Running Head: HYBRID TRAIT EVOLUTION

**Trait Evolution on Phylogenetic Networks**

Dwueng-Chwuan Jhwueng1,2, and Brian C. O’Meara3,4

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*

***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 for a Brownian motion model. Our method seeks to separately or jointly detect a bias in trait value coming from hybridization () and a burst of variation at the time of hybridization () associated with the hybridization event, as well as traditional Brownian motion parameters of ancestral state and rate of evolution (), as well as measurement error at 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. The new method is available on CRAN - Package: *BMhyb*.

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

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 structure (Arnold 1996; Doolittle 1999; Otto and Whitton 2000; Linder and Rieseberg 2004; Huson et al. 2010; Nakhleh 2011), and this reality is currently not yet accommodated 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; Solís-Lemus and Ané 2016). We thus stand at a point where phylogenetic networks will increasingly be inferred but we lack comparative methods to properly use this information. There is a notable exception, however: the work of Pickrell and Pritchard (2012), which developed a model for allele frequency data. Our model was developed independently of this earlier model but shares some similarities.

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



Figure 1: Three taxon networks with extant, sampled species X, R and Y. The left plot shows the simplest possible network. Species O leads to two daughter species, A and C. They hybridize leading to new species B after time *t*1+ *t*2 from the root. Species A, B, and C then evolve independently over the next time period (of duration *t*3) to form species X, R, and Y, respectively. The black lines show the “tree” structure, while the gray line shows the gene flow from the hybridization event (of course, the coloring on the AB and CB edges could be switched). The center plot shows a slightly more complicated network. Species A splits at height *t*1 above the root to form the lineage that leads to species X and the lineage that leads to species D and its descendant species E. The hybridization to form B occurs at height *t*1+ *t*2 above the root. However, E is not sampled (perhaps due to extinction or difficulty acquiring it). Thus, the immediate parents of B are not A and C but D and C: the evolutionary changes from A to D are reflected in B but not in X. The right hand plot shows how this would appear to a scientist using this sampling. It appears that genes flow forward in time from A to B, but this is just due to the unsampled lineage.

The center plot in Figure 1 represents a scenario where at height , 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 height , 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 or incomplete sampling it can look like flow forward in time: the shared history of R with X is not from when D exchanged genes (height ) but earlier, height : 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. Under Brownian motion model, the corresponding variance-covariance matrix for the extant species X, R, Y for the tree model (black edges only) is given by the matrix as following

.

Now consider the trait evolution includes the gene flow (dashed arrow in the central plot). Let the trait value of the root state *O* be . By assuming species evolve under Brownian motion (variables are typically measured in log scale in comparative analysis), the trait values at species D and species C are and , respectively, where and can be regarded as error terms that follow a normal distribution with zero mean and variance . Under our model, the hybrid species *B*, at the moment of hybridization assumed the value , can be defined as

.

In particular, follows a normal distribution with mean and variance . The parameter measures the proportion of the hybrid trait value inherited from parent D while measures the proportion the hybrid inherits from parent C (γ 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, might be much closer to one. In particular, indicates inheriting the trait equally from both parents. An extension of the model would allow a different, user-specified at each hybridization event (this can be estimated using population genetics data, but not from a single univariate continuous trait), but for this work, γ is assumed to always be 0.5. Note that while we use γ, which has become common in phylogenetics (i.e. Solís-Lemus and Ané 2016), Pickrell & Pritchard (2012) in their related model use *w*. 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 example, if there exists widespread heterosis, with hybrids being on average 20% larger than their parent species, *β* would be of value 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, , to represent log scale bias for the hybrid at formation. The variance for , , is calculate as

Here we assume that and were in log scale already (i.e. the representation in raw scale is ). To model a process like transgressive segregation, where a hybrid can deviate from the range of parental values but without a particular bias, a non-negative variance is added to lengthen the hybrid branch, equivalent to adding a burst of variation due to the hybridization event. Therefore, we have *V* and *V*. To allow hybrid species to have different rates of evolution than non-hybrid species, one would just require modifying this variance to be *V*, where the new term is the rate of evolution in hybrid species. We currently limit the model to the one where the hybrid and non-hybrid species have the same rate (). The corresponding variance-covariance matrix for the species for the network model under the assumption of the BM process for trait evolution is given by the matrix as following

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |

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, , to the diagonals of to represent measurement error. The final thus becomes:

Note that above construction of the variance-covariance matrix using the phylogenetic network in Figure 1 only assumes that X and Y are independent where the there is no covariation between two species at any time, in particular, . In general, when X and Y share an evolutionary history, the variance of their hybrid at the time of hybridization can be adjusted accordingly by

.

And the covariance of the hybrid and other species at the time of hybridization is

where .

 Given traits (in log scale) for species some of which are hybrids, the column vector can be treated as a multivariate normal random variable given an assumption of Brownian motion. i.e. where and or is the mean for non-hybrid species or the mean for hybrid species, respectively, and **V*R*** is calculated as above given the tree (with branch lengths) ***T****,* structure of gene flow , and parameters . The negative log likelihood function given theseis

where is the determinant of and is the transpose of . contains information on gene flow that indicates donors (), recipients (), and the time of hybridization events ( and , respectively). Note that must be provided by the user.

*Numerical problems*

It is known that the variance covariance matrix of a phylogeny can be ill-conditioned under some conditions: 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. In other fields, there are approaches to deal with this by modifying the variance-covariance matrix (Ardia et al. 2011; Rebonato and Jäckel 2000; Brissette et al., 2007; Higham 2002; Schafer and Strimmer 2005). We have explored using those, with various tests for matrices that perform adequately and trying transformations (lengthening terminal branch lengths, extrapolating likelihood from better conditioned matrices that we slowly transform to resemble the estimated matrix, even dropping taxa not involved in hybridization to try to make a more workable variance covariance matrix). These remain as options within our software, but we do not recommend modifying the defaults. The approach we seemed most useful overall was to use the function nearPD from the package Matrix (Bates and Maechler 2017), which implements the correction of Higham (2002) to find the nearest positive definite matrix to the given matrix given the parameter values. Simply using a numerically convenient value near but not at the actual value to examine is a problem (see the “streetlight effect” (Freedman, 2010)), and so taking the likelihood near an examined point but not at it should come at a cost. We implemented a penalty in such cases: if a nearPD correction resulted in an adjusted matrix not equal to the initial matrix, we penalized the likelihood under the adjusted matrix by 10 log likelihood units plus the Euclidean distance between the original and adjusted matrices. This biases the search to look in regions where the parameters lead to tractable results while still incurring a penalty for using an approximation. Another approach to deal with problematic regions of parameter space is to start analyses from a set of Latin hypercube samples (Carnell 2016) (by default, 5000 points with a range based on quick estimates of parameter values from Geiger (Harmon et al. 2008) and information in the data. In many cases, all these points will have finite likelihoods calculated, and the best one serves as a starting point for an optimization step. However, for some networks, many of these points result in a likelihood that cannot be calculated. In such cases, it is possible that the true optimal parameter values are in the region of space that cannot be feasibly explored. We have visualizations of the points explored with finite likelihoods so users can examine this. Future work could include implementing the model here into an approximate Bayesian computation framework where points are simulated to get estimates directly without computing likelihood.

It is worth noting that this problem, while common in our simulated networks and one empirical case (see below), is not limited to phylogenetic networks only. This can occur even in the Brownian motion on a tree case where tips are not coeval and there are some small branch lengths, even though the analysis is performed under the same model as used to generate the data. It is unclear how often these issues are seen in practice on phylogenetic trees. We note that programs implementing the Ho and Ané (2014) approach, which does not involve matrix inversion, are free from this problem, but many programs that use the multivariate normal to calculate likelihoods directly will have issues. See supplemental R script (SupplementalAnalysisForBrownianMotion.R) illustrating this problem.

*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 structure of hybridization, (iv) the value of (1 or 2), and (v) the value of (0, 0.1 times the variance that comes from evolution on the tree, or equal to the variance that comes from evolution on the tree). 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 1, 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 original hybrid. Other parameter values fixed in the model were , and . All simulations were carried out by the *BMhyb* package. Failed simulation attempts (where the simulated attachment point was not located correctly) were repeated; simulations that led to numerical problems were retained. 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)).

*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 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). Note that these are used as test cases for making sure the method operates sufficiently well and gives reasonable estimates for these data — any biological conclusions should be re-examined by specialists using more modern phylogenies and data.

*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. Their tree is not available in TreeBase or OpenTree, so to use their topology but estimate branch lengths, we downloaded their sequences (they used NADH dehydrogenase subunit 2 gene (ND2)) from GenBank (Benson et al. 2005). The sequences were aligned by MAFFT v. 7.300b (Katoh and Standley 2013). A backbone constraint was made from Kobmüller et al. (2007)’s overall hybrid tree in Mesquite (Maddison and Maddison 2011) and used in all subsequent searches. A constrained likelihood search was performed in RAxML 8.2.9 (Stamatakis 2014) using GTR amma and only one partition (since this is a small dataset with only one gene). The phylogram was converted to a chronogram using treePL (Smith and O’Meara 2012), with a fixed age of 4.19 MY used for the divergence of *Lamprologus callipterus* and *Neolamprologus calliurus* based on the midpoint for the range of this divergence given in Kobmüller et al. (2007). Note that due to an apparent numerical issue in treePL, we used a fixed age of 419 and then rescaled the tree within R. The point of this tree inference was to recover the tree and inferred network of Kobmüller et al. (2007), not to make new inferences regarding cichlid evolution. Trait values (standard length, in cm) were collected from FishBase (Froese and Pauly 2010) using the R package rfishbase (Boettiger et al. 2012). Taxa absent from the tree or with missing trait data were removed using *Geiger* (Harmon et al. 2008).

While our method can, in theory, estimate measurement error, in practice most good empiricists will have estimated this in the course of their study. Fishbase does not typically have a way to get estimates of uncertainty in maximum length. Instead, we approximated this by using observed ranges from the U. of Michigan Museum of Zoology fish collection. We downloaded all specimen information for all genera in our study, extracted the range of standard lengths, used the square root of the range divided by two as an estimate of the standard error, after first excluding records where the minimum was less than half the maximum (which could indicate juveniles included (S. Borstein, pers. comm)) and where there was only one observation. We then did a linear regression between this estimate and log transform the standard length. This regression was used to predict the standard error for the total length measurements in Fishbase. This is only an approximation, but the resulting standard errors seem biologically reasonable at typically 10% of the log-transformed measurement.

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 unsampled (perhaps extinct, though not necessarily) lineages. As demonstrated in Figure 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, again with the goal of replicating the original study tree, 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 unsampled 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 extracted from their Table 2. We note that this number is a proportion, thus not quite meeting the expectations of Brownian motion (unbounded traits); we log transformed it, but this is still an imperfect fix. We used iPlant TNRS (Boyle et al. 2013) to convert the taxon names from both datasets to the same taxonomy, and Geiger (Harmon et al. 2008) to prune the tree and data to the same taxon set. The evolutionary relationships for cichlid (Figure 2 left plot) and tobacco (Figure 2 right plot), respectively, are shown by the evolutionary tree with the relevant gene flow. The donor-recipient relationship among the hybrids and their parents in the cichlid and *Nicotiana* datasets can be found in supplemental material.



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

*Model Selection and Parameter Estimation*

We tried four different types of models for each empirical dataset. All fix the gene flow and allow and to be optimized. They differ in the settings for mean change in the hybrid and the hybrid variation at formation . Model 1 fixes at 1 but allows to vary; Model 2 allows to vary but fixes at 0; model 3 fixes at 1 and at 0 and model 4 allows both to vary. For the cichlid dataset, we tried each of these models with three settings for SE: fixed at zero, estimated as a free parameter, and predicted from museum samples. For tobacco, there are no good ways to get non-phylogenetic predictions of SE (the original paper just reports means) so we used SE fixed at zero and estimated as free only. For the cichlid dataset, 5000 points were sampled in a Latin hypercube design centered on estimates from the empirical dataset and the best of these used to start optimization for searching maximum likelihood estimates of parameters. For the tobacco dataset, numerical issues prevented effective optimization. Instead, we sampled 100,000 points to get parameter estimates, then an additional 100,000 points centered around the best of these to get better estimates of the uncertainty in the parameter values. However, given parts of parameter space were inaccessible numerically, one must be cautious in interpreting results.

*Adaptive confidence intervals sampling*

Uncertainty in parameter estimates can be substantial. One way of estimating this can be looking at the curvature of the surface at the maximum likelihood optimum, but this is known to be problematic when the likelihood function is not regular (Pawitan 2013). A different approach, advanced by Edwards (1992) is to look at a confidence region of all points that generate a log likelihood within a certain range (often, set to be a difference within 2 log likelihood units) of the maximum likelihood. One approach to calculate this would be to vary each parameter on its own while holding the others constant. This is convenient and fast to implement, but can result in artificially small confidence intervals. For example, if two parameters *a* and *b* covary such that the likelihood is the same as long as , the likelihood changing just *a* or just *b* would drop off very quickly, but there is a ridge containing a wide array of *a* and *b* values that would not affect the likelihood. Thus, we chose to examine varying all parameters at once, so that if there is a ridge or other structure for the likelihood surface we do not overestimate our certainty. While there are many algorithms to find the peak of a surface, there are fewer to find the entirety of a region two 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 *BMhyb* 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 doi:10.5061/dryad.6jn8s.

**RESULTS**

*Simulation for Assessing General Performance of Models*

Parameter estimates were obtained using searches of (i) optimization (ii) doing grid only, (iii) grid first then optimization. Overall we were not pleased with performance: looking at just the simulations with the best chance for good results (flow rate of 0.5, 100 non-hybrid taxa, 10 hybrid taxa) there was often quite poor correlation between the estimates of parameters and and the true values (SE, , and performed far better, but they are not interesting parts of our model). Experimenting with the results suggested that model averaging or just taking the parameter estimates from the best model had about the same performance; using models where only varied performed better than models where varied. We investigated this, and other elements of the simulation, by doing a regression where we found importance and coefficients of parameters by dredging (using all subsets of the global model, and using Akaike weights to calculate importance). This suggested that choosing model-averaged or best model only parameter estimates did not matter much (importance of 0.27) but using models with fixed was important for estimating (importance of 1.00), less so for (importance of 0.27) [and important for the other three free parameters], but the sign of the coefficient, negative for both and , suggested that limiting models to those with fixed reduced error in both and, oddly, . However, inspecting the parameter estimates coming from trees with each simulation scenario also suggested they performed very similarly. Summaries later in the paper thus merge trees of both types. Flow and number of hybrid and nonhybrid taxa were found to be important.

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



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

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

*Assessing Model Identifiability through Jointly Estimating Parameters*

The shape of the likelihood surface provides the ability to estimate parameter values: if the surface is flat, there is little support for a parameter estimate. In some cases, like the trend parameter for Brownian motion with a trend for coeval taxa, no amount of data is adequate to estimate the parameter: this parameter is formally non-identifiable. There is also a softer definition of identifiability: given a particular dataset, is there enough data to estimate a parameter? We investigated both of these by creating contour maps of the likelihood surface for pairs of parameters under the cichlid and *Nicotiana* data sets (see Figure 4). For these empirical datasets, parameters appear distinguishable: there are no ridges in the likelihood surface, even though the confidence intervals are wide. Thus, the parameters are formally 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 black contour shows the ∆2 log likelihood unit region, and the outer gray contour shows the ∆5 log likelihood unit region. Note the lack of ridges but presence of wide intervals.

*Model Selection and Parameter Estimation for the Empirical Data*

For cichlids we ran all four models under three approaches: assuming zero measurement error, estimating measurement error as part of the analysis, or using information from museum specimens to predict measurement error for each species and using this a known value in the analysis. Likelihood of models where measurement error was fixed at zero was worse than for equivalent models where a single measurement error across everything was estimated (which makes sense, given their extra degree of freedom). Under AICc the best models in every case had measurement error predicted from the museum specimen regression (these models had 86% of the Akaike weight overall). The best model overall was model 3 with museum-predicted measurement error, which has both and fixed. Model averaging across all the models suggested an estimate of of 0.91 (with bounds of 0.82 to 1.09), estimate of of 1.3e-9 log cm (effectively zero) with bounds of 1.3e-9 to 0.14 log cm. The estimate of was 0.009 (bounds of 0.005 to 0.045) (log cm)2 / MY. The best model for cichlids had no burst of variation at hybridization events nor a bias in hybrid size (Table 1). In models where was allowed to vary, the confidence interval included 1, and for , the confidence interval (and MLE in all models) included zero, again suggesting lack of evidence for evolutionary speed ups or mean changes with hybridization. Measurement error was estimated to be substantial; in fact, the model estimated no effective Brownian motion at all ( estimates of zero, variance at the tip due to the evolutionary process also near zero) with all observed variance being due to just measurement error at the tips. For *Nicotiana*, the best model (with OOO of the Akaike weight) had as a free parameter and constrained to be zero. There is strong evidence from this model that beta is greater than one (point estimate from best model is 2.358, the 95 percent confidence interval (CI) is 1.522 to 2.748; model averaged estimate is 2.206), suggesting that hybrid species have higher success rates as seedlings under drought conditions than do their parents. There is again little evidence for increased variance at hybridization events. Given the tree height and its rate, we expect variance at the tips to be 0.636; from measurement error, there is an additional 0.344 variance (both in units of log((seedling survival)2)), suggesting meaningful Brownian motion on the tree but still quite important measurement variance.

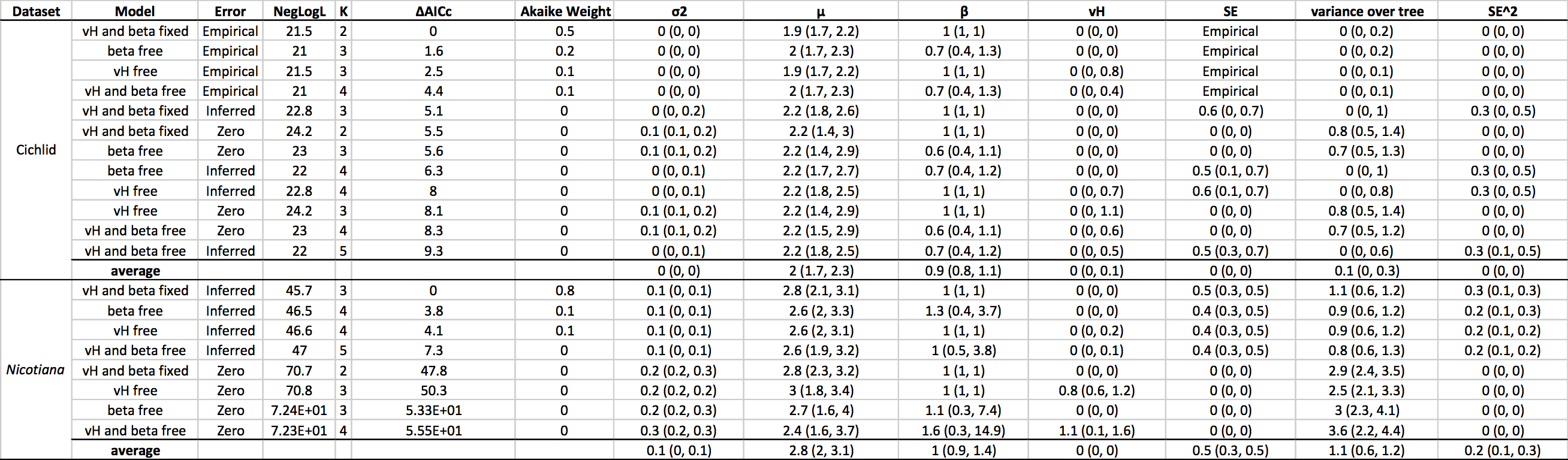


TABLE 1: Model fitting for the cichlid and tobacco data. In this table, NegLogL is the negative log likelihood of the model, is the number of free parameters, is calculated by subtracting the AICc value from the lowest AICc value among the four models, Akaike weight is calculated by and then normalizing the weights. SE is the estimated standard error; variance over the tree is σ2 times tree height. The relative ratio of standard error and the variance coming from the evolutionary process gives an indication of the importance of each. Parameter estimates and their adaptive confidence intervals are reported. The model averaged-parameter estimates (Burham and Anderson 2004) are reported where is the Akaike weight for the *i*th model.

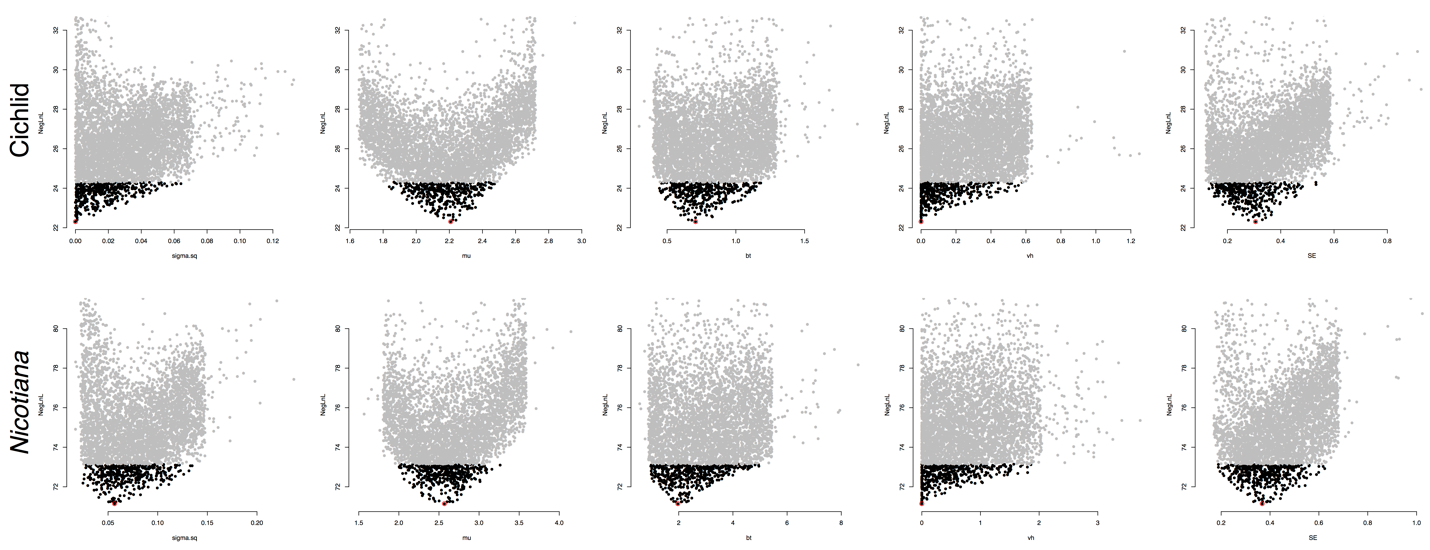


Figure 5. The adaptive confidence intervals for the free parameters of the best model for cichlids (top) and for *Nicotiana* (bottom). In each plot the vertical axis represents the negative log likelihood value and the horizontal axis represents a wide parameter range. The red dot at the lowest y-axis 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 dot. Dots in black represent the desired likelihood value (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 make attempt to use comparative methods on general networks rather than just trees and it works well for estimating general Brownian motion parameters such as evolutionary rateand root state . There are also hybridization-specific parameters where it performs variably. The method can estimate hybridization bias, 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 ; we would recommend caution interpreting any estimates of . 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 measurement error, but not enough to wipe out the phylogenetic history. The results suggest that hybrids perform better in droughts than their parent species, though this is not statistically significant given the confidence interval. However, it might point the way to further studies about drought tolerance, an area that will be of increasing importance.

Several approaches have been proposed for inferring different rates along the branch for a given phylogenetic tree (McPeek 1991; O'Meara et al. 2006; Revell 2008; and Beaulieu et al. 2012), and it would be useful to extend our work to allow for this heterogeneity. Another possible extension is to use a more parameter-rich Markovian process. The model can be extended to allow trait evolution following the OU process (Hansen 1997; Butler and King 2004; Beaulieu et al. 2012) or to model the evolution of interacting population (Bartoszek et al. 2017). Putting the approach in a Bayesian context is also possible. In this case, parameters of multiple selective regimes, multiple rates of variation, and multiple rates of constraining forces could be embedded in the model. Developing a more complex model of this type could be very useful when analyzing fairly large data sets of hundreds of species or more, where heterogeneity is expected and there may be power to provide estimates for many parameters.

In nature, approximately 10% of animal species and 25% of plant species hybridize (Mallet 2005, 2007), suggesting that there is widespread gene flow between “species.” Some of this gene flow may lead to hybrid speciation in the manner assumed in our method, and hybrid speciation is widely suspected in many groups (Arnold 1996, Welch and Riesberg 2002). However, even in the absence of hybrids formed from two distinct parent species, such ongoing gene flow suggests a need for a network metaphor, as suggested by Morrison (2014). Our method cannot currently deal with this sort of gene flow: we represent the hybrid as being the result of a single event 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.

A key input to our method is a phylogenetic network. There are an increasing number of

approaches to inferring these (Solís-Lemus and Ané 2016), some of which can infer unsampled lineages (i.e., hybridization events that seem to move forward in time). As with other comparative methods, quality of the input phylogeny (or in our case, phylogenetic network) may affect results. For models like this, which are essentially based on rescaling the variance covariance matrix, branch lengths as well as topology matter.

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 straightforward to use the existing approach to estimate ancestral states on a network (as implemented in recent verisons for a related model in the PhyloNetworks package (Solís-Lemus and Ané 2016). 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, Cécile Ané, Paul Bastide and an anonymous reviewer for their suggestions to improve this manuscript, especially Paul Bastide for finding a mathematical error in an earlier version. 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 D-C J and for summer visitor support, as well as a grant to D-C J from Ministry of Science and Technology of Taiwan, R.O.C., MOST 105-2119-M-035-001.

**REFERENCES**

Ardia, D., Boudt, K., Carl, P., Mullen, K.M., Peterson, B.G. 2011​.​ Differential Evolution with​ ​DE-optim​:​ An Application to Non-Convex Portfolio Optimization.​ ​The R Journal, 3(1), 27-34.

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

Bartoszek K., Glémin S.,Kaj I. and Lascoux M. 2017. Usingthe Ornstein–Uhlenbeck process to model the evolution of interacting populations, Journal of Theoretical Biology. doi: 10.1016/j.jtbi.2017.06.011

Bates D. and Macheler M. 2017. *Package* ‘*Matrix*’, Reference manual, version 1.2-11. Available: https://cran.r-project.org/web/packages/Matrix/Matrix.pdf

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.

Brissette F.P., Khalili M. and Leconte R. 2007. Efficient stochastic generation of multi-site synthetic precipitation data. Journal of Hydrology 345:121-133.

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.

Cardillo M., Mace G.M., Gittleman J.L., Purvis A. 2006. Latent extinction risk and the future battlegrounds of mammal conservation. Proceedings of the National Academy of Sciences. 103(11) 4157-4161.

Carnell R. 2016. lhs: Latin Hypercube Samples. R package version 0.13. https://CRAN.R-project.org/package=lhs

Chase M.W., Knapp S., Cox A.V., Clarkson J.J., Butsko Y., Joseph J., Savolainen V., and Parokonny A.S. 2003. Molecular systematics, GISH and the origin of hybrid taxa in *Nicotiana* (Solanaceae). Annals of Botany 92: 107-127.

Clarkson J.J., Lim K.Y., Kovarik A., Chase M.W., Knapp S. and Leitch A.R. 2005. Long-term genome diploidization I allopolyploid *Nicotiana* section Repandae(Solanaceae). New Phytologist 168:241-252.

Doolittle W.F. 1999. Phylogenetic classification and the universal tree. Science 284: 2124-9.

Duftner N., Kobmüller S., Sturmbauer C. 2005. Evolutionary relationships of the Limnochromini, a tribe of benthic deepwater cichlid fish endemic to Lake Tanganyika, East Africa. J Mol Evol 2005, 60:277-289.

Dutta H. 1994. Growth in fishes. Gerontology. 40:97-112.

Edward A.W.F. 1992. Likelihood. Expanded Edition. Johns Hopkins University Press.

Felsenstein J. 1985. Phylogenies and the comparative method. The American Naturalist 125:1-15.

Felsenstein J. 2008. Comparative methods with sampling error and within-species variation: contrast revisited and revised. The American Naturalist 171: 713-725.

Freedman, D.H. 2010. Why scientific studies are so often wrong: The streetlight effect. Discover Magazine 26.

Froese R., Pauly D. 2010. FishBase. World Wide Web electronic publication. [www.fishbase.org](http://www.fishbase.org).

Gittleman J.L., Kot M. 1990. Adaptation: statistics and a null model for estimating phylogenetic effects. Systematic Zoology 39: 227-241.

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

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

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

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

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

Higham N. 2002. Computing the nearest correlation matrix - a problem from finance; IMA Journal of Numerical Analysis 22, 329–343.

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

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

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

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

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

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

Koblmüller S., Duftner N., Sefc K.M., Aibara M., Stipacek M., Blanc M., Egger B., Sturmbauer C. 2007. Reticulate phylogeny of gastropod-shell-breeding cichlids from Lake Tanganyika: the result of repeated introgressive hybridization. BMC Evolutionary Biology 7:7.

Komori T., Myers P.N., Yamada S., Kubo T., and Imaseki H. 2000. Comparative study of the *Nicotiana* species with respect to water deficit tolerance during early growth. Euphytica 116:121-130.

Kubatko L.S. 2009. Identifying hybridization events in the presence of coalescence via model selection. Syst. Biol. 58:478-488.

Linder C.R., Rieseberg L.H. 2004. Reconstructing patterns of reticulate evolution in plants. American Journal of Botany 91:1700-1708.

Maddison, W.P. and Maddison D.R. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75 <http://mesquiteproject.org>

Maddison, W.P., Midford, P.E., and Otto, S.P. 2007. Estimating a binary character’s effect on speciation and extinction. Systematic Biology 56(5) 701-710.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends in Ecology and Evolution 20:229-237.

Mallet J. 2007. Hybrid speciation. Nature 445:279-283.

Martins E.P. and Hansen T.F. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. The American Naturalist. 149:646-667.

McPeek M.A. 1995. Testing hypotheses about evolutionary change on single branches of a phylogeny using evolutionary contrasts. American Naturalist 145:686-703.

Meng C., Kubatko L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theoretical Population Biology 75:35-45.

Moret B.M.E., Nakhleh L., Warnow T., Linder C.R., Tholse A., Padolina A., Sun J., Timme. R. 2004. Phylogenetic networks: modeling, reconstructibility, and accuracy. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1:13-23.

Morrison D.A. 2014. Is the tree of life the best metaphor, model, or heuristic for phylogenetics? Systematic Biology 63:628-638.

Nakhleh L. 2011. Evolutionary phylogenetic networks: model and issues. Problem solving handbook in computational biology and bioinformatics. P125-158.

O'Meara B.C., Ané C., Sanderson M.J., Wainwright P.C. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922-933.

Otto S.P., Whitton J. 2000. Polyploid incidence and evolution. Annual Review of Genetics 24: 401-437.

Paradis E, Claude J. & Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinfromatics 20:289-290.

Pawitan Y. 2013. In all likelihood: statistical modeling and inference using likelihood. Oxford University Press.

Pickrell J.K. and Pritchard J.K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8(11): e1002967.

Rebonato R. and Jäckel P. 2000. The most general methodology to create a valid correlation matrix for risk management and option pricing purposes. J. Risk, 2(2):17-26.

R Development Core Team. 2008. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.

Revell L.J. 2008. On the analysis of evolutionary change along single branches in a phylogeny. American Naturalist 172: 140-147.

Revell L.J. 2012. phytools: An R package for phylogenetic comparative biology(and other things). Methods Evol. Evol. 3:217-223. doi:10.1111/j.2041-210X.2011.00169.x

Rhymer J.M. and Simberloff D. 1996. Extinction by hybridization and introgression. Annual Review of Ecology and systematics 27:83-109.

Rieseberg L.H., Archer M.A., Wayne R.K. 1999. Transgressive segregation, adaption and speciation. Heredity 83: 363-372.

Rieseberg L.H. 2006. Hybrid speciation in wild sunflowers. Ann. Missouri Bot. Gard. 93: 34-48.

Sang T., Zhong Y. 2000. Testing hybridization hypotheses based on incongruent gene trees. Syst. Biol. 49:422-434.

Schafer J and Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implication for functional genomics. Statistical Application in Genetics and Molecular Biology 4(1):32

Schluter, D., Price T., Mooers A.O., Ludwig D. 1997. Likelihood of ancestor states in adaptive radiation. Evolution 51(6): 1699-1711.

Smith S.A. and O’Meara B.C. 2012. treePL: Divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics doi: 10.1093/bioinformatics/bts492.

Solís-Lemus C. and Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS genetics*12.3 (2016): e1005896.

Stadler T. TreeSim: simulating trees under the birth-death model. R package version 2.1.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post analysis of large phylogenies. Bioinformatics doi: 10.1093/bioinformatics/btu033.

Swofford, D.L. 2003. PAUP\*. Phylogenetic Analysis Using Parsimony(\* and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Wang J, Guo M, Xing L, Che K, Liu X, Wang C, 2013. BIMLR: A method for constructing rooted phylogenetic networks from rooted phylogenetic trees. Gene 527:344-351.

Weigel D.E., Peterson J.T., Spruell P. 2002. A model using phenotypic characteristics to detect introgressive hybridizations in wild west slope cutthroat trout and rainbow trout. Transactions of the American Fisheries Society 141:389-403.

Welch ME and Riesberg LH. 2002. Habitat divergence between a homoploid hybrid sunflower species, Helianthus paradoxus (Asteraceae), and its progenitors. American Journal of Botany 89:472-478.

Willson SJ, 2013. Reconstruction of Certain Phylogenetic Networks from Their Tree-Average Distances. Bulletin of Mathematical Biology 75:1840-1878.

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