REVIEW ARTICLE
https://doi.org/10.1038/s41564-018-0156-0

nature microbiology

Methods for phylogenetic analysis of microbiome data

Alex D. Washburne[1,6]*, James T. Morton[2,3,6], Jon Sanders[3], Daniel McDonald[3], Qiyun Zhu[3], Angela M. Oliverio[4,5] and Rob Knight[2,3]

How does knowing the evolutionary history of microorganisms affect our analysis of microbiological datasets? Depending on the research question, the common ancestry of microorganisms can be a source of confounding variation, or a scaffolding used for inference. For example, when performing regression on traits, common ancestry is a source of dependence among observations, whereas when searching for clades with correlated abundances, common ancestry is the scaffolding for inference. The common ancestry of microorganisms and their genes are organized in trees—phylogenies—which can and should be incorporated into analyses of microbial datasets. While there has been a recent expansion of phylogenetically informed analytical tools, little guidance exists for which method best answers which biological questions. Here, we review methods for phylogeny-aware analyses of microbiome datasets, considerations for choosing the appropriate method and challenges inherent in these methods. We introduce a conceptual organization of these tools, breaking them down into phylogenetic comparative methods, ancestral state reconstruction and analysis of phylogenetic variables and distances, and provide examples in Supplementary Online Tutorials. Careful consideration of the research question and ecological and evolutionary assumptions will help researchers choose a phylogeny and appropriate methods to produce accurate, biologically informative and previously unreported insights.

High-throughput sequencing yields information about microbial communities in quantities that outstrip our ability to make sense of it. Most microbial taxa have never been cultivated or experimentally characterized. For many, we have only sequence fragments, whole genome sequences for a few distant relatives, and a tree capturing the microorganisms' evolutionary histories. How can we organize and analyse the deluge of information about uncharacterized microorganisms and their sequence fragments?

Two essential tools for organizing the diversity of life are the taxonomy and the phylogeny. The taxonomy classifies a microorganism based on a hierarchy of taxonomic names ranging from three domains (Bacteria, Archaea and Eukarya) to several million species. The phylogeny is an estimation of the microorganisms' evolutionary history and classifies every organism by a series of splits corresponding to estimated events in which a most recent common ancestor speciated to form two daughter species.

Microbial taxonomy and phylogeny may eventually be equivalent, with every clade in the phylogeny having a taxonomic name. However, contemporary taxonomic classification is coarse; modern taxonomic labels categorize a small fraction of the branches in the phylogeny. For the time being, the phylogeny is a more detailed scaffold for microbial classification.

Phylogenies are a tool to organize and understand the microbial world[1,2]. Because related organisms tend to have similar characteristics, phylogenies can incorporate those characteristics into our analyses even if we cannot measure them directly. Phylogenies are a scaffold to classify lineages and infer functional ecological traits, even for lineages that have not been classified taxonomically or physiologically. Microbial ecology can be accelerated by high-throughput classification and inferences made possible with phylogenies. Resource consumption[3], habitat associations[4] and species interactions[5,6] are causes and consequences of traits, and using phylogenies to infer or implicitly work with traits may enhance our ability to manipulate microbial communities to impact human health[7], biogeochemistry[8] and climate change[9].

How can a phylogeny assist analyses of microbiome data? Different research questions require different considerations about how to amend statistical analyses using a phylogeny. For example, tests of associations between traits should consider the phylogeny as a source of dependence among observations, whereas studies looking for simpler ways of binning species should consider the phylogeny a scaffold for possible bins. There is a vast and growing literature on methods for analysing phylogenetically structured data, methods with subtle yet consequential differences in the questions they seek to answer. There is a need to simplify the diverse field into a set of conceptually distinct classes of methods and thereby provide a framework for instruction, comparison and development of methods for analysing phylogenetically structured data.

In this Review Article, we organize the field of phylogenetically structured data analysis by discussing the major classes of methods. We first emphasize a fundamental issue in the field: the imperfection of estimated phylogenies. We then define four classes: (1) comparative methods; (2) ancestral state reconstruction and descendant trait imputation; (3) phylogenetic variables; and (4) phylogeny-aware distances (Table 1); and provide Supplementary Online Tutorials with examples (https://knightlab-analyses.github.io/phylogenetic-tutorials/). Most statistical tools can be revisited for phylogeny-aware analyses, but the categories we cover capture the most commonly used and actively developed classes of methods.

[1]Department of Microbiology and Immunology, Montana State University, Bozeman, MT, USA. [2]Department of Computer Science, University of California San Diego, La Jolla, CA, USA. [3]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA. [4]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA. [5]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA. [6]These authors contributed equally: Alex D. Washburne and James T. Morton. *e-mail: alex.d.washburne@gmail.com

Table 1 | Different classes of methods for using the phylogeny in data analysis address different questions

Class of methods	Brief description	Specific example	General formula	Highlighted methods
Comparative methods	Find associations between traits, controlling for evolution on phylogeny	Is 16S rRNA gene copy number associated with growth rates in vivo?	$\mathbf{y}_i \sim g(Y) + \varepsilon$ where $\mathrm{Cov}[\varepsilon] = f(P)$	PGLS[18] Paired *t*-test[23]
Ancestral state reconstruction	Impute trait values for historical lineages in the phylogeny and use ancestral traits to impute trait values for contemporary species	What is the best estimate of 16S rRNA gene copy number of an OTU based on the 16S rRNA copy numbers of its relatives?	Infer features of $P \mid \mathbf{y}$ Impute $y_{i,j} \mid \mathbf{y}_j$	PICRUSt[29]
Phylogenetic variables	Use the phylogeny to construct variables that are biologically interpretable (for example, a clade's abundance) and simplify/summarize features in the community	Which interior edges in *P* separate taxa with different habitat associations? How does Faith's phylogenetic diversity change with pH?	Define variables: $\mathbf{v}_i = f_i(\mathbf{x},P)$ Analyse, interpret and combine $\{\mathbf{v}_i\}$	Diversity analyses Taxonomic analyses Phylofactorization[34] EdgePCA[33] PhILR[35]
Phylogeny-aware distances	Use the phylogeny to construct distances between samples, which can then be used to modify statistical tools for classification, regularized regression and more	How different are two microbial communities?	Define distance: $d[\mathbf{x}_i,\mathbf{x}_j] = K(\mathbf{x}_i,\mathbf{x}_j,P)$ Analyse and use to modify various statistical methods	UniFrac[9,45–47] Inner product methods[48,49]

These methods can be summarized based on their use for a given dataset of abundance vectors, **x**, a dataset (*Y*) of observed or imputed trait values, **y**, and the phylogeny, *P*. Comparative methods are used when traits, $\mathbf{y}_i$, are response variables for a regression model $g(Y)$. Ancestral reconstruction imputes features on the phylogeny given trait data at the tips of the tree. Phylogenetic variables, $\mathbf{v}_i$, can be defined by various transforms of abundance (or trait) data using the phylogeny, $f(\mathbf{x},P)$. Distances between two vectors of community composition, $d[\mathbf{x}_i,\mathbf{x}_j]$, are defined using a kernel, (*K*), that incorporates features of community composition and the phylogeny. Variables can be defined for a single abundance vector, whereas distances are defined through pairwise comparisons of abundance vectors.

We discuss challenges of phylogenetically aware analysis of microbiome data, including horizontal gene transfer (HGT) and the choice of which genes to use when building phylogenies. By partitioning the literature into distinct conceptual classes of methods, we provide a common framework for the development and implementation of these important methods in microbiome data analysis. See Box 1 for a glossary of relevant terms.

Phylogenetic inference

The tree of life is not known—it is estimated, and accurate phylogenies improve accuracy of phylogenetically structured data analysis. Microbial phylogenies are commonly estimated by collecting gene sequences, aligning sequences based on homologies, and using models of mutation to infer most-likely evolutionary histories. The estimated phylogeny can vary depending on which genes are sequenced, how sequence positions are aligned, which model of evolution is used, and the method for inferring histories. Errors in phylogenetic inference can propagate to errors in phylogenetically structured data analysis. Here, we discuss the interplay between phylogenetic inference and phylogenetically aware analyses; for a review of methods for phylogenetic inference, readers can consult focused reviews of that topic[10,11].

One can construct a phylogeny for any gene; different genes vary in the number of species containing the gene, the resolution of the phylogeny, and phylogenetic signal of various traits. The 16S rRNA gene is commonly used for phylogenetic inference in Bacteria and Archaea, but one could also construct a phylogeny for other genes such as β-lactamases and their relatives, yielding a phylogeny with edges along which antibiotic resistance traits arose[10]. Microbial eukaryotes likewise have many genes that can be used for phylogenetic inference, the 18S rRNA gene being most commonly used[12].

The genes chosen for phylogenetic inference determine the set of traits correlated with the phylogeny. Bacterial genome trees generally correlate with the 16S rRNA gene (16S)-derived phylogenies[13], but the correlation between a 16S tree and gene content varies over lineages and phylogenetic depths[14]. HGT disrupts the correlation between 16S trees and gene content by allowing bacteria with distant 16S genes to share common and consequential traits, such as pathogenicity islands and antibiotic resistance genes[15,16]. Moreover, the 16S sequence has multiple variable regions, and can vary among multiple copies within a single genome, complicating phylogenetic inference[17]. More complicated scenarios, such as when epistasis underlies a functional ecological trait and one of the epistatic genes can be horizontally transmitted, prohibit a clear prescription for which gene's tree to use.

Different methods for analysing phylogenetically structured data use different features of the phylogeny. Distances and phylogenetic comparative methods (PCMs) that aggregate information over many branches in the phylogeny are more robust to errors in phylogenetic inference[18,19]. Methods which rely on a few branches are more sensitive to errors in phylogenetic inference[20]. For methods relying on a few internal nodes or branches, uncertainty in phylogenetic inference—particularly the bootstrap support for the monophyly of critical branches[21]—may be important to incorporate into downstream data analysis. Where monophyly is crucial and polytomies allowable, researchers can collapse resolved nodes into polytomies to improve the bootstrap support across the tree. A more certain yet coarse-grained phylogeny may be preferable to a less certain yet fully resolved phylogeny.

Phylogenetic comparative methods

PCMs are used when comparing multiple traits across organisms. Closely related organisms often have similar traits due to inheritance from a common ancestor; the dependence of traits across organisms can affect tests of trait:trait and trait:habitat associations.

For example, an association between 16S copy number (trait) and pH preference (habitat) could be found through a correlation between 16S copy number and a measure of pH preference across 1,000 species of microorganisms (Fig. 1a). However, the significance of the association could be exaggerated if the taxa surveyed consist of a set of closely related Acidobacteria with low 16S copy number and low pH preference, and a set of closely related Fusobacteria with high 16S copy number and a high pH preference[17]. Intuitively, the phylogenetic signal of these traits reduces our sample size because the observed traits represent samples from two lineages, not 1,000 independent samples. Robust tests of trait-associations are done using PCMs[22,23] (Fig. 1b).

Box 1 | Glossary of terms

Ancestral state. The traits of the ancestral species, typically an estimate of the phenotype and the genotype of the ancestral organism.

Ancestral state reconstruction. Imputing the ancestral states at various points in the phylogeny.

Bayesian inference. Given a prior set of beliefs about the ancestral states, and observed phenotypes/genotypes of existing species, Bayesian methods will attempt to obtain a more informed estimate of the ancestral states, along with confidences of the prediction.

Blomberg's *K*. A more common, modern measure of phylogenetic signal compared to Pagel's λ (see below), ranging from 0 to infinity, which indicates the extent of acceleration or deceleration of evolution over time.

Bootstrapping (phylogenetics). Repeated, stochastic reconstruction of a phylogeny proposed by Felsenstein[21], often used to assess the percentage of reconstructions in which each clade is found.

Brownian motion. A continuous random walk where jumps are normally distributed random variables. Commonly used in PCMs as a null model of continuous trait evolution from the ancestral node towards the tips of the tree, where the random walk branches with those in the phylogeny. Under a Brownian motion model of evolution, the covariances between species' observed traits is proportional to the branch length of their shared ancestry.

Classification. Regression or other efforts to predict categorical dependent variables.

Clustering. Creation of classifiers (categorical variables) identifying groups of variables, such as groups of species with high within-group similarity and low between-group similarity.

DNA amplicons. DNA products of artificial amplification events, such as the resultant products of polymerase chain reaction amplification of 16S rRNA genes that are later sequenced and counted to assemble microbiome datasets. Amplicons may sometimes be used to construct accurate phylogenies for microorganisms.

Edge. A structure in a phylogeny representing a hypothesized distinct, unbroken lineage during a point in time.

Edge lengths. Edge lengths may represent either the time over which an historical lineage persisted or the number of mutation events separating its ancestral from daughter nodes.

EdgePCA. A method which performs principal component analysis on a set of variables, v_i, corresponding to differences of abundances along each edge, i.

Epistasis. When two or more genetic loci interact to determine a phenotypic trait.

Evenness. A general term for a variety of metrics indicating how close a community is to having equal abundances across all species.

ILR. Isometric log-ratio. A standardized difference of arithmetic means of log-transformed data. Often used in microbiological datasets that are more appropriately analysed on a log scale, but an analogous difference for non-log-transformed data is the t-statistic for a two-sample t-test.

Maximum likelihood. Maximum likelihood methods treat ancestral states as unknown parameters. Given a probabilistic model of evolution, maximum likelihood methods will attempt to optimize these parameters to try to find the most likely ancestral states that yield the traits that we observe in known species in the present.

Maximum parsimony. Maximum parsimony attempts to reconstruct ancestral states by minimizing the number of trait changes between the ancestor and the present descendants.

Monophyletic. A set of species is called 'monophyletic' relative to a larger set of species if their most recent common ancestor has no other descendants besides those within the set of species.

Node. A structure in a phylogeny representing a hypothesized timing of speciation, when one lineage splits into two or more distinct lineages.

Pagel's λ. A measure of phylogenetic signal, ranging between 0 and 1, which indicates the relative extent to which a traits' correlations among close relatives match a Brownian motion model of trait evolution.

PhILR. Phylogenetic isometric log-ratio. A transform of the data requiring a fully resolved phylogeny (that is, no polytomies). Instead of representing data with one variable for each species, the PhILR transform represents the data with one variable for each node in the phylogeny. Variables are constructed using the ILR transform to contrast sister clades descending from each node.

Phylofactorization. A method of choosing variables by a generalized graph-partitioning algorithm. Variables are constructed by first considering contrasts along edges, such as differences or ILRs contrasting birds and non-birds, and then finding out which variable maximizes a researcher's objective function. The phylogeny is partitioned along that edge and the process is repeated, limiting contrasts only to sub-phylogenies in which the edges are found (for example, after partitioning birds/non-birds, the edge separating doves/non-doves instead separates doves/non-dove-birds).

Phylogenetic comparative methods. Statistical methods which correct for correlations of trait observations among close relatives, to be used whenever traits, broadly defined as heritable features at the tips of a phylogeny, are a dependent variable or when testing differences between two traits. PCMs often use models of evolution to calculate correlations between observations of close relatives expected under random evolution.

Phylogenetic distance. The sum of edge lengths along the path connecting two species in a phylogeny.

Phylogenetic inference. The estimation of the evolutionary history of a set of genes.

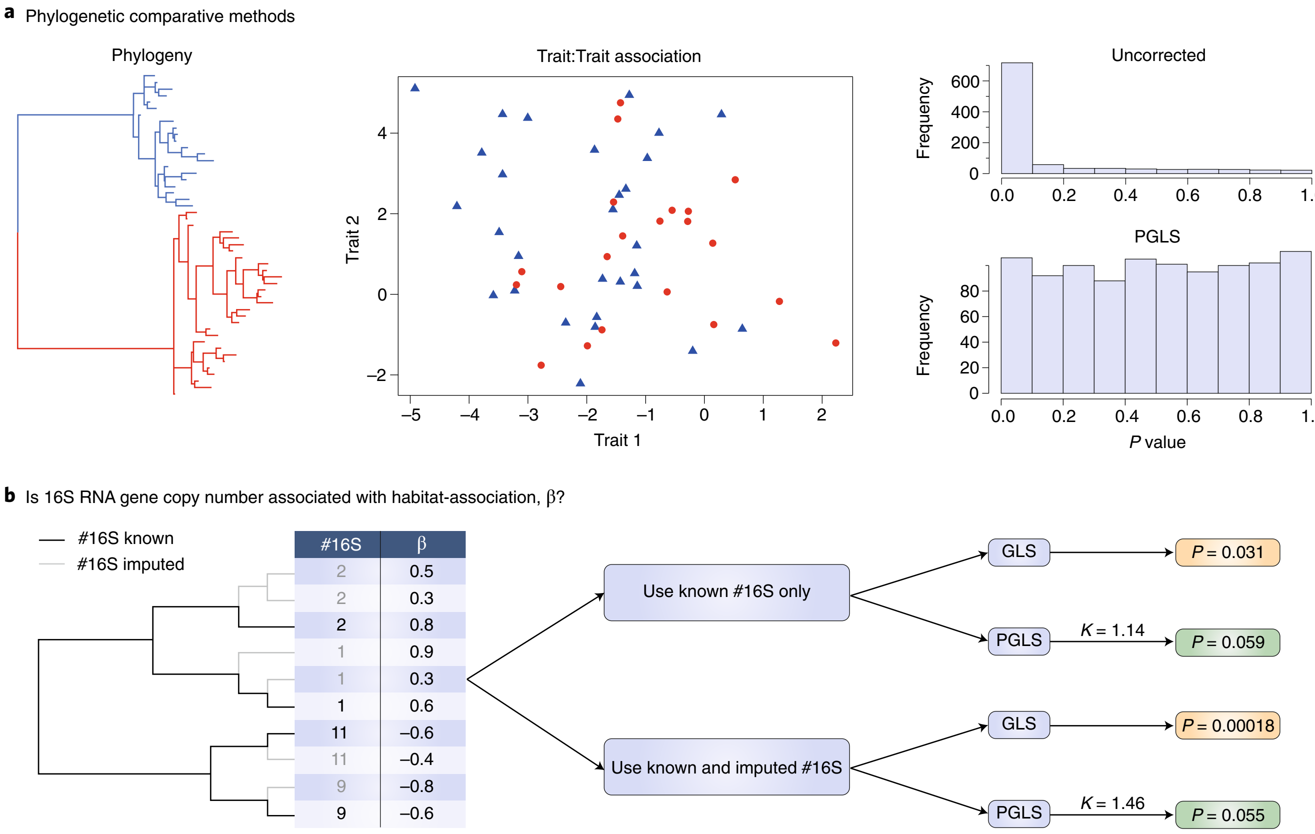
Phylogenetic variables. Variables constructed with the aid of a phylogeny (including the star phylogeny in which all species originate from the same polytomy). In contrast to phylogenetic distances, variables indicate directions and curves along which variation has biological meaning.

Phylogeny. A diagrammatic hypothesis of the evolutionary history of a set of genes. The phylogeny can be rooted, implying knowledge of the most basal common ancestor of the set of genes, or unrooted.

Polytomy. A node with more than two daughter lineages. Often, polytomies represent uncertainty about the precise timing of historical speciation events.

Regression. A predictive mathematical model that will attempt to estimate relationships between variables.

Shannon diversity. A particular measure of evenness, H, defined by a set of relative abundances, p_i, summing to 1: $H = -\text{sum}(p_i \log(p_i))$.



#16S	β
2	0.5
2	0.3
2	0.8
1	0.9
1	0.3
1	0.6
11	−0.6
11	−0.4
9	−0.8
9	−0.6

Fig. 1 | PCMs control for the statistical dependence among traits resulting from evolution of traits along the phylogenetic tree. a, An exaggerated phylogeny with two distantly related clades. If trait evolution is simulated as a random walk on the phylogeny, the two distantly related clades will drive covariances between traits. Failing to correct for the effects of random trait evolution can lead to a high false-positive rate. Methods such as PGLS correct for the residual covariance expected under random trait evolution and produce more accurate statistical tests of association. **b**, We illustrate PGLS using an imaginary test between the 16S copy number (#16S) and regression coefficient with environmental metadata, such as disturbance frequency (β). PGLS should be used when testing associations between traits, including trait quantities such as regression coefficients from abundance:metadata associations. To implement PGLS, a model of trait evolution needs to be assumed. We estimate Blomberg's *K* to test for phylogenetic signal; the values of $K > 1$ indicate that the relatives are more similar to one another than expected under a Brownian motion model. We then use corBlomberg from the R package phytools to control for residual covariance structure. PGLS should be used regardless of whether the traits used are known or imputed through ancestral state reconstruction. *P* values from F tests demonstrate that the significance of the association depends on whether or not one accounts for covariance structure expected under random evolution of traits.

Generalized least squares (GLS) can control for dependence among observations when performing regression. In GLS, residuals—the difference between predictions and observations—are expected to covary and the covariance matrix is used to modify least-squares calculations. Random evolution produces close relatives whose observed traits will covary due to the shared variation acquired during their shared ancestry[24]. Phylogenetic generalized least squares[22] (PGLS; Fig. 1), a tool for trait:trait and trait:habitat associations, implements GLS with residual covariances defined by a model of evolution.

A common first step in PCMs is to estimate and test the phylogenetic signal against a null model of no phylogenetic signal. The λ of Pagel[25] or the *K* of Blomberg et al.[26] are commonly used test statistics for phylogenetic signal. For PGLS, one must assume an evolutionary model; a Brownian motion, that is, a branching, random walk of trait values from an ancestral value at the root to the tips of the tree, is the default. The evolutionary model defines a covariance matrix for the residuals (Fig. 1b). Under a Brownian motion model of evolution, the covariance between the residuals of two species' trait values is proportional to the amount of shared evolutionary history; more closely related species have more closely related traits even under a null model of random evolution. For more complicated models of evolution, one can jointly estimate the parameters for the evolutionary model and the regression coefficients[27].

PCMs extend to many statistical tests. Testing whether the volume of bacterial spores is smaller than the volume of daughter cells would involve a paired *t*-test, absent of phylogenetic signal. A phylogenetic paired *t*-test[28] was developed to account for phylogenetic signal in such tests. There are many models of trait evolution, metrics of phylogenetic signal, and methods to control for phylogenetic signal when comparing traits. A recent scholarly edition of modern PCMs provides a review of the field and directions of current research[29].

PCMs are not commonly used in microbiome studies, although a recent study[30] has employed PCMs to identify genes associated with colonization of the human gut (trait:habitat). Failure to correct for phylogenetic dependence in tests of trait:trait and trait:habitat association can yield a high false-positive rate (Fig. 1b). To amend this, we recommend researchers familiarize themselves with and utilize PCMs. Many methods can be implemented through the R packages ape[31], phangorn[32], phytools[33], picante[34], caper[35], Geiger[36] and phylolm[37]. In the Supplementary Online Tutorial, we use these packages to simulate trait evolution, test associations between traits, and illustrate the sensitivity of these methods to HGT.

Ancestral state reconstruction

Estimating, or reconstructing, ancestral trait values assists imputation of traits in uncharacterized species and identification of lineages in which major trait differences arose. In microbiology, ancestral state reconstruction is commonly used to estimate genetic and metabolic profiles of extant communities using a set of reference genomes. In microbiome studies, this is commonly performed using PICRUSt[38], which uses ancestral state reconstructions to impute trait values, such as genes encoding glycoside hydrolase activity, for taxa whose traits are unknown.

PICRUSt operates on a 16S gene phylogeny connecting sequenced genomes and environmental sequences. Trait information in the sequenced genomes is used to infer ancestral trait profiles. Ancestral profiles are then used to impute the profiles of organisms in an environmental sample. An input sample's predicted metagenomic profile is estimated by adding the product of operational taxonomic unit (OTU) abundances in the sample and their corresponding imputed profiles. Because PICRUSt relies heavily on the reference database and available sequenced genomes, it underperforms in environments where few or no genome data is known. Conversely, PICRUSt predicts the profiles of whole genome shotgun sequencing in human faecal samples with a Spearman $r^2 > 0.9$ (ref. [38]).

The methodology underlying ancestral state reconstruction is related to PCMs, as both require models of evolution[39]. Three main types of algorithms are used for inference: maximum parsimony, maximum likelihood and Bayesian inference[40]. Maximum parsimony reconstructs ancestral states by minimizing the number of trait changes between the ancestor and the present descendants. This approach assumes that trait changes are slow, and does not account for scenarios involving rapid evolution. In addition, maximum parsimony treats all branches the same and minimizes the number of changes on each branch; this can be problematic, particularly if not all of the species have been observed[41]. Maximum likelihood and Bayesian inference improve on maximum parsimony by incorporating explicit models of evolution—such as a Brownian motion model of trait evolution along the tree—into the estimation of ancestral states. Rather than assuming that changes are rare, these methods can account for some changes occurring more frequently than others—for example, assuming synonymous substitutions are more frequent than non-synonymous substitutions—and fit parameters to these models given an estimated phylogeny. However, maximum likelihood often underestimates the number of changes within a single branch and can generate suboptimal results, particularly if the rate of evolution changes across the phylogeny[42]. Bayesian approaches can compute evolutionary parameters across a deep sampling of possible evolutionary trees and evaluate more complex models of evolution that account for non-uniform rates of evolution. While Bayesian methods can generate more accurate results than maximum parsimony or maximum likelihood, they are computationally expensive. Consequently, PICRUSt estimates microbial ancestral states using maximum parsimony or maximum likelihood. As for PCMs, estimates of ancestral states can be quickly confounded by HGT (see the Supplementary Online Tutorial), and thus applications of these methods to microbial datasets should be performed with consideration of the observed rates of transfer for the gene families of interest.

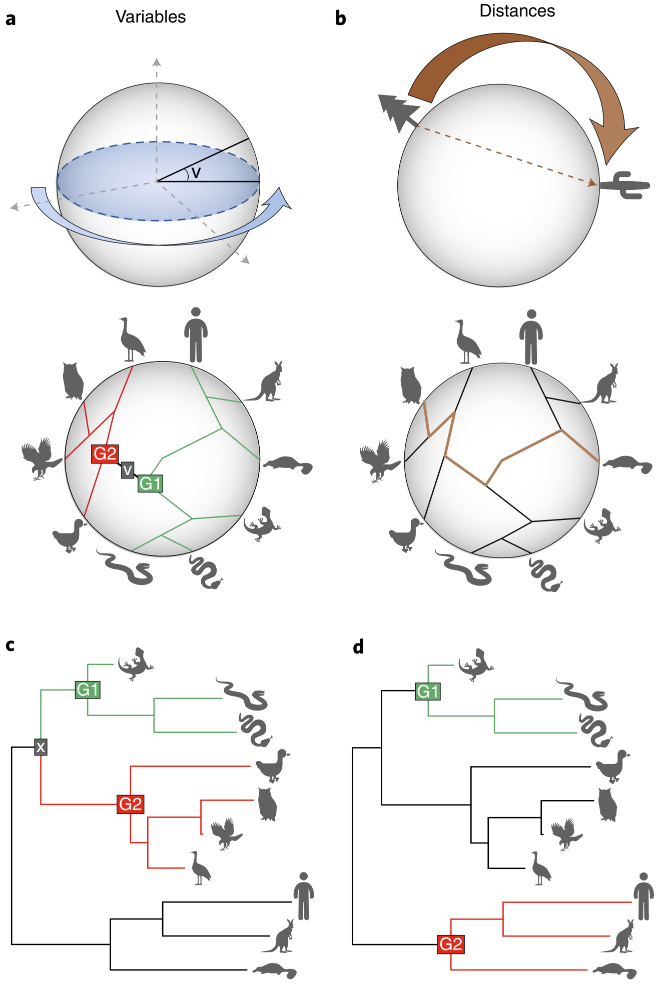


Fig. 2 | Phylogenies define the geometry of community ecological data, much like a sphere defines the geometry of GPS data. a, Changing variables can allow more natural descriptions of complex topologies. A spherical Earth indicates spherical coordinates. Phylogenetic variables use the tree as a scaffolding for constructing coordinates corresponding to phylogenetic features. Phylofactorization constructs coordinates for contrasting groups, G1 and G2, separated by edges where traits, such as flight, arose. **b**, A default path between two points is a straight line, but a more meaningful path on a sphere is a geodesic—that is, the shortest path along the surface of the sphere. Likewise, phylogeny-aware distances such as UniFrac define evolutionary paths and their distances between one community and another. **c**, PhILR constructs coordinates between contrasting sister clades. **d**, The space of possible phylogenetic variables and distances is infinitely large. Ratios between distant clades, as illustrated here, are viable but currently unused phylogenetic variables. Researchers should consider the biological interpretability of novel variables and distances, and their ability to inform future studies. Icons (kangaroo, platypus, lizard, outstretched snake, hawk, owl and crane) made by Freepik from www.flaticon.com.

Analysis of phylogenetic variables

Locations on the Earth's surface can be described with three Cartesian (xyz) coordinates, but they are more naturally described using two spherical coordinates (latitude and longitude). A phylogeny, similar to a sphere, suggests natural coordinates. Phylogenetic variables are used to reduce the dimension of community ecological data, simplify calculations of distances, and describe meaningful features and directions of change in communities (Fig. 2).

We coin the term 'phylogenetic variables' to describe variables constructed using features in the phylogeny to aggregate, contrast and summarize data of species in the phylogenetic tree (Fig. 2). Variables and distances are related but contain distinct information: saying the city is east does not indicate how far it is, and saying it is 80 km away does not indicate which direction it is. Directions are described through phylogenetic variables (Fig. 2a), and the magnitude of change is measured through distances (Fig. 2b). Phylogenetic variables include taxonomic abundances, diversity metrics, differences of abundance along all edges[43], differences of abundances between clades (Fig. 2a–d)[43,44], and more.

Phylogenetic variables simplify microbiome datasets by reducing the dimension of the data to a few variables carrying biological information. If a few monophyletic clades explain the majority of a microbiome dataset's variance along an environmental gradient, then there may be a few traits shared among members of each clade which determine abundance and underlie the observed community compositional changes along the environmental gradient.

The set of possible phylogenetic variables is infinitely large. Consequently, researchers must be deliberate when choosing phylogenetic variables—what are important directions of change that carry implications for further research? Community changes along the direction of a phylogenetic variable, such as alpha-diversity, do not necessarily convey useful biological information or readily suggest future research directions. Two common challenges in the analysis of phylogenetic variables can guide their selection and development: statistical dependence and biological interpretability.

Statistical independence, or well-characterized dependence, facilitates robust multivariate statistical analyses. For instance, when testing associations between species' abundances and environmental metadata, and repeating the process for genera, families, orders, classes and phyla, the variables analysed have a nested dependence: if one taxon increases in abundance, all else being equal, it will increase the abundance of all higher taxonomic groups in which it is found. For another example, if every sequence is a novel species, the Shannon diversity of n sequences will be $H = \log(n)$, and the species richness and evenness will be correlated. Failing to account for dependence among phylogenetic variables can increase error rates when performing multiple hypothesis tests.

Phylogenetic variables with clear biological interpretation can facilitate future study design and theory development. Changes in the abundance of a monophyletic clade suggest a heritable trait driving changes in abundance; future experiments can focus on the clade to search for possible functional ecological traits. In macroscopic ecology, theoretical arguments justify the utility of various diversity metrics as proxies for extinction rates, island biogeographic processes, ecosystem stability and conservation goals[45–47]. Theoretical justification of phylogenetic variables connects the analysis of phylogenetic variables (for example, associations between diversity and metadata) with experimental design and biological theory.

Two recently developed methods—phylogenetic isometric log-ratio (PhILR)[44] and phylofactorization[43]—illustrate the challenges of phylogenetic variables analysis. Given the compositional nature of sequence count data[48,49], both methods construct variables through average log ratios of abundances between two clades in the phylogeny. PhILR variables measure the difference between sister clades (Fig. 2c), and phylofactorization iteratively constructs variables measuring the difference between clades separated by edges in the tree (such as those in Fig. 2a,d).

Changes in a PhILR coordinate suggest that a trait is differentiating sister clades, whereas changes in coordinates from phylofactorization suggest that a trait arose along the identified edge. In both methods, significant associations between phylogenetic variables and metadata encourage future work comparing genomes of two clades to search for functional traits. PhILR can be used for the comparison of sister clades (for example, placental mammals to marsupials, or birds to crocodiles), whereas phylofactorization would compare clades separated by edges (for example, birds to non-birds). In the Supplementary Online Tutorial, we illustrate these two methods for phylogenetic variables analysis, show how to construct these variables, and compare them to EdgePCA, a method that performs principal components analysis on variables corresponding to differences in abundance across edges. To illustrate these methods, we analyse a simulated dataset where rRNA gene copy number drives associations with disturbance frequency in soils[50], and interpret the results.

The goal of analysing phylogenetic variables is to identify meaningful directions of change in microbiome data. Whereas principal components analysis identifies major directions/axes of variation in a dataset, phylogenetic variables identify directions of change in microbiome data, which explain variance in community composition and have implications for extinction risk, which organisms to cultivate, which genomes to compare, and more.

Using phylogeny-aware distances

Quantifying the dissimilarity between different species and between different communities comprising these species can facilitate accurate classification of metadata (such as whether a patient has a disease), clustering of samples, and inferences of community function. Trees in forests sequester carbon in wood, whereas grasses do not. Consequently, measures of distance between communities containing trees and those containing grasses may be indicative of differences in the ecosystem function of forests and grasslands. For the microbial world, traits driving ecosystem function are often unknown, yet accurate classification of disease states can have major consequences for human health and when heritable traits analogous to woody biomass underlie habitat associations, incorporating the phylogeny into distance measures can aid classification (Fig. 3). Phylogeny-aware distances translate a dataset (Fig. 3a) into a distance matrix between samples (Fig. 3b), which can be used to classify samples (Fig. 3c).

One of the most widely used methods for phylogeny-aware analysis of microbiome data is the analysis of UniFrac distances between samples[51]. The UniFrac distance is used as a more biologically meaningful distance between communities than standard Euclidean and Bray–Curtis distances. The intuition behind UniFrac, and most phylogeny-aware distances, is that communities containing more phylogenetically distinct species are more different than communities with more closely related species. Incorporating phylogenetic distances along which functional changes occur may better quantify functional differences between communities.

Many extensions of UniFrac have been explored with the aim of controlling statistical artifacts in count data and tuning the importance of abundance in UniFrac distances. If counts are randomly distributed among species, clades with more species will have higher variances in total counts and thus have a greater impact on UniFrac distances than clades with fewer species. To remedy this effect, variance adjusted weighted (VAW)-UniFrac[52] stabilizes the variance of UniFrac distances. VAW-UniFrac was extended by the generalized UniFrac distance[53], which contains a tuneable parameter to increase or decrease the importance of abundance in the distances between communities.

There have been a number of other phylogenetically informed distance metrics, such as Sorensen's index, Rao's D and Rao's H, that differ in how they incorporate evolutionary information[54]. Further, standard statistical techniques, such as linear regression, can be augmented to penalize differences between close relatives[55–57]. The phylogeny is a scaffold for many variables, and can serve as the basis for many useful distance metrics. Which distance(s), of the possible distances, are of interest to a microbiologist?

There are two non-exclusive categories of research objectives for the construction and use of phylogeny-aware distances: improving sample-site classification or visualization, and providing biologically meaningful community differences. Explicit awareness of these research objectives can increase the impact of analyses for phylogenetic distances.

If sample-site classification or visualization is the goal of an experiment, a researcher may be inclined to search through a space of possible distances until finding one that looks the best, irrespective of the biological interpretability of the distance. However, searching too many distances risks dredging the data and presenting statistically significant patterns, which were obtained by testing multiple candidates without proper corrections for multiple hypothesis tests performed. Correcting for such multiple tests will

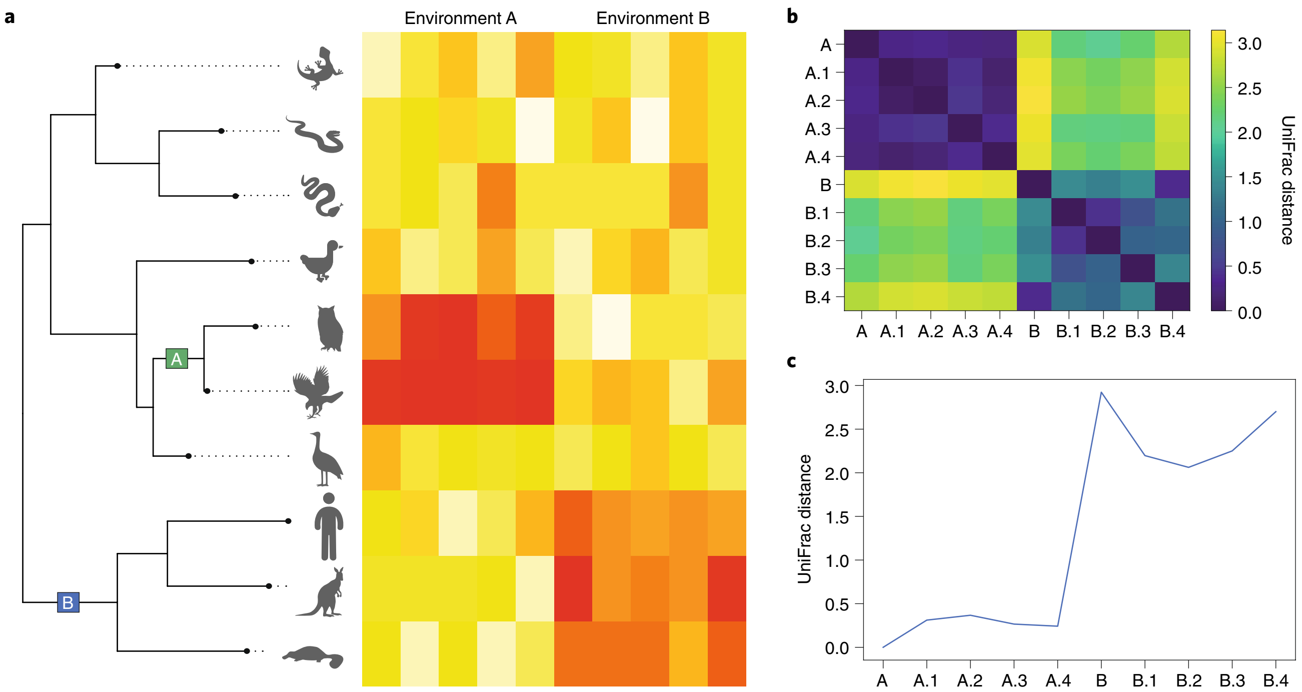


Fig. 3 | Phylogeny-aware distances. a, A heatmap of species abundances with red indicating high abundance and yellow indicating low abundance across different environments. The evolutionary history is represented by the phylogenetic tree, and the main differences between environment A and environment B are being driven by the abundances in clade A and clade B. **b**, While variables contain information for each sample, distances relate two samples. Plotted are the pairwise UniFrac distances between the samples; distances between samples from environment A and samples from environment B are larger compared to distances between samples from environment A or distances between samples from environment B. **c**, The UniFrac distance between a sample from environment A to all other samples illustrates how distances can be useful for sample-site classification. A high UniFrac distance indicates a high degree of phylogenetic novelty or a major change in phylogenetic representation between two communities. Icons (kangaroo, platypus, lizard, outstretched snake, hawk, owl and crane) made by Freepik from www.flaticon.com.

face the same challenges of unclear dependence among tests that arise in the analysis of multiple phylogenetic variables.

While many existing distances can successfully classify and visually separate samples across a range of site categories and clinical variables, the biological implications of discovered differences are often unclear. Does a larger distance indicate greater difficulty in bioremediation of one community into another? Does a larger distance imply a larger difference in ecosystem function or patient morbidity? What follow-up experiments should one conduct to better understand the biochemical and microbiological causes of community differences, given a large UniFrac distance?

Construction of new phylogeny-aware distances and their use in modified statistical methods should consider the performance gains relative to existing methods and whether they provide a new interpretation of discovered differences. Careful justification of new distances can improve the biological interpretation of results. For instance, macroscopic ecologists debate how beta-diversity can be used for conservation[46]. Such discussions can improve the interpretation of existing and newly developed phylogeny-aware distances and help researchers to understand any implications of high or low distances between communities. In addition, a high-quality tree is critical for revealing ecologically relevant patterns[58]. As with PCMs and phylogenetic variables, phylogeny-aware distances benefit from explicit consideration of ecological and evolutionary models to aid the biological interpretation of their results.

Challenges of phylogenetic analysis

There are challenges to phylogenetically structured data analysis, including HGT, the choice of which gene tree to use, the sensitivity to errors in phylogenetic inference, and the explicit consideration of ecological and evolutionary models. Here, we discuss broader challenges of phylogenetic analysis; for challenges especially relevant to microbial and microbiome datasets, see Box 2. HGT between microbial genomes complicates the evolutionary story of vertical transmission captured in a phylogenetic tree[59]. HGT raises the question of which phylogeny to use and how informative the phylogeny is for the research question. For PCMs, HGT can lead to improper corrections and poorly calibrated statistical tests (illustrated in the Supplementary Online Tutorials). HGT of a major trait driving variation in the data can reduce the appropriateness of the phylogenetic variables or distances being used. It is favourable to choose gene families that are insensitive to HGT for inferring phylogenies. Studies have evaluated the chance of HGT based on functional and ecological features[59,60], providing guidelines for this task. Perhaps there is no gene absolutely HGT-free throughout the tree of life, including 16S (ref. [61]). Using multiple genes in phylogenetic inference can minimize the negative impact of HGT[62], and reveal genes influenced by HGT within the selected range of taxa[56]. Computational tools are available for assessing the probability of putative HGT events based on species or gene tree reconciliation[63]. Exploration of genomic context, sequence signature and atypical homology search results also help tracking HGTs[64].

HGT does not invalidate phylogeny-aware analyses of microbiome data. HGT of functional traits could be hypothesized through phylogeny-aware analyses by strong effects with little phylogenetic signal[65]. If phylofactorization identifies an unusually large number of tips of a tree associated with antibiotic exposure, HGT may be driving variation in the data and can be further tested by comparison of genomes among the phylogenetic factors identified. Nonetheless, HGT requires consideration when

Box 2 | Challenges to phylogenetic analysis of microbiome data

Horizontal gene transfer. HGT disrupts the correlation between evolutionary histories of genes and raises important questions about which gene trees to use for phylogenetic analysis. While the 16S gene tree correlates to the bulk of genomic content in microorganisms, important horizontally transmitted genes such as β-lactamases have phylogenies that are different from the 16S. Analysing a β-lactamase gene tree will allow analysis of β-lactamase traits—such trees may be appropriate for studying the composition of antibiotic resistant genes in the environment. We discuss this further in the 'Challenges of phylogenetic analysis' section of the text.

Phylogenetic inference. The 16S rRNA gene tree is most commonly used, but other genes, such as β-lactamase genes, can be used to make phylogenies. Regardless of the gene, phylogenetic inference is an estimate of evolutionary history and the estimate is most accurate with large and even taxon sampling[69]. Uneven taxon sampling can produce erroneous phylogenies resulting in similar traits being misinterpreted as homologies. Phylogenetic reconstruction using skin and skeletal structure of many species of lizards, one species of bird and one species of bat may incorrectly estimate that birds and bats are sister taxa, whereas a more complete sampling of taxa to include mammals may correctly group bats with mammals.

Building trees de novo within each study site, with the limited taxon sampling of each locale, risks producing many erroneous trees that are difficult to compare across studies. Global consensus trees built from commonly used genes, even taxon sampling and standardized methods for adding new sequences to the existing tree can ensure that researchers make comparable inferences on the same, reasonably accurate scaffold of microorganisms' evolutionary history.

Ancestral state reconstruction. As with phylogenetic inference, sparse taxon sampling can increase the error rate of ancestral state reconstruction. Methods such as PICRUSt, which draw on genomes and traits of organisms from relatively well-sampled environments such as the human microbiome, will probably have high error rates for organisms in less well-sampled environments.

Vast number of species. A recent study estimates there to be upwards of a trillion microbial species[70]. While large datasets for macroscopic organisms exist, the regularity of species-rich datasets in microbial ecology and the ease of collecting many samples warrant special consideration of the computational costs, visualization and interpretation of methods often developed for smaller datasets. Parallelization, emphasis on lineages of common knowledge or importance, common knowledge of phylogenetics among microbiologists, and simplified representations of phylogenies by collapsing clades may allow researchers to perform thorough phylogenetic analysis of microbial big data.

Evolutionary model for microorganisms. There is no microbial fossil record. In macroecology, fossil records are used to calibrate evolutionary rates necessary for phylogenetic inference and ancestral state reconstruction[71,72]. While we know that different genes within different species have different mutation rates[73,74], the calibration and validation of evolutionary models for microorganisms is still an open area of research. The correct evolutionary model can produce accurate effect sizes and measurements of uncertainty (significance, confidence intervals, and so on), ensuring accuracy and reproducibility of inferences in phylogenetically structured data analysis.

analysing phylogenetically structured data. The sensitivity of many methods to the horizontal transfer of functional traits is currently understudied.

Finally, all methods face the challenge of being interpretable and advancing our knowledge of microbiological systems. To that end, new methods should explicitly consider ecological and evolutionary models for how traits evolve and drive patterns in the data. One study simulated trait evolution on a tree and compared PGLS with phylogenetic eigenvector regression methods[66], which use eigenvectors from phylogenetic distance matrices as explanatory variables and do not correspond to a clear evolutionary model. The study found that PGLS produced more reliable and better-calibrated statistical results[67]. Considering evolutionary and population genetic models in method development promotes accurate understanding of the assumptions under which a phylogeny-aware analysis performs well, and informs the interpretation of findings in the context of underlying biological processes[68]. The interpretability of phylogenetic analyses is important. As novel and complicated methods are developed, researchers should be aware of the tradeoffs between machine learning and human understanding: the former may produce more accurate predictions in the short term, whereas the latter produces theory that can generate more accurate and generalizable predictions in the long term.

Discussion

The common ancestry of microorganisms can be a source of confounding variation in our data, or a scaffolding on which we make inferences. There are many existing and emerging methods for analysing microbiome datasets in light of evolution, and choosing the right method requires precise statements of the research question (Table 1).

First, decide which tree to use. Commonly, microbiome studies use the 16S tree for Bacteria and Archaea and the 18S tree for microbial eukaryotes, but there is a phylogeny for every gene and some questions are better analysed with trees from other genes. The phylogeny obtained will be an estimate, and uncertainty in phylogenetic inference can translate to uncertainty in downstream phylogenetically structured data analysis.

If the research question uses a trait as a response variable, the phylogeny may be a source of confounding variation. PCMs, such as PGLS, correct for dependence among traits one expects under null models of evolution along the tree.

To study historical trait values, or edges along which major trait differences arose, ancestral state construction is needed. If testing associations between imputed traits, researchers need to combine ancestral state reconstruction for imputation of missing traits with PCMs, which correct for confounding variation.

To simplify patterns of community composition, the phylogeny is a scaffolding that can be used to produce biologically informative variables and directions of change. The choice of variables should be made according to their ability to capture features in data, their statistical dependence, and their biological interpretation.

To distinguish samples of microbial communities, the phylogeny can define distances between samples. By re-defining distances, the phylogeny can be used to modify virtually any statistical method, but the choice of which distance to use should be based on the research goals of sample-site classification or biological interpretation of differences.

Phylogenetic analysis of microbiome data can allow researchers to categorize unclassified microorganisms, test evolutionary hypotheses about trait associations or traits driving habitat associations, and better understand how microbial communities differ and how they change over time, space and treatments. There are several possible classes of methods for analysing microbiome data in light of evolution. Careful consideration of the research question and the ecological and evolutionary assumptions enables researchers to identify existing methods or produce novel methods that address their research question and produce previously undescribed, accurate and biologically informative insights. The deluge of information about microbial sequences is producing phylogenetically structured data which, given the right tools, can accelerate our understanding of microbial community structure and function.

Received: 22 May 2017; Accepted: 27 March 2018;
Published online: 24 May 2018

References

1. Martiny, J. B. H., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: a phylogenetic perspective. *Science* **350**, aac9323 (2015).
2. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
3. Tilman, D. *Resource Competition and Community Structure* (Princeton Univ. Press, Princeton, 1982).
4. MacArthur, R. H. Environmental factors affecting bird species diversity. *Am. Nat.* **98**, 387–397 (1964).
5. May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, Princeton, 2001).
6. Arditi, R. & Ginzburg, L. R. *How Species Interact: Altering the Standard View on Trophic Ecology* (Oxford University Press, Oxford, 2012).
7. Consortium, H. M. P. et al. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
8. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
9. Bardgett, R. D., Freeman, C. & Ostle, N. J. Microbial contributions to climate change through carbon cycle feedbacks. *ISME J.* **2**, 805–814 (2008).
10. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, Oxford, 2000).
11. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* **13**, 303 (2012).
12. Hillis, D. M. & Dixon, M. T. Ribosomal DNA: molecular evolution and phylogenetic inference. *Q. Rev. Biol.* **66**, 411–453 (1991).
13. Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108–110 (1999).
14. Zaneveld, J. R., Lozupone, C., Gordon, J. I. & Knight, R. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res.* **38**, 3869–3879 (2010).
15. Hall, B. G. & Barlow, M. Evolution of the serine β-lactamases: past, present and future. *Drug Resist. Updat.* **7**, 111–123 (2004).
16. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**, 2226–2238 (2002).
17. Větrovský, T. & Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* **8**, e57923 (2013).
18. Lozupone, C. A., Hamady, M., Kelley, S. T. & Knight, R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* **73**, 1576–1585 (2007).
19. Stone, E. A. Why the phylogenetic regression appears robust to tree misspecification. *Syst. Biol.* **60**, 245–260 (2011).
20. Riesenfeld, S. J. & Pollard, K. S. Beyond classification: gene-family phylogenies from shotgun metagenomic reads enable accurate community analysis. *BMC Genomics* **14**, 419 (2013).
21. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
22. Grafen, A. The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **326**, 119–157 (1989).
23. Martins, E. P. & Hansen, T. F. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* **149**, 646–667 (1997).
24. Blomberg, S. P., Lefevre, J. G., Wells, J. A. & Waterhouse, M. Independent contrasts and PGLS regression estimators are equivalent. *Syst. Biol.* **61**, 382–391 (2012).
25. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
26. Blomberg, S. P., Garland, T. Jr, Ives, A. R. & Crespi, B. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**, 717–745 (2003).
27. Lavin, S. R., Karasov, W. H., Ives, A. R., Middleton, K. M. & Garland, T.Jr. Morphometrics of the avian small intestine compared with that of nonflying mammals: a phylogenetic approach. *Physiol. Biochem. Zool.* **81**, 526–550 (2008).
28. Lindenfors, P., Revell, L. J. & Nunn, C. L. Sexual dimorphism in primate aerobic capacity: a phylogenetic test. *J. Evol. Biol.* **23**, 1183–1194 (2010).
29. Garamszegi, L. Z. *Modern Phylogenetic Comparative Methods and their Application in Evolutionary Biology: Concepts and Practice* (Springer, London, 2014).
30. Bradley, P. H., Nayfach, S. & Pollard, K. S. Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. Preprint at https://www.biorxiv.org/content/early/2017/09/16/189795 (2017).
31. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
32. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
33. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
34. Kembel, S. W. et al. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464 (2010).
35. Orme, D. The Caper Package: Comparative Analysis of Phylogenetics and Evolution in R. R Package v.5 (CRAN, 2013).
36. Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E. & Challenger, W. GEIGER: investigating evolutionary radiations. *Bioinformatics* **24**, 129–131 (2007).
37. Tung Ho, Ls & Ané, C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.* **63**, 397–408 (2014).
38. Langille, M. G. I. et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).
39. Cunningham, C. W., Omland, K. E. & Oakley, T. H. Reconstructing ancestral character states: a critical reappraisal. *Trends Ecol. Evol.* **13**, 361–366 (1998).
40. Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T. & Poon, A. F. Y. Ancestral reconstruction. *PLoS Comput. Biol.* **12**, e1004763 (2016).
41. Kuhner, M. K. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459–468 (1994).
42. Joy, J. B., Liang, R. H., Mccloskey, R. M., Nguyen, T. & Art, F. Ancestral reconstruction. *PLoS Comput. Biol.* **112**, e1004763 (2016).
43. Washburne, A. D. et al. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ* **5**, e2969 (2017).
44. Silverman, J. D., Washburne, A. D., Mukherjee, S. & David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* **6**, e21887 (2017).
45. Socolar, J. & Washburne, A. Prey carrying capacity modulates the effect of predation on prey diversity. *Am. Nat.* **186**, 333–347 (2015).
46. McCann, K. S. The diversity-stability debate. *Nature* **405**, 228 (2000).
47. Socolar, J. B., Gilroy, J. J., Kunin, W. E. & Edwards, D. P. How should beta-diversity inform biodiversity conservation? *Trends Ecol. Evol.* **31**, 67–80 (2016).
48. Aitchison, J. *The Statistical Analysis of Compositional Data* (Chapman and Hall, London, 1986).
49. Gloor, G. B. & Reid, G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.* **62**, 692–703 (2016).
50. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* **66**, 1328–1333 (2000).
51. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
52. Chang, Q., Luan, Y. & Sun, F. Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics* **12**, 118 (2011).
53. Chen, J. et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. **28**, 2106–2113 (2012).
54. Swenson, N. G. Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. *PLoS ONE* **6**, e21264 (2011).
55. Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D. & Li, H. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* **14**, 244–258 (2013).
56. Purdom, E. Analysis of a data matrix and a graph: metagenomic data and the phylogenetic tree. *Ann. Appl. Stat.* **5**, 2326–2358 (2011).
57. Fukuyama, J. et al. Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Comput. Biol.* **13**, e1005706 (2017).

58. Hamady, M., Lozupone, C. & Knight, R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.* **4**, 17 (2010).
59. Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* **3**, 679 (2005).
60. Cohen, O., Gophna, U. & Pupko, T. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer research article. **28**, 1481–1489 (2011).
61. Kitahara, K. & Miyazaki, K. Natural and experimental evidence for horizontal gene transfer of 16S rRNA revisiting bacterial phylogeny. **3**, e24210 (2013).
62. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 2304 (2013).
63. Than, C., Ruths, D. & Nakhleh, L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* **9**, 322 (2008).
64. Ravenhall, M., Škunca, N., Lassalle, F. & Dessimoz, C. Inferring horizontal gene transfer. *PLoS Comput. Biol.* **11**, e1004095 (2015).
65. Lozupone, C. A. & Knight, R. Species divergence and the measurement of microbial diversity. *FEMS Microbiol. Rev.* **32**, 557–578 (2008).
66. Diniz-Filho, J. A. F., Sant'Ana, C. E. R. & Bini, L. M. An eigenvector method for estimating phylogenetic inertia. *Evolution* **52**, 1247–1262 (1998).
67. Gloor, G. B. & Reid, G. Compositional analysis: a valid approach to analyze microbiome high throughput sequencing data. *Can. J. Microbiol.* **62**, 692–703 (2016).
68. Freckleton, R. P., Cooper, N. & Jetz, W. Comparative methods as a statistical fix: the dangers of ignoring an evolutionary model. *Am. Nat.* **178**, E10–E17 (2011).
69. Heath, T. A., Hedtke, S. M. & Hillis, D. M. Taxon sampling and the accuracy of phylogenetic analyses. *J. Syst. Evol.* **46**, 239–257 (2008).
70. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl Acad. Sci. USA* **113**, 5970–5975 (2016).
71. Hipsley, C. A. & Müller, J. Beyond fossil calibrations: realities of molecular clock practices in evolutionary biology. *Front. Genet.* **5**, 138 (2014).
72. Forest, F. Calibrating the tree of life: fossils, molecules and evolutionary timescales. *Ann. Bot.* **104**, 789–794 (2009).
73. Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367–372 (1996).
74. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756 (2011).

Acknowledgements

A.D.W. received support from Duke University Biology Department's provision of start-up funds for D. Nemergut (deceased) and the Defense Advanced Research Projects Agency (DARPA) grant D16AP0013. This paper is published in the spirit of D. Nemergut's contagious love of science.

Competing interests

The authors declare no competing interests.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to A.D.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.