COMMENT

On the accuracy and theoretical underpinnings of the multiple variance Brownian motion approach for estimating variable rates and inferring ancestral states

J. B. SMAERS[1]* and C. S. MONGLE[1,2]

[1]*Department of Anthropology, Stony Brook University, Circle Road, Stony Brook, NY 11794-4364, USA*
[2]*Interdepartmental Doctoral Program in Anthropological Sciences, Stony Brook University, Circle Road, Stony Brook, NY 11794-4364, USA*

In a recent contribution, Griffin & Yapuncich (2016) ('G&Y') report on purported inaccuracies and lack of theoretical underpinnings of our recently proposed ancestral estimation procedure. Our method is designed to provide a good overall fit with the data when different branches may be subject to different rates of change (multiple variance Brownian motion, or 'mvBM'). Here we demonstrate that G&Y's theoretical concerns stem from a misinformed account of basic statistical concepts and procedures, a misinterpretation of the primary literature and a circular adherence to a restrictive model of evolution (standard Brownian motion, or 'standard BM') whose usefulness has long been considered inappropriate for modelling branch-specific evolutionary patterns. We further apply a series of simple tests that falsify G&Y's claims on every account. Finally, we show that including a range of sample sizes (from 4 through 100) to G&Y's own suggested simulation procedure further substantiates what they purport to falsify: the validity of mvBM when modelling potential deviations from standard BM in trait evolution.

KEYWORDS: ancestral estimation, rate estimation, phylogenetic comparative method.

BACKGROUND

In the last 30 years, phylogenetic comparative methods have come to dominate the field of comparative biology. These methods combine a phylogeny with observed data points in order to explore a range of different questions regarding to the evolution of traits. Throughout this period of methodological development, the type of methods being proposed has diversified to accommodate different types of questions. Initially, the focus was on estimating co-evolution of traits by accounting for phylogenetic covariance (Felsenstein, 1985; Grafen, 1989). Gradually, methods have expanded towards trying to answer more explicitly evolutionary questions: What is the rate of evolution in particular traits? Are rates of evolution similar across different lineages? Can shifts in trait values be observed? If so, where and when in phylogenetic space did such shifts occur?

As questions have become more specific, the evolutionary models that support these methods have become more complex. Whereas the initial focus was on the standard Brownian motion ('BM') model, most researchers recognized that despite its advantages for hypothesis testing, its assumptions are unrealistic for the purposes of explicitly modelling evolutionary patterns. This notion was aptly summarized by Harvey & Purvis (1991): '*no evolutionary biologist actually believes that Brownian motion provides a realistic model of evolution*' (p. 623).

In search of more realistic models than the standard BM model, some have proposed to relax the constant variance assumptions of standard BM so as to allow for changes in the BM rate parameter (σ^2) (O'Meara *et al.*, 2006; Venditti, Meade & Pagel, 2011). Others have proposed the use of Ornstein-Uhlenbeck ('OU') models in order to increase the complexity of the standard BM model by adding parameters to the estimation procedure that reflect mean phenotype value and selection strength

*Corresponding author. E-mail: jeroen.smaers@stonybrook.edu

(Hansen, 1997). Butler & King (2004) subsequently put Hansen's method into an information-theoretic framework that allows for hypothesis testing among alternative evolutionary scenarios. Recently, the OU modelling framework has further expanded towards estimating shifts in trait values without *a priori* designation of selective regimes (Ingram & Mahler, 2013; Ho & Ané, 2014; Uyeda & Harmon, 2014; Khabbazian *et al.*, 2016).

These methods are now the consensus approach in phylogenetic comparative biology. This does not mean, however, that these methods are infallible. Several authors have cautioned that despite the considerable power of these approaches (Cressler, Butler & King, 2015), there remains uncertainty over the fundamental assumptions they make about the evolutionary process (e.g. Pennell *et al.*, 2015; Pyron, 2015). Because all these comparative methods rely on specific models of evolution, it is crucial to assess the degree to which the parameters assumed by the model accurately estimate patterns of evolutionary change and whether the actual patterns of change are expected to align with the assumptions of the model. Accordingly, several authors have highlighted incompatibilities between model parameters and the evolutionary concepts they purport to represent (Pennell & Harmon, 2013) and inherent limitations in both BM and OU models to accurately reflect evolutionary trait changes (Harmon *et al.*, 2010; Pennell *et al.*, 2015). This has led some to conclude that '*In general, available models such as BM and OU appear to be relatively inadequate and (reveal) poor absolute fits to real clades when examined from an empirical perspective*' (Pyron, 2015: 386).

In this light, several authors have reiterated the usefulness of more flexible models of evolution that account for the possibility of rate-heterogeneity (e.g. most recently Chira & Thomas, 2016). It is also in this context that Smaers, Mongle & Kandler (2016) proposed an exploratory method of ancestral estimation that accounts for different rates of change along individual lineages. The aim of this method is to adequately capture deviations from standard BM in trait evolution when the evolutionary assumptions include the occurrence of different rates along different branches of the tree. We emphasize here the importance of appropriately contextualizing the different contributions of different methods within the expanding toolbox of phylogenetic comparative methods. Ignoring such context may lead to misrepresentation and misinterpretation of methods, which, as we demonstrate, is the case for G&Y's criticism on mvBM.

The mvBM method proposed by Smaers *et al.* (2016) consists of two separate steps. A first step rescales the branches of the phylogeny according to estimates of lineage-specific trait changes, while a second step parameterizes this rescaled tree using standard BM. *The mvBM procedure is thus a two-step procedure* (for a step-by-step description of the algorithm, see Smaers *et al.*, 2016).

The first step serves to change the weight with which particular tips and branches contribute to node estimation. Because branch length and trait variance in the possible outcomes of trait evolution have a linear relationship in a BM model (Felsenstein, 1973), the first step provides an appropriate springboard to parameterize in a second step a BM model that includes different estimated rates for different lineages (i.e. a multiple variance BM model, or 'mvBM'). In mvBM, this adjustment of lineage-specific rates is accomplished by correcting for the baseline assumption that phylogenetic relatedness is an accurate proxy of trait change over time. This correction is achieved by leveraging a global estimate of nodal values with a local estimate. The global estimation assumes that all trait change can be accurately proxied by phylogenetic distance. This assumption is equivalent to the assumption that trait change is proportional to (the square root of) time which directly results from assuming a standard BM model. The local estimate assumes that components other than phylogenetic relatedness may influence trait change and that the effect of such local components is best approximated by observed trait changes in closely related lineages. The procedure to leverage global with local components hereby introduces a correction factor that estimates the extent to which branch-specific patterns of trait change deviate from a baseline BM assumption.

The mvBM procedure deviates from traditional phylogenetic comparative approaches in that it does not derive an algorithm from a statistical model, but rather comprises a heuristic algorithm that is based on the reasonable idea that both global and local factors influence trait diversity. The first step thus comprises a heuristic algorithm that rescales the branches of the phylogeny according to estimated trait change. The second step parameterizes this rescaled tree in a formal BM framework. Smaers *et al.* (2016) validated this approach in a simulation, the results of which confirmed the validity of mvBM by demonstrating that mvBM produces results equivalent to BM under BM conditions and outperforms BM when bursts are incorporated by identifying the location of the burst and reducing estimation error at this location.

The criticisms by G&Y on this method are based on several misreadings of the original work. First, G&Y criticize mvBM as inappropriate for confirmatory hypothesis testing purposes. The original work, however, explicitly states that mvBM is an ancestral *estimation* procedure that is suitable for exploring evolutionary hypotheses, not for testing them. The original work argues that such data explorations have become a crucial step in most research designs concerned with inferring evolutionary patterns based on a phylogeny and observed data, as evidenced in the widespread use of exploratory visualization techniques such as phylomorphospaces and phenograms. Second, G&Y criticize the fact that mvBM is

not based on a statistical model. It should, however, be clear that there is no intrinsic need for ancestral estimation algorithms to be derived from a statistical model. Any model (statistical or not) is a proxy of how evolution is assumed to work. From this perspective, all models are wrong and one can certainly dispute the validity of any model. The much more crucial question, ignored by G&Y, is whether an approach validly describes what it is proposed to describe. Third, G&Y do not acknowledge that the central aim of the mvBM procedure is to identify and accommodate regions in phylogenetic space where trait evolution *deviates* from standard BM when bursts are assumed to underpin trait diversification. Not recognizing this central feature leads G&Y to misrepresent how mvBM is calculated and posit untenable claims as to how it should be evaluated (e.g. that it is inappropriate and uninformative to compare mvBM to standard BM under any circumstance).

Here we provide detailed comments on each of the criticisms put forth by G&Y. We find that G&Y's claims are unfounded and based on a series of misconceptions about basic statistical concepts (iterativity vs. circularity, exploratory v. confirmatory analyses, increased parameterization vs. overfitting, overall vs. relative goodness-of-fit), misrepresentations of the primary literature and ill-conceived simulations that actually substantiate what they purport to falsify. We demonstrate that a series of simple tests falsify G&Y's claims on each account. Lastly, we performed a series of new simulations that expand G&Y's own suggested simulation procedure towards including a wider range of simulation conditions and sample sizes. These new simulations provide further evidence that mvBM is indeed a useful approach when exploring trait evolution under the assumption that rates may be different across different branches of a phylogeny. Specifically, these new simulations further substantiate that irrespective of sample size, mvBM produces results equivalent to standard BM when trait evolution is simulated under BM. These results are fundamental because they validate the mvBM procedure. Moreover, these new simulations further substantiate that mvBM's accuracy advantage over standard BM when bursts are incorporated in the simulation procedure is due to the identification and accommodation of bursts of change in the ancestral estimation procedure. These results demonstrate that mvBM does what it is proposed to do: identify and accommodate bursts of change in ancestral estimation.

TECHNICAL COMMENTS ON G&Y'S CRITICISM

PURPORTED CIRCULARITY

Circular reasoning can be defined as a logical fallacy in which a supposition and conclusion directly depend on one another. G&Y argue that the mvBM procedure is circular because, according to G&Y, the 'starting and ending points of the algorithm are the same (i.e. estimated trait values at internal nodes)'. In other words, G&Y argue that the circular nature of the procedure stems from mvBM's assumption that branch-specific rates of change can be estimated by leveraging global and local information. By all accounts, the procedure to start with a baseline assumption and subsequently adjust estimates based on a more complex assumption is an iterative procedure, not a circular one. There is not a single set of steps in mvBM's procedure whose starting points and ending points are the same. G&Y hereby confuse circularity with iterativity. That is not to say that mvBM is a fully iterative procedure. As described in Smaers *et al.* (2016) and also above, mvBM is a two-step procedure.

The statement that mvBM is circular also reveals a lack of understanding of the consequences of circularity. If mvBM were indeed circular, it would not produce results equivalent to BM under BM conditions; rather, it would identify bursts where there are none and therefore produce more error than BM under BM conditions. G&Y fail to recognize that mvBM is validated by the fact that mvBM produces results equivalent to BM under BM conditions, and further fail to recognize that this result demonstrates that there is no circularity. This result was made clear in Smaers *et al.* (2016: fig. 3) and is further substantiated in an expanded set of simulations here (Figs 1a and 2a).

PURPORTED OVERFITTING AND MODELLING OF 'RANDOM NOISE'

G&Y argue that mvBM overfits the data based on their observation that the mvBM procedure produces a higher absolute fit between the data and the tree than standard BM when a standard BM model underlies the simulations. According to G&Y, 'the only way a model can fit data better than the generating model is by overfitting'. Unfortunately, this argument is based on misconceptions about the effect of increased parameterization. It should be obvious that a common consequence of statistical approaches that use nested models is that a more parameter-rich model has a higher likelihood (i.e. overall fit) than a simpler model with fewer parameters – even on data simulated under the simpler model. Furthermore, it should also be clear that a consequence of overparameterization is that data structure is identified where there is no structure. Applied to our approach, this would mean that mvBM would identify bursts where there are none. Again here, G&Y fail to understand the consequence of the fact that mvBM produces results equivalent to BM under BM conditions (Smaers *et al.*, 2016: fig. 3,
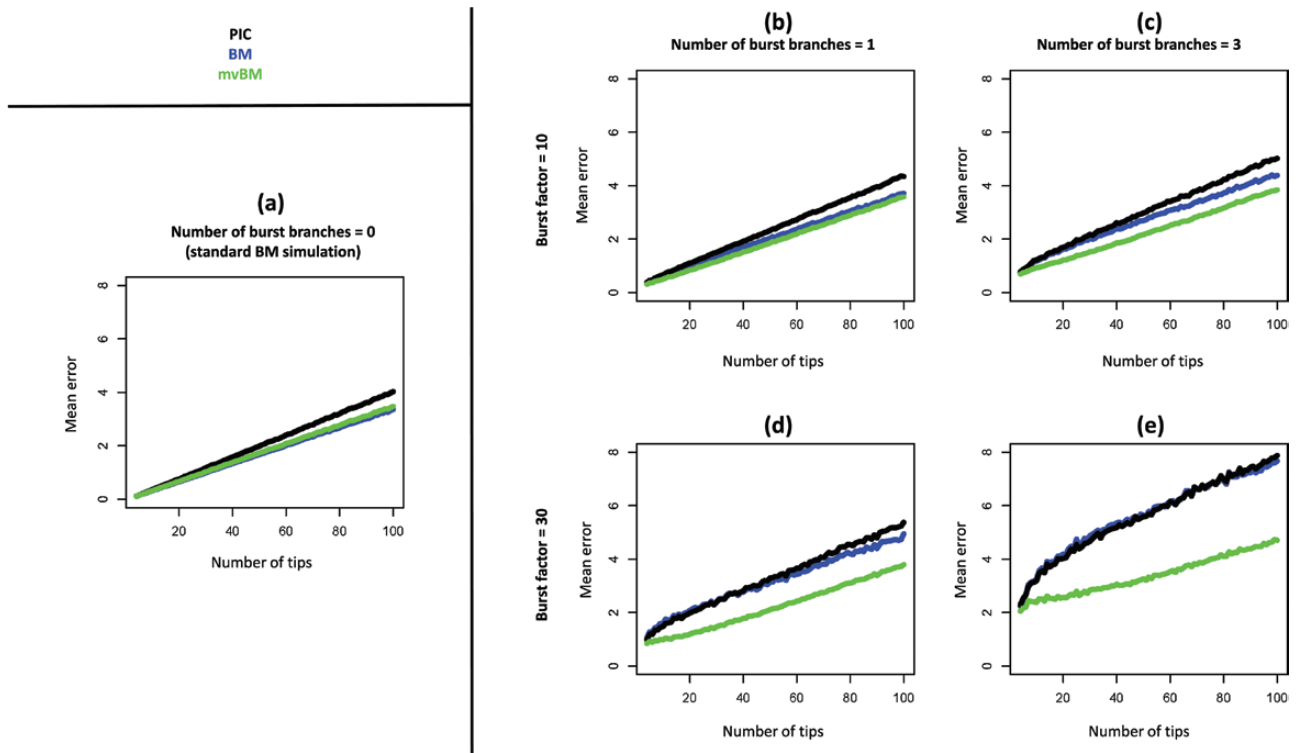
Figure 1. Results from simulations that expand G&Y's own suggested simulation procedure (randomizing bursts across the tree) by including more than a single phylogeny and a single set of tips. A range of different simulation conditions were used (described in detail in the main text) across sample sizes ranging from 4 through 100. For each simulation condition, one thousand trees and data sets were simulated for sample sizes ranging from 4 to 100 (1000 simulations for a 4-tip phylogeny, 1000 simulations for a 5-tip phylogeny, …, 1000 simulations for a 100-tip phylogeny). For each simulation, ancestral estimates were calculated using standard BM, mvBM and PICs. We note that the use of PICs for the purpose of ancestral estimation is inappropriate. PIC is not a method of ancestral estimation and produces valid estimates for the root only under strict standard BM conditions only. We use PICs here only to demonstrate that (counter to what G&Y claim) mvBM's accuracy advantage is not due to similarities with PICs. Results demonstrate that, expanding G&Y's own simulation procedure to more appropriate conditions and sample sizes substantiates what they purport to falsify: mvBM successfully reduces estimation error when bursts of change are assumed to underpin trait diversification.

and Figs 1a and 2a here). This validates the mvBM procedure by showing that it does not detect bursts where there are none, and provides evidence that the mvBM procedure does not overfit. It is correct that high parameterization *may* result in overfitting, but our simulations demonstrate that this is not the case for our approach because mvBM does not detect bursts where there are none and validly reduces the error at locations where there is indeed a burst.

In this context, G&Y further claim that mvBM's 'transformed phylogeny largely captures random noise' and therefore that it 'does not provide meaningful insight about the data'. G&Y, however, again do not provide any test to back up their claim, nor do they explain how they define 'random noise', much less specify what is meant by 'largely'. Moreover, this argument is again the result of G&Y's failure to understand the consequence of the fact that mvBM produces results

equivalent to BM under BM conditions, which nullifies the argument that mvBM models random noise. If mvBM would indeed model random noise, it would detect bursts where there are none. As mentioned, both the original work by Smaers *et al.* (2016: fig. 3) and an expanded set of simulations provided here demonstrate that this is not the case (Figs 1a and 2a). It should thus be clear that these unsupported claims are untenable and constitute a misrepresentation of the original work.

Furthermore, a logical consideration of the nature of their claim that mvBM models random noise also reveals that it is untenable. The mvBM rescaled phylogeny successfully captures deviations from a baseline BM assumption (as evidenced in the fact that mvBM produces equivalent results as standard BM when data is simulated according to BM and outperforms BM when bursts are included in the simulation procedure). Therefore, by suggesting that mvBM's
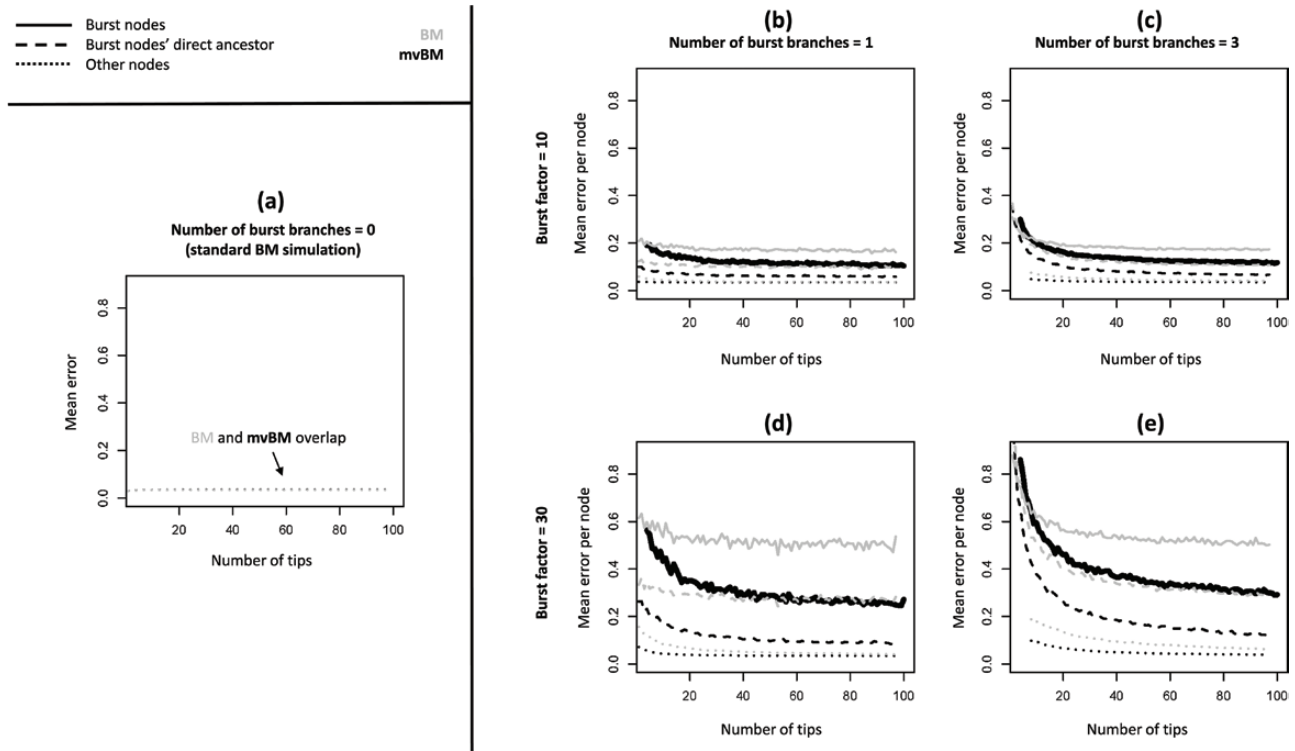
Figure 2. Using the same simulations as presented in Figure 1, separate results on mean error per node are presented for the burst node, the burst nodes' direct ancestor and other nodes. These results demonstrate that mvBM's accuracy advantage is due to improving the identification and accommodation of bursts of change in the ancestral estimation procedure.

rescaled phylogeny captures random noise, G&Y characterize deviations from BM as random noise. By defining noise as what is or is not captured by a model (in this case BM), the authors fall into the circular fallacy that meaningful evolutionary signal is measured by the model, and what the model measures is meaningful signal. It is ironic that G&Y falsely claim circularity by making circular arguments. Such circular argumentation reveals an inappropriate implicit adherence to a model of evolution (standard BM) that has long been considered as inappropriate for modelling lineage-specific patterns of trait change (e.g. Harvey & Purvis, 1991).

PURPORTED STATISTICAL NON-INTERPRETABILITY

G&Y further claim that because the mvBM procedure does not account for the number of parameters it assumes, it cannot be used for statistical inference. This argument confounds the fundamental difference between overall and relative goodness-of-fit and how these measures relate to exploratory estimation and confirmatory hypothesis testing. Overall fit measures the fit of the data to the model. Relative fit corrects the overall fit relative to the number of assumed parameters (e.g. the use of information criteria such as AIC).

Measures of overall fit can be used to provide exploratory best-*estimates* of data structure. Measures of relative fit should be used in the context of confirmatory hypothesis *testing*. The crucial misconception by G&Y is that any ancestral estimation method that provides explicit estimates for all nodal values in a tree would be appropriate for confirmatory hypothesis testing. It should be self-evident that *estimates* of ancestral values are *estimates*, not observations, and therefore are not suitable for hypothesis *testing*. This does not, however, mean that these ancestral estimation methods are not useful. Such methods are clearly useful for the purposes of exploring a description of the data that comprises the highest possible fit between data and model (i.e. high overall fit).

In this context, G&Y also claim that the mvBM procedure does not represent a recognizable statistical model. As mentioned, mvBM's first step uses a heuristic algorithm that rescales the tree according to estimates of lineage-specific trait changes, while a second step parameterizes this rescaled tree using standard BM. There is no intrinsic need for the first step to be derived from a statistical model. In fact, mvBM's procedure is specifically designed to be different from traditional statistical models for reasons discussed at length in Smaers *et al*. (2016) and concisely reviewed

again here in the Background section. Importantly, by parameterizing mvBM's rescaled tree with standard BM parameters in a second step, the eventual output of mvBM is undeniably a statistical model.

G&Y also argue that 'mvBM represents a major departure from basic principles underlying modern PCMs (phylogenetic comparative methods)'. Our procedure is indeed intended to be a departure from current PCMs, but G&Y's argument fails to acknowledge that the idea about how evolution is expected to produce diversity that underlies mvBM is consistent with that which underlies more traditional theoretical models of evolution. Specifically, the idea that both global and local components influence trait diversification is similar to the idea that underlies a phylogenetic mixed effects model and an OU model. Phylogenetic mixed effect models (Lynch, 1991; Housworth, Martins & Lynch, 2004) assume that evolutionary variation comprises both a phylogenetic component ('heritable' or 'additive' components) and a component specific to the nodes in question ('residual', or 'nonheritable' components). The OU model assumes that populations climb *local* adaptive peaks, effectively assuming that *local* selective pressures direct trait space towards certain peaks (Hansen, 1997). It should be clear that the mvBM approach proposed by Smaers *et al.* (2016) is not a phylogenetic mixed effect model, nor an OU model, but it does share the central idea that trait evolution is affected by both global and local components.

PURPORTED INACCURACY AND BIAS

G&Y argue that mvBM contains 'inaccuracies' and 'bias' in its ancestral estimates. They supported their argument with three simulations. Their first simulation comprised a single 3-tip phylogeny comparing the accuracy of standard BM, mvBM and PIC ('phylogenetic independent contrasts'). A second simulation used a single 100-tip phylogeny and considers only mvBM. A third simulation used a single 30-tip phylogeny and also considers only mvBM.

G&Y's first simulation comprised a single 3-tip phylogeny and a burst of evolution along one branch. This simulation is purported to show that mvBM's accuracy stems from similarities with PICs. Their simulation is problematic in a number of ways. First, G&Y misrepresent mvBM's central aim of capturing *deviations* from a global estimate (the calculation of PICs has some similarities with mvBM's global estimates in this respect). Because mvBM's global estimate reflects the standard BM assumption that trait change is proportional to phylogenetic distance, the similarities between PICs and mvBM's global estimate are expected because standardized PICs can and frequently are used to parameterize a

constant-rate BM model. G&Y, however, do not recognize that in mvBM the global estimate must be adjusted using a local estimate before calculating rates in order to fulfil mvBM's assumption that a constant-rate BM may be wrong. The claim that mvBM's global estimate is representative of its ancestral estimates (and therefore representative of PIC ancestral estimates) is therefore false and constitutes a clear misrepresentation of the original work. To exemplify how mvBM ancestral estimation differs from PIC ancestral estimation, we provide a new set of simulations that compare the accuracy of standard BM, PICs and mvBM across a wide range of samples (4 through 100). These simulations unequivocally demonstrate that PICs produce by far the least accurate ancestral estimations, even when data are simulated according to standard BM (Figs 1a–e; more details below). Furthermore, in designing their first simulation, G&Y do not acknowledge that a PIC analysis does not produce ancestral estimations because the contrasts have no natural 'root' that represents an ancestral condition. PIC is simply not a method of ancestral estimation. Rather, PIC is a method for removing redundancies due to shared ancestry for the purposes of trait correlation analysis or rate analysis. The *only* value that can be considered to be correctly estimated using PICs is the root value, which is equal to the estimate found under a standard constant-rate BM model. *No other* PIC ancestral estimate besides the root can be considered to be a correct estimate under a standard constant-rate BM model. Our simulations exemplify this property by demonstrating that PIC ancestral estimates indicate more error than BM under BM conditions (thus invalidating their use). It is, of course, possible to repeatedly reroot the tree to obtain proper estimates of each of the nodes using PIC algorithms (Garland & Ives, 2000), but this produces the same results as the generalized least-squares ('GLS') procedure. The GLS procedure is preferred over repeatedly rerooting the tree because GLS allows for a wider application (Smaers & Rohlf, 2016). Also, the notion that a 3-tip phylogeny provides comprehensive information about the accuracy of a multirate method is unreasonable. In order to reliably estimate parameters, all models require enough information. A 3-tip phylogeny clearly does not provide this. Finally, and most paradoxically, G&Y's results in this narrow simulation procedure actually substantiate what they purport to falsify: mvBM's accuracy relative to standard BM.

In a second simulation, G&Y used a single 100-tip phylogeny and vary the location of the burst branch across all branches 1000 times. G&Y claim that this simulation exemplified mvBM's 'bias' and 'inaccuracy'. The 'bias' G&Y refer to relates to the observation that

'reconstructions at the base of the burst branch are strongly biased in the direction of the burst, and the degree of bias depends on the size of the burst'. Again, G&Y do not test their claim that the degree of autocorrelation prohibits accurate detection of trait shifts, but rather describes the error as 'quite large'. G&Y do not acknowledge that Smaers *et al.* (2016: 88) accurately reported that, per construction, mvBM exhibits a certain degree of autocorrelation. Moreover, Smaers *et al.* (2016) provide clear results indicating that the degree of autocorrelation is in fact not prohibitive of accurately detecting bursts. G&Y's further assessment that mvBM error is 'quite large' is based only on the observation that mvBM's estimates do in fact show error. G&Y do not make any comparison with any other method, nor with a baseline expectation in order to assess whether mvBM in fact shows less error than other methods and is therefore more useful. The claim that statistical estimation is wrong because it contains error is naive at best. It should be self-evident that *any* ancestral estimation method produces error. The much more important question ignored by G&Y is whether or not the amount of error in mvBM's estimates renders it a useful approach. Smaers *et al.* (2016) provide ample evidence that mvBM is, in fact, useful. These results are further substantiated here in a comprehensive set of simulations presented in Figs 1–2.

In a third simulation, G&Y used a single 30-tip phylogeny to demonstrate that the R^2 values (from a regression of estimated to observed ancestral values drawn from a series of simulations) reported in Smaers *et al.* (2016) are inflated. G&Y used a different procedure for simulating bursts across the tree (randomly across all branches) than the one used by Smaers *et al.* (2016) (who used fixed locations). Following the different simulation procedure, G&Y found a different R^2 value than the one reported by Smaers *et al.* (2016) [one that is lower: median of 0.81 while a mean of ~0.97 was reported in Smaers *et al.* (2016)]. G&Y used their result to argue that this constitutes evidence of mvBM's 'inaccuracy'. G&Y here do not seem to realize that changing the procedure of any simulation will lead to different results for any method. Although this is of course self-evident, it is nonsensical to use this attribute to claim that mvBM is 'inaccurate'. G&Y further do not acknowledge that Smaers *et al.* (2016) did not claim that the reported R^2 values apply to all possible simulation conditions. Rather, Smaers *et al.* (2016) argue that mvBM's reported R^2 values are *higher* than those observed for standard BM when bursts are simulated and equivalent to standard BM when simulations are according to BM, something which G&Y do not test. G&Y's argumentation for systematic inaccuracy based on the observation that different results are found when using different simulation procedures is nonsensical.

PURPORTED NON-INFORMATIVE NATURE OF THE COMPARISON BETWEEN MVBM AND STANDARD BM

Likely because G&Y do not acknowledge that mvBM's central aim is to identify and accommodate *deviations* from standard BM, they make the untenable claim that it is inappropriate to compare mvBM's accuracy to that of standard BM. G&Y make the correct statement that standard BM is appropriate under conditions that do not violate its assumptions, but that BM is inappropriate when its assumptions are violated. G&Y, however, go on to argue that this renders standard BM invalid as a baseline comparison to test whether mvBM does what it is proposed to do. G&Y thereby do not acknowledge that the central feature of mvBM is to identify where in the phylogeny standard BM does not fit when bursts do occur and to provide an accommodation of such instances when estimating ancestral values. In comparing mvBM to standard BM results, Smaers *et al.* (2016) do not set out to show that standard BM is not a good method, nor that it does not produce good estimates when its assumptions are violated. Rather, Smaers *et al.* (2016) set out to demonstrate that mvBM produces ancestral estimates equivalent to standard BM when trait change occurs according to BM, and produces more accurate results than standard BM at those phylogenetic locations where bursts occur and standard BM does not fit the model assumptions. Testing these proposals is crucial, because if mvBM would not produce equivalent results as standard BM when data are simulated under standard BM, and not outperform standard BM when it is not, then mvBM would indeed model noise and therefore not provide insightful information about the data. But it does do what it is proposed to do, so it does not model noise, and it does provide insightful information about the data. The notion that it is non-informative and inappropriate to test that mvBM performs as it is proposed to perform is absurd.

VERIFIABLY FALSE AND MISLEADING STATEMENTS BY G&Y

One of G&Y's central claims is that Smaers *et al.* (2016) only report 'aggregated results across the entire tree' in order to hide purported bias and inaccuracy in mvBM's estimates. However, G&Y do not acknowledge Table 2 in Smaers *et al.* (2016: 86) in which separate accuracy measures are reported for all nodes, for the burst node, for nodes closely related to the burst node, for nodes distantly related to the burst node and for the root. These results clearly show that mvBM's accuracy advantage relative to BM is highest around the burst node. Moreover, these results demonstrate that mvBM's accuracy advantage for the burst node

increases with the size of the burst and with the number of simulated bursts. These results clearly falsify G&Y's claims of inaccuracy and bias. Using the purported absence of Table 2 to fuel false suggestions of inaccuracy and bias is misleading. In an expanded set of new simulations, here we again provide node-specific results that again confirm that mvBM's accuracy advantage is due to improved identification and accommodation of bursts of change when estimating ancestral values (Figs 2b–e; details below).

G&Y further claim that 'Smaers *et al*. suggest their method relates to existing statistical models of evolution, particularly the multiple-rate BM model of Venditti *et al*.'. Such a suggestion was never made. The similarities and differences with Venditti *et al*.'s approach are more nuanced than G&Y imply. The underlying idea about how evolution is expected to produce trait diversity (i.e. deviations from constant-rate BM) is indeed similar between these approaches. However, the manner in which these ideas are formalized is very different (through a heuristic algorithm in mvBM vs. by deriving an algorithm from a statistical model by Venditti *et al*.). Lastly, both approaches constitute two procedural steps. A first step rescales the phylogeny, and a second parameterizes the rescaled phylogeny using a standard BM. G&Y's claim thus confounds the fundamental difference between a method's underlying idea, its formalization and its procedure.

G&Y also often refer to a previous contribution of theirs (Griffin & Yapuncich, 2015) in which they make very similar statements as reported in their current criticism, but related to a previous method proposed by Smaers (Smaers & Vinicius, 2009). However, the same theoretical misconceptions and fallacies reported here underlie their previous criticism. Using these arguments to back up their criticism of Smaers *et al*. (2016) is therefore misleading.

<div align="center">

EXPANDING G&Y'S OWN SIMULATION
PROCEDURE SUBSTANTIATES MVBM'S VALIDITY

</div>

G&Y claimed that the simulation conditions reported in Smaers *et al*. (2016) were favourable to mvBM. Here we extend G&Y's own suggested simulation procedure (i.e. randomly varying bursts across the tree) towards including more than a single phylogeny and a single sample size. Specifically, we simulated 1000 trees and data sets for phylogenies with a number of tips ranging from 4 to 100 (1000 simulations for a 4-tip phylogeny, 1000 simulations for a 5-tip phylogeny, …, 1000 simulations for a 100-tip phylogeny). As suggested by G&Y, bursts were assigned randomly across all branches of the tree. To explore mvBM's accuracy relative to the number of bursts, we ran separate simulations using a different set number of bursts (1 or 3 burst branches,

Figs 1b–e, 2b–e). To explore the effect of the size of the burst, we further ran separate simulations in which bursts were defined as a factor increase (either 10 or 30, Figs 1b–e, 2b–e) relative to a baseline rate of 0.01.

Results indicate that irrespective of sample size, mvBM produces equivalent results to standard BM when data are simulated according to standard BM [Figs 1a and 2a, see also Fig. 3 in Smaers *et al*. (2016)]. When bursts of change are assumed to underpin trait diversification, results confirm that mvBM's accuracy advantage relative to standard BM is irrespective of sample size and increases with the number of bursts and the strength of the bursts (Figs 1b–e). When separating the mean error per node for the burst nodes, the burst nodes' direct ancestor, and other nodes, results demonstrate that mvBM's accuracy advantage over standard BM is due to the identification and accommodation of bursts of change in the ancestral estimation procedure (Fig. 2b–e). It is evident that for both mvBM and standard BM, the mean error at the burst node and its direct ancestor are larger than in other nodes. This demonstrates that bursts indeed echo error down the tree to a certain extent [as accurately reported in Smaers *et al*. (2016)]. However, and most crucially, for both the burst node and its direct ancestor, mvBM's mean error is significantly reduced. This demonstrates that mvBM successfully reduces the error caused by bursts of change. Moreover, these results also demonstrate that mvBM successfully reduces the bursts' echoing effect down the tree. In other words, results unequivocally demonstrate that mvBM does precisely what it is proposed to do: identify and accommodate bursts of change in ancestral estimation.

These results provide further confirmation of the accuracy advantage of mvBM relative to standard BM and affirm its validity and usefulness when modelling lineage-specific patterns of change under the assumption that trait diversification may comprise different rates of evolution in different lineages. Results also demonstrate that including a range of different sample sizes (from 4 through 100) in G&Y's own suggested simulation procedure (randomizing bursts across the phylogeny) substantiates what they purport to falsify.

<div align="center">

CONCLUSION AND DISCUSSION

</div>

We conclude that G&Y's criticism is based on a series of unsupported claims that are based on fundamental misconceptions about circularity and overparameterization. G&Y misrepresent the original work by not recognizing several crucial features of the method that they criticize (e.g. rates are calculated only after adjusting a global estimate with a local estimate) and providing a series of misleading simulations that,

when expanded to a more reasonable range of sample sizes, actually support what they claim to falsify.

The only valid discussion point that could be drawn from G&Y's criticism relates to whether a statistical or a nonstatistical model is expected to better represent how evolution works. Although G&Y do not recognize that this discussion underlies their argumentation, G&Y clearly favour statistical models in this regard. This is not, however, a criticism of mvBM per se. It should be clear that any model (statistical or not) is a proxy of how evolution is assumed to work, and one can certainly have many philosophical discussions about what is the best assumption. But one should not confound philosophical assumptions with empirical validity. On this account, it is unequivocal that mvBM is an empirically valid procedure because, irrespective of sample size, it successfully identifies and accommodates bursts of change under assumptions of multirate trait diversification and produces equivalent results to standard BM under assumptions of constant rate trait diversification.

Another discussion point that G&Y touch upon, but only tangentially, is sample size. Unfortunately, also here their argument is untenable. G&Y suggest that mvBM's accuracy is lower with lower sample size and go on to suggest that the original work by Smaers *et al.* (2016) used a large tree in order to 'dilute the effect of the burst'. This serious accusation is, however, unfounded. The untenable nature of this accusation would have been apparent if only G&Y had done the appropriate tests to back it up. Our expanded set of simulations (which are no more than an expansion of G&Y's own suggested simulation procedure) explores the effect of sample size on mvBM estimation and the 'effect of the burst'. Results unequivocally demonstrate the validity of mvBM estimation irrespective of sample size (Fig. 2).

A further point of discussion that could be re-emphasized is mvBM's potential application in a phylogenetic comparative research design. G&Y seem to argue that mvBM is not a blanket solution and therefore must be wrong. It should be self-evident that *any* ancestral estimation procedure is not a blanket solution for all that is phylogenetic comparative methodology. But that does not imply that mvBM does not have use in contributing information about trait evolution. By providing results equivalent to BM under BM conditions and outperforming BM when bursts occur by reducing error at the location of the burst, the mvBM procedure is useful in that it provides a better estimate of trait evolution. For example, despite the common acknowledgement that BM is not up to the task when different lineages indicate different rates of evolution, phenograms and phylomorphospaces continue to be used in conjunction with constant-rate BM ancestral estimation. mvBM here provides an alternative that better identifies and accommodates bursts of change in trait evolution. mvBM can thus validly be used when the aim is to provide best-estimates of trait evolution. However, as with any other ancestral estimation approach, mvBM should be combined with other methods that are appropriate for hypothesis testing whenever the data allow. As mentioned in the Background section, several such methods are widely available (O'Meara *et al.*, 2006; Khabbazian *et al.*, 2016). We also hope that future work will continue to improve upon ancestral estimation procedures in order to further reduce estimation error and better accommodate for a wide range of possible patterns of trait diversification.

ACKNOWLEDGEMENTS

REFERENCES

Butler MA, King AA. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *American Naturalist* **164:** 683–695.

Chira AM, Thomas GH. 2016. The impact of rate heterogeneity on inference of phylogenetic models of trait evolution. *Journal of Evolutionary Biology*, **29**(12): 2502-2518s.

Cressler CE, Butler MA, King AA. 2015. Detecting adaptive evolution in phylogenetic comparative analysis using the Ornstein-Uhlenbeck model. *Systematic Biology* **64:** 953–968.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* **25:** 471–492.

Felsenstein J. 1985. Phylogenies and the comparative method. *American Naturalist* **125:** 1–15.

Garland Jr T, Ives AR. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *American Naturalist* **155:** 346–364.

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

Griffin RH, Yapuncich GS. 2015. The independent evolution method is not a viable phylogenetic comparative method. *PloS One* **10:** e0144147.

Griffin RH, Yapuncich GS. 2016. A critical comment on Smaers et al. (2016). *Biological Journal of the Linnean Society* **118:** 74–98.

Hansen TF. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* **51:**1341–1351.

Harmon LJ, Losos JB, Jonathan Davies T, Gillespie RG, Gittleman JL, Bryan Jennings W, Kozak KH, McPeek MA, Moreno-Roark F, Near TJ, Ricklefs RE, Schluter

D, Schulte JA, Seehausen II O, Sidlauskas BL, Torres-Carvajal O, Weir J T, Mooers AO. 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution; International Journal of Organic Evolution* **64:** 2385–2396.

Harvey PH, Purvis A. 1991. Comparative methods for explaining adaptations. *Nature* **351:** 619–624.

Ho LST, Ané C. 2014. Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution* **5:** 1133–1146.

Housworth EA, Martins EP, Lynch M. 2004. The phylogenetic mixed model. *American Naturalist* **163:** 84–96.

Ingram T, Mahler DL. 2013. SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. *Methods in Ecology and Evolution* **4:** 416–425.

Khabbazian M, Kriebel R, Rohe K, Ané C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. *Methods in Ecology and Evolution* **7:** 811–824.

Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution* **45:** 1065–1080.

O'Meara BC, Ané C, Sanderson MJ, Wainwright PC. 2006. Testing for different rates of continuous trait evolution using likelihood. *Evolution* **60:** 922–933.

Pennell, MW, FitzJohn RG, Cornwell WK, and Harmon LJ. 2015. Model adequacy and the macroevolution of angiosperm functional traits. *The American Naturalist* **186:** E33-E50.

Pennell MW, Harmon LJ. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences* **1289:** 90–105.

Pyron RA. 2015. Post-molecular systematics and the future of phylogenetics. *Trends in Ecology & Evolution* **30:** 384-389.

Smaers JB, Mongle CS, Kandler A. 2016. A multiple variance Brownian motion framework for the estimation of ancestral states and rates of evolution. *Biological Journal of the Linnean Society* **118:** 78–94.

Smaers JB, Rohlf FJ. 2016. Testing species' deviation from allometric predictions using the phylogenetic regression. *Evolution* **70:** 1145–1149.

Smaers JB, Vinicius L. 2009. Inferring macro-evolutionary patterns using an adaptive peak model of evolution. *Evolutionary Ecology Research* **11:** 991–1015.

Uyeda JC, Harmon LJ. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Systematic Biology* **63:** 902–918.

Venditti C, Meade A, Pagel M. 2011. Multiple routes to mammalian diversity. *Nature* **479:** 393–396.