Reply to Bodare *et al.* "Origin and demographic history of the endemic Taiwanese spruce (*Picea morrisonicola*)"

April 10, 2013

1 Summary

This study aims at measuring and explaining contemporary genetic variation and assessing long-term genetic stability of Taiwan spruce (*Picea morrisonicola*) endemic to the island of Taiwan. Approximate Bayesian computation (ABC) is applied to compare alternative models for the demographic history of the species and to infer parameters (effective population size and its variation over time, ...). Using previously published sequences of four supposedly closely related species (including *P. wilsonii*), the authors establish the phylogenetic origin of *P. morrisonicola*, suggesting closest relationship to *P. wilsonii*. When ignoring variation in N_e (see above), the authors date the split to 4–8 million years ago, which coincides with the formation of the island of Taiwan. When accounting for variation in N_e, the split between *P. morrisonicola* and *P. wilsonii* is dated to a much more recent point in time.

[Say something about the relevance, impact and writing.]

- What about assessing the genetic stability? Why not put a sentence in the Abstract?
- *P. morrisonicola* suffers from overexploitation from logging. Conservational interest of the paper?
- Is there evidence for there being habitat suitable for conifers immediately after formation of Proto-Taiwan? In other words, does 4–8 mya sound like a realistic time for the split between *P. morrisonicola* and *P. wilsonii*?
- Are power and false positive rate of the ABC-type model comparison procedure appropriately assessed? Are they assessed with respect to sampling size only or also with respect to the choice of statistics?
- How do the authors deal with and discuss the main issues of ABC: i) choice of summary statistics, ii) choice of algorithm (rejection vs. ABC-MCMC), iii) choice of distance metric?

Main goals of the study:

1. Compare extant genetic diversity to that of other spruce species.
2. Assess the impact of climate change on the trajectory of N_e. This uses ABC, comparing alternative demographic models.
3. Estimate divergence time and rates of gene flow between *P. morrisonicola* and four related spruce species found on mainland China. Two Bayesian approaches (ABC, MIMAR) are used and compared.

2 General issues

1. Judging from a map of Taiwan, it is not obvious if the sampling locations used in this study cover the whole distributional area of *P. morrisonicola* or actually only a small percentage of it (*cf.* Figure 1). Incomplete sampling of the species range could lead to artefacts that are not discussed. Although conifer species usually show little population substructure, I would like to have some support for or against the idea that the samples used by the authors are representative of the whole island. Moreover, as the samples are confined to a small area, micro rather than macro climatic conditions might have played just as big a role. In the lack of a 'control' location, it is not possible to disentangle potential effects and interactions.
2. The supporting data (sample locations, DNA sequences and input files for MIMAR and STRUCTURE should have been available to the reviewers. Judging from the intention of the authors to make these data available on DRYAD, I suggest the data be made available for a potential second round of review.
3. The choice of summary statistics in known to be crucial in ABC, but the authors do not mention this. Given that several approaches for choosing them have recently been published (JOYCE and MARJORAM, 2008; WEGMANN *et al.*, 2009; NUNES and BALDING, 2010; FEARNHEAD and PRANGLE, 2012; AESCHBACHER *et al.*, 2012), it would seem appropriate that studies applying ABC should at least properly discuss their choice, if not apply at least one of these approaches.
4. Related to the previous point, ABC-based model comparison suffers from the lack of sufficient statistics, even more so than ABC-based inference of parameters for a specific model. Although the authors have performed a simulateion-based power analysis (which is much appreciated), they should explicitly mention the issue and refer to the respective literature (*e.g.* ROBERT *et al.*, 2011). See l.266–271.

3 Specific comments

3.1 Abbreviations used

Q Question

C Comment

S Suggestion (mostly style)

R Re-formulation or change needed (usually followed by a suggestion)

→ Suggested change/correction

3.2 Introduction

l.39 **S:** Comma after 'In the face of global warming'

l.40 **S:** Ditto after 'For some species'

l.50–51 '. . . is subjected to very different climatic pressures than its mainland or boreal relatives' → . . . is subject to climatic conditions very different from those experienced by its mainland or boreal relatives. **S:** In brackets, give examples (in terms of temperature, humidity. . .).

l.51 **Q:** I guess 'boreal' is an established term. If so, why do you quote it? If no, please define it. Is there a difference between 'cool-termperature' and 'boreal' species (*cf.* l.55–56)?

l.90–92 **C:** It was not clear to me what was meant to be said in this sentence. A population genetic approach does not *per se* justify the use of more independent loci compared to increasing the number of sampled individuals. This depends on the question of interest (*e.g.* fine-scale study of population structure *vs.* medium- to long-terms studies of population history/admixture/growth). I guess the authors want to make sure that the relatively small sample size (actually, both in

terms if individuals and loci) is not a major limitation to the study. Analysis of power and false positive rate are done, which is fine. It's just that this one sentence reads confusing. Why not directly state that sample sizes are rather small for a relatively large area? In orther to really 'profit' from a large number of independent loci, the number of loci might have to be well beyond 15, depending on the question of interest.

l.112–113 'This method [ABC] is approximate and assumes... relevant to the model'. **R:** Either drop 'approximate and' because 'approximate' is already part of the name of the method and hence redundant, or make sure that the reader does not get the impression that the use (and choice) of summary statistics is the only approximation involved. The others being i) the choice of a distance metric (rejection tolerance) and ii) the fact that a finite number of simulation is used, whereas convergence to the true ABC posterior would only be reached with an infinite number of simulations (see one of the recent ABC reviews, *e.g.* Beaumont, 2010).

l.116 No comma after 'Although'.

l.120 'supplement' → complement (?)

3.3 Materials and Methods

As a more general comment, I suggest inverting the order of subsections 'Within-species ABC models' and 'Between-species ABC models' (and accordingly of Figures 2 and 3). This would avoid switching forth and back between the 'within' and 'between' species context.

l.130–136 C: As all fifteen loci were initially called in species other than *P. morrisonicola*, some of them not very closely related to *P. morrisonicola*, I missed a justificiation for why the authors think that the markers they used are appropriate (*e.g.* no ascertainment bias, putative neutrality, map distances / physical linkage structure). The assumption of no physical linkage seems particularly important (*cf.* l.161) and should be stated and justified.

l.138 Q: What is the maximum number of copies tolerated? Is copy number variation for a given gene comparable across the spruce species concerned here?

l.163–165 C: A potentially considerable amount of information is thrown out here (62 out of 192 SNPs are kept) in order to have putatively low physical linkage (or background LD) between markders. This is not a criticism of the paper, but rather represents the state of current inference methods and shows the need for methods that can account for between-locus physical linkage (or, from another perspective, within-locus recombination). **C:** Nevertheless, you might want to justify the choice of 50bp for the minimum distance between retained SNPs. In their Discussion, Falush *et al.* (2003) suggest two ways of doing so: i) combining historical information about likely admixture times and knowledge of between-locus recombination rates or ii) post-hoc inspection of Structure output with, in this case, varying choices of between-marker distance.

l.187 Q: Is it justified to assume symmetric gene flow between *P. morrisonicola* and *P. wilsonii*? In any case, you should discuss and if possible justify the assumption.

l.202–204 C: The argumentation here seems somewhat circular. Not only do estimates of divergence time depend on mutation rates, but estimates of population mutation rates / substitution rates are themselves dependent on (non-genetic) estimates of divergence time.

l.227–228 R: For consistency between main text and figures: '...at t coalescent units...' → at $t_0 = t$ coalescent units...; '...that persist for 0.2 coalescent time units...' → ...that persist for $t_1 = 0.2$ coalescent time units...

l.229–230 S: 'choose' → chose (?).

l.234–235 S: '...against a 'null' constant effective population size model...' → ...against a null model with constant effective population size...

l.235–237 S: I suggest using a different formulation for the priors: 'The priors for the model parameters were chosen to be uniform as follows: $\theta \sim U(0, 0.01)$, $\rho \sim U(0, 0.02)$, $t \sim U(0, 1.5)$, ...' **R:** The prior for α, $U(1, 1.5)$, did not make sense to me, at least when following Figure 2, where it says that the size of the bottlenecked population is αN. I would thus expect $\alpha < 1$ throughout if one wants to consider bottlenecks only. From l.272 it seems, however, that the lower bound should be 0 instead of 1 and that the authors would like to include the potential of population growth, too (upper bound of 1.5).

l.237–238 R: The choice of summary statistics should be appropriately discussed (see general issues 3 and 4 above).

l.247 S: Comma after 'In the simplest scenario'

l.269–271 See general issue 4 above.

l.271–272 S: Again, I suggest using standard notation for the priors (see comment to l.235–237 above).

3.4 Results

In analogy to 'Materials and Methods', I suggest inverting the order of the subsections on within- *vs.* between-species ABC.

l.xx 'strongest effect ... is persistent founder effects' → strongest effect ... is caused by persistent founder effects

l.xx 'few population founders' → a few population founders

3.5 Discussion

l.xx 'strongest effect ... is persistent founder effects' → strongest effect ... is caused by persistent founder effects

l.xx 'few population founders' → a few population founders

3.6 Figures and tables

l.713 C: Figure 1 definitely needs a more comprehensive caption, with a delineation of the species range and a code mapping colors to altitudes. It would also be desirable to have a distinction between woodland and open areas, if applicaple. I further missed a declaration of the cartographic source in the caption (rather than the main text, see l.131–130). The figure needs a compass rose, a scale and, ideally, an inset figure showing the larger geographic context including parts of mainland China.

.715–719 C: I suggest using the same symbol for the effective population size in main text and figures. Specifically, I suggest replacing N in Figures 2 and 3 by N_e.

4 Supporting Information

4.1 Supporting figures

l.xx 'strongest effect ... is persistent founder effects' → strongest effect ... is caused by persistent founder effects

l.xx 'few population founders' → a few population founders

References

Aeschbacher, S., M. A. Beaumont, and A. Futschik, 2012 A novel approach for choosing summary statistics in approximate Bayesian computation. Genetics **192**: 1027–1047.

Beaumont, M. A., 2010 Approximate Bayesian computation in evolution and ecology. Annu. Rev. Ecol. Evol. Syst. **41**: 379–406.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics **164**: 1567–1587.

Fearnhead, P. and D. Prangle, 2012 Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. J. R. Stat. Soc. B **74**: 419–474.

Joyce, P. and P. Marjoram, 2008 Approximately sufficient statistics and Bayesian computation. Stat. Appl. Genet. Mol. Biol. **7**.

Nunes, M. A. and D. J. Balding, 2010 On optimal selection of summary statistics for approximate Bayesian computation. Stat. Appl. Genet. Mol. Biol. **9**.

Robert, C. P., J.-M. Cornuet, J.-M. Marin, and N. S. Pillai, 2011 Lack of confidence in approximate bayesian computation model choice. Proc. Natl. Acad. Sci. U.S.A. **108**: 15112–15117.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics **182**: 1207–1218.