## Review of USYB-2015-180

The bifurcating phylogenetic species tree requires that evolutionary traits are inherited ancestrally, but never horizontally between species as might occur through hybridization. Jhwueng and O'Meara extend the Brownian motion model to operate on phylogenetic hybridization networks, where hybridization may induce a burst of phenotypic variation ($v_H$) or a rescaled ($\beta$) mixture of hybrid phenotypes ($m$). The authors approach this by adopting the multivariate normal representation of a tree-dependent Brownian motion, but modifying how the covariance terms are defined in terms of hybrid related variation. Overall, I think the idea is a biologically reasonable one, but I have some concerns about the formulation of the model.

First, the covariance matrix definition may be incorrect. The simplest example showing this concern would be to consider two taxa $X_1$ and $X_2$ whose MRCA is the root. They hybridize at the present with $m = 0.5$, $\beta = 1.0$, and $v_H = 0.0$ to create $X_H$. $X_1$ and $X_2$ can be considered iid samples from a Normal distribution (i.e. $Cov(X_1, X_2) = 0$), but $X_H$ is the sum of weighted values, $mX_1 + (1-m)X_2$.

Figure 1 shows that all contemporaneous extant species, since they have each evolved for the same amount of time since their common origin, have the same variance. Assuming $\sigma^2 = 1$ can be suppressed, this shows that $Var(X_H) \neq Var(X_1)$

$$\begin{aligned} Var(X_H) =& Cov(X_H, X_H) \\ =& m \cdot m \cdot Cov(X_1, X_1) + m \cdot (1-m) \cdot Cov(X_1, X_2) \\ & (1-m) \cdot m \cdot Cov(X_2, X_1) + (1-m) \cdot (1-m) \cdot Cov(X_2, X_2) \\ =& 0.5 \cdot 0.5 \cdot 1 + 0.5 \cdot 0.5 \cdot 0 + 0.5 \cdot 0.5 \cdot 0 + 0.5 \cdot 0.5 \cdot 1 \\ =& 0.5 \end{aligned}$$

Or a small proof through simulation

```
> m=1.0;n=100000;var(m*rnorm(n)+(1-m)*(rnorm(n)))
[1] 1.009986
> m=0.5;n=100000;var(m*rnorm(n)+(1-m)*(rnorm(n)))
[1] 0.5010948
```

Now, if the hybridization event occurred immediately after $X_1$ and $X_2$ originated, then $Var(X_H) \approx Var(X_1)$, since it would effectively manifest as another iid draw from the Normal.

This is a simple example (simpler than shown in Figure 1), and the implications of hybridization on covariance structure is more complicated in general. Since I did not see the work cited, the authors should be interested to learn of the work of Pickrell and Pritchard (2012) made available in the program TreeMix. They follow Cavalli-Sforza and Edwards (1967) in modeling allele frequency diffusion as a Brownian motion on a bifurcating tree, but allow admixed populations to inherit some proportion of alleles from two ancestral populations. In their Supporting Information document, equations 12 and 13 show how to compute the covariance for an arbitrary DAG. Using these equations, I believe $Var(R) = (m^2+(1-m)^2)(t_1+t_2+t_3)+2m(1-m)t_3$ for Figure 1, but the authors should verify this. The algorithm would be simple enough to modify to include $v_H$ and $\beta$ terms.

From the Discussion, the authors conveyed some disappointment in the performance of their method. If they agree the covariance matrix is incorrect, they may find hope in the fact that fixing $m = 0.5$ maximizes the error, so they stand to gain the largest improvements with the correction (Line 295). Additionally, the correction will account for some missing covariance that arises when seemingly unrelated species covary due to complex admixture/covariance histories (Line 199 – 10 hybrids). On the other hand, if the covariance matrix is constructed identically for simulation and for inference, then performance may not improve by much. I am hopeful!

Aside from the covariance matrix, I worry about the log transformation of the data and its interpretation during hybridization events. The authors take some care with interpreting $\beta$ in a log scale, but not $m$. When $X_1$ and $X_2$ are on a linear scale $X_H = mX_1+(1-m)X_2$ is the $m$-weighted average of parent species and is intuitive. On the log scale, the mixture implies $X_H = \exp(\log(X_H)) = \exp(m\log(X_1) + (1-m)\log(X_2)) = X_1^m X_2^{(1-m)}$, but it's not clear what process this represents. Note, Pickrell and Pritchard (2012) don't face this issue because they use a linear scale for allele frequencies and concede the model is poorly defined for boundary conditions (near 0 and 1). If the authors choose to remain in the log scale for traits, they must provide an interpretation for $m$.

In light of what I view as fundamental flaws in the model, I cannot recommend the paper be published as is. The covariance matrix appears readily remedied, but I am not sure what to do about the hybridization parameter $m$

other than to use a linear scale for $X_i$. If the authors choose to address these points, a fair portion of the manuscript will require rewriting. Although the flow and style were good, the content may need to change substantially, so I spent little time providing per-line corrections (typos, grammar, etc.).

Thank you for the invitation to review this paper.

Pickrell and Pritchard (2012):
http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002967
Pickrell and Pritchard (2012) Supporting Info:
http://journals.plos.org/plosgenetics/article/asset?unique&id=info:doi/10.1371/journal.pgen.1002967.s016