bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Trait Evolution on Phylogenetic Networks

Dwueng-Chwuan Jhwueng[1,2], and Brian C. O'Meara[3,4]

[1]*Department of Statistics, Feng-Chia University*

No. 100, Wenhwa Rd., Seatwen, Taichung, Taiwan 40724, R.O.C.

[2]*Email:dcjhwueng@fcu.edu.tw*

[3]*Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, 37996-1610, USA*

[4]*Email: bomeara@utk.edu*

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Species may evolve on a reticulate network due to hybridization or other gene flow rather than on a strictly bifurcating tree, but comparative methods to deal with trait evolution on a network are lacking. We create such a method, which uses a Brownian motion model. Our method seeks to separately or jointly detect a bias in trait value coming from hybridization (β) and a burst of variation at the time of hybridization (v_H) associated with the hybridization event, as well as traditional Brownian motion parameters of ancestral state (μ) and rate of evolution (σ^2) of Brownian motion, as well as measurement error of the tips (SE). We test the method with extensive simulations. We also apply the model to two empirical examples, cichlid body size and *Nicotiana* drought tolerance, and find substantial measurement error and a hint that hybrids have greater drought tolerance in the latter case. The new methods are available in CRAN R package *BMhyd*.

Keywords: hybridization, gene flow, phylogenetic comparative methods, Brownian motion, network.

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Various comparative methods have been proposed to deal with the fact of non-independence of species due to shared history on a phylogenetic tree (Felsenstein 1985, 2008; Cheverud et al. 1985; Grafen 1989; Gittleman and Kot 1990; Hansen 1997; Lynch 1991; Housworth et al. 2004; Butler and King 2004; O'Meara et al. 2006; Hansen et al. 2008; Beaulieu et al. 2012) as well as many other problems, ranging from ancestral state estimation (Schluter et al. 1997) to estimating the effect of traits on diversification rates (Maddison et al. 2007) to predicting extinction risk (Cardillo et al. 2006). However, hybridization, a fairly common process (Mallet 2005, 2007; Riesberg 2006), results in species being related in a phylogenetic network rather than a tree (Arnold 1996; Doolittle 1999; Otto and Whitton 2000; Linder and Rieseberg 2004; Huson et al. 2010; Nakhleh 2011), and this reality is not yet accomodated by existing comparative methods. Due to the importance of hybridization as a process, numerous methods have been developed to infer phylogenetic reticulate networks (for simplicity, we refer to these as "networks") rather than trees (e.g., Sang and Zhong 2000; Weigel et al. 2002; Bryand and Moulton 2004; Moret et al. 2004; Huson and Bryant 2006; Joly et al. 2009; Kubatko 2009; Meng and Kubatko 2009; Wang et al., 2013; Willson, 2013; Wu, 2013). We thus stand at a point where phylogenetic networks will increasingly be inferred but we lack comparative methods to properly use this information.

A species formed as a hybrid of two parental species can differ from its parents in important ways. Due to transgressive segregation (Rieseberg et al. 1999), hybrids may have trait values outside those of their parental species. If we consider species' trait values evolving through time in a Brownian motion random walk, this transgressive segregation can be modeled in a few different ways. For example, if hybrids are on average 10% larger than their parent species, this

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

could be modeled by a shift in mean trait value associated with hybridization. If processes like transgressive segregation lead to difference from parents but with no consistent trend in direction across many hybrid events, this could be modeled as a burst of variation at the time of the hybrid event. Hybrids may also evolve at different rates than parental species, especially if they are formed from polyploidization (Ainouche et al. 2008). This could be reflected in a different rate parameter for hybrid species than for non-hybrids. Finally, hybrids may not be formed equally from both parental species. For example, one "hybrid" species may have formed through regular allopatric speciation of a single species, plus a few genes introgressed from a neighboring species. If a phenotypic trait value represents the additive result of multiple quantitative loci, an appropriate model would treat the hybrid trait mean as being a weighted average of the two parental species' means, with the weighting based on the relative genetic contributions of each parent.

In this work, we propose a new comparative method to study trait evolution under phylogenetic networks. This method allows for estimation of traditional evolutionary parameters such as rate and overall mean under a Brownian motion model while also allowing investigation of trait evolution occurring as a result of the hybridization process. We test our model with simulations and investigate two empirical datasets of cichlid and *Nicotiana* (tobacco and relatives).

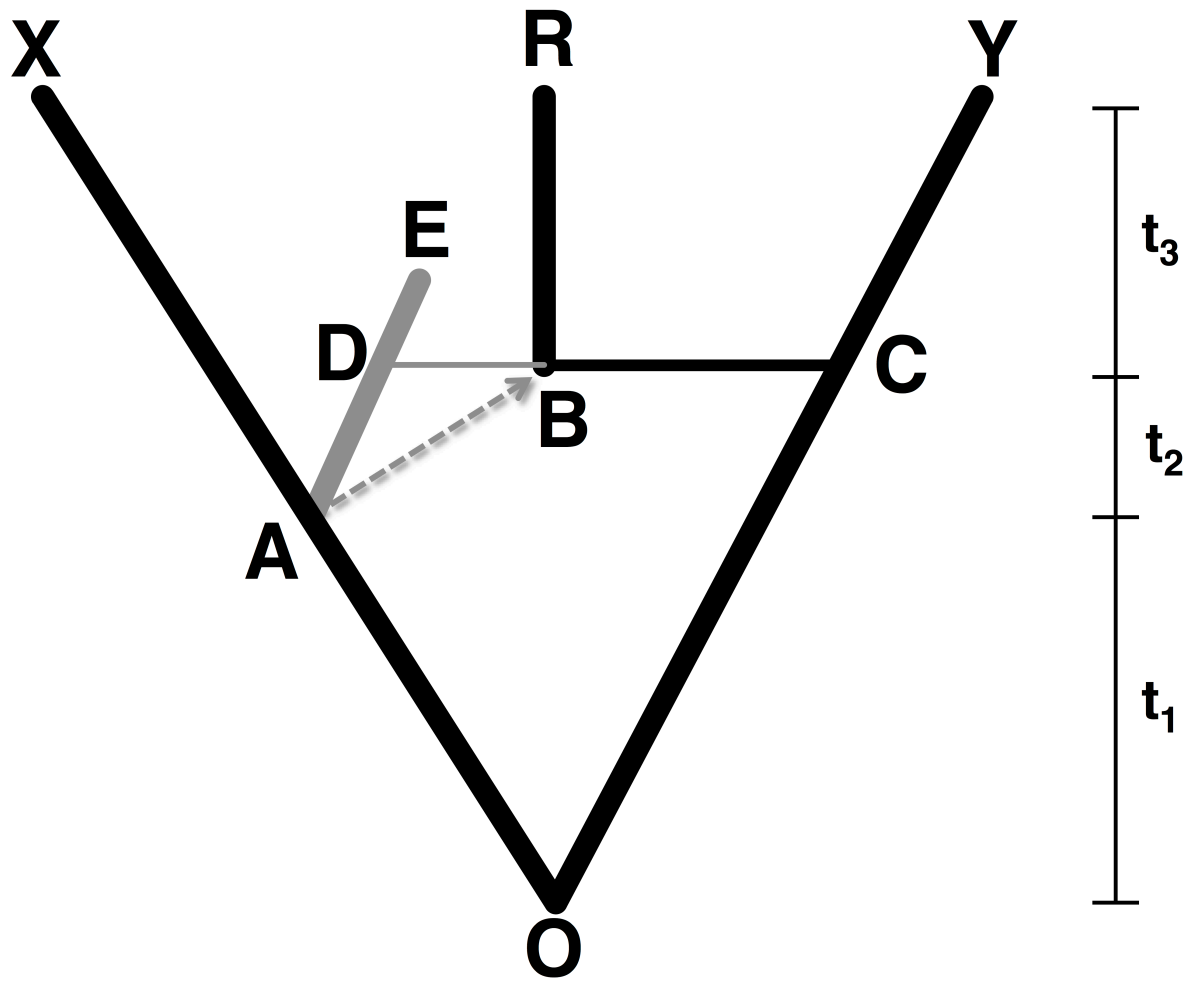
METHODS

Brownian Motion for Trait Evolution

Brownian motion (BM) is a general model for unbounded continuous trait evolution commonly used in phylogenetics (Felsenstein 1985). Biologists often incorrectly believe this

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

77 is only a model for traits evolving under genetic drift, but in fact a variety of biological
78 mechanisms can lead to this same model, such as selection towards an optimum that changes
79 due to multiple factors through time, drift-mutation balance, an evolutionary trend, as well as
80 pure genetic drift (Hansen & Martins 1996). Under Brownian motion, the variance of a trait
81 is proportional to evolutionary time multiplied by the rate of evolution, σ^2. Therefore, given
82 a phylogenetic tree the covariance among species can be represented by the shared branched
83 length on the phylogenetic tree. Figure 1 shows a three-taxon phylogenetic network with
84 gene flow.



85

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1: Three taxon network with extant species X, R and Y. Species C and species D are the immediate parents of species B; in this example, it gets 75% of its genes from C, 25% from D. This network can be thought of as a tree (the black edges) combined with gene flow from A via D into B (the gray edges). D's descendant, E, is not in the final tree that is used (X,(R,Y)) either through extinction or not being sampled. Thus, even though gene flow must occur between coeval species, the effective path of gene flow from the X lineage to the R lineage occurs starting at time t_1, not time $t_1 + t_2$, the actual origin of the hybrid: the changes occurring on the X branch in the t_2 time interval cannot be shared with the hybrid B. The dashed arrow starting at time t_1 and ended at time $t_1 + t_2$ thus indicates the effective gene flow from A to B.

This represents a scenario where at time t_1, there was a speciation event: one branch led to X, and the other led to a species D that eventually went extinct (E) or was otherwise unsampled in this analysis. However, at time $t_1 + t_2$, species D exchanged genes with the species at C to form a hybrid species, B, which survived to be sampled species R. Though gene flow only occurs between taxa occurring at the same point in time (D and C), due to extinction it can look like flow forward in time: the shared history of R with X is not from when D exchanged genes (time $t_1 + t_2$) but earlier, time t_1: changes on the A to D branch are not shared between X and R. Thus, the dashed line shows the effective path leading to the covariance of the observed tips, rather than the path from A to D to B and thus to R. The corresponding variance-covariance matrix for the extant species X, R, Y for the tree model (black edges only) is given by the matrix $\mathbf{V_s}$ as following

$$\begin{matrix} & X & R & Y \end{matrix}$$

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\mathbf{V}_s = \begin{array}{c} \\ X \\ R \\ Y \end{array} \begin{bmatrix} \sigma^2(t_1+t_2+t_3) & 0 & 0 \\ 0 & \sigma^2(t_1+t_2+t_3) & \sigma^2(t_1+t_2) \\ 0 & \sigma^2(t_1+t_2) & \sigma^2(t_1+t_2+t_3) \end{bmatrix}.$$

Now consider the trait evolution includes the gene flow (dashed arrow). Let the trait value of the root state *O* be μ. By assuming species evolve under Brownian motion (variables are measured in log scale in comparative analysis), the trait values at species D and species C are $\mu_D = \mu + \varepsilon_D$ and $\mu_C = \mu + \varepsilon_C$, respectively, where ε_D and ε_C can be regarded as error terms that follow a normal distribution with zero mean and variance $\sigma^2(t_1 + t_2)$. Under our model, the hybrid species *B*, at the moment of hybridization, assumes the value μ_B, defined as

$$\mu_B = \left[m\mu_D + (1-m)\mu_C \right] + \log\beta$$

which follows a normal distribution with mean $\mu + \log\beta$ and variance $\sigma^2(t_1 + t_2)$. The parameter m measures the proportion of the hybrid trait value inherited from parent D while $1 - m$ measures the proportion the hybrid inherits from parent C (m is bounded between zero and one). If the hybrid species is formed mostly from individuals of species D, with only some gene flow from species C, then for a polygenic quantitative trait, m might be much closer to one. In particular, $m = 0.5$ indicates inheriting the trait equally from both parents. In the absence of information to the contrary, 0.5 represents a reasonable value to use. The parameter β governs the possible bias in trait value as a result of hybridization. If there is a bias that leads to greater fitness, this is often called heterosis or hybrid vigor; if there is a bias that leads to lower fitness, this may be called outbreeding depression. Here we care about trait values, not their fitness effects, but hybrid means may be thought of in the same way, in that they may be something other than the average of their parents. For

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

example, if there exist widespread heterosis, with hybrids being on average 20% larger than their parent species, β would be 1.2. The natural lower bound for β is zero and the upper bound is arbitrary; a value of 1 indicates that the hybrid is just a weighted average of its parents. Brownian motion assumes that an increase or decrease by a certain amount has the same probability regardless of a trait value. This is often not the case for raw measurements: an increase or decrease of mass by 1 kg over a million years is far likelier for an elephant species than for a mouse species. However, they both might be equally likely to increase or decrease their mass by 1%. It is thus typical to log transform raw values to meet this assumption. Therefore, we add the log parameter, $\log \beta$, to represent log scale bias for the hybrid at formation. Here we assume that μ_A, μ_B and μ_C were in log scale already (i.e. the representation in raw scale is $e^{\mu_B} = \beta e^{m\mu_D + (1-m)\mu_C}$). To model a process like transgressive segregation, where a hybrid can deviate from the range of parental values but without a particular bias, non-negative variance v_H is added to lengthen the hybrid branch, equivalent to adding a burst of variation due to the hybridization event. Therefore, we have $Var(X) = Var(Y) = \sigma^2(t_1 + t_2 + t_3)$ and $Var(R) = \sigma^2(t_1 + t_2 + t_3) + v_H$. To allow hybrid species to have different rates of evolution than non-hybrid species, one would just require modifying this variance to be $Var(R) = \sigma_H^2 t_3 + \sigma^2(t_1 + t_2) + v_H$, where the new term σ_H^2 is the rate of evolution in hybrids. We currently limit the model to the one where the hybrid and non-hybrid species have the same rate ($\sigma_H^2 = \sigma^2$). The corresponding variance-covariance matrix for the species X, R, Y for the network model under the assumption of the BM process for trait evolution is given by the matrix $\boldsymbol{V_R}$ as following

$$\begin{matrix} & X & R & Y \end{matrix}$$

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\begin{array}{c} X \\ R \\ Y \end{array}\begin{bmatrix} \sigma^2(t_1+t_2+t_3) & \sigma^2 m t_1 & 0 \\ \sigma^2 m t_1 & \sigma^2(t_1+t_2+t_3)+v_H & \sigma^2(1-m)(t_1+t_2) \\ 0 & \sigma^2(1-m)(t_1+t_2) & \sigma^2(t_1+t_2+t_3) \end{bmatrix}$$

Furthermore, measurement error can be substantial, and it has the effect, if ignored, of leading to larger estimates of rates on tip branches. We deal with this (following a suggestion in O'Meara et al. (2006)) by adding a parameter, SE, to the diagonals of $\mathbf{V_R}$ to represent measurement error. The final $\mathbf{V_R}$ thus becomes:

$$\begin{array}{c} \\ X \\ R \\ Y \end{array}\begin{array}{c} \begin{array}{ccc} X & R & Y \end{array} \\ \begin{bmatrix} \sigma^2(t_1+t_2+t_3)+\mathrm{SE} & \sigma^2 m t_1 & 0 \\ \sigma^2 m t_1 & \sigma^2(t_1+t_2+t_3)+v_H+\mathrm{SE} & \sigma^2(1-m)(t_1+t_2) \\ 0 & \sigma^2(1-m)(t_1+t_2) & \sigma^2(t_1+t_2+t_3)+\mathrm{SE} \end{bmatrix} \end{array}$$

Given traits $y_1, y_2, \cdots, y_n$ (in log scale) for n species some of which are hybrids, the column vector $\mathbf{Y} = (y_1, y_2, \cdots, y_n)^T$ can be treated as a multivariate normal random variable given an assumption of Brownian motion. i.e. $\mathbf{Y} \sim MVN(\boldsymbol{\mu}, \mathbf{V_R})$ where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^T$ and $\mu_i = \mu$ or $\mu + \log\beta$ is the mean for usual species or the mean for hybrid species, respectively, and $\mathbf{V_R}$ is calculated as above given the tree (with branch lengths) $\boldsymbol{T}$, structure of gene flow $\boldsymbol{D}_{t_d,t_r,m}$, and parameters $\mu, \sigma^2, v_H,$ and β. The negative log likelihood function given these is

$$-\log L\left(\mu, \sigma^2, v_H, \beta \middle| \mathbf{Y}, \boldsymbol{T}, \boldsymbol{D}_{t_d,t_r,m}\right) = \frac{n}{2}\log 2\pi + \frac{1}{2}\log|\mathbf{V_R}| + \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{V_R}^{-1}(\mathbf{Y}-\boldsymbol{\mu})$$

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where $|\mathbf{V_R}|$ is the determinant of $\mathbf{V_R}$ and $(\mathbf{Y}-\boldsymbol{\mu})^T$ is the transpose of $(\mathbf{Y}-\boldsymbol{\mu})$. $\mathbf{D}_{t_d,t_r,m}$ contains information on gene flow. It must be provided by the user, and indicates donors, recipients, and the time of hybridization events.

Numerical problems

It is known that the variance covariance matrix of the phylogeny can be ill-conditioned: the matrix can be effectively singular, which makes dealing with its inverse, as required to calculate the likelihood, numerically difficult. Anecdotally, this seems to occur more frequently for phylogenetic networks than trees. One approach would be to just prohibit analyses if this is the case. This is practically problematic: a biologist spending years gathering data and a tree could find her or his analysis stymied just due to the structure of the network. Instead, we opted for an approximate solution a user may choose to apply (though the default is not to do this). We evaluate the condition by taking the variance covariance matrix and calculating the determinant of it, of it multiplied by 1000, and of it multiplied by 0.0001. If all three determinants are positive, the matrix is more likely to be well-conditioned, no matter the parameter estimates (which, in the case of σ^2, v_H and SE, may change the magnitude of entries). If it fails this test, we lengthen terminal branch lengths slightly and try again, repeating until the maximum number of tries (by default, 100, causing the terminal branch lengths, likely to be in units of millions of years, to be increased by just 0.001). If it is still ill-conditioned, the analysis aborts; otherwise, it continues, but will only return an approximate answer. This adjustment is done to the original tree, before fitting parameters; regardless of whether that was done, the matrix may also be poorly conditioned for some parameter combinations, resulting in numerical errors in calculation of likelihood. During a

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

run, the log of the condition number of the final variance covariance matrix is measured and compared to a set precision; if this value exceeds the precision, numerical problems may ensue. We approximate what the likelihood would be given the problematic variance covariance matrix by calculating the likelihood with a series of better conditioned ones (by decreasing the magnitude of off diagonal elements while increasing the magnitude of diagonal elements), and then predicting the likelihood with no such transformation from a smooth spline extrapolation of the likelihood from matrices with decreasing magnitude of transformations. This approximation is turned off by default, but can be helpful if the maximum likelihood estimates occur in a problematic region. The transformation of the mean vector proceeds similarly.

Assessing the general performance of the model

We assessed the performance of our model, varying (i) the number of non-hybrid taxa (30 or 100), (ii) the number of hybrids (1, 5, or 10), (iii) the proportion of flow of hybrid inherited from the parents (all from one parent, 10% from one and 90% from the other, or equally from both parents), (iv) the structure of hybridization, (v) the value of β (0.1, 1, or 10), and (vi) the value of v_H (0, 10, or 100). For each replicate we simulated a bifurcating tree of 30 or 100 taxa using TreeSim (Stadler 2009), with a birth rate of 1, death rate of 0.5, sampling frequency of 0.5, and tree height of 50. We then added 1, 5, or 10 species of hybrid ancestry to the tree in one of two ways. The first was to attach those taxa randomly around the tree, but forcing each to arise from its own hybridization event: that is, no hybrid species subsequently speciated. The second was to have just one hybridization event on the tree, and have this lead to the observed number of species of hybrid origin through speciation of the

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

original hybrid. Different simulations differed in the ratio of genes coming from the two parents: it could be all from one parent (a flow of 0: this is equivalent to a tree model); 10% from one parent and 90% from the other; or 50% from each parent. Other parameter values fixed in the model were $\mu = 1,\ \sigma = 0.01$, and SE = 0. All simulations were carried out by the *BMhyd* package. After the runs were done, we looked at deviation (absolute value of the difference between the observed and true parameter value, divided by the true parameter value (if it is nonzero)). We regressed this for each parameter of interest against the values used in our simulation: tree type, flow magnitude, number of non-hybrid taxa, number of hybrid taxa, and how a summary of the parameters could be calculated (using only the best of four models, doing model-averaging across all four models, or doing model-averaging across the models with v_H fixed). The two 3-state values (flow and summary approach) were converted into two binary variables each. The package MuMIn (Barton 2015) was used to estimate the Akaike importance and coefficients for each of these parameters. The goal of this was to allow us to focus figures and discussion on the simulation parameters that had a strong effect on the relative error, rather than try to plot the seven dimensions of variation used in the simulations.

Empirical test cases

Simulations are essential in examining a new method to verify that it is working well enough in choosing models and estimating parameters. However, it can also be useful to run empirical datasets, both to make new discoveries using a new method and to verify that a method operates smoothly on real, messy data. Unfortunately, as the true model or parameter

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

estimates are not known from empirical data, information about accuracy may only come from the simulations, but empirical results can show other problems. In this paper, we look at hybridization in two examples: cichlid body size evolution, using a network from Kobmüller et al. (2007) and data from FishBase (Froese and Pauly 2010), and tobacco and relatives, using a network from Chase et al. (2003) and drought tolerance data from Komori et al. (2000).

Cichlids

Cichlids are notorious for widespread hybridization; their phylogeny is difficult due to presumed hybrid origin or ongoing gene flow. In fact, some may be going extinct due to merging through hybridization (Rhymer and Simberloff 1996). They thus reflect a good test case for this method. Kobmüller et al (2007) developed a phylogeny for cichlids which, importantly, included information about hybrid species and their presumed direction of ancestry. To replicate their tree, we downloaded their sequences from GenBank (Benson et al. 2005). The sequences were aligned by MAFFT (Katoh and Standley 2013) and subsequently inspected in Mesquite (Maddison and Maddison 2011) to trim ends of sequences for only a small subset of taxa. A backbone constraint was made from Kobmüller et al (2007)'s overall hybrid tree in Mesquite and used in all subsequent searches. The aligned sequences were then analyzed under the software PAUP (Swofford 2003) for parsimony tree search. The searched best-rooted tree was taken and was used to set to likelihood of GTR + Γ substitution model with clock. We filtered for best under this model and then did a search limited to 10 hours to resolve branch lengths. Trait values (total body

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

length of cichlids) were collected from FishBase (Froese and Pauly 2010) using the R package rfishbase (Boettiger et al. 2012).

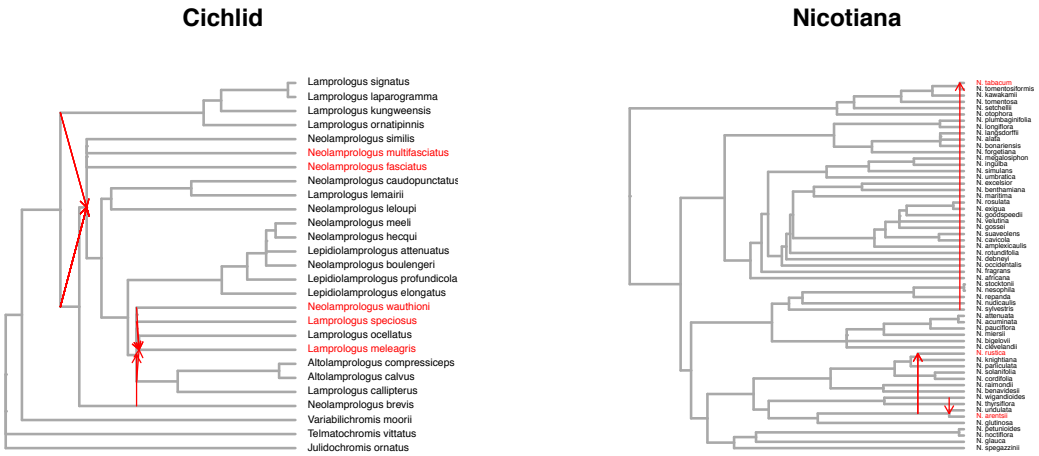
In Kobmüller et al. (2007), the cichlid data contains 27 species where five species are putative hybrids. Three of the five hybrid species (*Neolamprologus wauthioni*, *Lamprologus speciosus*, and *Neolamprologus fasciatus*) are inferred to have arisen due to mating between extant species and two of them (*Lamprologus meleagris*, *Neolamprologus multifasciatus*) are inferred to be formed as a result of hybridization between extinct lineages. As in Fig. 1, this results in the gene flow appearing to be forward in time from at least one relative of the parental species.

Nicotiana

This group contains tobacco and relatives. Their relationships were long suspected to be reticulate (Godspeed 1954), and this was supported by Chase et al. (2003) in a work based on internal transcribed spacer region (ITS) and in situ hybridization. We followed the same procedure as for the cichlid dataset in returning a chronogram, except that we did not use Chase et al.'s parsimony trees as constraints. The crown age was set to 15.3 MY, following Clarkson et al. (2005). Taxa of hybrid origin and the placement of hybridization events were pulled from Chase et al.'s results; timing of events came from branch lengths on the chronogram, where the donor and recipient times were set to be equal (thus, no postulate of extinct intermediate hybrid parents) and to occur at the origin of the hybrid taxon. The relative seedling growth under mannitol treatment dataset from Komori et al. (2000) was

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

277 extracted from their Table 2. We note that this number is a proportion, thus not quite meeting
278 the expectations of Brownian motion (unbounded traits); we log transformed it, but this is
279 still an imperfect fix. We used iPlant TNRS (Boyle et al. 2013) to convert the taxon names
280 from both datasets to the same taxonomy, and Geiger (Harmon et al. 2008) to prune the tree
281 and data to the same taxon set. Figure 2 represent the evolutionary tree and the gene flow for
282 both cichlid and tobacco. The donor-recipient relationship among the hybrids and their
283 parents in the cichlid and *Nicotiana* datasets can be found in Supplemental material.
284



285

286 Figure 2: The empirical networks. Species in red are of putative hybrid origin. Red arrows
287 show movement from a parent to a new hybrid lineage; in cases where only one arrow is
288 shown leading to a lineage, it is because the hybrid lineage comes from its sister species on
289 the tree plus the source of the arrow. Arrows appearing to move forward in time show
290 transfer via an unsampled lineage (see explanation on Fig. 1).
291
292

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

293 *Model Selection and Parameter Estimation*

294

295 We tried four different models for each empirical dataset. All fix the gene flow $m = 0.5$ and
296 allow μ, σ^2, and SE to be optimized. They differ in the settings for mean change in the
297 hybrid β and the hybrid variation at formation v_H. Model 1 fixes β at 1 but allows v_H to
298 vary; Model 2 allows β to vary but fixes v_H at 0; model 3 fixes β at 1 and v_H at 0 and model
299 4 allows both to vary. We fit those models to both cichlid and tobacco datasets.

300

301 *Adaptive confidence intervals sampling*

302 Uncertainty in parameter estimates can be substantial. One way of estimating this can be
303 looking at the curvature of the surface at the maximum likelihood optimum, but this is
304 known to be problematic when the likelihood function is not regular (Pawitan 2013). A
305 different approach, advanced by (Edwards 1992) is to look at a confidence region of all
306 points that generate a log likelihood within a certain range (often, set to be a delta of 2 log
307 likelihood units) of the maximum likelihood. One approach to calculate this would be to
308 vary each parameter on its own while holding the others constant. This is convenient and fast
309 to implement, but can result in artificially small confidence intervals. For example, if two
310 parameters *a* and *b* covary such that the likelihood is the same as long as $a = 0.7b$, the
311 likelihood changing just *a* or just *b* would drop off very quickly, but there's a ridge
312 containing a wide array of *a* and *b* values that would not affect the likelihood. Thus, we
313 chose to examine varying all parameters at once, so that if there is a ridge or other structure
314 for the likelihood surface we do not overestimate our certainty. While there are many
315 algorithms to find the peak of a surface, there are fewer to find the entirety of a region two

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

log likelihood units below the peak. We thus developed a Monte Carlo method to estimate this. We start by simulating points using a multivariate uniform centered on the maximum likelihood estimates. The likelihood at each of these points is calculated. The algorithm periodically checks to make sure half the points are within the region and half are outside. If too many are within the cutoff of the peak likelihood, there is not good enough sampling of the boundaries of the confidence region and the sampling width is increased; if there are too many that have values too far from the optimal likelihood, the sampling width decreases. For a given parameter value, we thus calculate the likelihood over a range of values for the other parameters, giving a more realistic, less conservative confidence interval. Note, however, that this merely examines uncertainty due to flatness of the likelihood surface: there can be substantial additional sources of uncertainty from tree topology or branch length uncertainty, problems with measurements beyond what a fixed measurement error can address, or other issues.

Software and Data

We have implemented this model in R (R team 2015) in the *BMhyd* package (on CRAN). It is open source, and includes functions for fitting models on networks, visualizing gene flow on networks, and even simulating random networks. It uses functions or code from *Geiger* (Harmon et al. 2008), *phytools* (Revell 2012), TreeSim (Stadler 2014), ape (Paradis 2004). All relevant R code and data files in this work can be found at Dryad Digital Repository LINK NEED.

RESULTS

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

339 *Simulation for Assessing General Performance of Models*

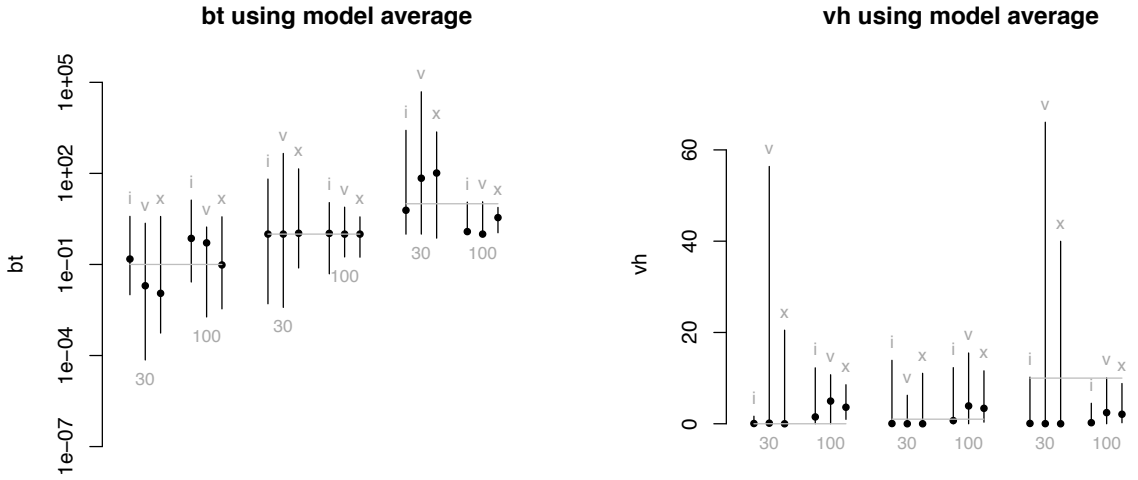
340

341 Overall we were not pleased with performance: looking at just the simulations with the best
342 chance for good results (flow rate of 0.5, 100 non-hybrid taxa, 10 hybrid taxa) there was often
343 quite poor correlation between the estimates of parameters v_H and β and the true values (SE, μ,
344 and σ^2 performed far better, but they are not interesting parts of our model). Experimenting with
345 the results suggested that model averaging or just taking the parameter estimates from the best
346 model had about the same performance; using models where only β varied performed better than
347 models where v_H varied. We investigated this, and other elements of the simulation, by doing a
348 regression where we found importance and coefficients of parameters by dredging (using all
349 subsets of the global model, and using Akaike weights to calculate importance). This suggested
350 that choosing model-averaged or best model only parameter estimates did not matter much
351 (importance of 0.27) but using models with v_H fixed was important for estimating v_H
352 (importance of 1.00), less so for β (importance of 0.27) [and important for the other three free
353 parameters], but the sign of the coefficient, negative for both v_H and β, suggested that limiting
354 models to those with v_H fixed reduced error in both β and, oddly, v_H. Tree type was not
355 important. This surprised us: for ten hybrid taxa, having them the outcome of ten independent
356 hybrid events should give much more information about the hybridization process than ten
357 hybrids descended from one event in the past. However, inspecting the parameter estimates
358 coming from trees with each simulation scenario also suggested they performed very similarly.
359 Summaries later in the paper thus merge trees of both types. Flow and number of hybrid and
360 nonhybrid taxa were found to be important.

361

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

362 Model-averaged parameter estimates for the most relevant parameters β and v_H in simulations
363 are shown in Fig.3; following the information on importance, above, we grouped results
364 regardless of hybridization type, and only show results with equal flow from both parent species
365 (though results were similar with flow of 10% from one parent and 90% from the other, as well
366 as, oddly, flow from only one parent). The estimates for all parameters, using model averages
367 from the set of all models and from the set of models that had v_H fixed at zero, are shown in
368 Supp Fig 1.
369
370



371
372 Figure 3 caption: Shown are estimates for β (left) and v_H (right) under the simulations. Gray bar
373 shows the true values, dot shows the median model averaged estimate across simulations, and
374 error bars show the range for 95% of the simulations. “i”, “v”, and “x” represent simulations
375 with 1, 5, or 10 hybrid taxa; 30 or 100 refers to the number of non-hybrid taxa. Note β estimates
376 often centered on the true value, while v_H was often very wrong.

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

377

378 Even for β, however, there is cause for concern. Though the median estimates across simulations
379 were fairly close to the true values, the range across simulation replicates was still quite extreme.
380 For example, for 100 taxon trees with 10 hybrids, the "best case" scenario we examined, if β was
381 truly 1.0, indicating no expected bias between the hybrid and the mean of its parents, estimates
382 of beta ranged from 0.13 to 5.14; that is, if we were looking at an organism whose parent species
383 were each 100g, and we used log(mass) as the trait undergoing Brownian motion, the expected
384 mass of the hybrid would be anywhere from 13g (=exp(log(100) + log(0.13)) to 514g according
385 to the estimates. Across the range of the 675 simulations that completed (some with 30
386 nonhybrid taxa and only one hybrid taxon), where the true value of β was 1, the expected mass
387 of a hybrid of two 100g species based on the estimated β could be anywhere from 0.00001 g to
388 10,707,208 g.

389

390 *Assessing Model Identifiability through Jointly Estimating Parameters*

391 The shape of the likelihood surface provides the ability to estimate parameter values: if the
392 surface is flat, there is little support for a parameter estimate. In some cases, like the trend
393 parameter for Brownian motion with a trend for coeval taxa, no amount of data is adequate to
394 estimate the parameter: this parameter is formally non-identifiable. There is also a softer
395 definition of identifiability: given a particular dataset, is there enough data to estimate a
396 parameter. We investigated both of these by creating contour maps of the likelihood surface for
397 pairs of parameters under the cichlid and *Nicotiana* data sets. Results are shown in Figure 4. For
398 these empirical datasets, parameters appear distinguishable: there are no ridges in the likelihood
399 surface, even though the confidence intervals are wide. Thus, the parameters are formally

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

identifiable. The large confidence intervals, though, suggest that they can often be practically problematic.

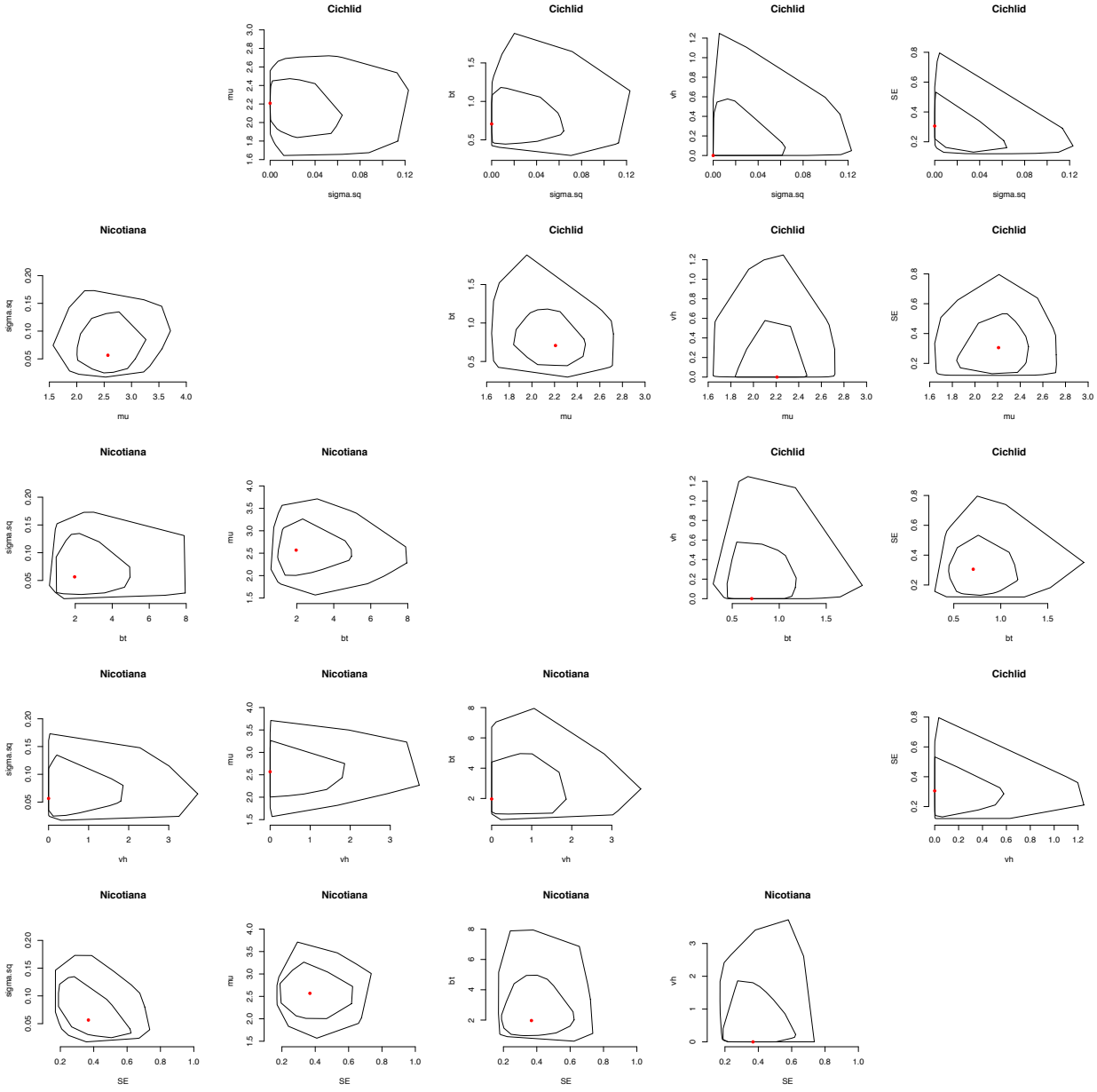


Figure 4: Likelihood surfaces for pairs of traits. Results from cichlids are shown above the diagonal, *Nicotiana* below. The red dot represents the maximum likelihood estimate; the inner

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

406 circle shows the Δ2 log likelihood unit region, and the outer shows the Δ5 log likelihood unit
407 region. Note the lack of ridges and the wide intervals.
408
409 *Model Selection and Parameter Estimation for the Empirical Data*
410
411 The best model for cichlids had no burst of variation at hybridization events nor a bias in hybrid
412 size (Table 1). In models where β was allowed to vary, the confidence interval included 1, and
413 for v_H, the confidence interval (and MLE in all models) included 0, again suggesting lack of
414 evidence for evolutionary speed ups or mean changes with hybridization. Measurement error was
415 estimated to be substantial; in fact, the model estimated no effective Brownian motion at all (σ^2
416 estimates of 0) with all observed variance being due to just measurement error at the tips. For
417 *Nicotiana*, the best model (but only with 0.39 of the Akaike weight) had β as a free parameter
418 and v_H constrained to be 0. There is weak evidence from this model that beta is greater than one
419 (point estimate from best model is 1.97, but CI is 0.89 to 4.52; model averaged estimate is 1.49),
420 suggesting that hybrids have higher success rates as seedlings under drought conditions than do
421 their parents. There is again little evidence for increased variance at hybridization events. Given
422 the tree height and its σ^2 rate, we expect variance at the tips to be 0.50; from measurement error,
423 there is an additional 0.37 variance (both in units of log((seedling survival)^2)), suggesting
424 meaningful Brownian motion on the tree but still quite important measurement variance.
425

Dataset	vH parameter	β parameter	NegLogL	K	deltaAICc	AkaikeWeight	σ2	μ	β	vh	SE
Cichlid	**fixed**	**fixed**	**23.077**	**3**	**0.000**	**0.524**	**0 (0, 0.095)**	**2.152 (1.741, 2.466)**	**1 (fixed)**	**0 (fixed)**	**0.324 (0.091, 0.567)**
	free	fixed	23.077	4	2.775	0.131	0 (0, 0.09)	2.152 (1.759, 2.418)	1 (fixed)	0 (0, 0.989)	0.323 (0.112, 0.556)
	fixed	free	22.304	4	1.230	0.283	0 (0, 0.076)	2.216 (1.813, 2.501)	0.708 (0.434, 1.205)	0 (fixed)	0.306 (0.112, 0.536)
	free	free	22.307	5	4.275	0.062	0 (0, 0.06)	2.208 (1.848, 2.482)	0.707 (0.424, 1.225)	0 (0, 0.545)	0.306 (0.132, 0.497)
Cichlid model average							0 (0, 0.087)	2.174 (1.77, 2.471)	0.899 (0.804, 1.072)	0 (0, 0.163)	0.317 (0.102, 0.552)
Nicotiana	fixed	fixed	72.365	3	0.134	0.365	0.048 (0.012, 0.112)	2.591 (1.955, 3.232)	1 (fixed)	0 (fixed)	0.424 (0.244, 0.767)
	free	fixed	72.275	4	2.258	0.126	0.053 (0.019, 0.128)	2.581 (1.955, 3.171)	1 (fixed)	0.229 (0.152, 5.143)	0.392 (0.195, 0.713)
	fixed	**free**	**71.146**	**4**	**0.000**	**0.390**	**0.057 (0.014, 0.149)**	**2.569 (1.895, 3.252)**	**1.968 (0.889, 4.529)**	**0 (fixed)**	**0.367 (0.178, 0.681)**
	free	free	71.146	5	2.391	0.118	0.056 (0.025, 0.148)	2.568 (1.916, 3.24)	1.967 (0.827, 4.286)	0 (0, 1.969)	0.368 (0.191, 0.623)
Nicotiana model average							0.053 (0.015, 0.133)	2.578 (1.927, 3.233)	1.492 (0.936, 2.766)	0.029 (0.019, 0.882)	0.391 (0.206, 0.71)

426

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TABLE 1: Model fitting for the cichlid and tobacco data. In this table, NegLogL is the negative log likelihood of the model, K is the number of free parameters, ΔAICc is calculated by subtracting the AICc value from the lowest AICc value among the four models, Akaike weight is calculated by $w = \exp(-0.5\Delta\text{AICc})$ and then normalizing the weights. Parameter estimates and their adaptive confidence intervals are reported. The model-averaged parameter estimate $\hat{\bar{\theta}} = \sum_{i=1}^{4} w_i \hat{\theta}_i$ are reported where w_i is the Akaike weight for the *i*th model.

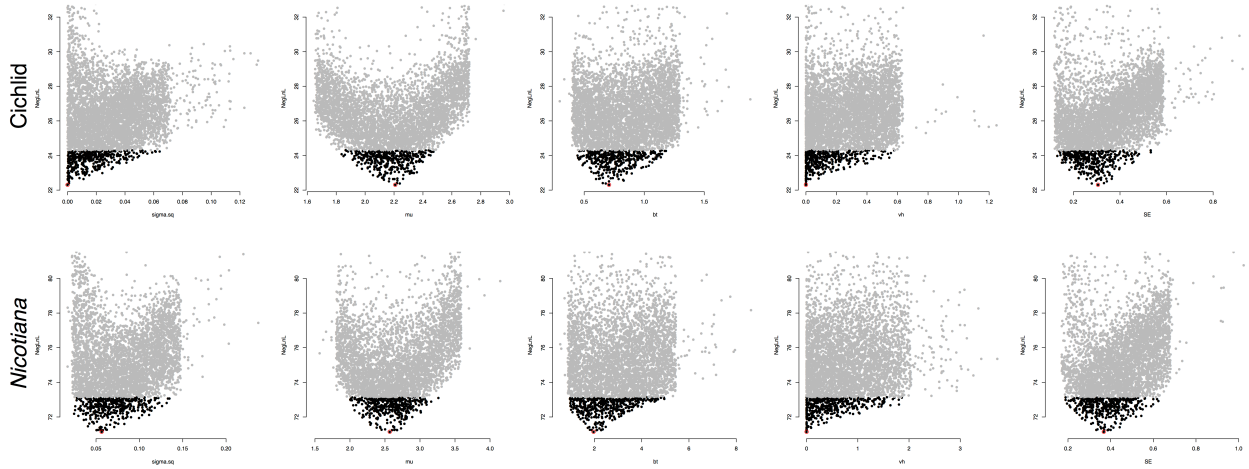


Figure 5. The adaptive confidence intervals for the free parameters of the best model for cichlids (top) and for *Nicotiana* (bottom). In each plot the vertical axis represents the negative log likelihood value and the horizontal axis represents a wide parameter range. The red dot at the lowest y-axis value represents the MLE estimated from the overall search. Sampling works by proposing a set of parameters and estimating the likelihood for this set. This likelihood value is then plotted versus each of the parameters used in that set in a different subplot as a gray or black dots. Dots in black represent the desired likelihood value

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(taken as those no more than 2 log likelihood units away from the maximum) and the width of the adaptive confidence intervals are measured by the left most and right model black dots.

DISCUSSION

Our new approach allows analysis of trait data on a phylogenetic network. We can now use comparative methods on general networks rather than just trees and it works well for estimating general Brownian motion parameters such as evolutionary rate σ^2 and root state μ. There are also hybridization-specific parameters where it performs variably. The method can estimate hybridization bias β, the consistent increase or decrease in trait value upon hybridization that may lead to hybrid vigor or outbreeding depression, though with substantial uncertainty. It performs surprisingly poorly for estimating a burst of variance associated with hybridization v_H; we would recommend caution interpreting any estimates of v_H. Unfortunately, this may be the most biologically interesting question.

Our empirical results for the cichlid dataset do not provide great biological insights, other than 1) the feasibility of running a model given a network and 2) quite extensive measurement uncertainty in the body length measurements. This latter could reflect real measurement uncertainty (fish have indeterminate growth (Dutta 1994)), so the notion of a true species mean for this trait is problematic) but errors in the tree topology or branch lengths would tend to result in this appearing as measurement error in this model, as well. *Nicotiana* also had substantial

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

467 measurement error, but not enough to wipe out the phylogenetic history. The results suggest that
468 hybrids perform better in droughts than their parent species, though this is not statistically
469 significant given the confidence interval. However, it might point the way to further studies
470 about drought tolerance, an area that will be of increasing importance.
471
472 Several approaches have been proposed for inferring different rates along the branch for a given
473 phylogenetic tree (McPeek, 1991; O'Meara et al., 2006; Revell, 2008; and Beaulieu et al. 2012),
474 and it would be useful to extend our work to allow for this heterogeneity. Another possible
475 extension is to use a more parameter-rich Markovian process. The model can be extended to
476 allow trait evolution following the OU process (Hansen 1997; Butler and King 2004; Beaulieu et
477 al. 2012). Putting the approach in a Bayesian context is also possible. In this case, parameters of
478 multiple selective regimes, multiple rates of variation, and multiple rates of constraining forces
479 could be embedded in the model. Developing a more complex model of this type could be very
480 useful when analyzing fairly large data sets of hundreds of species or more, where heterogeneity
481 is expected and there may be power to provide estimates for many parameters.
482
483 In nature, approximately 10% of animal species and 25% of plant species hybridize (Mallet 2005,
484 2007), suggesting that there is widespread gene flow between "species." Some of this gene flow
485 may lead to hybrid speciation in the manner assumed in our method, and hybrid speciation is
486 widely suspected in many groups (Arnold 1996, Welch and Riesberg 2002). However, even in
487 the absence of hybrids formed from two distinct parent species, such ongoing gene flow suggests
488 a need for a network metaphor, as suggested by Morrison (2014). Our method cannot currently
489 deal with this sort of gene flow: we represent the hybrid as being the result of a single event

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

between two parent lineages (though we do allow for one or both parents to be missing from the tree, making the event appear as if it is going forward in time from the nearest sampled relative(s)). Gene flow over continuous time periods is thus not modeled yet, though it would be a basic extension.

This approach, especially the creation of the modified variance covariance matrix given hybridization and the potential for a modified matrix of expected species values, could form the core for multivariate approaches, in the same way the traditional Brownian motion tree model lies at the heart of methods as various as PGLS (Martins and Hansen 1997), PGLM (Ives and Helmus 2011), independent contrasts (Felsenstein 1985, 2008), phylogenetic linear regression (Ho and Ané 2014) and more. It is also trivial to use the existing approach to estimate ancestral states on a network. While network inference methods are advancing, it is important to make sure that comparative methods using these networks keep pace, of which this work is a start that we hope can be built upon.

ACKNOWLEDGEMENTS

We thank Luke Harmon and an anonymous reviewer for their suggestions that improves this manuscript. We also thank the National Institute for Biological and Mathematical Synthesis (NIMBioS), an Institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Award #EF-0832858, with additional support from The University of Tennessee, Knoxville for funds for a postdoc for

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

513 D-C J and for summer visitor support, as well as a grant to D-C J from National Science Council

514 grant of Taiwan, ROC, NSC-102-2118-M-035-004.

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

Ainouche M.L., Fortune P.M., Salmon A., Parisod C., Grandbastien M.-A., Fukunaga, K., Ricou M., Misset M.-T. 2008. Hybridization, polyploidy and invasion: Lessons from Spartina (Poaceae). Biological Invasions 11: 1159-73.

Arnold M.L. 1996. Natural hybridization and evolution. Oxford University Press, New York, USA.

Barton K. 2015. MuMIn: Multi-Model Inference. R package version 1.15.1.

Beaulieu M.J., Jhwueng D.-C., Boettiger C., O'Meara B.C. 2012. Modeling stabilizing selection: Relaxing the Ornstein-Uhlenbeck model of adaptive evolution. Evolution 66: 2369-2383.

Benson D.A., Karsch-Mizrachi L., Lipman D.J., Ostell J., and Wheeler D.L. 2005. GenBank. Nucleic Acids Res. 33:D34-D38.

Boettiger C, Land D.T., and Wainwright P.C. 2012. rfishbase: exploring, manipulating and visualizing FishBase data from R. Journal of Fish Biology 81:2030-2039.

Boyle B., Hopkins N., Lu Z., Garay J.A.R., Mozzherin D., Rees T., Matasci N., Narro M.L., Piel W.H., Mckay S.J., Lowry S., Freeland C., Peet R.K., Enquist B.J. 2013. The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics. 14:16. Doi:10.1186/1471-2105-14-16.

Bryant D., Moulton V. 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255-265.

Burham K.P., Anderson D.R. 2004. Model selection and multimodel inference. Springer-Verlag New York.

Butler M.A., King A.A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. The American Naturalist 164:683-695.

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

538 Cardillo M., Mace G.M., Gittleman J.L., Purvis A. 2006. Latent extinction risk and the future
539 battlegrounds of mammal conservation. Proceedings of the National Academy of Sciences.
540 103(11) 4157-4161.
541 Chase M.W., Knapp S., Cox A.V., Clarkson J.J., Butsko Y., Joseph J., Savolainen V., and
542 Parokonny A.S. 2003. Molecular systematics, GISH and the origin of hybrid taxa in
543 *Nicotiana* (Solanaceae). Annals of Botany 92: 107-127.
544 Clarkson J.J., Lim K.Y., Kovarik A., Chase M.W., Knapp S. and Leitch A.R. 2005. Long-term
545 genome diploidization I allopolyploid *Nicotiana* section Repandae(Solanaceae). New
546 Phytologist 168:241-252.
547 Doolittle W.F. 1999. Phylogenetic classification and the universal tree. Science 284: 2124-9.
548 Dray S., Dufour A.-B. 2007.The ade4 package: implementing the duality diagram for ecologists.
549 Journal of Statistical Software. 22:1-20.
550 Duftner N., Koblmüller S., Sturmbauer C. 2005. Evolutionary relationships of the
551 Limnochromini, a tribe of benthic deepwater cichlid fish endemic to Lake Tanganyika, East
552 Africa. J Mol Evol 2005, 60:277-289.
553 Dutta H. 1994. Growth in fishes. Gerontology. 40:97-112.
554 Edward A.W.F. 1992. Likelihood. Expanded Edition. Johns Hopkins University Press.
555 Felsenstein J. 1985. Phylogenies and the comparative method. The American Naturalist 125:1-15.
556 Felsenstein J. 2008. Comparative methods with sampling error and within-species variation:
557 contrast revisited and revised. The American Naturalist 171: 713-725.
558 Froese R., Pauly D. 2010. FishBase. World Wide Web electronic publication. www.fishbase.org.
559 Gittleman J.L., Kot M. 1990. Adaptation: statistics and a null model for estimating phylogenetic
560 effects. Systematic Zoology 39: 227-241.

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

561 Godspeed T.H. (1954) The genus *Nicotiana*. Chronica Botanica Company, Waltham, MA.

562 Grafen A. 1989. The phylogenetic regression. Philosophical Transactions of the Royal Society of
563 London. Series B, Biological Sciences 326:119-157.

564 Hansen T.F., Martins E.P. 1996. Translating between microevolutionary process and
565 macroevolutionary patterns: the correlation structure of interspecific data. Evolution
566 50:1404-1417.

567 Hansen T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution
568 51:1341-1351.

569 Harmon L.J., Weir J.T., Brock C.D., Glor R.E., and Challenger W. 2008. GEIGER: investigating
570 evolutionary radiations. Bioinformatics 24:129-131.

571 Ho L.S.T. and Ané C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait
572 evolution models. Systematic Biology. 63:397-408.

573 Huson D.H., Bryant D. 2006. Application of Phylogenetic Networks in Evolutionary Studies,
574 Mol. Biol. Evol. 23:254-267.

575 Huson D.H., R. Rupp, Scornavacca C. 2010. Phylogenetic networks: concepts, algorithms and
576 applications. Cambridge University Press.

577 Ives A.R. and Helmus M.R. 2011. Generalized linear mixed models for phylogenetic analyses of
578 community structure. Ecological Monographs. 81:511-525.

579 Joly S., McLenachan P.A., Lockhart P.J. 2009. A statistical approach for distinguishing
580 hybridization and incomplete lineage sorting. American Naturalist 174:E54-70.

581 Katoh K. and Standley D.M. 2013. MAFFT Multiple sequence alignment software version 7:
582 improvements in performance and usability. Molecular Biology and Evolution. 30:772-780.

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

583 Koblmüller S., Duftner N., Sefc K.M., Aibara M., Stipacek M., Blanc M., Egger B., Sturmbauer
584 C. 2007. Reticulate phylogeny of gastropod-shell-breeding cichlids from Lake Tanganyika:
585 the result of repeated introgressive hybridization. BMC Evolutionary Biology 7:7.
586 Komori T., Myers P.N., Yamada S., Kubo T., and Imaseki H. 2000. Comparative study of the
587 *Nicotiana* species with respect to water deficit tolerance during early growth. Euphytica
588 116:121-130.
589 Kubatko L.S. 2009. Identifying hybridization events in the presence of coalescence via model
590 selection. Syst. Biol. 58:478-488.
591 Linder C.R., Rieseberg L.H. 2004. Reconstructing patterns of reticulate evolution in plants.
592 American Journal of Botany 91:1700-1708.
593 Maddison, W.P. and Maddison D.R. 2011. Mesquite: a modular system for evolutionary analysis.
594 Version 2.75 http://mesquiteproject.org
595 Maddison, W.P., Midford, P.E., and Otto, S.P. 2007. Estimating a binary character's effect on
596 speciation and extinction. Systematic Biology 56(5) 701-710.
597 Mallet J. 2005. Hybridization as an invasion of the genome. Trends in Ecology and Evolution
598 20:229-237.
599 Mallet J. 2007. Hybrid speciation. Nature 445:279-283.
600 Martins E.P. and Hansen T.F. 1997. Phylogenies and the comparative method: a general
601 approach to incorporating phylogenetic information into the analysis of interspecific data.
602 The American Naturalist. 149:646-667.
603 McPeek M.A. 1995. Testing hypotheses about evolutionary change on single branches of a
604 phylogeny using evolutionary contrasts. American Naturalist 145:686-703.

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

605 Meng C., Kubatko L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage
606 sorting using gene tree incongruence: a model. Theoretical Population Biology 75:35-45.
607 Moret B.M.E., Nakhleh L., Warnow T., Linder C.R., Tholse A., Padolina A., Sun J., Timme. R.
608 2004. Phylogenetic networks: modeling, reconstructibility, and accuracy. IEEE/ACM
609 Transactions on Computational Biology and Bioinformatics 1:13-23.
610 Morrison D.A. 2014. Is the tree of life the best metaphor, model, or heuristic for phylogenetics?
611 Systematic Biology 63:628-638.
612 Nakhleh L. 2011. Evolutionary phylogenetic networks: model and issues. Problem solving
613 handbook in computational biology and bioinformatics. P125-158.
614 O'Meara B.C., Ané C., Sanderson M.J., Wainwright P.C. 2006. Testing for different rates of
615 continuous trait evolution using likelihood. Evolution 60:922-933.
616 Otto S.P., Whitton J. 2000. Polyploid incidence and evolution. Annual Review of Genetics 24:
617 401-437.
618 Paradis E, Claude J. & Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R
619 language. Bioinfromatics 20:289-290.
620 Pawitan Y. 2013. In all likelihood: statistical modeling and inference using likelihood. Oxford
621 University Press.
622 R Development Core Team. 2008. R: A language and environment for statistical computing. R
623 foundation for statistical computing, Vienna, Austria.
624 Revell L.J. 2008. On the analysis of evolutionary change along single branches in a phylogeny.
625 American Naturalist 172: 140-147.
626 Revell L.J. 2012. phytools: An R package for phylogenetic comparative biology(and other
627 things). Methods Evol. Evol. 3:217-223. doi:10.1111/j.2041-210X.2011.00169.x

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

628 Rieseberg L.H., Archer M.A., Wayne R.K. 1999. Transgressive segregation, adaption and
629 speciation. Heredity 83: 363-372.
630 Rieseberg L.H. 2006. Hybrid speciation in wild sunflowers. Ann. Missouri Bot. Gard. 93: 34-48.
631 Rhymer J.M. and Simberloff D. 1996. Extinction by hybridization and introgression. Annual
632 Review of Ecology and systematics 27:83-109.
633 Sang T., Zhong Y. 2000. Testing hybridization hypotheses based on incongruent gene trees.
634 Syst. Biol. 49:422-434.
635 Schluter, D., Price T., Mooers A.O., Ludwig D. 1997. Likelihood of ancestor states in adaptive
636 radiation. Evolution 51(6): 1699-1711.
637 Stadler T. TreeSim: simulating trees under the birth-death model. R package version 2.1.
638 Swofford, D.L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony(* and Other Methods).
639 Version 4. Sinauer Associates, Sunderland, Massachusetts.
640 Wang J, Guo M, Xing L, Che K, Liu X, Wang C, 2013. BIMLR: A method for constructing
641 rooted phylogenetic networks from rooted phylogenetic trees. Gene 527:344-351.
642 Weigel D.E., Peterson J.T., Spruell P. 2002. A model using phenotypic characteristics to detect
643 introgressive hybridizations in wild west slope cutthroat trout and rainbow trout.
644 Transactions of the American Fisheries Society 141:389-403.
645 Welch ME and Riesberg LH. 2002. Habitat divergence between a homoploid hybrid sunflower
646 species, Helianthus paradoxus (Asteraceae), and its progenitors. American Journal of Botany
647 89:472-478.
648 Willson SJ, 2013. Reconstruction of Certain Phylogenetic Networks from Their Tree-Average
649 Distances. Bulletin of Mathematical Biology 75:1840-1878.

bioRxiv preprint doi: https://doi.org/10.1101/023986; this version posted August 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

650 Wu Y, 2013. An Algorithm for Constructing Parsimonious Hybridization Networks with
651 Multiple Phylogenetic Trees. Journal of Computational Biology 20:792-804.
652