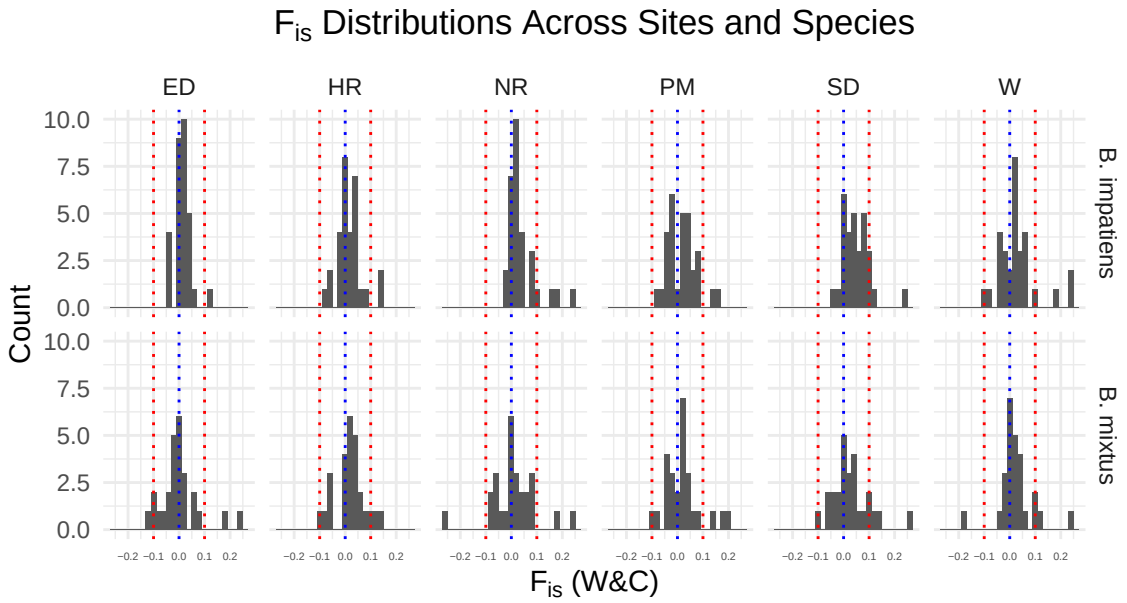
Remove siblings before marker assessment

We used an iterative approach to assign workers and queens to their natal colonies. First, using the pedigree reconstruction software COLONY 2.0 (Jones and Wang 2010), we assigned full siblingships based on all available microsatellite data, assuming male and female monogamy and no inbreeding. A single run was carried out for each species and year, using the software's full-likelihood approach and no siblingship size scaling or priors. Siblingships were maintained at this stage only if P(full sibling dyad) = 1. A single individual from each putative colony (including non-circular colonies) was maintained for downstream analyses of microsatellite locus quality.

Check distribution of F_{is} estimates across loci

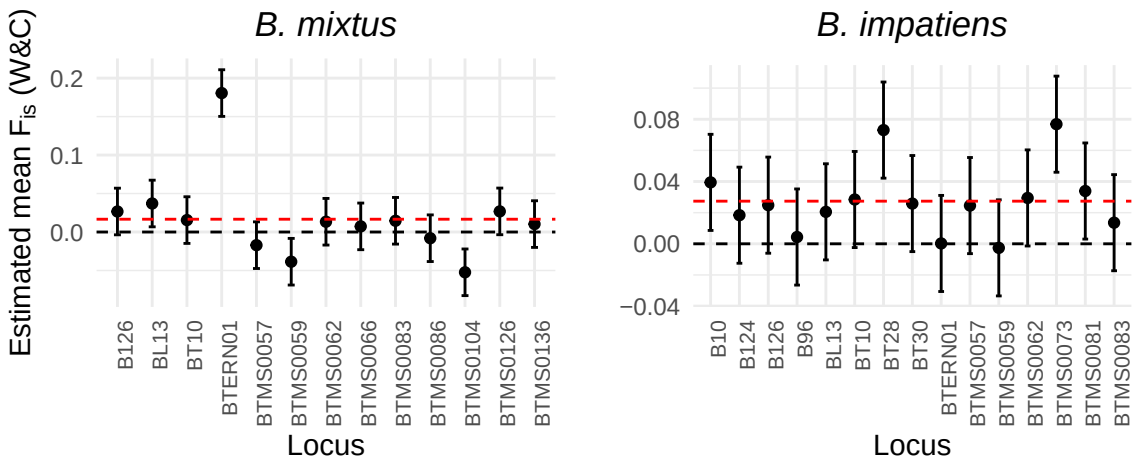
While COLONY 2.0 can account for inbreeding at the population level, locus-specific estimates of F_{is} should generally be similar to one another, reflecting their shared evolutionary history. Loci with F_{is} estimates that deviate significantly from the species mean (across loci) may suffer from null alleles or other types of scoring errors that can bias siblingship assignment. We therefore tested individual locus deviations from population mean F_{is} as a criterion for marker inclusion/exclusion.

Single locus F_{is} estimates were computed for all site, year, species groups following Weir and Cockerham (1984) in the package *genepop* (Rousset 2008). The distributions of F_{is} estimates for each species at each site are shown below. The blue dotted line indicates $F_{is} = 0$, and the red dotted lines indicate $F_{is} = \pm\ 0.1$.



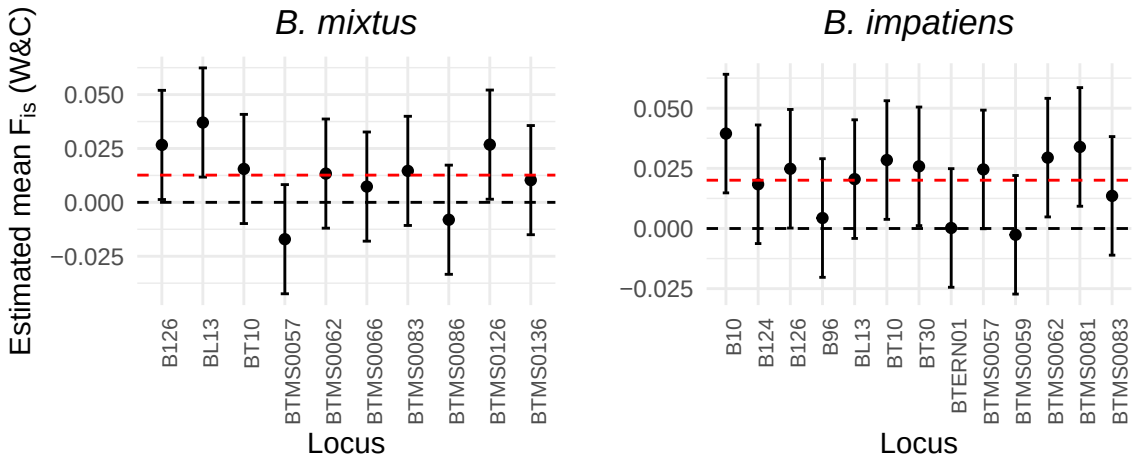
We fit linear models to F_{is} estimates with locus and site as fixed predictors. We used sum-to-zero coding so that model intercepts represented mean F_{is} across all loci, and calculated the estimated marginal mean of

each locus using the *emmeans* package (Lenth 2024). Marginal means are plotted below; the dashed black line denotes $F_{is} = 0$, and the dashed red line denotes mean F_{is} for each species.



We iteratively removed loci with F_{is} significantly different from the species mean F_{is}, starting with the locus with the greatest deviation, and re-running the model after each removal (i.e., because removing a locus with a high or low inbreeding coefficient will change the species mean estimate and therefore all comparisons to the mean). We did not apply an adjustment for multiple-hypothesis testing, but instead utilized a relatively stringent p-value ($\alpha = 0.01$) for removal of loci.

This process resulted in the removal of loci BTERN01, BTMS0104, and BTMS0059 from downstream analyses for *B. mixtus* and removal of BTMS0073 and BT28 for *B. impatiens*. F_{is} estimates for the remaining loci (n = 10 for *B. mixtus*, n = 13 for *B. impatiens*) are shown below.



Check markers for linkage disequilibrium

Next, we checked locus pairs for linkage disequilibrium (LD). Marker linkage can lead to non-independent assortment, a condition which violates the assumptions of COLONY. LD was calculated for each locus pair in each population (site, year, species groups) in *genepop*. We applied a Bonferroni correction for multiple hypothesis testing, and flagged locus pairs which showed significant deviations in > 2 populations.

[[Is there a better way of approaching this? It seems odd to me to correct for multiple hypothesis testing and then only drop loci if they exceed the Bonferroni-correct p-value in *multiple* cases–it's common practice in the literature, but perhaps it would be better to first assess F_{st} between sites and years and then test LD of loci across *all* samples rather than individually for each population?]]

There was not strong support for linkage disequilibrium between locus pairs in *B. mixtus*. Four pairs showed signs of LD, but each occurred at only a single site in a single year. For *B. impatiens*, there was strong evidence for linkage disequilibrium between BTMS0057 and BL13 (7 out of 12 populations showed significant LD following Bonferroni correction). To determine which locus to maintain, we calculated the polymorphic information content (PIC) using the package *PopGenUtils* (Tourvas 2025).

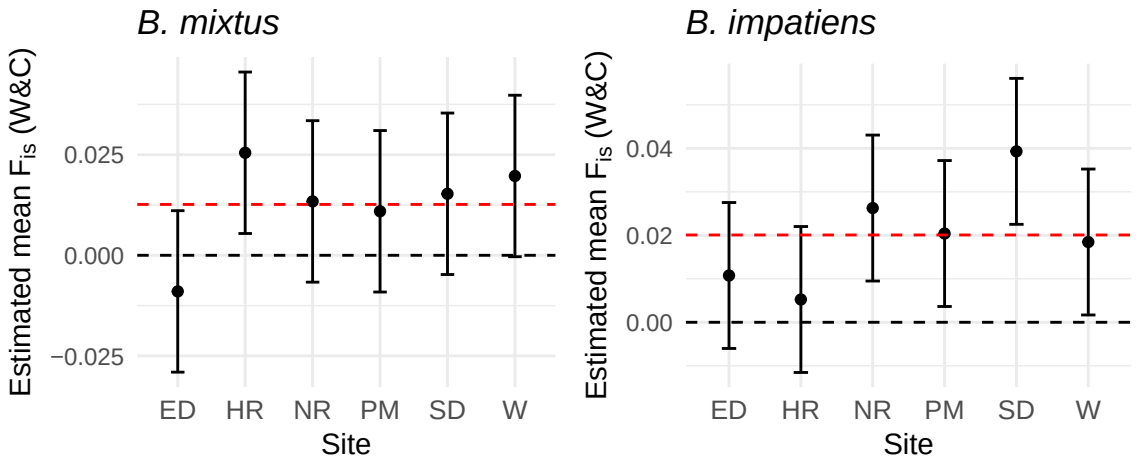
We found that BTMS0057 had higher PIC in both years (PIC = 0.78) compared to BL13 (PIC = 0.47 in 2022 and 0.45 in 2023). For this reason, we chose to maintain BTMS0057 and remove BL13 from further analyses for *B. impatiens*.

Calculate F_{st} and F_{is} between sites

After locus quality screening, we calculated global and pairwise F_{st} following Nei (1987) in the package *hierfstat* (Goudet 2005).

For both species and years, estimates of global F_{st} were less than 0.005. Pairwise F_{st} ranged from -0.002** to 0.01, indicating little or no genetic differentiation between surveyed sub-populations. [[**I read that F_{st} can sometimes be negative if between population variance is estimated to be slightly smaller than within population variance, as an artifact of finite sampling–would you agree, and if so is it better to report like this or round to 0?]]

Finally, we computed the marginal mean F_{is} for each site to determine whether there was evidence for inbreeding following locus removal, and whether it varied between sites.



Finalize siblingship assignments

We used our updated marker sets to more stringently assign siblingships. Jones and Wang (2010) state that the inbreeding model in COLONY 2.0 should only be implemented for dioecious species when there is strong evidence for high levels of inbreeding. Because we found evidence for only minor inbreeding (e.g., $F_{is} < 0.05$) we ran all final colony assignments using the no-inbreeding model. Given the very low estimates of global and pairwise F_{st}, we chose to combine sites for siblingship assignments to maximize the accuracy of allele frequency estimation and to validate our methodology (described below) for excluding spurious siblingships.

[[I have not run the finalized colony assignments yet, bu the method I have come up with for excluding the spurious siblingships is to run COLONY 5 times for each population and only maintain the siblingships which occur with high probability in all of them. I will need to develop a simple heuristic for dealing with non-circular families (e.g., A related to B, B related to C, A not related to C). Currently I am thinking: (1) check if A is related to C in at least 3/5 runs of COLONY, (2) if yes to (1), keep the whole family. If no to (1), randomly draw a sib pair to maintain (AB or BC). This seems to happen only rarely from what I can see, so hopefully it won't be hugely problematic.]]

Goudet, Jérôme. 2005. "Hierfstat, a Package for r to Compute and Test Hierarchical F-Statistics." *Molecular Ecology Notes* 5 (1): 184–86. https://doi.org/10.1111/j.1471-8286.2004.00828.x.

Jones, Owen R., and Jinliang Wang. 2010. "COLONY: A Program for Parentage and Sibship Inference from Multilocus Genotype Data." *Molecular Ecology Resources* 10 (3): 551–55. https://doi.org/10.1111/j.1755-0998.2009.02787.x.

Lenth, Russell V. 2024. "Emmeans: Estimated Marginal Means, Aka Least-Squares Means." Manual. https://CRAN.R-project.org/package=emmeans.

Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press. https://books.google.ca/books?id=UhRSsLkjxDgC.

Rousset, François. 2008. "Genepop'007: A Complete Re-Implementation of the Genepop Software for Windows and Linux." *Molecular Ecology Resources* 8 (1): 103–6. https://doi.org/10.1111/j.1471-8286.2007.01931.x.

Tourvas, Nikolaos. 2025. "PopGenUtils: A Collection of Useful Functions to Deal with Genetic Data in R." Manual.

Weir, B. S., and C. Clark Cockerham. 1984. "Estimating F-Statistics for the Analysis of Population Structure." *Evolution* 38 (6): 1358–70. https://doi.org/10.2307/2408641.