Peer review on manuscript

"Determining causes of genetic differentiation in [redacted] to identify the need for conservation actions"

by Peer 552

ADDED INFO ABOUT FEATURED PEER REVIEW	**This peer review is written by Dr. Miguel Navascués, Researcher at The Center for Biology and Management of Populations, France** **PEQ = 4.7 / 5** **Peer reviewed by 2 Peers**

Introduction

This study [1] focuses on [redacted] . This is a peripheral population which has been shown to be significantly differentiated from the surrounding [redacted] populations and to have a lower genetic diversity in a previous study [2]. The authors attempt to determine the demographic history of BP population that has lead it to such genetic isolation and reduced size. In order to do so they proposed a novel simulation-based framework which they consider to be applicable to other species.

Revision Recommendations	
Question:	Minor
Data:	Accept
Methods:	Major
Inference:	Major
Writing:	Major

Merits

(1) The study [1] addresses the case of a peripheral [redacted] population with lower genetic diversity. According to the authors, their research is relevant for other carnivore species under similar circumstances (particularly in [redacted]). Arguments in the introduction seem sound to me and I agree that this could be an interesting study case once the methodological problems described below are addressed.

Critique

(2) The major weakness in [1] is their choice of methods and models, particularly with the proposed simulation-based approach:

(2.1) From a population genetics point of view the problem addressed by the authors is the classical case of two diverging populations with gene flow (a.k.a. "Isolation with Migration" model, IM). Authors [1] are trying to make inferences on the time of divergence, presence of gene flow and changes in population sizes. The software IMa [3,4] implements a likelihood based MCMC method under the IM model for estimating those parameters. This method has been largely used and tested [5] and its weakness and strengths are well known [6]. I do not see any reason for not using it in this case.

(2.2) Alternatively, methods based on simulations are also available. In the last 10-15 years a new statistical inference framework based on simulations has been developed which is usually known as Approximate Bayesian Computation (ABC, [7]), and there is software that implements the IM model [e.g. 8] under this framework. The simulation-based scheme proposed in [1] ignores all these developments. Two important features of the ABC framework are missing in [1]:

(2.2.1) In ABC, parameter values are taken from a prior probability distribution. Instead, simulations in [1] take a single or few parameter values into consideration, based on estimates that simulations in [1] take a single or few parameter values into consideration, based on estimates that might not be even close to the true values. Particularly, effective population size (Ne) parameter was based on some census population size estimate; however, effective and census sizes rarely match (See also point (2.4) regarding the joint estimation of Ne and μ). Also, mutation rate (μ) was based on some average mammalian rate; but the mutation rate can vary few orders of magnitude even for the same type of microsatellite in the same species [e.g. 9]. Using prior distributions allows to use previous information about the parameters but taking into account the uncertainty of such estimates and to explore a reasonable range of values for those parameters for which there is no previous information. From the simulations presented in [1] it is not possible to know if some of the rejected scenarios could reproduce the same genetic diversity patterns using some other combinations of parameter values.

(2.2.2) In ABC, several developments have been made to ensure that the comparison of the simulated particles and the real data is done in the most objective and informative way. Summary statistics are chosen as to be informative in key parameters of the model (why did not [1] use FST as summary statistics when models include migration?), and are standardize [7] or transformed [10] so there are not differences in their weight and they are chosen on their informativeness. Simulation particles are weighted depending on their distance to the observation, so inferences are made from the simulations most resembling the real data. The scheme presented in [1] has not clearly defined criteria for model choice (what is the role of the STRUCTURE results in model choice? it is not mentioned in Materials and methods section but it seems to be used as a criterion in the results). Also there is no measure on how good or bad the fit of the model is, so it is not possible to know which of the models that were not rejected offers the best fit to the data.

(2.3) Population genetic simulations can be done under the coalescent model (backwards in time) or forward in time (individual-based models). Each method has its merits and its limitations. The main

limitation of coalescent modeling is the simulation of natural selection. However, all the features of simulations performed in [1] can be done under the coalescent (including sex-biased dispersal, but see also point (2.5) regarding estimation of migration rates for males and females). Coalescent simulations can be used to simulate different effective population sizes, mating systems (with some limitations on software availability, see http://popmodels.cancercontrol.cancer.gov/gsr/), migration and mutational models contrary on what seems to be suggested in [1, lines 162-165]. In addition, coalescent simulations have several advantages over forward-time simulations regarding simulation-based statistical inference:

(2.3.1) Under the coalescent only a sample of the population is simulated which allows direct comparison with observed data.

(2.3.2) Coalescent simulation finishes with the most recent common ancestor of the sample, which allows to tackle models in disequilibrium (such as those of diverging populations or populations with population size changes). In contrast forward-time simulations are forced to an equilibrium starting point [1, Appendix I, lines 83-86].

(2.3.3) Coalescent simulations are much faster than forward-time simulations. Considering that large numbers of simulations (at least 1 million per scenario under the ABC framework) are required to obtain meaningful inferences, this is a very important point.

(2.4) The models and inferences made in [1] separate the population genetic parameter $\theta=2Ne\mu$ into two parameters Ne and μ. However, population genetic data is usually only informative on θ (as evidenced in the equations discussed in [1], lines 254-257). Assumptions made on the mutation rate value are unfounded, and thus are the estimates of Ne based on heterozygosity. In order to obtain estimates of Ne dissociated from μ, only the methods based on linkage disequilibrium (from unlinked loci, so the recombination rate is known) or in temporal samples are valid. Those estimates are for the current Ne and for the Ne between sampled times respectively.

(2.5) Simulations with different sex-biased dispersal were used in [1] in their simulation-based inference scheme. However, data (autosomal microsatellite) used in that inference is not informative about sex specific parameters. In order to study such features, markers with contrasting inheritance (autosomal, sexual chromosomes, mitochondrial) should be used. However, given the low diversity of the mitochondrial and being a single loci I would not recommend the author to try. In conclusion, adding such a complex feature as sex-bias dispersal in the model is unnecessary.

(3) Presentation of the materials and methods is very poor. There is important information missing and some key information only appears at the end of the article or in the appendices:

(3.1) Additional information on the molecular markers used should be present. How many microsatellite loci? Include a reference for those loci (reference that first described them and contains primer sequence). Also, which mitochondrial locus was sequenced?

(3.2) In the Materials and methods section it is stated that some "genetic structure analysis" was performed (line 220). There is no reference concerning the method (which we learn is the clustering analysis implemented in STRUCTURE at the Results section) nor to any of the options used in the analysis (Which model was used: F model? correlated frequencies? Which length of the MCMC was used? How many replicates? How many different K values were explored? etc.). The only reference is a previous study on the same species which might not be available to the reader: the minimum relevant information about the analysis must appear also in [1].

(3.3) Predictions on the future of BP population under different management treatments were studied using forward-time simulations. The initial conditions of those simulations are not properly described: the manuscript says that initial conditions are set using "information provided by our BP samples" (what information? Allele frequencies? Actual genotypes? Which parameters or initial conditions are determined and to which values based on what information?). Then parameters (note the plural, as in [1, line 332]) are said to be the same as in scenario 3, which is described in the appendix. Some additional information seems to be obtained from the appendix but also new questions are opened. We can guess that the initial conditions are set by using the genotypes from the BP given the text in appendix I, lines 124-125. However the effective population size is said to be set to 220 individuals and the sample size of BP is of 139 individuals. Also, what are the other parameters mentioned in [1, line 332]?

(3.4) There are several points that would require some clarification or should be moved to the materials and methods section instead of appearing in an appendix regarding the simulations. However, because I consider the analysis unsatisfactory in their current state (as discussed in points (2.1), (2.2) and (2.4)) I will not discuss them in detail.

Discussion

This study [1] is based on previously published data [2] on [redacted] (the only new data, [redacted], are not particularly relevant for the aim of the work). Therefore, the publication interest resides mainly on the new questions addressed and the new methodology presented. I agree that the inference of the demographic history of an endangered population has some conservation interests, particularly when management strategies are discussed (whether this is of general or local interests should be better judge by a researcher with more experience in zoology and conservation than me). However, the proposed new approach of simulation-based inference is very unsatisfactory. Authors do not seem to be aware of the latest developments of the field of population genetic statistical inference and their proposal is not new and clearly inferior to currently available methods. I suggest to reanalyze the data with the software IMa [3,4], or alternatively under the ABC framework. I have appended few additional suggestion that I think they will improve the article:

(5) Bottleneck analyses (HE excess and M-ratio) in [1] are relatively old methods less powerful than likelihood-based methods developed more recently such as those implemented in MSVAR [11] and BEAST [12] or ABC methods such as implemented in DIYABC [13]. What is more, since these methods are based on summary statistics we could argue that they have been superseded by the ABC. As an example, Bottleneck software authors (i.e. Cornuet & Estoup team) do not longer maintain or use that software having changed to DIYABC. There is nothing wrong with those bottleneck analysis but I would recommend using more modern approaches to any researcher. Note, in any case, that for the present case the most appropriate software are those which implement the IM model (see points (2.1) and (2.2)).

(6) If the Bottleneck (HE excess) analysis is kept, I would recommend to drop the TPM. This model might be more realistic for some loci; however, its properties regarding the expected HE are intermediate between the SMM and the IAM. This makes unnecessary to test also the TPM: if both IAM and SMM are significant TPM will be significant, if none are significant TPM will not be significant, if one is significant and the other is not, there will be a value of parameter pg for which TPM changes from significant to no significant. None of that is really informative since the actual mutational model is unknown and distract the reader from more interesting results (large tables and unnecessary text). Regarding the M-ratio test, I would advice to restrict to the SMM, since any parameter value used for the TPM (even those "recommended values" found in the literature) are based on little evidence and certainly not on evidence of the species and loci studied in [1] (i.e. there is not such a thing as some magical universal values applicable to any species, population or locus). Using the wrong parameter values will lead to incorrect inference about the bottleneck, as noted in [14].

(7) I think genetic diversity indexes are important enough for the article as to appear in the main text rather that in an appendix. Also, for mitochondrial sequences I suggest to point to the relevant Genebank references rather that (or in addition to) some codes used in another publication on the same species.

References

[1] Anonymous authors (2013) Determining causes of genetic differentiation in [redacted] to identify the need for conservation actions (unpublished manuscript) - Peerage of Science

[2] [redacted]

[3] Nielsen R. & Wakeley J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158(2):885-96.

[4] Hey J. & Nielsen R. (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. PNAS 104(8):2785-90.

[5] Pinho C. & Hey, J. (2010) Divergence with Gene Flow: Models and Data. Annual Review of Ecology, Evolution, and Systematics 41:215-230

[6] Strasburg J.L. & Rieseberg L.H. (2010) How Robust Are "Isolation with Migration" Analyses to Violations of the IM Model? A Simulation Study. Mol Biol Evol 27(2):297-310.

[7] Beaumont M.A., Zhang W. & Balding D.J. (2002) Approximate Bayesian Computation in Population Genetics. Genetics 162(4):2025-2035.

[8] Lopes J.S., Balding D. & Beaumont M.A. (2009) PopABC: a program to infer historical demographic parameters. Bioinformatics 25(20):2747-2749.

[9] Burgarella C. & Navascués M. (2011) Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data. European Journal of Human Genetics 19(1):70- 75.

[10] Wegmann D., Leuenberger C. & Excoffier L. (2009) Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood. Genetics 182(4):1207-1218.

[11] Beaumont M.A. (1999) Detecting Population Expansion and Decline Using Microsatellites. Genetics 153(4):2013-2029.

[12] Wu C.-H. & Drummond A.J. (2011) Joint Inference of Microsatellite Mutation Models, Population History and Genealogies Using Transdimensional Markov Chain Monte Carlo. Genetics 188(1):151-164.

[13] Cornuet J.-M., Ravigné V. & Estoup A. (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0) BMC Bioinformatics 11:401.

[14] Peery M.Z., Kirby R., Reid B.N., Stoelting R., Doucet-Bëer E., Robinson S., Vásquez- Carrillo C., Pauli J.N. & Palsbøll P.J. (2012) Reliability of genetic bottleneck tests for detecting recent population declines. Mol Ecol (14):3403-18.

Additional comments for authors

Line 291: it should read SMM instead of SSM.