Looking at high FST and highly negative Tajima’s D (which suggests a loss in intermediate frequency alleles) should, in theory, give us regions with less variation in the post-invasion samples than in the pre-invasion samples. FST finds regions where variation within one or both populations is lower than variation between the two populations. This includes instances that we want (low variation in post-invasion samples but more variation in pre-invasion samples) but also instances that we don’t want, including regions with high variation in post samples and low variation in pre, which could happen through mechanisms like gene flow or relaxed selection.

It would also include regions with low variation in post-invasion samples and distinct but still low variation in pre-invasions samples. This would be unlikely but could occur due to random chance and low sampling effort. It could also occur if a region experienced a selective sweep pre-invasion and then ALSO experienced a second selective sweep on a different haplotype in the post invasion (highly unlikely). This latter circumstance with low variation in pre and low variation in post would result in very high FST values which we do not see.

Consider only the former two options that could drive high relative FST – low variation in post and high variation in pre (which I’ll call the SSV option for selection on standing variation) and high variation in post but low variation in pre (which I’ll call the EXP option for expansion). I think that we can distinguish between these two options using ∆ Tajima’s D. This statistic is sensitive to changing levels of intermediate frequency alleles. We expect to see a loss of intermediate frequency alleles during a selective sweep, which leads to a negative ∆ Tajima’s D. We expect to see an increase of intermediate frequency alleles during relaxed selection, which leads to a positive ∆ Tajima’s D.

If FST is biased but ∆Tajima is so well tuned to disentangling selection from other forces in short periods, why not only use ∆Tajima? Tajima could decrease due to a variety of forces as well. I think the most convincingly possible force is that purifying selection may strengthen in many regions of the genome that are indirectly affected by *Philornis*. A low-frequency deleterious allele that slightly reduces the fitness of a bird under normal circumstances could be more quickly weeded out under circumstances with a more global selective effect on body condition. This changes the SFS through linked regions with the deleterious allele and skews Tajima’s D. But these would not change FST, because the frequency of the beneficial allele, which is already high, might not change all that much.

Other options include: Gene flow followed by drift could cause a decrease in Tajima’s D, but would only result in a high FST if the gene flow were followed by positive selection (which would still be a region of interest). Reduced balancing selection would drive a negative Tajima’s D, as would a population expansion. Given the boom-bust cycles of the Galápagos and a lack of population size data, I don’t think we can yet rule out stochastic expansion with no relevance to the parasite in a way that could cause some regions of the genome to decrease in Tajima’s D at a randomly higher rate than others.

To spot check this theory, I took the top 0.1% of windows with a high FST and a ∆Tajima’s D of less than -1 and performed a GO test. This identified three genes with involvement in angiogenesis: HPSE, ANGPT1, and ITGA2B. I looked at a PCA of the windows that contained these genes and identified them as outliers. All three align with predictions, with post-invasion samples clustering more tightly than pre-invasion samples. This suggests an ongoing sweep of these genes. In each, blue is “post-invasion” and red is “pre-invasion”.

Importantly, the post-invasions individuals that form clusters are distinct individuals across genes. This suggests that it isn’t inbreeding driving these trends, or some element of genomic sequencing error (average depth of coverage, for instance). These windows input into the PCA are 50-65kb long. Looking at Manhattan plots of the SNP FST, the signal for HPSE and ANGPT1 appears to be within the gene, while the signal of differentiation for ITGA2B appears to be immediately following the gene (potentially gene regulation). I think the evidence of an ongoing selective sweep is strongest in ITGA2B with 8/9 samples from the post-invasion forming a cluster, second strongest in HPSE (6/9), and weakest in ANGPT1 (5/9).

Mean ∆Tajima: -0.338

Mean FST: 0.038

Mean abs(∆AF): 0.011

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gene** | **Window** | **FST** | **∆Tajima** | **∆AF** |
| HPSE | 59800000 | 0.158 | -1.845 | -0.033 |
|  | 59812500 | 0.165 | -2.098 | -0.043 |
| ANGPT1 | 133612500 | 0.237 | -1.747 | -0.023 |
|  | 133600000 | 0.245 | -1.564 | -0.015 |
| ITGA2B | 1175000 | 0.284 | -1.300 | -0.019 |

HPSE (Heparanase)

A graph with red and blue dots

AI-generated content may be incorrect.

ANGPT1 (Angiopoietin-1)

A graph with red and blue dots

AI-generated content may be incorrect.

ITGA2B (Integrin Subunit Alpha 2b)

A graph with red and blue dots

AI-generated content may be incorrect.