Accounting for within-species variation in continuous trait evolution on a phylogenetic network

Benjamin Teo[1], Jeffrey P. Rose[2,3], Paul Bastide[4], Cécile Ané[1,3*]

December 15, 2022

ABSTRACT
Within-species trait variation may be the result of genetic variation, environmental variation or measurement error for example. In phylogenetic comparative studies, failing to account for within-species variation has many adverse effects, such as increased error in testing hypotheses about evolutionary correlations, biased estimates of evolutionary rates, and inaccurate inference of the mode of evolution. These adverse effects were demonstrated in studies that considered a tree-like underlying phylogeny. Comparative methods on phylogenetic networks are still in their infancy. The impact of within-species variation on network-based methods has not been studied. Here, we introduce a phylogenetic linear model in which the phylogeny can be a network, to account for within-species variation in the continuous response trait assuming equal within-species variances across species. We show how inference based on the individual values can be reduced to a problem using species-level summaries, even when the within-species variance is estimated. Our method performs well under various simulation settings, and is robust when within-species variances are unequal across species. When phenotypic (within-species) correlations differ from evolutionary (between-species) correlations, estimates of evolutionary coefficients are pulled towards the phenotypic coefficients, for all methods we tested. Also, evolutionary rates are either underestimated or overestimated, depending on the mismatch between phenotypic and evolutionary relationships. We applied our method to morphological and geographical data from *Polemonium*. We find a strong negative correlation of leaflet size with elevation, despite a positive correlation within species. Our method can explore the role of gene flow in trait evolution, by comparing the fit of a network to that of a tree. We find marginal evidence for leaflet size being affected by gene flow, and support for previous observations on the challenges of using individual continuous traits to infer inheritance weights at reticulations. Our method is freely available in the `Julia` package `PhyloNetworks`.

Keywords: gene flow, hybridization, Polemonium, measurement error, phylogenetic regression, PGLS, REML

Author Affiliations
[1] Department of Statistics, University of Wisconsin - Madison, USA
[2] Department of Biology, University of Nebraska at Kearney, Kearney, NE 68849, USA
[3] Department of Botany, University of Wisconsin - Madison, USA
[4] IMAG, Université de Montpellier, CNRS, Montpellier, France

Correspondence*
Cécile Ané: cecile.ane@wisc.edu

1 Introduction

Phylogenetic Comparative Methods (PCM) are used to test hypotheses about the evolution of traits, using a time-scaled phylogeny to account for shared ancestry among species. For example, we consider here whether the evolution of leaflet size was correlated with biogeography, notably elevation and latitude, in the plant genus *Polemonium*. To address this question, we need to account for the correlation between species using their phylogenetic relationships. In this work, we deal with two complications: gene flow occurred in *Polemonium* [Rose et al., 2021], and leaflet size, elevation and latitude vary greatly among individual plants within a species.

Within-species trait variation is conventionally referred to as "measurement error" [e.g. Ives et al., 2007, Silvestro et al., 2015], which is a misleading term because it is too narrow. Models for trait evolution consider the mean value of a trait across a species, but this mean is usually calculated from a sample of individuals, not from the whole population. For most traits, individuals vary within a species, so the sample mean inevitably differs from the true species mean. Within-species trait variation can be due to many factors such as genetic differences, plasticity and environmental variation within a species, variation within the lifespan of an individual, or error in the act of measurement.

When the phylogeny is a tree, failure to account for within-species trait variation can lead to increased type I error [Harmon and Losos, 2005], biased and imprecise parameter estimates [Ives et al., 2007], and model selection biased towards more parameter-rich models [Silvestro et al., 2015, Cooper et al., 2016].

The impact of ignoring within-species trait variation has not been documented when the phylogeny is a network, with reticulations that can represent events such as gene flow or hybrid speciation. This is understandable since both theory and implementation of PCMs for networks are still in their infancy [Solís-Lemus et al., 2017, Bastide et al., 2018]. However, as patterns of reticulate evolution are increasingly being tested to explain phylogenetic discordance, it is crucial that the current suite of network-PCMs be expanded to account for within-species trait variation, and that the impact of this variation on inference be quantified in the context of reticulate evolution. In addition to better accounting for phylogenetic relatedness than tree-based PCMs, PCMs on networks can address new questions. For example, we quantify here the evidence that leaflet length was influenced by gene flow.

1.1 Existing approaches

We restrict our attention to regression PCMs, in which a response trait y (such as leaflet size) is modeled as a linear function of one or more predictor traits (such as elevation and latitude):

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,. \tag{1}$$

In this traditional model [Martins and Hansen, 1997], $\boldsymbol{y}$ and $\boldsymbol{x}$ contain the species means of the response and predictor traits. The residual terms in $\boldsymbol{\epsilon}$ capture the phylogenetic correlation between species, based on a given phylogeny and an evolutionary model. Under the Brownian motion (BM) model on a tree, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma}_b^2\boldsymbol{V})$ where $\boldsymbol{V}$ contains the times of shared ancestry as determined from the phylogeny, and σ_b^2 is a rate of variance accumulation [Harmon, 2019]. On a network, a hybrid's traits are taken to be a weighted average of its parents traits (see Discussion), and this may define multiple paths from the root to a given species. In this case, $\boldsymbol{V}$ contains the expected length of the shared paths, which can be computed efficiently without having to enumerate all the paths [Bastide et al., 2018]. Beyond the BM, more flexible models can provide a continuum between low and high phylogenetic correlation like the Ornstein-Uhlenbeck model [Hansen and Martins, 1996] or Pagel's λ [Pagel, 1999]. For these models, the phylogenetic covariance between species $\boldsymbol{V}_{\boldsymbol{\theta}}$ is parametrized by model parameters $\boldsymbol{\theta}$.

In (1), each species contributes a single value for each trait. Typically, a species' trait value is taken to be the value from one individual, or the mean over a sample of individuals, and this sample mean is effectively treated as the true species mean. To model within-species variation, we can expand (1) to model individual values rather than species means:

$$y_{ij} = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i + \delta_{ij} \tag{2}$$

where y_{ij} is the response of individual j in species i, $\boldsymbol{x}_i$ contains the predictors for species i, and δ_{ij} is the difference between y_{ij} and the mean in species i. These δ_{ij} values capture within-species variation in the response trait, and are typically assumed to be independent and normally distributed. Like in (1), the ϵ_i values capture between-species correlation due to shared ancestry. This model was used by Ives et al. [2007], whose approach is now implemented in the R package `phytools` for instance [Revell, 2012]. In their approach, the within-species error variance is estimated separately for each species and supplied by the user (as-is or via a sample from each species). These estimated variances are then "plugged-in" as

true population variances, ignoring their estimation error.

As an alternative to this "plug-in" approach, a joint estimation is used by several methods when a single observation per species is available ($j = 1$ only in (2)), such as the phylogenetic mixed model (PMM) [Lynch, 1991, Housworth et al., 2004] or `phylolm` [Ho and Ané, 2014]. With a single value per species, the covariance of the total error term $\epsilon_i + \delta_{i1}$ includes between-species correlations (for ϵ) plus an independent error variance (for δ) assumed equal across all species. All variance components are then estimated jointly, most often by maximum likelihood. With a single observation per species, there is no direct information about the variability within a species. Consequently, the "within-species variation" captured by this approach includes in fact any other variation that is independent across species, not already accounted for by the between-species model. On an ultrametric tree, this approach is equivalent to Pagel's λ model [Housworth et al., 2004, Leventhal and Bonhoeffer, 2016]. With more than one observation per species and the same individuals observed across all variables (response and predictors), an alternative to regression (2) is a correlation framework to model within-species variation in all variables [Ives et al., 2007, Felsenstein, 2008]. In that approach, a model is assumed for the evolution of all variables (rather than for residuals only) and between-species relationships are represented by phylogenetic covariances between traits instead of the $\boldsymbol{\beta}$ coefficients. In addition, within-species relationships are represented by a multivariate phenotypic covariance matrix, whereas (2) has univariate within-species variation in the response only.

For a list of implementations that account for within-species variation, see Table 7.1 of Garamszegi [2014]. All of them assume the phylogeny to be a tree. The `Julia` package `PhyloNetworks` [Solís-Lemus et al., 2017] is currently the only available implementation for PCMs on a network. Prior to this work, `PhyloNetworks` could not account for within-species variation in the response trait other than indirectly via Pagel's λ model.

1.2 Our contributions

We derived and implemented methods for model (2) along a phylogenetic network, to estimate between-species and within-species variation in the response trait, linear regression parameters, and allow for possible reticulations in the phylogeny. Our method requires that at least one species has multiple individual observations. As other regression methods based on (2), we focus on estimating the evolutionary (between-species) relationships expressed by the $\boldsymbol{\beta}$ coefficients. Phenotypic (within-species) relationships between traits are not modeled, as (2) uses the predictors via their species means only.

Our method differs in three main aspects from the most widely used implementations for trees. First, our method allows for one or more species to have a single observation, as is frequent in empirical data sets. This flexibility is linked to our assumption that all species share the same within-species variance of the response trait. We find that our method is robust to a violation of this assumption. Second, we do not assume that the error in sample means is perfectly known. Instead, the true within-species variance is estimated jointly with all other parameters. Finally, our implementation uses restricted maximum likelihood (REML) by default, as an alternative to maximum likelihood (ML) [Patterson and Thompson, 1971, Harville, 1974, LaMotte, 2007]. REML is known to help correct the underestimation of variance components typical of ML. For instance, Housworth et al. [2004] and Ives et al. [2007] showed that REML provides a less biased estimate of the total phenotypic variance and phylogenetic signal. Our implementation takes in individual-level data. This suggests a high computational cost. For example, with 30 species and 10 individuals per species, the input has 300 rows instead of 30 if the data were summarized by species means. As covariance matrices scale with the square of the number of rows, the cost of dealing with a much larger covariance matrix may be problematic. In this work, we show that the calculations for jointly estimating all parameters can be reduced to a computational complexity that scales with the number of species only. In fact, we show that the likelihood and restricted likelihood can be computed directly from the sample means, sample sizes, and sample standard deviations of each species. Hence, our implementation also admits this set of species-level information as input.

In the rest of the paper, we first explain the model, its assumptions, our derivations to lower the computational complexity of the (restricted) likelihood, and derivations for the reconstruction of species means. We present a thorough simulation study to assess the method's accuracy and robustness to assumption violations, and then illustrate the method to study leaflet size evolution in *Polemonium*, whose history was shown to involve reticulation [Rose et al., 2021].

2 Methods

2.1 Model

Model (2) models within-species variation in the response variable via the δ_{ij} term specific to individual j in species i. Let n be the number of species. We assume that we have data on m_i individuals from species i, and that $m_i \geq 2$ for at least one species. We focus on the evolutionary correlation between the response and predictor(s), that is, the correlation over evolutionary time between the response and predictor species means. However, phenotypic (within-species) correlation can differ considerably from evolutionary correlation [Felsenstein, 1988, Goolsby et al., 2017]. For example, longevity tends to increase with body mass across species, but decrease with body mass within a species [Garamszegi, 2014]. To capture evolutionary correlations specifically, (2) uses the predictors' species means, ignoring within-species variation for predictors.

We assume a time-scaled phylogeny including the n species of interest. If the phylogeneny is a network (also called admixture graph), each reticulation appears as a node with multiple parent edges, to represent an admixed population with genetic material from multiple parental lineages. The population inherits from each parent a proportion of genes, and inheritance proportions are assumed to be known. This punctual event is a simplified model for processes such as hybridization, horizontal gene transfer or gene flow, that can happen over a period of time [Huang et al., 2022].

For trait evolution, we assume a Gaussian model for the species-level residuals, such as Brownian motion (BM), Ornstein-Uhlenbeck (OU), or Pagel's λ (Pλ). For now the OU model is not implemented in `PhyloNetworks`, though the subsequent derivations would equally apply. From this model and the phylogeny, we get the n-by-n unscaled covariance matrix $\boldsymbol{V}$, which may depend on some parameters, like the selection strength α (or phylogenetic half-life) for the OU process. For the BM on a tree, the unscaled covariance V_{ik} between species i and k is the length of the shared path from the root to the most recent common ancestor of i and k. To extend the BM to networks, we follow Bastide et al. [2018]. At a reticulation, the mean of the admixed population is taken to be the weighted average of the parental populations' means. The inheritance proportions are used as weights, as reasonable for polygenic traits controlled by many additive genes. Under this model, $\boldsymbol{V}$ can be calculated in linear-time with a single traversal of the network [Bastide et al., 2018].

The evolutionary relationships are captured by a model on the true species means:

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is the vector of the n true but unobserved species means for the response trait, $\boldsymbol{x}$ is an $n \times p$ matrix of predictors (including a column of ones for the intercept), $\boldsymbol{\beta}$ is the vector of the p regression coefficients, and $\boldsymbol{\epsilon}$ are species-level residuals, assumed to be phylogenetically correlated: $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\mathrm{b}}^2 \boldsymbol{V})$. Under a BM, σ_{b}^2 is the variance rate for the between-species residuals.

But $\boldsymbol{y}$ is unobserved. Instead, we observe a larger vector $\boldsymbol{Y}$ containing the response trait of sampled individuals. With $N = m_1 + \cdots + m_n$ individuals total, $\boldsymbol{Y}$ is a vector of length N, built by stacking the values from each species above one another. We can similarly stack the δ_{ij} values from (2) into a vector $\boldsymbol{\Delta}$ of length N, starting with the m_1 values from species $i = 1$ followed by the m_2 values from species $i = 2$ and so on. We can then write (2) in matrix form as follows:

$$\boldsymbol{Y} = \boldsymbol{Z}\ (\boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}) + \boldsymbol{\Delta} \tag{3}$$

where $\boldsymbol{Z}$ is the $N \times n$ model matrix that lifts a vector of species values into a vector of individual values, by repeating the species i value m_i times. Namely, $\boldsymbol{Z}$ is made of n blocks stacked above one another, with block i of size $m_i \times n$, filled with zeros except for column i filled with ones. More specifically, $Z_{k,i} = 1$ if $m_1 + \cdots + m_{i-1} < k \leq m_1 + \cdots + m_i$ and $Z_{k,i} = 0$ otherwise.

Like earlier, we assume that the deviations from the linear relationship are phylogenetically correlated at the species-level: $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\mathrm{b}}^2 \boldsymbol{V})$. We further assume that the added within-species variation is independent with a common within-species variance σ_{w}^2: $\boldsymbol{\Delta} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\mathrm{w}}^2 \boldsymbol{I}_N)$ where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. We can then write the full $N \times N$ covariance matrix of the total residual $\boldsymbol{Z}\boldsymbol{\epsilon} + \boldsymbol{\Delta}$ as

$$\sigma_{\mathrm{b}}^2 \boldsymbol{W}_\eta \quad \text{with } \boldsymbol{W}_\eta = \boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}' + \eta \boldsymbol{I}_N$$

where $\eta = \sigma_{\mathrm{w}}^2 / \sigma_{\mathrm{b}}^2$.

2.2 Parameter estimation

If we knew η and any evolutionary parameters for $\boldsymbol{V}$, then (3) would be a standard linear model with known covariance and the following generalized least squares estimator for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}(\eta) = \left(\boldsymbol{X}'\boldsymbol{W}_\eta^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{W}_\eta^{-1}\boldsymbol{Y} \tag{4}$$

where $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{x}$. The above expression is rather unwieldy since it involves inverting and multiplying the

large $N \times N$ matrix $\boldsymbol{W}_\eta$. Fortunately, we show in appendix A that this expression can be simplified to:

$$\hat{\boldsymbol{\beta}}(\eta) = \left(\boldsymbol{x}'\boldsymbol{V}_\eta^{-1}\boldsymbol{x}\right)^{-1}\boldsymbol{x}'\boldsymbol{V}_\eta^{-1}\overline{\boldsymbol{y}} \tag{5}$$

where $\overline{\boldsymbol{y}}$ are the observed species means of the response trait and $\boldsymbol{V}_\eta$ is $n \times n$ (much smaller than $\boldsymbol{W}_\eta$) given by

$$\boldsymbol{V}_\eta = \boldsymbol{V} + \eta\boldsymbol{D}^{-1} \tag{6}$$

where $\boldsymbol{D}$ is the $n \times n$ diagonal matrix with the sample sizes on its diagonal: $D_{ii} = m_i$. Note that $\boldsymbol{V}_0 = \boldsymbol{V}$ corresponds to no within-species variation.

The estimation of the variance components σ^2_{b}, σ^2_{w} (hence their ratio η) and any parameters in $\boldsymbol{V}$ can be done via ML or REML. This is done by optimizing the corresponding likelihood criterion (twice the negative log likelihood) as a two-step approach. First, we fix η (and any parameters for $\boldsymbol{V}$) and optimize the other parameters, to obtain the *profile* criterion. In appendix A, we show that this profile criterion can be expressed using the smaller matrix $\boldsymbol{V}_\eta$ instead of the larger matrix $\boldsymbol{W}_\eta$ to lower the computational task:

$$\begin{aligned}\ell_{\mathrm{ml}}(\eta) = {} & d_{\mathrm{ml}}\log 2\pi + d_{\mathrm{ml}}\log\hat{\sigma}^2_{\mathrm{b}}(\eta) + d_{\mathrm{ml}} \\ & + (N-n)\log\eta + \sum_{i=1}^{n}\log m_i + \log|\boldsymbol{V}_\eta|\end{aligned} \tag{7}$$

$$\begin{aligned}\ell_{\mathrm{reml}}(\eta) = {} & d_{\mathrm{reml}}\log 2\pi + d_{\mathrm{reml}}\log\hat{\sigma}^2_{\mathrm{b}}(\eta) + d_{\mathrm{reml}} \\ & + (N-n)\log\eta + \sum_{i=1}^{n}\log m_i + \log|\boldsymbol{V}_\eta| \\ & - \log|\boldsymbol{x}'\boldsymbol{V}_\eta^{-1}\boldsymbol{x}|\end{aligned} \tag{8}$$

where $\hat{\sigma}^2_{\mathrm{b}}(\eta)$, defined in (9) below, depends on the criterion via the corresponding degree of freedom: $d_{\mathrm{ml}} = N$ or $d_{\mathrm{reml}} = N - p$.

We first estimate $\hat{\eta}$ (and other parameters for $\boldsymbol{V}$) as the value that minimizes ℓ above. Then, we plug $\hat{\eta}$ in the estimate of σ^2_{b} given η:

$$\hat{\sigma}^2_{\mathrm{b}}(\eta) = (\eta^{-1}\mathrm{SSW} + \|\overline{\boldsymbol{y}} - \boldsymbol{x}\hat{\boldsymbol{\beta}}(\eta)\|^2_{\boldsymbol{V}_\eta})/d \tag{9}$$

where $\mathrm{SSW} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_i)^2$ is the sum of squared residuals within species, d is d_{ml} or d_{reml}, and $\|\boldsymbol{u}\|^2_{\boldsymbol{M}} = \boldsymbol{u}'\boldsymbol{M}^{-1}\boldsymbol{u}$. Finally, we use $\hat{\sigma}^2_{\mathrm{b}} = \hat{\sigma}^2_{\mathrm{b}}(\hat{\eta})$ to estimate the within-species variance $\hat{\sigma}^2_{\mathrm{w}} = \hat{\eta}\hat{\sigma}^2_{\mathrm{b}}$, and plug $\hat{\eta}$ in (5) to estimate the regression coefficients.

For inference about phylogenetic coefficients β_k, we implemented confidence intervals and hypothesis tests based on $n - p$ degrees of freedom for σ^2_{b}. These are approximate, because η needs to be estimated (see appendix B). For more general model comparisons, we also implemented likelihood ratio tests.

2.3 Species means reconstruction

Conditional on our estimate of η (and of other parameters for $\boldsymbol{V}$), we can use our model to estimate the true species mean for any species in the phylogeny. For ancestral species, this task is traditionally called "ancestral state reconstruction". This task also applies to extant species, to predict their mean based on their predictor values and data from closely related species.

Recall that $\boldsymbol{y}$ and $\overline{\boldsymbol{y}}$ denote the true (unobserved) and the observed means, for the species with data. We further consider the true means $\boldsymbol{y}_0$ of a set of species for which a prediction is desired. This set may include ancestral species, extant species with missing response data, or species with observed data. We assume that we know their predictor values, which we call $\boldsymbol{x}_0$. For ancestral species, this is a very strong assumption, although it is reasonable if the predictor set is limited to the intercept column of ones or to discrete predictors that evolve sufficiently slowly for a reliable prediction in clades without variation. Another caveat, if using more than an intercept, is that the evolutionary regression (3) fitted to present-day species may not apply to past species. For example, consider a predictor X evolving according to a BM and a response Y adapting to it via an OU process with optimum $b_0 + b_1X$ that varies over time, as X evolves. Due to the lag time for adaptation, the "evolutionary slope" for X in $\boldsymbol{\beta}$ is attenuated from the "optimal slope" b_1, by a factor that depends on the strength of adaptation and the height of the phylogeny [Hansen et al., 2008]. As this attenuation depends on the time from the root, inertia affects ancestral species more than present-day species under this BM-OU model, and extrapolating our regression model to ancestral states should be taken cautiously. We recommend using an intercept only for predicting ancestral states. Using other predictors should be limited to predict the mean of present-day species, or when there is evidence of fast adaptation and reliable knowledge of the predictors ancestral states.

Based on our model (3) we have that:

$$\begin{bmatrix}\boldsymbol{y}_0\\ \overline{\boldsymbol{y}}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}_0\\ \boldsymbol{x}\end{bmatrix}\boldsymbol{\beta}, \sigma^2_{\mathrm{b}}\begin{bmatrix}\boldsymbol{V}_0 & \boldsymbol{V}_c\\ \boldsymbol{V}_c' & \boldsymbol{V}_\eta\end{bmatrix}\right)$$

where $\boldsymbol{V}_\eta$ is given in (6), $\boldsymbol{V}_0$ is the phylogenetic covariance among species for which prediction is sought, and $\boldsymbol{V}_c$ is the cross-covariance of the true means, between the set of species to predict and the set of species with data. Given knowledge of $\overline{\boldsymbol{y}}$, the conditional distribution of $\boldsymbol{y}_0$ is also Gaussian with mean

$$\boldsymbol{\mu}_0(\boldsymbol{\beta}) = \boldsymbol{x}_0\boldsymbol{\beta} + \boldsymbol{V}_c\boldsymbol{V}_\eta^{-1}(\overline{\boldsymbol{y}} - \boldsymbol{x}\boldsymbol{\beta})$$

and (co)variance, or prediction error variance:

$$\boldsymbol{\Sigma} = \sigma_{\mathrm{b}}^2(\boldsymbol{V}_0 - \boldsymbol{V}_c\boldsymbol{V}_\eta^{-1}\boldsymbol{V}_c')\,.$$

Conditional on η and parameters for $\boldsymbol{V}$, $\boldsymbol{\mu}_0(\hat{\boldsymbol{\beta}})$ is the best linear unbiased predictor for $\boldsymbol{y}_0$. The prediction error $\boldsymbol{y}_0 - \boldsymbol{\mu}_0(\hat{\boldsymbol{\beta}})$ has variance equal to $\boldsymbol{\Sigma}$ plus an extra term due to estimating $\boldsymbol{\beta}$ [Christensen, 2001] given in appendix C.1. If this prediction variance is ψ_i for species i (which may be an ancestral or extant), then an approximate prediction interval for the true mean of that species is $\mu_i(\hat{\boldsymbol{\beta}}) \pm t\sqrt{\psi}_i$ where t is the quantile corresponding to the desired confidence level, from the T-distribution with degree of freedom associated with $\hat{\sigma}_{\mathrm{b}}^2$ (see appendix C.2).

We note that if we have data for species i, then $\mu_i(\hat{\boldsymbol{\beta}})$ is not necessarily equal to the sample mean $\overline{y}_i$. This is because the prediction is influenced by data from closely related species. Appendix C.3 illustrates this on a simple 3-species example. If many individuals are observed for a given species, then the prediction of the true mean for that species is very close to its sample mean. If few individuals are observed instead, the predicted mean is also influenced by the linear relationship with the predictors for that species, and by observations from closely related species.

2.4 Within-species variation in predictors

The model described above ignores within-species variation for predictors, to focus on their evolutionary relationship with the response trait. Indeed, the evolutionary (between-species) and phenotypic (within-species) relationships can be different. At the extreme, two traits can be negatively correlated within each individuals species, yet positively correlated evolutionarily [Garamszegi, 2014, Fig. 7.2]. However, if the phenotypic and evolutionary relationships are similar, then some information about this common relationship is lost when ignoring within-species (phenotypic) variation in predictors.

If one is willing to assume that the regression coefficients $\boldsymbol{\beta}$ are shared within and between species and if there is within-species variation in one or more predictors, then it is appropriate to consider the individual values for each predictor without summarizing the predictor data to a single average value per species. Accordingly, we consider the following model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\epsilon} + \boldsymbol{\Delta} \tag{10}$$

where $\boldsymbol{Z}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\Delta}$ are as before. Here, the matrix of predictors $\boldsymbol{X}$ contains values at the individual level, where different individuals from the same species may have different values. In contrast, (3) imposes the constraint that $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{x}$. Note that (10) has the same parameters as (3) and the same variance $\sigma_{\mathrm{b}}^2\boldsymbol{W}_\eta$ for the total residual $\boldsymbol{Z}\boldsymbol{\epsilon} + \boldsymbol{\Delta}$, where $\boldsymbol{W}_\eta = \boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}' + \eta\boldsymbol{I}_N$.

If we assume that $\boldsymbol{V}$ is from a BM and has no extra parameter, then this model is equivalent to Pagel's λ model [Pagel, 1999] on an expanded network that has one leaf per individual, if the network is ultrametric (all leaves are equidistant from the root). Indeed, consider the expanded network constructed from the species network as follows: for each i, change the tip for species i in the original network into an internal node, then graft on this node m_i external edges of length 0 (creating a polytomy of $m_i \geq 3$ or a degree-2 node if $m_i = 1$) and label each new tip with an individual sampled from species i. This zero-length extension is similar in spirit to the approach by Felsenstein [2008]. Then, the BM covariance under this expanded network is exactly $\boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}' = \boldsymbol{W}_0$. Bastide et al. [2018] described Pagel's λ model on a network, which requires that the network be time-consistent (any two paths from the root to the same end node have the same length). If the distance from the root to every leaf is h (for height), then Pagel's λ covariance is

$$\sigma_\lambda^2\left(\lambda\boldsymbol{W}_0 + (1-\lambda)h\boldsymbol{I}_N\right) \tag{11}$$

where σ_λ^2 controls the total variance from the root to the tips, and λ is the proportion explained by the phylogeny. No phylogenetic signal corresponds to $\lambda = 0$ with independent observations, while $\lambda = 1$ corresponds to the BM. The variance from (10) equals that from Pagel's λ in (11) if we reparametrize the variance components as follows: $\sigma_{\mathrm{b}}^2 = \lambda\sigma_\lambda^2$ and $\sigma_{\mathrm{w}}^2 = (1-\lambda)h\sigma_\lambda^2$, hence $\eta = h(1-\lambda)/\lambda$.

In practice, we can fit (10) by expanding the network and using the routine developed by Bastide et al. [2018] under Pagel's λ (which we expanded to allow for the REML criterion), then re-expressing the variance parameters in terms of between and within-species variances.

Since (10) is used with a BM model, and the coefficients $\boldsymbol{\beta}$ are assumed to apply both between species and within species, this model corresponds to a BM with a phenotypic relationship constrained to match the evolutionary relationship. Therefore, we abbreviate this model as $\mathrm{BM}_{\mathrm{pheno}}$ later.

It is worth noting that the degrees of freedom for testing hypotheses about $\boldsymbol{\beta}$ is larger in model (10) than (3), because $\boldsymbol{\beta}$ is an individual-level parameter in (10) as opposed to a species-level parameter in (3). Intuitively, (10) makes a stronger assumption with respect to $\boldsymbol{\beta}$ and accordingly, it allows for a more powerful statistical test. Note that in both models,

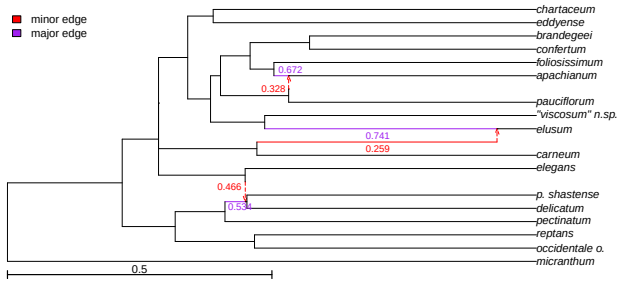

Figure 1: Calibrated 17-taxon SNaQ network. Edge lengths are normalized so that the network height is 1. The dotted vertical components of minor edges indicate the destination of gene flow, and do not contribute to length. Hybrid edges are labelled with their inheritance weights. The *major tree* of the network is obtained by deleting the minor (red) edges and setting the weights for the major (purple) edges to 1. The *minor tree* is obtained by deleting the major edges and setting the weights of the minor edges to 1 (e.g. *P. elusum* is sister to *P. carneum*, not *P. "viscosum" n.sp.*, in the minor tree).

these tests are only approximate because the variance ratio η is estimated. Tools for mixed linear models, such as Satterthwaite's or Kenward-Roger's approximation [see e.g. `R` package `lmerTest`, Kuznetsova et al., 2017] or bootstrap approaches could provide more accurate confidence intervals for fixed effects and variance parameters.

2.5 Simulations

To quantify the performance of our method and its robustness to assumptions, we used `PhyloNetworks` to simulate trait data on a network with 3 reticulations. We used the 17-taxon network on the flowering plant genus *Polemonium* estimated by Rose et al. [2021]. We calibrated it following the approach described in Bastide et al. [2018], to obtain branch lengths proportional to time instead of branch lengths in coalescent units. The resultant network is shown in Figure 1.

We describe here the most general form of our simulation model, which allows for model violation via within-species variation in the predictor and possible phenotypic correlation. Since the simulated phenotypic and evolutionary correlations may differ, our simulation model is similar to the PMM [Lynch, 1991], which has separate trait covariances for the heritable and non-heritable components (but uses a single value per species).

We simulated one predictor X with a BM with variance rate $\sigma^2_{b,x}$ and within-species variance $\sigma^2_{w,x}$:

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{x} + \boldsymbol{\Delta}_x \quad \text{with } \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{b,x}\boldsymbol{V}) \qquad (12)$$

and with $\boldsymbol{\Delta}_x \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{w,x}\boldsymbol{I}_N)$. We then simulated the response Y as a linear function of the true species mean of X, an additional phylogenetic component between species, and within-species variation possibly correlated with the within-species variation in X:

$$\boldsymbol{Y} = \boldsymbol{Z}(\beta_1\boldsymbol{x} + \boldsymbol{\epsilon}_y) + \beta_2\boldsymbol{\Delta}_x + \boldsymbol{\Delta}_y \qquad (13)$$

with $\boldsymbol{\epsilon}_y \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{b,y}\boldsymbol{V})$ from a BM with rate $\sigma^2_{b,y}$ and $\boldsymbol{\Delta}_y \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$ is within-species variation independent of the predictor, using $\boldsymbol{D} = \sigma^2_{w,y}\boldsymbol{I}_N$ as in our estimation model, unless otherwise noted. In some simulations, we set $\boldsymbol{D}$ to be diagonal with different entries for different species, that is, unequal within-species variances. With these notations, the true species means for the response are $\boldsymbol{y} = \beta_1\boldsymbol{x} + \boldsymbol{\epsilon}_y$.

Our model (3) allows for an intercept, which we fixed to $\beta_0 = 0$ in our simulations. Our model does not make any assumption on X, but assumes that the species means $\boldsymbol{x}$ are observed. This is the case if $\sigma^2_{w,x} = 0$, which implies that $\boldsymbol{\Delta}_x = \boldsymbol{0}$ and β_2 becomes irrelevant. If $\sigma^2_{w,x} > 0$, then $\boldsymbol{x}$ is unobserved, and the sample species means for X need to be used for estimation instead. In that case, our simulations violate the assumptions of our model. The phylogenetic and phenotypic relationships are equal if $\beta_1 = \beta_2$, as assumed in (10) by our λ-model on the expanded network.

In all of our simulations, we set $\sigma^2_{b,x} = 2$, $\sigma^2_{b,y} = 1$ and $\beta_1 = 1$. We set the sample sizes m_i and other parameters according to various settings, as described below, and simulated 500 data sets for each combination of parameters.

We then estimated the model parameters using various methods and ML or REML. Namely, we used the BM or Pagel's λ model that use species means, which we abbreviate as $\mathrm{BM_n}$ and $\mathrm{P}\lambda_\mathrm{n}$ (where "n" stands for "no" within-species variation). We also used model (3) under a BM, which we abbreviate as $\mathrm{BM_y}$ as it accounts for within-species variation in Y, but not in predictors. Finally, we used the $\mathrm{BM_{pheno}}$ model (10), which accounts for within-species variation in both the response and predictors but constrains the phenotypic relationship.

2.5.1 Impact of ignoring within-species variation

To assess the impact of accounting for within-species variation, we used equal sample sizes $m_i = m$ with $m = 3$ or 8, $\sigma^2_{w,y}$ in $\{0.4, 0.6, 0.8\}$, and no model violation: $\sigma^2_{w,x} = 0$. We then compared the estimates obtained with ML versus REML, and the methods that ignore or account for within-species variation.

2.5.2 Impact of unequal within-species variances

Our model (3) assumes equal variances within species. To assess the robustness of our method, we used $\boldsymbol{D}$ diagonal with entry $\sigma^2_{w,y,i}$ for species i. For each simulated data set, $\sigma^2_{w,y,i}$ was set to a "low" value σ^2_{lo} for 9 species and to a "high" value σ^2_{hi} for 8 species. Species were randomly re-assigned to a low or high variance for each simulated data set. Variances $(\sigma^2_{\text{lo}}, \sigma^2_{\text{hi}})$ were set to $(0.2, 0.2)$, $(0.2, 0.4)$ or to $(0.2, 0.8)$. We then compared the BM methods that account for within-species variation, with either ML or REML.

We again used equal sample sizes $m_i = m$ with $m = 3$ or 8 and $\sigma^2_{w,x} = 0$.

2.5.3 Impact of unequal sample sizes

In most empirical data sets, sample size variation can be substantial, with one or a few individuals from rare species, to hundreds of individuals from abundant species. To assess the impact of sample size variation, we simulated data under the same settings as in 2.5.1 except for the sample sizes m_i. Like in 2.5.1, the average sample size $\overline{m}$ was set to either 3 or 8. For $\overline{m} \approx 3$, species were randomly assigned a sample size such that 5 species had $m_i = 1$, 6 species had $m_i = 3$ and 6 species had $m_i = 5$, leading to a total of 53 individuals and $\overline{m} = 3.12$. Species were re-assigned to sample sizes for each simulated data set. For $\overline{m} \approx 8$, species were similarly randomly assigned such that 6 species had $m_i = 2$, 6 species had $m_i = 8$ and 5 species had $m_i = 15$, for a total of 135 individuals and $\overline{m} = 7.94$.

We compared the ML and REML methods that account for within-species variation in this setting to the setting when all $m_i = 3$ or all $m_i = 8$ from earlier.

2.5.4 Within-species variation in the predictor

Our method assumes no within-species variation in X. If present, this variation is ignored in practice and the sample species means are used for X. To assess robustness to a violation of this assumption, we simulated data as in 2.5.1 except that we set $\sigma^2_{w,x}$ to be non-zero. Specifically, we set $\sigma^2_{w,x} = \sigma^2_{w,y}$. Within-species variation in X was uncorrelated with Y, that is, we set $\beta_2 = 0$ to simulate the absence of phenotypic correlation. We then used the methods that ignore or account for within-species variation in the response Y, with ML or REML.

2.5.5 Impact of phenotypic correlation

We ran the same basic settings as in 2.5.1, except that we simulated within-species variation in X and phenotypic correlation by setting β_2 to be in $\{-1, 1, 2\}$, so that within each species Y is correlated with X with regression coefficient β_2. When β_2 is set to 1, the phenotypic and evolutionary coefficients are equal, as assumed by the λ-model on the expanded network. When β_2 is set to -1, the phenotypic and evolutionary relationships are opposite. Also, we set $m = 8$ and $\sigma^2_{w,x} = 4\sigma^2_{w,y}$ with $\sigma^2_{w,y}$ set in $\{0.1, 0.15, 0.2\}$. We used smaller values for $\sigma^2_{w,y}$ here than in 2.5.1 because the total within-species variance in Y is $\beta_2^2\sigma^2_{w,x} + \sigma^2_{w,y}$, with values comparable to that in previous settings.

We then compared the estimates obtained with REML for methods ignoring within-species variation (BM_n and $\text{P}\lambda_\text{n}$), and accounting for within-species variation in Y (BM_y) or in both Y and X (BM_pheno).

2.6 *Polemonium* leaflet size evolution

2.6.1 Objectives

We applied our method on morphological and geographical data from the flowering plant genus *Polemonium* (Polemoniaceae). *Polemonium* is widespread in North America and northern Eurasia, occurring across a broad latitudinal range from central Mexico to northern Alaska. Within its range, species of *Polemonium* can be found from sea level to the alpine zone of mountains. Vegetatively, leaves are deeply dissected (compound) into multiple leaflets. Attendant with the broad ecological amplitude of the genus is extreme variation in leaflet size across species, giving an opportunity to explore the relationship between leaflet traits and ecological predictors while accounting for phylogenetic correlation among species and trait variation within and among species.

An overall trend well-demonstrated in the ecological literature is a decreased size of vegetative structures within and across species at increasing elevations. It is thought that the wider boundary layer of large leaves (or their functional analogues) makes heat exchange more difficult and therefore large leaves are more susceptible to frost damage than small leaves [Körner et al., 1989, Wright et al., 2017]. Any relationship between morphological traits and elevation may be confounded by latitude, as high latitude communities are expected to be more ecologically similar to high elevation communities at low latitudes. Specifically, we hypothesized that leaflet size would tend to be larger in species found in low elevation, low latitude communities and smaller in species from high elevation or high latitude communities.

Because Rose et al. [2021] found evidence for reticulate evolution in *Polemonium*, we additionally investigated whether leaflet size is a trait that could have

been carried along with any gene flow events. Specifically, we can test if hybridization is useful to explain residual variation, beyond the variation explained by geographical predictors.

Finally, we sought to quantify how modeling choices may impact conclusions for this dataset. As described below, choices included (1) using ML versus REML, (2) using a tree that ignores reticulation but has more taxa (therefore more data) versus using a network that better represents the phylogenetic signal but has fewer taxa, and (3) accounting for or ignoring within-taxon variation.

2.6.2 *Polemonium* phylogeny

We conducted two sets of analyses, each using one of two phylogenies from Rose et al. [2021]: a 17-taxon species network inferred with SNaQ from 325 nuclear genes (Fig. 1), and a 48-taxon species tree inferred with ASTRAL from 316 genes (Fig. 2). The taxa in the network are a subset of the taxa in the tree (tips in blue in Fig. 2) because network inference methods are limited in the number of taxa they can handle. We pruned all outgroups from the ASTRAL tree. We then calibrated each phylogeny following the approach described by Bastide et al. [2018] to obtain branch lengths proportional to time. This approach uses the branch lengths in substitutions per site in the gene trees, while accounting for gene tree discordance. In total, these two trees contained up to 30 ingroup accessions that represent unique taxa (a named species or infraspecific taxon defined by morphological traits), which we will hereafter refer to as "morphs". These morphs may or may not be monophyletic based on molecular data.

All 17 taxa in the network corresponded to a unique morph. The tree contained 18 morphs represented by a single tip while 12 morphs were represented by two or more tips, yielding 27 tips that could not be uniquely mapped to a morph. Because our morphological data is at the species and not population level (see next) we could not assign trait values to individual tips of morphs containing multiple samples, we selected a single tip per morph in all possible ways. For 8 duplicated morphs, all tips formed a monophyletic group in the tree. Since the tree is ultrametric, the choice of the representative tip did not affect the resulting pruned tree, so we chose one tip and deleted the others. For the 4 non-monophyletic morphs (*eximium*, *pulcherrimum p.*, *chartaceum*, *californicum*), each one was represented by 2 tips. Because the choice of tip affects the covariance matrix, we therefore considered the $2^4 = 16$ trees obtained by choosing one of the 2 tips to represent each morph,

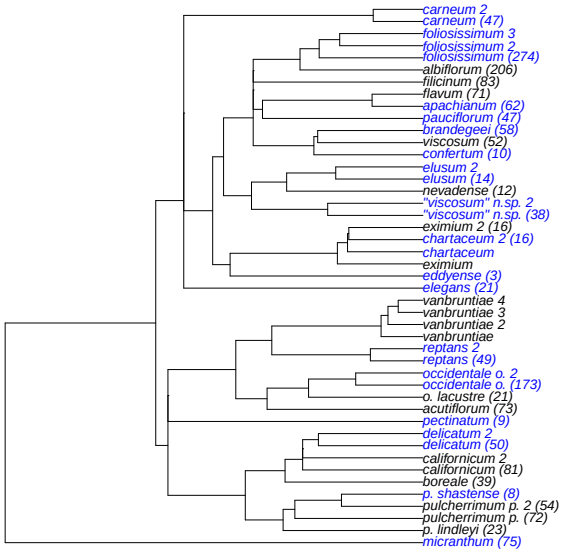

Figure 2: Calibrated 45-taxon *Polemonium* tree, after removing outgroups from the ASTRAL tree from Rose et al. [2021]. Edge lengths are proportional to time and normalized to a tree height of 1. The tips are labelled with their morph name, possibly with an extra index when multiple tips are from the same morph (e.g. 4 tips are from *vanbruntiae*). Morphs sampled in the network (Fig. 1) are in blue. Specimen counts for each morph are shown in parentheses, and indicate the tips retained to prune the tree to one tip per morph in Tables 1, 3 and 4.

pruning the other one from the tree. Each analysis was repeated on the 16 trees, each with a single tip per morph. One of these trees is shown in Figure 2.

To assess the signature of reticulation on leaflet evolution, we further considered two trees displayed in the 17-taxon network. First, we considered the major tree obtained by keeping all 3 major hybrid edges (which contributed a proportion of genes $\gamma > 0.5$ to their child hybrid node) and deleted the 3 minor hybrid edges (with $\gamma < 0.5$) from the network. Second, we considered the "minor" tree obtained by keeping the minor edges and deleting the major hybrid edges from the network (Fig. 1). The SNaQ network and the ASTRAL tree are mostly in agreement. The major tree differs from the ASTRAL tree in the placement of *P. pectinatum* and *P. pauciflorum* (Fig. 2).

2.6.3 Morphology and geography data

We obtained leaflet length and width, latitude, and elevation data for all 30 *Polemonium* morphs with molecular data. Data previously published for a subset of morphs [Rose, 2021] were combined with newly generated data obtained from imaged specimens from the Consortium of Intermountain Herbaria[1], Consor-

[1] `https://intermountainbiota.org/portal/`

tium of Pacific Northwest Herbaria[2], Consortium of California Herbaria[3], or loans from other herbarium collections made to JPR. Images were measured using Fiji [Schindelin et al., 2012], measuring multiple leaflets per specimen when feasible, and then averaged to obtain a specimen mean. For imaged specimens, if coordinates were present but elevation was missing, elevation was extracted from the WorldClim 2 elevation shapefile at 30-arc-second resolution [Fick and Hijmans, 2017] using the R package raster [Hijmans, 2020]. Leaflet data was obtained for between 3 to 275 specimens per morph (Fig. 2, 1757 specimens total). For the $\mathrm{BM_y}$ model, we used leaflet size from all 1757 imaged specimens, and we used the median latitude and elevation for each morph, calculated using all specimens, imaged or not ($>$ 11000 specimens total, from 3 (latitude) and 5 (elevation) to $>$ 1300 per morph). For the $\mathrm{BM_{pheno}}$ model, we only used imaged specimens for which latitude and elevation data could be extracted (997 specimens total, 2-218 per morph).

2.6.4 Comparative analyses

For phylogenetic regression, we considered the following response variables: log leaflet width, log leaflet length, or log leaflet area, where the area a was estimated from the length ℓ and width w assuming an ellipsoid shape: $a = \pi\ell w/4$. We used the natural log, a choice that impacts the interpretation of regression coefficients. We log-transformed these variables because their within-morph variance was strongly positively correlated with the mean, violating the equal-variance assumptions of our regression model. After the log transformation, the within-morph variance was stable across morphs and uncorrelated with the mean response (Fig. S1).

Using both elevation and latitude as predictors, we analyzed each measure of leaflet size using $\mathrm{BM_y}$ with REML on all phylogenies, to investigate leaflet size evolution and the signature of gene flow.

To assess the impact of model choice, we ran more extensive analyses on leaflet length, since all 3 measures showed strong positive correlation among themselves (Fig. S2). For leaflet length we used 6 methods on the full data set: ignoring ($\mathrm{BM_n}$, $\mathrm{P}\lambda_\mathrm{n}$) or accounting for ($\mathrm{BM_y}$) within-morph variation in leaflet size, using either ML or REML. We then restricted the data to specimens that had both morphological and geographical data (997 specimens), so as to use $\mathrm{BM_{pheno}}$. We also used $\mathrm{BM_y}$ on this data subset, to see if differences between analyses were driven by model choice or data reduction.

For each model, we recorded the coefficient estimates and their p-values, the estimated variance-component(s) and the Akaike information criterion (AIC) [Akaike, 1974]. For $\mathrm{P}\lambda_\mathrm{n}$, we conducted a likelihood ratio test of $\lambda = 1$, by comparing $\mathrm{P}\lambda_\mathrm{n}$ to the simpler $\mathrm{BM_n}$ model.

To study the impact of within-morph variation, we repeated the above analyses 100 times, each time using only a subset of 3 specimens per morph, randomly sampled without replacement from each morph.

2.6.5 Leaflet size reconstruction

We demonstrate reconstruction of species means using the 17-taxon *Polemonium* network, and under the $\mathrm{BM_y}$ model with REML. To assess the effects of sample size and of model predictors on the prediction of the true mean for morphs with observed data, we predicted log leaflet length for the two morphs with smallest and the greatest number of specimens, and using or ignoring elevation and latitude as predictors. To assess the impact of node age on the uncertainty of the predicted log leaflet length, we measured the length of the prediction interval at nodes of various ages: the hybrid node ancestor to *elusum*, its minor parent, and the root (Fig. 1). For predicting ancestral states, we used a model restricted to an intercept only, as recommended above.

3 Results

3.1 Simulations

3.1.1 No within-species variation in the predictor

Under settings without within-species variation in X (sections 2.5.1 to 2.5.3), $\hat{\beta}_1$ was unbiased (based on testing $\mathbb{E}\hat{\beta}_1 = 0$ from 500 observed replicates using a t-test: $p > 0.01$ in all settings). The accuracy of $\hat{\beta}_1$ was comparable across different models, even with unequal within-species variances or varying sample sizes (Figs. 3 to 5, top).

Bias in $\hat{\sigma}^2_{b,y}$ showed more sensitivity across different settings (Figs. 3 to 5, bottom). Namely, $\mathrm{BM_y}$'s estimate of $\hat{\sigma}^2_{b,y}$ with REML was unbiased, even when sample sizes were variable or when the within-species variance varied across species. Ignoring within-species variation resulted in overestimating $\hat{\sigma}^2_{b,y}$, more so at smaller sample sizes. Using ML instead of REML resulted in lower estimates of $\hat{\sigma}^2_{b,y}$, especially at smaller sample sizes. This underestimation was slightly exacerbated when sample sizes were variable.

[2] https://www.pnwherbaria.org/
[3] https://ucjeps.berkeley.edu/consortium/

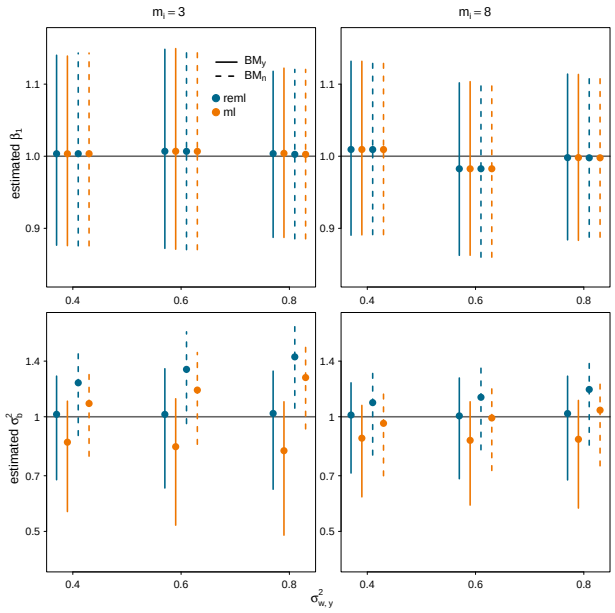

Figure 3: Simulations with within-species variation in Y, from section 2.5.1. Top: Estimated slope $\hat{\beta}_1$. The true slope $\beta_1 = 1$ is indicated by a horizontal line. Bottom: Estimated between-species variance rate on a logarithmic scale. The true value $\sigma^2_{b,y} = 1$ is indicated by a horizontal line. For each within-species variance $\sigma^2_{w,y}$ and sample size m_i per species, the dots and vertical bars respectively indicate the mean and 25th-75th percentile of estimates.

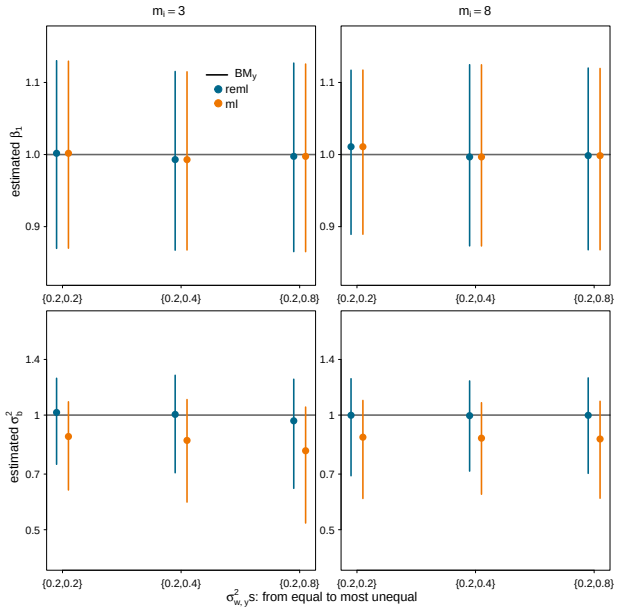

Figure 4: Simulations with unequal within-species variances for Y, from section 2.5.2. About half of species had a "low" variance and the other half had a "high" variance. Estimation used models accounting for within-species variation, but assuming equal variances.

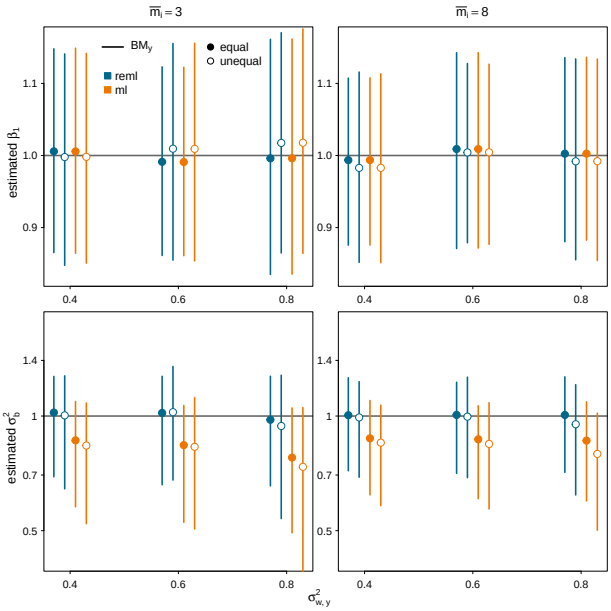

Figure 5: Simulations with unequal numbers of individuals per species, with average $\overline{m}_i$ across species, from section 2.5.3. Filled dots: all species had equal sample sizes. Empty dots: species had a sample size of 1, 3 or 5 when $\overline{m}_i = 3$, and of 2, 8, or 15 when $\overline{m}_i = 8$.

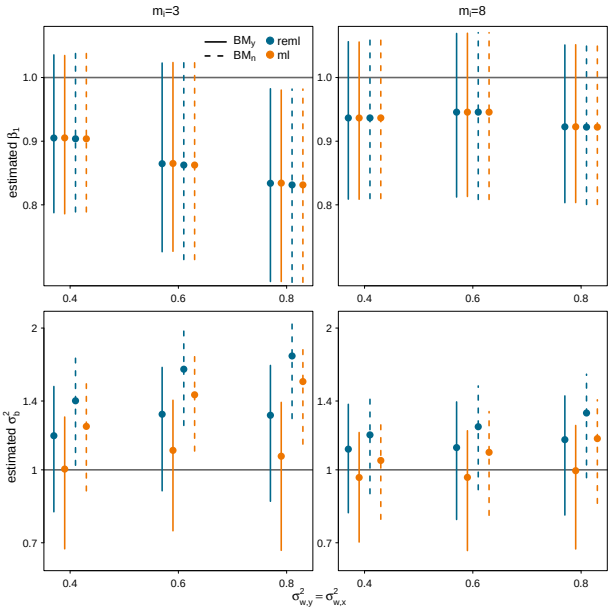

Figure 6: Simulations with variation within species in both Y and X but no phenotypic correlation, from section 2.5.4.

3.1.2 Within-species variation in the predictor

When within-species variation was simulated for X (sections 2.5.4 and 2.5.5), $\hat{\beta}_1$ was pulled towards the true value of β_2. With no phenotypic correlation, $\beta_2 = 0$ so $\hat{\beta}_1$ was attenuated towards 0. Since $\beta_1 = 1$, all methods underestimated β_1, especially so with a smaller sample size (Fig. 6, top). In general, the pull towards β_2 was similar across methods that ignore within-species variation in X ($\mathrm{BM_n}$, $\mathrm{P}\lambda_\mathrm{n}$, and $\mathrm{BM_y}$, see Figs. 6 to 7, top) and extremely high under $\mathrm{BM_{pheno}}$ (Fig. 8, top).

Like before, using ML instead of REML leads to a smaller estimate of the evolutionary variance rate $\hat{\sigma}^2_{b,y}$, and ignoring within-species variation leads to a larger estimate (Figs. 6 to 7, bottom). However, our method $\mathrm{BM_y}$ gave a biased estimate of $\hat{\sigma}^2_{b,y}$ in the presence of within-species variation in X, with an upward bias when $\beta_2 = -1$, little bias when $\beta_2 = \beta_1 = 1$ and downward bias when $\beta_2 = 2$. This bias was exacerbated as $\sigma^2_{w,x}$ increased.

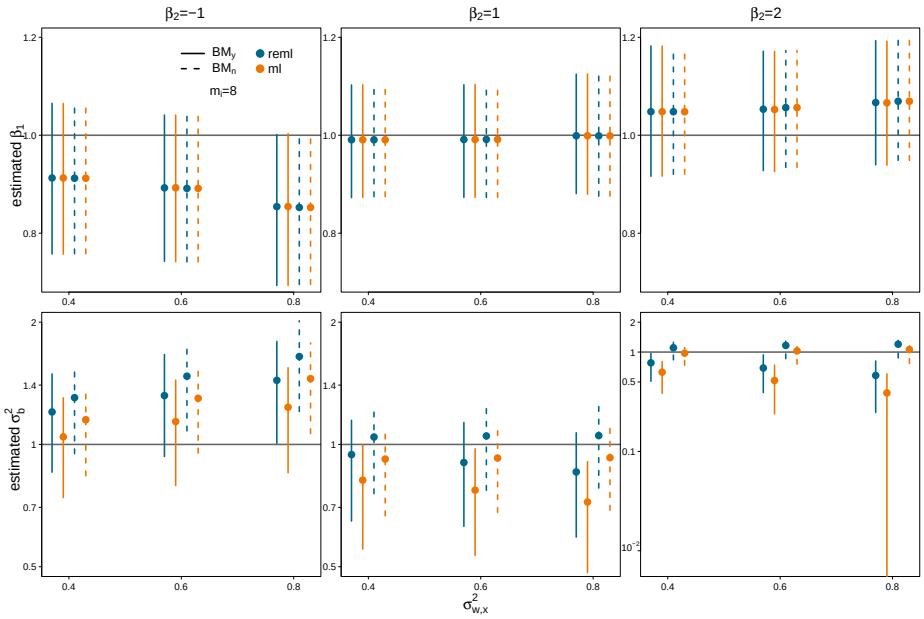

Figure 7: Simulations with phenotypic correlation (slope β_2 within species), from section 2.5.5. Top: opposite phenotypic and evolutionary relationships ($\beta_2 = -\beta_1$). Middle: identical phenotypic and evolutionary relationships. Bottom: stronger phenotypic relationship ($\beta_2 = 2\beta_1$).

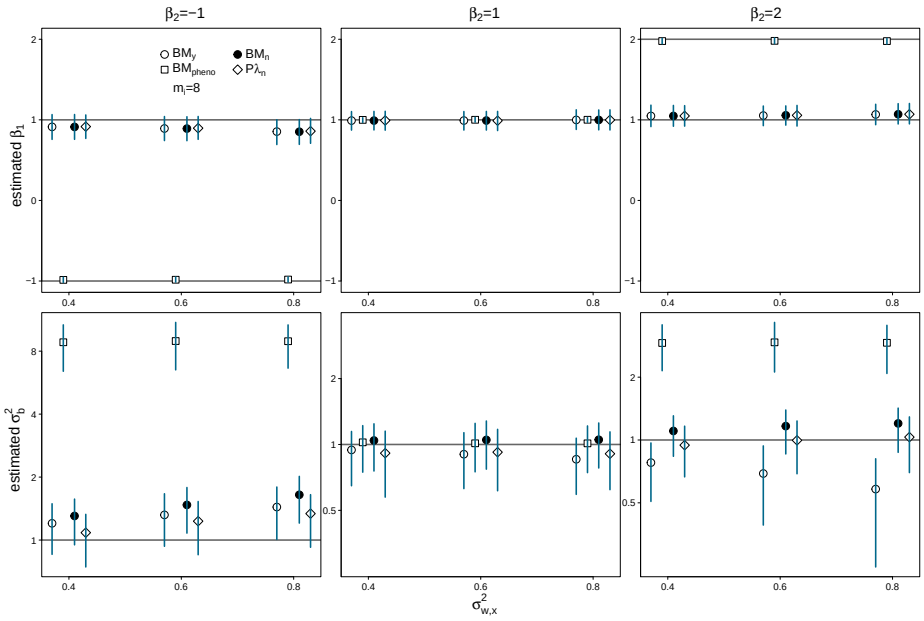

Figure 8: Simulations with phenotypic correlation as in Figure 7, and its impact on $\mathrm{BM_{pheno}}$, which assumes identical phenotypic and evolutionary relationships. The REML criterion was used for all methods: $\mathrm{BM_n}$, $\mathrm{P}\lambda_\mathrm{n}$, $\mathrm{BM_y}$ and $\mathrm{BM_{pheno}}$.

3.1.3 Impact of within-species variation in X

To theoretically explain the bias in $\hat{\beta}_1$ and $\hat{\sigma}^2_{b,y}$ when within-species variation in X or phenotypic correlation is misspecified by the model, we derived the true distribution of $\boldsymbol{Y}$ conditional on the observed species means $\overline{\boldsymbol{x}}$, under our simulation settings. In appendix D, we show exact expressions that simplify, when m is large or $\sigma^2_{w,x}$ is low, to:

$$\begin{aligned} \mathbb{E}(\boldsymbol{Y} \mid \overline{\boldsymbol{x}}) &= \boldsymbol{Z}\mathbb{E}(\overline{\boldsymbol{y}} \mid \overline{\boldsymbol{x}}) \text{ with} \\ \mathbb{E}(\overline{\boldsymbol{y}} \mid \overline{\boldsymbol{x}}) &\approx \beta_1\overline{\boldsymbol{x}} + (\beta_2 - \beta_1)u\boldsymbol{V}^{-1}\overline{\boldsymbol{x}} \end{aligned}$$

where $u = \sigma^2_{w,x}/(\sigma^2_{b,x}m)$. This relationship explains why our assumed evolutionary slope β_1 is correctly specified if $m \to \infty$ or $\sigma^2_{w,x} = 0$ or $\beta_2 = \beta_1$. It also shows that the bias in $\hat{\beta}_1$ is expected to be in the direction of $\beta_2 - \beta_1$, hence the pull towards β_2.

For the residual variance, appendix D shows that

$$\begin{aligned} \text{var}(\boldsymbol{Y} \mid \overline{\boldsymbol{x}}) &= \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}' + (\beta_2^2\sigma^2_{w,x} + \sigma^2_{w,y})\boldsymbol{I}_N \\ \text{var}(\overline{\boldsymbol{y}} \mid \overline{\boldsymbol{x}}) &= \boldsymbol{\Sigma} + (\beta_2^2\sigma^2_{w,x} + \sigma^2_{w,y})/m\ \boldsymbol{I}_n \\ \text{where } \boldsymbol{\Sigma} &\approx \sigma^2_{b,y}\boldsymbol{V} + \beta_1(\beta_1 - 2\beta_2)\frac{\sigma^2_{w,x}}{m}\boldsymbol{I}_n\ . \end{aligned}$$

In comparison, our model BM_y assumes $\boldsymbol{\Sigma} = \sigma^2_{b,y}\boldsymbol{V}$. Therefore, if we focus on the diagonal terms in $\boldsymbol{V}$ (which are the largest) and their average $\overline{v}$, then our model expects these terms to be around $\sigma^2_{b,y}\overline{v}$, while the data will provide values around $\sigma^2_{b,y}\overline{v} + \beta_1(\beta_1 - 2\beta_2)\frac{\sigma^2_{w,x}}{m}$. Based on these diagonal terms only, we can expect $\hat{\sigma}^2_{b,y}$ to be around $\sigma^2_{b,y} + \beta_1(\beta_1 - 2\beta_2)\frac{\sigma^2_{w,x}}{m\overline{v}}$, which explains a positive or negative bias, depending on how β_1 compares to $2\beta_2$. If $\beta_2 = 0$ for example, then we expect a positive bias (overestimation) as observed in 2.5.4 (Fig. 6, bottom). If $\beta_2 = -\beta_1$, then we also expect a positive bias. But if $\beta_2 = 2\beta_1$, then we expect a negative bias (underestimation). This is indeed what we observed in 2.5.5 (Fig. 7, bottom).

3.2 *Polemonium* leaflet size

3.2.1 Small leaflets correlate with high elevation

The 16 ASTRAL subtrees provide extremely similar results, with parameter estimates that do not exceed 1% difference among one another (Table S1). Table 1 shows the results from one of these subtrees, selected because it gave the lowest AIC for all of the leaflet size variables.

Elevation and latitude correlate negatively with leaflet size regardless of the measure used for leaflet size (length, width or area) or of the phylogeny (Table 1). Using the network and its 17 taxa, the elevation coefficient is negative with strong evidence for area and length, and moderate evidence only for width. The latitude coefficient is negative with moderate evidence for all 3 measures of leaflet size. Using the tree and its larger set of taxa, both elevation and latitude are negative with very strong evidence for area and length, and strong evidence for width. The p-values are smaller on the tree than on the network. This is unsurprising since the tree has almost twice the number of taxa (from 17 to 30) and specimens (from 954 to 1757). Therefore, if a true relationship exists, then the tree is expected to have more power to detect it, unless model misspecification due to using a tree causes a decrease in power (if a true relationship exists) or an increase in type-1 error (if no relationship exists). Here, the effect of phylogenetic placement is expected to be minor because the network and tree are in good agreement: the tree pruned to 17 taxa is displayed in the network, except for a small change in the position of *pectinatum*.

	elevation	latitude	$(\hat{\sigma}^2_\text{b},\ \hat{\sigma}^2_\text{w})$
17-taxon network			
area	-0.91, 0.0051	-0.16, 0.030	(1.6, 0.40)
length	-0.59, 0.0026	-0.078, 0.069	(0.54, 0.11)
width	-0.32, 0.060	-0.085, 0.04	(0.51, 0.12)
30-taxon subtree			
area	-1.0, 9.2×10^{-5}	-0.11, 5.3×10^{-4}	(1.6, 0.36)
length	-0.65, 2.1×10^{-5}	-0.069, 2.1×10^{-4}	(0.53, 0.099)
width	-0.36, 0.0052	-0.042, 0.011	(0.48, 0.11)

Table 1: Results from fitting our model BM_y with REML on *Polemonium* data, to explain variation in leaflet area, length and width (each log-transformed) using elevation and latitude as predictors simultaneously. In the elevation and latitude columns, the first value is the estimated regression coefficient followed by the p-value to test that the coefficient is 0, colored in red when < 0.05.

The coefficients are very stable across the two phylogenies. They remain negative across all three responses. The results for area are consistent with the results for length and width. On the network for instance, the elevation coefficients for log(area), log(length) and log(width) in Table 1 translate to an expected decrease in area, length and width by approximately 40%, 55% and 73% for a 1000-meter increase in elevation.[4] These changes are consistent with one another since $0.40 \approx 0.55 \cdot 0.73$.

The estimated variance components (σ^2_b and σ^2_w) are also very stable across the two phylogenies (within 10% of each other). The ratio of within-to-between species variances, $\hat{\sigma}^2_\text{w}/\hat{\sigma}^2_\text{b}$ is relatively stable across measures of leaflet size (ranging from 0.18 to 0.26).

[4] $e^{-0.908531} \approx 0.40$, $e^{-0.589785} \approx 0.55$, $e^{-0.319361} \approx 0.73$.

tree	elevation	latitude	$(\sigma^2_{\mathrm{b}},\ \sigma^2_{\mathrm{w}})$	Δ_{AIC}
area				
major	-0.91*	-0.16*	(1.6, 0.40)	0.56
minor	-0.93*	-0.18*	(1.6, 0.40)	0.28
length				
major	-0.60*	-0.078	(0.53, 0.11)	0.40
minor	-0.56*	-0.087	(0.54, 0.11)	0.18
width				
major	-0.31	-0.082	(0.51, 0.12)	0.87
minor	-0.36*	-0.089*	(0.50, 0.12)	0.13

Table 2: Same as Table 1, but without reticulation: using either the major or minor tree displayed in the 17-taxon network. The elevation and latitude columns contain the estimated regression coefficient, with a star (*) to indicate an associated p-value < 0.05. The last column contains the change in AIC resulting from removing reticulations: $\Delta_{\mathrm{AIC}} = \mathrm{AIC(tree)} - \mathrm{AIC(network)}$. The positive values indicate that the network provides a better fit than the tree in all cases.

3.2.2 Gene flow may explain residual variation

To test the importance of gene flow in leaflet size evolution, we compared the fit of the network to the fit of trees on the same 17-taxon data. We considered two tree models, using the "major" and the "minor" trees displayed in the network, representing the largest and the smallest proportions of the genome respectively, based on inheritance along hybrid edges in the network.

Regression coefficients estimated from these trees are fairly similar to those from the network, and the qualitative conclusions about evolutionary correlations with elevation and latitude are mostly unchanged (Table 2 and upper rows of Table 1 highlighted in cyan). The percent change in the trees' estimates compared to the network's estimates ranges from 0.09-13% for elevation, 0.93-11% for latitude, and 0.88-2.0% for σ^2_{b}.

The change in AIC from the network to a tree $\Delta = \mathrm{AIC(tree)} - \mathrm{AIC(network)}$ is positive regardless of which tree or response variable is used, supporting our hypothesis that gene flow explains residual variation: a reticulate network is a better representation of leaflet size evolution than a tree. However, $\Delta < 1$ in all cases, meaning that the network is not especially helpful for explaining residual variation in the model beyond what can already be explained using either tree. Similarly, the minor tree's AIC is better than but close to the major tree's AIC, suggesting that the leaflet data is only marginally better explained by the minor tree than by the major tree.

3.2.3 Impact of modeling choices

For log leaflet length, we explored the impact of three modeling choices: using ML instead of REML, ignoring within-species variation, and fitting $\mathrm{BM_{pheno}}$. We begin by addressing the first two, which involve comparing $\mathrm{BM_y}$, $\mathrm{BM_n}$ and $\mathrm{P}\lambda_{\mathrm{n}}$, fitted with ML and REML. The estimated coefficients for elevation and latitude were very stable across these methods, as were the magnitude of their associated p-values (Table 3). The estimated within-species variance $\hat{\sigma}^2_{\mathrm{w}}$ was also fairly stable across methods that estimated it. This may be due to around a third of the morphs having large sample sizes (>50 specimens). The method choice had most impact on the estimated evolutionary variance rate, σ^2_{b}.

Using ML instead of REML caused a large decrease in $\hat{\sigma}^2_{\mathrm{b}}$ under $\mathrm{BM_y}$, by 18% on the network and 10% on the tree (Table 3). The smaller $\hat{\sigma}^2_{\mathrm{b}}$s resulted in smaller standard errors for the coefficients under $\mathrm{BM_y}$, and hence smaller p-values. In our study, this moderate decrease (<0.03) in p-values changed the qualitative conclusions for latitude only. But in other studies, a decrease in p-values may result in more (or more drastic) qualitative changes in conclusions, and possibly inflated type I error rates from using ML compared to REML.

Using $\mathrm{P}\lambda_{\mathrm{n}}$, the choice of ML versus REML can lead to an extreme difference in the estimated λ and apparently contradictory conclusions: from no phylogenetic correlation when $\hat{\lambda} = 0$ under ML, to almost full phylogenetic correlation as expected from the BM when $\hat{\lambda} = 0.95 \approx 1$ under REML. This large change occurs on the network only (Table 3), which has only about half of tree's taxa, and is therefore less informative about phylogenetic correlation. The contradiction disappears when we use a likelihood ratio test. Using the network and either ML or REML, the likelihood is rather flat, such that there was no evidence to reject the hypothesis of a BM ($\lambda = 1$), and also no evidence to reject the lack of phylogenetic signal ($\lambda = 0$). From the larger taxon set, using either ML or REML, there was no evidence to reject $\lambda = 1$ but strong evidence to reject $\lambda = 0$. This finding supports our use of the BM model when accounting for within-species variation ($\mathrm{BM_y}$).

That $\hat{\lambda} > 1$ on the tree is surprising, because it means that within-species variation in leaflet length is estimated at 0. This may reflect error in the estimated tree topology or branch lengths. This may also be due to large sample sizes, leading to small standard errors in the estimated species means: 13 of the 30 morphs in the tree have over 50 specimens. On the network, fewer morphs (5 out of 17) have > 50 specimens, and $\hat{\lambda}$ is smaller. In data sets with small sample sizes, within-species variation causes greater error in species means, and we expect Pagel's λ to capture within-species variation as part of the non-phylogenetic signal. Indeed, $\hat{\lambda}$ was smaller when subsetting our data with only 3 specimens per morph.

method	elevation	latitude	$\hat{\sigma}^2_b$	$\hat{\sigma}^2_w$	$\hat{\lambda}$
17-taxon network					
BM_y, REML	-0.59	-0.078	0.54	0.11	
	0.0026	*0.069*			
BM_y, ML	-0.59	-0.078	0.45	0.11	
	0.0012	*0.047*			
BM_n, REML	-0.59	-0.078	0.55		
	0.0025	*0.069*			
BM_n, ML	-0.59	-0.078	0.45		
	0.0025	*0.069*			
$P\lambda_n$, REML	-0.59	-0.079	0.53		0.95
	0.0024	*0.066*			*0.94*
$P\lambda_n$, ML	-0.61	-0.088	0.32		0.0
	0.0016	*0.047*			*0.30*
30-taxon subtree					
BM_y, REML	-0.65	-0.069	0.53	0.099	
	2.1×10^{-5}	*2.1×10^{-4}*			
BM_y, ML	-0.65	-0.069	0.48	0.099	
	9.8×10^{-6}	*1.1×10^{-4}*			
BM_n, REML	-0.65	-0.069	0.54		
	2.0×10^{-5}	*2.1×10^{-4}*			
BM_n, ML	-0.65	-0.069	0.49		
	2.0×10^{-5}	*2.1×10^{-4}*			
$P\lambda_n$, REML	-0.64	-0.070	0.64		1.11
	1.9×10^{-5}	*1.9×10^{-4}*			*0.19*
$P\lambda_n$, ML	-0.64	-0.070	0.57		1.11
	1.9×10^{-5}	*1.9×10^{-4}*			*0.21*

Table 3: Analysis of leaflet length (log-transformed) on the full data set. BM_n and $P\lambda_n$ ignore within-species variation. In the elevation and latitude columns, the first value is the estimated coefficient. The second italicized value is the associated p-value, in red when < 0.05. The maximal possible λ (1.08 for the network and 1.14 for the subtree) may exceed 1 depending on the terminal branch lengths in the phylogeny. $\lambda > 1$ means greater phylogenetic correlation than under a BM. In the $\hat{\lambda}$ column, the first value is the estimate and the second italicized value is the p-value from the likelihood ratio test of $\lambda = 1$, which corresponds to the BM model.

For instance, using REML, the mean $\hat{\lambda}$ across 100 subsets was 0.85 (vs 0.95) under the network.

In simulations, $\hat{\sigma}^2_b$ tended to be substantially larger when within-species variation was ignored, in which case $\hat{\sigma}^2_b$ needs to compound between- and within-species variation. For leaflet length however, modeling versus ignoring within-species variation had little impact on $\hat{\sigma}^2_b$. This again may be due to the large number of specimens. Using REML, when only 3 specimens were subsampled per morph, the $\hat{\sigma}^2_b$s for 100 such subsets were on average about 9% larger (9.48% for the tree, 8.54% for the network) when within-species variation was ignored.

To fit BM_{pheno} and assess the impact of assuming equal phenotypic and phylogenetic relationships, we needed to reduce the data to specimens with both leaflet size and geographic data (997 out of 1757). Reducing the data had little effect on estimates and conclusions using BM_y (Table 3 for the full data, Table 4 for the reduced data). Using BM_{pheno}, however, had a quite drastic effect compared to using BM_y.

method	elevation	latitude	$\hat{\sigma}^2_b$	$\hat{\sigma}^2_w$
17-taxon network				
BM_{pheno}	-0.089	0.023	0.89	0.091
	0.0039	*5.0×10^{-4}*		
BM_y	-0.59	-0.062	0.57	0.096
	0.0035	*0.13*		
fixed	-0.078	0.025		0.091
effects	*0.013*	*2.0×10^{-4}*		
30-taxon subtree				
BM_{pheno}	-0.12	0.0083	0.95	0.087
	$<10^{-6}$	*0.082*		
BM_y	-0.61	-0.051	0.60	0.090
	$<10^{-4}$	*0.0039*		
fixed	-0.11	0.011		0.087
effects	*$<10^{-5}$*	*0.025*		

Table 4: Analysis of leaflet length (log-transformed) with REML on the subset of specimens with both morphological and geographical data. BM_{pheno} assumes equal phenotypic and phylogenetic relationships. BM_y accounts for phenotypic variation in the response but not in the predictors. The model with fixed effects uses a standard linear regression on the individual-level data, with morph means (or intercept) estimated as fixed-effect parameters, and does not estimate σ^2_b. In the elevation and latitude columns, the first value is the estimated coefficient. The second italicized value is the associated p-value, in red when < 0.05.

The coefficient for elevation estimated using BM_{pheno} was still significantly negative, but much smaller in magnitude (Table 4). More importantly, BM_{pheno} estimates latitude to correlate positively with leaflet length, with strong evidence on the network and weak evidence on the tree. This is in contradiction with the negative correlation found using BM_y.

This discordance suggests conflicting phenotypic and phylogenetic relationships. To estimate the phenotypic relationships alone, we fit a linear regression with morph means modeled as fixed effects (i.e. separate intercepts for each morph) and estimated as parameters, instead of estimating an evolutionary variance rate. We found a positive phenotypic coefficient for latitude, of magnitude greater than that estimated by BM_{pheno}, on the network and on the tree (Table 4). This behavior matches our simulations under a phenotypic coefficient opposite to the evolutionary coefficient, in which the BM_{pheno} estimate was heavily biased towards the phenotypic coefficient, increasingly so with more specimens. The BM_y estimate was only slightly biased, and less so with more specimens. Therefore, we consider the BM_{pheno} estimates to be misleading for studying the evolution of leaflet size in *Polemonium*, because the phenotypic and evolutionary coefficients for latitude appear to be opposite, strongly violating the BM_{pheno} assumption.

3.2.4 Leaflet size reconstruction

The predicted true mean of log leaflet length was obtained for *eddyense* (3 specimens) and *foliosissimum* (274 specimens). For *eddyense*, the observed mean log leaflet length was -1.22. The true morph mean was predicted at -1.16 and -1.18 with and without elevation and latitude as predictors, representing 5.9% and 4.3% increases from the observed mean on the original scale. The prediction intervals (both of width 0.78) encompassed the observed mean. For *foliosissimum*, the observed mean was 0.622. The predicted means were both very close (0.621, a 0.1% decrease on the original scale) and the prediction intervals were narrow (width < 0.085). This illustrates that predictions for species with smaller sample sizes are more influenced by other species' data, and less certain. Regardless of sample size, model predictors had little influence on the predictions.

At internal nodes, the width of prediction intervals increased with age: from 1.07 for the hybrid node directly ancestral to *elusum*, 1.72 for its minor parent, to 2.05 for the root. This pattern of ancestral state uncertainty increasing with distance from the tips was already known on trees [Ané, 2008].

4 Discussion

We presented a method to account for within-species trait variation on phylogenetic networks, a task with a long history on trees, and whose importance has been stressed by many authors. We reiterate the importance of this for avoiding overestimating the evolutionary variance. Intuitively, ignoring within-species variation is compensated for by an inflated evolutionary variance rate.

4.1 Methodology

Our approach is the first to allow for both within-species trait variation and reticulation, and to estimate within-species variance simultaneously with other model parameters instead of considering within-species variances as known without error. Our method assumes equal variances within species, and is robust to a violation of this assumption based on simulations. This assumption is also made by the PMM when used to account for within-species variation. It is suggested by Ives et al. [2007] as an option when the sample size per species is small, via the estimation of a pooled variance. When traits are hard to measure experimentally, typical sample sizes per species are very low. Moen et al. [2022] highlight this challenge for studies of adaptation, and the advantage of assuming equal variances in this context. Future method development could consider relaxing the assumption of homogeneous within-species variance for species with many sampled individuals.

Our implementation currently assumes that each hybrid node has exactly two parent lineages in the network, but the method allows for polytomies where a node has three or more children. Our implementation is currently limited to the BM, although our theory applies to more complex evolutionary models for the evolutionary covariance $\boldsymbol{V}$, at the cost of optimizing extra parameters. For example, Pagel's λ model would include an independent component at the species level, beyond the variation between individuals or measurement error. It would also be interesting to allow for separate evolutionary rates along different parts of the phylogeny. In this case, $\boldsymbol{V}$ depends on the different rates and their mapping along the phylogeny [O'Meara et al., 2006]. As more complex models are developed, conclusions about evolutionary rates and phylogenetic signal could rely on likelihood ratio tests, although these tests are approximate. Tests based on bootstrapping procedures could be a possible future development to perform more accurate model comparisons.

Our simulations highlight the advantages of using REML instead of ML, especially with models that have multiple variance parameters, or to answer questions about character rate evolution. For *Polemonium* leaflet length for example, switching from ML to REML can sway the estimate of phylogenetic signal from 0 to 1. The advantages of REML are well known [Ives et al., 2007] yet many software tools use ML only (e.g. `geiger`, Pennell et al. [2014] or `phylolm`, Ho and Ané [2014]). Unfortunately, REML is typically not an option for non-Gaussian generalized linear models, such as for phylogenetic logistic regression.

4.2 Importance of gene flow for traits

As of now, methods to estimate species networks scale poorly with the number of taxa. To detect gene flow and represent reticulations in a network, most studies focus their questions on a subsample of 20 taxa or so, a scale that methods such as SNaQ and PhylonetMPL can handle [Hejase and Liu, 2016]. Downstream comparative analyses then face a dilemma: should they use more taxa on an approximate phylogeny without reticulation, or fewer taxa on a more accurate representation of the group's phylogeny? Our case study on *Polemonium* suggests that using more taxa is advantageous and more powerful, as the increase in data quantity (and signal) can be substantial, out-

weighing the approximation to the phylogenetic covariance using a tree phylogeny. A caveat is that model mis-specification caused by ignoring reticulation on a tree may decrease power or increase type-1 error. We hope that this dilemma will disappear as network inference methods improve.

At each reticulation, one may ask which parent contributed to a given trait. Was the trait value inherited solely from one of the parents? Our BM model assumes that both parents contributed, such that the trait value at the reticulate node is a weighted average from the two parent values. The weights are the inheritance proportions (γ): the proportion of genes inherited from each parent. This is a legitimate prior for quantitative traits that are controlled by many genes of small and additive effects. But at each reticulation, one may ask if this null model is adequate for the trait under study.

To this end, Bastide et al. [2018] proposed a test for transgressive evolution after reticulation. This test can readily be used with our method to account for within-species trait variation.

Another approach is to compare the network model with a tree model, in which we assume that a trait is inherited from a single parent only, although model choice would need to account for the large number of options (up to 2^h) with an increasing number h of reticulations. More generally, one may seek to optimize the weights (γ) of the two parents at each reticulation, to best match evidence from the trait data. Bastide [2017] took this approach. Optimizing all h inheritance parameters could be too many, so he used a single parameter to scale the weights of all major hybrid edges simultaneously. Even with this amortized inference strategy, simulations showed that a single continuous trait variable had low information about the inheritance weights at reticulations.

Our findings in *Polemonium* are consistent. For all measures of leaflet size, the network model with inheritance values from genetic data was preferred over a tree model, in which the trait was forced to be inherited from a single parent (corresponding to inheritance values set to 0 or 1). However, the preference for the network was only very slight: the morphological signal is consistent with the genetic signal, but tenuous.

Multiple continuous traits would need to be combined to estimate the morphological signal for gene flow. As for trees, combining morphological traits is complicated by trait correlations. This caveat is especially important if we want to assume that traits share a common signal of gene flow. The traits more likely to have been inherited together through gene flow are the traits that share a genetic basis or form an integrated morphological component, and can be highly correlated with each other.

The inheritance signal may be stronger from discrete traits than from continuous traits, if the discrete trait is evolving slowly enough for accurate ancestral reconstruction. For example, Karimi et al. [2020] found support that flower color was introgressed during the evolution of baobabs in Madagascar. It would be interesting to extend our method for within-species variation to the study of discrete characters.

4.3 Phenotypic correlation

Our simulations highlight an important bias affecting many widely-used methods, when there is within-species variation in the predictors. The regression coefficient describing the historical evolutionary relationships are pulled towards the phenotypic coefficients. This bias is traditionally named "attenuation" when variation in the predictor is solely due to measurement error, uncorrelated with the other sources of variation [Fuller, 1987]. This pull decreases as the within-species sampling effort increases, for methods ignoring within-species variation in predictors.

For these methods, within-species variation in predictors causes a complex bias in estimating evolutionary variance rates. If a phenotypic relationship is absent or opposite to the evolutionary relationship, then σ^2_{b} is overestimated. If the phenotypic relationship is equal or stronger than the evolutionary relationship, then σ^2_{b} is underestimated. This interplay between phenotypic relationships (most often ignored for the study of long-term evolutionary patterns) and inference of evolutionary rates has not been identified before, to the best of our knowledge.

When predictors are available for the same set of individuals as the response trait, the $\mathrm{BM}_{\mathrm{pheno}}$ model can be applied to account for within-species variation in predictors. However, $\mathrm{BM}_{\mathrm{pheno}}$ assumes shared evolutionary and phenotypic relationships, such that the pull towards the phenotypic coefficients strengthens with more sampling effort, and the bias becomes extreme. We observed this for *Polemonium* leaflet size, where discordant evolutionary and phenotypic relationships led to opposite conclusions about the direction of correlation between leaflet length and latitude. For this reason, we recommend using this method with caution, and in combination with an assessment of the method's assumption regarding phenotypic relationships. To estimate phenotypic correlations, standard linear models can be used with species as a fixed factor. Future work could tackle the question of rigorously testing whether phenotypic and evolutionary relationships are equal, extending the

methods by Revell and Harmon [2008] and Goolsby et al. [2017] to reticulate phylogenetic networks and to a linear regression context (rather than correlation).

New methods are needed to handle the case when predictors are available on a different set of individuals than the response trait, if one wishes to use all individual values to best account for within-species trait variation, and to eliminate the pull of evolutionary coefficients towards phenotypic coefficients.

Funding

This work was supported in part by the National Science Foundation (DMS-1902892 and DMS-2023239) and by an NSF doctoral dissertation improvement grant (DEB 1501867) to JPR.

Acknowledgements

We thank Cathy Cao for technical help setting up simulations. We also thank Joe Felsenstein and two anonymous reviewers for their thorough feedback, which helped improve the structure and clarity.

Software and Data Availability

Our method is implemented in the `Julia` package `PhyloNetworks` available at `https://github.com/crsl4/PhyloNetworks.jl` starting with v0.14.0. Data and code for all simulations and analyses are available from the Dryad Digital Repository: `https://doi.org/10.5061/dryad.9ghx3ffkc`.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

C. Ané. Analysis of comparative data with hierarchical autocorrelation. *The Annals of Applied Statistics*, 2(3):1078 – 1102, 2008. doi: 10.1214/08-AOAS173.

P. Bastide. *Shifted stochastic processes evolving on trees: application to models of adaptive evolution on phylogenies*. Theses, Université Paris Saclay (COmUE), Oct. 2017. URL `https://tel.archives-ouvertes.fr/tel-01629648`.

P. Bastide, C. Solís-Lemus, R. Kriebel, K. William Sparks, and C. Ané. Phylogenetic comparative methods on phylogenetic networks with reticulations. *Systematic Biology*, 67(5): 800–820, 2018. doi: 10.1093/sysbio/syy033.

R. Christensen. *Linear models for spatial data: Kriging*, pages 269–311. Springer New York, New York, NY, 2001. ISBN 978-1-4757-3847-6. doi: 10.1007/978-1-4757-3847-6_6.

N. Cooper, G. H. Thomas, C. Venditti, A. Meade, and R. P. Freckleton. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biological Journal of the Linnean Society*, 118(1):64–77, 2016. doi: 10.1111/bij.12701.

E. Demidenko. *Mixed Models: Theory and Applications*. John Wiley & Sons, Inc., 2004. doi: 10.1002/0471728438.

J. Felsenstein. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, 19(1):445–471, 1988. doi: 10.1146/annurev.es.19.110188.002305.

J. Felsenstein. Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *The American Naturalist*, 171(6): 713–725, 2008. doi: 10.1086/587525.

S. E. Fick and R. J. Hijmans. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315, 2017. doi: 10.1002/joc.5086.

W. A. Fuller. *A single explanatory variable*, pages 1–99. John Wiley & Sons, 1987. ISBN 978-0-4718-6187-4. doi: 10.1002/978-0-4703-1666-5.

L. Z. Garamszegi. Uncertainties due to within-species variation in comparative studies: measurement errors and statistical weights. In L. Z. Garamszegi, editor, *Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice*, pages 157–199. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-662-43550-2. doi: 10.1007/978-3-662-43550-2_7.

E. W. Goolsby, J. Bruggeman, and C. Ané. Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, 8(1):22–27, 2017. doi: 10.1111/2041-210X.12612.

T. F. Hansen and E. P. Martins. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific

data. *Evolution*, 50(4):1404–1417, 1996. doi: 10.1111/j.1558-5646.1996.tb03914.x.

T. F. Hansen, J. Pienaar, and S. H. Orzack. A comparative method for studying adaptation to a randomly evolving environment. *Evolution*, 62(8): 1965–1977, 2008. doi: 10.1111/j.1558-5646.2008.00412.x.

L. Harmon. *Phylogenetic comparative methods*. Luke J. Harmon, 2019. URL `https://lukejharmon.github.io/pcm/`.

L. J. Harmon and J. B. Losos. The effect of intraspecific sample size on type i and type ii error rates in comparative studies. *Evolution*, 59(12):2705–2710, 2005. doi: 10.1111/j.0014-3820.2005.tb00981.x.

D. A. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61 (2):383–385, 1974. doi: 10.1093/biomet/61.2.383.

H. A. Hejase and K. J. Liu. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics*, 17(1):422, 2016. doi: 10.1186/s12859-016-1277-1.

C. R. Henderson, O. Kempthorne, S. R. Searle, and C. M. von Krosigk. The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15(2):192, June 1959. doi: 10.2307/2527669.

R. J. Hijmans. *raster: geographic data analysis and modeling*, 2020. URL `https://cran.r-project.org/package=raster`. R package version 3.4-5.

L. S. T. Ho and C. Ané. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Systematic Biology*, 63(3):397–408, 2014. doi: 10.1093/sysbio/syu005.

E. A. Housworth, E. P. Martins, and M. Lynch. The phylogenetic mixed model. *The American Naturalist*, 163(1):84–96, 2004. doi: 10.1086/380570.

J. Huang, Y. Thawornwattana, T. Flouri, J. Mallet, and Z. Yang. Inference of gene flow between species under misspecified models. *Molecular Biology and Evolution*, 39(12), 2022. doi: 10.1093/molbev/msac237.

A. R. Ives, P. E. Midford, and J. Garland, Theodore. Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology*, 56(2):252–270, 2007. doi: 10.1080/10635150701313830.

N. Karimi, C. E. Grover, J. P. Gallagher, J. F. Wendel, C. Ané, and D. A. Baum. Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (*Adansonia*; bombacoideae; malvaceae). *Systematic Biology*, 69(3): 462–478, 2020. doi: 10.1093/sysbio/syz073.

C. Körner, M. Neumayer, S. P. Menendez-Riedl, and A. Smeets-Scheel. Functional morphology of mountain plants. *Flora*, 182(5):353–383, 1989. ISSN 0367-2530. doi: 10.1016/S0367-2530(17)30426-7.

A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82: 1–26, 2017. doi: 10.18637/jss.v082.i13.

L. R. LaMotte. A direct derivation of the reml likelihood function. *Statistical Papers*, 48(2):321–327, 2007. doi: 10.1007/s00362-006-0335-6.

G. E. Leventhal and S. Bonhoeffer. Potential Pitfalls in Estimating Viral Load Heritability. *Trends in Microbiology*, 24(9):687–698, 2016. doi: 10.1016/j.tim.2016.04.008.

M. Lynch. Methods for the Analysis of Comparative Data in Evolutionary Biology. *Evolution*, 45(5): 1065–1080, 1991. doi: 10.1111/j.1558-5646.1991.tb04375.x.

E. P. Martins and T. F. Hansen. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149 (4):646–667, 1997.

D. S. Moen, E. Cabrera-Guzmán, I. W. Caviedes-Solis, E. González-Bernal, and A. R. Hanna. Phylogenetic analysis of adaptation in comparative physiology and biomechanics: overview and a case study of thermal physiology in treefrogs. *Journal of Experimental Biology*, 225(Suppl.1), 2022. doi: 10.1242/jeb.243292.

B. C. O'Meara, C. Ané, M. J. Sanderson, and P. C. Wainwright. Testing for different rates of continuous trait evolution using likelihood. *Evolution*, 60 (5):922 – 933, 2006. doi: 10.1554/05-130.1.

M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, 1999. doi: 10.1038/44766.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971. doi: 10.1093/biomet/58.3.545.

M. Pennell, J. Eastman, G. Slater, J. Brown, J. Uyeda, R. Fitzjohn, M. Alfaro, and L. Harmon. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, 30:2216–2218, 2014.

J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer Science & Business Media, 2006. ISBN 9780387227474.

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

L. J. Revell and L. J. Harmon. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. *Evolutionary Ecology Research*, 10:311–331, 2008.

J. P. Rose. Taxonomy and relationships within *Polemonium foliosissimum* (Polemoniaceae): untangling a clade of colorful and gynodioecious herbs. *Systematic Botany*, 46(3):519–537, 2021. doi: 10.1600/036364421X16312067913372.

J. P. Rose, C. A. P. Toledo, E. M. Lemmon, A. R. Lemmon, and K. J. Sytsma. Out of sight, out of mind: widespread nuclear and plastid-nuclear discordance in the flowering plant genus *Polemonium* (Polemoniaceae) suggests widespread historical gene flow despite limited nuclear signal. *Systematic Biology*, 70(1):162–180, 2021. doi: 10.1093/sysbio/syaa049.

J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 2012. doi: 10.1038/nmeth.2019.

D. Silvestro, A. Kostikova, G. Litsios, P. B. Pearman, and N. Salamin. Measurement errors should always be incorporated in phylogenetic comparative analysis. *Methods in Ecology and Evolution*, 6(3):340–346, 2015. doi: 10.1111/2041-210X.12337.

C. Solís-Lemus, P. Bastide, and C. Ané. Phylonetworks: a package for phylogenetic networks. *Molecular Biology and Evolution*, 34(12):3292–3298, 2017. doi: 10.1093/molbev/msx235.

I. J. Wright, N. Dong, V. Maire, I. C. Prentice, M. Westoby, S. Díaz, R. V. Gallagher, B. F. Jacobs, R. Kooyman, E. A. Law, M. R. Leishman, Ü. Niinemets, P. B. Reich, L. Sack, R. Villar, H. Wang, and P. Wilf. Global climatic drivers of leaf size. *Science*, 357(6354):917–921, 2017. doi: 10.1126/science.aal4760.

Appendix

A Parameter estimation

We prove here that (4) can be simplified to (5). Intuitively, (5) comes from reducing the model to species averages. The formula for $\hat{\boldsymbol{\beta}}$ in (4) involves the inverse of $\boldsymbol{W}_\eta$, so we first show that this large $N \times N$ matrix can in fact be inverted using smaller $n \times n$ matrices. Similar developments have reduced the computation complexity of some classes of mixed models [Demidenko, 2004, section 2.2.3], sometimes referred to as Henderson's formula (1959). Our framework differs due to the phylogenetic correlation between species ("clusters" in the classical context). To express $\boldsymbol{W}_\eta^{-1}$, we apply the Woodbury matrix identity[5]:

$$\begin{aligned}\boldsymbol{W}_\eta^{-1} &= (\eta \boldsymbol{I}_N + \boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}')^{-1} \\ &= \eta^{-1}\boldsymbol{I}_N - \eta^{-2}\boldsymbol{Z}(\boldsymbol{V}^{-1} + \boldsymbol{Z}'(\eta\boldsymbol{I}_N)^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}' \\ &= \eta^{-1}\boldsymbol{I}_N - \eta^{-2}\boldsymbol{Z}(\boldsymbol{V}^{-1} + \eta^{-1}\boldsymbol{D})^{-1}\boldsymbol{Z}' .\end{aligned}$$

To invert $\boldsymbol{V}^{-1} + \eta^{-1}\boldsymbol{D}$, we apply the Woodbury matrix identity to $\eta\boldsymbol{I}_n + \boldsymbol{V}\boldsymbol{D}$:

$$(\eta\boldsymbol{I}_n + \boldsymbol{V}\boldsymbol{D})^{-1} = \eta^{-1}\boldsymbol{I}_n - \eta^{-2}(\boldsymbol{V}^{-1} + \eta^{-1}\boldsymbol{D})^{-1}\boldsymbol{D} .$$

Using $\boldsymbol{V}_\eta$ from (6), $\boldsymbol{V}_\eta\boldsymbol{D} = \eta\boldsymbol{I}_n + \boldsymbol{V}\boldsymbol{D}$ and we get:

$$(\boldsymbol{V}^{-1} + \eta^{-1}\boldsymbol{D})^{-1} = \eta\boldsymbol{D}^{-1} - \eta^2\boldsymbol{D}^{-1}\boldsymbol{V}_\eta^{-1}\boldsymbol{D}^{-1} .$$

Combining the above equations, we get

$$\boldsymbol{W}_\eta^{-1} = \frac{1}{\eta}(\boldsymbol{I}_N - \boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{Z}') + \boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{V}_\eta^{-1}\boldsymbol{D}^{-1}\boldsymbol{Z}' . \quad \text{(S1)}$$

We are now ready to simplify (4), recalled here:

$$\hat{\boldsymbol{\beta}}(\eta) = \left(\boldsymbol{X}'\boldsymbol{W}_\eta^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_\eta^{-1}\boldsymbol{Y}$$

where $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{x}$ and $\overline{\boldsymbol{y}} = \boldsymbol{D}^{-1}\boldsymbol{Z}'\boldsymbol{Y}$. Using (S1) and $\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{D}$, we have:

$$\begin{aligned}\boldsymbol{X}'\boldsymbol{W}_\eta^{-1}\boldsymbol{X} &= \boldsymbol{x}'\boldsymbol{V}_\eta^{-1}\boldsymbol{x} \\ \boldsymbol{X}'\boldsymbol{W}_\eta^{-1}\boldsymbol{Y} &= \boldsymbol{x}'\boldsymbol{V}_\eta^{-1}\boldsymbol{D}^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \boldsymbol{x}'\boldsymbol{V}_\eta^{-1}\overline{\boldsymbol{y}} .\end{aligned}$$

Combining the above equations gives (5).

We now turn to simplifying the profile likelihood criterion to be maximized, for the estimation of variance parameters ($\hat{\sigma}_{\mathrm{b}}^2$, $\hat{\eta}$ and any parameters for $\boldsymbol{V}$), to

[5] $(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{I}-\boldsymbol{U}(\boldsymbol{C}^{-1}+\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1})$.

prove (7), (8) and (9). As usual, we instead write and seek to minimize twice the negative log (restricted) likelihood, denoted as $\ell_{\mathrm{ml}}(\sigma_{\mathrm{b}}^2, \eta)$ for ML and $\ell_{\mathrm{reml}}(\sigma_{\mathrm{b}}^2, \eta)$ for REML:

$$\begin{aligned}\ell_{\mathrm{ml}} &= N\log 2\pi + \log|\sigma_{\mathrm{b}}^2\boldsymbol{W}_\eta| + \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\eta)\|^2_{\boldsymbol{W}_\eta}/\sigma_{\mathrm{b}}^2 \\ \ell_{\mathrm{reml}} &= \ell_{\mathrm{ml}} - p\log 2\pi + \log|\boldsymbol{X}'(\sigma_{\mathrm{b}}^2\boldsymbol{W}_\eta)^{-1}\boldsymbol{X}|\end{aligned}$$

where we recall that $\|\boldsymbol{u}\|^2_{\boldsymbol{M}} = \boldsymbol{u}'\boldsymbol{M}^{-1}\boldsymbol{u}$. We now show how each term involving $\boldsymbol{W}_\eta$ can be simplified using smaller matrices.

First, we use Sylvester's determinant identity[6] to express $|\boldsymbol{W}_\eta|$ in terms $|\boldsymbol{V}_\eta|$.

$$\begin{aligned}|\boldsymbol{W}_\eta| &= \eta^N|\boldsymbol{I}_N + \eta^{-1}\boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}'| = \eta^N|\boldsymbol{I}_n + \eta^{-1}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{V}| \\ &= \eta^{N-n}|\eta\boldsymbol{I}_n + \boldsymbol{D}\boldsymbol{V}| = \eta^{N-n}|\boldsymbol{D}\boldsymbol{V}_\eta| \\ &= \eta^{N-n}|\boldsymbol{V}_\eta|\prod_{i=1}^n m_i\,. \qquad \text{(S2)}\end{aligned}$$

Next we use (S1) and (5) to simplify $\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2_{\boldsymbol{W}_\eta}$.

$$\begin{aligned}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\eta)\|^2_{\boldsymbol{W}_\eta} &= \boldsymbol{Y}'\boldsymbol{W}_\eta^{-1}\boldsymbol{Y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{W}_\eta^{-1}\boldsymbol{Y} \\ &\quad + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{W}_\eta^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \eta^{-1}\boldsymbol{Y}'(\boldsymbol{I}_N - \boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{Z}')\boldsymbol{Y} + \overline{\boldsymbol{y}}'\boldsymbol{V}_\eta^{-1}\overline{\boldsymbol{y}} \\ &\quad - 2\hat{\boldsymbol{\beta}}'\boldsymbol{x}'\boldsymbol{V}_\eta^{-1}\overline{\boldsymbol{y}} + \hat{\boldsymbol{\beta}}'\boldsymbol{x}'\boldsymbol{V}_\eta^{-1}\boldsymbol{x}\hat{\boldsymbol{\beta}} \\ &= \eta^{-1}(\boldsymbol{Y}'\boldsymbol{Y} - (\boldsymbol{Z}'\boldsymbol{Y})'\boldsymbol{D}^{-1}\boldsymbol{Z}'\boldsymbol{Y}) + \|\overline{\boldsymbol{y}} - \boldsymbol{x}\hat{\boldsymbol{\beta}}\|^2_{\boldsymbol{V}_\eta}\,.\end{aligned}$$

Recalling that $\mathrm{SSW} = \sum_{i=1}^n\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_i)^2$ captures the sum of squared residuals within species, we get:

$$\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\eta)\|^2_{\boldsymbol{W}_\eta} = \eta^{-1}\mathrm{SSW} + \|\overline{\boldsymbol{y}} - \boldsymbol{x}\hat{\boldsymbol{\beta}}(\eta)\|^2_{\boldsymbol{V}_\eta}\,. \qquad \text{(S3)}$$

Next, we optimize σ_{b}^2 analytically as a function of η to profile ℓ_{ml} and ℓ_{reml} as functions of η only. If we fix η (and any other potential parameters for $\boldsymbol{V}$) and substitute (S2) and (S3) into ℓ_{ml} and ℓ_{reml}, then we obtain the optimal value for σ_{b}^2 given in (9), which depends on the criterion via the degree of freedom $d = N$ for ML and $d = N - p$ for REML. Plugging $\hat{\sigma}_{\mathrm{b}}^2$ from (9) into ℓ_{ml} and ℓ_{reml} above, we obtain the profiled ML and REML criteria given in (7) and (8).

B Parameter inference

To test hypotheses about a coefficient β_k, we use its estimated standard error SE_k with

$$\mathrm{SE}_k^2 = \hat{\sigma}_{\mathrm{b}}^2(\hat{\eta})\left[\left(\boldsymbol{x}'\boldsymbol{V}_{\hat{\eta}}^{-1}\boldsymbol{x}\right)^{-1}\right]_{kk}.$$

[6] If $\boldsymbol{A}$ is $m\times n$ and $\boldsymbol{B}$ is $n\times m$, then $|\boldsymbol{I}_m + \boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{I}_n + \boldsymbol{B}\boldsymbol{A}|$.

If the true η were known and used in the definition of SE_k, then $(\hat{\beta}_k - \beta_k)/\mathrm{SE}_k$ would follow a T-distribution with $N - p$ degrees of freedom. But η is unknown. We approximate the distribution of $(\hat{\beta}_k - \beta_k)/\mathrm{SE}_k$ by a T-distribution with $n - p$ degrees of freedom, being conservative by taking into account the number of species instead of the total number of observations. This approximation is exact in some classical contexts with balanced experiments, such as for the estimation of a population mean from n samples, each with m subsamples. More generally, a similar approximation is used for mixed models and has been shown to be superior to likelihood ratio tests for fixed effects [see e.g. Section 2.2.4 in Pinheiro and Bates, 2006]. Confidence intervals for regression coefficients also use this approximation, assuming $n-p$ degrees of freedom associated with $\hat{\sigma}_{\mathrm{b}}^2$.

C Predicting species means

This task is traditionally called "ancestral state reconstruction", but we favor the term "prediction", as this task can be applied to present-day species. We use here the notations from section 2.3. In particular, $\boldsymbol{y}_0$ denotes the true mean of species for which prediction is sought.

C.1 Prediction variance

Section 2.3 gives the conditional mean $\boldsymbol{\mu}_0(\boldsymbol{\beta})$ and variance $\boldsymbol{\Sigma}$ of $\boldsymbol{y}_0$ given $\overline{\boldsymbol{y}}$, in the case when $\boldsymbol{\beta}$ is known. When $\boldsymbol{\beta}$ is estimated from $\overline{\boldsymbol{y}}$, then the best prediction is $\boldsymbol{\mu}_0(\hat{\boldsymbol{\beta}})$ but has variance larger than $\boldsymbol{\Sigma}$. Namely, the prediction variance $\mathrm{var}(\boldsymbol{y}_0 - \boldsymbol{\mu}_0(\hat{\boldsymbol{\beta}}))$ is then given by:

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma} + \sigma_{\mathrm{b}}^2\boldsymbol{M}(\boldsymbol{x}'\boldsymbol{V}_\eta^{-1}\boldsymbol{x})^{-1}\boldsymbol{M}' \qquad \text{(S4)}$$

where $\boldsymbol{M} = \boldsymbol{x}_0 - \boldsymbol{V}_c\boldsymbol{V}_\eta^{-1}\boldsymbol{x}$ [Christensen, 2001].

One might ask if conditioning on the individual-level data $\boldsymbol{Y}$ provides more information about $\boldsymbol{y}_0$ than can be gained from the taxon-level means $\overline{\boldsymbol{y}}$. We show that both reduce to the same estimator, so that $\overline{\boldsymbol{y}}$ is sufficient for predictive purposes:

$$\begin{bmatrix}\boldsymbol{y}_0 \\ \boldsymbol{Y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x} \\ \boldsymbol{Z}\boldsymbol{x}\end{bmatrix}\boldsymbol{\beta}, \sigma_{\mathrm{b}}^2\begin{bmatrix}\boldsymbol{V}_0 & \boldsymbol{V}_c\boldsymbol{Z}' \\ \boldsymbol{Z}\boldsymbol{V}_c' & \boldsymbol{W}_\eta\end{bmatrix}\right)$$

By applying (S1) and $\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{D}$ we have:

$$\begin{aligned}\mathbb{E}(\boldsymbol{y}_0 \mid \boldsymbol{Y}) &= \boldsymbol{x}_0\boldsymbol{\beta} + (\boldsymbol{V}_c\boldsymbol{Z}')\boldsymbol{W}_\eta^{-1}(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{x}\boldsymbol{\beta}) \\ &= \boldsymbol{x}_0\boldsymbol{\beta} + \boldsymbol{V}_c\boldsymbol{V}_\eta^{-1}\boldsymbol{D}^{-1}\boldsymbol{Z}'(\overline{\boldsymbol{y}} - \boldsymbol{Z}\boldsymbol{x}\boldsymbol{\beta}) \\ &= \mathbb{E}(\boldsymbol{y}_0 \mid \overline{\boldsymbol{y}}) = \boldsymbol{\mu}_0(\boldsymbol{\beta})\,.\end{aligned}$$

C.2 Prediction interval

The prediction error $\boldsymbol{e}^* = \boldsymbol{y}_0 - \boldsymbol{\mu}_0(\hat{\boldsymbol{\beta}})$ has distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^*)$ with $\boldsymbol{\Sigma}^*$ given in (S4), so the error $\boldsymbol{e}_i^*$ for the ith species to be predicted satisfies

$$u_i = \frac{\boldsymbol{e}_i^*}{\sigma_{\mathrm{b}}\sqrt{\Sigma_{ii}^*/\sigma_{\mathrm{b}}^2}} \sim \mathcal{N}(0,1)$$

Note that the formula for $\boldsymbol{\Sigma}^*/\sigma_{\mathrm{b}}^2$ involves η but not σ_{b}^2. If η is known, then $\hat{\boldsymbol{\beta}}(\eta)$ and $\hat{\sigma}_{\mathrm{b}}(\eta)$ are independent by Cochran's theorem. However, $\boldsymbol{e}_i^*$ is not guaranteed to be independent of $\hat{\sigma}_{\mathrm{b}}$. Nevertheless, we may use $\psi_i = \Sigma_{ii}^*(\hat{\eta}, \hat{\sigma}_{\mathrm{b}}(\hat{\eta}))$ to estimate the variance of $\boldsymbol{e}_i^*$ We then approximate the distribution of $\boldsymbol{e}_i^*/\sqrt{\psi_i}$ by a T-distribution with $n-p$ degrees of freedom, as done above for testing fixed coefficients about between-species relationships.

Consequently, to build a $100(1-\alpha)\%$ prediction interval for the ith species mean, we first find the $(1-\alpha/2)$ quantile q of the T-distribution with $n-p$ degrees of freedom and then use

$$\boldsymbol{\mu}_0(\hat{\boldsymbol{\beta}})_i \pm q\sqrt{\psi_i}\,.$$

Recall here that formula (S4) for $\boldsymbol{\Sigma}^*$ (hence $\sqrt{\psi_i}$) was obtained assuming that we know η and any other parameters for $\boldsymbol{V}$, and then simply plugging in their estimates in (S4). Doing so does not account for the extra uncertainty due to estimating η, which is hard to quantify [Christensen, 2001]. Hence, the prediction interval above should be considered as liberal.

C.3 Example: Influence of other species information on reconstructed mean

We consider here the task of predicting the true mean for a species for which we do have data. In a simple example, we show that the prediction $\hat{\boldsymbol{y}}_0 = \boldsymbol{\mu}_0(\hat{\boldsymbol{\beta}})$ can be different from the observed species mean $\overline{\boldsymbol{y}}$, especially if a species has few sampled individuals. This example provides an intuition for what affects the prediction.

Suppose we have three taxa with sample sizes m_1, m_2, m_3, and that the unscaled covariance matrix constructed from their phylogeny is:

$$\boldsymbol{V} = \begin{bmatrix} v_{11} & v_{12} & 0 \\ v_{21} & v_{22} & 0 \\ 0 & 0 & v_{33} \end{bmatrix}$$

with taxon 1 and 2 sister to each other. Applying equations from section 2.3 and focusing on predicting the means for species 1 and 3, we get that

$$\hat{y}_1 = \frac{1}{K}\big(k_1(\boldsymbol{x}\hat{\boldsymbol{\beta}})_1 + k_2\overline{y}_1\big) + \frac{v_{12}\eta}{Km_1}(\overline{\boldsymbol{y}} - \boldsymbol{x}\hat{\boldsymbol{\beta}})_2$$
$$\hat{y}_3 = \frac{\eta}{\eta + v_{33}m_3}(\boldsymbol{x}\hat{\boldsymbol{\beta}})_3 + \frac{v_{33}m_3}{\eta + v_{33}m_3}\overline{y}_3$$

where $k_1 = \dfrac{v_{22}\eta}{m_1} + \dfrac{\eta^2}{m_1 m_2}$, $k_2 = \dfrac{v_{11}\eta}{m_2} + v_{11}v_{22} - v_{12}v_{21}$ and $K = k_1 + k_2$. Therefore, $\hat{y}_i$ does not necessarily equal the sample mean $\overline{y}_i$. Instead, $\hat{y}_i$ is pulled towards $(\boldsymbol{x}\hat{\boldsymbol{\beta}})_i$, which depends on data across all species, and represents the ancestral state at the root if $\boldsymbol{x}$ is reduced to the intercept only. The pull is strong if m_i is small or if within-species variation (η) is large. If m_i is large, then the pull disappears and $\hat{y}_i \approx \overline{y}_i$.

Beyond the weighted average of $(\boldsymbol{x}\hat{\boldsymbol{\beta}})_1$ and $\overline{y}_1$, $\hat{y}_1$ has an additional term proportional to the residual of its sister species 2. This term shows how information from closely related species is borrowed to influence prediction, in this simple example. As expected, this term vanishes when m_1 increases.

D Phenotypic correlation model

We study here the simulation model described in section 2.5, in which the within-species (or phenotypic) relationship between the response and the predictor traits differs from the between-species (or phylogenetic) relationship. We derive the distribution of the full response data $\boldsymbol{Y}$ and of the species means $\overline{\boldsymbol{y}}$ conditional on the observed predictor's species means $\overline{\boldsymbol{x}}$. Since all variables are Gaussian, we simply need to derive the conditional means and variances. To do so, we repeatedly use the standard conditional distribution formulae for Gaussian processes.

Using (12) to simulate the predictor, (13) to simulate the response, the expression $\overline{\boldsymbol{x}} = \frac{1}{m}\boldsymbol{Z}'\boldsymbol{X}$ for species means, and the fact that $\boldsymbol{Z}'\boldsymbol{Z} = m\boldsymbol{I}_n$, we get

$$\begin{aligned} \mathrm{var}(\overline{\boldsymbol{x}}) &= \sigma_{b,x}^2 \boldsymbol{V}\left(\boldsymbol{I}_n + \frac{\eta_x}{m}\boldsymbol{V}^{-1}\right) \\ \mathrm{cov}(\boldsymbol{Y}, \overline{\boldsymbol{x}}) &= \beta_1\sigma_{b,x}^2 \boldsymbol{Z}\boldsymbol{V}\left(\boldsymbol{I}_n + \frac{\beta_2}{\beta_1}\frac{\eta_x}{m}\boldsymbol{V}^{-1}\right) \end{aligned}$$

where $\eta_x = \sigma_{w,x}^2/\sigma_{b,x}^2$. Therefore

$$\mathbb{E}(\boldsymbol{Y} \mid \overline{\boldsymbol{x}}) = \beta_1 \boldsymbol{Z}\tilde{\boldsymbol{x}}$$

where we further define $u = \eta_x/m$ and

$$\begin{aligned} \tilde{\boldsymbol{x}} &= \left(\boldsymbol{I}_n + \frac{\beta_2}{\beta_1}u\boldsymbol{V}^{-1}\right)\left(\boldsymbol{I}_n + u\boldsymbol{V}^{-1}\right)^{-1}\overline{\boldsymbol{x}} \qquad \text{(S5)} \\ &= \left(\boldsymbol{I}_n + \frac{\beta_2 - \beta_1}{\beta_1}u\big(\boldsymbol{V} + u\boldsymbol{I}_n\big)^{-1}\right)\overline{\boldsymbol{x}} \\ &\approx \left(\boldsymbol{I}_n + \frac{\beta_2 - \beta_1}{\beta_1}u\boldsymbol{V}^{-1}\right)\overline{\boldsymbol{x}} \quad \text{if } u = \frac{\eta_x}{m} \to 0\,. \end{aligned}$$

If $\beta_1 = \beta_2$ or $\sigma_{w,x}^2 = 0$ or $m \to \infty$, then this simplifies to $\tilde{\boldsymbol{x}} = \overline{\boldsymbol{x}}$, so that $\mathbb{E}(\boldsymbol{Y} \mid \overline{\boldsymbol{x}}) = \beta_1\boldsymbol{Z}\overline{\boldsymbol{x}}$ as assumed by our estimation model.

Next, $\mathrm{var}(\boldsymbol{Y}) - \mathrm{cov}(\boldsymbol{Y}, \overline{\boldsymbol{x}})\mathrm{var}(\overline{\boldsymbol{x}})^{-1}\,\mathrm{cov}(\overline{\boldsymbol{x}}, \boldsymbol{Y})$ gives us

$$\mathrm{var}(\boldsymbol{Y} \mid \overline{\boldsymbol{x}}) = \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}' + (\beta_2^2\sigma_{w,x}^2 + \sigma_{w,y}^2)\boldsymbol{I}_N$$

where

$$\begin{aligned}\boldsymbol{\Sigma} &= \sigma_{b,y}^2\boldsymbol{V} + \beta_1^2\sigma_{b,x}^2\boldsymbol{V}\Big(\boldsymbol{I}_n - \big(\boldsymbol{I}_n + \frac{\beta_2}{\beta_1}u\boldsymbol{V}^{-1}\big) \quad \text{(S6)}\\ &\qquad \big(\boldsymbol{I}_n + u\boldsymbol{V}^{-1}\big)^{-1}\big(\boldsymbol{I}_n + \frac{\beta_2}{\beta_1}u\boldsymbol{V}^{-1}\big)\Big)\\ &= \sigma_{b,y}^2\boldsymbol{V} - \sigma_{b,x}^2\Big(\beta_1(2\beta_2 - \beta_1)u\boldsymbol{I}_n\\ &\qquad + (\beta_2 - \beta_1)^2u^2\big(\boldsymbol{V} + u\boldsymbol{I}_n\big)^{-1}\Big)\\ &\approx \sigma_{b,y}^2\boldsymbol{V} + \beta_1(\beta_1 - 2\beta_2)\frac{\sigma_{w,x}^2}{m}\boldsymbol{I}_n \text{ if } u = \frac{\eta_x}{m} \to 0\,.\end{aligned}$$

If $\sigma_{w,x}^2 = 0$ or $m \to \infty$, then $\boldsymbol{\Sigma}$ simplifies to $\sigma_{b,y}^2\boldsymbol{V}$ and the residual variance $\mathrm{var}(\boldsymbol{Y} \mid \overline{\boldsymbol{x}})$ is as assumed in our estimation model.

For methods that ignore within-species variation, the conditional distribution of $\overline{\boldsymbol{y}}$ is relevant. From $\overline{\boldsymbol{y}} = \frac{1}{m}\boldsymbol{Z}'\boldsymbol{Y}$ and our results above, we get

$$\begin{aligned}\mathbb{E}(\overline{\boldsymbol{y}} \mid \overline{\boldsymbol{x}}) &= \beta_1\tilde{\boldsymbol{x}}\\ \mathrm{var}(\overline{\boldsymbol{y}} \mid \overline{\boldsymbol{x}}) &= \boldsymbol{\Sigma} + (\beta_2^2\sigma_{w,x}^2 + \sigma_{w,y}^2)/m\;\boldsymbol{I}_n\end{aligned}$$

where $\tilde{\boldsymbol{x}}$ is as in (S5) and $\boldsymbol{\Sigma}$ is as in (S6).

E *Polemonium* leaflet analyses

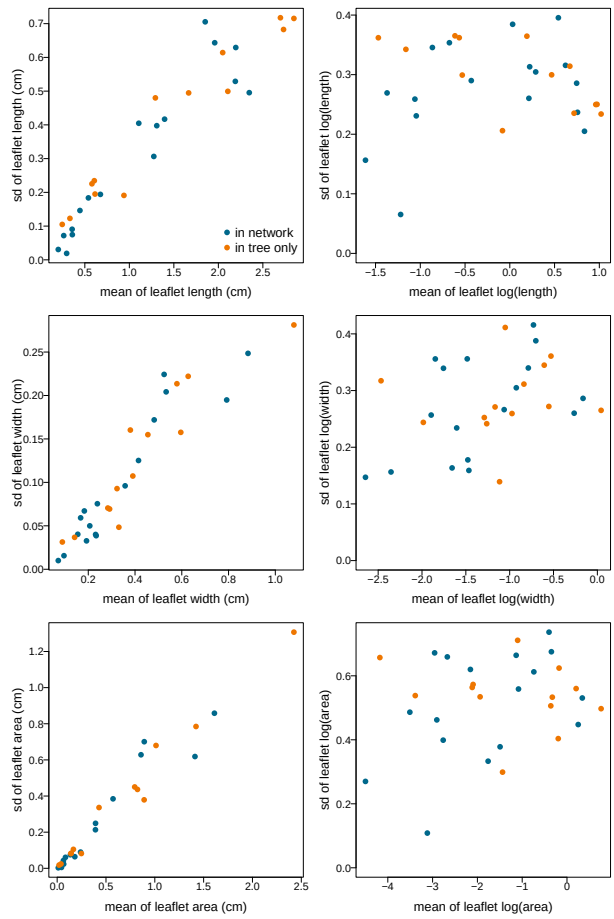

Figure S1: Log-transforming the response stabilizes within-morph variation and decorrelates it from mean response. The sample standard deviation (SD) in leaflet size is positively correlated with mean leaflet size across morphs (left), but not after transformation with the natural log (right). The spread of sample SDs also becomes more compact, reflecting a decrease in the relative variation of sample SDs across morphs.

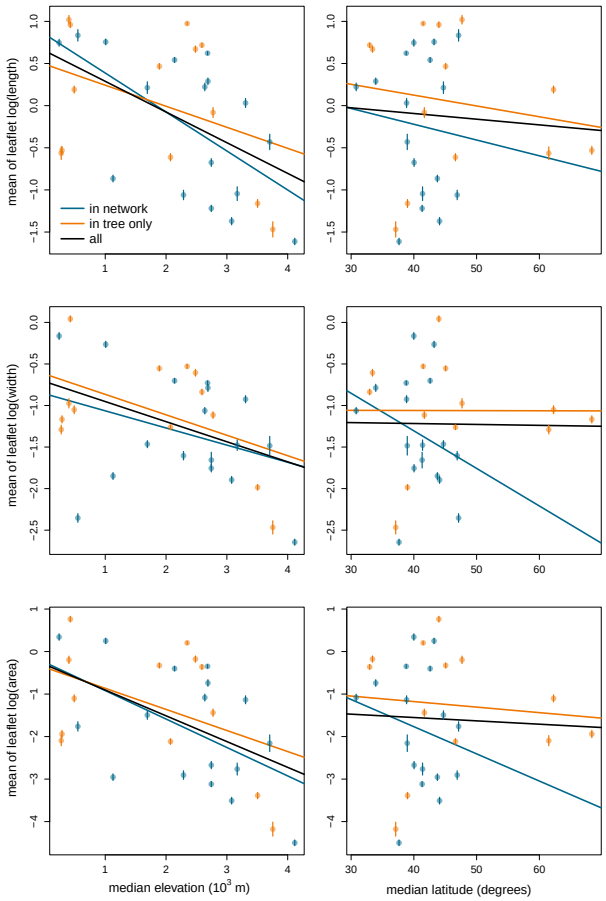

Figure S2: Leaflet size (log-transformed) versus elevation (left) and latitude (right). Each point represents a different morph. Colors indicate sampling across phylogenies: morphs in orange are in both the network and the tree. Morphs in blue are only in the tree. Vertical lines show ±1 standard error. The lines are based on ordinary (non-phylogenetic) simple linear regression using a single predictor, and either the orange or blue points only (orange and blue lines) or all points (black line).

elevation	latitude	$(\hat{\sigma}^2_{\mathrm{b}},\ \hat{\sigma}^2_{\mathrm{w}})$	AIC
(1, 1, ·, ·)			
-0.651 *1.98× 10^{-5}*	-0.0694 *2.15× 10^{-4}*	(0.536, 0.0986)	1072.74
(1, 2, ·, ·)			
-0.649 *1.98× 10^{-5}*	-0.0695 *2.08× 10^{-4}*	(0.535, 0.0986)	1072.39
(2, 1, 1, ·)			
-0.649 *2.06× 10^{-5}*	-0.0693 *2.18× 10^{-4}*	(0.536, 0.0986)	1072.62
(2, 1, 2, ·)			
-0.649 *2.07e-5*	-0.0693 *2.18× 10^{-4}*	(0.535, 0.0986)	1072.6
(2, 2, 1, ·)			
-0.647 *2.06× 10^{-5}*	-0.0694 *2.10× 10^{-4}*	(0.534, 0.0986)	1072.27
(2, 2, 2, ·)			
-0.647 *2.07× 10^{-5}*	-0.0694 *2.11× 10^{-4}*	(0.534, 0.0986)	1072.25

Table S1: Results from fitting $\mathrm{BM_y}$ with REML on the 16 subtrees for log leaflet length. The subtrees are partitioned into 6 groups. Results are identical for all subtrees within the same group. Each group is represented by a 4-tuple, in which the first element is 1 (resp. 2) if *eximium* (resp. *eximium 2*) is selected; the second element is 1 (resp. 2) if *pulcherrimum p.* (resp. *pulcherrimum p. 2*) is selected, and the third element is 1 (resp. 2) if *chartaceum* (resp. *chartaceum 2*) is selected. The last element corresponds to the choice of the *californicum* accession. As it did not affect the results, both choices 1 and 2 are grouped and are represented by a dot. The last group (2, 2, 2, ·) was used in Tables 1, 3 and 4.