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

April 11, 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, population subdivision / migration). Between-species analyses are complemented by applying MIMAR. Using previously published sequences of four closely related species (including *P. wilsonii*), the authors employ STRUCTURE to 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 (mya), 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 (1.1 to 2.2 mya, depending on the generation time assumed).

[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
- The relationship to climate change is speculative. There are no climatic data shown or referred to, and if so, the relationship proposed could be correlational, not causal.
- Why splitting the whole inference problem into so many chunks? If necessary, could at least the uncertainty in the estimate of T be incorprotated in the 'within-species' analysis?
- ...

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.
5. I have a reservation against the interpration (l.301–309) of the results presented in Table 1. Although it is true that the average across loci of Tajima's D is (weakly?) positive (0.281), two out of eight loci show negative D and the only significant value is also a negative one (locus SE1427). Given that (linkage to sites under) positive/purifying (balancing) selection could also cause negative (positive) Tajima's D for some loci, I think the authors should be more careful in taking their data as evidence for population decline. Considering Fay & Wu's H and Zeng's Z, the pattern is even more against the author's interpretation: more than half of the loci show negative values (two of them significant for H), which is suggestive of positive selection (local adaptation?) and/or population expansion. To summarise, as the number of loci is small and the distribution of neither of the 'neutrality' statistics is univoqually in favour of population decrease, I think there is little support for a reduction in population size.
6. Large parts of the analyses conducted seem sound and correct within their scope, but i) the interpretation of some of the results is questionnable and ii) it is a pity that the link from between-species to within-speeces analyses was not made more tight. For instance, comparing the MIMAR between-species analysis and the ABC within-species analysis, there is a gap between roughly 160,000 and 11,000 generations ago, during which the effective population size, according to the estimates presented, must have increased from ∼5,000 to ∼124,000. Unfortunately, this is neither modeled nor discussed. Regarding the ABC within- and between-species analyses, I

think the split is even more artificial and I wonder why the whole process (between- and within-species events) was not modelled in one entity and analysed using ABC. This would still allow for various scenarios to be compared, but avoid the artificial split into two phases that seem difficult to link *a posteriori*. Along these lines, I am uneasy about the difference between the estimates of N_1 in Table 5 for the within- and between-species ABC analysis (123,841 *vs.* 43,737, an almost threefold difference).

7. I found the listing of the main aims of the study given in the first paragraph of the Discussion inconsistent with the listing given in the Introduction (l.79–103). In particular, assessment of whether small sample sizes give reliable results was not stated as one of the main goals at the beginning of the paper. Overall, it is not fully clear whether genetic data should be used to infer unknown demographic events, or whether the goal was to find genetic signatures of past events that the authors think must have happened. This way, directions of causality and inference are blurred.

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 Abstract

l.1 Q: Are both 'Taiwan' and 'Taiwanese' Spruce correct species names? Is the inconsistent use between title and main text intended?

3.3 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-temperature species' 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-term study 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 (actually, seven out of them turned out to be monomorphic in the focal species), 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 method name 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 use of a non-zero 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 (?)

3.4 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 (see the MIMAR section before).

l.130–136 C: As all 15 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 range of spruce species considered 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. Strictly speaking, the choice of the minimum distance between retained SNPs is expected to influence the inference of admixture proportions, and so a sensitivity analysis would seem appropriate.

l.174–175 C: 'average estimated poterior probabilities of data' is a slightly puzzling term, as posterior probabilities are usually used in connection with parameters or models, not with data. Do you mean 'likelihoods' (*cf.* caption to Figure 4)?

l.187 Q: Is it justified to assume symmetric gene flow between *P. morrisonicola* and *P. wilsonii*? Would it not seem more likely that gene flow was directional, from *morrisonicola* to *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... **R:** In this context, make sure that 'coalescent time unit' ($= 4N_e$ generations) is defined at the right place. Currently, it is not defined until l.249, which would be fine if the order of the within- and between-species subsections were inverted, as suggested above.

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

3.5 Results

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

l.301–309 See general issue 5 above.

l.325 C: Figure 6 seems to be missing.

l.327–328 Q: Are these really confidence intervals? I suspect these are highest posterior density (HPD) intervals. Please check and correct if necessary.

l.328 R: According to Table 2, the lover bound of the 90% confidence (or HPD?) interval should be 5.5 mya, not 5.6.

l.329–330 C: It is hard to imagine that the founder population of *P. morrisonicola* was of effective size $N_e = 5000$. Strictly speaking, this is an effect of the model assumption that N_e was constant after the split. Yet, an estimate of N_e in this context may at most be interpreted as a long-term effective population size, certainly not as the effective number of founders. I strongly recommend a reformulation of this.

l.334–336 C: '. . . that whilst the retained model fits the data well, it does not accommodate all aspects on the demography of the species.' This reads contradictory (either the model fits the data well or not). In accordance with general issue 5 above, could the deviation between observation and model output be explained by (some) loci being (linked to sites) under selection?

l.360–361 C: Again, this could be a hint towards (linkage to sites under) positive selection / local adaptation to conditions on the island. See general issue 5.

l.381–385 C: Similar to the previous comment, I wonder whether the poor coverage of observed values of D in Supplementary Figure 2 could be indicative of signals of selection. I therefore find the conclusion in l.384–385 premature.

l.389–392 Low sample size is not the only reason for why power and false positive rate should be analysed: the lack of suffuciency of summary statistics for model comparison with ABC is at leas as important and should be mentioned (see general issue 4 above).

l.493 S: Comma after 'In the Pleistocene'

l.524 'be lost to genetic drift' → be lost due to genetic drift (?)

3.6 Discussion

l.417–420 C: If I interpret the estimates of effective population sizes correctly, it cannot be excluded that there was actually a period of substantial expansion between the split from *P. wilsonii* and the start of the bottleneck putatively caused by climatic changes (*cf.* general issue 6 above).

l.432 S: Comma after 'In this species'

l.444 Q: As mentioned above, are eight loci enough to provide the effect the authors decribe here?

l.463–466 S: Add explicit references to the respective figures.

481–484 C: This raises the question of why (directional) gene flow was not included as a factor in the models analysed using ABC (see earlier comment to Material and Methods above).

3.7 Figures and tables

As a general comment, I suggest denoting parameter estimates in tables with a $\hat{}$ to distinguish them from parameters.

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.

l.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.

Table 2 Caption: **R:** 'Mode, 5% and 95% condence [confidence?] intervals of...' → Mode, 5% and 95% quantiles of...; **S:** 'population mutation rate' → the population mutation rate; 'symmetrical migration rate' → the symmetrical migration rate; 'split time' → the split time.

Table 2 and Supp. Fig. 3 Q: Is it correct that averages across the 8 polymorphic loci are shown (*cf.* Table 1)? If so, please state this in the captions.

Tables 4 and 5 Please remind the reader that lower-case t is used for time on the coalescent time unit and upper-case T for corresponding times in years or generations. I missed an explicit definition of N_0 and N_1. It seems that the mode was used as a point estimator of θ when using the values in Supplementary Tables 4 and 5 to compute $\hat{N}_1$ in the first and second row of Table 5, respectively. Please state this. Please also indicate that the mean and mode of the θ posterior reported in Supplementary Table 4 differed by almost a factor of 2. Last, I suggest providing Supplementary Figures showing the full posterior distributions for all parameters, including the prior distribution. I would like to see their shape.

Suppl. Table 4 and 5 R: Please define in the caption what exactly is meant by 2.5% and 95%. Are these the limits of the 95% HPD interval?

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.