Transcending data gaps: a framework to reduce inferential errors in ecological analyses

Abstract
The analysis of functional diversity (FD) has gained increasing importance due to its generality and utility in ecology. In particular, patterns in the spatial distribution and temporal change of FD are being used to predict locations and functional groups that are immediately vulnerable to global changes. A major impediment to the accurate measurement of FD is the pervasiveness of missing data in trait datasets. While such prevalent data gaps can engender misleading inferences in FD analyses, we currently lack any practical guide to handle missing data in trait datasets.

Here, we identify significant mismatches between true FD and values derived from datasets that contain missing data. We demonstrate that imputing missing data with a phylogeny-informed approach reduces the risk of misinterpretation of FD patterns, and provides baseline information against which central questions in ecology can be evaluated.

Keywords
Functional biogeography, functional diversity, functional trait, imputation, trait database.

INTRODUCTION
Variability in functional traits and their responses to environmental filters provide raw material for a variety of research areas from community ecology (McGill et al. 2006) and population ecology (Santini et al. 2013), to conservation (Cardillo et al. 2008). In particular, functional diversity (FD) is increasingly recognised as an important indicator of ecosystem functioning, and the spatial distribution of and temporal change in FD are frequently used to identify locations and functional groups that are vulnerable to global changes (Díaz & Cabido 2001; Bellwood et al. 2003; Díaz et al. 2007). A fundamental challenge in FD analyses is that compilation of trait information of numerous taxa on a large geographic scale requires prohibitively extensive fieldwork (Baraloto et al. 2010; Swenson 2014). Consequently, global trait datasets for many organisms remain sparse (Yoccoz et al. 2001; Kéry & Schmid 2004; Roth et al. 2017). This lack of knowledge hampers the accurate identification of locations and groups with unique patterns of FD (Kattge et al. 2011; González-Suárez et al. 2012; Swenson 2014; Madin et al. 2016). Despite the pervasiveness of missing values in trait datasets, present tools for FD analysis do not provide data gap-filling methods. Instead, listwise deletion of taxa with missing data, or complete case analysis is performed as a default setting (e.g. dbFD function in FD package in R; Laliberté et al. 2014). Such reduction in the size of a dataset leads to erroneous or biased inferences (Moorcroft 2006; Nakagawa & Freckleton 2008; van Buuren 2012). For trait datasets, a reduction in a dataset undermines FD and biogeography analyses (Pakeman 2014; Májeková et al. 2016; Borgy et al. 2017; van der Plas et al. 2017). Given the benefits of FD analyses and urgency to utilise their applications under rapid global change, a general gap-filling framework relevant across taxonomic domains and ecological disciplines is urgently needed to avoid misinterpretation of FD patterns (di Marco et al. 2014; Swenson 2014; Madin et al. 2016).

Three numerical treatments can complement arduous, expensive, and time-consuming fieldwork without performing complete case analyses: single imputation, likelihood-based approaches, and multiple imputation (van Buuren 2012; Newman 2014). Whilst previous studies have adopted various modifications of the three approaches to tackle the missing data problem, most were limited to exclusively numerical approaches without accounting for biological, ecological, or evolutionary features of species with missing information; such approaches lack a foundation as to why any particular approach is suitable for trait data (Penone et al. 2014). One of the most fundamental and striking patterns in ecology and evolution befits the alternative theoretical framework to purely numerical methods; the resemblance in trait expression among closely related taxa, or strong phylogenetic signal in biological and ecological traits, is common across taxonomic domains from microbes, terrestrial plants and mammals, to marine invertebrates, and can be implemented in an imputation framework as phylogenetic imputation to estimate missing trait values (Blomberg et al. 2003; Donoghue 2008; Baird et al. 2009; Cavender-Bares et al. 2009; Harrison 2011; Kerr et al. 2011; Swenson 2011, 2014; Martiny et al. 2013; Penone et al. 2014). Despite the apparent need to avoid complete case analyses and advantages of phylogenetic imputation, studies using phylogenetic imputation are rare, and the degree to which phylogenetic imputation benefits ecological analyses over complete case analyses remains unexplored.

Here, we test if a trait dataset amended by imputing missing trait data produces a better estimation of true FD than a trait dataset constructed by removing missing data. We investigate a suite of scenarios with various patterns in missing data and types of species occurrence data for both theoretical generality and relevance to empirical data. We use phylogenetic
eigenvector regression models (Diniz-Filho et al. 1998; Kühn et al. 2009) as the base model for phylogenetic imputation, with the expectation that phylogenetic imputation estimates accurate trait values, and reduces inferential errors in subsequent ecological analyses. Together, our findings underscore the risk of complete case analyses in ecological studies. In addition, our approach to estimating expected values for missing trait information is relevant across taxonomic domains and ecological disciplines, and provides advances in our understanding of biological and ecological traits through transcending pervasive data gaps in current trait databases.

MATERIALS AND METHODS

Dataset

We first generated 1000 phylogenetic trees, each with 200 taxa. For each phylogeny, we simulated a total of six traits, comprising three continuous and three discrete traits. To avoid using phylogenies with the same tempo and mode of evolution for the phylogenetic relationship among taxa and trait simulations, we rescaled the generated phylogenies for each trait using the delta model of trait evolution prior to trait simulation (Pagel 1999). Continuous traits were simulated with a rejection sampling algorithm to exhibit a gradient of phylogenetic signals in traits. Discrete traits were binary, and were simulated using a Markov model with randomly variable transition probabilities between binary states to achieve a gradient of phylogenetic signals in traits (Paradis et al. 2004).

We selected Blomberg’s $K$ to measure phylogenetic signal in continuous traits. Blomberg’s $K$ is a scaled ratio of the variance among observed trait values over the variance among expected trait values (Blomberg et al. 2003). $K$ values can range from 0 to infinity; $K = 0$ indicates no phylogenetic signal; $K = 1$ suggests that the trait evolved as expected under Brownian Motion; $0 < K < 1$ indicates that trait values among closely-related species are less similar than expected under Brownian Motion; $K > 1$ suggests that trait values among closely related species are more similar than expected under Brownian Motion. For discrete traits, we adopted a phylogenetic signal metric for binary traits, $D$, from Fritz & Purvis (2010). $D$ is a scaled ratio of the observed sum of state differences between sister clades over the mean expectations from a model with stochastic distribution of binary traits across a phylogeny. $D$ is not limited by a range boundary; $D = 0$ indicates that the trait is conserved as expected under Brownian Motion; more negative $D$ values suggest stronger phylogenetic signals; more positive $D$ values indicate weaker phylogenetic signals. We report the additive inverse of $D$ values in results and figures to render the inferences of $D$ value boundaries analogous to those of $K$. Phylogeny and trait data were simulated using ‘ape’ (Paradis et al. 2004), ‘caper’ (Orme et al. 2013), ‘geiger’ (Harmon et al. 2008), and ‘phytools’ (Revell 2012) packages in R (R Core Team 2017).

For species occurrence data, we generated two sets of 30 communities for all simulated trait data replicates. Each community harboured a random subset of the species pool available in the phylogenetic tree. The first set of simulated communities included species presence-absence data, whereas the second set of simulated communities contained log-normally distributed species abundance information. Each community was simulated independently to avoid multicollinearity in species occurrence among communities.

Sources of error in FD measurements

Functional diversity measurements can be compromised by missing entries in species trait or occurrence data. In empirical data, the availability of trait information is often systematically biased; missing data are more common among both rare taxa and taxa with particular trait expressions (Yoccoz et al. 2001; Kéry & Schmid 2004; González-Suárez et al. 2012; Sandel et al. 2015; Roth et al. 2017). Therefore, we first identified 5–15% of taxa, in increments of 5%, that represented locally rare (low local population size) and geographically restricted (limited geographic range) taxa (Rabinowitz 1981), and removed their trait values to simulate taxonomic bias in trait measurements. Subsequently, we removed an equal proportion of trait entries to simulate measurement bias in trait data (Sandel et al. 2015). For continuous traits, we removed trait entries that exhibited the lowest end of the trait expression spectrum. For discrete traits, we first identified a trait expression with lower frequency within the binary states, and an equal proportion of entries was randomly removed from the taxa pool with the trait value of lower frequency. Consequently, the total proportion of taxa with missing trait entries ranged between 10 and 30%, in increments of 10%. For theoretical generality, we also generated datasets where missing trait entries were randomly dispersed in trait matrices. We decided not to generate datasets with over 40% of missing information because too few taxa with complete trait data remained in trait matrices when missing entries were randomly dispersed. In total, we generated 12,000 replicates of species occurrence and trait data.

Phylogenetic imputation

We selected multiple imputation as the basis for our imputation framework over other approaches because multiple imputation resolves overly optimistic standard error ranges associated with other imputation methods, such as single imputation and likelihood-based approaches (van Buuren 2012; Newman 2014). Among the available multiple imputation methods, multiple imputation by chained equations (hereafter ‘MICE’) offers users the flexibility to implement various multiple imputation models, resulting in minimal error and bias (Ambler et al. 2007). Here, we use MICE with the recursive partitioning technique as it effectively handles highly dimensional data with nonlinear structures, commonly found in trait datasets (Santini et al. 2013; Doove et al. 2014). In addition, MICE with the recursive partitioning technique recognises variability in end nodes of regression and classification trees, and reflects such variability in the range of imputed estimates (Burgette & Reiter 2010). This property allows the incorporation of intraspecific trait variability, which is crucial in evaluating central themes in ecology, such as adaptation, acclimation, and biological interactions (Albert et al. 2011;
Phylogenetic imputation was included in multiple imputation as phylogenetic eigenvectors (Diniz-Filho et al. 1998; Kühn et al. 2009; Swenson 2014). Phylogenetic eigenvectors were extracted from principal component analyses on pairwise distance matrices generated from the original phylogenetic trees including all taxa. To avoid over- and under-fitting of phylogenetic eigenvector regression models, we included eigenvectors that represented at least 95% of the variation in the phylogenetic structure of each tree. It is important to note that eigenvector regression modelling has been criticised for its inability to incorporate all eigenvectors (Rohlf 2001; Freckleton et al. 2011). Phylogenetic eigenvector regression modelling is required to exclude a number of eigenvectors in order to leave extra degrees of freedom for and account for variance in trait data (Freckleton et al. 2011). Consequently, phylogenetic eigenvector regression modelling fails to capture the complete phylogenetic structure, theoretically reducing performance of phylogenetic imputation (Swenson 2014). However, empirical studies show no evidence that the phylogenetic eigenvector regression approach is devalued for phylogenetic imputation; rather, trait value imputation can be effectively achieved using phylogenetic imputation with eigenvectors (Penone et al. 2014; Swenson 2014). Previous studies have also included phylogenetic information in the form of a taxonomic hierarchy to predict trait values (e.g. BHPMF – Schrodt et al. 2015). While the novelty of this approach is uncontroversial, Linnaean hierarchy does not and cannot completely represent the true phylogeny (Ereshefsky 2000). In addition, taxonomic polyphyly is common in many species groups from bacteria (Gugger & Hoffmann 2004), corals (Luck et al. 2013), fish (Hubert et al. 2012), to crabs (Tsang et al. 2018), and this undermines the suitability of this approach as a general phylogenetic imputation framework.

The performance of phylogenetic imputation was assessed by measuring effect sizes of the association between the original and imputed trait values. Pearson correlation coefficients were computed as the measure of association in continuous traits. Cramér’s V was calculated as the measure of association in discrete traits. Like other common correlation coefficients, a value closer to 1 indicates a stronger correlation, and a value closer to 0 suggests a weaker correlation for both metrics. Correlation metrics were calculated for each trait individually, and mean and standard error values for each trait were obtained for each multiple imputation set, resulting in a total of 12,000 sets of coefficient terms for each trait. Effects of phylogenetic signal and proportion of missing data on the performance of phylogenetic imputation were assessed with random forest implemented in the ‘randomForest’ package (Liaw & Wiener 2002) and using linear mixed-effects model in the ‘nlme’ package (Pinheiro et al. 2017) in R (R Core Team 2017). The performance of phylogenetic imputation based on raw simulation data and the predicted performance values were compared using a Student’s t-test.

Functional biogeography simulation

We analysed FD of simulated communities using widely used FD metrics: functional richness (FRic; Villéger et al. 2008), functional evenness (FEve; Villéger et al. 2008), functional divergence (FDiv; Villéger et al. 2008), functional dispersion (FDis; Laliberté & Legendre 2010), and Rao’s quadratic entropy (RaoQ; Rao 1982; Botta-Dukát 2005). We computed FD metrics for the generated communities using the complete, imputed, and missing-data-removed (i.e. complete case, hereafter ‘deleted’) datasets. We also measured functional uniqueness (Violle et al. 2017) of each taxon in each occurrence replicate, and evaluated how loss of functional uniqueness from taxa with missing information contributes to error in FD analyses. FD and functional uniqueness metrics were calculated using the ‘FD’ (Laliberté & Legendre 2010) and ‘funrar’ (Grenié et al. 2017) packages in R (R Core Team 2017). Divergence matrices, and ANOSIM and PERMANOVA statistics were computed using the ‘vegan’ package (Oksanen et al. 2017) in R (R Core Team 2017). We used a parallelised version of the dbFD function (see Supporting Information).

The absolute values of FD metrics may highlight potential issues of a particular data manipulation practice, yet provide little ecological information (Taugourdeau et al. 2014). In contrast, their rankings across communities signify ecological and conservation values of the communities (Díaz et al. 2007; Taugourdeau et al. 2014). Therefore, we computed Kendall’s tau rank correlation coefficients between complete and imputed, and complete and deleted datasets, and compared their coefficients. A tau value of 1 indicates that FD rankings of communities in the complete and analysed datasets are in perfect agreement, whereas a tau value of −1 suggests that FD rankings of communities in the complete and analysed datasets are a direct inversion of each other. A tau value of 0 implies a lack of association. Effects of the FD metric, data manipulation technique, and dataset attributes on the tau coefficient were assessed, using linear mixed-effects models in R (‘nlme’ package; Pinheiro et al. 2017). Post hoc and linear hypotheses tests were performed using the ‘lsmeans’ (Lenth 2016) and ‘multcomp’ (Hothorn et al. 2008) packages in R (R Core Team 2017).

RESULTS

Performance of phylogenetic imputation

Correlation coefficients between the original and imputed datasets (hereafter ‘performance’) varied widely depending on the strength of the phylogenetic signal in traits (Fig. 1; Table S1). Overall, a stronger phylogenetic signal in a trait enhanced performance (Fig. 1; Table S1), and a higher proportion of missing data resulted in an inferior performance (Fig. 1; Table S1). When both the phylogenetic signal and proportion of missing data were accounted for, the effect of
phylogenetic signal on performance scaled with proportion of missing data (Fig. 1; Table S1). Mean performance values predicted by random forest models were significantly different from the raw performance values (Fig. 1; Table S2). Mean performance values predicted by random forest models always exceeded the raw values for continuous traits (Table S2). The gap between mean performance values predicted by random forest models and raw values decreased with an increasing proportion of missing data for discrete traits (Table S2).

**Functional biogeography simulation**

Both similarity statistics $R$ and pseudo-$F$ were lower for the complete and imputed dataset pairs than for the complete and deleted dataset pairs (Fig. 2; Table S3). The dissimilarity between dataset pairs (i.e. complete and imputed, and complete and deleted) scaled with an increasing proportion of missing information in both types of community data (Fig. 2; Table S3). This pattern was more pronounced when the species occurrence data included presence and absence information (Fig. 2; Table S3). The gap in dissimilarity between dataset pairs was smaller when the distribution of missing entries in trait data was concentrated among rare species (Fig. 2; Table S3).

Overall, Kendall’s tau rank correlation coefficients between the original and deleted/imputed datasets (hereafter ‘tau’) varied with the types of FD metric, data manipulation technique, proportion of missing data, and presence of systematic bias in

---

**Figure 1** Raw and predicted performance of phylogenetic imputation. Performance of phylogenetic imputation is assessed by measuring correlation between complete and imputed datasets with varying proportions of missing data and strengths of phylogenetic signal: (a) raw and predicted performance of phylogenetic imputation for continuous traits, (b) raw and predicted performance of phylogenetic imputation for discrete traits. Higher Blomberg’s $K$ and additive inverse of $D$ (i.e. $-D$) values indicate stronger phylogenetic signal. Solid lines indicate locally weighted scatterplot smoothing fits of the predicted performance of phylogenetic imputation for each dataset based on random forest models. The shaded areas correspond to the 95% confidence interval for the fits.

© 2018 John Wiley & Sons Ltd/CNRS
trait data (Fig. 3; Table S4). Tau values uniformly decreased with an increasing proportion of missing data (Fig. 3; Table S4). FD values computed from imputed datasets were more similar to those of the complete dataset than FD values calculated from deleted datasets, regardless of the type of FD metric (Fig. 3; Table S4). Community data with abundance information produced slightly better estimations of true FD values (Fig. 3; Table S4). The presence of systematic bias in trait data improved the estimation of true FD values for deleted (i.e. complete case) datasets (Fig. 3; Table S4). The effect of systematic bias in trait data waned for imputed datasets (Fig. 3; Table S4). The accuracy and precision in FD measurements among imputed datasets were only marginally improved with the presence of systematic bias in the trait data (Fig. 3; Table S4). Among the FD metrics, FDis and RaoQ values resulted in higher tau values (Fig. 3; Table S4). Loss of functional uniqueness from missing taxa showed mixed results depending on the type of FD metrics (Fig. 4; Table S5). The accuracy of FD measurements declined for FRic, and FEve computations when using species presence/absence information, with an increase in functional uniqueness loss. In contrast, FEve computed from species abundance information, FDiv, FDis, and RaoQ were all unaffected by the loss of functional uniqueness from missing taxa (Fig. 4; Table S5).
DISCUSSION
Global FD analyses are uncommon compared to local and regional-scale studies despite the availability of several global trait databases, and even though FD can provide insights into a suite of ecosystem features, including ecosystem functioning, ecosystem services, and resilience to disturbances and climate change (Walker et al. 1999; Díaz et al. 2007; Roscher et al. 2012; Mouillot et al. 2013a). This may well be a function of a fundamental lack of agreement in the literature as to whether global trait databases contain an appropriate amount of information for accurate FD measurements. Missing information plagues trait databases (Kattge et al. 2011; González-Suárez et al. 2012; Penone et al. 2014; Swenson 2014; Madin et al. 2016), and great caution must be exercised in handling missing information because downplaying missing information results in misleading projections (Moorcroft 2006; Nakagawa & Freckleton 2008; Purves & Pacala 2008; van Buuren 2012; Pukeman 2014; Borgy et al. 2017; van der Plas et al. 2017). Global FD analysis is likely to become more accessible when the missing data problem is addressed.

Phylogenetic imputation is a viable solution to estimate missing trait values, especially for traits with a strong phylogenetic signal. In particular, our findings suggest that a strong phylogenetic signal in traits offsets the crippling effect of an increasing proportion of missing information on the
performance of phylogenetic imputation (Fig. 1; Tables S1 and S2). This is a critical advance over findings from previous phylogenetic imputation studies, which were restricted to bivariate associations between the proportion of missing data and performance of phylogenetic imputation (Penone et al. 2014), and between the strength of the phylogenetic signal and the performance of phylogenetic imputation (Goberna & Verdú 2016). Our findings also render phylogenetic imputation an attractive alternative to data removal in global trait analyses because a strong phylogenetic signal in traits is common across taxa (Cavender-Bares et al. 2009; Swenson 2011; Martiny et al. 2013; Goberna & Verdú 2016), and supports the notion that an appropriate phylogenetic imputation approach can counterbalance the detrimental effects of missing trait values in trait databases. Nevertheless, we emphasise that phylogenetic imputation must be used with caution.

We foresee two potential misuses of phylogenetic imputation: use of single imputation approaches, and the erroneous reduction of multiple imputations to a single imputation framework. Single imputation approaches produce the same imputed values every time, and are incapable of providing error associated with imputed estimates; therefore, the precision of parameter estimates is overly optimistic (van Buuren 2012), and ecological analyses based on single imputation approaches are subject to errors, especially if the proportion of missing data in the original dataset is high (e.g. McWilliam et al. 2018). Multiple imputation resolves this issue by

**Figure 4** Rank correlation between data manipulation approaches with distinct types of species occurrence data, and mean loss of functional uniqueness from taxa with missing trait entries for: (a) functional richness (FRic; Villéger et al. 2008), (b) functional evenness (Villéger et al. 2008), (c) functional divergence (Villéger et al. 2008), (d) functional dispersion (Laliberté & Legendre 2010), and (e) Rao’s quadratic entropy (Rao 1982; Botta-Dukát 2005). Rank correlations are measured using Kendall’s tau. The regression lines indicate predicted tau values for a given loss of functional uniqueness based on linear mixed-effects model. Only significant relationships are depicted. The shaded areas correspond to the standard error for the fits. Note the lack of species occurrence treatment for FRic because FRic cannot handle species abundance information.
generating multiple different estimates for each missing entry, creating multiple distinctive imputed datasets. After analysing each imputed dataset separately, parameter estimates are subsequently combined across datasets using Rubin’s rules to incorporate both the within and among analysis sources of variation (Rubin 1987). Other measures to reduce imputed estimates from multiple imputation (e.g. taking mean/median of imputed estimates) defeat the very purpose of multiple imputation, and need to be avoided. It is also important to note that multiple imputation assumes that missing data are dispersed randomly in a dataset, or other variables in the datasets are associated with the pattern of missing data (van Buuren 2012). A violation of these assumptions can reduce the effectiveness of imputation (van Buuren 2012). Missing entries are more common among rare species in empirical data. Therefore, for empirical data, species occurrence information ought to be included in the dataset to be imputed, so that the probability of missingness can be computed based on species occurrence during imputation.

While phylogenetic imputation showed an excellent ability to estimate missing trait values in past studies, it was unclear as to whether the degree of error in subsequent ecological analyses based on an imputed trait dataset was more or less tolerable than the error stemming from removing records (i.e. taxa) with missing information (Penone et al. 2014; Swenson 2014; Taugourdeau et al. 2014). Our results show that imputed trait datasets produce better estimations of true FD than datasets that have been constructed by deleting records. This finding is attributable to the computational elements of the FD metrics. All FD metrics comprise terms associated with species occurrence (presence-absence or abundance) and traits of taxa within a community; therefore, the completeness of species occurrence and trait matrices is directly linked to the degree of error in FD computations (Rao 1982; Botta-Dukát 2005; Villéger et al. 2008; Laliberté & Legendre 2010). FD computations based on imputed datasets only suffer errors in terms associated with traits because terms associated with species occurrence for imputed datasets are identical to those of the complete dataset. Therefore, the degree of error in FD analyses using imputed datasets is dictated by the factors that influence the performance of imputation, namely the strength of phylogenetic signal in traits of interest, and the proportion of missing data in the trait dataset. On the other hand, complete case analyses (i.e. analyses where species with missing trait values are removed) are prone to errors in both species occurrence and trait associated terms because taxa without complete trait information are removed and do not contribute to both the trait and community data. These errors result in arbitrary FD rankings among communities, and weaker rank associations between the true and computed FD values than their imputed counterparts, regardless of dataset attributes. Our findings corroborate and quantify the recommendations against data removal in ecological analyses (Ellington et al. 2014; Pakeman 2014).

Our results are particularly important for estimating the suitability of trait databases for global analyses. Taxonomic, spatial, and measurement biases are ubiquitous in global trait databases, often heavily skewed toward abundant and widespread taxa with a limited range of trait expressions (Yoccoz et al. 2001; Kéry & Schmid 2004; Swenson 2014; Sandel et al. 2015). Given their low abundance with restricted distribution, the consequences of losing rare species have been deemed minimal, and the ecological significance of losing rare species has long been overlooked (Grime 1998). However, recent findings emphasise that rare species also make significant and unique contributions to FD (Mouillot et al. 2013b; Jain et al. 2014; Chapman et al. 2018), and the loss of rare species can impair long-term and large-scale ecosystem functioning (Lyons et al. 2005).

Previous studies reported that systematic bias in a dataset compromises ecological analyses, yet associations between the degree of systematic bias and the accuracy of FD measurements, and the distribution of taxa with missing trait values in trait space and the accuracy of FD measurements, remain unexplored (Pakeman 2014; Májeková et al. 2016; Borgy et al. 2017; van der Plas et al. 2017). Our results suggest that FD measurements in complete case analyses are substantially affected by the degree of systematic bias in trait data. This is not surprising because the amount of data removed in a complete case analysis scales with the degree of stochasticity in the distribution of missing entries (van Buuren 2012). In contrast, systematic bias in trait data only caused marginal effects on datasets amended by phylogenetic imputation. This attribute renders phylogenetic imputation an applicable framework across taxonomic domains as the degree of systematic bias in empirical data varies widely between the two extreme scenarios we tested (random placement of missing trait values vs. rarest taxa with missing trait values; Kéry & Schmid 2004). While a trend across a wider range of loss in functional uniqueness is more desirable, our findings also indicate that non-random sampling of rare species in functional space can result in misleading interpretation of FD patterns for FRic, and FEve computed with species presence and absence information (Fig. 4). Given the fact that rare species can support the most unique trait combinations (Mouillot et al. 2013b) and it is impossible to gauge functional uniqueness of missing species in empirical data, the use of these metrics is more likely to result in erroneous inferences, compared to other FD metrics.

As we have shown, the accuracy of FD measurements is heavily influenced by the completeness of trait and species occurrence data. The completeness of trait and species occurrence data is largely dependent upon taxonomic and spatial coverage of trait data (Borgy et al. 2017); the difficulty in completing trait and species occurrence data escalates with an increase in taxonomic and spatial coverage (MacKenzie et al. 2002; Jarzyna & Jetz 2016). Consequently, there must be an inevitable trade-off between taxonomic and spatial coverage, and the accuracy of FD measurements (Borgy et al. 2017). The accuracy of FD measurements can be improved in two ways: by restricting taxonomic and spatial resolution for the completeness of raw trait and spatial data, or by selecting taxonomic and spatial resolutions for optimal trait and spatial imputations. Both of these approaches entail compilation of taxa with low proportions of missing data, and traits with strong phylogenetic signal at a local or regional spatial scale, to ensure the accuracy of imputations and FD measurements (e.g. Ding et al. 2013; Denis et al. 2017). In short, FD studies at the global scale are likely to suffer considerable errors with the amount of data available in current trait databases.
Orchard & Woodbury (1972) noted that the best method for handling missing data is to have no missing data at all. In spite of the apparent advantages of complete datasets, missing data are common in ecology and evolution, and perhaps even more prevalent in global trait databases (Nakagawa & Freckleton 2008; Kattge et al. 2011; Penone et al. 2014; Swenson 2014; Gobena & Verdú 2016). Reviews across disciplines caution against neglecting missing data (cancer research – Burton & Altman 2004; ecology – Pakeman 2014; Borgy et al. 2017; van der Plas et al. 2017; epidemiology – Klebanoff & Cole 2008; political science – Lall 2016; statistics – Little & Rubin 2002; van Buuren 2012), yet statistical software and modules quietly eliminate cases with missing data (i.e. complete case analysis). Complete case analysis is also a norm in FD analysis tools (Lal–van Buuren 2012), yet statistical software and modules quietly eliminate cases with missing data (i.e. complete case analysis). Our findings indicate that downplaying missing data distorts FD measurements, and results in misinterpretation of patterns in FD. While the most preferable solution to the ubiquitous problem of missing values in trait databases is to complement current databases with further trait measurements, complete trait databases with no missing data are improbable in the near future. Alternatively, an appropriate use of phylogenetic imputation provides powerful tools to estimate expected values for missing trait data, and the resultant imputed datasets reduce inferential errors in FD analyses regardless of systematic bias in trait data. It is evident that FD will continue to gain traction because of its perceived generality and ability to expound ecosystem functions (Díaz & Cabido 2001; Bellwood et al. 2004; Hooper et al. 2005). Nevertheless, numerous gaps in trait data are conspicuous, and thus stimulated multiple interests in approaches to overcome missing data and impact assessments of missing data in ecological analyses (Pakeman 2014; Penone et al. 2014; Swenson 2014; Borgy et al. 2017; van der Plas et al. 2017). Here, we have provided a unified approach that addresses the risk of complete case analyses, and highlights a general framework that alleviates ubiquitous missing data problems in ecology. The simplicity, generality, and utility of our framework offer a unique avenue for ecological studies across disciplines and taxonomic domains to tackle fundamental questions in ecology using trait data.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their insightful comments that greatly improved this manuscript. This research was funded by the Korea Institute of Marine Science and Technology research grant (PM60522), International Postgraduate Research Scholarship and the University of Queensland Centennial Scholarship to SWK, and funding from the Australian Research Council Centre of Excellence for Coral Reef Studies grant (CE140100020) to JMP and others.

AUTHORSHIP

SWK, SPB and JMP conceived the ideas. SWK and SPB designed methodology. SWK collected data. SWK and SPB analysed output data. SWK wrote the first draft of the manuscript. All authors contributed to subsequent revisions.

DATA ACCESSIBILITY

This study does not use empirical data. R script to scale up FD analyses is available from the following Figshare repository: https://doi.org/10.6084/m9.figshare.6167903.v1.

REFERENCES


© 2018 John Wiley & Sons Ltd/CNRS