

## Second Review of Korfman et al. for PCI Evol Biol

I find the replies mostly convincing, and the made changes are welcome. However, I still have a serious problem with the scaling of the Beta-coalescent and it's limiting behaviour for  $\alpha \rightarrow 2$ .

(Note: All line numbers refer to the document with tracked changes)

Most importantly, I do not understand, why the Beta-coalescent is not *exactly* transitioning to the Kingman coalescent for  $\alpha = 2$ . The classical definition of the lambda-coalescent defines a coalescent rate for  $k$  out of  $b$  lineages as (see [1]):

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k} \Lambda(dx)$$

Now, as the authors define the Beta-coalescent we set:

$$\Lambda(x) = \text{Beta}(2 - \alpha, \alpha)$$

to be a Beta-Distribution with a parameter  $\alpha$ .

As far as I know, we have the limiting behaviour:

$$\lim_{\alpha \rightarrow 2} \text{Beta}(2 - \alpha, \alpha) = \delta_0$$

which is a Delta-Distribution at  $x = 0$ . Putting that into the formula for the lambda-coalescent rate yields:

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k} \delta(0) dx = \begin{cases} 1 & \text{for } k=2 \\ 0 & \text{otherwise} \end{cases}$$

which is exactly the Kingman-coalescent behaviour, isn't it? Why is there a scaling constant?

The authors's reply is unhelpful, I'm afraid, as they only explain why for  $\alpha < 2$ , the dynamics are different, which I can believe. But what the authors show in their plots is that even for  $\alpha = 1.9$ , there is a massive difference between the

population sizes estimated by PSMC/MSMC and the one put into the simulation.

Doesn't this show that there is some scaling problem of the mutation rates in your simulation?

I can make this more concrete:

1.) The explicit formulas for the scaling-factor are incomplete: In the formula for the so-called "scaling constant" on Line 64, there appears a  $\beta$ , which has not been introduced or defined as a parameter.

2.) The quotations after these formulas are unhelpful, at least to me. I took a look at all three papers (refs. 8, 55 and 56), and while I admit I didn't read them in all detail, I could not really find these formulas. Perhaps these formulas could be derived for the reader (with references) in a short Supplementary Chapter or a methods paragraph. They can then be taken out of the text in lines 62-64, actually, where they are a bit overwhelming I think.

3.) The authors' response about my critique of their figure 2 is partly convincing. I get that you want to make the point that indeed the population size inference gets wrong if the assumptions break down. *But*, coming back to my main point above, this point only comes across if you actually show that the discrepancy between expectation and fit actually vanishes for  $\alpha \rightarrow 2$ . I find it hard to believe that for  $\alpha = 1.9$ , the violation of the Kingman-coalescent assumption is already so stark that the population size is mis-estimated by a factor 100, which is what I see in Figure 2a. To repeat myself: I think there is something wrong with that. What I would have expected from that figure is a fit which looks *very good* for, say,  $\alpha = 1.99$ , perhaps marginally worse for  $\alpha = 1.9$ , and then perhaps increasingly bad for lower values. Instead, what I see in your Figure 2 is a terrible fit in all four cases, with a discrepancy ranging from a factor 100 to 1000.

Minor point: In Line 62 there is a typo, I think. It says  $Beta(2\alpha, \alpha)$ , but I think it should be  $Beta(2 - \alpha, \alpha)$

## References

- [1] Jim Pitman. Coalescents with multiple collisions. *The Annals of Probability*, 27(4):1870–1902, October 1999.