The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy

Ahmed Ragab Nabhan and Indra Neil Sarkar

Submitted: 22nd November 2010; Received (in revised form): 25th February 2011

Abstract

Over the past two decades, there has been a long-standing debate about the impact of taxon sampling on phylogenetic inference. Studies have been based on both real and simulated data sets, within actual and theoretical contexts, and using different inference methods, to study the impact of taxon sampling. In some cases, conflicting conclusions have been drawn for the same data set. The main questions explored in studies to date have been about the effects of using sparse data, adding new taxa, including more characters from genome sequences and using different (or concatenated) locus regions. These questions can be reduced to more fundamental ones about the assessment of data quality and the design guidelines of taxon sampling in phylogenetic inference experiments. This review summarizes progress to date in understanding the impact of taxon sampling on the accuracy of phylogenetic analysis.

Keywords: Phylogenetics; taxonomic sampling; bioinformatics

INTRODUCTION

Phylogenetics is the reconstruction of the evolutionary relationships (a ‘phylogeny’) of a group of taxa, such as species [1]. The process of phylogenetic inference generally involves two main phases: first, homologous characters are identified across a given data set (e.g. a sampling of sequences from species) and second, the evolutionary history of organisms is inferred through the comparison of these characters using tree reconstruction methods [2]. The general process for phylogenetic inference is illustrated in Figure 1. There are two major approaches for phylogenetic inference: Distance-based (or ‘phenetic’) approaches, which include methods such as neighbor-joining [3–6] and character-based (which includes methods like maximum parsimony [7–9], maximum likelihood [10, 11] and Bayesian inference [12–14]). Details of these methods have been described by others [15–17]. A set of common terms and phrases used in phylogenetic inference studies is shown in Table 1.

As more phylogenetically informative data become available (e.g. due to high-throughput sequencing), evidence for new relationships among a given set of taxa can emerge and placement of species within clades can change. Incongruent phylogenies of previous studies or availability of new data can thus motivate researchers to formulate new hypotheses about new potential relationships among taxa as determined by subsequent phylogenetic analyses. In pursuit of more complete phylogenetic analyses, researchers often choose the strategy of increasing the sampling of taxa to include for a given study. Taxon sampling can thus greatly influence hypotheses supported by phylogenetic inference. Assessment of phylogenetic accuracy has long been discussed in the phylogenetic inference

Ahmed R. Nabhan is a PhD student in the Department of Computer Science at the University of Vermont, USA, as well as an Assistant Lecturer in the Faculty of Computers and Information, Fayoum University, Egypt.

Indra Neil Sarkar is the Director of Biomedical Informatics in the Center for Clinical and Translational Science, as well as an Assistant Professor in Microbiology and Molecular Genetics and Computer Science at the University of Vermont. His research involves the development of biomedical informatics methods across the entire spectrum of life, from molecules to populations.

Corresponding author. Indra Neil Sarkar, Center for Clinical and Translational Science, 89 Beaumont Avenue, Given Courtyard N309, Burlington, VT 05405, USA. Tel: +1-802-656-8283; Fax: +1-802-656-4589; E-mail: neil.sarkar@uvm.edu

© The Author 2011. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com
literature [26–32]. Criteria for assessing phylogenetic accuracy include [33, 34]: consistency (convergence on the correct tree as available data tend to infinity), efficiency (rate of convergence on the correct tree as more data are included) and robustness (sensitivity to violation of assumptions). Assumptions about the evolutionary process can be violated in real data and therefore can affect accuracy [35, 36]. Examples of these assumptions include [20]: character evolution under a Markov model, symmetric transition probability, as well as identical and independently distributed characters. Taxon sampling can affect the degree to which these types of assumptions are violated. Careful taxon sampling can address some potential problems with real data and can alleviate other problems that can be attributed to inference methods. For example, rapidly evolving sequences are not generally useful to infer phylogenetic relationships. In this case, sampling slowly evolving sequences can produce more informative data sets. Taxon sampling has been an issue for studies concerning both gene trees and species trees. In the case of gene trees, it has been pointed out that genes can be found at different loci and therefore, ‘the effect of taxon sampling can be locus dependent.’ [37] This review focuses on exploring taxonomic sampling issues that concern the phylogenetic inference of species trees with particular emphasis on molecular data. Nonetheless, it is important to note that the approaches and issues raised about taxonomic

**Table 1:** Basic phylogenetic concepts

Bootstrap: A procedure that involves resampling with replacement of the characters of the phylogenetic matrix to reproduce a number of matrices. Phylogenetic trees are then inferred from each resampled phylogenetic matrix. The number of times that a node appears in each of the resampled matrices is the ‘bootstrap value’ of the node [18].

Character: A feature such as form, structure and trait. Character states are the set of values that characters can take (for instance, if eye color is a character, then brown is a state of this character and in molecular sequences a character is an individual site in a sequence and nucleotides G, C, A and T are possible states). Characters can be primitive (e.g. they may already exist before a speciation event occurs) or homologous (see ‘Homologous characters’). The latter type is the one used in phylogenetic analysis studies. A character in a phylogenetic analysis context can be of many types, including: morphological structures, ultrastructural characteristics of cells, biochemical pathways, genes, amino acids and nucleotides [2]. Taxa are represented as sets of sequences of character states.

Homologous characters: Characters that are descended from a common ancestor.

Homoplasy: Having some character states that arise more than once in a given phylogenetic tree [15]. This concept describes the presence of some characters that give conflicting information about relationships among a set of organisms [19].

Inconsistency of estimation: ‘Inconsistency of phylogenetic estimations refers to the property of certain estimation methods to converge on the positively wrong estimate with increasing amounts of data.’ [20]

Leaf stability: Sometimes, leaves that represent taxa can wander between two clades, reflecting instability. This can affect the bootstrap value for an entire branch in the tree. Leaf stability is measured by calculating the occurrences of triplets of taxa in trees computed using an optimality criteria such as likelihood [21, 22].

Long branch attraction (LBA): ‘The erroneous grouping of two or more long branches [in a phylogenetic tree] as sister groups due to methodological artifacts’ [23]. This type of error is common with rapidly evolving taxa.

Phylogenetic matrix: A matrix where rows represent taxa and columns represent characters (bases of the molecular sequences).

Phylogenetic signal: The resemblance of taxa to one another based on the agreement or conflict across data types (characters as defined above).

Phylogenetic tree: A graph that represents the branching patterns of evolution and relationships among organisms.

Phylogenomics: Application of phylogenetic inference on genome-scale data.

Reference phylogeny: A published phylogeny derived via an explicit, reproducible analysis [24].

Tree diameter: Maximum distance between all pairs of taxa in a phylogenetic tree [25].
sampling for species trees can be discussed within context of phylogenetic inference of gene trees.

The present review does not aim to provide a comprehensive overview of phylogenetic inference methods. Instead, the focus of the discussion is on the problem of taxon sampling for its deep impact on the accuracy of phylogenetic inference. This review begins with an overview of the nature of taxa and its implications on the resulting phylogenies. Next, methods for assessing informativeness of taxonomic samples are discussed. A number of important studies are then summarized, with focus on the impact of design principles and methodologies on taxon sampling. Following a discussion of genome-scale approaches to taxon sampling, conclusions are then presented about the current state of the taxon sampling problem in light of phylogenetic inference.

THE NATURE OF MOLECULAR SEQUENCE-BASED TAXA

Taxon sampling of molecular sequence data can be subject to an array of issues. In this section, we discuss three issues that can be observed in molecular data-sets: (i) conflicting signals, (ii) inadequate rate of sequence evolution and (iii) violation of method assumptions. Gaining insight on these types of issues may help identify more appropriate data sets to be used in phylogenetic inference.

Conflicting phylogenetic signals (Table 1) can cause lack of resolution and incongruent phylogenies [38-40]. This problem can arise when a sample of taxa consists of multiple concatenated loci. Inadequate sampling can further magnify the effect of this phenomenon. For the same gene, phylogenetic signals in data sets depend on the specific taxonomic group under study. For instance, samples drawn from 18S rDNA have been found to have conflicting phylogenetic signals for some taxonomic groups [41], while for other taxonomic groups it has been shown that 18S and 28S contain more signal compared to other loci [42, 43]. This problem is relevant for both gene and species trees. It can also be found within a single locus as well as across multiple loci that have been concatenated into ‘super matrices’ [38, 44]. Taking this problem into consideration, it can be crucial to decide on which set of genetic marker(s) are suitable for inclusion in phylogenetic analysis. Such a decision can depend on previous phylogenetic inference studies that describe the behavior of particular genetic markers regarding existing phylogenetic signals. This issue can be addressed in different ways. Rokas and Carroll [45] suggested increasing the number of loci to overwhelm potential conflicting signals. In other cases, one can make use of different types of molecular data (i.e. nucleotide versus amino acid). In cases where nucleotide sequences are difficult to align, it has been shown that amino acid sequences can be used to infer distant phylogenetic relationships [44, 46].

Sampling sequences of genetic markers that evolve at an adequate pace can be important to develop a phylogenetically informative dataset. In this context, Aguinaldo et al. [47] addressed the problem of constructing a phylogeny of nematodes using ribosomal DNA. Given that 18S sequences evolve too rapidly, they used only the slowest evolving sequences. Their results showed significant differences between phylogenies constructed using rapidly or slowly evolving sequences. The merits of Aguinaldo et al.’s approach to guide taxon sampling have been discussed in a number of further works [48-50]. Inspecting taxa for site variability and sequence evolutionary rate has become a necessary step in empirical studies performed on real data and that address evolution history problems. For instance, Evans et al. [51] found that removing a number of sites within molecular sequences can result in changing tree topology and consequently, support a different hypothesis. In particular, they highlighted the importance of model selection and taxon sampling based on the characteristics of taxa. However, sampling more taxa can sometimes introduce a ‘long branch attraction’ (defined in Table 1) problem [52]. This challenge can be addressed by sampling more slowly evolving taxa. With rapid advances of genome sequencing technology and the increasing availability of genetic databases, choosing the best gene(s) and the most informative set of taxa is a major issue in phylogenetic inference problems.

A third issue that can arise is when an observed set of nucleotides in a set of molecular sequences violates fundamental assumptions made by phylogenetic inference methods. To address this scenario, a variety of models for sequence evolution can be embedded in phylogenetic inference methods [10, 53, 54], where each model may have specific defined assumptions (e.g. assuming a uniform distribution of nucleotides). Moreover, there can be variability in evolutionary rates across sites as well as between lineages [55]. Nonetheless, there will likely still be cases
where real data may simply violate one or more assumptions about these models. Assumption violations can occur for gene as well as species trees.

As an example of a violation of model assumptions, consider an inference method that uses a model of sequence evolution and assumes that the nucleotide bases in a given set of molecular sequences are equally likely to occur. For this method, if one uses a taxon sample where sequences have GC bias (i.e. percentage of G and C bases are higher than A and T bases), then the assumption of uniform distribution of characters is clearly violated and a phylogenetic inference method can become statistically inconsistent [56–59]. Some phylogenetic inference methods also make use of relaxed molecular clock models that do not assume a constant rate of molecular evolution across lineages [60–62]. The actual rate of base substitution per site for a given sample can be expected to change every time a single taxon is added or removed. These different models and hypotheses about sequence evolution should be taken into consideration while performing taxon sampling.

QUANTIFYING TAXON INFORMATIVENESS

In the preceding section, discussion was presented about how a given set of taxa can disrupt the expected performance of inference methods in different ways. In this section, the discussion shifts to the problem of finding relevant criteria to select the most informative set of taxa that will lead to the expected performance of inference methods. The lack of objective criteria to select taxa for a phylogenetic inference experiment has motivated researchers to investigate the problem of quantifying taxon informativeness. Indeed, many studies have addressed sampling impact on phylogenetic inference, both analytically and by data simulation, without giving explicit means to quantify the informativeness of characters. Heuristics relating to gene selection have been used (e.g. using genes that evolve at adequate pace); however, such heuristics can perform poorly [63]. Acquisition of new characters that are most informative in terms of resolving polytomes or breaking long branches should be quantitatively measured [63, 64].

Two decades ago, missing data was a major problem for phylogenetic inference and there was a debate about its impact on accuracy [20, 65–70]. One could argue that large molecular data sets can now efficiently be obtained thanks to high-throughput sequencing technologies. However, there are a number of studies that highlight the challenge of determining which combination of loci or taxa are most informative for phylogenomic analysis [2, 38, 45, 71]. In some cases, the molecular sequence content of taxa for a given sample can affect the parameter estimation process of model based approaches. Taxon sampling has been shown to affect the estimation of rate heterogeneity parameters of maximum likelihood models [72]. Inaccurate estimation of phylogenies may be attributed not to the method used, rather to taxon sampling. When a taxon sample contains homoplastic characters, then different characters would support different phylogenies and therefore the support of each phylogeny would depend on which characters are present in the taxon sample [17]. Nonetheless, some studies propose that homoplastic characters are beneficial and can increase the phylogenetic structure [73]. Selection of taxa and characters thus needs some guiding measurements.

Goldman [74] presented a theory to address experimental design issues in phylogenetics, using Fisher information [74–76] to quantify information of sequences with respect to parameters such as branch length or position of internal nodes in trees. Goldman further illustrated how to calculate the amount of increase in information for a single node in a phylogeny as more taxa or more characters are added to a given sample. A second issue that was addressed in that work was the choice of the most informative relative evolutionary rate if a new gene marker is used [74].

Further highlighting the disconcordance between taxon sampling and experimental design is the situation when different conclusions are drawn for the same set of data [38, 65, 68, 77]. Until recently, little attention has been given to developing experimental design criteria for choosing taxa to be included in phylogenetic inference. Geuten et al. [78] highlight this problem:

The lack of general guidelines on how to design a phylogenetic inference experiment has led to a noticeable gap in many research articles: details and discussion on how taxa were chosen are often absent from the Materials and Methods section, the choice perhaps being guided more by intuition than any concrete criterion.
Towards guidelines for designing experiments, they extended the work of Goldman [74] through using the Fisher information matrix [74–76] to compare between different candidate positions for adding a given taxon to a fixed phylogeny. Three different criteria were used for designing experiments based on the amount of information gained when adding a taxon to a phylogeny. The behavior of these criteria was studied under the condition of adding one branch at different positions of a fixed phylogeny. This approach allows for assessing different scenarios when adding new taxa to a sample.

As an example of efforts to quantify phylogenetic informativeness of taxa and how to assess the power of genes to resolve phylogenies, we present the work of Townsend [63] and Townsend and Lopez-Giraldez [64]. They addressed the question of identifying characters that can increase the inferential power and developed a theory to predict the optimal rate of change for a given phylogenetic character [63, 64]. The theory was applied to develop a diagnostic tool for assessing how good specific genes are for resolving phylogenies. Starting with Poisson probability distribution, the theory was developed to profile the amount of information in a set of characters, independent of a specific inference method. To profile the phylogenetic informativeness for a sequence of \( n \) characters, Townsend used the following quantity:

\[
P(T; \lambda_1, \ldots, \lambda_n) = \sum_{i=1}^{n} 16 \lambda_i^2 T e^{-4 \lambda_i T}
\]

To explain this function, assume that a set of taxa forms a polytomy in an arbitrary phylogenetic tree. For this polytomy, \( T \) is the time at which taxa diverge and \( \lambda_1 \ldots \lambda_n \) are the respective character evolutionary rates. The value of \( P(T; \lambda) \) indicates the power of characters 1, 2, \ldots, \( n \) to resolve polytomies at sequential depth of a phylogeny. The character evolutionary rates \( \lambda_1 \ldots \lambda_n \) should be estimated from the taxon sample.

To calculate the overall phylogenetic informativeness of character sets over a specific period of time starting at time \( h_1 \) and ending at time \( h_2 \), the above equation is integrated over the time period specified:

\[
\int_{h_1}^{h_2} \sum_{i=1}^{n} P(T; \lambda) dT
\]

This function integrates over all \( T \) (divergence time at polytomies).

How can this formula be used to compare genes based on the power to resolve phylogenetic trees? For each gene, one needs to calculate evolutionary rates of characters \( (\lambda_1 \ldots \lambda_n) \) of the set of taxa. Then, for the inferred phylogenetic tree, one can use the integral function above to evaluate \( P(T; \lambda) \) over the time period interval (\( h_1 \) up to \( h_2 \)) as indicated by the phylogenetic tree.

Despite continuing efforts to quantify informativeness of taxa, there are no standards or concrete guidelines to ‘engineer’ the taxon sampling process. Gene selection strategies for genome-scale studies are not generally based on informative criteria to indicate which gene is better or which set of genes do not have conflicting signals. In many cases, taxon sampling is still guided by intuition and limited by time and budget constraints. To explore the difficulties in identifying phylogenetically informative taxa, the following sections describe research that involve searching for guidelines for developing good taxon samples. The conditions under which phylogenetic analyses are performed have changed over time. In the years 1990–99, the major concern was the effect of missing data; in the first decade of the twenty first century, the concern was more about making best use of a huge amount of molecular data.

**THE 1990S: A DECADE OF DEBATE**

Most studies that address the taxon sampling problem apply simulation techniques and analytical approaches. The measurements used in these studies are functions of phylogenetic tree properties such as the percentage of correct trees and the percentage of correct branches in association with the number of taxa or characters. A potential issue with such tree-dependent taxon sampling studies can be that there is no consensus about what a ‘correct’ phylogenetic tree might be for a given set of taxa. Another major problem with such studies can be that the informativeness of a set of taxa is not measured according to the taxa only. Instead, the results of phylogenetic inference methods (i.e. a particular phylogeny and its properties) are also used to infer taxonomic informativeness.

A review of the informativeness of genes among vertebrates was done by Graybeal [79], where it was shown that using simple measures of evolutionary rates to make predictions about gene informativeness has some risks. Graybeal further demonstrated the usefulness of using topological and branch length
information of phylogenetic simulation studies. Kim [20] addressed the problem of determining inconsistency conditions for a sample of taxa using mathematical models that were evaluated via data simulation. Parameter conditions that lead to inconsistency of estimation were defined quantitatively. It was shown that long-branches could be a ‘poor predictor of inconsistent conditions.’ Through simulation studies, it was also demonstrated that the probability of inconsistency can increase as more taxa are added when the likelihood of a change in a given long-branch is low. A paradoxical four-taxon sample is an example of when such an inconsistency may occur, even though there are only three possible fully resolved trees (Figure 2). This four-taxon example of inconsistency was used by Kim [20] to show that adding more taxa would not change the inconsistency condition. Kim also warned that adding more taxa can decrease the accuracy of phylogenetic inference and gave a counter example to show that inconsistency can also stem from short-branches. However, a number of authors have since argued that the long-branch attraction effect can be minimized when taxa with slower evolution rates are sampled to subdivide the long branches [52, 67, 80].

In 1998, the effect of taxon sampling on phylogenetic analysis was discussed by a number of authors in a special edition of *Systematic Biology* [81]. To suggest an approach for suitable sampling, Hillis [81] classified studies according to real [82, 83] and simulated [67] data sets, in the light of theoretical considerations [84]. It was clear that there was a controversy about the feasibility of increasing taxon sampling to yield better phylogenetic inference. This controversy may be reduced to the problem of missing data, because during that decade obtaining sufficient data was hard, both for bioinformatics and experimentally driven studies. Under these circumstances, importance was given to the analysis of accuracy of phylogenetic inference based on if more characters were sampled versus if more taxa were added.

Different strategies have been presented for assessing the importance of taxonomic sampling on phylogenetic inference [67, 81, 83, 84]. Graybeal [67]

![Figure 2: A sampling of four-taxon phylogenies. This figure represents examples of topology and branching pattern for trees consisting of four taxa (it does not depict all possible tree arrangements for four taxa): (A), (B) and (C) are three possible trees for a four taxon sample, (D) is an example of long-branch attraction for an unrooted tree and (E) is an example of long-branch attraction for a rooted tree (branch lengths here are proportional to evolutionary changes, which means that rapidly evolving taxa will have long plotted branches).](https://academic.oup.com/bib/article-abstract/13/1/122/219054)
addressed the question of whether it is better to add more characters or more taxa in order to increase phylogenetic accuracy. Data simulation was used to test different scenarios (e.g. increasing taxa versus increasing characters), but the measurements used to quantify the effects were all functions of phylogenetic trees (e.g. number of branches and percentage of correct trees). It was found that increasing both taxa and characters could improve phylogenetic accuracy. By keeping the total number of characters constant, accuracy tends to be higher when characters are distributed over a large number of taxa [67]. Hillis summarized different scenarios (strategies) for taxon sampling and presented relevant works that discussed each scenario. He concluded that taxonomic sampling can have a deep impact on phylogenetic inference and therefore, suggested that systematists (scientists who specialize in phylogenetics) should take into account decisions about taxon sampling and should describe the strategy used to sample taxa [81]. He called for more work on the theoretical level to evaluate competing sampling strategies.

The years 1990–99 witnessed strong emphasis on the use of simulation and analytical studies to explore the impact of taxon sampling on performance of phylogenetic inference methods given limited availability of molecular sequence data. There was no consensus about effect of different taxon sampling strategies on phylogenetic accuracy. Although there was no definite answer to the ‘more taxa or more characters’ question, these studies established strong understanding about the performance of methods under different scenarios of data availability. The conclusions drawn from a number of studies made during that decade were conditioned on the observations that emerged from simulations. As an example, Kim [20] argued that increasing the number of taxa could increase inconsistency conditions. Of course, this depends on the assumptions that a chosen method makes. If the method assumes a uniform nucleotide distribution model for sequence evolution, then increasing the number of taxa can introduce bias of nucleotide content and thus violate the uniformity assumption. But this may also happen while increasing the number of characters. The bottom line is that the effect of taxon sampling depends on the nature of molecular sequences as discussed above, which can also be impacted by the composition of the taxonomic group(s) that are the focus of study.

THE 2000s: YET ANOTHER DECADE OF CONTROVERSY

Throughout the first decade of the new millennium, the controversy about the effect of taxonomic sampling on the accuracy of phylogenetic inference continued. Rosenberg and Kumar [25] highlighted the difference between conditions of taxon sampling for bioinformatics versus experimentally driven contexts. This distinction can be important when evaluating phylogenetic inference performance. In bioinformatics studies, taxon sampling generally involves searching genetic databases for sequences of interest. As a result, reliable sampling can be biased by the availability of relevant sequences. Consequently, this can affect the performance of phylogenetic inference methods. On the other hand, taxon sampling in experimentally driven contexts is planned and not necessarily restricted to contents of genetic database resources (although there can be time and budget constraints). Researchers can choose which taxa and loci (e.g. genes) to add (or remove) for a given taxon sample and thus can guide the sequencing strategy according to knowledge about loci and species (e.g. rate of evolution of the same gene can differ from one clade to the other). The question of whether it is better to add more taxa or more characters to improve phylogenetic accuracy, especially in light of potential budget and time constraints, has long been discussed [25, 45, 66, 67].

Rosenberg and Kumar [65] used a simulation technique to test the effect of taxonomic sample size on accuracy, covering a large set of genes. Starting with a sample that was used to infer some phylogeny, taxa were sub-sampled (i.e. different subsets of the full taxon sample were drawn) to construct other phylogenies. Then, by comparing the resultant phylogenies to a reference phylogeny (Table 1), they found that the phylogenetic error per internal branch was similar for each generated phylogeny. They concluded that incomplete taxon sampling does not degrade accuracy of phylogenetic inference. The simulation study performed by Rosenberg and Kumar was different from the studies presented earlier in this review because their simulation process started with a large phylogenetic tree and then simulated the effect of sampling taxa by taking sub-samples from the larger tree. Earlier, simulation studies did the opposite: they started with the famous paradoxical four-taxon tree and then tested the effect of sampling by adding more taxa to extend the tree [67, 82, 84].
Pollock et al. [68] did a reanalysis of the same data set used in the study by Rosenberg and Kumar [65]. In their study, Pollock et al. removed sequences with extremely high or low base substitution rates and re-examined the relationship between sequence length and phylogenetic error. This led to ‘a sharp decrease in error as more data are added to shorter sequences’ [68]. According to their study, increasing the number of taxa decreases percent error and phylogenetic inference is negligibly affected by substitution rate.

Towards explaining why different conclusions were obtained from the same data such as the above described studies, Zwickl and Hillis [85] suggest five aspects of study design that can lead to different perspectives regarding taxon sampling effects: (i) calculating expected phylogenetic error in randomly selected trees must be taken into consideration when studying taxon sampling, (ii) considering the phylogenetic scope of problems, as represented by tree diameter (Table 1) and average divergence among taxa, d (iii) determining the necessity of performing a more complete tree space search, (iv) simulating different models of sequence evolution (variable complexity) indicate that higher levels of error occur when complex models (as opposed to simpler models such as ones that adopt assumptions like a uniform distribution of bases) of sequence evolution are employed by an inference method and (v) reflecting that sampling studies concern the optimality criterion of a particular phylogenetic inference method.

Hillis et al. [69] followed-up to focus on the taxon sampling problem by addressing two major questions: (i) is phylogenetic error independent of taxon sample size? and (ii) is it better to add more taxa or more characters to the sample? The answer to the first question is no, as described by Hillis et al. [69]:

We all now appear to agree that phylogenetic error is strongly and negatively correlated with taxon sample size and that phylogenetic error is strongly and negatively correlated with character sample size (number of characters examined per taxon).

The answer to the second question is not as simple. Whether to add more taxa or more characters inherently depends on the context of where and how phylogenetic inference is applied. Hillis et al. disagreed with the argument that adding more characters is necessarily a better strategy than more taxa for the same characters.

Rosenberg and Kumar [25] represented their statement about taxon sampling strategies through clarifying the conditions of limited resources when considering sampling additional taxa or characters. They asserted that the relevance of either ‘characters or taxa’ was questionable when performing phylogenetic inference in experimentally driven contexts where sequencing strategies are needed to collect data. As such, an important distinction was drawn between taxon sampling for bioinformatics studies versus for studies that are experimentally driven. With more characters (e.g. nucleotide bases), more accurate phylogenies can be constructed using a variety of methods (distance-based, maximum likelihood, maximum parsimony and Bayesian inference). Using data simulation, the effect of adding more characters was measured while fixing the number of taxa. They found that the percentage of time inter-ordinal branches were reconstructed correctly increases as characters increase [25]. Additionally, a strong positive correlation was identified between the number of characters added and the measure used to assess accuracy, while keeping number of taxa fixed.

The incomplete taxa problem was examined by Wiens [66] from a different perspective. Wiens [66] explored the underlying mechanisms that cause incomplete taxa to be problematic in phylogenetic inference studies by measuring accuracy as a function of ‘the number of shared clades between phylogenies’. The reduced accuracy pertaining to incomplete taxa was shown to be due to missing characters of certain sites (vertically across the sample), not as a result of the absence of some sites from the sample (horizontally by using longer sequences). This implied that increasing the character-per-site quantity meant increasing the coverage for key characters, which is necessary for accurate phylogenetic inference.

Hedtke et al. [39] discussed the utility of taxon sampling in the resolution of phylogenetic conflicts for large data sets. They identified a source of inconsistency that stems from different rates of evolution between genes. Another source of phylogenetic conflict was attributed to branch-length asymmetry. An important conclusion of their study was that high support values for individual clades can reflect robustness, but may also support misleading results because of inconsistency of the chosen method.
Heath et al. [86] performed a comprehensive review on phylogenetic inference accuracy, including an exploration of the effect and strategies of taxon sampling. Notably, they presented a set of strategies for taxon sampling (based on information theory and taxon addition, detecting long-branch attraction, outgroup sampling, ingroup sampling and adding taxa with missing data). The effect of the density of taxon sampling on accuracy and on parameter estimation was described, especially when assumptions about sequence evolution models are violated. They concluded that there are many benefits of dense taxon sampling and advised biologists to appreciate taxon sampling when interpreting phylogenetic analysis results to explain any unexpected patterns. In related work, Lemmon et al. [87] used simulation studies to test the effect of ambiguous taxa on phylogenetic inference. Their study demonstrated a scenario where ambiguous data can affect phylogenetic inference and lead to a wrong conclusion.

Throughout the first decade of the twenty-first century, advances in high-throughput sequencing technologies provided a plethora of opportunities to conduct empirical studies addressing taxon sampling effects. As described above, these studies highlighted many issues about potential effects of taxon sampling on phylogenetic accuracy and consequently about the hypotheses being tested. There is strong evidence that, in general, increased taxon sampling is beneficial to phylogenetic inference accuracy, because it will lead to better estimates for evolutionary rates of characters. But, it should be clear that the nature of molecular taxa is also important to consider (a typical example is rapidly evolving sequences). The addition of more data can actually increase the violation to assumptions of phylogenetic methods (specifically in cases where a parametric approach involves an assumed evolutionary rate). In some cases (e.g. different genetic markers), more data can yield conflicting phylogenetic signal. It thus, remains necessary to study the nature of molecular sequences and the prior knowledge about the relatedness of taxa in taxonomic groups. The common wisdom is to use as many informative (e.g. adequate rate of sequence evolution) and unambiguous data (e.g. exclusion of conflicting signals) as possible.

**GENOME-SCALE APPROACHES**

The increased availability of genomic sequences stimulated a number of studies examining the effect of phylogenetic approaches that made use of genome-scale data (such studies can be referred to as ‘phylogenomic’). Rokas et al. [38] investigated the impact of using genome-scale data on phylogenetic inference. It was found that using a single or small number of loci could increase the potential of supporting conflicting topologies. On the other hand, using a full genome-wide set of concatenated genes can yield a single tree with highest support.

Rokas et al. motivated a number of subsequent studies that incorporated genome-wide data on taxa to enhance phylogenetic inference. Gatesy et al. [77] focused on this approach and started to address a common question: ‘if the availability of genome-wide data for a set of taxa can enhance phylogenetic inference, how many genes should be included to achieve the desired accuracy?’ This question is very important for a specific reason: if a phylogenetic inference is inconsistent (e.g. due to violation of some assumption about model of substitution rate), then no matter how many characters are added to the sample (a whole genome in the extreme case), one cannot guarantee an accurate phylogenetic inference [20]. Using the same data set used by Rokas et al. [38], Gatesy and colleagues did experiments on sub-samplings (analyzing different subsets of the full taxon sample) to show that the set of taxa studied by Rokas et al. were not representative of most phylogenetic inference studies. In fact, the strategy employed by Gatesy et al. (i.e. sub-sampling) to test the generality of conclusions of Rokas et al. resembles the strategy used by Rosenberg and Kumar [65].

Kubatko and Degnan [88] further addressed the notion of using multiple genes for phylogenetic inference. Their approach took into consideration the theoretical result that the concatenation of multiple genes would increase variation in single gene history and would thus lead to inconsistency of the method. Three conditions were identified under which using concatenated sequences of genes collected from multiple loci could lead to poor performance of phylogenetic inference methods. These conditions were identified as [88]: ‘(i) evolution according to phylogenetic and coalescent assumptions, (ii) widespread incomplete lineage sorting (which can lead to a gene tree with a topology different from that of the species tree [89]) due to species tree branch lengths that are short relative to effective population size and (iii) sampling one individual per species.’ It should be noted that inconsistency between gene and species trees could stem from reasons other than
incomplete lineage sorting. Gene duplication, loss and transfer can also lead to gene trees that differ from species trees [17, 90].

Based on the successes of studies demonstrating the feasibility of genome-scale taxa to resolve phylogenetic inference conflicts [38, 91], many studies have made use of genome-scale data to address one of the most fundamental aspirations of phylogenetic studies: to build the tree of life [2, 21, 40, 71, 92]. Towards achieving this goal, Dunn et al. [71] used broad sampling techniques to improve the resolution of phylogeny. Their study resolved some issues where there was conflicting support from earlier studies under conditions of less available data. Stability of taxa was assessed according to leaf stability. Leaves of phylogenetic trees should be stable [i.e. a given leaf (taxon) does not alternate between two clades]. If a leaf alternates between clades, this will affect the overall support values. Leaf stability is measured by calculating the occurrences of triplets of taxa in trees computed using an optimality criteria such as likelihood [21, 22]. Quantitative criteria were used to remove unstable taxa from the sample to get more consistent results. In the context of genome-wide sampling of taxa, De La Torre-Bárcena et al. leveraged an automated approach [93], to investigate the impact of missing data, outgroup choice, number of genes used and partitioning schemes on major seed plant phylogenetics [94].

It was concluded that, when dealing with genomic samplings of data, inadequate taxon sampling could lead to inaccurate phylogenetic inference. Understanding the impact of taxon sampling on tree construction can affect our understanding of the problems associated with assembling the tree of life. The problem of constructing the tree of life in light of various phylogenomic approaches has been discussed extensively in the literature [2, 38, 71, 91, 95]. It has been suggested that using a complete set of genes can help increase the phylogenetic signal (and potentially overwhelm conflicting signals) hence providing a basis for diverse species to be phylogenetically analyzed [91]. However, there are a number of ways in which genome-wide taxon sampling may yield misleading results. For example, lack of resolution in a phylogeny that creates a monophyletic group may be due to missing data, taxon sampling, or an artifact of the method used [40]. The activity of taxon sampling and careful analysis of taxa can also guide the selection of the evolutionary model that suits the taxon sample.

ARRIVING AT A CONSENSUS AFTER TWO DECADES OF CONTROVERSY?

After two decades of controversy, a strong consensus has emerged on the importance of taxon sampling to produce congruent and well-resolved phylogenies. As Baurain et al. [40] suggest, ‘it is no longer worthwhile to argue on the relative benefits of gene versus taxon sampling’ and that efforts should be directed towards developing better models of sequence evolution. Nonetheless, there still remains some debate over whether one should sample more taxa or characters. Agnarsson et al. [96] emphasize the importance of dense taxon sampling relative to accuracy and represent the argument that adding more taxa is as important as adding more characters. Conversely, Lemmon et al. [87] call for efforts to develop robust diagnostics for assessing the effect of ambiguous characters on taxon sampling relative to the accuracy of phylogenetic inference. Havird et al. [97] emphasized that inadequate taxon sampling can make a significant difference regarding phylogenetic accuracy and hence impact the ultimate interpretation of resulting trees.

Studies such as these demonstrate the relative benefits and risks of using multiple loci for species tree construction. However, it remains unclear to what degree multi-gene data violate possible model assumptions or also introduce conflicting phylogenetic signal. More efforts are needed towards the development of (automated) locus selection strategies that ensure minimization of both conflicting phylogenetic signal and violation of possible model assumptions (including those that reflect sequence evolution). As taxonomic sampling can violate (to different degrees) assumptions made by models of sequence evolution employed by phylogenetic inference methods, more efforts are needed to develop more realistic models of sequence evolution. The need for such realistic (and probably complicated) models of sequence evolution is even more required in the case when taxa are composed of multiple gene markers.

CONCLUSION

This review summarized recent knowledge about taxon sampling issues and their relative impact on performance and consistency of phylogenetic analysis, with a particular emphasis on work done in the last two decades. Earlier studies showed…
controversy regarding the effect of adding either taxa or characters on phylogenetic inference performance. Further issues have emerged regarding the use of multiple sequence types (e.g. from multiple loci) in phylogenetic analysis. Mathematical models, in addition to data simulation, are key possible approaches that can help quantify the effects of taxon sampling on a given phylogenetic analysis. In light of recent advances in high-throughput sequencing, taxon sampling will become increasingly important when considering the development of hypotheses that involve phylogenetic inference.

Key Points
- Taxon sampling can play an important role in phylogenetic inference accuracy and thus can also affect the hypothesis being tested.
- Taxon sampling has been studied in a range of scenarios to explore its affect on phylogenetic inference.
- The effect of taxon sampling will be important to understand in light of whole-genome or multi-locus phylogenetic studies.
- Quantification of information in taxonomic samples, independent of a particular phylogenetic method, may be a good approach to addressing challenges of taxon sampling.

FUNDING
The General Missions Program, Egypt (to A.R.N.); the National Institutes of Health (grant R01LM009725 to I.N.S.).

References


