Assessing the Impact of Continuous Traits on the Evolution of Discrete Traits: The Ancestral Condition Test.

Heath Blackmon and Richard H. Adams

**ABSTRACT**

Co-evolutionary relationships where the value of a continuous trait affects the evolution of a discrete trait remain difficult to study. We present an R package, EvobiR, that includes a novel comparative method that tests whether a discrete trait transitions when a continuous trait has values more extreme than expected if both traits were evolving independently. We show that this approach has acceptable type I error rates but has low power unless the sample size is large. We also briefly describe functions to facilitate comparative and population genetic analyses included in our package. Finally we discuss three pedagogical tools included as part of EvobiR. These components use R to produce interactive web pages that can be used for teaching important concepts in evolutionary biology either as demonstrations or to allow students to explore the concepts independently.

**INTRODUCTION**

One of the central goals of evolutionary biology is to understand the evolution of traits among species. With the advent of statistically robust methods of phylogenetic tree inference, we can now fit mathematical models to the evolution of both discrete traits (Lewis 2001; Felsenstein 2012) and continuous traits (Felsenstein 1985; Butler & King 2004; O'Meara *et al.* 2006; Landis, Schraiber & Liang 2013). A variety of methods that test for correlations between multiple continuous traits have been developed (reviewed in Felsenstein 2004). However, there are fewer methods available to identify and analyze a co-evolutionary relationship where a continuous trait affects the evolution of a discrete trait (but see Hadfield 2010; Felsenstein 2012).

In a recent study of sex chromosome evolution, we were interested in testing whether loss of the Y chromosome occurred in taxa with a significantly higher total number of chromosomes than would be expected by chance (Blackmon & Demuth 2015). A number of existing approaches might be employed in such a situation. First, by discretizing chromosome number into a low and high category, we could apply Pagel’s test for the correlation of two binary characters (Pagel 1994). This approach is problematic, however, as defining threshold values for high and low classes involves arbitrary decisions that may not reflect biological reality, and if several threshold values are tested, p-values should be corrected for multiple comparisons. Additionally, by discretizing a continuous variable we effectively reduce the amount of data being applied to the problem and decrease our power to detect a significant relationship. Another approach would be to assume that our discrete state (presence or absence of a Y chromosome) could be modeled as if it were underlied by an unobserved continuous trait that we cannot observe. If this assumption is justified, we could employ the threshold model (Wright 1934; Felsenstein 2012). This approach would allow us to estimate the correlation between chromosome number and presence or absence of the Y chromosome. A third approach would be to use a phylogenetic mixed model treating chromosome number as a predictor variable and presence or absence of the Y chromosome as a response variable (Hadfield 2010).

Of these three approaches, the latter two are perhaps the most frequently applied in such situations, but both can suffer from the same problem: correlation does not provide information about the direction of causality. For example, does high chromosome number lead to Y chromosome loss or does Y chromosome loss lead to high chromosome number?

In light of these issues, we have developed a flexible statistical test to infer significant relationships between the value of a continuous trait and transitions in the state of a discrete trait. The flexibility of our approach comes from the fact that we can choose to reconstruct the continuous trait using all data or in cases where there is a clear derived and ancestral condition we can restrict our reconstruction of the continuous trait to only use data from species that have retained the ancestral condition of the discrete trait. This approach allows us to control for the biological reality that their may be feedback where the derived condition of the discrete trait can impact the evolution of the continuous trait.

Briefly, our approach estimates the mean value of the continuous trait when the discrete character transitions and we evaluate significance of this mean value by comparison to a null distribution that is expected if there was no relationship between the traits being studied. We apply our ancestral condition test to simulated datasets, and we find that our method is conservative but has relatively low power unless the number of taxa included is large. We believe that our approach offers a valuable addition applicable in many cases where the evolution of a discrete character may be influenced by the state of a continuous trait.

**METHODS**

Our approach determines whether there is evidence that nodes subtending the transitions of a discrete character occur under extreme values for a continuous character under study. We refer to this method as the ancestral condition test (AncCond), which involves four steps described below:

1) Reconstruct the value of the continuous character while optionally pruning data from those species that exhibit the derived state of the discrete character. By pruning data from taxa exhibiting the derived state of the discrete character, we create a more conservative test that will be applicable in cases where transitions into the derived state of the discrete character is expected to lead to changes in the selective forces acting on the continuous character. Ancestral state reconstruction is accomplished under a Brownian motion model using existing functions (Revell 2012) (Fig. 1A).

2) Next, we reconstruct the evolution of the discrete character with stochastic trait mapping to build a distribution of possible evolutionary histories (Fig. 1B) (Huelsenbeck, Nielsen & Bollback 2003; Bollback 2006). Our test allows users to fix a single state at the root of the tree in those cases where this is appropriate. For instance, when transitions are expected to occur in only one direction (cite us haplodiploidy paper). Stochastic mapping is performed using existing functions in R (Revell 2012) (Fig. 1B).

3) We then process the stochastic mappings to classify all nodes in the tree into three groups: 1) Nodes exhibiting state 1 of the discrete character character and no transitions in discrete character in immediate daughter branches (Fig. 1C, red nodes), 2) Nodes exhibiting state 2 of the discrete character and no transitions in discrete character in immediate daughter branches (Fig. 1C, blue nodes), 3) producing nodes exhibiting one state of the discrete character but with one of the immediate daughter branches showing a transition into the other state of the discrete character (Fig. 1C, green nodes). We then parse producing nodes to select those that produce transitions from state 1 to state 2 or from state 2 to state 1. Next, we calculate the mean of the continuous trait at each group of producing nodes identified across stochastic mappings and (if available) a distribution of trees, incorporating both phylogenetic and ancestral state reconstruction uncertainty. This value is the estimate of the mean value of all nodes that produce the derived state. We refer to this value as the “ancestral condition” for each of the states of the discrete character (Fig. 1C).

4) We construct a null distribution by using a markov model to simulate the evolution of a discrete character that evolves with transition rates inferred from the empirical data. Using this simulated discrete character we then calculate the ancestral condition for each state of this neutral character. This is repeated 1000s of simulated discrete characters to produce a null distribution of ancestral conditions. Finally, we use this null distribution to calculate an empirical p-value for the observed ancestral condition of the discrete character (Fig. 1D).

*Simulated data*

Scenario one

Unidirectional change in the discrete character

Scenario two

Bidirectional change in the discrete character

Low rates of evolution

High rates of evolution

Scenario three

Empirical phylogeny

To test our approach, we simulated 200 trees using a birth death model with a birth rate of 1.0 and death rate of 0.5 allowing trees to grow until 200 extant species were reached. Trees were simulated using the function sim.bdtree in Geiger (Harmon *et al.* 2008). Next, we simulated a continuous character evolving by Brownian motion with a rate parameter of 0.2 and a starting mean of 1.0 on each tree using the function sim.char in Geiger (Harmon *et al.* 2008). We used a branch scaling approach to generate simulated discrete characters. We first identified those branches that had a mean value for the continuous trait in the upper or lower quartile. Branches in the lower quartile were scaled by a factor of 1/x while branches in the upper quartile were scaled by a factor of x, and we repeated this process for ten values of x ranging from 1 to 10 on each of our 200 trees. Next, we evolved a discrete character on these scaled trees, setting the root to state to 1 (ancestral condition) and allowing the trait to evolve under a Mk2 model where transition to state 2 (derived condition) were allowed but not back transitions to state 1 (matching our expectations for the type of characters we believe this approach may be most useful) using the sim.char function from Geiger. Various transition rates from state 1 to state 2 were evaluated and we found that a rate of 0.02 was sufficient to insure that two or more transitions occurred in all simulated datasets. Thus, our total of 2,000 simulated datasets consisted of a set 200 trees with 10 replicates with branch scaling factors varying from 1 to 10.

When the branch scaling factor was equal to one, there was no significant relationship between our characters and thus allowed us to test the type I error rate of our method. The remaining nine levels of the scaling factor allowed us to evaluate the power of this approach under an increasingly strong relationship between the continuous and discrete character: branches that had high continuous values had more time to experience a transition in the discrete character, while branches inferred to have low continuous character values were less likely to experience a transition. This matches the hypothesis we tested for Y chromosome loss.

To assess the relationship between the number of taxa included in an analysis and the statistical power of our approach to detect significant co-evolutionary relationships, we simulated datasets with a scaling factor of 5 and pruned from 0 to 180 randomly selected tips. If a pruned dataset did not include at least two taxa in each state of the discrete character, it was discarded and a new dataset was generated. A total of 10 of these reduced taxa datasets were created for each of 100 trees randomly drawn from the 200 described above for a total of 1000 simulated datasets.

Below we show the performance of this approach with the simulated data, reporting both type I error, power, and sample size (number of taxa available for analysis). Tests were considered statistically significant at a p-value ≤ 0.05. All analyses were completed using our software package EvobiR version 1.0 loaded from GitHub with RStudio version 0.98.976 running R version 3.02 on a MacBookPro with 4GB of 2600MHz RAM and a 2.5GHz processor (RStudio 2012; R Development Core Team 2013).

**RESULTS**

Analysis of the 200 simulated datasets with a scaling factor of 1 (no relationship between characters) showed a standard type I error rate of 5% (figure 2). Analysis of the datasets with the scaling factor ranging from 2-10 allowed us to assess the power of this approach. We found that under the simulation conditions our approach had a power ranging from 15% (scaling factor of 2) to 76% (scaling factor of 10).

The number of taxa included in comparative analyses often varies by orders of magnitude. We found that our ancestral condition test requires moderately large sample sizes to reliably detect a significant relationship. Figure 3 shows that when 20 taxa where included in our simulated datasets we were only able to detect a significant relationship 16% of the time, a reduction in power of approximately 70% relative to when all 200 taxa were present. Only when we had over 160 taxa did the power of our approach reach 50%.

**DISCUSSION**

Our simulation study demonstrates the conservative nature of our ancestral condition test, which entails acceptable type I error rates but relatively low power. In situations where no true relationship exists between the origin of a derived state of a discrete trait and the value of a continuous trait (scaling factor of 1), our test will incorrectly identify a significant relationship 5% of the time – typical of most statistical tests of inference.

However, with a scaling factor of 5, branches with a high continuous value effectively have a transition rate into the derived state of the discrete trait that is 5 times higher than intermediate branches and 25 times higher than branches with a low continuous value. Arguably a very strong relationship, but even in this case we detect the relationship in only 55% of simulated datasets.

While current approaches for jointly analyzing continuous and discrete traits are effective in finding correlations, our novel approach provides an important and useful extension to current comparative methods as it explicitly provides information about the order in which traits evolved across a phylogenetic tree. An additional strength of our approach is that it is robust to effects that the derived state of discrete character may have on the evolution of the continuous trait. By only using continuous trait values from taxa that exhibit the ancestral version of the discrete state we are explicitly asking if there is evidence in these taxa that the species exhibiting the derived state originated among lineages with continuous trait values that are significantly different from what we would expect if both traits were evolving independently. Like many comparative methods though the performance of the ancestral condition test is dependent on having a sufficiently large tree with multiple origins of the derived state of the discrete character.

Our approach assumes an Mk2 model for the evolution of the discrete trait and a Brownian motion model for the evolution of the continuous trait. The Mk2 model implemented assumes a discrete character that has an ancestral and derived state and does not experience reversion. It would be straightforward to extend our approach to allow for other models for the discrete trait; however, more complex models will often lead to greater uncertainty in ancestral state reconstructions and thus lower power to detect significant relationships. The adequacy of these underlying models should be evaluated prior to using the ancestral condition test. Model adequacy of the continuous trait can be accomplished in a number of ways (Garland, Harvey & Ives 1992; Purvis & Rambaut 1995; Pennell *et al.* 2014). Perhaps the most robust of these methods involves first calculating test statistics on the phylogenetic independent contrasts from the observed data. Simulations are then performed and these same test statistics are calculated for each simulation - generating null distributions. The benefit of this approach is that deviation of observed test statistics can not only determine if a model is adequate, but also may provide information about what type of alternative model might perform better (Pennell *et al.* 2014). Evaluation of model adequacy for discrete traits is currently less developed. One option is to use a Monte Carlo approach to determine if the model and parameter estimates are able to generate data with similar proportions of taxa in each discrete state (Price *et al.* 2012; Blackmon & Demuth 2014). An additional approach is to examine the number of state changes expected under parsimony for the observed data and compare this to the number of state changes expected under parsimony for simulated datasets (Beaulieu, O'Meara & Donoghue 2013).

Many comparative methods can suffer from pseudoreplication (Maddison & FitzJohn 2014), returning significant results when some would argue that they should not (increased type I error rates). Our ancestral condition test could suffer from such shortcomings. For example, one could imagine a situation where the most recent common ancestor of a clade evolved a higher value for the continuous trait, followed by the evolution of an additional trait that increased the probability of transitions to the derived state of the discrete state. Such a pattern could conceivably produce a significant result despite having only a single origin for the high continuous trait value. We believe the best approach to avoiding the errors associated with pseudoreplication is thorough data exploration prior to applying nearly any comparative method, including our ancestral condition test. There are many tools now available that allow simultaneous visualization of trees and data that can be useful in early stages of an analysis to alert researchers of possible issues. An alternative approach to solving the pseudoreplication problem is to perform what is effectively a meta-analysis . With this approach the original dataset is divided into a number of independent data sets and the results of all of the independent tests can be evaluated (Mayrose *et al.* 2011).

Beaulieu, J.M., O'Meara, B.C. & Donoghue, M.J. (2013) Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Syst Biol,* **62,** 725-737.

Blackmon, H. & Demuth, J.P. (2014) Estimating Tempo and Mode of Y Chromosome Turnover: Explaining Y Chromosome Loss With the Fragile Y Hypothesis. *Genetics,* **197,** 561-572.

Blackmon, H. & Demuth, J.P. (2015) The fragile Y hypothesis: Y chromosome aneuploidy as a selective pressure in sex chromosome and meiotic mechanism evolution. *Bioessays,* **37**.

Bollback, J.P. (2006) SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics,* **7,** 88.

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

Durand, E.Y., Patterson, N., Reich, D. & Slatkin, M. (2011) Testing for ancient admixture between closely related populations. *Mol Biol Evol,* **28,** 2239-2252.

Felsenstein, J. (1985) Phylogenies and the Comparative Method. *The American Naturalist,* **125,** 1-15.

Felsenstein, J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland Ma.

Felsenstein, J. (2012) A comparative method for both discrete and continuous characters using the threshold model. *Am Nat,* **179,** 145-156.

Garland, T., Harvey, P.H. & Ives, A.R. (1992) Procedures for the Analysis of Comparative Data Using Phylogenetically Independent Contrasts. *Systematic biology,* **41,** 18-32.

Hadfield, J.D. (2010) MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software,* **33,** 1-22.

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

Huelsenbeck, J.P., Nielsen, R. & Bollback, J.P. (2003) Stochastic Mapping of Morphological Characters. *Systematic biology,* **52,** 131-158.

Landis, M.J., Schraiber, J.G. & Liang, M. (2013) Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. *Systematic biology,* **62,** 193-204.

Lewis, P.O. (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology,* **50,** 913-925.

Maddison, W.P. & FitzJohn, R.G. (2014) The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters. *Systematic biology*.

Mayrose, I., Zhan, S.H., Rothfels, C.J., Magnuson-Ford, K., Barker, M.S., Rieseberg, L.H. & Otto, S.P. (2011) Recently Formed Polyploid Plants Diversify at Lower Rates. *Science,* **333,** 1257-1257.

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

Pagel, M. (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* **255,** 37-45.

Pennell, M., FitzJohn, R.G., Cornwell, W.K. & Harmon, L.J. (2014) Model adequacy and the macroevolution of angiosperm functional traits. *bioRxiv***,** 004002.

Price, S.A., Hopkins, S.S., Smith, K.K. & Roth, V.L. (2012) Tempo of trophic evolution and its impact on mammalian diversification. *Proc Natl Acad Sci U S A,* **109,** 7008-7012.

Purvis, A. & Rambaut, A. (1995) Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Computer applications in the biosciences: CABIOS,* **11,** 247-251.

R Development Core Team (2013) R: A Language and Environment for Statistical Computing. Vienna, Austria.

Revell, L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution,* **3,** 217-223.

RStudio (2012) RStudio: Integrated development environment for R (Version 0.98.976). Boston, MA.

Wright, S. (1934) An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics,* **19,** 506.

**Table 1 Summary of EvobiR functions**

|  |  |
| --- | --- |
| Function | Description |
| *Applied Phylogenetics* | |
| AncCond | Tests whether a derived state of a binary character originates when a continuous character has extreme values. |
| PPSDiscrete | Produce posterior predictive datasets based on log files from programs such as BayesTraits or diversitree. |
| *Population Genetics* | |
| CalcD | Test of introgression implementing an algorithm described in (Durand *et al.* 2011) |
| WinCalcD | Sliding window version of CalcD |
| CalcPopD | Test of introgression implementing an algorithm described in (Durand *et al.* 2011) |
| *Miscellaneous* | |
| ResSel | For use in selection experiments identifies those individuals for high or low selection lines after regression of one trait on another. |
| SampleTrees | Processes large nexus files, removing burn-in, randomly sampling, and saving in various formats. |
| FuzzyMatch | Identifies records in trees and trait dataset that may be lost due to differences in spelling. |
| Even | Tests whether a number is odd or even |
| Mode | Returns the most common value in a numeric or character vector |
| SlidingWindow | Applies any function that operates on a vector to a sliding window across a vector |
| AICc | Calculates the small sample size corrected version of the Akaike information criteria based on the log likelihood, number of model parameters and sample size. |
| SuperMatrix | Combines multiple alignments with varying taxa sets into a single supermatrix. |
| *Pedagogical-Shiny Apps – each produces an interactive html page* | |
| ViewEvo | |
| wf.model | Performs population genetics simulations, and allows users to vary populations size, mutation rates, fitness, etc. |
| bd.model | Produces a set of phylogenies using a birth-death model and user selected parameters to illustrate variability in this generating process. |
| dist.model | Allows users to explore a variety of statistical distribution with reactive graphs that change as parameters are varied |