nature portfolio

Peer Review File

Genomic signatures of adaptation in native lizards exposed to human-introduced fire ants

Corresponding Author: Dr Braulio Assis

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this manuscript, Assis and colleagues investigate genomic signature of adaptation in a native species (the fence lizard) that has been exposed to an invasive species (the red fire ant) representing both a venomous predator and a toxic prey. The authors survey three fence lizard populations (only one of which was invaded by fire ants; n=20 per population). For this focal set, they relied on whole genome sequencing at high coverage (~13x), which they used to reconstruct patterns of population structure, and to scan the genomes for the signature of natural selection. The authors also obtained low-coverage (~4x) whole genome sequencing data for an additional 361 individuals from the invaded site. This low-coverage dataset was used to understand the genetic architecture of limb length (a candidate adaptive trait in the context of fire ant invasion), via GWAS.

Results provided evidence for 101 genomic regions with a signature of natural selection (as confirmed using at least two of three metrics used), across the three populations. Of these regions, 42 were detected only in the invaded population. These candidate adaptive loci were enriched for several functional categories, such as myofibril, complement and coagulation cascades, or osteoblast signalling. The same patterns of functional enrichment were generally not observed in the two non-invaded populations. Finally, when combining genome scan results with GWAS results, there were two candidate limb length loci (out of the 24 loci revealed by GWAS) that overlapped with genomic regions inferred to have been shaped by natural selection in the invaded site. The authors interpret these results considering the adaptive potential of natural populations that are exposed to abrupt human-mediated change.

I enjoyed reading the manuscript, and I think it tackles a key topic (evolutionary responses of native-range populations to introduced invasive species) that deserves more study. As well, the manuscript is well written. That said, I think there are important limitations, which I detail below.

1. The sampling available for this study is a major limitation, in my view. Because only one invaded site was included, we cannot confidently exclude the possibility that genomic signatures of natural selection detected here are related to some other aspect of the environment that is different between the Alabama (invaded) site and the Arkansas and Tennessee (noninvaded) sites. Indeed, one can assume important differences related to latitude (e.g. climate) are also at play. IA sampling design with some level of replication could have mitigated this confounding factor.

2. It is not entirely clear why the authors chose those specific three metrics for inferring the action of natural selection. Additional explanation would be helpful, especially since two of the three metrics (saltilassi and LSBL) are not widely used. How exactly do these metrics complement one another, and why are they the best metrics to be used here, given expectations on the genetics of adaptation? Presumably, because adaptation in this context is rapid (and occurring in a population within its ancestral range), soft selective sweeps on standing genetic variation would predominate. Are the metrics used here reliable for detecting such sweeps?

3. The use of very low coverage sequencing for the association study (~4x) means that genotyping errors are more likely, unless the authors took steps to confirm that indeed such errors are rare, and their genotype calls are reliable (I did not see mention of this). While missing genotype data is imputed, how do we know that this imputation step did not introduce errors? Are there replicates available (e.g., individuals sequenced repeatedly) that could be used to check the genotyping errors (before and after the imputation step)? In some of the associations detected (e.g., the chromosome 2 locus; Fig. 3B),

sampling is highly uneven between the three genotype classes, a scenario that is known to be prone to false positives (and could be further exacerbated by genotyping errors).

4. The discussion focused on gene ontology detracts from the readability in my view, as it is largely speculative, and includes terms (e.g., gene names) that many readers of the paper would not be familiar with. Because no additional functional experiments are conducted to verify these associations (and considering the limited population sampling used in the genome scans and the low coverage used in the GWAS), it is possible that are least some of these results are due to experimental noise. Thus, I recommend that the authors try to condense this section.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
In their manuscript “Genomic signatures of adaptation in native lizards exposed to human-introduced fire ants”, Assis and colleagues describe analyses designed to detect genomic signatures of selection in an Eastern Fence Lizard population subject to a novel selection pressure – the introduction of non-native fire ants 70 years ago. The authors perform a series of population genomic analyses to identify regions where selective sweeps may have occurred in one lizard population impacted by fire ants and two unaffected populations then contrast summaries of these results among populations. They further perform a GWAS on a much larger sample of animals all from the impacted population to map putative genomic regions associated with hindlimb length a trait known to impact ant removal by fence lizards. While I believe this a very interesting system and the data collected are appropriate to answers the author’s questions, I have some major concerns regarding the manner in which these analyses were performed. I detail those concerns as well as more minor comments below.

Major Concerns
My concerns are largely limited to the population genomic inferences made with the 20 sample-per-population, 13x coverage, population genomic dataset. The comments that follow do not apply to the larger sample set used for GWAS analyses (although I am less well-versed in these analyses and reserve more careful scrutiny to other reviewers).

First, I believe the authors initially sought to use the two non-invaded populations as near and more distance outgroups to polarize their analyses of the AL population. Upon discovering that there was not a straightforward relationship among these populations they were forced to analyze their data in a pairwise fashion. I think that is the correct approach, but I remain a bit confused about exactly how these comparisons were performed. The authors describe a few different filtering schemes but the methods do not make clear which filtered dataset was used for the single population analyses (Tajima’s D and saltilassi), nor the pairwise LSBL estimates built upon FST estimates. It is a bit difficult to assess those results without that knowledge. I was also confused a bit by how comparison of populations was incorporated into the functional enrichment analyses. Did the authors include all genes within and near candidate windows for each population or did they remove any windows that overlapped with other populations? Since the major focus of this paper is the AL population, it would make sense to polarize results in this manner, but I am not certain if it was.

Second, many estimators of population genomics parameters can be biased by using only variant sites as done by the authors. Specifically, programs like VCFTools will assume any site that is not recorded as a SNP in the vcf in a given window has been confidently scored as invariant, however it is often the case that these windows include a mixture of both called invariant sites as well as sites where there is insufficient coverage to call the nucleotide at that position. In other words, missing data. This can seriously bias measures of nucleotide diversity and divergence and downstream summary statistics that rely on them, like Tajima’s D (see Korunes and Samuk 2021 for a detailed explanation of this issue). Tools that have been developed to account for this potential bias (pixy, angsd, mop) should be used.

Finally, third, the values of pi reported for these fence lizard populations are exceptionally high, particularly for the AL population. Buffalo (2021 eLife, Fig 2 and supplemental data) cataloged nucleotide diversity estimates from published studies across 170 animal phyla and found a maximum pi of 0.083. If accurate, the AL and TN populations would have greater diversity than every species sampled. I strongly suspect biases introduced by the removal of invariant sites is to blame here, however the bias I describe above typically results in reduced estimates of pi, not increased. Nevertheless, these estimates make me deeply concerned about each of the three approaches used by the authors to identify candidate regions under selection as each relies to some degree on the accurate estimation of diversity within and between populations.

Minor comments
Line 24: Remove extra space after “investigate.”
Line 79: Change “induce” to either “induced” or “induces.”
Line 102: Inconsistent use of uppercase/lowercase letter for figure referencing. Seems that most often a lowercase letter is used.
Line 104: I don’t think 13x fits the general definition of “high coverage”. Please consider “moderate coverage” instead
Line 135, 137, etc.: Inconsistent capitalization of “Fig”
Line 137, 141, 155, etc.: Inconsistent use of “Supp fig” or “Supp Figure.”
Line 142: Inconsistent use of “Fig” or “Figure”

Line 186: Change “Fig 2x” to the appropriate figure panel.
Line 187: Inconsistent use of “Supp fig” or “Supplementary Figures.”
Line 188: “Table” should be plural.
Line 188 and 192: Inconsistent use of “Supplementary Table” or “Supp Table.”
Line 198, 217, etc.: Inconsistent capitalization of “supp.”
Line 211: Reference the correct panel in Figure 2.
Line 273: “Fig” should be “Table” or the figure number is incorrect.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In the revised version of this manuscript, the authors have dealt with some of the concerns I previously had. However, for two major comments (detailed below), I do not find the changes made to be particularly compelling.

1. Regarding the lack of replication in the invaded sites, the answer provided by the authors did not really address my concern. Even if there was (likely) only one invasion source for fire ants in the US, there has been subsequent spread. Therefore, more than one fence lizard population with exposure to fire ants could have been investigated, allowing for replication. Even if those fence lizard populations are exposed to a (relatively) genetically homogenous fire ant invader, they are still exposed to a range of other environmental variables. Thus, replicability in this case is important, and would have allowed adaptation to fire ant exposure to be better disentangled from adaptation to other environmental variables.

2. The test for accuracy of missing data imputation indicated that errors are close to 8-9%. While the authors point out that these rates are smaller than those reported in previous studies of non-model organisms, I think the analyses should take these errors into account. For example, for the lead GWAS association for hind limb length, how many of the rare genotypes (G/G) were imputed due to missing data? Could the errors rates reported here have influenced this result?

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have addressed each of my major concerns in a very detailed manner, as well as those of the other reviewer. I have no further comments or concerns to share.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The comment regarding lack of replicability of populations was not addressed. I am not surprised by this, as it would involve substantial additional work. Nonetheless, I maintain that this represents a major limitation of the study, and could represent an important source of false positives for the genome scans (i.e., you are actually recovering the signal of adaptation to another environmental variable, that simply co-varies with fire ant exposure; this scenario is likely, I think, because of the limited number of populations included in the analysis).

The comment regarding the effect of missing data imputation was addressed via simulating a phenotype (and carrying out GWAS on the simulated data). These results indicate that indeed missing data imputation was likely not a strong source of false-positive associations. Nonetheless, I would have still liked to see a mention of how many of the rare genotypes at the lead GWAS locus (there seem to be ~12 G/G individuals in Fig. 3C) actually derive from imputed missing data.

(Remarks on code availability)

We are thankful for the thorough and constructive assessment of our manuscript by the reviewers. Addressing these great comments has helped us improve the manuscript and its ultimate impact on readers. We provide point-by-point responses below to address each concern. Text edits and additions in the manuscript itself are identified with red text.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this manuscript, Assis and colleagues investigate genomic signature of adaptation in a native species (the fence lizard) that has been exposed to an invasive species (the red fire ant) representing both a venomous predator and a toxic prey. The authors survey three fence lizard populations (only one of which was invaded by fire ants; n=20 per population). For this focal set, they relied on whole genome sequencing at high coverage (~13x), which they used to reconstruct patterns of population structure, and to scan the genomes for the signature of natural selection. The authors also obtained low-coverage (~4x) whole genome sequencing data for an additional 361 individuals from the invaded site. This low-coverage dataset was used to understand the genetic architecture of limb length (a candidate adaptive trait in the context of fire ant invasion), via GWAS.

Results provided evidence for 101 genomic regions with a signature of natural selection (as confirmed using at least two of three metrics used), across the three populations. Of these regions, 42 were detected only in the invaded population. These candidate adaptive loci were enriched for several functional categories, such as myofibril, complement and coagulation cascades, or osteoblast signalling. The same patterns of functional enrichment were generally not observed in the two non-invaded populations. Finally, when combining genome scan results with GWAS results, there were two candidate limb length loci (out of the 24 loci revealed by GWAS) that overlapped with genomic regions inferred to have been shaped by natural selection in the invaded site. The authors interpret these results considering the adaptive potential of natural populations that are exposed to abrupt human-mediated change.

I enjoyed reading the manuscript, and I think it tackles a key topic (evolutionary responses of native-range populations to introduced invasive species) that deserves more study. As well, the manuscript is well written. That said, I think there are important limitations, which I detail below.

1. The sampling available for this study is a major limitation, in my view. Because only one invaded site was included, we cannot confidently exclude the possibility that genomic signatures of natural selection detected here are related to some other aspect of the environment that is different between the Alabama (invaded) site and the Arkansas and Tennessee (noninvaded) sites. Indeed, one can assume important differences related to latitude (e.g. climate) are also at play. IA sampling design with some level of replication could have mitigated this confounding factor.

Response: **We agree that the (effectively) single comparison represented in our study is suboptimal. This design, however, reflects the natural history of fire ants in the Southeastern United States, i.e. with a single invasion site and spread from there. As such, our study aims to advance understanding of this important system and evolutionary process without necessarily representing the final word on the topic. In our revised manuscript, we have added this explicit acknowledgment in our Discussion, accordingly:**

"The natural history of the fire ant invasion – with a single invasion in Southern Alabama followed by range expansion from there – precludes us from disentangling biogeographical vs. invasion-related evolutionary effects on the patterns of genetic variation in fence lizards that we observed in the absence of further experimental confirmation. That said, given that latitudinal patterning of fence lizard hind limb length variation reversed from prior to fire ant invasion vs. today (Thawley et al. 2019) and that longer hind limbs are associated with increased survival against swarming ants (Langkilde 2009), the genetic variants detected in our study may plausibly constitute some of the adaptations underlying fence lizard limb length variation."

2. It is not entirely clear why the authors chose those specific three metrics for inferring the action of natural selection. Additional explanation would be helpful, especially since two of the three metrics (saltilassi and LSBL) are not widely used. How exactly do these metrics complement one another, and why are they the best metrics to be used here, given expectations on the genetics of adaptation? Presumably, because adaptation in this context is rapid (and occurring in a population within its ancestral range), soft selective sweeps on standing genetic variation would predominate. Are the metrics used here reliable for detecting such sweeps?

Response: **Thank you for this comment. We agree that more clarification/ justification regarding our choice of methods is needed. One reason is that while research on the effects of human activity on natural ecosystems needs to extend to non-model organism studies, this excludes certain types of genomic analysis. The lack of a fence lizard linkage map, for instance, precludes the use of many well-established analytical tools in evolutionary genomics. Fortunately, new tools to circumvent such limitations continue to be developed. Our revised manuscript now explains our selection of this statistic:**

"The saltilassi statistic computes likelihood ratio tests on the haplotype frequency spectrum of a given population to identify haplotypes with high frequencies compared to the genome-wide expectation. This statistic implements a model that allows for one or more sweeping haplotypes (i.e. hard or soft sweeps), making it ideal to test the hypothesized rapid adaptation of invaded fence lizards. The ad hoc null distribution in saltilassi is generated from the genome-wide average haplotype frequency spectrum, thus not requiring a linkage map for its calculation. This feature makes saltilassi valuable for studies of non-model taxa without this information, as in our case"

Regarding LSBL, the confusion regarding its use is understandable. LSBL is in fact the precursor to the widely used PBS statistic, but in which pairwise F_{ST}'s are not log-transformed, which has advantages, as we now describe:

"LSBL is the precursor of the Population Branch Statistic (PBS), with the difference that F_{ST}*'s are not log-transformed. This feature allowed us to retain sites with* F_{ST}*=0 (i.e., not log-transformable) in one pairwise comparison, which otherwise would exclude SNPs in our dataset that are invariant in two of the three populations."*

3. The use of very low coverage sequencing for the association study (~4x) means that genotyping errors are more likely, unless the authors took steps to confirm that indeed such errors are rare, and their genotype calls are reliable (I did not see mention of this). While missing genotype data is imputed, how do we know that this imputation step did not introduce errors? Are there replicates available (e.g., individuals sequenced repeatedly) that could be used to check the genotyping errors (before and after the imputation step)? In some of the associations detected (e.g., the chromosome 2 locus; Fig. 3B), sampling is highly uneven between the three genotype classes, a scenario that is known to be prone to false positives (and could be further exacerbated by genotyping errors).

Response: **Great suggestion. Indeed, because twenty individuals were sequenced independently in the two sequencing runs (~13x and 4x), this allowed us to robustly assess the concordance rate in allelic calls at 4x coverage (by holding each test individual out of the reference set, in turn) in our revised manuscript. We now present the below text in the Methods ("Genotype imputation" sub-section) with corresponding new Supplementary Figures S28 and S29:**

"Because the 20 reference individuals were also sequenced in our low coverage data generation step, this provided an opportunity to assess imputation accuracy. Specifically, we removed individual i from the MC reference panel (20-1 = 19) to then impute chromosome 6 of individual i based only on that individual's LC data. This process was repeated in turn for all 20 individuals, yielding a mean rate of allele concordance between MC genotyping and LC imputation of 91.3% (Supplementary Figure S28). We also imputed all chromosomes (1-11) for one individual using this approach, resulting in a genome-wide mean concordance rate of 91.7% (Supplementary Figure S29). These rates outperform multiple other studies on non-model organisms and simulations across various methods (Pei et al. 2008; Druet et al. 2014; Tsai et al. 2017; Brian L. Browning et al. 2018). Because these tests derived from smaller reference panels than the one used in our actual analysis (i.e., 19 individuals rather than 20), these accuracy estimates are conservative."

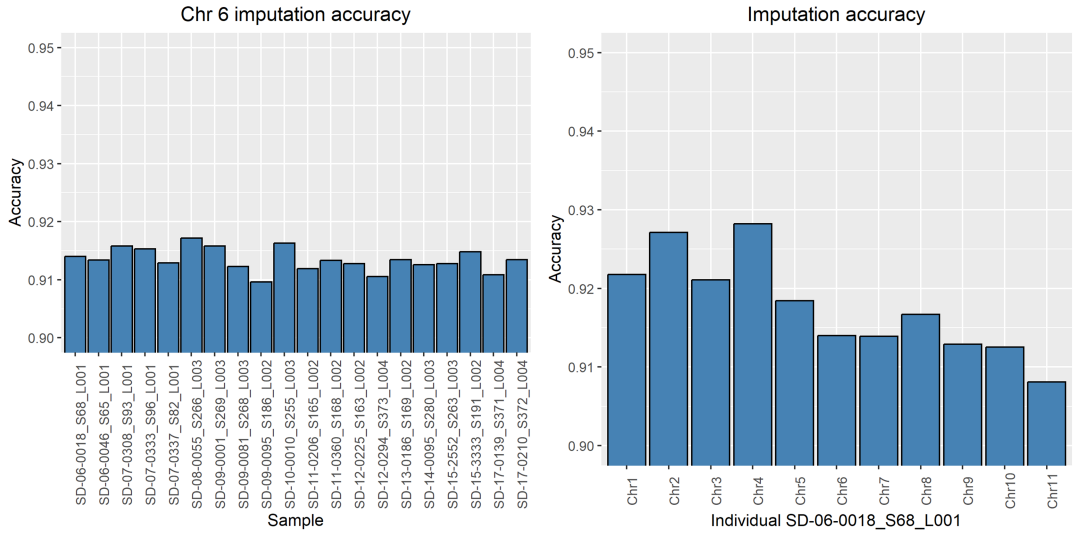

Figure S28: For chromosome 6, and for each of 20 individuals, concordance rates for alleles sequenced at 13X, and imputed alleles based on a reference panel that did not contain that individual.
Figure S29: For chromosomes 1 through 11, concordance rates for alleles of an individual sequenced at 13X and imputed alleles based on a reference panel that did not contain that individual.

4. The discussion focused on gene ontology detracts from the readability in my view, as it is largely speculative, and includes terms (e.g., gene names) that many readers of the paper would not be familiar with. Because no additional functional experiments are conducted to verify these associations (and considering the limited population sampling used in the genome scans and the low coverage used in the GWAS), it is possible that are least some of these results are due to experimental noise. Thus, I recommend that the authors try to condense this section.

Response: **We appreciate this sentiment and comment, and we took an incredibly hard look. Ultimately, we think it is a valuable reporting of the results of our study, even if the enrichment results alone are not (and cannot ever be) definitive, so we prefer to keep this text in the Results.**

Reviewer #2 (Remarks to the Author):

In their manuscript “Genomic signatures of adaptation in native lizards exposed to human-introduced fire ants”, Assis and colleagues describe analyses designed to detect genomic signatures of selection in an Eastern Fence Lizard population subject to a novel selection pressure – the introduction of non-native fire ants 70 years ago. The authors perform a series of population genomic analyses to identify regions where selective sweeps may have occurred in one lizard population impacted by fire ants and two unaffected populations then contrast summaries of these results among populations. They further perform a GWAS on a much larger sample of animals all from the impacted population to map putative genomic regions associated with hindlimb length a trait known to impact ant removal by fence lizards. While I believe this a

very interesting system and the data collected are appropriate to answers the author's questions, I have some major concerns regarding the manner in which these analyses were performed. I detail those concerns as well as more minor comments below.

Major Concerns
My concerns are largely limited to the population genomic inferences made with the 20 sample-per-population, 13x coverage, population genomic dataset. The comments that follow do not apply to the larger sample set used for GWAS analyses (although I am less well-versed in these analyses and reserve more careful scrutiny to other reviewers).

First, I believe the authors initially sought to use the two non-invaded populations as near and more distance outgroups to polarize their analyses of the AL population. Upon discovering that there was not a straightforward relationship among these populations they were forced to analyze their data in a pairwise fashion.

Response: Your inference is correct; we did indeed modify our approach when initial analyses indicated the relationship between these populations was not straightforward. To be more transparent about how our research methodology developed, we now discuss this in our revised Results.

"Taking all these results together, despite the Mississippi River being a putative biogeographic barrier for AR fence lizards, we did not observe a clear, strong genetic separation between AR vs. AL+TN, precluding the definitive assignment of any population as a true outgroup for downstream evolutionary analyses. For this reason, we focused on population-specific and AL vs. (AR and TN) statistics."

I think that is the correct approach, but I remain a bit confused about exactly how these comparisons were performed. The authors describe a few different filtering schemes but the methods do not make clear which filtered dataset was used for the single population analyses (Tajima's D and saltilassi), nor the pairwise LSBL estimates built upon FST estimates. It is a bit difficult to assess those results without that knowledge.

Response: **We apologize for the confusion. We have now included new text that clarifies which dataset was used for each set of analyses. To summarize, we used our 3-population filtering scheme for 3-way demographic analyses: PCA, admixture, and the neighbor-joining tree, lines 138-139.**

"Next, we restricted the set of SNPs to loci that were variable between at least two of the three populations (11,437,455 SNPs)"

For 2-population analyses (FST and consequently LSBL) the same filtering scheme was applied, but only for the two populations being analyzed, lines 134-137.

"For this analysis, we restricted the set of SNPs to loci that were either variable in both populations, or for which the minor allele in one population was fixed in the second. Average F_{ST} values were AL-TN = 0.210 (16,075,146 SNPs), AL-AR = 0.141 (33,225,824 SNPs), and TN-AR = 0.223 (16,383,584 SNPs; Figure 1B)"

For within population analyses (pi, Tajima's D, saltilassi) we used the original SNP dataset that was only filtered for standard quality control measures (lines 494-514, 46.9M SNPs).

I was also confused a bit by how comparison of populations was incorporated into the functional enrichment analyses. Did the authors include all genes within and near candidate windows for each population or did they remove any windows that overlapped with other populations? Since the major focus of this paper is the AL population, it would make sense to polarize results in this manner, but I am not certain if it was.

Response: Yes, the selection signatures observed in the AL population were exclusive to that population. Thanks to this suggestion, we now emphasize in lines 206-207:

"None of these candidate regions overlapped between populations."

Second, many estimators of population genomics parameters can be biased by using only variant sites as done by the authors. Specifically, programs like VCFTools will assume any site that is not recorded as a SNP in the vcf in a given window has been confidently scored as invariant, however it is often the case that these windows include a mixture of both called invariant sites as well as sites where there is insufficient coverage to call the nucleotide at that position. In other words, missing data. This can seriously bias measures of nucleotide diversity and divergence and downstream summary statistics that rely on them, like Tajima's D (see Korunes and Samuk 2021 for a detailed explanation of this issue). Tools that have been developed to account for this potential bias (pixy, angsd, mop) should be used.

Response: **This is an important point. Because this concern is raised again in the next point by the reviewer, we refer to our response below.**

Finally, third, the values of pi reported for these fence lizard populations are exceptionally high, particularly for the AL population. Buffalo (2021 eLife, Fig 2 and supplemental data) cataloged nucleotide diversity estimates from published studies across 170 animal phyla and found a maximum pi of 0.083. If accurate, the AL and TN populations would have greater diversity than every species sampled. I strongly suspect biases introduced by the removal of invariant sites is to blame here, however the bias I describe above typically results in reduced estimates of pi, not increased. Nevertheless, these estimates make me deeply concerned about each of the three approaches used by the authors to identify candidate regions under selection as each relies to some degree on the accurate estimation of diversity within and between populations.

Response: **We appreciate that you brought this issue to our attention. Upon investigation, we discovered that the high values we reported were products of our mistaken choice to report population means on a per-SNP rather than a window basis. The studies used by Buffalo (2021) report π values calculated in 100Kb, 500Kb, or 1Mb windows, which would include invariant as well as missing loci in the per-window**

calculation. We now calculate mean π in 100Kb windows, and obtain population means of π = 0.323%, 0.241%, and 0.196% (0.0032, 0.0024, and 0.0020) for AL, TN, and AR, respectively, values that are well in line with observations from other taxa:

"The AL population was the most genetically diverse (AL genome-wide mean pairwise nucleotide distance in 100Kb windows, π = 0.323&; TN π = 0.241%; AR π = 0.196%)"

Next, the reviewer expressed concern for the reliability of our downstream analysis in the face of inflated nucleotide diversity (now corrected, as per the above; we also note that π is only involved in the Tajima's D statistic, which was already calculated in 100Kb windows and therefore was not affected by the otherwise inflated π values originally reported in the manuscript).

Lastly, we address whether expanded consideration of invariant sites in our analyses is necessary. We submit that the results and conclusions from our specific analyses of the three selection statistics used in our study (LSBL, saltilassi, Tajima's D) would not be meaningfully affected by a different approach.

For LSBL, all invariant sites would be LSBL = 0. Though there could theoretically be false-negative high-LSBL sites, this is unlikely, and in any case such an occurrence would not be disproportionately biased with respect to any region of the genome.

For saltilassi, the algorithm uses the truncated haplotype frequency spectrum, limiting the analysis to the K most frequent observed haplotypes (in this analysis, K = 20). Results would be influenced substantially only if large amounts of missing data occurred (*i.e.*, for more than half of the individuals in our analysis). We show in the figures below that, throughout the three regions reported in figure 2, the number of haplotypes scored for a given window was never less than 20, which is the maximum number of haplotypes considered by the analysis (K = 20).

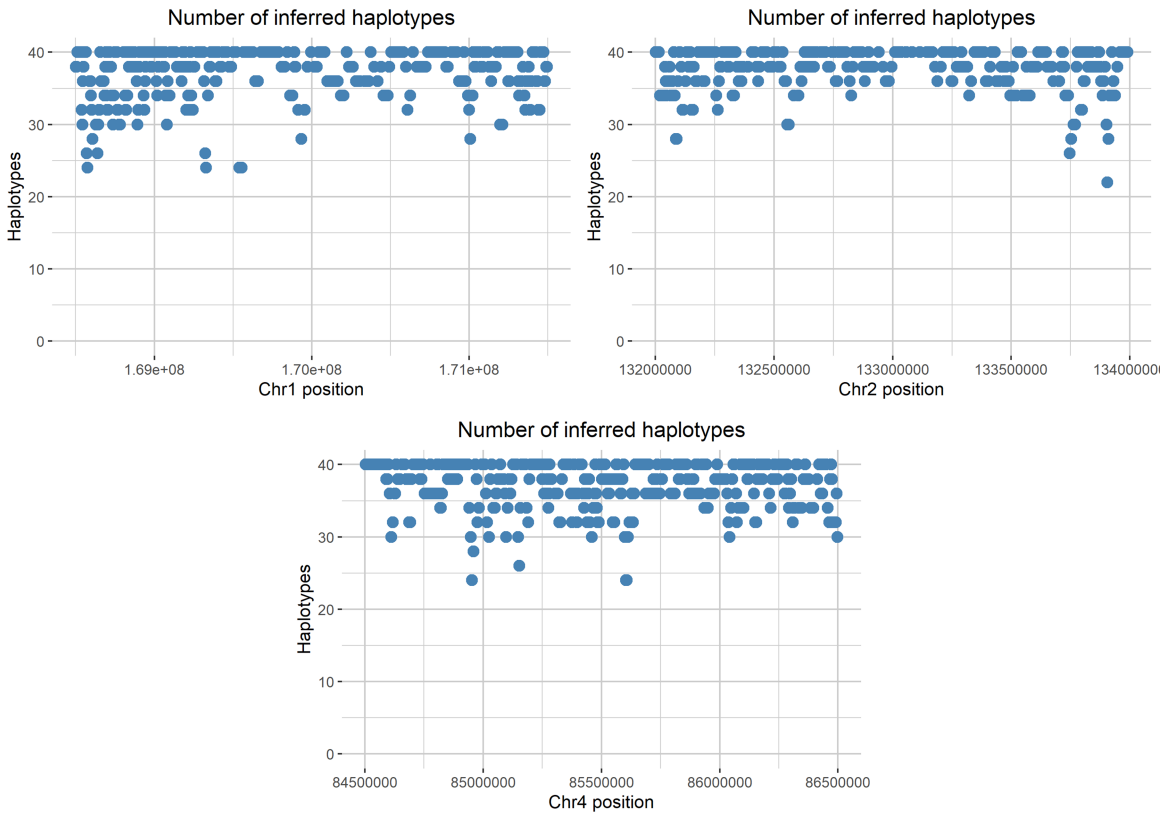

Figure 1: Number of haplotypes inferred for three candidate regions under positive selection according to saltilassi statistic for Alabama fence lizards. Mean haplotype length is 19Kb for chromosome 1, 14Kb for chromosome 2, and 10Kb for chromosome 4.Throughout all regions, more haplotypes were identified than the number considered by the analysis (K = 20).

For Tajima's D, it is true that missingness of data could influence the window-based statistic. However, any systematic missingness would occur consistently throughout the genome and not be biased to a particular region, and our approach for identifying candidate regions was not based on an absolute value cutoff for the statistic, but on relative values for the genome-wide distribution of each population (top 0.5th percentile).

Minor comments

Line 24: Remove extra space after "investigate."

Line 79: Change "induce" to either "induced" or "induces."

Line 102: Inconsistent use of uppercase/lowercase letter for figure referencing. Seems that most often a lowercase letter is used.

Line 104: I don't think 13x fits the general definition of "high coverage". Please consider "moderate coverage" instead

Line 135, 137, etc.: Inconsistent capitalization of "Fig"

Line 137, 141, 155, etc.: Inconsistent use of "Supp fig" or "Supp Figure."

Line 142: Inconsistent use of "Fig" or "Figure"

Line 186: Change "Fig 2x" to the appropriate figure panel.

Line 187: Inconsistent use of “Supp fig” or “Supplementary Figures.”
Line 188: “Table” should be plural.
Line 188 and 192: Inconsistent use of “Supplementary Table” or “Supp Table.”
Line 198, 217, etc.: Inconsistent capitalization of “supp.”
Line 211: Reference the correct panel in Figure 2.
Line 273: “Fig” should be “Table” or the figure number is incorrect.

Response: Thank you for identifying and reporting each of these mistakes. We have fixed every one of them in the revised manuscript.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In the revised version of this manuscript, the authors have dealt with some of the concerns I previously had. However, for two major comments (detailed below), I do not find the changes made to be particularly compelling.

1. Regarding the lack of replication in the invaded sites, the answer provided by the authors did not really address my concern. Even if there was (likely) only one invasion source for fire ants in the US, there has been subsequent spread. Therefore, more than one fence lizard population with exposure to fire ants could have been investigated, allowing for replication. Even if those fence lizard populations are exposed to a (relatively) genetically homogenous fire ant invader, they are still exposed to a range of other environmental variables. Thus, replicability in this case is important, and would have allowed adaptation to fire ant exposure to be better disentangled from adaptation to other environmental variables.

***Response*: We appreciate your point. While we agree that increased sampling may have provided new insights and that it would be better than not having increased sampling (even if one could not truly treat the multiple invaded sites as independent, as per our previous response), such new sampling efforts are beyond our limits for this present study.**

2. The test for accuracy of missing data imputation indicated that errors are close to 8-9%. While the authors point out that these rates are smaller than those reported in previous studies of non-model organisms, I think the analyses should take these errors into account. For example, for the lead GWAS association for hind limb length, how many of the rare genotypes (G/G) were imputed due to missing data? Could the errors rates reported here have influenced this result?

***Response*: Thank you for this continuing discussion. In response, we have now taken several steps to model the potential impact of imputation error in our study, using the particular association result reported in Figure 3 as a case study. Specifically, we performed a series of simulations to estimate the potential effects of our imputation error rate. The text and figures below have now been added to the Methods and to the Appendix I (figures S30-S33). In summary, our simulation results suggested that false positive associations with such properties would be extremely unlikely, whereas we would expect to observe an increased proportion of false-negative (and otherwise, reduced significance) results, due to the imputation error rate. This is conservative with respect to the results we present in our paper.**

Because we verified that genotype imputation of the low coverage sequences resulted in a small rate of genotyping error (see Genotype imputation) we assessed the potential impact of this error rate on the results reported for the GWAS. Following a series of simulations, we did

not observe any false-positive result on the same level of significance as that reported for locus Chr2: 129,424,718 (Figure 3), although false-negative results may have occurred. More details can be found in Appendix I, figures S30-S33.

All simulations involved a randomly generated phenotype: a normally distributed variable with the same n, mean, and variance as relative limb length in our study. Additionally, for all simulations, we applied a genotyping error rate similar to that observed in the 100Kb window surrounding the Chr2: 129,424,718 locus, 7.29% (27 genotypes out of 381 in our sample).

In our first set of simulations, we quantify the probability that this error rate would produce a false positive association with the phenotype in our study. We generated a population genotype distribution equivalent to the locus in question (minor allele frequency, MAF = 0.1). For each iteration, a new genotype distribution with MAF = 0.1 was generated, and 7% of the genotypes were randomly drawn and switched to another genotype. Finally, we tested the association between this new genotype distribution (with 7% error introduced) on our simulated phenotype. This process was repeated 4,500,000 times (compared to 4,245,544 SNPs tested in our study).

There was no occurrence of an association near our significance threshold ($p < 5e-8$). In 4,500,000 simulations, $p < 1e-5$ occurred 40 times, and $p < 1e-6$ occurred only once (6.8e-7).

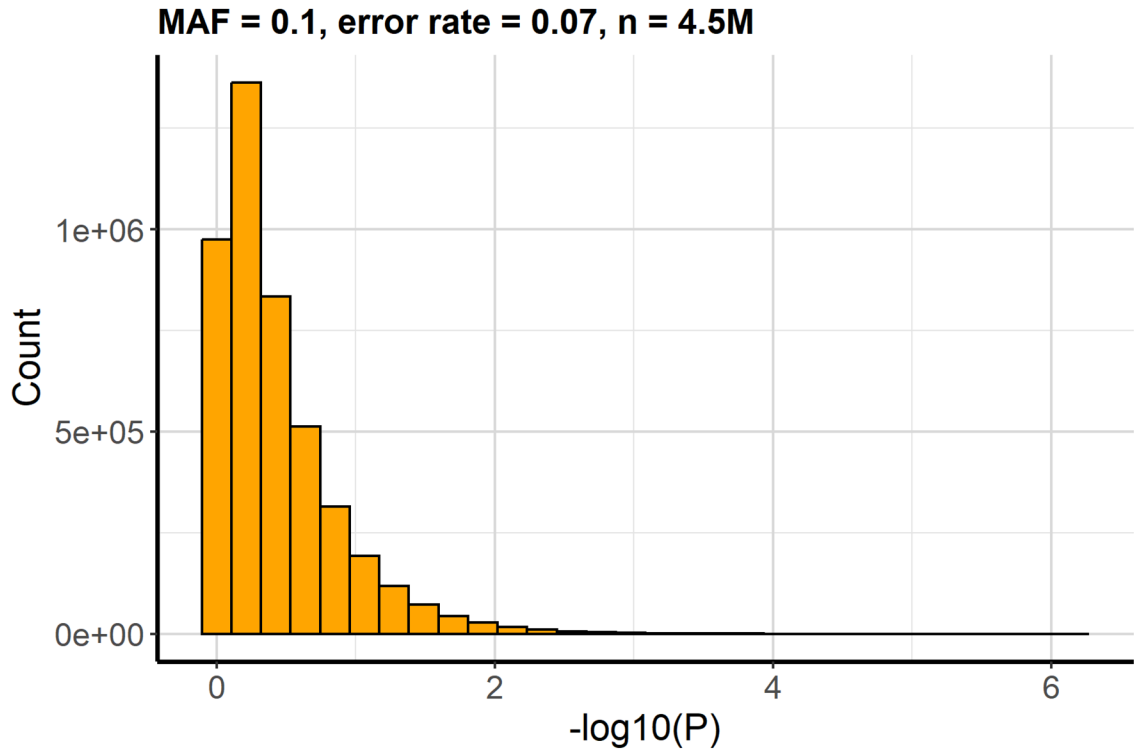

Figure S30: For each of 4,500,000 simulations, a population with n=381 and randomly generated genotypes (minor allele frequency = 0.1) had 7% of genotypes randomly sampled and changed to another genotype. The association between each genotype distribution and a randomly generated phenotype with the same mean and variance as relative limb length in this study did not produce a significance level below the threshold for this study (< 5e-8).

Next, we repeated this simulation but this time starting with different genotype distributions: MAF = 0.05, 0.1, 0.15, 0.2, and 0.25 (for each, n = 100,000). The smallest p-value observed across any of these simulations was p = 3.8e-6.

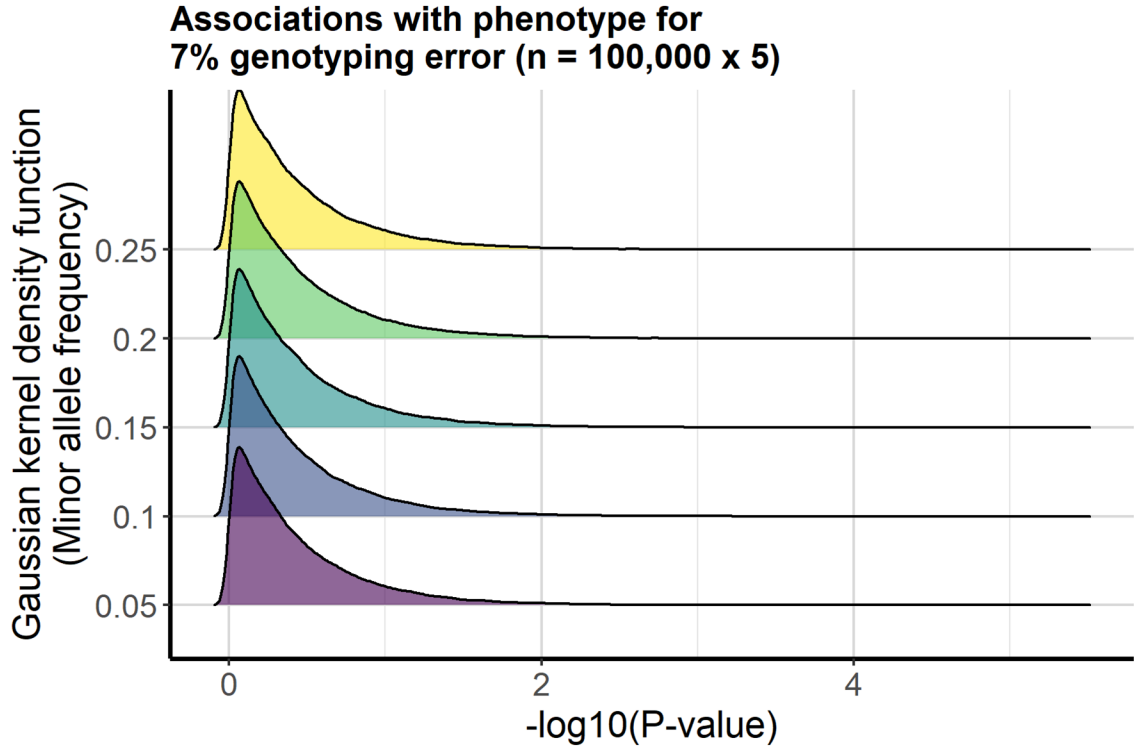

Figure S31: Density plots with distribution of p-values for the association between randomly generated genotype distributions of varying minor allele frequencies and the randomly generated phenotype (population size of n=381). For each of 100,000 iterations per MAF, 7% of genotypes were randomly sampled and changed to another genotype.

In summary, these simulations did not produce a false-positive result at the same significance as the genome-wide association study results reported in our study.

We then assessed the extent to which this imputation error rate may have produced false negative results. Here, we randomly generated a genotype distribution with the same allele frequency as the locus in question (MAF = 0.1), assigned to the alternate allele an effect size on the randomly generated phenotype equivalent to the locus in question (ß = 0.1), and then repeated this simulation 100,000 times. As expected, this allele distribution and effect size combination produced numerous significant association results among the permutations.

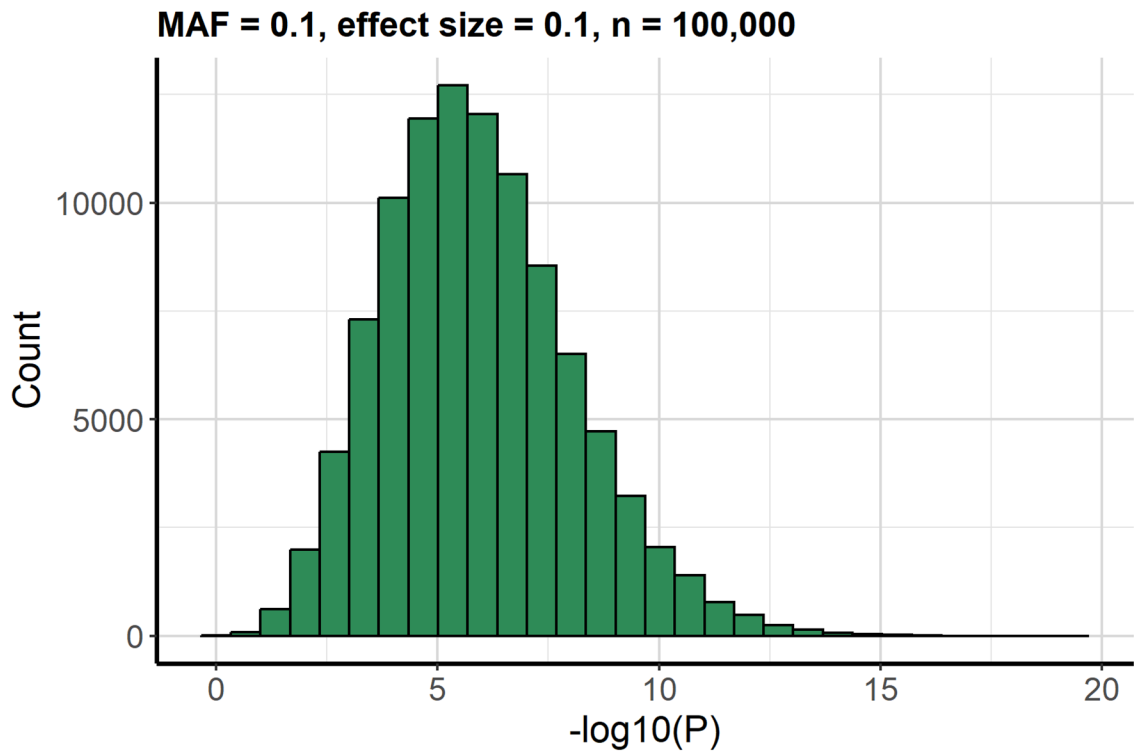

Figure S32: For each of 100,000 iterations, a population of size 381 had a randomly generated genotype distribution with a minor allele frequency of 0.1 and for which one allele had an assigned effect size of 0.1. With the assigned effect size, numerous significant associations (P < 5e-8) with the randomly generated phenotype occurred.

Then, we repeated this process with multiple genotype distributions and effect sizes, but now introducing 7% random error. While many significant associations do still occur among these permutations, there is an understandable shift in the p-value distribution, representing an increased proportion of false-negative associations.

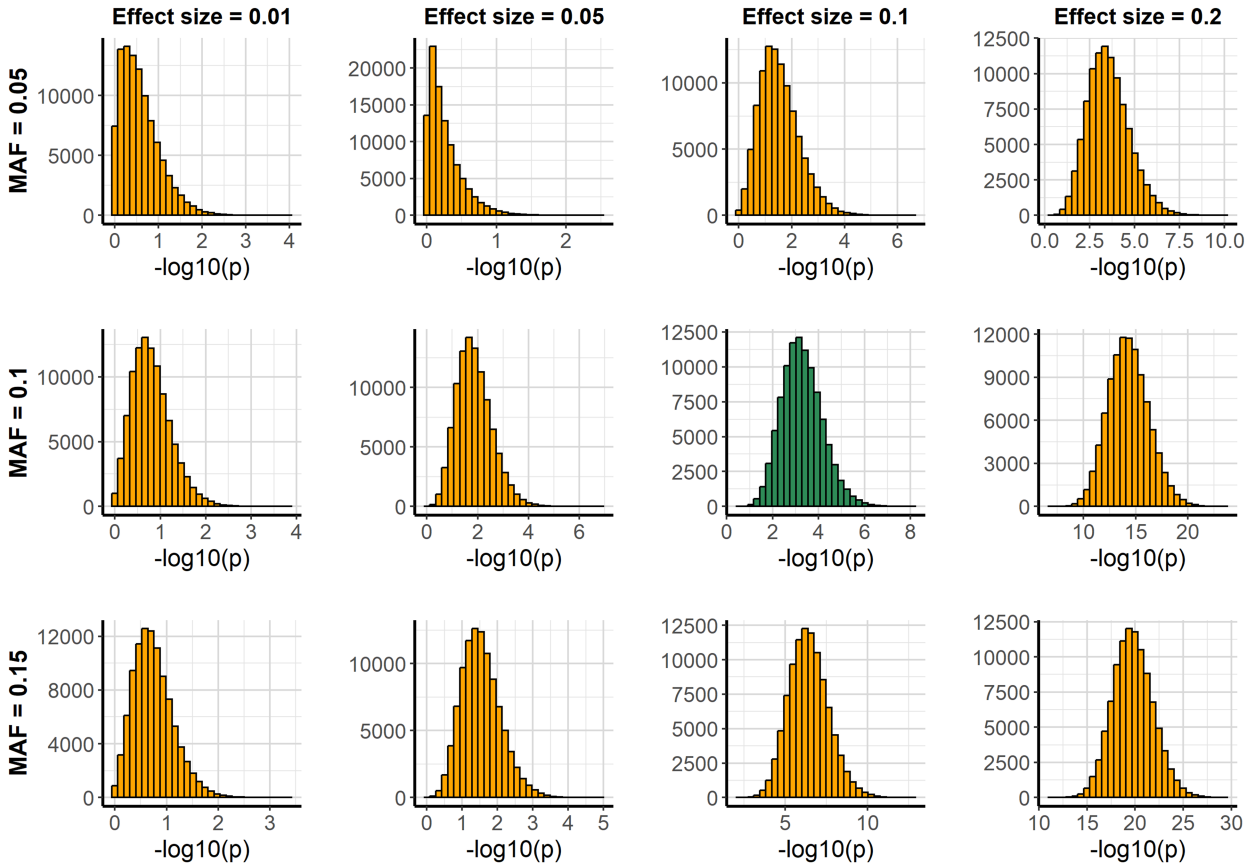

Figure S33: With a population of size n=381, randomly generated genotype distributions with varying allele frequencies (MAF of 0.1 to 0.15) and allele effect sizes (0.01 to 0.2) had 7% of genotypes randomly sampled and assigned to a different genotype for 100,000 iterations. With the assigned effect sizes, significant associations with he randomly generated phenotype occurred despite the introduction of simulated errors.

Reviewer #2 (Remarks to the Author):

The authors have addressed each of my major concerns in a very detailed manner, as well as those of the other reviewer. I have no further comments or concerns to share.

***Response*: We greatly appreciate your positive feedback.**

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The comment regarding lack of replicability of populations was not addressed. I am not surprised by this, as it would involve substantial additional work. Nonetheless, I maintain that this represents a major limitation of the study, and could represent an important source of false positives for the genome scans (i.e., you are actually recovering the signal of adaptation to another environmental variable, that simply co-varies with fire ant exposure; this scenario is likely, I think, because of the limited number of populations included in the analysis).

***Response*: Thanks to this discussion and comments by the editor, we now acknowledge this and other potential limitations of our sampling design in the abstract, introduction, and discussion of this revision.**

The comment regarding the effect of missing data imputation was addressed via simulating a phenotype (and carrying out GWAS on the simulated data). These results indicate that indeed missing data imputation was likely not a strong source of false-positive associations. Nonetheless, I would have still liked to see a mention of how many of the rare genotypes at the lead GWAS locus (there seem to be ~12 G/G individuals in Fig. 3C) actually derive from imputed missing data.

***Response*: We appreciate this discussion, as it motivated us to explore in multiple ways the accuracy of our imputation procedure. For the specific locus in question, both the common and the rare alleles (A and G) in this population were imputed based on linkage disequilibrium with genotyped loci. For this reason, the potential bias for inaccurate imputation of rare alleles was not applicable to this locus, and we apologize for overseeing this question in the previous revision. While imputation is not a perfect tool, it remains a cost-effective method for achieving greater statistical power with satisfactory results in the field of evolutionary ecology, as indicated by our simulations.**