Optimal Selection of Gene and Ingroup Taxon Sampling for Resolving Phylogenetic Relationships

JEFFREY P. TOWNSEND1,2,* AND FRANCESC LOPEZ-GIRALDEZ1

1Department of Ecology and Evolutionary Biology and 2Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA;
*Correspondence to be sent to: Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA;
E-mail: Jeffrey.Townsend@Yale.edu.

Received 29 July 2009; reviews returned 5 October 2009; accepted 19 March 2010

Abstract.—A controversial topic that underlies much of phylogenetic experimental design is the relative utility of increased taxonomic versus character sampling. Conclusions about the relative utility of adding characters or taxa to a current phylogenetic study have subtly hinged upon the appropriateness of the rate of evolution of the characters added for resolution of the phylogeny in question. Clearly, the addition of characters evolving at optimal rates will have much greater impact upon accurate phylogenetic analysis than will the addition of characters with an inappropriate rate of evolution. Development of practical analytical predictions of the asymptotic impact of adding additional taxa would complement computational investigations of the relative utility of these two methods of expanding acquired data. Accordingly, we here formulate a measure of the phylogenetic informativeness of the additional sampling of character states from a new taxon added to the canonical phylogenetic quartet. We derive the optimal rate of evolution for characters assessed in taxa to be sampled and a metric of informativeness based on the rate of evolution of the characters assessed in the new taxon and the distance of the new taxon from the internode of interest. Calculation of the informativeness per base pair of additional character sampling, for included taxa versus additional character sampling for novel taxa can be used to estimate cost-effectiveness and optimal efficiency of phylogenetic experimental design. The approach requires estimation of rates of evolution of individual sites based on an alignment of genes orthologous to those to be sequenced, which may be identified in a well-established clade of sister taxa or of related taxa diverging at a deeper phylogenetic scale. Some approximate idea of the potential phylogenetic relationships of taxa to be sequenced is also desirable, such as may be obtained from ribosomal RNA sequence alone. Application to the solution of recalcitrant unresolved nodes in an otherwise well-known phylogeny is the most obvious application. We validate the theory by analysis of its predictions regarding the phylogenetic informativeness for taxon addition of 46 amino acid alignments of 21 fungal taxa. Gene and taxon sampling according to the theory herein and following a “deepest ingroup” heuristic are shown to provide significantly improved resolution of specified deep internodes.

[Character; cost-effectiveness; experimental design; fungi; information; phylogeny; power.]

A major and long-standing controversy in phylogenetic experimental design is whether phylogenetic accuracy is increased by sampling more characters or more taxa (Kim 1996, 1998; Graybeal 1998; Hillis 1998; Poe 1998; Rannala et al. 1998; Bremer et al. 1999; Sullivan et al. 1999; Berbee et al. 2000; Pollock and Bruno 2000; Rosenberg and Kumar 2001, 2003; Pollock et al. 2002; Zwickl and hillis 2002; Rokas and Carroll 2005; Hedtke et al. 2006). The recent advent of whole-genome sequencing has made the question even more pressing (Dacks and Doolittle 2001; Rokas et al. 2003; Deluc et al. 2005), as analyses of whole-genome sequenced taxa tend to provide some robust phylogenetic results and some controversial results as well as revealing a diverse suite of orthologous genes that could be individually sequenced on a broad taxonomic scale to resolve difficult nodes in the phylogeny. Analytical approaches have been advanced to identify the most phylogenetically informative genes. Goldman (1998) advanced a promising approach toward optimizing inference based on the Fisher information that sums over observed information to predict future information for every possible site pattern that can occur. Townsend (2007) advanced an approach that facilitates the use of data on molecular evolutionary pattern of genes in related taxa to infer the phylogenetic utility for specified epochs. This phylogenetic “informativeness” approach yields a normalized asymptotic probability density for a true synapomorphy occurring in an asymptotically short, deep internode at a time T of a quartet of taxa under an infinite-states Poisson model of character evolution.

Analytical approaches to evaluating the utility of increasing taxonomic sampling would expand the scope of phylogenetic experimental design, but any approach must critically depend on the chronology of ancestral linkages of the historical lineages of the taxa added to the data set (Fiala and Sokal 1985; Huelsenbeck 1991; Kim 1996, 1998; Poe 2003; Debruy 2005). Geuten et al. (2007) derived a variant approach to characterizing the Fisher information to that of Goldman (1998), which characterized the utility of taxon addition to a topology, but did not yet integrate both topology and character evolutionary rate. Such an integrated theory would be especially applicable when one or a few very broadly sequenced loci (e.g., ribosomal DNA, internal transcribed spacer, and mitochondrial DNA) have been sequenced with dense taxonomic sampling, providing some phylogenetic hypotheses without providing significant resolution, and many other sequences have been characterized for just a few taxa. In the taxon versus character sampling debate, however, it has been...
a challenge to place tree topology, rate of evolution of the character, and novel taxon position within a single, evaluative analytical framework within which all calculations are tractable. Quantitative procedures for selection of characters from additional ingroup taxa that exhibit appropriate rates of evolution to resolve soft polytomies are needed.

Here, we demonstrate that the utility of sequencing novel taxa for incorporated characters (characters already present in the data matrix for other taxa) may be quantified. To usefully address the debate, potential phylogenetic utility should be assessed in a common framework both for increasing resolution by novel character sampling in incorporated taxa and for increasing resolution by increased sampling of incorporated characters in a novel taxon. Therefore, we show that the informativeness per base pair of additional character sampling for included taxa versus additional character sampling for novel taxa can be compared given estimated rates of evolution of the relevant characters and an estimate of the time of divergence of the novel taxon from the internode of interest. This quantitative formulation may play a helpful, objective role in resolving the taxon sampling versus character sampling controversy. We hypothesize that combining acquisition of character data for new taxa that branch close to the time of a specified polytomy with acquisition of new characters that are most informative about that time period will yield the greatest phylogenetic resolution. We test our theoretical predictions and validate them by computational analysis of a 21-taxon 46-gene amino acid sequence data set extracted from Marthey et al. (2008).

**THEORY**

### Phylogenetic Informativeness

Consider a quartet phylogeny rooted at a common ancestor at time $T$. When the synapomorphies that are essential to parsimony and likelihood methods of phylogenetic inference are used to select an optimal tree, only a character that changes along an internode between two sister clades (Fig. 1a,b; segments $t_1 + t_2$) will provide nonhomoplasious signal about the actual branching order underlying the polytomy. Moreover, an informative character that changes during the ancestral internode must thereafter remain unchanged during the subsequent evolution of the four taxa. The optimal rate of evolution of a character, $\hat{\lambda}_c$, has been characterized for inference of the relationships within a quartet of taxa diverging at time $T$ as

$$\hat{\lambda}_c = \frac{1}{4T}$$  \hspace{1cm} (1)

(Townsend 2007), a result subsequently arrived at in a different analytical context by Fischer and Steel (2009). Furthermore, the phylogenetic utility of an added character has been characterized as a function of its rate of evolution, $\lambda$, as

$$\rho(T ; \lambda) = 16\lambda^2T e^{-4\lambda T}$$  \hspace{1cm} (2)

(Fischer and Steel 2009). This characterization is an asymptotic result for the case in which the internode approaches zero length. Likelihood methods of inference also depend critically on the same synapomorphies for true phylogenetic signal, but feature probabilistic models that facilitate inference of their presence despite subsequent change. Here, theory is presented to characterize the utility of adding taxa to this canonical phylogenetic quartet and to compare the relative utility of additional taxon sampling and additional character sampling as they are dependent upon the evolutionary rate of the characters to be sampled.

### The Optimal Rate of Change for a Character Added from a Novel Ingroup Taxon that Meets the Quartet along a Subtending Branch

We assess utility by accounting only for patterns of state that provide resolution of the original short, deep internode of interest (labelled $t_1 + t_2$ in Fig. 1) as in Townsend (2007). Taxon addition can be useful in resolving an internode if the quartet is unresolved by the current character state pattern and if the novel ingroup lineage meets the quartet tree along one of the subtending branches. Four distinct topologies exist without labelling the four tips, but including the newly sampled taxon as well as the unsampled outgroup in the topology (Fig. 2). If there were only a single change of character state on the entire 5-taxon topology, the additional taxon sampling would not be informative with regard to the original short, deep internode. In each informative case shown, one change of state occurs on the short, deep internode (providing an apomorphy), whereas another change of state occurs toward the tip of the lineage that is sister to the novel taxon. Thus, the novel taxon can bypass a derived character state to reveal a synapomorphy that resolves the original quartet.

Thus, for the newly sampled state of the novel ingroup taxon to be of use in resolving the original quartet, the newly sampled character from the novel ingroup lineage must have experienced no changes in state since diverging from the original quartet. It would then be identical by descent with the state at the tip of one of the original four taxa. That additional character then could distinguish two pairs of taxa, featuring
Ultrametric trees depicted in Figure 2 are labelled with symbols $T$ and $t_0$ (as in Fig. 1), but also feature assorted internodes labelled $t$. A dot substitutes for a subscript that has been omitted because denoting each subscript will be unnecessary. All such internodes will be short ($t << T$) if the following assumptions apply: 1) the internode $t_0$ is short and deep, 2) the tree is approximately clocklike, and 3) the root lineage meets the original quartet at or near the short, deep internode. In the following derivations, all branches labelled $t$ have negligible effects on probability of informativeness in the sense of Townsend (2007) because they are paired with $T$ in sums or differences. Thus, as a rule, these negligible variables are suppressed.

For changes of state to map onto the quartet as in Figure 2, one or more changes must have occurred on the short internode, with a Poisson-distributed probability of $1 - e^{-h\lambda}$. Furthermore, four subtending lineages must have experienced no changes of state, with probability $e^{-4T\lambda}$. Lastly, the sister lineage to the novel ingroup lineage must have experienced one or more changes of state during $T - \tilde{t}$, with probability $1 - e^{-(T-\tilde{t})\lambda}$. Thus, sampling the character state of a new ingroup lineage would be informative with the probability

$$e^{-4T\lambda}(1 - e^{-h\lambda})(1 - e^{-(T-\tilde{t})\lambda}).$$

(3)

Note that parameterization more typical of maximum likelihood phylogenetic inference does not conceptually disambiguate rate of evolution and branch length; Equation 3 may be rewritten in this notation by substituting $b = T\lambda$, $x_0 = t_0 / T$, and $\tilde{t} = \tilde{t} / T$.

To visualize the implications of Equation 3 with regard to utility for phylogenetic inference, one can standardize $T$ arbitrarily at 1, take $t_0 << T$, and plot contours describing the increasing magnitude of probability (Fig. 3). Optimal phylogenetic utility is predicted at low $\tilde{t}$ and at a rate of change intermediate between 0 and 1. The exact optimal rate of change for a character sampled from a new ingroup lineage maximizes Equation 3. The derivative with respect to $\lambda$ of Equation 3 may be written out, but it is a transcendental function. It thus does not succumb to algebraic solution even via asymptotic analysis in the limit of small $t_0$ and/or $\tilde{t}$. However, numerical evaluation of the peak of Equation 3 (or of the zeroes of its derivative) when $t_0 = T$ and $\tilde{t} = T$ reveals an asymptotic optimum. That optimal rate for sampling known characters from a novel ingroup lineage is

$$\lambda_{\tilde{t}} \approx \frac{1}{(2.2)T}.$$  

(4)

$\lambda_{\tilde{t}}$ is faster than $\lambda_{\tilde{t}}$, the optimal rate of $1/4T$ predicted for novel characters sampled exclusively from the original quartet of lineages (Townsend 2007). This faster optimal rate for additional taxon sampling is consistent with a higher ratio of mutations to branch length in the informative state (Figs. 1 and 2) as well as being consistent with observations that characters
exhibiting higher rates of evolution become more informative for phylogenetic inference with increased taxon sampling (Mooers et al. 1995; Graybeal 1998; Hillis 1998; Heath et al. 2008). Each additional ingroup taxon sampled provides an additional information on the historical states occupied by a character, so the ability of additional taxon sampling to take advantage of faster rates of change should be consistent with intuition.

A more parameterized case for the utility of sampling a new ingroup may be formulated when the internode is long enough such that the two pairs of sister lineages diverge at two different times, \( T_1 \) and \( T_2 \). When the new ingroup derives from one of the two lineages diverging at time \( T_1 \), sampling the state of a character from the new ingroup lineage would be informative with probability

\[
e^{-2(T_1+T_2)\lambda}(1-e^{-t_0\lambda})(1-e^{-(T_1-i)\lambda}). \tag{5}
\]

Below, we refer to the result of this equation as the phylogenetic informativeness for taxon addition (PITA). For a set of \( n \) characters evolving at rates \( \lambda_1, \ldots, \lambda_n \) over branches with times \( T_1, T_2, t_0 \), and \( i \), PITA was calculated as

\[
\sum_{i=1}^{n} e^{-2(T_1+T_2)\lambda_i}(1-e^{-t_0\lambda_i})(1-e^{-(T_1-i)\lambda}). \tag{6}
\]

The calculated PITA was used to rank-order gene–taxon pairs from highest priority to lowest priority for our sampling in our analyses of the large data sets below.

**A Suggested Heuristic: Sampling the Deepest Branching Ingroup Lineages**

From the perspective of experimental design, \( T \) and \( t_0 \) in Equation 3 are properties of history and are in no way under the control of the investigator, leaving potential for optimal design to the selection of characters evolving at the appropriate rate, \( \lambda \), that diverge from the quartet at an optimal time as parameterized by \( i \) (Fig. 3). For small \( i \), it was possible to derive an optimal \( \lambda \) (Eq. 4). Some of the simulations of Poe (2003), as well as the specific findings of Goldman (1998) and statements in Townsend (2007), imply that an appropriate heuristic for taxon sampling would be to prioritize sampling the deepest branching ingroup lineage. It is possible to evaluate this prioritization via analysis of Equation 3. Taking the derivative with respect to \( i \) yields

\[
-\lambda e^{-4\lambda T_1}(1-e^{-t_0\lambda})(e^{-(T_1-i)\lambda}). \tag{7}
\]

This derivative is always negative, Equation 3 always decreases with increasing \( i \), implying decreasing utility for phylogenetic inference. This result suggests a heuristic for selecting taxa for sampling when the goal is to resolve a specified internode at time \( T \). The heuristic is, for any character or gene, to sample the deepest branching ingroup lineage first. Prioritizing via the more sophisticated quantification in Equation 6 can only potentially improve on this heuristic by favoring informative genes over uninformative genes, as it likewise prioritizes the deepest branching ingroup when the \( \lambda_i \) are held constant.

**Comparing the Utility of Sampling States of Novel Characters from Incorporated Taxa to the Utility of Sampling the Character States of Incorporated Characters from Novel Taxa**

For sampling of the states of novel characters from known taxa (Townsend 2007) and for sampling character states of known characters from novel taxa (herein), informativeness has been quantified in an equivalent fashion. Namely, sampling efforts that are informative have been characterized as those that localize a true synapomorphy to a short, deep interior branch. By choosing to sample characters and novel taxa with an optimal evolutionary rate, the likelihood of this outcome and the cost-effectiveness of sampling efforts may be maximized. Moreover, with a common informativeness measure and estimates of the \( \lambda_i \) and \( i \), the 2 divergent strategies for increasing sampling may be directly compared.

A contour plot displaying the decision space for a single character is shown in Figure 4. The contour distinguishes the area of parameter space where sampling...
novel characters from the four taxa of the nominal quartet (equation 3 from Townsend 2007) has four times more informativeness than sampling the character state for a novel taxon (Eq. 5 herein). This “4-fold” contour is the isocline for cost-effectiveness when cost is accounted as per base pair of sequencing; taxon addition to an existing quartet requires assessment of one additional character state, whereas character addition to a quartet requires assessment of four additional character states. The contour demonstrates that the optimal sampling strategy depends critically on the rate of evolution of the character as well as the historical divergence time of the novel taxon. The optimal evolutionary rate for incorporated characters to be sampled from novel taxa (Eq. 4) is faster than the optimal evolutionary rate for novel characters to be sampled from the nominal quartet of taxa (equation 6 from Townsend 2007). The optimal sampling strategy when constrained to sample fast evolving characters is to increase novel taxon sampling. This result is in accord with the conventional wisdom that increasing novel taxon sampling helps to break up long branches and disambiguate convergent evolution. In contrast, the optimal sampling strategy when constrained to slowly evolving characters is to sample more of them for the taxa of the nominal quartet. This result reflects that when homoplasy is not an issue, novel taxon sampling has little potential benefit, and novel character sampling will inevitably provide increased data for resolution of the problem.

To validate the theory developed here, we applied it to a data set of 46 fungal single-copy orthologous amino acid sequences across 21 taxa extracted from the assembly by Marthey et al. (2008). Our three main goals in analyzing the two independent clades represented in this data set are as follows.

First, we demonstrate the utility of the simple heuristic for phylogenetic experimental design derived with Equation 7 above, which was presaged by some of the simulations of Poe (2003) as well as the theory of Goldman (1998), Geuten et al. (2007), and Townsend (2007). This heuristic is that “the optimal additional taxon to select to resolve a deep internode is the taxon that branches off closest to the internode.” This heuristic does not facilitate resolution of the position of the new taxon or support of its placement in the new tree. The placement of the new taxon will likely be poorly supported, and the average support measures for the larger tree may well decline. Rather, the heuristic is designed to maximize the induced difference in support between the optimal tree that is consistent with the true 4-taxon statement and the optimal tree that is inconsistent with the true 4-taxon statement for the original quartet. According to the heuristic, the difference in support will increase the most by addition of the taxon branching closest to the deep internode. Thus, if resolution of that internode without regard to the new taxon, rather than production of a fully resolved tree is the goal, then optimal taxon selection should be of taxa that branch closest to that internode.

Second, we demonstrate that the methods derived herein provide a valid prediction for optimal taxon sampling, that is, that despite the stochasticity and complexity of the phylogenetic process that plagues phylogenetic inference, our metric for the phylogenetic informativeness of taxon addition correlates positively with the degree of additional support for the well-established tree (Hibbett et al. 2007; Aguileta et al. 2008).

Third, we demonstrate that our quantitative metric predicts optimal gene–taxon pairs significantly better than sequencing taxon–gene pairs haphazardly and also better than sequencing genes haphazardly while using the simple deepest ingroup heuristic mentioned above.

**METHODS**

The 46 amino acid alignments were selected from FUNYBASE database (Marthey et al. 2008), a database of fungal single-copy orthologous genes extracted from available fungal genomes sequences. FUNYBASE was constructed with the goal of providing a reliable resource of orthologous gene families to perform comparative and phylogenetic analyses in fungi, such as the one performed herein. We chose all single-copy
orthologous genes present in 28 of 30 species available in FUNYBASE (46 genes total). Stagonospora nodorum and Aspergillus oryzae were excluded due to their weakly supported phylogenetic placement (Aguilera et al. 2008). These 28 species supplied sufficient taxon sampling to estimate a rate of evolution for individual amino acid sites. Amino acid sequences were aligned using MUSCLE v3.6 (Edgar 2004) with default settings. GBlocks v0.91b (Castresana 2000) was used to remove ambiguously aligned positions from the alignments. The minimum number of sequences for a flank position was set to 16 and only sites in which more than half of sequences consisted of gaps were treated as a gap position; otherwise, default settings were applied. The resulting alignments are available in a NEXUS format as Supplementary Data File 1 available from http://www.sysbio.oxfordjournals.org.

To compute the rates of evolution, we specified an evolutionary tree. The concatenated protein sequences (13,082 amino acids) were used to estimate the phylogeny with MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003). We performed a mixed-model Markov chain Monte Carlo (MCMC) analysis including invariant sites and specifying a gamma-shaped rate variation with four rate categories. All parameters were unlinked; thus, the models and parameters were estimated during the analysis separately for each locus. Phylogenetic reconstruction was carried out with 10 independent runs using four MCMC chains and random starting trees of 500,000 generations each. We sampled trees at every 100 generations. The first 100,000 generations were discarded as burn-in, long after the log likelihood reached apparent stationarity.

We obtained a calibrated chronogram by passing the phylogenetic tree with the highest likelihood to the r8s software v1.71 (Sanderson 2003). The tree was calibrated by fixing the split of Debaryomyces hansenii and Candida albicans from the other yeast at 272 My (Miranda et al. 2006). Divergence times were estimated by the penalized likelihood method in r8s v1.71 with a truncated Newton algorithm. The smoothing parameter was set to 0.01. The optimization of the smoothing parameter was obtained following the r8s instructions for cross-validation. Using the 28-species chronogram and the 46-gene alignment, molecular evolutionary rates for each gene at each alignment position were inferred by maximum likelihood by the program Rate4site (Mayrose et al. 2005) with a Jones, Taylor, and Thornton model (Jones et al. 1992).

To validate predictions of the theory, we analyzed 5-taxon sequence alignments for 21 of the 28 species—those belonging to two clades, Saccharomycetes and Pezizomycotina. Saccharomycetes taxa Saccharomyces cerevisiae and Yarrowia lipolytica and Pezizomycotina taxa Neurospora crassa and Coccidioides immitis were maintained as the nominal quartet throughout all analyses. Thus, the 4-taxon internode (I branch; Fig. 2) corresponds to the branch connecting Saccharomycetes and Pezizomycotina species, and T1 and T2 correspond to the ancestral branches common to all lineages within these clades. The new internode generated by sampling the fifth taxon (I branch; Fig. 2) corresponds to that created by addition of one of the Saccharomycetes species: Saccharomyces paradoxus, S. mikatae, S. kudri, S. bayanus, S. castelli, C. glabatra, Ashbya gossypii, Kluveromyces lactis, S. kluyveri, D. hansenii, or C. lusitaniae, or one of the Pezizomycotina species: Chaetomium globosum, Magnaporthe grisea, Trichoderma reesei, Fusarium graminearum, Sclerotinia sclerotiorum, Botrytis cinerea, or Coccidioides immitis.

For each gene–taxon pair sampling, we parameterized Equation 6 with the rates inferred by Rate4site and the length of the new internode (I) to calculate the PITA values (Equation 6) using Mathematica (Wolfram Research, Inc. 2008). We then calculated the improvement (or difference) in support for the 4-taxon internode between each 5-taxon tree containing that internode and the nominal 4-species tree. Two measures of support for the 4-taxon internode were calculated. First, we calculated double the absolute difference in log likelihood (DLL) between the optimal tree that is consistent with the true 4-taxon statement for the original four taxa and the optimal tree that is inconsistent with the true 4-taxon statement for the original four taxa. TREE-PUZZLE v5.2 (Schmidt et al. 2002) provided this support value by calculating log likelihoods for all possible topologies (i.e., 15 topologies for 5 species) under an accurate algorithm using a Whelan and Goldman substitution model (Whelan and Goldman 2001) and equal amino acid proportions. Additionally, in the context of maximum parsimony, the decay index, also known as Bremer support, was calculated. Decay index denotes the number of parsimony steps from the best tree to the next best tree without our branch of interest. As in the maximum likelihood analysis, decay index was calculated using the parsimony steps for all possible topologies in PAUP* (Swofford 2003) with exhaustive search. Differences between our results for the two measures of support based on the two optimality criteria were trivial, therefore only the maximum likelihood results are presented below.

For three cases with two taxa (D. hansenii/C. lusitanae, S. sclerotiorum/B. cinerea, and T. reesei/F. graminearum) and for one case with three taxa (A. gossypii/K. lactis/S. kluveromyces) shared ancestry led to nonindependent observations of support. In these cases, support levels were averaged across the correlated taxa for all results. All alignment manipulation, data parsing, and communication among the software applied were performed in Perl.

**Results**

Our analyses showed that within theSaccharomycetes and the Pezizomycotina, sequences of genes sampled from taxa branching close to the internode of interest tended to have greater DLL than sequences sampled from taxa branching close to the tips of the tree (Fig. 5, panels a–c; Supplementary Tables S1 and S2). For instance, within the Saccharomycetes, additional gene sequences sampled from the four most recently
FIGURE 5. Scatter plot of the DLL between the most likely tree that is consistent with the true 4-taxon statement for the nominal quartet, and the most likely tree that is inconsistent with the true 4-taxon statement for the nominal quartet. Each line corresponds to a suite of orthologues, and each point corresponds to a taxon–gene pair. Points lining up vertically correspond to a diversity of genes associated with a single taxon. Lines connecting data points are colored on a scale depicted in the legend to reflect the sum of the DLLs of the gene across taxa. The y-axis scale was limited to maximize readability while minimizing the number of outlying points; in nearly all cases, the y-axis value of outlying points may be determined by projecting the line to the next time of divergence. a) DLL versus distance from the most recent common ancestor with the most closely related taxon of the nominal quartet within the Saccharomycetes. b) DLL versus distance from the most recent common ancestor with the most closely related taxon of the nominal quartet within the Pezizomycotina. c) Chronogram for the complete data set. d) Scatter plot of the DLL versus PITA within the Saccharomycetes. e) DLL versus PITA within the Pezizomycotina.
diverging taxa provided little DLL for the basal internode on average. In contrast, sequences sampled from the deeper branching lineages *S. castelli*, *A. gossypii*/*K. lactis*/*S. kluyveromyces*, and *D. hansenii*/*C. lusitanea* showed an increasing frequency of a positive DLL, an increasing range of DLL, and an increasing potential to yield high DLL (Fig. 5a). The Pezizomycotina data set was composed of fewer recently derived taxa than was the Saccharomycetes data set, but otherwise, a similar pattern was obtained (Fig. 5b). Analysis of the data set using parsimony and Bremer support yielded strikingly similar results (Supplementary Fig. S1, panels a–c).

Within both the Saccharomycetes and the Pezizomycota, taxon–gene sequences exhibiting high PITA (Equation 6) also tended to have greater DLL than sequences exhibiting low PITA (Fig. 5, panels d and e; Supplementary Tables S3 and S4). Taxon–gene sequences with low PITA exhibited low to negative DLL. Taxon–gene sequences with high PITA exhibited a wider range of DLL and trended toward greater DLL with greater PITA. Comparing the DLL for the 5-taxon topology to the DLL of the base 4-taxon topology for each gene, addition of the putatively optimal taxon improved the DLL by 20% in many cases and in some cases by 50% or more. Analysis using parsimony and Bremer support again yielded strikingly similar results (Supplementary Fig. S1, panels d and e).

Following the optimal taxon sampling criterion led to improved support values across intermediate levels of gene–taxon sampling (Figs. 6a and 7, panels a and b). Within the Saccharomycetes, sampling genes haphazardly following the deepest ingroup heuristic yielded dramatic improvement in DLL for the basal internode compared with wholly haphazard sampling (Fig. 6). Prioritizing sequences featuring highest PITA outperformed the expectation of haphazard deepest ingroup taxon sampling. As noted above, little to no improvement of resolution of the basal internode resulted from sampling the four taxa most recently diverged from *S. cerevisiae*. The relative utility of PITA versus a deepest ingroup or versus a wholly haphazard experimental design remained the same when these tip lineages were excluded from the analysis (Fig. 7a). Within the Pezizomycota, prioritizing sequences from taxon–gene pairs featuring highest PITA outperformed the expectation of haphazard deepest ingroup taxon sampling (Fig. 7b). Here as well, the analysis using parsimony and Bremer support yielded strikingly similar results (Supplementary Figs. S2 and S3).

As expected based on its definition (Equation 5), PITA tended to rank priorities moderately similarly to the deepest ingroup heuristic. To provide a brief comparison, the deepest ingroup heuristic would indicate to sample all 46 genes from the deepest available lineage within the Pezizomycota (*S. sclerotiorum*/*B. cinerea*) first (Fig. 7b). PITA, in moderate contrast, only indicated that 21 of the first 46 sequences added to the matrix should be from *S. sclerotiorum* or *B. cinerea*, whereas 14 of the first 46 sequences should be sequences of genes from *T. reesei* or *F. graminarum* that had already been sequenced in *S. sclerotiorum* or *B. cinerea* and 9 of the first 46 sequences should be sequences of *M. grisea* genes that had already been sequenced in *T. reesei* or *F. graminarum* and *S. sclerotiorum* or *B. cinerea*, and two of the first 46
sequences should be sequences of \textit{C. globosum} genes that had already been sequenced in all the other available independent lineages in the Pezizomycota. Correspondingly, the deepest ingroup heuristic indicates that all the final 46 sequences should be sequences of genes from \textit{C. globosum}, whereas PITA indicated that 26 of the last 46 sequences added to the matrix should be from \textit{C. globosum}, 11 of the last 46 sequences should be from \textit{M. grisea}, 7 of the last 46 sequences should be from \textit{T. reesei} or \textit{F. graminarum}, and only two of the last 46 sequences should be genes from \textit{S. sclerotiorum} or \textit{B. cinerea}. All sequencing of those two genes in any of the lineages in the Pezizomycota was deferred by the PITA ranking until the last 46 sequences.

**DISCUSSION**

We examined the impact of taxon sampling on a canonical 4-taxon case to derive the impact of adding a fifth taxon that branches from the quartet at a given phylogenetic depth. We also derived the optimal rate of evolution for the additional characters to be sampled from that taxon. Our theory provides quantitative characterization of the potential inferential power of adding characters from a novel taxon to the canonical 4-taxon case. It also provides a metric for comparing the relative utility of sampling states of novel characters from known taxa to the utility of sampling the character states of known characters from novel taxa. Our example case—profiling the informativeness of 46 genes for resolving the fungal phylogeny—demonstrated the application of the method and validated its predictions under parsimony and maximum likelihood optimality criteria. Support for the true 4-taxon statement generally increased the most with additional sampling of the deepest ingroup—a heuristic predicted by the theory. Moreover, support for the true 4-taxon statement was predicted even better by the full theory, incorporating the rates of evolution of the characters, than by this heuristic. This theory provides a key component of phylogenetic experimental design and promises to facilitate phylogenetic research that is more cost-effective and less error-prone.

Our results support a deepest ingroup sampling priority for experimental design and are consistent with the examples studied in simulation by Poe (2003) and via theory by Goldman (1998) and Geuten et al. (2007). Although quartets featuring dramatic departures from clocklike evolution (such as quartets in the Felsenstein zone) introduce additional complications (Goldman 1998; Poe 2003; Geuten et al. 2007), the trend in simulation analyses is that deepest ingroups provide the greatest signal. Goldman (1998) and Geuten et al. (2007) advocated application of the likelihood-based Fisher information to phylogenetic experimental design, which in the examples analyzed consistently returned the deepest ingroup as the optimal sampling strategy even for data sets with large numbers of taxa. However, the Fisher information approach as implemented in Goldman (1998) is computationally intensive and does not allow for addition of novel taxa and that of Geuten et al. (2007) did not yet incorporate parameters relating to the rate of character evolution.

Although a limitation of this study is its formal domain over only 4- to 5-taxon phylogenies, the 4-taxon case used to derive the analytical results represents a fundamental unit of the tree graph (Bandelt and Dress 1986) and furnishes a tractable and versatile framework for theoretical study of phylogeny (Felsenstein 1978; Huelsenbeck and Hillis 1993; Gaut and Lewis 1995) with considerable utility for revealing optimal phylogenetic methodology for larger data sets. Results
based on analysis of the 4-taxon case may often be readily extrapolated to trees of more taxa (Cummings et al. 2003; Townsend et al. 2008; Schoch et al. 2009). Nonetheless, the results provided here only demonstrate the validity of the prediction with regard to the addition of a fifth taxon. In particular, it should be noted that after addition of the deepest ingroup branching from a given lineage, further sampling of additional recently derived taxa on that same lineage will be less informative than predicted here because along one subtending lineage of the quartet, the potential confounding mutations (Fig. 2) are not necessarily independent. Therefore, despite the 10–20% or greater improvement in resolution demonstrated by PITA over the deepest ingroup heuristic across the experimental time courses depicted in Figures 7 and 8, simply applying the deepest ingroup heuristic might be sufficient as a current design principle. An important goal will be to develop the theory further for increasingly parameterized and informative application to phylogenies with larger numbers of taxa.

With regard to this goal, several generalizations may be conjectured without detailed theoretical analysis. First, the theory described should apply fairly well to the resolution of any internode whose near neighborhood resembles that of the canonical 4-taxon case (Fig. 2). For instance, additional shallow sampling of lineages diverging at the tips of a quartet will do little to change the dynamics of phylogenetic informativeness. To motivate this claim, consider an asymptotic case of Equation 3. As $t \rightarrow T$, Equation 3 $\rightarrow 0$ and the informativeness of the additional taxon sampling becomes negligible. This negligible informativeness also implies that such sampling of shallow taxa will have little effect on how informative sampling of taxa that diverge deeper in the phylogeny would be (i.e., sampling of taxa for which $t \ll T$).

It is also useful to note the topologies where these calculations would approximate the informativeness of sampling particularly poorly. In particular, dense sampling near the internode of interest, by the converse argument from that constructed in the previous paragraph, will presumably dramatically affect the dynamics of the informativeness of further sampling, such that predictions based upon the 4-taxon subset considered here are unlikely to perform optimally. Based on simulations of such cases, the most useful additional taxon sampling may consist of adding characters from lineages that diverge from the middles of the longest branches (Poe 2003). It would be valuable to analytically characterize the effects of such sampling. Perhaps the effects could be characterized by deriving asymptotic results for situations where taxon sampling is arbitrarily dense. In any case, it will remain essential to simultaneously consider both the tree topology of the taxon addition and the rate of evolution of the characters added to come to any robust conclusions regarding phylogenetic utility.

To provide an assessment of optimal taxon sampling, our theory requires some prior idea of the phylogeny of the taxa to be sequenced. In many cases, where a deep internode is unknown and characterization of potential additional taxa is being contemplated, there will be little to no information about the topological relationship of the ingroup taxa that may be chosen for sampling. However, this lack of resolution is not always the case; consider two extremes. At one extreme, nearly all the phylogeny is already well established, and the goal is to resolve recalcitrant nodes. Many phylogenetic problems fall in this category; numerous deep internodes across the tree of life remain unresolved with many gene sequences applied, even though many clades within them have become resolved with just a few (e.g., eukaryotes: Baldauf et al. 2000, fungi: James et al. 2006, or seed plants: Burleigh and Mathews 2007). In this case, approximate values for the relevant internode ($\bar{t}$) are easy to arrive at based on prior data featuring better or worse support. It is important to realize that any uncertainty will only bring the utility of the sampling strategy toward the haphazard sampling utility in proportion to the degree of uncertainty. In other words, even if the uncertainty in estimating these parameters is high, it would not be cost effective to put aside quantitative efforts at experimental design. The haphazard sampling that would ensue would likely yield considerably worse outcomes. This reasoning is no less true for the opposite extreme case, where there is little to no information about the topological relationships of the ingroup taxa that can be chosen for sampling. To the extent that an expert can predict (even very approximately) the phylogeny that will be revealed, these guidelines for experimental design will assist in achieving sampling that is more cost effective than haphazard sampling and provide more robust results for the research effort invested. Phylogenetic experimental design has always been subject to the issue of de novo taxon and character sampling; formalizing the process by which intuition guides sampling should only improve the yield of intuition.

Formalizing that intuition and quantifying the assumptions underlying phylogenetic experimental design should both improve the power of phylogenetic studies and provide for richer, more thoughtful, and productive discussion of results. We have provided here a theory based on taxon addition to quartets that quantifies the optimal rate of change for addition of characters from new taxa, provides a guide as to the efficacy of additional taxon sampling versus character sampling, and facilitates optimal taxon selection for the resolution of challenging internodes. Furthermore, application of quantitative and analytical approaches should clarify long-standing debates regarding experimental approaches and aids in the future interpretations of the power of data to inform applications of phylogenetic research.

**Supplementary Material**

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.
FUNDING
This work was supported by a postdoctoral fellowship support through the “Beatriu de Pinós” from the Departament d’Universitats, Recerca i Societat de la Informació, Generalitat de Catalunya (ref. 2006-BPA 10081 to F.L.-G.). Additional funding for this research was provided by Yale University.

ACKNOWLEDGMENTS
Thanks to Zheng Wang for providing valuable references, to David Hillis and an anonymous reviewer for helpful comments on the manuscript, and to Associate Editor Cecile Ane for additional comments and for suggesting Figure 3.

REFERENCES


