PeerJ Preprints NOT PEER-REVIEWED

A simple and general method for simultaneously accounting for phylogenetic and species sampling uncertainty via Rubin's rules in comparative analysis

Shinichi Nakagawa[1,2,4*] and Pierre de Villemereuil[3,,4]

[1]Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia

[2]Diabetes and Metabolism Division, Garvan Institute of Medical Research, Sydney, NSW 2010, Australia

[3]CEFE-CNRS, UMR 5175, 1919 route de Mende, 34293 Montpellier 05, France

[4]Both authors contributed equally

Running head: Rubin's rules may find missing trees and traits

*Correspondence: S. Nakagawa & P. de Villemereuil

e-mail: s.nakagawa@unsw.edu.au & bonamy@horus.ens.fr

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

Abstract

Phylogenetic comparative methods (PCMs), especially ones based on linear models, have played a central role in understanding species' trait evolution. These methods, however, usually assume that phylogenetic trees are known without error or uncertainty, but this assumption is most likely incorrect. So far, Markov chain Monte Carlo, MCMC-based Bayesian methods have mainly been deployed to account for such 'phylogenetic uncertainty' in PCMs. Here, we propose an approach with which phylogenetic uncertainty is incorporated in a simple, readily implementable and reliable manner. Our approach uses Rubin's rules, which are an integral part of a standard multiple imputation procedure, often employed to recover missing data. We see true phylogenetic trees as missing data under this approach. Further, unmeasured species in comparative data (i.e. missing trait data) can be seen as another source of uncertainty in PCMs because arbitrary sampling of species in a given taxon or 'species sampling uncertainty' can affect estimation in PCMs. Using two simulation studies, we show our method can account for phylogenetic uncertainty under many different scenarios (e.g. uncertainty in branching and branch lengths) and, at the same time, it can handle missing trait data (i.e., species sampling uncertainty). A unique property of the multiple imputation procedure is that an index, named 'relative efficiency', could be used to quantify the number of trees required for incorporating phylogenetic uncertainty. Thus, by using the relative efficiency, we show the required tree number is surprisingly small (~50 trees). However, the most notable advantage of our method is that it could be combined seamlessly with PCMs that utilize multiple imputation to handle simultaneously phylogenetic uncertainty (i.e. missing true trees) and species sampling uncertainty (i.e., missing trait data) in PCMs.

Keywords – Bayesian statistics; comparative analysis, data augmentation; information theory; model averaging, phylogenetics,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

(Introduction)

Phylogenetic comparative methods, PCMs, have been playing a central role in investigating trait evolution across species (reviewed in Garamszegi 2014). The most popular methods in comparative biology are based on linear regression such as independent contrasts (Felsenstein 1985), phylogenetic generalized least squares (PGLS; Grafen 1989), or phylogenetic (generalized) linear mixed models (Lynch 1991, Hadfield and Nakagawa 2010). When one phylogenetic tree is used in analysis, all these methods assume that the phylogeny of organisms is known without error.

However, no phylogenetic trees (or hypotheses) are known without error. Errors come in the form of uncertainty in branch length, topology, and also in the model of assumed character evolution (Cooper, et al. 2016, Cornwell and Nakagawa 2017). Researchers have been investigating the impact of these types of uncertainty on statistical inference (e.g., Diaz-Uriarte and Garland 1996, Symonds 2002). These studies generally suggest the importance of incorporating 'phylogenetic uncertainty' in PCMs; note that by using one tree, point estimates (e.g. regression coefficients) are not necessarily biased (Stone 2011), but uncertainty estimates (e.g. standard error or confidence intervals) are not accurate. Therefore, a number of methods have been proposed to include phylogenetic uncertainty (e.g. Losos 1994, Martins 1996, Huelsenbeck, et al. 2000, Housworth and Martins 2001, Rangel, et al. 2015). Among these methods, probably the best one is to use Bayesian Markov Chain Mote Carlo, MCMC (Huelsenbeck, et al. 2000, Huelsenbeck and Rannala 2003, de Villemereuil, et al. 2012); the Bayesian MCMC methods utilize phylogenetic trees sampled from posterior tree set obtained from Bayesian phylogenetic tree estimation programs such as BEAST (Drummond and Rambaut 2007) and MrBayes (Ronquist and Huelsenbeck 2003).

Nonetheless, these methods are not always met with enthusiasm in the evolutionary biology community (cf. Pagel, et al. 2004, Pagel and Meade 2006). Difficulties we see are two-fold: (i) currently, few easy-to-use implementations for such Bayesian MCMC methods are widely available, at least, for regression-based PCMs (but see Hadfield 2010, de Villemereuil, et al. 2012);

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

and (ii) even if implemented, Bayesian MCMC-based analysis may take a long time to process many phylogenetic trees (e.g., see Figure 6 in de Villemereuil, et al. 2012). More recently, Garamszegi and Mundry (2014) have proposed a readily implementable frequentist solution, which employs model averaging with Akaike information criterion (AIC) in PGLS incorporating many phylogenetic trees (see also Mahler et al. 2010). Such a method overcomes the aforementioned difficulties. However, Garamszegi and Mundry (2014) acknowledge the lack of theoretical basis for this proposal, and that theoretical or simulation-based confirmation of their method is necessary.

Here, we propose another solution to account for phylogenetic uncertainty. Our method is simple, generally applicable, and, what is more, it is fairly reliable and readily implementable (see below). Also, it is firmly based on missing data theory (reviewed in Little and Rubin 2002), and utilizes Rubin's rules, which have been proposed as a part of the multiple imputation procedure (Rubin 1987). Evolutionary biologists and ecologists have just recently recognized the usefulness of techniques based on missing data theory (reviewed in Nakagawa and Freckleton 2008, Nakagawa 2015). Also, the importance of these missing-data methods has been discussed in the phylogenetic comparative literature (e.g. Garamszegi and Moller 2011, de Villemereuil and Nakagawa 2014). Notably, multiple imputation has been successfully employed in a number of comparative studies to handle missing data (e.g. Fisher, et al. 2003, Gonzalez-Suarez, et al. 2012, Liker, et al. 2014, Pollux, et al. 2014). Yet, so far, nobody seems to have made a use of Rubin's rules to deal with phylogenetic uncertainty. We note, however, that Martins' work (1996) is conceptually and practically very similar, if not identical, to the proposed method (see also Rangel, et al. 2015).

Paterno et al. (2018) recently discussed three main sources of uncertainty which affect PCMs: 1) phylogenetic uncertainty, 2) species sampling uncertainty, which can be seen as a missing-data problem (because one can see unsampled species as missing data; Nakagawa and Freckleton 2008), and 3) data uncertainty, which include measurement error and within-species variation (see also Rangel, et al. 2015, Cooper, et al. 2016, Cornwell and Nakagawa 2017). Once we could show Rubin's rules can be used for accounting for phylogenetic uncertainty, there is a highly practical

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

possibility that we could seamlessly combine multiple imputation with PCMs to handle missing trait data, thus, addressing species sampling uncertainty simultaneously. There are two ways of imputing missing phenotypic data. The one is that we directly use a phylogenetic correlation (variance-covariance) matrix in the multiple imputation process (e.g., Bruggeman, et al. 2009, Goolsby, et al. 2017; see below for more details). The other is that we employ (phylogenetic) eigenvectors from a phylogenetic correlation (or variance-covariance) matrix (Penone, et al. 2014). These two approaches, surprisingly, have never been systematically compared in terms of performance in augmenting missing comparative data.

Below, we first describe Rubin's rules associated with multiple imputation, and explain the rationale and potential advantages of our proposed method. Then, we conduct two simulation studies: 1) using 12 phylogenetic trees covering different taxa, we compare the performance of our proposed method to other methods such as methods using only one phylogenetic tree and the AIC-based method; and 2) we test how the proposed method can perform with different degrees and types of missing data, when used with the two types of multiple imputation methods (i.e., the one using a phylogenetic correlation matrix and the other phylogenetic eigenvectors).

Multiple imputation and Rubin's rules

Multiple imputation is a three-step process: imputing data, analyzing imputed data and pooling results. In the first step, *m* copies of 'complete' data sets are generated from an incomplete original data set. Popular techniques for the imputation steps use EM/EMB (expectation maximization with bootstrap) and MCMC algorithms, both of which are implemented in R packages such as Amelia (Honaker, et al. 2011), mice (van Buuren and Groothuis-Oudshoorn 2011) and mi (Su, et al. 2011); for more details regarding the algorithms, see Schafer (1997), Enders (2010) and van Buuren (2012). In the second step (analysis), we run separate statistical analyses on *m* data sets. In the final step (pooling), we use Rubin's rules (see below) to aggregate *m* sets of results to produce parameter estimates along with their uncertainty.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

As an example of applying this three-step process to PCMs, let us assume that we have complete data for species traits (see Discussion for cases where missing data exist). Then, what remains missing is the 'true phylogenetic tree'; note that this is the central reason for us using (a part of) multiple imputation to account for phylogenetic uncertainty. Currently, a standard approach to creating candidate trees is to use Bayesian phylogenetic methods, as mentioned above, such as BEAST and MrBayes, which yield a posterior distribution of phylogenetic trees (for a guidance on building phylogenetic trees, see Garamszegi and Gonzalez-Voyer 2014). Alternatively, we can use published Bayesian tree sets as in Jetz et al. (2012) for birds, and Arnold et al. (2010) for primates. We consider this tree generation stage as our imputation step (the first step). The second step can be conducted using any frequentist or Bayesian statistical procedures including PCMs, such as independent contrasts, PGLS and phylogenetic mixed models. Say, we will run PGLS with m randomly sampled phylogenetic trees from a Bayesian posterior tree set, which will result in m sets of results. Then, by combining these result sets via Rubin's rules (the final step), we will have integrated phylogenetic uncertainty in our estimates from PGLS.

Rubin's rules are a set of formulas for combining multiple statistical results, and they are as follows (Rubin 1987). With m imputations, parameters can be estimated by:

$$\bar{\mathbf{b}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{b}^{j}, \quad (1)$$

where $\bar{\mathbf{b}}$ is a k length vector and an average of $\mathbf{b}^j$, and $\mathbf{b}^j$ is the jth set (of m) of k parameter estimates (e.g. regression coefficients). An overall variance-covariance matrix of $\bar{\mathbf{b}}$ is obtained by:

$$\mathbf{V} = \overline{\mathbf{W}} + \left(1 + \frac{1}{m}\right)\mathbf{B}, \quad (2)$$

$$\overline{\mathbf{W}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{W}^{j}, \quad (3)$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

$$\mathbf{B} = \frac{1}{m-1}\sum_{j=1}^{m}(\mathbf{b}^j - \bar{\mathbf{b}})(\mathbf{b}^j - \bar{\mathbf{b}})^T, \quad (4)$$

where **V** is the overall (total) variance(-covariance) matrix for $\bar{\mathbf{b}}$, the within-imputation variance(-covariance) matrix, $\overline{\mathbf{W}}$ is the average of the variance-covariance matrix, $\mathbf{W}^j$ for **b**, and **B** is the between-imputation variance(-covariance) matrix for $\mathbf{b}^j$; note that the standard error of the *i*th parameter (of *k*) is $\sqrt{\mathbf{V}_{ii}}$ (subscript denotes the *i*th row and *i*th column, or *i*th diagonal element). Also, the term, (1+1/*m*) in Equation (2) can be seen as a correction for *m* not being infinite. An important concept in multiple imputation is called, 'fraction of missing information', usually denoted by γ and given by:

$$\bar{\gamma} = \left(1 + \frac{1}{m}\right)\mathrm{tr}\left(\mathbf{B}\mathbf{V}^{-1}\right)\frac{1}{k}, \quad (5)$$

where $\bar{\gamma}$ is the initial estimate of the fraction of missing information, ranging from 0 to 1 (see below; cf. Equation (12)), and the term, tr($\mathbf{BV}^{-1}$) denotes the trace of the resulting matrix from $\mathbf{BV}^{-1}$. We can appreciate why $\bar{\gamma}$ is termed 'the fraction of missing information' because it represents a proportion of the between-imputation variance to the total (overall) variance (note that it may be easier to see this in Equation (8) below). In other words, it represents the proportion of the parameter uncertainty due to using different trees. We can obtain statistical significance and confidence intervals based on *t* distributions with the degrees of freedom of the following:

$$\bar{v} = (m-1)\frac{1}{\bar{\gamma}^2}, \quad (6)$$

where $\bar{v}$ is the degrees of freedom to be used for *t* values ($\mathbf{b}_i / \sqrt{\mathbf{V}_{ii}}$). However, since the parameters will not be influenced equally by the phylogenetic uncertainty, it is probably better to obtain a fraction of missing information value for each parameter ($\mathbf{b}_i$) rather than omnibus values as in

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

Equations (5 and 6) (Lipsitz et al. 2002). Such separate values of the degree of freedom (v_i) can be obtained by:

$$\gamma_i = \left(1+\frac{1}{m}\right)\left(\frac{\mathbf{B}_{ii}}{\mathbf{V}_{ii}}\right), \quad (7)$$

$$v_i = (m-1)\frac{1}{\gamma_i^2}. \quad (8)$$

However, the formulation of v_i or $\overline{v}$ assumes a very large sample size, n (which is the length of data when no data are missing; Rubin and Schenker 1986, Rubin 1987). Barnard & Rubin (1999) proposed the following adjustment in the degrees of freedom (cf. Lipsitz, et al. 2002):

$$v_i^* = \left(\frac{1}{v_i}+\frac{1}{v_{\text{obs}(i)}}\right)^{-1}, \quad (9)$$

$$v_{\text{obs}(i)} = (1-\gamma_i)\left(\frac{v_{\text{com}}+1}{v_{\text{com}}+3}\right)v_{\text{com}}, \quad (10)$$

$$v_{\text{com}} = n-k \quad (11)$$

where v_i^* is the degrees of freedom for ith parameter, especially suitable when sample size, n is small. The degrees of freedom, v_{obs} denotes the observed degrees of freedom, whereas v_{com} denotes the complete degrees of freedom (i.e. the degrees of freedom for the complete data set assuming no missing data). In the next section, we will compare the performance of both v_i (hereafter denoted "original df") and v_i^* (hereafter denoted "corrected df").

Once we have an estimate of the corrected degrees of freedom, we can obtain a refined estimate of the fraction of missing information, γ_i^* for each parameter:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

$$\gamma_i^* = \left(1+\frac{1}{m}\right)\frac{\mathbf{B}_{ii}}{\mathbf{V}_{ii}} + \frac{2}{(\nu_i^*+3)\mathbf{V}_{ii}}. \quad (12)$$

Then, we can use γ_i^* to find a very useful quantity called 'relative efficiency', which is given by:

$$\varepsilon_i = \left(1+\frac{\gamma_i^*}{m}\right)^{-1}, \quad (13)$$

where ε_i is relative efficiency of the *i*th parameter and ranges from 0 to 1. Relative efficiency represents the efficacy of multiple imputation process, compared to the case of *m* being infinite. In other words, this number can be used to assess how many imputations (*m*) are needed to account for uncertainty due to missing data. In our case, relative efficiency can indicate how many phylogenetic trees we should use for analysis (typically, the number of required trees to account for phylogenetic uncertainty are chosen arbitrarily). Notably, to achieve fairly high relative efficiency, the required number of *m* is surprisingly low, even when the fraction of missing information is relatively large. For example, with $\gamma = 0.5$ and $m = 5$, relative efficiency is 90.91%, while it is 95.24% when $\gamma = 0.5$ and $m = 10$. Rubin's (1987) initial recommendation of *m* was low (3-10) probably due to computational limitation at that time, but current thinking is to use much larger *m*, aiming at over 99% relative efficiency (e.g. Graham, et al. 2007, von Hippel 2009, Nakagawa 2015). As you see in Equation (13), we obtain a relative efficiency value (ε_i) for every parameter and such values vary among parameters. For assessing efficiency of a model, we will use the relative efficiency (ε^*) that is obtained from the largest value of the fraction of missing information, following McKnight et al. (2007); that is:

$$\varepsilon^* = \left(1+\frac{\max(\gamma_i^*)}{m}\right)^{-1}, \quad (14)$$

where $\max(\gamma_i^*)$ denotes the maximum (largest) value of γ_i^*; the use of the maximum value of γ_i^* ensures all parameters will achieve at least a certain relative efficiency level or above. We can

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

194 easily automate calculations involving the above formulae with currently available R packages for
195 multiple imputation such as mice (reviewed in Nakagawa and Freckleton 2011; see also Penone, et
196 al. 2014).

197 Simulation studies

198 *Incorprating phylogenetic uncertainty as missing data*

199 In order to assess the overall quality of our new method and compare it to existing ones, we
200 performed a simulation study using 12 trees extracted from TreeBase (the number of tips ranging
201 from 67 to 174; www.treebase.org, see Supplementary Table 1). We simulated data sets in which a
202 variable *y* was linearly predicted from a variable *x*, with an intercept of 5 and a slope of 2. The error
203 structure of this relationship was constrained by the phylogenetic tree chosen among the 12 trees
204 (hereafter called the 'true tree'), following a Brownian motion model. Different residual standard
205 deviations were used (sigma, $\sigma = 2, 5, 10$ or 15). From the true tree, a distribution of trees was
206 created by altering branch lengths and topology. To alter branch lengths, random noise drawn from
207 a uniform distribution centered around 0 was added to the true value. The maximum level of that
208 noise varied between 0% (no branch length noise), 10%, 20%, 40%, 70% or 90% of the true branch
209 length. To alter topology, we randomly "swapped" branches belonging to a focal clade to a sister
210 clade. To choose the branch to swap, a tip was chosen at random, and a "threshold" was chosen
211 from a uniform distribution with the thresholds of [0.1, 1]. The node just below this threshold in the
212 path from the tip chosen to the root was swapped. We used several levels of topological noise (no
213 swaps, i.e. no topological noise, or 1, 2, 5, 10, 20, 30 swaps in the tree). To construct the
214 distribution of trees, the probability of each swap was set to 0.5. For each set of parameters (true
215 tree, level of branch noise, level of topological noise), we constructed a distribution of 100 trees and
216 replicated the analysis 100 times. This resulted in 2016 conditions, hence 201,600 different
217 analyzes. Using the simulated phenotypes and tree distributions, we compared GLS using the true
218 tree or two types of consensus trees (majority rule or consensus), with both multiple GLS with

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

pooling of the results using AIC averaging (as in Garamszegi and Mundry 2014) and pooling with Rubin's rules as described above (either using the original degrees of freedom, df, or the corrected df as in Equation (9)).

The accuracy of the intercept and slope were only slightly influenced by the different parameters (Table 1 and Fig. S1, S2 and S3). On the contrary, the estimation of the residual standard deviation depended strongly on the method used (as well as, trivially, the true parameter sigma, and to a far lesser extent, all of the other parameters, see Table 1). Notably, the estimation of residual standard deviation was biased upward for the two methods using consensus trees (strict or majority rule, see Fig. S1, S2 and S3).

The coverage of the confidence interval for the slope was heavily influenced by the method used and more marginally by other parameters (except the true parameter sigma which had negligible influence, Table 1). The coverage was correctly calibrated when using the true tree (True GLS, Fig. 1) and heavily mis-calibrated when using consensus trees (strict and majority rule consensus GLS, Fig. 1). Accounting for uncertainty yielded better-calibrated coverages. AIC averaging was the closest to correct calibration. It was, however, slightly but consistently too liberal (Fig. 1). Using Rubin's rule yielded conservative coverages. Contrary to AIC averaging, the coverage was sensitive to the level of branch length and/or topological noise, decreasing when the noise increased (thus being even more conservative, Fig. 1).

In order to assess the behavior of the proposed method using Rubin's rules to account for phylogenetic uncertainty, we also conducted a study using different sample size for the trees (T = 10, 20, 50 or 100) and computed the relative efficiency as shown in Equation (14). This analysis revealed two interesting patterns (Fig. 2). First, no efficiency lower than 0.90 was recorded for a total of 806,400 simulated data sets, even for a sample size of trees as low as T = 10. Second, the relative efficiency depended strongly on the number of trees used (Fig. 2 and Table 1). It also depended on the level of branch length noise, and to a lesser extent, on the level of topological noise

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

(Fig. 2 and Table 1), as well as, even more marginally, on the nature of the true tree (Table 1). Third, in order to reach a relative efficiency over 0.99, on average, only 50 trees were necessary even with high levels of branch length and topological noise. With 100 trees, the relative efficiency was always over 0.99.

Incorporating both phylogenetic uncertainty and missing trait data

We then investigated the possibility to combine the ability of multiple imputation to account simultaneously for phylogenetic uncertainty and missing phenotypic values. To do so, we conducted a study with parameters fixed to the following values: the residual standard deviation σ was set to 5, the branch length noise to 20% and topological noise to 2 swaps. For simulated data according to these parameters, we deleted records of phenotypic values at various proportions (10%, 30% and 50%) and according to three mechanisms inspired from Penone *et al.*(2014): values were missing completely at random (MCAR), missing at random according to the environmental variable (MARvar) or missing at random according to the phylogeny (MARphylo). For more details of missing data mechanisms (e.g. MCAR, MAR), see Little and Rubin (2002; see also Nakagawa and Freckleton 2008). The multiple imputation of the missing phenotypic values were handled using two different methods: on the one hand, we used an R implementation of the method PhyloPars (Bruggeman, et al. 2009), called Rphylopars (Goolsby, et al. 2017), to impute the missing values according to both the phylogeny and environmental (non-missing) data (hereafter, the matrix method). On the other hand, we used the method described in Penone *et al.* (2014) using the information contained in phylogenetic eigenvectors (Diniz, et al. 1998; see also Guenard, et al. 2013) to impute the missing vales (hereafter, the eigenvector method).

The results of our simulations show that the matrix method (RphyloPars) yielded estimates with little bias (Fig. 3A, especially when missing values are missing according to the phylogeny, MARphylo), while using eigenvectors resulted in a stronger bias, strongly increasing with the proportion of missing values. Overall, the level of bias strongly depended on the characteristics of the true tree and the method used, and only slightly on the rate of missing values (Table 2).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

Coverage analysis of the confidence intervals (Fig. 3B) show that the matrix method is slightly too liberal when values are missing completely at random (MCAR) or missing at random according to the environmental variable (MARvar), but slightly conservative when they are missing at random according to the phylogeny (MARphylo). By contrast, the eigenvector method produced the coverage too liberal to be useful, although, interestingly, decreasing with the proportion of missing values. Overall, the coverage depended mostly the true tree and method used, and only marginally on the mechanism and rate of missing values (Table 2). The strong influence of the true tree on the estimate and its coverage is mainly driven by a strong instability of the eigenvector method regarding a particular tree (Tree #11 in Figure S4 and Table S1). Removing this tree from the analysis does not qualitatively impact the results shown in Fig. 3. However, this example makes an interesting point about the eigenvector method being potentially very sensitivity to the nature of a phylogenetic tree.

Discussion

The aim of this article is to introduce a simple and readily implementable method (i.e. Rubin's rubles) to account for phylogenetic uncertainty in phylogenetic comparative methods, PCMs. More practically, we explored the use of Rubin's rules simultaneously handling phylogenetic uncertainty and species sampling uncertainty (i.e. missing trait data; see Paterno, et al. 2018). Via a simulation study using a simple PGLS, we compared the proposed method using Rubin's rules with other existing methods across different levels of branch length and topological noise, and we also assessed the number of trees required to accurately account for phylogenetic uncertainty. Further, we tested the practicality of our method to handle missing trait data under different imputation procedures and missing-data mechanisms. Four main results have emerged from our simulation study.

First, in terms of error rate, methods ignoring phylogenetic uncertainty performed poorly and had a bad coverage for the slope confidence interval (CI). These findings are concordant with the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

previous work by de Villemereuil et al. (2012) comparing different methods. Both our proposed methods using Rubin's rule and the AIC-based method were much closer to the expected results using a PGLS with the true tree. Hence, using a consensus tree (either being a strict consensus or a majority rule based one) will yield too narrow CI, meaning that any test framework linked to it (e.g. slope significance testing) will yield an uncontrolled type I error rate.

The second main result is that the behavior of the methods accounting for phylogenetic uncertainty differed between them and depends on the level of phylogenetic noise in the tree distribution. Whereas the AIC-based method was consistently slightly too liberal, our proposed method using Rubin's rule was, by contrast, slightly conservative. The method assuming infinite sample size ("original df") was less conservative than the method correcting for small sample size ("corrected df"). This conservative behavior depended on the level of noise: our proposed method became more conservative as the level of phylogenetic noise increased. The AIC-based method was, on the contrary, less sensitive to the level of noise.

The third main result is that the number of phylogenetic trees needed to correct for phylogenetic uncertainty is surprisingly low. The required number of trees is far less than 1000 (as in Garamszegi and Mundry 2014), and probably less than 100 (as in de Villemereuil, et al. 2012). It is likely to be a matter of dozens. In our simulation, sets of 50 randomly selected trees achieved almost always over 99% relative efficiency; in other words, using 50 trees should be almost as good as using an infinite number of trees. For low to medium levels of noise, even a sample size as low as 10 trees almost always yielded over 99% relative efficiency. As a whole, we recommend the use of over 50 phylogenetic trees in a PCM to account for phylogenetic uncertainty. However, for any given analysis and tree set, we recommend checking the number of trees needed to reach a relative efficiency of 99% (Nakagawa 2015). In practice, indeed, the required number of trees required to achieve high efficiency will strongly depend on the phenotypic data (e.g., phylogenetic signal), the complexity of the model and the variability in the tree estimates (e.g. strong topological and branch length uncertainty). We note that the statistical literature has discussed other criteria apart from the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

relative efficiency to determine how many imputations one requires (see Graham, et al. 2007, White, et al. 2011).

As mentioned, the AIC-based method (Garamszegi and Mundry 2014) accounted for phylogenetic uncertainty performed well, although with slightly liberal CIs. Therefore, the AIC-based method is definitely an option to correct for phylogenetic uncertainty. The method based on Rubin's rules (or multiple imputation), despite being slightly conservative, has the advantage of being a theoretical founded, yet simple method (we note that being conservative is probably preferred to being slightly liberal). This is, given that the imputation step is 'proper', which is the case here as long as the trees come from a Bayesian posterior distribution and the estimates are Maximum Likelihood Estimators (e.g. BEAST/PGLS combination, for example; for the definition on proper multiple imputation, see Rubin 1987, Nielsen 2003). However, there is another clear benefit of using the proposed method.

This leads to our fourth point, that is, multiple imputation can simultaneously handle missing trait data (species sampling uncertainty) and phylogenetic uncertainty in a comparative data set. Especially, using the matrix method (PhyloPars; Bruggeman, et al. 2009; implemented as Rphylopars by Goolsby, et al. 2017) to account for missing phenotypic values, while accounting for the phylogenetic uncertainty at the same time, yields estimate with little bias on the slope and almost calibrated coverage of the confidence interval. Using the eigenvector method, as suggested in Penone *et al.* (2014) does not seem to yield satisfying results, however. The sensitivity of the matrix method (Rphylopars) to the rate and mechanism of missing data was relatively small, suggesting that the method should perform fairly well in many different circumstances. An exception to this is that when missing values are missing at random according to the phylogeny, the matrix method is slightly too conservative, while it is slightly too liberal for the two other missing-data mechanisms we tested here. Given the pervasive nature of missing data, we suggest multiple imputation may be useful for virtually every comparative data set (Nakagawa and Freckleton 2008, Garamszegi and Moller 2011). Note that Rphylopars is intended to produce point estimate of the missing phenotypic value with standard errors, which can be used to produce multiple imputation as

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

we did. However, this process might not conserve all the properties of the multiple imputation model (e.g., it might slightly decreased covariance between species in the multiple imputation). Work is being conducted on a more proper multiple imputation method using a matrix method for missing values in the context of phylogenetic comparative analysis (S. Blomberg, Pers. Comm., see also the package in development at https://github.com/pdrhlik/phylomice). We provide implementations of our method using R at GitHub repository (https://github.com/devillemereuil/SimulTrees).

It is notable that the procedure known as 'data augmentation' can also be used for dealing with missing data instead of multiple imputation. The term, data augmentation is used in a number of ways in the statistical literature, but here we follow the usage by McKnight et al. (2007); that is, in this procedure, uncertainty of missing data is incorporated in to parameter estimates during analysis (see the original usage of this term as in Tanner and Wing 1987). A data augmentation procedure is implemented, for instance, in MCMCglmm (Hadfield 2010). However, there is one disadvantage to data augmentation, which does not affect multiple imputation. Data augmentation assumes the use of just identified or over-identified models (Enders and Bandalos 2001, Enders 2010). That is, a particular model (for imputation) includes enough or more predictor variables, so that missing values can be recovered accurately from these predictors. In contrast, because multiple imputation separates the steps of data imputation and analysis, we do not need to clutter a statistical model for analysis (i.e. the analysis step) with many variables, which assist in recovering missing values (known as auxiliary variables; Enders 2010, Nakagawa 2015). Technically speaking, auxiliary variables are supported to make missing values to fulfill the assumption of missing at random, MAR (Little and Rubin 2002). In a multiple imputation procedure, we need add auxiliary variables only to a statistical model for imputation (i.e. the imputation step). For example, known data on species body size can be used during the imputation step to help recover missing data on species longevity, given the strong correlation between the two. However, because multiple imputation separates imputation and analysis, body size does not need to be a part of the final model. The use

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

of multiple imputation probably has wider applications over data augmentation. Most importantly, to integrate phylogenetic uncertainty in a comparative data set with missing data, one just needs to conduct extra imputations (e.g. more *m* as in Equation (1)) to include the adequate number of trees, which can be measured by the efficiency index as in Equation (13).

In conclusion, the method using Rubin's rules is readily usable for all comparative biologists. Clearly, the use of multiple imputation used with the matrix method is extremely useful not only for imputing missing trait data, but also for integrating phylogenetic uncertainty, even simultaneously, as we have shown above. We expect such a simultaneous use of these two aspects of multiple imputation to be common in phylogenetic comparative analyses in the near future.

Acknowledgements

We thank Will Cornwell, Travis Ingram, Losia Lagisz, Alistair Senior and Simon Blomberg for comments, which have improved the manuscript. We also thank Eric Goolsby who provided help with Rphylopars. SN was supported by an ARC Future Fellowship (FT130100268).

References

Arnold C, Matthews LJ, Nunn CL. 2010. The 10k trees website: A new online resource for primate phylogeny. Evol Anthropol, 19:114-118.

Barnard J, Rubin DB. 1999. Small-sample degrees of freedom with multiple imputation. Biometrika, 86:948-955.

Bruggeman J, Heringa J, Brandt BW. 2009. Phylopars: Estimation of missing parameter values using phylogeny. Nucleic Acids Res, 37:W179-W184.

Cooper N, Thomas GH, FitzJohn RG. 2016. Shedding light on the "dark side' of phylogenetic comparative methods. Methods Ecol Evol, 7:693-699.

Cornwell W, Nakagawa S. 2017. Phylogenetic comparative methods. Curr Biol, 27:R333-R336.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

396 de Villemereuil P, Nakagawa S. 2014. General quantitative genetic methods for comparative
397 biology. In: Garamszegi LZ editor. Modern phylogenetic comparative methods and their
398 application in evolutionary biology, Springer Berlin Heidelberg, p. 287-303.
399 de Villemereuil P, Wells JA, Edwards RD, Blomberg SP. 2012. Bayesian models for comparative
400 analysis integrating phylogenetic uncertainty. BMC Evol Biol, 12.
401 Diaz-Uriarte R, Garland T. 1996. Testing hypotheses of correlated evolution using phylogenetically
402 independent contrasts: Sensitivity to deviations from brownian motion. Syst Biol, 45:27-47.
403 Diniz JAF, De Sant'ana CER, Bini LM. 1998. An eigenvector method for estimating phylogenetic
404 inertia. Evolution, 52:1247-1262.
405 Drummond AJ, Rambaut A. 2007. Beast: Bayesian evolutionary analysis by sampling trees. BMC
406 Evol Biol, 7.
407 Enders CK. 2010. Applied missing data analysis. New York, Guilford Press.
408 Enders CK, Bandalos DL. 2001. The relative performance of full information maximum likelihood
409 estimation for missing data in structural equation models. Struct Equ Modeling, 8:430-457.
410 Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat, 125:1-15.
411 Fisher DO, Blomberg SP, Owens IPF. 2003. Extrinsic versus intrinsic factors in the decline and
412 extinction of australian marsupials. Proc R Soc Lond B, 270:1801-1808.
413 Garamszegi L, Gonzalez-Voyer A. 2014. Working with the tree of life in comparative studies: How
414 to build and tailor phylogenies to interspecific datasets. In: Garamszegi LZ editor. Modern
415 phylogenetic comparative methods and their application in evolutionary biology, Springer
416 Berlin Heidelberg, p. 19-48.
417 Garamszegi L, Mundry R. 2014. Multimodel-inference in comparative analyses. In: Garamszegi LZ
418 editor. Modern phylogenetic comparative methods and their application in evolutionary
419 biology, Springer Berlin Heidelberg, p. 305-331.
420 Garamszegi LZ. 2014. Modern phylogenetic comparative methods and their application in
421 evolutionary biology. New York, Springer, p. pages cm.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

422 Garamszegi LZ, Moller AP. 2011. Nonrandom variation in within-species sample size and missing
423 data in phylogenetic comparative studies. Syst Biol, 60:876-880.
424 Gonzalez-Suarez M, Lucas PM, Revilla E. 2012. Biases in comparative analyses of extinction risk:
425 Mind the gap. J Anim Ecol, 81:1211-1222.
426 Goolsby EW, Bruggeman J, Ane C. 2017. Rphylopars: Fast multivariate phylogenetic comparative
427 methods for missing data and within-species variation. Methods Ecol Evol, 8:22-27.
428 Grafen A. 1989. The phylogenetic regression. Philos T Roy Soc B, 326:119-157.
429 Graham JW, Olchowski AE, Gilreath TD. 2007. How many imputations are really needed? - some
430 practical clarifications of multiple imputation theory. Prev Sci, 8:206-213.
431 Guenard G, Legendre P, Peres-Neto P. 2013. Phylogenetic eigenvector maps: A framework to
432 model and predict species traits. Methods Ecol Evol, 4:1120-1131.
433 Hadfield J. 2010. Mcmc methods for multi-response generalized linear mixed models: The
434 mcmcglmm r package. J Stat Softw, 33:1-22.
435 Hadfield JD, Nakagawa S. 2010. General quantitative genetic methods for comparative biology:
436 Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J.
437 Evol. Biol., 23:494-508.
438 Honaker J, King G, Blackwell M. 2011. Amelia ii: A program for missing data. J Stat Softw, 45:1-
439 47.
440 Housworth EA, Martins EP. 2001. Random sampling of constrained phylogenies: Conducting
441 phylogenetic analyses when the phylogeny is partially known. Syst Biol, 50:628-639.
442 Huelsenbeck JP, Rannala B. 2003. Detecting correlation between characters in a comparative
443 analysis with uncertain phylogeny. Evolution, 57:1237-1247.
444 Huelsenbeck JP, Rannala B, Masly JP. 2000. Accommodating phylogenetic uncertainty in
445 evolutionary studies. Science, 288:2349-2350.
446 Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space
447 and time. Nature, 491:444-448.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

448 Liker A, Freckleton RP, Szekely T. 2014. Divorce and infidelity are associated with skewed adult
449 sex ratios in birds. Curr Biol, 24:880-884.
450 Lipsitz SR, Parzen M, Zhao LP. 2002. A degrees-of-freedom approximation in multiple imputation.
451 J Stat Comput Sim, 72:309-318.
452 Little RJA, Rubin DB. 2002. Statistical analysis with missing data. 2nd ed. Hoboken, N.J., Wiley.
453 Losos JB. 1994. An approach to the analysis of comparative data when a phylogeny is unavailable
454 or incomplete. Syst Biol, 43:117-123.
455 Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution,
456 45:1065-1080.
457 Martins EP. 1996. Conducting phylogenetic comparative studies when the phylogeny is not known.
458 Evolution, 50:12-22.
459 McKnight PE, McKnight KM, Sidani S, Figueredo AJ. 2007. Missing data: A gentle introduction.
460 New York, NY, The Guilford Press.
461 Nakagawa S. 2015. Missing data: Mechanisms, methods and messages In: Fox GA, Negrete-
462 Yankelevich S, Sosa VJ editors. Ecological statistics. Oxford, Oxford University Press, p. 81-
463 105.
464 Nakagawa S, Freckleton R. 2011. Model averaging, missing data and multiple imputation: A case
465 study for behavioural ecology. Behav Ecol Sociobiol, 65:103-116.
466 Nakagawa S, Freckleton RP. 2008. Missing inaction: The dangers of ignoring missing data. Trends
467 Ecol Evol, 23:592-596.
468 Nielsen SF. 2003. Proper and improper multiple imputation. Int Stat Rev, 71:593-607.
469 Pagel M, Meade A. 2006. Bayesian analysis of correlated evolution of discrete characters by
470 reversible-jump markov chain monte carlo. Am Nat, 167:808-825.
471 Pagel M, Meade A, Barker D. 2004. Bayesian estimation of ancestral character states on
472 phylogenies. Syst Biol, 53:673-684.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

473 Paterno GB, Penone C, Werner GD. 2018. Sensiphy: An r-package for sensitivity analysis in
474 phylogenetic comparative methods. Methods Ecol Evol.
475 Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, Brooks TM, Young BE,
476 Graham CH, Costa GC. 2014. Imputation of missing data in life-history trait datasets: Which
477 approach performs the best? Methods Ecol Evol, 5:961-970.
478 Pollux BJA, Meredith RW, Springer MS, Garland T, Reznick DN. 2014. The evolution of the
479 placenta drives a shift in sexual selection in livebearing fish. Nature, 513:233-+.
480 Pratt JW. 1987. Dividing the indivisible: Using simple symmetry to partition variance explained.
481 Proceedings of the second international Tampere conference in statistics, 1987, Department of
482 Mathematical Sciences, University of Tampere, p. 245-260.
483 Rangel TF, Colwell RK, Graves GR, Fucikova K, Rahbek C, Diniz JAF. 2015. Phylogenetic
484 uncertainty revisited: Implications for ecological analyses. Evolution, 69:1301-1312.
485 Ronquist F, Huelsenbeck JP. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed
486 models. Bioinformatics, 19:1572-1574.
487 Rubin DB. 1987. Multiple imputation for nonresponse in surveys. New York, NY, J. Wiley & Sons.
488 Rubin DB, Schenker N. 1986. Multiple imputation for interval estimation from simple random
489 samples with ignorable nonresponse. J Am Stat Assoc, 81:366-374.
490 Schafer JL. 1997. Analysis of incomplete multivariate data. London, Chapman & Hall.
491 Stone EA. 2011. Why the phylogenetic regression appears robust to tree misspecification. Syst Biol,
492 60:245-260.
493 Su YS, Gelman A, Hill J, Yajima M. 2011. Multiple imputation with diagnostics (mi) in r: Opening
494 windows into the black box. J Stat Softw, 45:1-31.
495 Symonds MRE. 2002. The effects of topological inaccuracy in evolutionary trees on the
496 phylogenetic comparative method of independent contrasts. Syst Biol, 51:541-553.
497 Tanner MA, Wing HW. 1987. The calculation of posterior distributions by data augmentation. J Am
498 Stat Assoc, 82:528-540.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

van Buuren S. 2012. Flexible imputation of missing data. Boca Raton, FL, CRC Press.

van Buuren S, Groothuis-Oudshoorn K. 2011. Mice: Multivariate imputation by chained equations in R. J Stat Softw, 45:1-67.

von Hippel PT. 2009. How to impute interactions, squares and other transformed variables. Sociol Methodol, 39:265-291.

White IR, Royston P, Wood AM. 2011. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med, 30:377-399.

Figure legends

Figure 1. Complementary of the coverage (1 - coverage) for 95% confidence intervals for the different estimation methods against the two types of noise (left: branch length noise, right: topological noise). Grey area is the zone of non-significance for a binomial test with a true probability of 0.05 (i.e. expected complementary coverage).

Figure 2. Relative efficiency distribution for different tree sample size (T) and different levels of branch length noise (BLN) and topological noise (Nb. Swap).

The boxes depict the 50% inter-quantile interval, the whiskers depict the 95% inter-quantile interval and the horizontal bar is the average estimate. The red lower dot is the minimal relative efficiency yielded during the simulations.

Figure 3. Estimate of the slope (A) and complementary of the coverage (1 - coverage) of its associated confidence interval (B) for the two methods of multiple imputation of missing phenotypic values (PhyloPars and Eigenvectors) according to the proportion of missing values in the data and mechanism of missing values: MCAR, missing completely at random; MARvar, missing at random according to the environmental variable: MARphylo, missing at random

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

according the phylogeny. Grey area in B is the zone of non-significance for a binomial test with a true probability of 0.05 (i.e. expected complementary coverage).

Figure S1. Average estimates of the intercept, slope and residual standard deviation for the different estimation methods and true vales for sigma, according to the level of branch length noise. The true value of the intercept is 5 and the true value for the slope is 2.

Figure S2. Average estimates of the intercept, slope and residual standard deviation for the different estimation methods and true vales for sigma, according to the level of topological noise (i.e. number of swaps). The true value of the intercept is 5 and the true value for the slope is 2.

Figure S3. Average estimates of the intercept, slope and residual standard deviation for the different estimation methods and true vales for sigma, according to the true tree used to construct the distribution of trees. The true value of the intercept is 5 and the true value for the slope is 2.

Figure S4. Average estimates according to the true tree, methods (PhyloPars or Eigenvectors), mechanisms (MCAR, MARvar, MARphylo; see the main text) and proportion of missing values. The true value of the slope is 2.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

PeerJ Preprints NOT PEER-REVIEWED

Table 1. Variance partitioning using a linear model to model the distribution of the inferred parameters, confidence interval coverage and efficiency. The total R^2 of the linear model is given, followed by the relative contribution (i.e. relative Pratt's measure; Pratt 1987) from each parameter to the total R^2. Relative contributions sum up to 1. "Number of trees" was available only for the study of efficiency.

Parameter Estimation	Model R^2	Parameter contribution to R^2					
		True Tree	Method	Sigma	Branch Length Noise	Topology Noise	Number of trees
Intercept	0.0075	0.51	0.018	0.29	0.062	0.12	–
Slope	0.007	0.8	0.041	0.027	0.11	0.026	–
Residual St. Dev.	0.79	0.043	0.3	0.66	0.00015	0.0017	–
CI Coverage							
Slope	0.66	0.013	0.98	3.4 x 10^-5	0.0019	0.0055	–
Efficiency analysis							
Efficiency	0.71	0.023	–	1.9 x 10^-7	0.37	0.037	0.58

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

PeerJ Preprints NOT PEER-REVIEWED

Table 2. Variance partitioning using a linear model to model the distribution of the inferred slope and confidence interval coverage in the simulation study on missing values. The total R^2 of the linear model is given, followed by the relative contribution (i.e. relative Pratt's measure; Pratt 1987) from each parameter to the total R^2. Relative contributions sum up to 1.

Parameter Estimation	Model R^2	Parameter contribution to R^2			
		True Tree	Method	Mechanism	Proportion of missing
Slope	0.39	0.41	0.43	0.0078	0.15
CI Coverage					
Slope	0.65	0.33	0.59	0.026	0.056

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

PeerJ Preprints NOT PEER-REVIEWED

Table S1. Information regarding the 12 TreeBase trees used in the simulation analysis.

No. Tree	No. Taxa	Date	Journal	Taxon info	First Author	Title
1	88	2010	*Evolution*	Plants (Legume)	Marazzi, Brigitte	Large-Scale Patterns of Diversification in the Widespread Legume Genus Senna and the Evolutionary Role of Extrafloral Nectaries.
2	102	2011	*Fungal Biology*	Fungi	Voglmayr, Hermann	The diversity of ant-associated black yeasts: Insights into a newly discovered world of symbiotic interactions
3	110	2011	*BMC Evolutionary Biology*	Animals (Fishes)	Nakatani, Masanori	Evolutionary history of Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation
4	67	2011	*Taxon*	Fungi	Justo, Alfredo	Phylogenetic classification of Trametes (Basidiomycota, Polyporales) based on a five-marker dataset
5	94	2011	*Nature*	Animals (Lizards)	Alfoldi, Jessica	The genome of *Anolis carolinensis*, the green anole lizard, and a comparative analysis with birds and mammals
6	146	2004	*Proceedings of the National Academy of Sciences (PNAS)*	Animals (Birds)	Barker, F. Keith	Phylogeny and diversification of the largest avian radiation.
7	147	2013	*Annals of the Missouri Botanical Garden*	Plants (Asterids)	Liede-Schumann, Sigrid	The Orthosiinae revisited (Apocynaceae, Asclepiadoideae, Asclepiadeae)
8	81	2011	*Molecular Phylogenetics and Evolution*	Plants (Monocots)	Nauheimer, Lars	Giant taro and its relatives: A phylogeny of the large genus Alocasia (Araceae) sheds light on miocene floristic exchange in the malesian region
9	93	2011	*Zoologica Scripta*	Animals (Squamates)	Heinicke, Matthew	Phylogeny of a trans-Wallacean radiation (Squamata, Gekkonidae, Gehyra) supports a single early colonization of Australia
10	75	2012	*American Naturalist*	Animals (Birds)	Claramunt, Santiago	Ecological opportunity and diversification in a continental radiation of birds: Climbing adaptations and cladogenesis in the Furnariidae

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

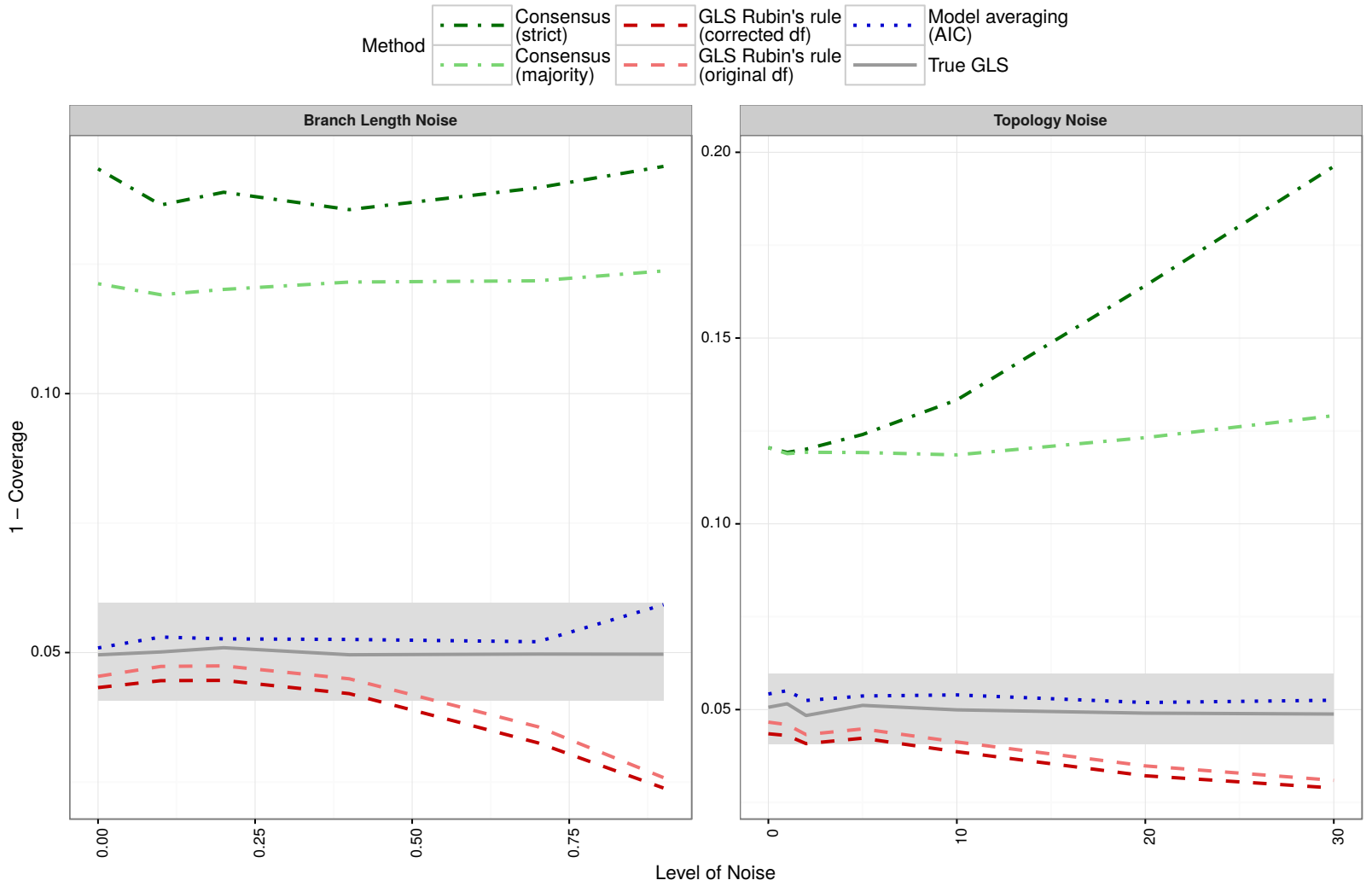
PeerJ Preprints NOT PEER-REVIEWED

11	139	2013	*Molecular Phylogenetics and Evolution*	Animals (Fishes)	Unmack, Peter	Phylogeny and biogeography of rainbow fishes (Melanotaeniidae) from Australia and New Guinea
12	102	2014	*Journal of Biogeography*	Plants (Umbellifers)	Spalik, Krzysztof	Recurrent short-distance dispersal explains wide distributions of hydrophytic umbellifers (Apiaceae tribe Oenantheae)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

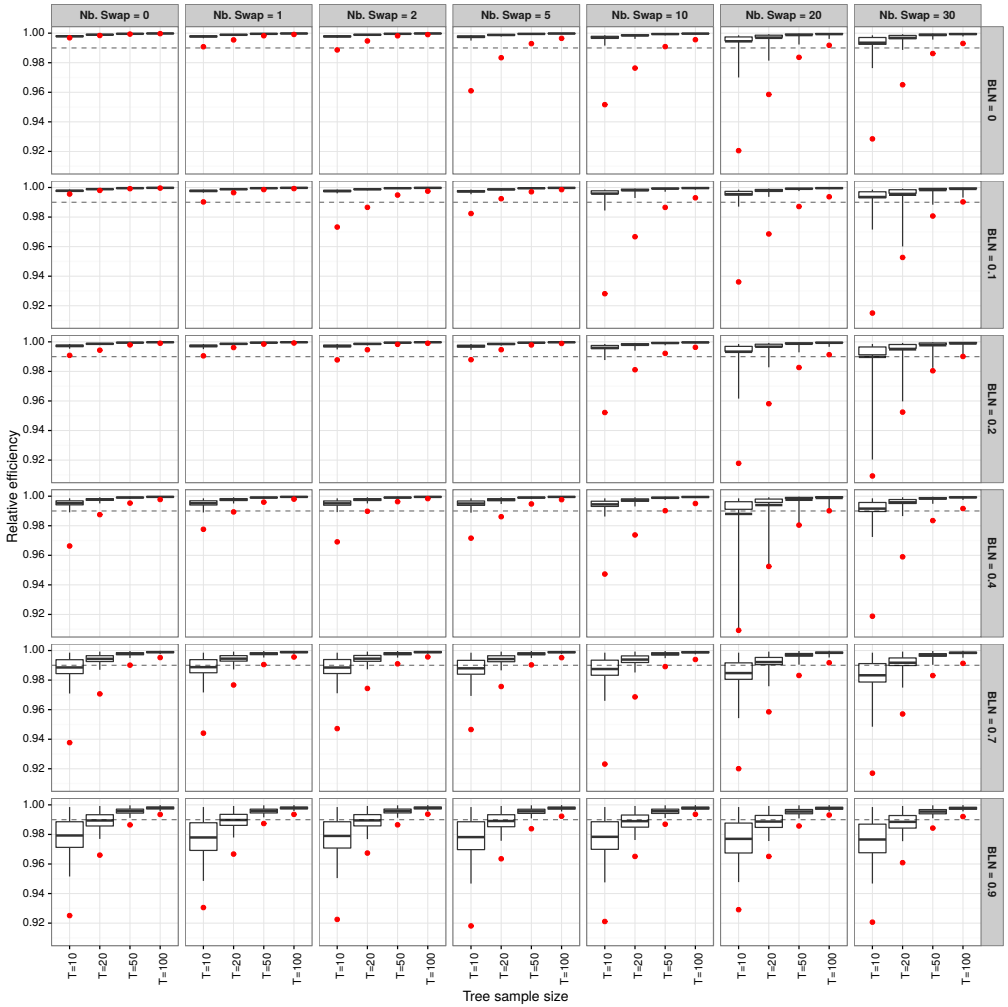
PeerJ Preprints NOT PEER-REVIEWED

Figure 1



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

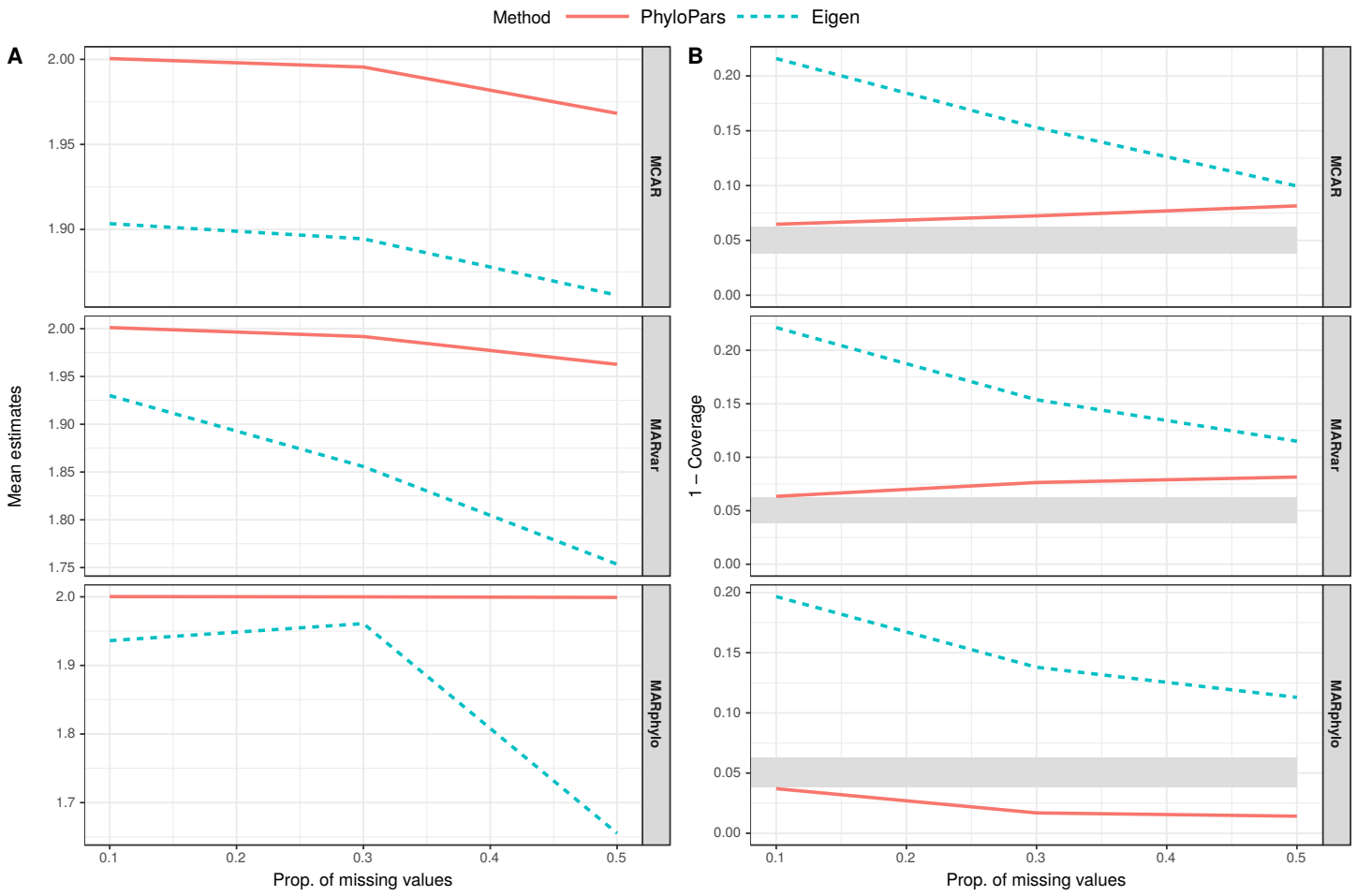
Figure 2



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

PeerJ Preprints NOT PEER-REVIEWED

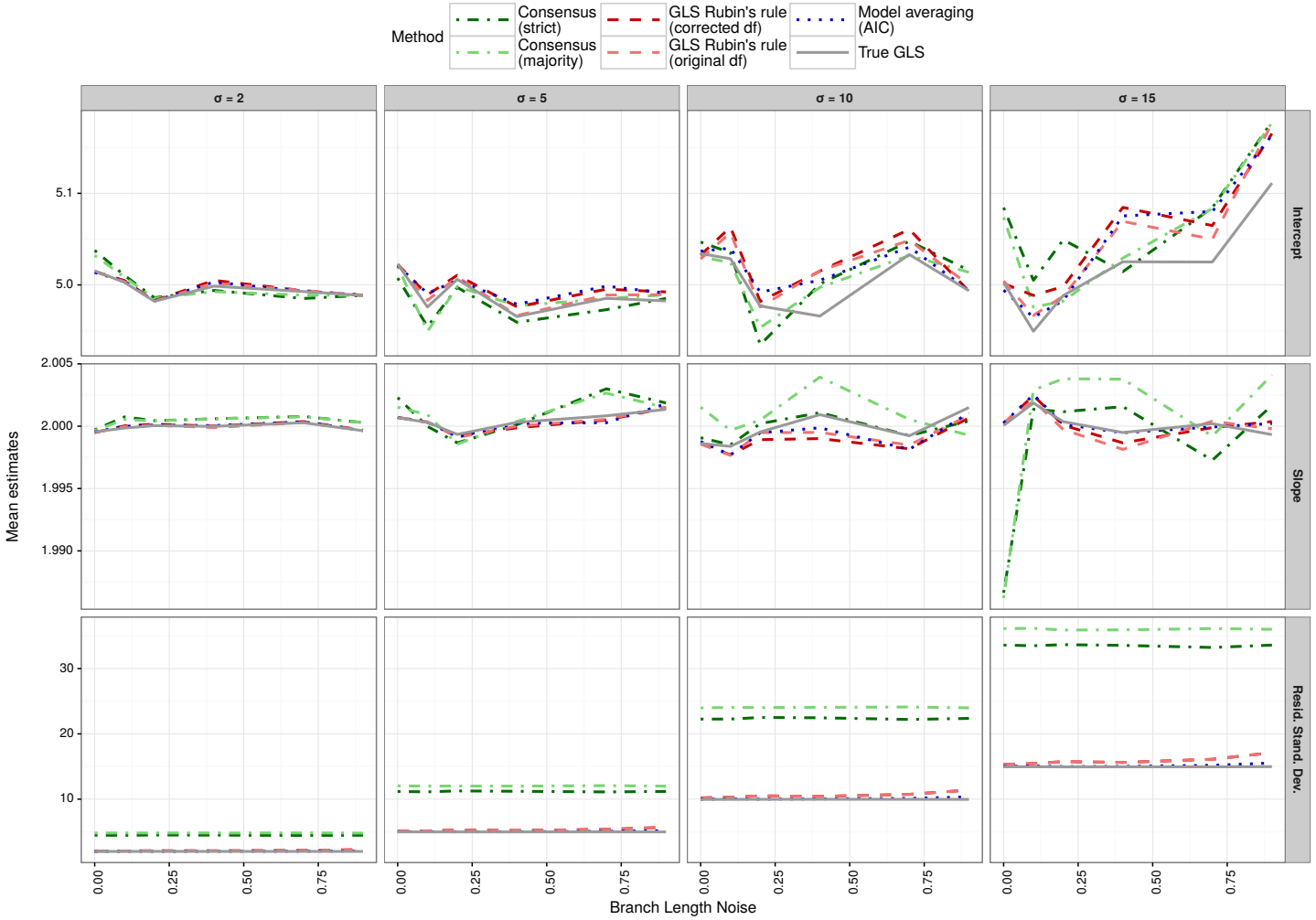
Figure 3



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

PeerJ Preprints NOT PEER-REVIEWED

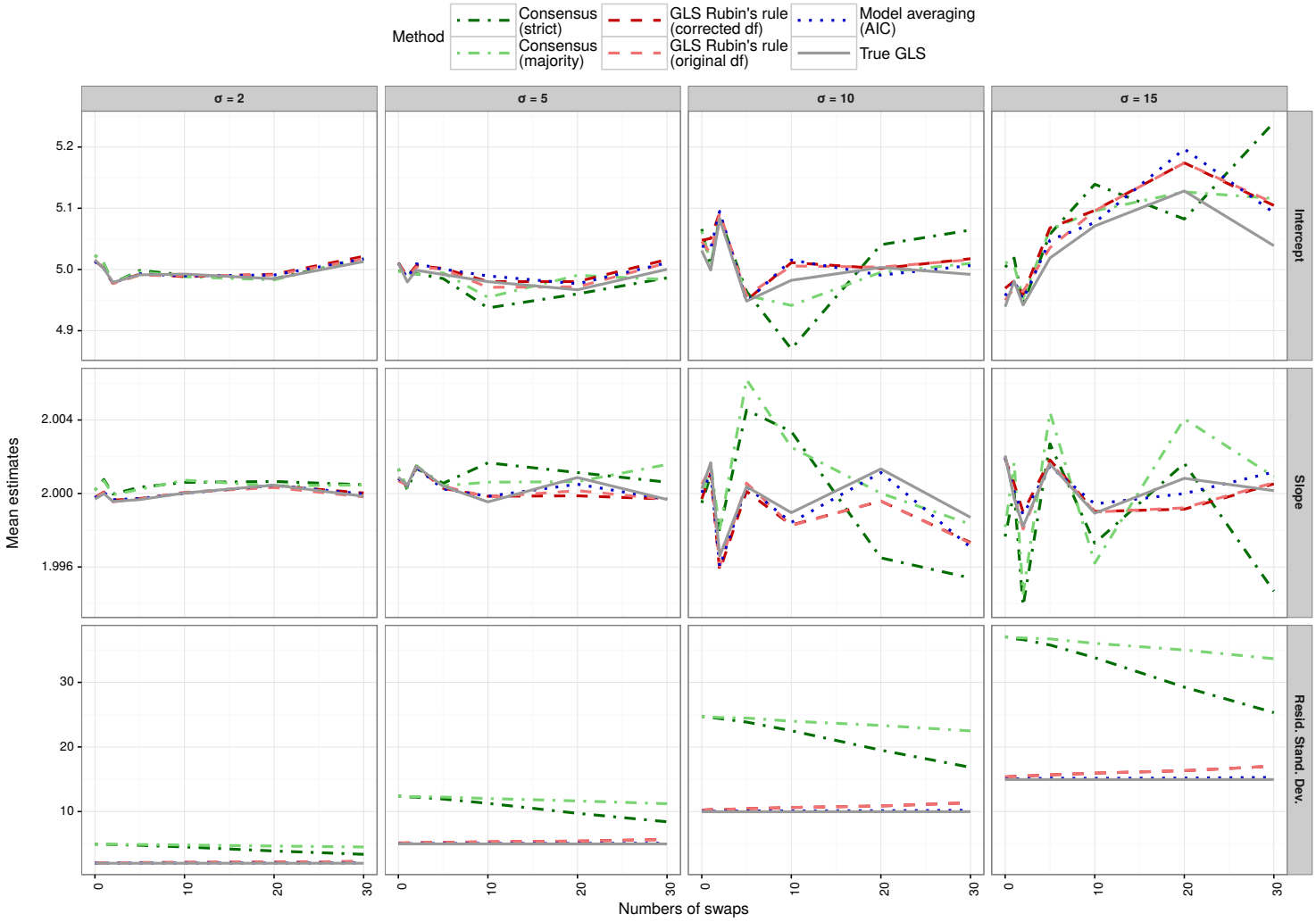
Figure S1



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

PeerJ Preprints NOT PEER-REVIEWED

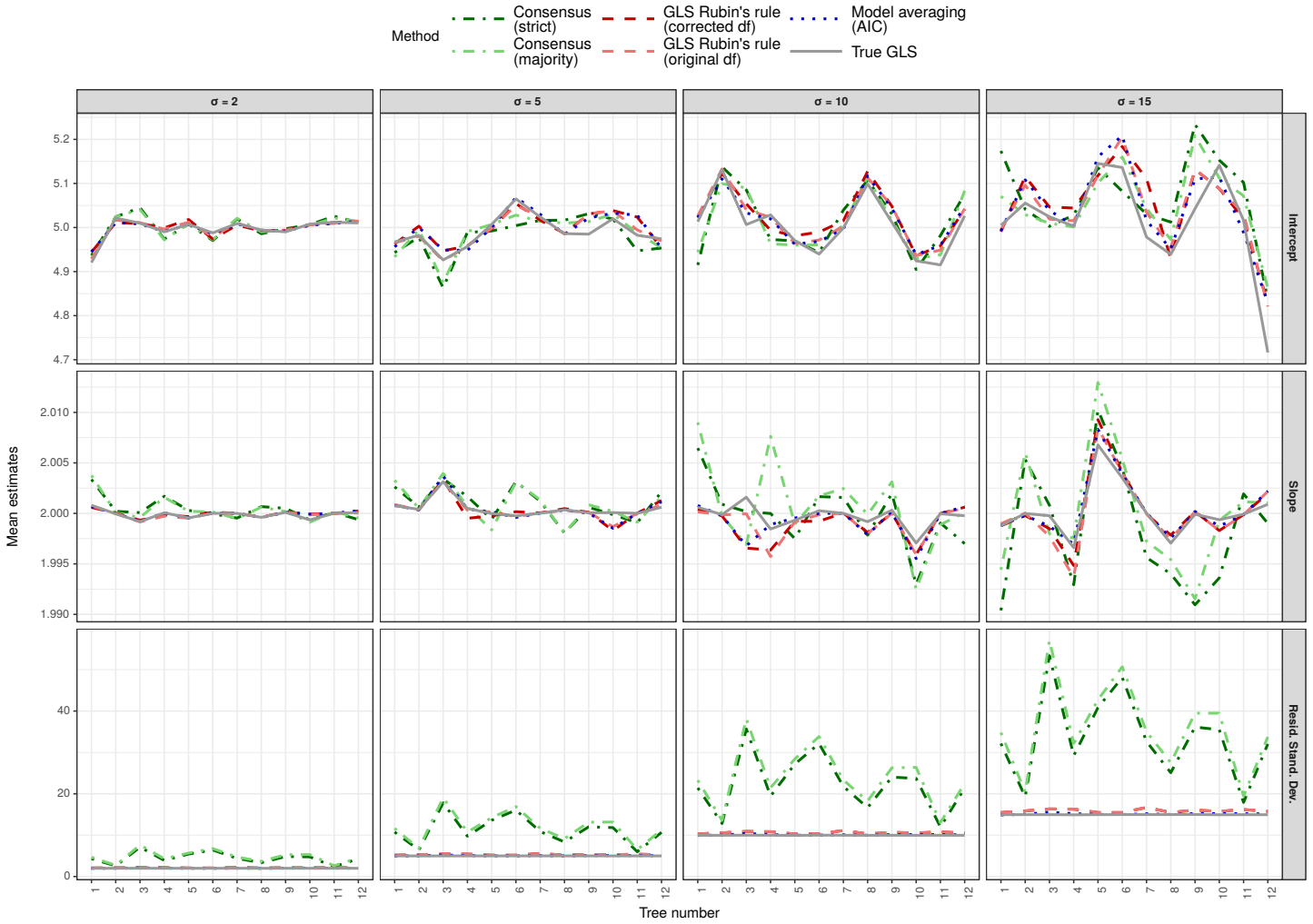
Figure S2



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

PeerJ Preprints NOT PEER-REVIEWED

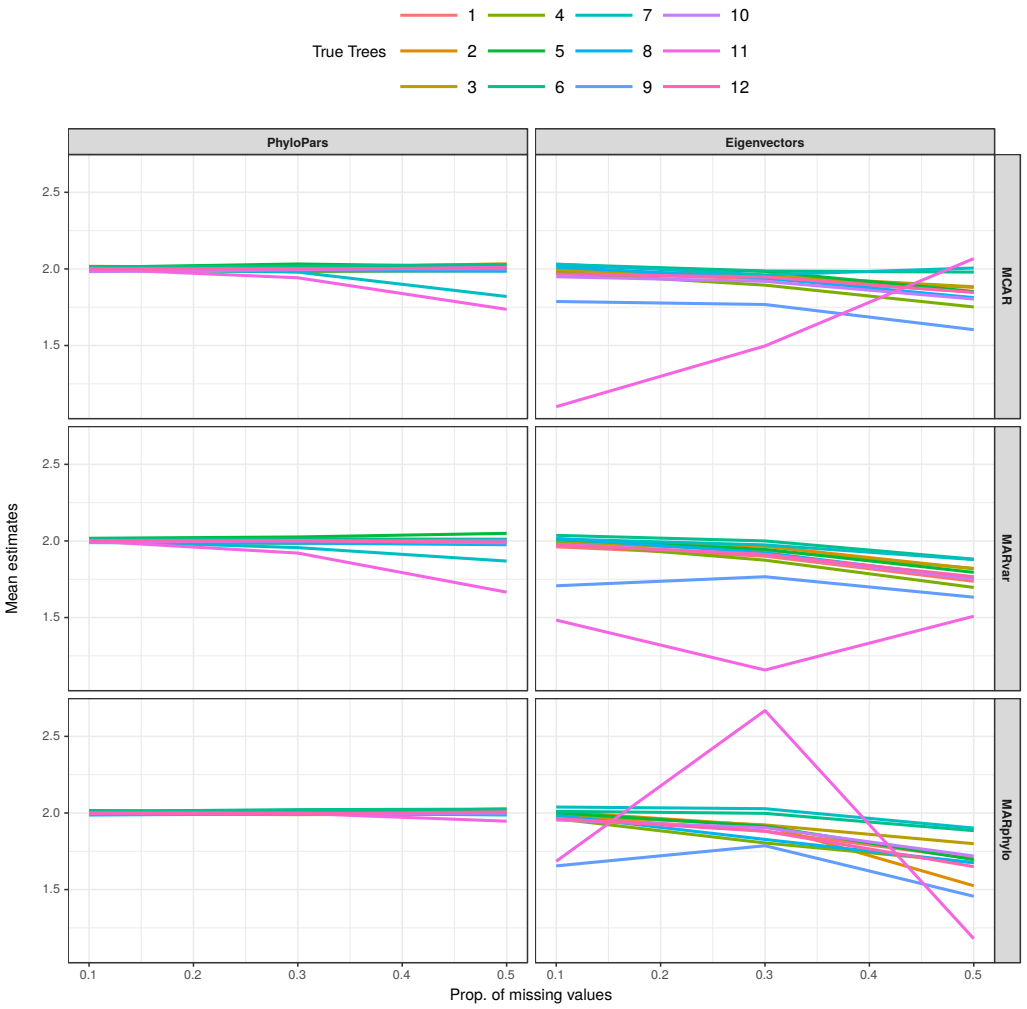
Figure S3



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018

PeerJ Preprints NOT PEER-REVIEWED

Figure S4



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1216v2 | CC BY 4.0 Open Access | rec: 14 Aug 2018, publ: 14 Aug 2018