Manuscript Details	
Manuscript ID:	JEB-2016-00117
Manuscript Type:	Research Papers
Keywords:	Theory, Ecological genetics, Adaptation, Natural selection
Date Submitted:	(blinded)
Manuscript Title:	Natural selection and the probability of parallel genetic evolution
Editor:	Gardner, Andy
Editor-in-Chief:	Ritchie, Mike

JEB aims to publish very good papers of broad interest to general evolutionary biologists. Papers that are of narrow interest, or are not original in scope should be rejected or referred to other journals such as Ecology & Evolution (which is partly supported by ESEB). Of course, papers that are scientifically flawed should be rejected, with advice for the authors. In your opinion, is this paper:

	Now	After appropriate revision
Well Above Average		
Above Average		
Average	✓	✓
Below Average		

Anonymity		
	Yes	No
req Do you wish to remain anonymous? If no, please sign your report for the authors.		✓

req Recommendation	
	Accept
✓	Revise
	Resubmit
	Reject

Comments

Confidential Comments to the Associate Editor

Dear Andy,

Thanks for the opportunity to review this article. I think this article will be fine for JEB after revision. It opens up a perspective for further work. Some of the results seem trivial, but it seems fair to publish them, as the links that are made here seem original and of broad enough interest.

Best regards,
Simon

Comments to the Author

Summary

The authors propose a model of parallel genetic evolution in response to environmental differences between the habitats of a common ancestral and multiple descendent populations. A single quantitative trait subject to directional selection is assumed, whereby the fitness of an individual is a linear function of the trait value; low trait values are favoured in the ancestral habitat, whereas high trait values are adaptive in the novel environment. The trait is determined by additive contributions from the underlying loci. A simple and straightforward expression is derived for the probability of fixation of the same derived allele in an arbitrary number of descendent populations from an initially low frequency in the ancestral population. The authors can easily extend this expression to multiple loci thanks to the assumption of weak selection and quasi-linkage equilibrium. As expected, the results show that parallel genetic evolution at a single locus is facilitated by strong phenotypic selection, a large effect on the trait, and a high initial frequency of the adaptive allele. With multiple loci, the probability of complete parallel evolution decreases, unless selection is very strong. The distribution of allelic effect sizes seems to have a minor effect unless selection is very strong. Although the authors mention the distribution of allelic effects as a factor of major interest, the discussion of their findings with respect to this could be more extensive.

Through their analytical work, the authors show that, for given initial allele frequency and allelic effect size, the probability of parallel evolution depends on a compound parameter defined by the product of the ratio of the slope an the intercept of the linear fitness function, times the effective population size. This parameter makes intuitive sense and serves as a means of quantifying the relative importance of adaptive forces underlying parallel genetic evolution, as opposed to genetic drift.

The authors consider two different schemes that are used in practice to identify candidate genes undergoing parallel evolution. It is suggested that the so-called "QTL method", where QTLs underlying divergence from the ancestral population are identified for each descendent population separately, is preferable. This is explained by the fact that loci where the beneficial allele is fixed only in a subset of descendent populations are also informative, but may often not be detected by the other, so-called "candidate gene" method.

This article is relevant as it addresses the inherent difficulty of distinguishing between parallel adaptive evolution and coincidental establishment of the same alleles in multiple populations. This challenge must be addressed in view of emerging multilocus data sets from multiple derived populations potentially undergoing adaptation to similar novel environments. The article is therefore of considerable interest to the community.

A second important link is drawn between the probability of parallel genetic evolution and the genetic architecture of the trait subject to selection. While the article makes the strong and potentially limiting assumption of quasi-linkage equilibrium, it does show that complete parallel genetic evolution at all affected loci is unlikely unless phenotypic evolution is very strong. Ultimately, a more comprehensive study of the importance of the genetic architecture is needed, but I think the current paper presents a good first step.

The authors devise a Bayesian inference procedure making use of a standard Metropolis-Hastings Markov chain Monte Carlo procedure to infer the compound parameter eta given fixed initial frequencies and allelic effects. This inference procedure is evaluated against Wright-Fisher type and individual-based simulations. The latter are used to test the sensitivity of the approach to several assumptions, including the ones of linear selection, known initial allele frequency and allelic effects, identical selection regimes, high recombination, and no gene flow from the ancestral population.

Overall, the article is well written, clean. The formula used for the fixation probability in equation (4) does not look familiar to me (see comment 1 below). If this is indeed wrong, then all the analyses need to be re-run and analytical results adjusted.

Many important results are only represented in tabular form (Tables S1 to S6), which made it difficult for me to follow the sometimes weak trends based on which the authors seem to base their conclusions. Better visualisation of these results would be appreciated and make the paper stronger. I suggest moving what is currently Fig. 5 to the supporting information instead, if figure space is limiting.

I have outlined some concerns below.

Comments
---

1) The expression used for the fixation probability in equation (4) does not look familiar to me. According to Kimura (1957), it should read Pfix = (1 - exp(-4Nsp_0))/(1 - exp(-4Ns)) for diploid organisms, where s is the advantage of a heterozygous carrier. As you seem to deal with haploid individuals, you should use Pfix = ((1 - exp(-2Nsp_0))/(1 - exp(-2Ns))), where s is the selective advantage of allele A over a. The factor of 2 is dropped because the variance in allele frequency is x(1-x)/N in a haploid population, rather than x(1-x)/(2N), where x is the allele frequency.

2) I suggest investigating the role of gene flow among descendent populations (not only from ancestral to descendent ones), as this could greatly inflate your estimates of parallel evolution.

3) Given that the distribution of allelic effect sizes is a focal key factor (l. 66) of the current analysis, I'd have expected a more comprehensive analysis of its effect and a more detailed discussion. Figure 3 shows that the shape and rate parameters of the gamma distribution do not have a strong influence on the probability of parallel genetic evolution unless selection (eta) is substantial and the initial allele frequency (p_0) about 10% or more. The qualitative differences between the three distributions evident in the bottom right panel need to be addressed. What feature of the distribution drives this pattern? Is there an effect of kurtosis, or the relative abundance of small vs. large-effect alleles?

4) Important results apparently supporting the robustness of the approach are currently presented in multiple supporting tables and it is hard to figure out trends and differences. It would be nice if the authors could work these out a bit more verbally, or even come up with a way of presenting them graphically. In particular, the meaning of the intercept (and strong differences between various settings) are not motivated and can lead to confusion. I also found it difficult to convince myself that the "GG" method outperforms the "GC" method.

5) I think it would be good to better explore the effect of the mutation rate. Currently, it is fixed to a single value for all simulations. But recurrent mutation (similar to gene flow among descendent populations; cf. comment 1) could largely inflate your estimates of parallel evolution through shared ancestry.

6) The link to previous works by Orr (2005) and Chevin et al. (2010) is made in the Introduction. It would be nice if the authors could come back to these articles in the Discussion and summarise what we learn from this novel analysis. I also missed a mentioning of Ralph and Coop (2015; PLoS Genet) who studied convergent evolution in a spatially more complex setting, allowing for gene flow as well as explicitly incorporating the effect of recurrent mutation.

7) For highly quantitative traits, a substantial level of adaptation may be reached even when not every single underlying locus is fixed for the favoured allele. Yet, the authors apply a strict definition of parallel genetic adaptation as the situation where all underlying alleles are fixed (l. 159-163). I wonder how the conclusions would be affected by two relaxations, namely that i) fixation does not need to occur at all, but only a proportion, of underlying loci, and ii) fixation does not need to be complete - which would be the case if there were gene flow from the ancestral population or other sources.

Minor comments
---

l. 8: Insert "that parallel genetic evolution" after "effective population size, and".

l. 19-10: Replace "how genomic architecture shapes adaptation" by "how genomic architecture impacts adaptation".

l. 27: Is this reference to Hohenlohe et al. (2010) at the correct position, i.e. is this really where "parallel evolution) is defined/reviewed for the first time?

l. 30: Replace "shape" by "influence". As a comment: the genetic architecture is also expected to evolve in response to selection pressure, but most likely on a longer time scale.

l. 32: Are all the k alleles adaptive? If so, please insert "adaptive" or "beneficial" after "possible".

l. 34: I suggest introducing "gene reuse" in a sentence; it is a bit confusing as it also applies to "allele reuse" in your context.

l. 49-51: It would be nice to have one more sentence summarising Chevin et al.'s (2010) findings on the

distribution of allelic effect sizes a bit more.

l. 55: Insert comma after "systems".

l. 62: I find "extends" not exactly matching and suggest "complements". Then I would replace "complementary" by "alternative" in l. 59 to avoid repetition.

l. 64: Insert "genetic" after "parallel".

l. 65-70: Given the outline of discussing the effects of the allelic effect size distribution and the importance of the experimental design, I would have expected more than only a sentence about these topics in the Results/Discussion.

l. 79: Insert comma after "example".

l. 87-88: Please explain "genetic complementation tests" or add a reference.

l. 88/93: Please introduce the abbreviations for the two tests eventually used later on in the text and tables.

l. 96: Please specify that you consider a haploid model.

l. 98: Insert "associated with allele A" after "b_i" to make clear that allele A is the one that increases the trait value.

l. 99-100: Delete ":" after "by" and insert comma immediately after the equation.

l. 104: It is a bit misleading that you set the initial frequencies equal for all descendent populations, but nevertheless use an index i for p_0. I realise that this is to clarify the notation when multiplying in equation (7). Perhaps mention that you keep the index i for clarity.

l. 107-108: Delete ":" after "expression" and insert comma immediately after the equation.

l. 118: Insert a full-stop immediately after equation (3).

l. 124: Delete ":" after "by".

l. 128-129: Delete ":" after "as" and insert comma immediately after the equation.

l. 122-129: This could be written more compactly; you only need to show what is currently equation (5); what is currently equation (4) is shown in the SI, which is sufficient.

l. 134: Delete the second "simple".

l. 143: Insert "genetic" after "parallel".

l. 148: Change "figure" to "Figure".

l. 169-170: An interpretation/explanation is missing. See comment 3) above.

l. 173: Overall, I think that the fact that the approach is Bayesian is overemphasised (e.g. it is unnecessary to say that it is Bayesian on l. 354), given that you do not explore alternative choices of the prior of eta. The main result of this paper is to provide the likelihood function.

l. 180: I agree that the limit of eta -> 0 theoretically corresponds to the case of no selection. However, the diffusion approximation states that drift will be dominating if eta < ~1. So, the biologically relevant threshold is not eta = 0, but eta ~ 1.

l. 193: Insert comma immediately after the equation.

l. 209 ff.: When describing the inference procedure, please state the maximum n (no. of loci) that you used (Figure 3 implies 10, but I was not sure). More importantly, I suspect that your approach scales very well with the number of loci and populations. If so, it would seem worth emphasising that (e.g. in the Discussion).

l. 214-215: Did you draw the allele frequencies from a uniform distribution between 0 and 0.1? Please clarify.

l. 220-222: It was not clear to me what you mean by this.

l. 222: Replace "effect" by "affect".

l. 228/229: Delete the apostrophes after "F1" and "QTL".

l. 230/233: You do not seem to return to n_CG and (CG or GC?) and n_QTL. Is it necessary to introduce these variables?

l. 232: "D under this method..." -> "Under this method, D...".

l. 235: The rejection criterion in the supplementary material suggests you are using a Metropolis-Hastings algorithm, not a Metropolis algorithm – unless the jump distribution is the same for all steps, in which case you should specify this in the supplementary material. Metropolis would start with an uppercase letter.

l. 241: Shouldn't this be the other way round, i.e the number of candidate genes under the QTL method is at least as large as the one under the candidate genes method?

l. 243: Insert "the" before "data".

l. 254: Fix the starting quotes for "reproduction", which are currently typeset as ending quotes.

l. 259-260: I suggest tracking the fixation times when you repeat the analyses, and report them. Is it realistic to assume that there is enough time for fixation in reality?

l. 263: See comment to l. 235.

l. 265: "Individual based" -> "individual based".

l. 270: Insert a comma immediately after the equation.

l. 271: "optima" -> "optimum" (singular).

l. 274: See comment to l. 271. No need to repeat "theta".

l. 284: Insert comma after "linear".

l. 293-295: Please also assess the sensitivity to gene flow among descendent populations (see comment 2 above).

l. 296: Please correct the references to Tables S4-S6 if necessary.

l. 299-303: This is because there is a likelihood function, not because the framework is Bayesian!

l. 301-307: I suggest thinking in terms of "biological", not statistical, significance, i.e. eta > 1 should be your criterion. See comment to l. 180 above.

l. 310: Perhaps "inherently" instead of "inexorably"?

l. 311: I wonder if you want to be a bit more conservative in your formulation, as you only partially formalise the connection between the genetic architecture and parallel genetic evolution. You basically ignore linkage, assuming quasi-linkage equilibrium.

l. 319-321: It is not clear what you mean by "another piece of genetic natural history". Do you mean "demographic history"? Please clarify.

l. 332: Insert comma after "In contrast".

l. 335-337: This result falls short of being trivial and basically directly follows from Barton and Turelli (1991) and Kirkpatrick et al. (2002).

l. 336: Insert "the" before "probability".

l. 337-340: Are you implying that in this stickleback example variation in initial frequency prevented detection of a signal? Please add a clarifying sentence.

l. 348: Delete "terribly".

l. 356: Add a comma after "For example".

l. 357: "8 total loci" -> "8 loci in total". Also, would it not make more sense to talk about "data points" rather than "loci" here?

l. 361: Please add references for the beach mice and cave fish examples.

l. 363: Insert a comma after "evolution".

l. 370: "be" -> "by".

l. 389: "estimate for" -> "estimate of".

l. 390: Delete "Bayesian".

l. 392: "our approach" -> "this method".

l. 392-393: "on parallel genetic evolution" -> "from multiple populations with a common ancestry."

l. 400: In all references, replace en-dashes in page ranges by em-dashes.

l. 404-405: Species names should be italic. Also applies to l. 447, l. 450-451, and l. 462-663.

l. 406: Use lowercase initial letters. Also applies to l. 414-415, l. 426-427, l. 432-433, l. 439, and l. 452.

l. 476/478: Remind the reader of the meaning of GC and GG; they have not been introduced before and I found it difficult to relate GC and GG to "candidate gene method/test" and "QTL method". Would, e.g., "CG" and "IQ" for "candidate gene" and "independent QTL test" perhaps be better options?

l. 483: The non-linear increase of the probability of parallel evolution is not surprising given the logistic form of Eq. (6).

l. 487: Delete comma after "n loci".

l. 488: Insert commas after "n = 50" and after "frequency".

l. 490: "dipicted" -> depicted

Figure 3: Please make clear that the first panel shows the three effect-size distributions that are considered. Please provide an interpretation of the pattern in the bottom-right panel (eta = 500, p_0 = 0.1). Why do the curves for different input distributions look the way they do?

Figure 5: It is not clear what the difference between panels A and B is in terms of parameters. Please specify the values of eta, b and p_0 in the caption.

Supplementary Material:

Middle of p. 1: Insert a comma after "Given this simplification".

Around equation S4: Replace ":" by a comma after "each locus". Add a full-stop after the equation. Replace "=" by an approximately equal sign.

First line after equation (S5): Delete "the alleles at"; "loci" -> "locus".

Around equation (S6): Delete ":" after "reveals that"; put the two equations on separate lines and typeset "and" in regular, not italic font.

Around equation (S7): Replace the full-stop after "Kirkpatrick et al. (2002)" by a comma; add a comma after the equation; change "Where" to "where", and typeset the locus indices i and j in italic.

Two lines above equation (S8): Insert "coefficient" after "selection".

Around equation (S8): Delete ":" after "in (S7) to"; add a comma after the equation; "which simplifies to..." -> "which further simplifies to..."; "upon" -> "after".

Equation (S10): add a comma immediately after the equation.

First line on page 4: Adjust according to the comment to l. 122-129 above.

Last sentence before "B: Markov Chain Monte Carlo Simulations of Posterior Distributions:" Please fix the following formulation: "the product a single locus parallel evolution events across loci".

First paragraph of "B: Markov Chain Monte Carlo Simulations of Posterior Distributions:": Insert an opening sentence. "At the conclusion of the individual based simulation we have..." -> "The individual based simulations provide...; "consists" -> "consist". "2 X n" -> "2 x n". Remove apostrophes after "0" and "1". If

appropriate, replace "Metropolis algorithm" by "Metropolis-Hastings algorithm". It would be good to tell the reader that details will follow later. E.g. insert "(see below for details)" after "...of the algorithm respectively". "can be computed" -> "can be quantified". Use another variable than m for the number of chains, as m is already used for the number of loci. Omit the entire part starting with "One important feature..." and ending with "... in a sequence (Gelman 2004)". Just say "We treated the first half of the sequences as burn-in period".

Before the description of the algorithm, add a transitional sentence.

Algorithm: Step 3: Please specify the jump distribution. Step 5: "now ranges between 1 and n" -> "now ranges from 1 to n". Add a comma after the equation for R; replace "Where:" by "where". Give references for the equations for B and W, derive them, or make clear they are also given in Gelman (2004).

Second sentence in "C: Sensitivity of the Bayesian Estimator to migration from the ancestral population": I disagree with "Although this is likely true in many well-studied cases". In most cases, there is probably quite some gene flow from the ancestral population, as well as admixture among derived populations.

Last two sentences on page 6: Omit the sentence starting with "Although this assumption is...". "We tested this possibility by..." -> "We assessed the effect of varying strengths of selection by...".

First paragraph of page 7: I found that quite interesting. At what stage do you see differences, i.e. an effect of different selection gradients among populations? Did you also compare different absolute magnitudes of the selection coefficient?

Section "E: Sensitivity of Bayesian Estimator to error in parameters": "to many violations" -> "to violations"; "with an assumption of their own" -> "under an overarching assumption"; "centered about" -> "centered around".

Section "F: References": Replace en-dashes in page ranges by em-dashes. Gelman (2004): The book title should be italic. The same applies to Hart and Clark (2007) and Karlin and Taylor (1981). Change initial letters to lowercase in the title of Nagylaki (1993).

Tables S1 to S6: Please make clearer in the captions that only Table S1 is based on the Wright-Fisher simulations, and all the others on the individual-based simulations. Please remind the reader of the meaning of the intercept, as it varies substantially among comparisons.

The caption to Table S4 seems to be missing.

The captions to Tables S5 and S6 seem to be confounded.

Print Close Window

© Thomson Reuters | © ScholarOne, Inc., 2015. All Rights Reserved.