Review of manuscript TPB-D-21-00048

Summary

In this manuscript, the authors apply population genetic theory to demonstrate the relationship between genetic differentiation, migration rates, and selection and dominance coefficients at migration—selection equilibrium. The authors use a deterministic one–locus model with two demes of infinite size, and so ignore genetic drift. The selection coefficients can be expressed in terms of the remaining model parameters and estimates of allele frequencies at equilibrium. Hence, the authors argue that their expressions could be used to estimate the strength of selection in natural populations. The authors also explore the relationship between selection coefficients and F_{ST} at equilibrium. Because F_{ST} is a ratio of allele–frequency variances, there is no unique path of inference from F_{ST} to selection coefficients. As the authors argue, empirical estimates of F_{ST} could be used to identify combinations of selection coefficients compatible with the observed data. The approach proposed here requires knowledge of quantities that are potentially difficult to estimate, including migration rates and dominance coefficients. The authors argue that migration rates are readily estimated in the genomics era and that uncertainty about dominance has minor effects as long as dominance is weak. In the Discussion the authors propose that dominance and selection coefficients could be estimated from within–generation changes in allele frequencies. In passing, the authors correct a mistake in the theory in a previous study by Hoekstra et al. (2004).

I strongly agree that our understanding of evolution, and local adaptation and speciation in particular, hinges on accurate estimates of the strength and mode of selection. Hence I think the paper pursues an important purpose. The approach proposed here for inferring selection from allele frequencies and population differentiation is straight–forward and obvious in principle. However, there are good reasons for why there is only a limited number of published applications of this and related approaches (Hoekstra et al. 2004 being one example). The manuscript currently ignores two important sampling effects that will likely limit their approach in practice: the estimation of allele frequencies and genetic drift. Both effects are well known to contribute substantial uncertainty to the estimation of population genetic parameters. The theory developed in this manuscript on it own is of limited novelty to justify its publication as a stand–alone TPB paper. In

combination with simulations and power analyses addressing the effects of allele frequency estimation and genetic drift, however, I could see this manuscript as a good fit to TBP. I am also concerned about the parameterisation chosen by the authors, which makes the model asymmetric and difficult to interpret. Last, I was not able to reproduce the allele frequency dynamics shown in Fig. 2 based on my re-implementation of the equations provided by the authors. This may be an easy point to fix.

Major issues

1. The main objective of this paper is to motivate an approach to estimate selection coefficients from population genetic data from natural populations. This approach is based on strong assumptions, most importantly the absence of genetic drift. It is well established that both allele frequencies and F_{ST} are very sensitive to genetic drift. The authors cite Jewett et al. (2016) as if that publication showed genetic drift has a minor effect on the estimation of selection in general. However, the focus of Jewett et al. (2016) was on estimation from allele-frequency *trajectories*, not from a single snapshot of allele frequencies at supposed equilibrium. The authors should in my opinion perform simulations that include genetic drift and assess the performance of their approach for various intensities of genetic drift.

2. The approach for estimating selection coefficients presented here also assumes that either allele frequencies or F_{ST} can be estimated from data. The authors state that uncertainty in allele frequency estimates could be factored into the uncertainty about estimated selection coefficients. While I agree, I think the authors should illustrate their idea by reporting how errors in allele frequency estimates would translate to uncertainty in selection coefficients for a set of combinations of deme sizes.

3. I re-implemented and double-checked all equations shown in the manuscript main text and found no substantial errors (see Minor comments below for a formal comment on Eq. 7). However, the allele-frequency dynamics shown in Fig. 2 seems to be wrong — or at least not correspond to the parameter values given in the figure caption. I attached a Mathematica notebook with my calculations and my version of Fig. 2. It seems as if allele frequencies in Fig. 2 do not approach the expected equilibrium ($\hat{p}_1 \approx 0.707, \hat{p}_2 \approx 0.681$) and that the approach to equilibrium is too fast. My suspicion is that selection coefficients used in the simulations shown in Fig. 2 might be higher in absolute value than reported in the caption.

4. As stated by the authors, a primary biological context in which an inference approach under migration–selection equilibrium could be successful is local adaptation. However, in this context, the parameterisation of fitness chosen by the authors seems suboptimal to me for two reasons. First, the parameterisation is asymmetrical with respect to the selective benefit of the locally favoured alleles. If alleles P and Q are locally beneficial in demes 1 and 2, respectively, the local fitness advantage of P in deme 1 is limited, whereas its fitness cost in deme 2 is theoretically unbounded. As a corollary, the local advantage of Q in deme 2 may be infinitely large, whereas its cost in deme 1 is bounded.

   Second, dominance under the current parameterisation refers to the effect of allele Q irrespectively of whether allele Q resides in deme 1 or 2. In the case of dominance, i.e. $0.5 < h <= 1$, the locally beneficial effect of allele P is (partially) shielded in heterozygotes in deme 1, while its selective disadvantage is overemphasised in heterozygotes in deme 2. In contrast, allele Q experiences a relatively larger benefit in deme 2 and a reduced disadvantage in deme 1 in the heterozygote state.

   Both of these asymmetries are unfortunate and render the interpretation of the results difficult. In my Mathematica notebook I suggested a parameterisation that is symmetrical with respect to the selective advantage and dominance effect of the respective locally beneficial alleles. My analytical results suggest that formulae do not become substantially more complicated under this alternative parameterisation.

Minor comments

Generic comments

- The parentheses in in-text citations are currently not correctly typeset.
- The notation used for the alleles seemed unconventional and confusing to me. I propose to call the two alleles of interest A_1 and A_2 instead of P and Q. My motivation for proposing this change also comes from the fact that P and Q (and R) are commonly used to denote the frequencies of genotypes A_1A_1 and A_1A_2 (and A_2A_2), respectively.

Abstract

- Lines 6–7: Move "is" to after "alleles".

Introduction

Page 2:

- Lines 4—7: Split this sentence; it is too long.
- Line 9: Replace en-dash by em-dash in "Wright–Fisher".
- Line 17: "Scaled selection coefficients are defined as the products" –> "The scaled selection coefficient is defined as the product".
- Line 18: "their effects" –> "the effects of these two quantities".
- Line 49: Remove spaces before and after the em-dash in "0.44 — 0.77".

Page 3:

- Line 12: Insert "divergent" after "strength of".
- Lines 25—26: The meaning of this sentence remained unclear to me.

Theory and Methods

Page 4:

- Caption to Fig. 1:
  - The statement about niches being contiguous or overlapping is potentially confusing. I suggest to clarify that even in the case of spatially contiguous or overlapping niches, random mating must be confined to discrete, non-overlapping demes. Otherwise, the assumptions of the model are violated.
  - The second sentence starts with parameter symbols, which I suggest to avoid.

Page 5:

- Line 16: "Setting this equal to p_2' as given by equation 1b we obtain:" –> "Setting equations 4a and 4b equal to equations 1a and 1b, and solving for s_1 and s_2, respectively,"
- Lines 19—22: I suggest to rearrange and rewrite this part to "Equation 5 shows how [...] equilibrium values of p_1 and p_2. If m_{12}, m_{21} and h are known, uncertainty [...] binomially distributed.".
- Eq. 7: There are "+" signs in the subscripts of q_2 and q_1 which appear to be typos.

Page 6:

- Line 6: Rephrase to "Equation 7 only provides implicit expressions for...".
- Line 14: The checks showing an agreement between simulations and numerical solution of Equation 7 should be visualised in the Supplementary Information, in my opinion.
- Caption to Fig. 2:
  - "...process for the case of $h = 1$, $m = 0.05$..." -> "...process according to equations 1 and 2 for $h = 1$, $m_{12} = m_{21} = m = 0.05$...".
  - Insert a comma after "In this case".
- Line 21: I suggest to not start a sentence with "F_{ST}".
- Lines 21–23: I suggest to change the tone so it becomes clear that the alternative interpretation of F_{ST} described here has been known for a long time. I also suggest to add an appropriate reference.
- Line 26: I suggest to amend the subsection header to not contain a formula.
- Line 27: "...how equilibrium values of F_{ST} are related..." -> "...how equilibrium genetic differentiation, measured by F_{ST}, is related...". Then, in lines 29–30, shorten "genetic differentiation, measured by F_{ST}," to "F_{ST}".
- Line 30: I suggest to avoid "erode" in the context of an equilibrium.
- Lines 30–31: The clause "when selection is maximal, total selection [...] in niche 2 (Fig. 3)." should be rephrased. It is unclear what "maximal" selection is under the current fitness parameterisation in deme 2, and "total" selection is ambiguous. "Maximum selection" is used also in the sentence following the one of concern here.

Page 7:

- Lines 2–4: I suggest not to express m in percentages, but as decimal fractions.
- Line 4: I suggest to add a note here to explain the effect of dominance (recessive locally deleterious alleles being maintained as they are partially shielded).
- Fig. 3:
  - The quality of the figure was not great in the version of the manuscript that I reviewed. I suggest to check the quality again. As mentioned above, I suggest to not express m as a percentage.
- Caption of Fig. 3:
  - First sentence: Insert "at migration–selection equilibrium" after "F_{ST}".
  - Second sentence: "when selection" -> "if selection". See also my previous comments about "total" selection and the asymmetric parameterisation of fitness, which shows it ugly head here.
  - Third sentence: I suggest to rephrase this to a full sentence.

- Lines 5—7: I suggest to split this sentence, as it is fairly long.
- Lines 9—11: Under the current parameterisation, selection is at its strongest in the limit of $s_2 \to \infty$, not if $s_2 = 1.00$ as is currently stated. This sentence should be corrected if the current parameterisation is retained.
- Line 11: "If selection is too low" -> "If selection is too weak relative to migration". Here, I missed a reference to existing theory formalising this point.
- Line 16: "that is" -> "the latter is".

Page 8:

- Caption of Fig. 4:
  - First sentence: Insert "at migration—selection equilibrium" after "F_{ST}". Change "...for three rates of migration between niches m,..." -> "...for three migration rates m,...".
- Line 1: Insert "equilibrium" after "If". "...known, Fig. 4 can be used to..." -> "...known, the relationship between F_{ST} and s_1 and s_2 in Fig. 4 can be used to...".
- Lines 1—3: "..., possible values of s_1 and s_2 for a given value of F_{ST} lie on the relevant contour..." -> "..., values of s_1 and s_2 compatible with a given value of F_{ST} lie on the respective contour..."
- Lines 3—4: "...that migration rate = 1 % and..." -> "...of $m = 0.01$ and...".
- Line 4: "...possible determining values..." -> "...combinations of compatible values...".
- Line 6: "...if they are not equal then the value of one or the other of them could be as low..." -> "...if $s_2 \neq |s_1|$ then s_2 or $|s_1|$ can be as low...".
- Caption of Fig. 5: "...for the case that migration rate = 1%..." -> "...for the case of a migration rate of $m = 0.01$". "These are the contours for..." -> "Contour plots shown correspond to...".

Discussion

Page 9:

- Line 10: "...in Figs. 3 — 5 as..." -> "...in Figures 3 to 5 as...".
- Lines 12—13: I suggest not to start a sentence with a symbol.
- Line 20: "migrants" -> "emigrants".
- Line 26: Insert a comma after "Furthermore".
- Line 15: I could not make sense of "simultaneous" in this context.