

# Evaluating Phylogenetic Congruence in the Post-Genomic Era

Jessica W. Leigh<sup>\*1</sup>, François-Joseph Lapointe<sup>2</sup>, Philippe Lopez<sup>3</sup>, and Eric Baptiste<sup>3</sup>

<sup>1</sup>Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand

<sup>2</sup>Département de sciences biologiques, Université de Montréal, Montréal, Québec, Canada

<sup>3</sup>UMR CNRS 7138 Systématique, Adaptation, Evolution, Université Pierre et Marie Curie, Paris, France

\*Corresponding author: E-mail: jleigh@maths.otago.ac.nz.

**Accepted:** 27 May 2011

## Abstract

Congruence is a broadly applied notion in evolutionary biology used to justify multigene phylogeny or phylogenomics, as well as in studies of coevolution, lateral gene transfer, and as evidence for common descent. Existing methods for identifying incongruence or heterogeneity using character data were designed for data sets that are both small and expected to be rarely incongruent. At the same time, methods that assess incongruence using comparison of trees test a null hypothesis of uncorrelated tree structures, which may be inappropriate for phylogenomic studies. As such, they are ill-suited for the growing number of available genome sequences, most of which are from prokaryotes and viruses, either for phylogenomic analysis or for studies of the evolutionary forces and events that have shaped these genomes. Specifically, many existing methods scale poorly with large numbers of genes, cannot accommodate high levels of incongruence, and do not adequately model patterns of missing taxa for different markers. We propose the development of novel incongruence assessment methods suitable for the analysis of the molecular evolution of the vast majority of life and support the investigation of homogeneity of evolutionary process in cases where markers do not share identical tree structures.

**Key words:** incongruence, lateral gene transfer, microbial evolution, phylogenetic networks, phylogenomics.

## A Brief History of Congruence in Evolutionary Biology

Congruence is a central yet polysemic notion in a fundamentally comparative science, such as evolutionary biology. In phylogenetics, analysis of the incongruence of evolutionary histories inferred from different data sets helps to address multiple essential questions. Historically, for a given taxonomic sample, congruence between the organismal phylogeny based on morphological characters and phylogenies of orthologous (single copy) genes was expected to provide no less than “the best evidence for evolution” (Zuckerlandl and Pauling 1965; Penny et al. 1982; Pisani et al. 2007). This application is limited to studies of macroorganisms harboring a sufficient number of morphological and ultrastructural features, but then serves to back up claims in favor of a genealogical relationship, and to erect a meaningful taxonomy (Gilbert and Rossie 2007; Jablonski and Finarelli 2009; Virgilio et al. 2009). For prokaryotes and many microbial eukaryotes, however, this sort of comparison cannot be

achieved, as no organismal tree based on morphological characters can be proposed. Hence, Woese (1987) thoughtfully proposed that congruence between independent gene phylogenies should be used to unravel the real evolutionary history of these organisms. Just as morphological and genetic features provided a cross-validation of phylogenetic inferences, the topological agreement between orthologous gene trees is considered strong independent evidence in favor of shared relationships.

Historically, congruence has also played a decisive role in critical phylogenetic analyses based on multiple markers. As independent data sets for phylogenetic analysis became increasingly available, two camps advocating different strategies for dealing with these data emerged (see de Queiroz et al. 1995; Huelsenbeck et al. 1996; Cunningham 1997; Lavesseur and Lapointe 2001). On one side of the argument, supporters of “taxonomic congruence” (sensu Mickevich 1978), or separate analysis, argued that a particularly strong argument could be made for phylogenetic relationships recovered with independent data. Thus, independent data sets

should be subjected to separate phylogenetic analysis, and the resulting tree topologies should be compared (Swofford 1991; Bull et al. 1993; Huelsenbeck et al. 1994; Miyamoto and Fitch 1995). The results of an analysis based on taxonomic congruence can then be summarized by combining the trees by consensus (de Queiroz 1993; Miyamoto and Fitch 1995). Taxonomic congruence is also at the heart of super-tree-based phylogenomic analyses (Sanderson et al. 1998; Bininda-Emonds 2004; Creevey et al. 2004; Pisani et al. 2007).

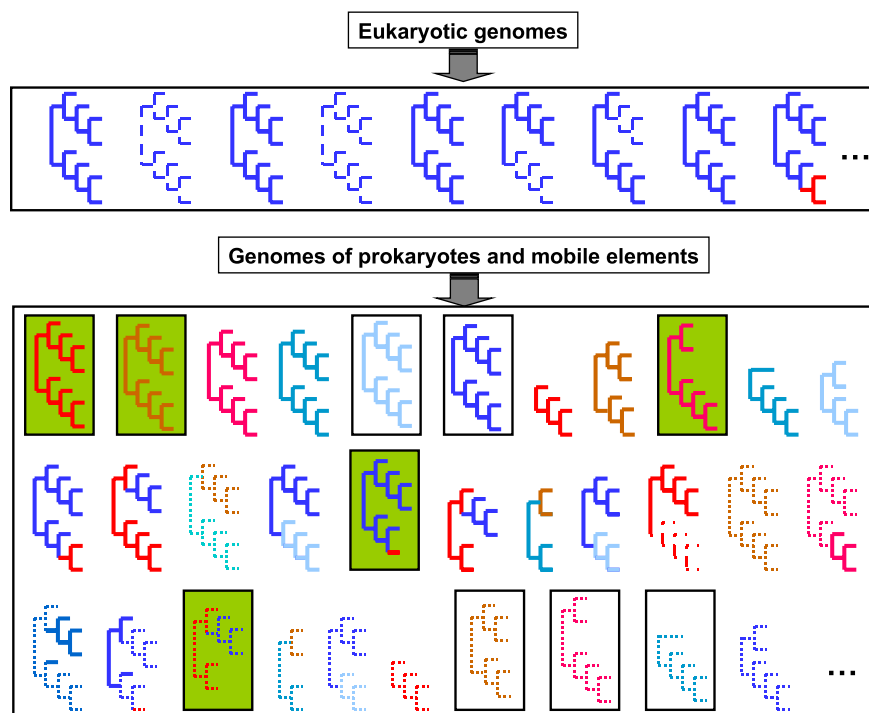
In the other phylogenetic analysis camp, scientists advocating “character congruence,” simultaneous or combined analysis, proposed that the principle of total evidence (sensu Kluge 1989, 1998; Barrett et al. 1991; Kluge and Wolf 1993; Rieppel 2005) should be applied to phylogenetic inference. Total evidence dictates that all available information should be concatenated in a supermatrix (de Queiroz and Gatesy 2007) to reconstruct their common phylogeny (Levasseur and Lapointe 2001). The extent to which characters in a data set are incongruent (i.e., disagree with one another), given the inferred phylogeny, can be assessed via a number of statistics, such as the consistency and retention indices (Kluge and Farris 1969; Farris 1989), as well as a number of statistics that have been developed specifically for partitioned data (Farris et al. 1994; Huelsenbeck and Bull 1996; Waddell et al. 2000).

A third camp of scientists argued that neither the taxonomic congruence nor character congruence method was always the best approach. Instead, they suggested “conditional data combination” (Huelsenbeck and Bull 1996). This strategy involves first testing the data to determine whether they are significantly heterogeneous (i.e., whether they reject the hypothesis that they evolved along the same tree). If the different data appear to be heterogeneous, they are then subjected to separate phylogenetic analysis using a taxonomic congruence approach. As with any taxonomic congruence analysis, the resulting trees are often then either inspected to identify discordant relationships implied by the different trees or tested statistically to evaluate whether they are more similar than expected by chance. If there is no evidence that the data evolved along different trees, they are instead combined using a character congruence approach.

Patterns of incongruence (or conversely, agreement between independent data) have also been extensively used to “expand our knowledge of evolutionary processes.” For instance, comparisons between the tree of hosts on the one hand and the tree of parasites (Hafner and Nadler 1988; Refregier et al. 2008; Wu et al. 2008; Garamszegi 2009) or symbionts (Nelsen and Gargas 2008) on the other hand provide insight about mechanisms of coevolution and about the mode of transmission—vertical or lateral—of symbionts and parasites. Likewise, the agreement between a gene tree and an accepted reference phylogeny—be it a concatenated gene tree (Lerat et al. 2003; Shi and Falkowski 2008), a ribosomal

tree (Shi and Falkowski 2008), or a consensus/supertree phylogeny (MacLeod et al. 2005)—is frequently used to argue that the gene followed the mainstream (accepted or average) evolutionary path (de Andrade Zanotto and Krakauer 2008). By contrast, the disagreement between a gene tree (in the absence of methodological artifact) and a reference phylogeny is frequently used to suggest cases of gene duplication events (Page and Charleston 1997) or lateral gene transfer (LGT; Beiko et al. 2005; Biedler et al. 2007). During this major evolutionary process, a host acquires DNA from a donor, although these two genetic partners are not in an ancestor-descendent relationship. Consequently, LGT can produce branching patterns in the gene tree, incongruent with the reference tree, when donors and hosts are not closest relatives in the reference tree.

Ultimately, although assessment and testing of incongruence is relevant to our understanding of evolutionary processes, most of the tests of incongruence used to date were elaborated on the basis of biological assumptions that are likely no longer valid for most evolving entities (prokaryotic cells or mobile genetic elements) and thus for most genes. The two main reservoirs of genetic diversity, the prokaryotic genomes and the genomes of mobile elements, evolve under much more complex evolutionary processes than was previously assumed. In addition to vertical inheritance (in combination with duplication/loss and variable evolutionary rates), most gene histories are also impacted by rampant LGT and recombination events (Dagan and Martin 2006; Hanage et al. 2006; Fraser et al. 2007; Brill et al. 2008; Dagan et al. 2008; Boucher and Baptiste 2009; Norman et al. 2009). Gene distribution in these genomes results from multiple (often conflicting) selective pressures, so we should not expect 1) that all the genes for a given set of genomes share an identical taxonomic distribution or 2) that they evolved along identical evolutionary histories (tree topologies) (fig. 1). For instance, drug resistance genes are not present in all the same taxa as genes encoding the photosynthetic system because independent and distinct rates of LGT have affected the organismal distribution of the genes coding for these features. Accordingly, as the sequences from genome projects accumulate, molecular data sets become massive and messy, with the majority of gene alignments presenting odd (patchy) taxonomic distributions and conflicting evolutionary histories. Yet tests used in conditional data combination to address the validity of a character congruence approach, which evaluate incongruence of character data, often perform poorly when data are highly heterogeneous. On the other hand, null models used in tests that compare tree topologies in a topological congruence analysis were elaborated on the basis of graph theory and standard statistics that do not reflect actual biological processes valid for most evolving entities (e.g., genes of prokaryotic cells or mobile genetic elements are constrained by some events of vertical



**FIG. 1.**—Scheme of the expected gene tree distributions for eukaryotic versus prokaryotic data sets. Each tree corresponds to an individual gene tree. The color of the tree indicates the phylogenetic history of the gene. Monochromatic gene trees have undergone a given phylogenetic history. Bichromatic trees have evidence of multiple distinct evolutionary histories. Trees, and branches, with similar colors have closer evolutionary histories. Solid trees are strongly resolved; trees with dashed branches are poorly resolved for those branches. Boxes around some trees indicate: 1) gene trees that were frequently transferred horizontally (green-filled boxes) or 2) gene trees that were very rarely transferred horizontally (uncolored boxes). The expected forest of gene trees from eukaryotes is very different—less variable and patchy—from that expected from prokaryotes and mobile elements.

inheritance and, therefore, do not evolve randomly according to independent statistical distributions).

In this review, we clarify the problems met by incongruence analyses in the face of such increasingly numerous data from genomes of prokaryotes and mobile elements. For these data sets, the expected proportion of genes with genuinely discordant evolutionary histories has increased from limited to substantial. Although our intent is not to comprehensively review existing congruence methods, we recall the anatomy of some currently widely used incongruence tests (summarized in table 1) to show how the complex evolution of prokaryotes and mobile elements should affect our methods to detect incongruence. We argue that these tests are only well suited to study the evolution of a minority of taxa and genes, as they lack some important requirements to critically analyze the majority of available phylogenomic data. Using a moderately large prokaryotic multigene data set, we also demonstrate the limited performance of some of the available tests, in terms of long computation time or hard-to-interpret results. Consequently, considering what incongruence analyses ought to do for evolutionary biology in the context of the complexity of molecular data, we propose an alternative—biologically and statistically grounded—theoretical approach for assessing gene incongruence, adapted to

massive and messy molecular data sets, and a notion of congruence based on homogeneity of process, which may be present even when markers have evolved along different true trees.

## Anatomy of Current Incongruence Tests

Statistical approaches to assess incongruence have been devised by both the character and taxonomic congruence communities. For the purpose of character congruence analyses, incongruence is assessed using tests that pose homogeneity as their null hypothesis; that is, that there exists a unique underlying tree and that the differences observed among gene trees are only due to sampling error. Thus, the null hypothesis of homogeneity is evaluated with respect to different randomization of the data using a relevant null model—that is, permutations (Farris et al. 1994), resampling methods (Shimodaira and Hasegawa 1999), or Monte Carlo simulations (Goldman et al. 2000).

Among proponents of the taxonomic congruence approach, another suite of statistical tests for assessing incongruence has been developed. These tests compare differences in tree topologies inferred from independent data

**Table 1**

Characteristics of Popular Congruence Tests

| Test  | H <sub>0</sub>          | Algorithmic Complexity <sup>a</sup> | Identification of Multiple Subsets? | Interpretation of Missing Taxa  |
|---|-------------------------|-------------------------------------|-------------------------------------|---------------------------------|
| MAST (Lapointe and Rissler 2005; de Vienne et al. 2007) | Incongruence            | $O(n)$                              | Yes <sup>b</sup>                    | Pruned and ignored <sup>b</sup> |
| CADM (Campbell et al. 2009)                             | Incongruence            | $O(n^2)$                            | Yes                                 | N/A                             |
| ILD (Farris et al. 1994)                                | Congruence              | $O(n)^c$                            | No                                  | N/A                             |
| Multiple ILD (Planet and Sarkar 2005)                   | Congruence              | $O(n^2)$                            | Yes                                 | Pruned and ignored              |
| LRT (Huelsenbeck and Bull 1996)                         | Congruence              | $O(n)^c$                            | No                                  | N/A                             |
| Concatenator hierarchical LRT (Leigh et al. 2008)       | Congruence              | $O(n^2)$                            | Yes                                 | Pruned and ignored              |
| LRT (Waddell et al. 2000)                               | Congruence              | $O(nm)$                             | No                                  | N/A                             |
| Likelihood-based topology tests                         | Congruence              | $O(nm)$                             | No                                  | Pruned and ignored              |
| Principal component analysis                            | Congruence              | $O(nm)$                             | No                                  | Pruned and ignored              |
| Heatmaps  | Congruence              | $O(nm)$                             | Yes                                 | Pruned and ignored              |
| Likelihood-based topology tests                         | Congruence <sup>d</sup> | $O(nm)$                             | No                                  | N/A                             |

<sup>a</sup> Algorithmic complexity refers to the main phylogenetic analysis and likelihood estimation steps of the tests;  $n$ , number of genes;  $m$ , number of topologies evaluated.

<sup>b</sup> MAST implementations differ. The implementation described by Lapointe and Rissler (2005) can be used to identify congruent subsets of markers and is able to accommodate differences in taxonomic composition among markers; the implementation of de Vienne et al. (2007) does not identify congruent marker subsets and requires that all taxa be represented in all markers.

<sup>c</sup> The Huelsenbeck and Bull (1996) likelihood ratio and ILD (Farris et al. 1994) were described as pairwise tests. Their algorithmic complexity is  $O(n)$  if used as a one-versus-all test, either iteratively or to test a single pair of genes.

<sup>d</sup> Although likelihood-based topology tests are not strictly congruence tests, they have been adapted to this purpose by several authors (e.g., Lerat et al. 2003; Baptiste et al. 2005). The null hypothesis of congruence is normally assessed on a per-gene basis by testing whether the median or global tree is within the confidence set of each gene.

sets, posing heterogeneity or incongruence as the null hypothesis. That is, tests address whether the trees being compared are uncorrelated, and a statistic is used to assess whether these trees are more similar than expected by chance alone (e.g., Lapointe and Legendre 1990, 1992a; Rodrigo et al. 1993; Miyamoto and Fitch 1995). Historically, tables of statistical significance for different tree distance metrics or consensus indices were generated for pairs of random trees (Day 1983; Shao and Rohlf 1983; Shao and Sokal 1986; Steel 1988; Lapointe and Legendre 1992b; Steel and Penny 1993), but recent tests are now based on Monte Carlo simulations, resampling, or permutational approaches.

The computation of character-based tests evaluates whether some of the genes reject a common or global tree. When a statistical test of this class is applied to a set of genes, the rejection of the null hypothesis indicates that these genes are incongruent as a set but does not indicate which (if any) of these genes are not incongruent and which are. The same problem applies to topology-based tests, in which rejection of the null hypothesis indicates that the trees are not altogether incongruent as a set. When many genes are compared, rejecting either one of the null hypotheses thus amounts to saying that at least some of the genes support different topologies (or in the case of tests used in taxonomic congruence, some genes share some patterns of inheritance). As a result, some authors have adapted these tests for this purpose by computing all pairwise comparisons (Planet and Sarkar 2005; Leigh et al. 2008), as described in Box 1. Another approach is to assess the contribution of each gene a posteriori (Campbell et al. 2009).

Within both the taxonomic and character congruence schools, different approaches to measuring incongruence have been developed. The statistical outcome of a given test

### Box 1:

#### Popular character-based congruence methods

Methods for assessing incongruence are sometimes classified as either topological or character-based (for a good review of both classes, see Planet 2006). Topological methods, in which trees are compared directly through statistics such as MAST, are generally used in the fields of phylogeography (e.g., Lapointe and Rissler 2005) and the study of coevolution between parasites and hosts (de Vienne et al. 2007), whereas CADM (Campbell et al. 2011) has also been proposed to test whether multiple trees are more similar than by chance alone. These methods are less relevant to phylogenomics and prokaryote genome evolution than are character-based methods, of which we will summarize some of the most popular. A summary of the more important features of these tests (as they apply to large, whole-genome prokaryotic data) is presented in table 1.

#### The ILD Test

Farris' ILD test (Farris et al. 1994), implemented in the popular phylogeny package PAUP\* (Swofford 2003), is undoubtedly the most highly cited of the incongruence tests. In the ILD, a parsimony tree is estimated for each marker, as well as for the entire concatenated data set. The number of additional steps required for the data under the concatenated tree (compared with marker-

specific trees) is calculated (eq. 1).

$$\text{ILD} = L_{\hat{T}_{AB}}^{AB} - (L_{\hat{T}_A}^A + L_{\hat{T}_B}^B). \quad (1)$$

Here,  $L_{\hat{T}_Y}^X$  indicates the length of the tree estimated from data set  $Y$ , imposed on data set  $X$ . This ILD test statistic is compared with a null distribution produced by repeatedly randomly partitioning sites of the data set to produce re-shuffled markers of the same sizes as the real markers; each time, the ILD is calculated. If the ILD for the true partition of the data set is greater than most of the null distribution, the markers are considered to be significantly incongruent.

### Likelihood Ratio Tests

LRTs for incongruence have been developed by two groups. Huelsenbeck and Bull (1996) proposed a LRT for phylogenetic heterogeneity (incongruence) between markers that is intuitively similar to the ILD. Rather than measuring the number of additional steps when topologies are separately inferred for each marker, they proposed calculating the increase in log-likelihood of the data when each marker is allowed its own topology (compared with the summed log-likelihood over all markers when all are forced to share a single topology; eq. 2).

$$\text{LR}_{AB} = \Lambda_{\hat{T}_A, \hat{\theta}_A}^A + \Lambda_{\hat{T}_B, \hat{\theta}_B}^B - (\Lambda_{\hat{T}_{AB}, \hat{\theta}_A}^A + \Lambda_{\hat{T}_{AB}, \hat{\theta}_B}^B), \quad (2)$$

Here,  $\Lambda_{\hat{T}_Y, \hat{\theta}_Z}^X$  indicates the log-likelihood of data set  $X$  under the topology estimated from data set  $Y$ , with parameters (edge lengths, rates across sites shape parameter, and other aspects of the substitution model) estimated from data set  $Z$ . Huelsenbeck and Bull (1996) proposed assessing the significance of this statistic by generating a null distribution of likelihood ratios from a series of parametric bootstraps. If the likelihood ratio from the real data set is larger than most of the bootstrap replicates, homogeneity is rejected. Their test statistic has been implemented as a pairwise hierarchical test, using instead a nonparametric bootstrap procedure to generate the null distribution, in which sites are sampled from only one of two markers or homogeneous subsets for each replicate (Leigh et al. 2008).

Another LRT was proposed by Waddell et al. (2000). The test statistic proposed by these authors is calculated for each tree,  $\hat{T}_m$ , in a large collection of trees (including at least the ML trees for all markers), and is the sum of likelihood ratios for each marker between the likelihood calculated under the ML tree and the tree in question (eq. 3).

$$\text{LR}_m = \sum_i (\Lambda_{\hat{T}_i, \hat{\theta}_i}^i - \Lambda_{\hat{T}_m, \hat{\theta}_i}^i). \quad (3)$$

The significance of the test statistics is validated through a nonparametric bootstrapping approach or

more quickly using RELL-based bootstrapping of site-wise log-likelihoods calculated under the different trees (Kishino et al. 1990). The bootstrapping involves a centering step, which causes the resampled log-likelihoods for the different trees to conform to a distribution that might be expected if all markers were homogeneous. If likelihood ratios for all trees are significantly larger than the corresponding bootstrap distribution, the null hypothesis is rejected: markers are heterogeneous (incongruent).

### Adapted Likelihood-Based Topology Tests and Data Exploration Methods

One of the most popular likelihood-based methods for assessing incongruence is an adaptation of the Shimodaira-Hasegawa (SH; Shimodaira and Hasegawa 1999) or AU (Shimodaira 2002) topology tests. These tests were designed to assess whether any given tree is a significantly better hypothesis than other trees. When these tests are used to assess incongruence, a pool of trees, normally including at least the ML tree for the entire concatenated data set and the ML gene trees, is evaluated with each marker. Any single marker able to reject the global tree is assumed to be incongruent with the global (vertical) history of the organisms (e.g., Lerat et al. 2003).

$P$  values from AU or SH tests, as well as raw tree likelihoods, have also been used in data set exploration methods. Rather than assessing incongruence via a statistical test that evaluates a probability for the data under a null hypothesis, these methods allow a visualization of various aspects of the data. Brochier et al. (2002) developed a method to assess incongruence by estimating the likelihoods for a pool of tree topologies with a large number of genes. They then used principal component analysis to visualize the genes as a 2D scatter plot, in which they argued that the genes that shared the dominant (vertical) phylogeny formed a cluster, whereas points representing incongruent genes were further away. Baptiste et al. (2005) and Susko et al. (2006) adapted this method, using AU or SH test  $P$  values in the place of raw likelihood values. These authors also proposed an alternative method for visualizing the variation in topological support in the same data. They presented the  $P$  value matrix as a heatmap, in which rows and columns are sorted according to clustering of genes according to their "responses" to trees and clustering of trees according to genes' responses to them. The whole matrix is presented as a color-coded image in which both the phylogenetic strength of individual markers and conflicting patterns of support for different topologies can easily be distinguished.

is likely to be affected by different aspects of the testing procedure, including 1) the test statistics, 2) the number of distinguishable representations of the null hypothesis, and 3) the null model itself (Lapointe 1998). For example, for topology-based tests used in taxonomic congruence, the comparison of trees or their corresponding path-length matrices (distance matrices derived from inferred trees; Campbell et al. 2009, 2011) can be assessed with various consensus indices (Shao and Rohlf 1983; Shao and Sokal 1986), and with a wide selection of tree distance metrics, such as the partition metric (Robinson and Foulds 1981; Penny and Hendy 1985), the nearest-neighbor interchange metric (Waterman and Smith 1978; Křivánek 1986), the subtree pruning and regrafting distance (Bordewich and Semple 2004; Wu 2009), the quartet distance (Estabrook et al. 1985), and maximum agreement subtrees (MAST; Bryant et al. 2003) among others (Steel and Penny 1993). This wealth of measures makes it critical to use different metrics to analyze data sets with different levels of incongruence, as the sensitivity varies among metrics. For example, it is well known that where partition metrics such as the Robinson–Foulds distance suggest that two trees are maximally distant, quartet-based distances may still find similarity (e.g., Adams 1986).

In addition to carefully selecting an appropriate tree distance metric, the population of trees from which random samples are drawn also needs to be defined. For example, the number of rooted trees is larger than the number of unrooted trees (Phipps 1975). Moreover, for the same population of trees, there exist different sampling distributions (e.g., each tree is equally likely [Simberloff et al. 1981] or each branching point is equally likely when growing the tree [Harding 1971; Lapointe and Legendre 1995]). In character-based tests, used to justify a character congruence approach, the phylogenetic inference method (e.g., parsimony [Farris et al. 1994] vs. distances [Zelwer and Daubin 2004]) and randomization method (e.g., nonparametric bootstrapping [Leigh et al. 2008] vs. parametric bootstrapping [Helsenbeck and Bull 1996]) also influences the statistical outcome of the test (see Planet 2006).

At the end of such analyses, the current statistical framework can only determine that some genes are homogeneous or that some trees are incongruent. Such a result (however interesting) does not suffice for researchers interested in the evolutionary mechanisms of prokaryotes and mobile element genomes, for reasons we will discuss below.

### Limits of Current Incongruence Tests for Most Phylogenomic Studies

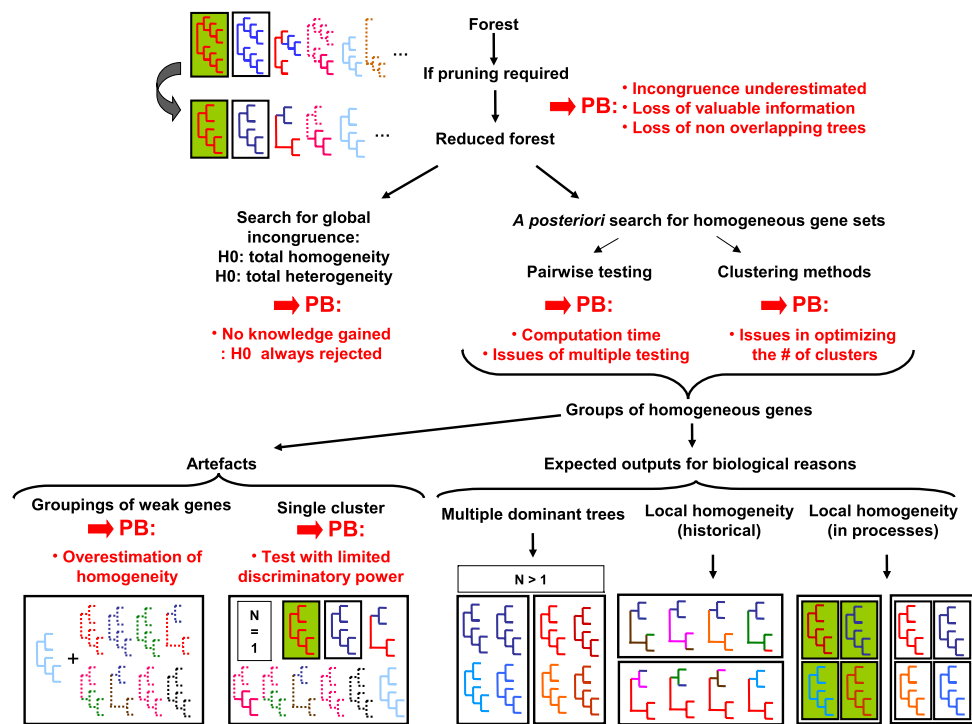
The growing interest in phylogenomic studies based on the large number of whole prokaryotic genome sequences requires a shift in the way we look at incongruence methods. With the expected high level of incongruence resulting from LGT and the increased number of genes available for

phylogenomic analysis, many existing tests have reached their limits for analysis of these data (fig. 2). We examine available tests and evaluate how they handle high levels of incongruence, patchy taxonomic distribution, and whether they perform in computation time that scales well with the size of the data set. Our goal is to stress the need to better take biology into account when designing incongruence analyses (but see Planet [2006] and Box 1 for a more detailed review of existing tests).

### Biological Reality Versus Null Hypothesis

As described above, tests for incongruence involve the postulation of a null hypothesis of either total lack of correlation of divergence patterns between markers (i.e., complete incongruence or heterogeneity) or identical underlying tree topologies among markers (i.e., agreement or homogeneity). The former hypothesis could obviously never reflect biological reality in the case of markers that evolved within the same set of genomes (but see, e.g., Puigbò et al. 2009). Even in the case of extreme LGT, we might expect some proportion of the genome (however small) to have followed a strictly vertical pattern of inheritance in some lineages over some portion of the time since the divergence of some operational taxonomic units in the data set, or at the very least, some markers might have followed the same LGT pattern. That is, some evolution is always homogeneous, and the evolution of genes that “coevolved” in the same genome is therefore correlated, at least in localized regions of the tree. The second null hypothesis, complete topological agreement among markers, likewise does not represent biological reality in most genome data (i.e., in prokaryotes, viruses, and mobile elements). This fact in itself is not necessarily a problem; identification of where and when the hypothesis is false is what makes a statistical test useful.

Another issue in incongruence analyses, exacerbated in presence of many heterogeneous (incongruent) markers, is caused by adaptation of pairwise tests (e.g., Planet and Sarkar 2005; Leigh et al. 2008; see Box 1) to larger data sets. When such statistical tests are repeated iteratively, or significance is assessed only for selectively chosen outliers, significance thresholds should be adjusted (Abdi 2007). If multiple testing corrections become an important aspect of the test, this can lead to either an overly liberal or overly conservative test, depending on the nature of the correction (Leigh et al. 2008). Typically, as the level of correction for multiple comparisons increases, the line between apparent heterogeneity and homogeneity is increasingly blurred. To control against this bias, an option that has not yet been explored for incongruence testing but is widely used in other cases with severe multiple testing problems (such as analysis of microarray data) is the false discovery rate (Benjamini and Hochberg 1995; Storey 2002), where the proportion of expected false positives is



**FIG. 2.**—Pitfalls and possible improvements in incongruence analyses of prokaryotic forests of gene trees. The main steps—and their respective limitations, in red—of most incongruence tests available currently, as described in main text. The color code for gene trees is the same than in figure 1. In the bottom right corner, we suggest some groups of concordant gene trees worth identifying to better analyze forests of prokaryotic gene trees and of mobile elements, which will however require refined incongruence analyses.

used to assess significance. However, in their current form, these tests are probably not appropriate for massive prokaryotic phylogenomic data sets, given the high level of resolution desired (i.e., accurate identification of incongruence at the individual gene level), and the large number of tests needed.

Yet another problem when phylogenetic homogeneity is used as the null hypothesis is that genes often genuinely lack strong phylogenetic signals (fig. 1). Consequently, many incongruence tests (see Box 1) will often fail to reject homogeneity between genes with weak phylogenetic signal and virtually any other gene (fig. 2, lower left; adapted likelihood-based topology tests are particularly sensitive to this problem). This is not to say that incongruence should be assumed even in the absence of evidence: conclusions about the tree-like nature of prokaryotic evolution based on methods that require strong phylogenetic signal should simply be approached with caution.

Although a small number of genes with weak phylogenetic signals may not have substantial adverse effects on a phylogeny inferred from a large number of markers, tests that can only identify incongruence with a reference topology for markers with strong phylogenetic signal (e.g., Lerat et al. 2003) can severely underestimate the level of LGT in prokaryotic data. In the case of adapted likelihood-based topology tests (see Box 1), which are particularly sensitive to this issue, the goal is to identify the “noisy” markers that

do not agree with the reference topology (assumed to be the vertical or species phylogeny). If phylogenomic analysis is the objective, these discordant markers are usually removed from the data set in order to improve resolution of the tree. However, in prokaryotic data, the evolution of genomes is frequently not tree like; in all likelihood, many (if not most) markers have undergone horizontal evolution at some point in their history (Dagan et al. 2008). In addition, some sets of markers may share the same pattern of horizontal acquisition along “gene-sharing highways” (Beiko et al. 2005; Pisani et al. 2007). As such, there may be a series of competing dominant tree topologies underlying the evolution of the data set and identifying which sets of markers share the same tree may be a more interesting (and reasonable) goal than pruning out the suspected few transferred genes.

## Patchy Taxonomic Distribution

In many data sets, the absence of a particular taxon indicates that the data for this taxon were simply not collected. With expressed sequence tag data, for example, the failure to sequence a marker is not necessarily indicative that the marker is not present in the genome of the taxon in question, just that it was not found. In these cases, the absence of a marker is not informative of the evolutionary process,

only of the choices or technical proficiency of scientists or effectiveness of available protocols.

However, the current post-genomic era offers a large number of complete genomes, which introduces another level of complication. When considering the evolutionary process of complete genomes, the absence of a marker for a taxon is actually informative with respect to the evolutionary process (Mira et al. 2010). That is, the absence of a marker indicates that the marker was either lost or gained in one of the two lineages (either as an unrecognizably diverged duplicated gene or through LGT). The presence/absence patterns of genes have indeed been used to study LGT by a number of authors (Lake and Rivera 2004; Rivera and Lake 2004; McInerney and Wilkinson 2005; Dagan et al. 2008), demonstrating the informative nature of missing data in truly phylogenomic data sets. Thus, existing tests for incongruence, which consider only the taxa that are shared between markers, fail to account for important evidence of heterogeneity between gene trees in prokaryotic data sets.

Consider, for example, the trees in figure 3. The marker whose tree appears in figure 3a is present in all taxa in the data set. The tree in figure 3b, however, has a taxonomic distribution that clearly indicates LGT: although the Eubacteria in the tree all fall within a single clan (i.e., there is a split that separates Eubacteria from Archaeobacteria), the presence of this marker in the genomes of only three members of Archaeobacteria strongly suggests that this tree represents a marker that was acquired by these taxa through LGT from a eubacterium. In figure 3c, though, where only Eubacteria are represented, it might be more plausible that the marker simply appeared in the ancestor of Eubacteria included in the analysis. It is not altogether clear whether the markers in figure 3a and c should be considered to disagree. We would say that they agree over a portion of their history

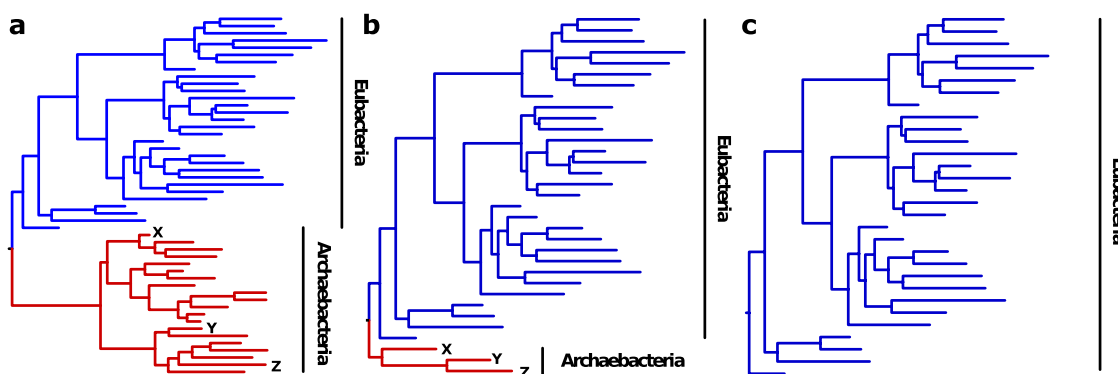
or are “locally homogeneous.” In any case, the interpretation of patchy distributions of taxa between markers should affect an assessment of incongruence in data sets based on complete genome sequences.

## Data set Size and Efficient Scaling

The quantity of data also highlights an unfortunate shortcoming of current incongruence analysis methods. The ever-growing sequence databases have made possible the move away from single-gene phylogeny in favor of phylogenomics, as well as leading to the recognition of the importance of horizontal evolution in shaping genomes. However, with more data comes a need for more efficient algorithms, and the last decade has seen the publication of a number of more efficient sophisticated phylogenetic analysis methods (e.g., Guindon and Gascuel 2003; Stamatakis 2006; Zwickl 2006; Lartillot et al. 2009; de Koning et al. 2010).

Still, increased data set size can pose a problem for congruence tests that involve pairwise comparison (Planet and Sarkar 2005; Leigh et al. 2008). If an exhaustive pairwise approach is used, the time to test all pairs increases with the square of the number of markers in the analysis. When phylogenetic analysis is involved in the pairwise analysis, the computation time can quickly become intractable (e.g., Leigh et al. 2008; see also below) as data sets grow to hundreds or even thousands of markers (table 2).

There are a number of “workarounds” to extend the workable data set size. Parallelization can be used effectively, particularly for independent phylogenetic analysis steps. Sometimes heuristics or short cuts, such as employing a faster phylogenetic analysis method to infer gene histories used for comparison, can be employed to decrease computation time, although this can decrease the performance of the test. Although the power of computational resources is constantly



**FIG. 3.**—Patchy taxonomic distributions and incongruence. In some cases, markers may appear homogeneous when only taxa appearing in both markers are considered when their true histories are clearly incongruent. In (a), all taxa in the analysis are present; (b) only a few members of one clan are present; (c) members of one clan are completely absent. It is highly unlikely that the patchy presence of marker (b) among Archaeobacteria can be explained by differential loss; it is more plausible that this marker was transferred from Eubacteria, then subsequently among archaeobacterial lineages. Thus, although there is a split separating archaeobacterial and Eubacterial lineages, the history of marker (b) is incongruent with that of marker (a). In the case of marker (c), its complete absence from Archaeobacteria suggests its emergence in Eubacteria following their divergence from Archaeobacteria.



**Table 2**

Summary of Results and Computation Time for Popular Congruence Tests with the NUTs Data set

| Test  | P Value            | Computation Time   | Number of Cores | Total CPU Time <sup>a</sup> |
|---|--------------------|--------------------|-----------------|-----------------------------|
| CADM (Campbell et al. 2009) <sup>b</sup>        | <0.001             | 1.3 h <sup>c</sup> | 1               | 1.3 h                       |
| ILD (Farris et al. 1994)                        | ≤0.01              | 14 days            | 1               | 14 days                     |
| LRT (Huelsenbeck and Bull 1996)                 | <0.01              | 5.5 days           | 16              | 88 days                     |
| Concaterpillar (Leigh et al. 2008) <sup>d</sup> | $1 \times 10^{-6}$ | 6.5 days           | 16              | 104 days                    |
| LRT (Waddell et al. 2000)                       | <0.001             | 12.5 h             | 16              | 9 days                      |

<sup>a</sup> Calculated as total computation time × number of cores used in parallel.<sup>b</sup> CADM's null hypothesis is incongruence.<sup>c</sup> Computation time for CADM includes time for distance matrix estimation (1 h 12 min). The time for CADM alone was less than 10 min.<sup>d</sup> Values given for Concaterpillar are for the point at which congruence was rejected and for the pruned (41-taxon) data set. This was the same data set used for other methods.

increasing, the development of tests that scale roughly linearly with data set size (i.e., tests that require only a single phylogenetic analysis step for each marker) is going to be increasingly important as data sets continue to grow (table 1).

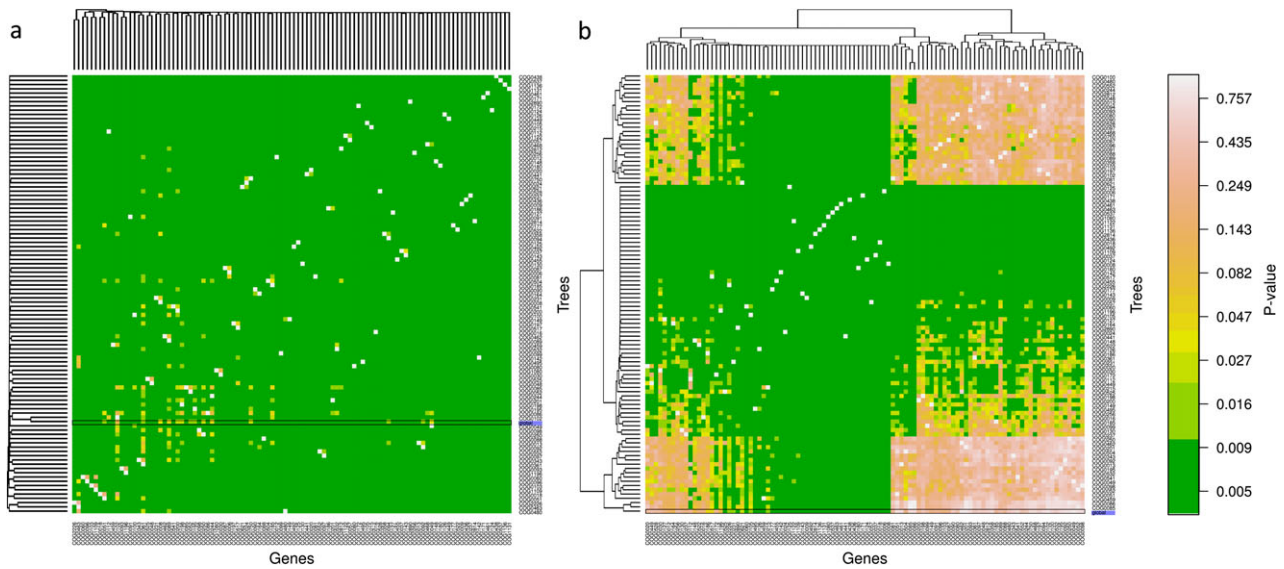
### Application of Existing Incongruence Tests to a Prokaryotic Multigene Data set

We evaluated the performance of a number of methods to assess incongruence in the “nearly universal trees” (NUTs) data set of Puigbò et al. (2009) to illustrate the various limits of analyses of incongruence for real data. The NUTs are a set of 102 amino acid markers for which at least 93 of the 100 taxa in their data set are represented in each marker, composed of 59 Eubacteria and 41 Archaeobacteria. We applied the incongruence length difference (ILD) test (Farris et al. 1994), two different likelihood ratio tests (LRTs; Huelsenbeck and Bull 1996; Waddell et al. 2000), the congruence among distance matrices (CADM) test (Campbell et al. 2011), Concaterpillar (Leigh et al. 2008), and the Approximately Unbiased (AU) (Shimodaira 2002) and SH (Shimodaira and Hasegawa 1999) likelihood-based topology tests. Because some of these tests require that all markers contain the same taxa, taxa missing for any marker were removed from the data set for all tests, leaving a total of 41 taxa. A second analysis of the data set with no taxa removed was performed using Concaterpillar, which can accommodate missing taxa by pruning them from markers as necessary during pairwise comparisons; when the algorithm fails to reject homogeneity of a pair of markers, the alignments are combined as a supermatrix in which taxa present in either marker are included.

The ILD test was performed using PAUP\* (Swofford 2003) with default parameters, except that only 100 repartitioning replicates were used to construct the null distribution. For LRTs, the AU test, and Concaterpillar, likelihoods and trees were calculated using RAxML (Stamatakis 2006) with the WAG (Whelan and Goldman 2001) +  $\Gamma$  model. Single-gene topologies, as well as the global tree inferred from the concatenated data set, were used in both the Waddell LRT and the AU test. The null distribution used to assess significance of Waddell's LRT statistics was produced using 1,000 RELL

bootstrap replicates (Kishino et al. 1990). For the Huelsenbeck and Bull LRT, significance was assessed from null distributions produced using two different methods: first, parametric bootstrapping was used, as recommended by the authors; second, the repartitioning method used by the ILD test was used. In both cases, null distributions were produced from 100 replicates. The CADM global test and a posteriori tests were performed in *R*, using the APE package (Paradis et al. 2004), with 999 permutations. Table 2 summarizes *P* values and computation times for all methods used.

Most of the incongruence tests agreed that genes within the NUTs had significantly different histories—a result that conflicts with the conclusion of Puigbò et al. (2009) that inheritance was generally vertical but is coherent with some of their results, as well as with Puigbò et al. (2010). For example, figure 4a shows the heatmap produced from the AU test *P* values. This plot associates colors with *P* values: dark green shows that a topology was rejected at  $P < 0.01$ . The large number of cells colored dark green indicates that most topologies were rejected by most markers. Even the global tree was rejected at  $P < 0.05$  by all but a single alignment; at  $P < 0.01$ , 15 markers did not reject this topology. These results suggest that the individual alignments in this data set are reasonably strong phylogenetic markers. Because the topologies tested were maximum likelihood (ML) trees for the individual markers, the rejection of most topologies by most other markers likely indicates a high level of pairwise incongruence. For comparison, we also produced a heatmap from the SH test *P* values (fig. 4b). The SH test is more conservative than the AU test, and predictably, many data sets rejected fewer tree topologies than with the AU test. However, even with the SH test, there were a number of markers that rejected nearly all topologies; these markers correspond to the nearly all-dark green columns toward the middle of this plot. The tree topologies that were not rejected by these markers, corresponding to white cells, are the gene trees for each of these markers, and were nearly always rejected by all other markers. This result suggests that these markers in particular are both highly incongruent to others and are strong phylogenetic markers.



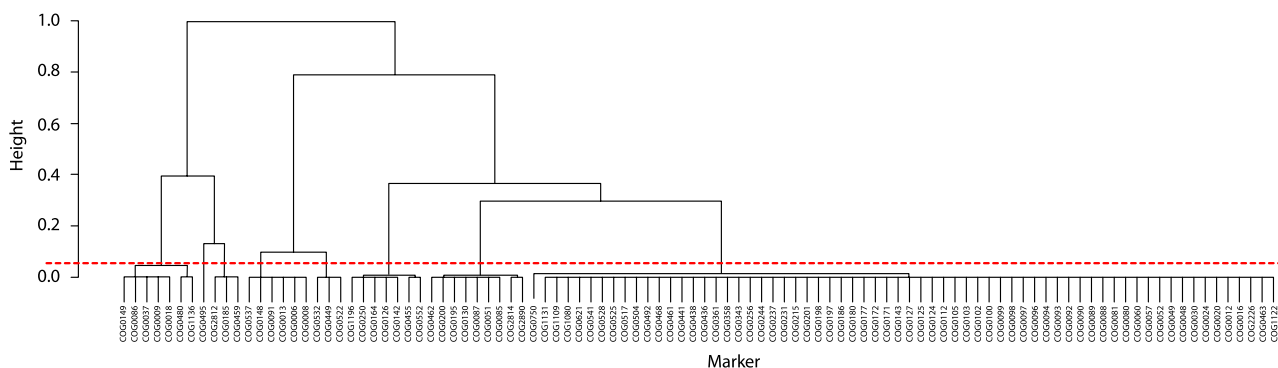
**FIG. 4.**—Heatmap showing AU and SH test results with NUTs and their gene trees. The AU and SH tests were used to assess the support of each marker in the 100-gene NUTs data set for the ML gene trees in the data set, as well as the global tree inferred by ML from the concatenated data set. (a) AU test  $P$  values and (b) SH test  $P$  values. Each row represents an individual tree topology, whereas each column represents an individual marker. Names of markers and trees corresponding to each row and column are indicated; the row corresponding to the global tree is indicated by the blue-highlighted name “global” and by a box around the row of cells. Rows and columns are sorted according to dendrograms above and to the left of the heatmap, which indicate similarity in patterns of  $P$  values. The cells of the heatmaps are themselves colored according to the  $P$  values from the AU or SH test, such that very small  $P$  values (indicating rejection of a particular tree topology with a particular gene) are shown in darker green shades, whereas larger  $P$  values are shown in yellow, orange, or white.

Likewise, all methods except CADM rejected homogeneity of the NUTs data set (table 2). The CADM global test rejected incongruence, indicating that at least one pair of markers is not completely incongruent (i.e., shared at least some local pattern of evolutionary relationships). A posteriori test were thus computed to detect which markers were not completely heterogeneous.

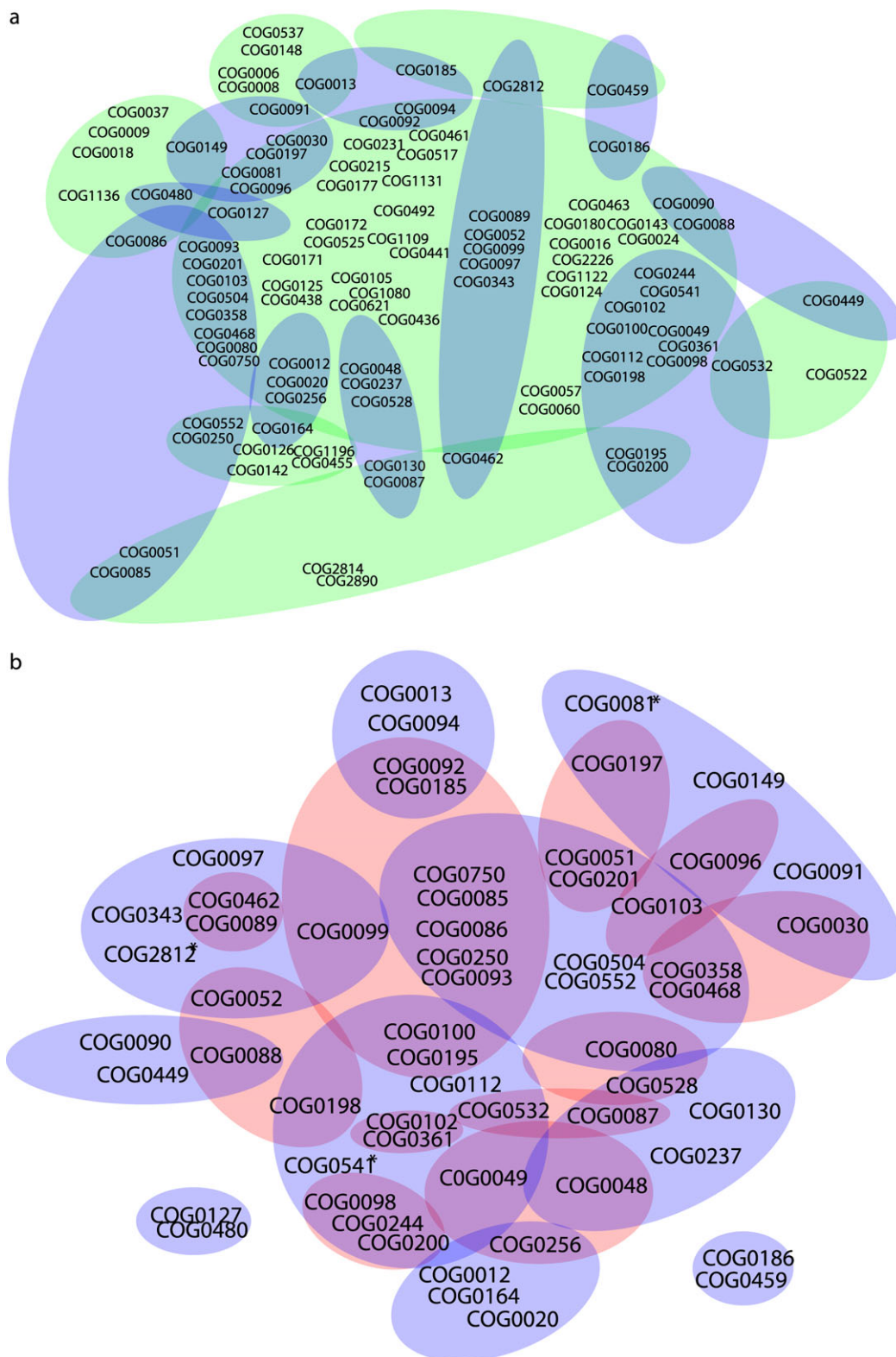
The subsequent search for the gene sets that may have a common history within the NUTs produced highly incompatible results. The ILD and LRTs were both able to detect incongruence but not to indicate whether any subsets of

markers were homogeneous. Both CADM and Concaterpillar were able to infer homogeneous subsets; for CADM, the  $P$  values of pairwise Mantel (1967) tests among all markers were clustered hierarchically with a complete linkage algorithm and those subsets that appear in clusters below the  $P = 0.05$  threshold were not considered incongruent or heterogeneous (fig. 5).

In figure 6a, the limited extent to which congruent subsets identified by CADM and Concaterpillar with the 41-taxon data set were in agreement is shown in a Venn diagram; figure 6b shows the Venn diagram produced from



**FIG. 5.**—Hierarchically clustered pairwise CADM test  $P$  values. The CADM test rejected global incongruence of the data set ( $P < 0.001$ ), indicating that at least one pair of markers was not incongruent over at least some part of their histories. We then assessed pairwise incongruence with Mantel tests and then clustered the  $P$  values hierarchically using a complete linkage algorithm. Those markers clustered above the threshold of 0.05 (indicated by a dashed red horizontal line) were considered homogeneous.



**FIG. 6.**—Similarity in homogeneous sets identified by CADM and Concaterpillar. (a) Venn diagram showing overlap in homogeneous sets identified by Concaterpillar (blue) and CADM (green) with the 41-taxon NUTs data set. (b) Venn diagram showing overlap in homogeneous sets identified by Concaterpillar with the 41-taxon (blue) and 100-taxon (red) data sets. One cluster was found in the 100-taxon data set but was incompatible with this Venn diagram; the members of this cluster (COG0081, COG0541, and COG2812) are indicated by an asterisk. Singletons (genes identified as incongruent to all others) identified by both methods are not shown.

the homogeneous subsets identified by Concaterpillar with the NUTs containing all taxa and the pruned NUTs containing 41 taxa. Qualitatively, it appears that there is very little overlap between the sets identified in these data sets with Concaterpillar. However, because the pattern of incongruence is likely to change depending on the taxa included in the analysis, this result is unsurprising. Interestingly, all 44 genes identified as incongruent to all others (singletons) by Concaterpillar with the 41-taxon data set were also identified as singletons with the 100-taxon data set; an additional 18 singletons were identified only in the 100-taxon data set. In addition, CADM and Concaterpillar are likely to identify different homogeneous subsets (see fig. 6a) because their definitions of incongruence differ (i.e., CADM will reject heterogeneity when there is more shared branching pattern than expected by chance, whereas Concaterpillar will reject homogeneity when there is sufficient evidence that gene trees are nonidentical).

Interestingly, analysis of a slightly expanded data set using a clustering-based method for detecting incongruence indicated that these markers were not homogeneous but produced only two subsets (Leigh et al. 2011). However, further analysis revealed that one of these subsets corresponded to most of the singletons identified in a Concaterpillar-based analysis of the same data. Additionally, the markers in this subset appeared to have undergone LGT events much more frequently than the others, and this subset was enriched in operational genes, whereas the other subset was enriched in informational genes. These results suggested that this clustering method identified some shared aspect of the evolutionary process other than a shared phylogenetic tree (e.g., the commonality of being subjected to higher rates of LGT, in this case).

Even for such a reduced data set, computational limits started to be observed (table 2). As we are not aware of a publicly available implementation of either the Huelsenbeck and Bull or Waddell LRTs, we implemented each of these tests (available by request) such that phylogenetic inference and likelihood estimation were calculated in parallel as much as possible; Concaterpillar also performs a number of steps in parallel. For this reason, an additional column was included in table 2 to indicate the approximate total CPU time used for each method, although this value is likely overestimated for methods using multiple CPU cores in parallel. For a data set of this size, the computation time for the Waddell LRT remained tractable. The ILD was much slower, although the total CPU time was comparable; the speed could be improved easily with a parallel implementation. The Huelsenbeck and Bull LRT was much slower, but most of the time was spent on parametric bootstrapping; this time was improved using a repartitioning method similar to that used in the ILD, which does not require inference of a global tree at each iteration (using 16 cores, repartitioning reduced

the computation time from 5.5 days to 38 h). For CADM, computation time was exceptionally fast: the entire analysis took less than 1.5 h, most of which was spent estimating distance matrices; the CADM global test itself was completed in under 2 min, whereas the a posteriori tests ran for just over 4 min. Computation time for each of these four methods (ILD, the two LRTs, and CADM) increases linearly with the number of markers in the data set, so even the slowest of these methods could reasonably be extended to larger data sets.

However, this is not the case with Concaterpillar. Its total computation time was less than a week, running in parallel over 16 cores. But because it scales with the square of the number of markers in the data set, the time for a data set with twice as many markers (around 200 genes) would be four times longer than for this data set, a total of 8 days. One can imagine that truly phylogenomic data sets could contain many more than 200 markers, and Concaterpillar would quickly become intractable.

Furthermore, the importance of parallelisation for these methods cannot be overstated. We intended to compare these results to those of the multiple ILD test (Planet and Sarkar 2005) but because the available implementation does not run any operations in parallel, completion of the analysis would have taken somewhere between 6 months and 8 years, depending on the point at which congruence is rejected. Likewise, had Concaterpillar been run on a single core, computation might have taken over 100 days.

## Building a Better Mousetrap: The Future of Congruence Tests

Our criticism of existing incongruence tests is not meant to deconstruct incongruence analysis in principle. Times have rarely been so exciting for phylogeneticists: there are now hundreds of whole-genome sequences, most of which are from prokaryotes, where phylogenetic disagreement between markers is of critical importance both to our understanding of the nature of genome evolution and to the meaning of phylogeny. Methodological progress is needed since existing incongruence methods show some serious limits in the post-genomic era, where data sets are increasing in size and phylogenetic complexity as sequence databases grow (fig. 2 and table 1). Methods that scale poorly with the number of markers in the data set (e.g., Leigh et al. 2008) or that are poorly suited to data sets where the level of heterogeneity is expected to be high (e.g., Brochier et al. 2002) are ill-suited to the data sets that are of growing interest.

Research is showing increasingly that, in terms of genome evolution, most of the “tree of life” is less a tree than a network (Brilli et al. 2008; Lima-Mendez et al. 2008; McInerney et al. 2008; Dagan and Martin 2009;

Ragan and Beiko 2009; Halary et al. 2010); that is, there is no common phylogenetic tree, with a few genes whose evolutionary history conflicts with that tree. Rather, there is a whole series of different trees, all of which are true trees for some parts of the genome. Some authors have avoided assessment of congruence altogether, opting instead to develop phylogenetic analysis methods that incorporate models that account for incongruence. Models have been proposed that explicitly account for incongruence among markers due to coalescence (Liu and Pearl 2007), LGT (Suchard 2005; Boussau and Daubin 2010), or generalized horizontal evolution (Bloomquist and Suchard 2010).

These methods are still in their infancy, and with the exception of the promising network-based method of Bloomquist and Suchard (2010), they effectively reconcile discordant gene tree evolution with a vertical species tree and can therefore be misleading in the case of prokaryote or viral evolution, where the existence of a species tree remains in question (Baptiste et al. 2009). Analyses of incongruence, on the other hand, can identify patterns of genes with identical, similar, or very different histories without attempting to merge heterogeneous information into a single tree. Thus, their range of utility is greater than that of any tree-based method because they make fewer assumptions to accommodate internal discrepancies in the data. Incongruence testing remains important both for testing whether combined phylogenetic analysis is appropriate and for exploring the evolutionary processes that shape genomic data. However, the fact remains that the vast majority of phylogenomic data have not evolved according to the same processes as those that shaped the data for which existing incongruence tests were conceived. We propose the development of methods for assessing incongruence that 1) accommodate both a high level of localized homogeneity and global incongruence; 2) appropriately account for and model patchy taxonomic distribution; and 3) scale reasonably well with the number of markers in the data set.

In order to be tractable for analysis of very large data sets, incongruence methods of the future will need to involve a phylogenetic analysis stage that scales linearly with the number of markers at worst. Clustering methods are promising in this regard (Leigh et al. 2011). Using an analysis method that produces a distribution of trees for each gene (e.g., a bootstrap distribution or a Bayesian posterior distribution), phylogenetic distances between distributions for all pairs of genes could be estimated and these then clustered (Nye 2008). The distances should somehow take differences in taxon representation into account rather than simply ignoring taxa missing from either of the two markers. Jackknifing of taxa or genes could potentially be used to assess the contribution of individuals to the perturbation of the recovered clusters of topologically homogeneous genes (e.g., if the removal of a particular taxon frequently causes two

clusters to merge, the phylogenetic position of this taxon is likely important to the incongruence of these sets of markers). In addition, an advantage of some clustering methods is that cluster membership need not be exclusive; fuzzy clustering (Bezdek and Ehrlich 1984) could allow a marker to belong to multiple clusters in cases where different regions of the gene have distinct evolutionary histories due to hybridization (gene conversion) events or where the marker in question shares local homogeneity with different clusters of genes due to independent LGT or gene recruitment in different lineages.

### Process Homogeneity: A Complementary Perspective on Incongruent Genes

In order to accommodate data sets in which evolution of genes along identical tree topologies is the exception, rather than the rule, it may prove useful to focus on identification of sets of genes that share more phylogenetic properties with each others than with other gene trees in the data set (e.g., congruence or homogeneity of evolutionary process), even if they are not themselves identical. More precisely, a homogeneous subset of gene trees need not share a single underlying tree, but could nonetheless share some remarkable evolutionary properties (i.e., a comparable rate of LGT). Elsewhere (Leigh et al. 2011), we have described these genes with significant evolutionary similarity as “evolutionary doppelgängers,” from the German word meaning “living double” or “walking double,” which usually refers to an identical “twin” who shares no literal relation to oneself. Genes sharing process homogeneity are similar to one another in significant ways but do not share the same pattern of inheritance (i.e., they do not share the same genealogy and, therefore, are incongruent or heterogeneous in the usual sense but share attributes of the evolutionary process such as similar rates of LGT, thus are homogeneous in this sense).

We feel that this type of congruence is relevant for microbial gene evolution, where many genes share process homogeneity. Consider for instance suites of genes within operons or other genetic modules that tend to be coinherited, at least between some taxa (Walsby 1994; Yellaboina et al. 2004; Watanabe et al. 2008; Iwasaki and Takagi 2009). Even though their trees might not be strictly identical, they will likely present some significant local regions of topological similarity, capturing real evolutionary processes uniting the evolution of these genes, and justifying their grouping into an evolutionarily meaningful set. In this case, the evolutionary history of prokaryotic genes is more accurately described by process homogeneity, where the notion of global phylogenetic homogeneity (identical trees) is too strict to describe local phylogenetic similarity between gene trees (fig. 2, lower right: “local homogeneity in history”).

Moreover, the term process homogeneity is flexible enough to include genes that have been subjected to similar evolutionary pressures, even if they do not share exactly the same pattern of inheritance. Such a group would typically be observed when genes fall into distinct classes of genes characterized by distinct rates of LGT. For example, according to the complexity hypothesis (Jain et al. 1999), genes fall into two classes: “informational” genes, supposedly less frequently transferred and “operational” genes, more frequently transferred. If the complexity hypothesis is correct, frequently transferred operational genes and rarely transferred informational genes have distinct evolutionary properties. Consistently, incongruence analyses could be designed to identify these two groups of markers (fig. 2, lower right: “local process homogeneity”). That the group of operational genes comprises multiple underlying histories does not make this grouping meaningless: the evolutionary resemblance between operational genes (i.e., their more frequent transfer relative to other genes), if correct, deserves recognition (Leigh et al. 2011). Although a method to detect process-homogeneous markers would be related to the notion of conditional data combination in that it would be based on incongruence analysis, it would not necessarily be used to evaluate combinability of data for the inference of a species tree; detection of these markers would be at least as useful for exploration of patterns of LGT frequency or of gene sharing highways (Beiko et al. 2005).

Simply put, we would argue that phylogenetic homogeneity should not exclusively mean shared, identical phylogenetic story, but should be expanded to include shared significant similarity in other aspects of evolutionary processes (e.g., when a group of genes presents a distinct rate of LGT relative to others and, consequently, distinct taxonomic/environmental distribution). As the latter resemblances occur in microbial evolution owing to the importance of LGT, the notion of process homogeneity could enrich the incongruence analysis tool kit.

## Conclusions

As our understanding of molecular evolution moves away from the tree metaphor (Baptiste et al. 2009; Dagan and Martin 2009; Ragan and Beiko 2009), the identification of incongruence will no doubt continue to prove useful for many areas of evolutionary biology and foster multiple novel important questions. How many separate histories do genomes of different lineages exhibit? Why do some sets of genes share patterns of not-strictly vertical evolution? Do genes whose products physically or functionally interact tend to share the same patterns of inheritance, encoding “molecular organs,” with their own evolutionary fate, as suggested by Forterre (2010)? Do genes tend to follow the same pattern of inheritance over the entire course of their histories or are some groups of genes only coinherited

at a certain evolutionary time? Has the rate of LGT for distinct functional categories varied over time, marking distinct adaptive stages of microbial evolution? The development of new, better tests, more grounded in biological knowledge, is crucial to address all these issues.

## Acknowledgments

We would like to thank Klaus Schliep and two anonymous referees for providing thoughtful comments on this manuscript. J.W.L. was supported by a postdoctoral fellowship from the Université Pierre et Marie Curie and is currently supported by a postdoctoral fellowship from the Natural Science and Engineering Research Council of Canada. F.J.L.’s work was partially funded by a visiting professorship from the Muséum National d’Histoire naturelle.

## Literature Cited

- Abdi H. 2007. In: Salkind NJ, editor. *Encyclopedia of measurement and statistics*. Thousand Oaks (CA): SAGE. p. 103–107.
- Adams EN. 1986. N-trees as nestings: complexity, similarity, and consensus. *J Classif.* 3:299–317.
- Baptiste E, et al. 2009. Prokaryotic evolution and the tree of life are two different things. *Biol Direct.* 4:34.
- Baptiste E, et al. 2005. Do orthologous gene phylogenies really support tree-thinking? *BMC Evol Biol.* 5:33.
- Baptiste E, et al. 2008. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. *Mol Biol Evol.* 25:83–91.
- Barrett M, Donoghue MJ, Sober E. 1991. Against consensus. *Syst Zool.* 40:486–493.
- Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A.* 102:14332–14337.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Biol Sci.* 289–300.
- Bezdek JC, Ehrlich R. 1984. FCM: the fuzzy c-means clustering algorithm. *Comput Geosci.* 10:191–203.
- Biedler JK, Shao H, Tu Z. 2007. Evolution and horizontal transfer of a DD37E DNA transposon in mosquitoes. *Genetics.* 177:2553–2558.
- Bininda-Emonds ORP, editor. 2004. *Phylogenetic supertrees: combining information to reveal the tree of life*. Computational biology, volume 4. Dordrecht (The Netherlands): Kluwer Academic Publishers.
- Bloomquist EW, Suchard MA. 2010. Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. *Syst Biol.* 59:27–41.
- Bordewich M, Semple C. 2004. On the computational complexity of the rooted subtree prune and regraft distance. *Ann Combin.* 8:409–423.
- Boucher Y, Baptiste E. 2009. Revisiting the concept of lineage in prokaryotes: a phylogenetic perspective. *Bioessays.* 31:526–536.
- Boussau B, Daubin V. 2010. Genomes as documents of evolutionary history. *Trends Ecol Evol.* 25:224–232.
- Brilli M, et al. 2008. Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. *BMC Bioinformatics.* 9:551.

- Brochier C, Bapteste E, Moreira D, Philippe H. 2002. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet.* 18:1–5.
- Bryant D, McKenzie A, Steel M. 2003. The size of maximum agreement subtree for random binary trees. In: Janowitz M, Lapointe FJ, McMorris FR, Mirkin B, Roberts FS, editors. *BioConsensus*. Providence (RI): American Mathematical Society. p. 55–66.
- Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ. 1993. Partitioning and combining data in phylogenetic analysis. *Syst Biol.* 42:384–397.
- Campbell V, Legendre P, Lapointe FJ. 2009. Assessing congruence among ultrametric distance matrices. *J Classif.* 26:103–117.
- Campbell V, Legendre P, Lapointe FJ. 2011. The performance of the Congruence Among Distance Matrices (CADM) test in phylogenetic analysis. *BMC Evol Biol.* 11:64.
- Creevey CJ, et al. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proc R Soc Lond B Biol Sci.* 271:2551–2558.
- Cunningham CW. 1997. Can three incongruence tests predict when data should be combined? *Mol Biol Evol.* 14:733–740.
- Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A.* 105:10039–10044.
- Dagan T, Martin W. 2006. The tree of one percent. *Genome Biol.* 7:118.
- Dagan T, Martin W. 2009. Getting a better picture of microbial evolution en route to a network of genomes. *Philos Trans R Soc Lond B Biol Sci.* 364:2187.
- Day WHE. 1983. Distributions of distances between pairs of classifications. In: Felsenstein J, editor. *Numerical taxonomy*. Berlin (Germany): Springer-Verlag. p. 127–131.
- de Andrade Zanotto PM, Krakauer DC. 2008. Complete genome viral phylogenies suggests the concerted evolution of regulatory cores and accessory satellites. *PLoS One* 3:e3500.
- de Queiroz A. 1993. For consensus (sometimes). *Syst Biol.* 42:368–372.
- de Queiroz A, Donoghue MJ, Kim J. 1995. Separate versus combined analysis of phylogenetic evidence. *Annu Rev Ecol Syst.* 26: 657–681.
- de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. *Trends Ecol Evol.* 22:34–41.
- de Vienne DM, Giraud T, Martin OC. 2007. A congruence index for testing topological similarity between trees. *Bioinformatics.* 23: 3119–3124.
- Estabrook GF, McMorris FR, Meacham CA. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst Zool.* 34:193–200.
- Farris JS. 1989. The retention index and the rescaled consistency index. *Cladistics.* 5:417–419.
- Farris JS, Källersjö M, Kluge AG, Bult C. 1994. Testing significance of incongruence. *Cladistics.* 10:315–319.
- Forterre P. 2010. Defining life: the virus viewpoint. *Orig Life Evol Biosph.* 40:151–160.
- Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. *Science.* 315:476–480.
- Garamszegi LZ. 2009. Patterns of co-speciation and host switching in primate malaria parasites. *Malar J.* 9:110.
- Gilbert CC, Rossie JB. 2007. Congruence of molecules and morphology using a narrow allometric approach. *Proc Natl Acad Sci U S A.* 104:11910–11914.
- Goldman N, Anderson JP, Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst Biol.* 49:652–670.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.
- Hafner MS, Nadler SA. 1988. Phylogenetic trees support the coevolution of parasites and their hosts. *Nature.* 332:258–259.
- Halary S, et al. 2010. Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci U S A.* 107:127.
- Hanage WP, Fraser C, Spratt BG. 2006. The impact of homologous recombination on the generation of diversity in bacteria. *J Theor Biol.* 239:210–219.
- Harding EF. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. *Adv Appl Probab.* 3:44–77.
- Huelsenbeck JP, Bull JJ. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. *Syst Biol.* 45:92–98.
- Huelsenbeck JP, Bull JJ, Cunningham CW. 1996. Combining data in phylogenetic analysis. *Trends Ecol Evol.* 11:152–158.
- Huelsenbeck JP, Swofford DL, Cunningham CW, Bull JJ, Waddell PJ. 1994. Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? *Syst Biol.* 43:288–291.
- Iwasaki W, Takagi T. 2009. Rapid pathway evolution facilitated by horizontal gene transfers across prokaryotic lineages. *PLoS Genet.* 5:e1000402.
- Jablonski D, Finarelli JA. 2009. Congruence of morphologically-defined genera with molecular phylogenies. *Proc Natl Acad Sci U S A.* 106:8262–8266.
- Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A.* 96:3801–3806.
- de Koning APJ, Gu W, Pollock DD. 2010. Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. *Mol Biol Evol.* 27:249–265.
- Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol.* 30:151–160.
- Kluge AG. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Syst Biol.* 38:7–25.
- Kluge AG. 1998. Total evidence or taxonomic congruence: cladistics or consensus classification. *Cladistics.* 14:151–158.
- Kluge AG, Farris JS. 1969. Quantitative phyletics and the evolution of anurans. *Syst Zool.* 18:1–32.
- Kluge AG, Wolf AJ. 1993. Cladistics: what's in a word? *Cladistics.* 9:183–199.
- Křivánek M. 1986. Computing the nearest neighbor interchange metric for unlabeled binary trees is NP-complete. *J Classif.* 3:55–60.
- Lake JA, Rivera MC. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol Biol Evol.* 21:681–690.
- Lapointe FJ. 1998. How to validate phylogenetic trees? A stepwise procedure. In: Hayashi C, et al. editors. *Data science, classification, and related methods*. Tokyo (Japan): Springer-Verlag. p. 71–88.
- Lapointe FJ, Legendre P. 1990. A statistical framework to test the consensus of two nested classifications. *Syst Zool.* 39:1–13.
- Lapointe FJ, Legendre P. 1992a. A statistical framework to test the consensus among additive trees (cladograms). *Syst Biol.* 41:158–171.
- Lapointe FJ, Legendre P. 1992b. Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees. *Syst Biol.* 41:378–384.
- Lapointe FJ, Legendre P. 1995. Comparison tests for dendrograms: a comparative evaluation. *J Classif.* 12:265–282.
- Lapointe FJ, Rissler LJ. 2005. Consensus, congruence, and the comparative phylogeography of codistributed species in California. *Am Nat.* 166:290–299.

- Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*. 25:2286–2288.
- Leigh JW, Schliep K, Lopez P, Baptiste E. Forthcoming 2011. Let them fall where they may: congruence analysis in massive, phylogenetically messy datasets. *Mol Biol Evol*.
- Leigh JW, Susko E, Baumgartner M, Roger AJ. 2008. Testing congruence in phylogenomic analysis. *Syst Biol*. 57:104–115.
- Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol*. 1:E19.
- Levasseur C, Lapointe FJ. 2001. War and peace in phylogenetics: a rejoinder on total evidence and consensus. *Syst Biol*. 50:881–891.
- Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. 2008. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol*. 25:762–777.
- Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol*. 56:504–514.
- MacLeod D, Charlebois RL, Doolittle F, Baptiste E. 2005. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evol Biol*. 5:27.
- Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res*. 27:209–220.
- McInerney JO, Cotton JA, Pisani D. 2008. The prokaryotic tree of life: past, present ... and future? *Trends Ecol Evol*. 23:276–281.
- McInerney JO, Wilkinson M. 2005. New methods ring changes for the tree of life. *Trends Ecol Evol*. 20:105–107.
- Mickevich MF. 1978. Taxonomic congruence. *Syst Biol*. 27:143–158.
- Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F. 2010. The bacterial pan-genome: a new paradigm in microbiology. *Int Microbiol*. 13:45–57.
- Miyamoto MM, Fitch WM. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Syst Biol*. 44:64–76.
- Nelsen MP, Gargas A. 2008. Dissociation and horizontal transmission of codispersing lichen symbionts in the genus *Lepraria* (Lecanorales: Stereocaulaceae). *New Phytol*. 177:264–275.
- Norman A, Hansen LH, Sørensen SJ. 2009. Conjugative plasmids: vessels of the communal gene pool. *Philos Trans R Soc Lond B Biol Sci*. 364:2275–2289.
- Nye TMW. 2008. Trees of trees: an approach to comparing multiple alternative phylogenies. *Syst Biol*. 57:785–794.
- Page RDM, Charleston MA. 1997. From gene tree to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*. 7:231–240.
- Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*. 20:289–290.
- Penny D, Foulds LR, Hendy MS. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature*. 297:197–200.
- Penny D, Hendy MD. 1985. The use of tree comparison metrics. *Syst Zool*. 34:75–82.
- Phipps JB. 1975. The numbers of classifications. *Can J Bot*. 54:686–688.
- Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol*. 24:1752–1760.
- Planet PJ. 2006. Tree disagreement: measuring and testing incongruence in phylogenies. *J Biomed Inform*. 39:86–102.
- Planet PJ, Sarkar IN. 2005. mlLD: a tool for constructing and analyzing matrices of pairwise phylogenetic character incongruence tests. *Bioinformatics*. 21:4423–4424.
- Puigbò P, Wolf YI, Koonin EV. 2009. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J Biol*. 8:59.
- Puigbò P, Wolf YI, Koonin EV. 2010. The tree and net components of prokaryote evolution. *Genome Biol Evol*. 2:745–756.
- Ragan MA, Beiko RG. 2009. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond B Biol Sci*. 364:2241.
- Refregier G, et al. 2008. Cophylogeny of the anther smut fungi and their Caryophyllaceae hosts: prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation. *BMC Evol Biol*. 8:100.
- Rieppel O. 2005. The philosophy of total evidence and its relevance for phylogenetic inference. *Pap Avulsos Zool (São Paulo)*. 45:1–31.
- Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431:152–155.
- Robinson D, Foulds L. 1981. Comparison of phylogenetic trees. *Math Biosci*. 53:131–147.
- Rodrigo AG, Kelly-Borges M, Bergquist PR, Bergquist PL. 1993. A randomization test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *N Z J Bot*. 31:257–268.
- Sanderson MJ, Purvis A, Henze C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends Ecol Evol*. 13:105–109.
- Shao K, Rohlf FJ. 1983. Sampling distributions of consensus indices when all bifurcating trees are equally likely. In: Felsenstein J, editor. *Numerical taxonomy*. Berlin (Germany): Springer-Verlag. p. 132–137.
- Shao K, Sokal RR. 1986. Significance tests of consensus indices. *Syst Zool*. 35:582–590.
- Shi T, Falkowski PG. 2008. Genome evolution in cyanobacteria: the stable core and the variable shell. *Proc Natl Acad Sci U S A*. 105:2510–2515.
- Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*. 51:492–508.
- Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol*. 16:1114–1116.
- Simberloff D, Heck KL, McCoy ED, Connor EF. 1981. There have been no statistical tests of cladistic biogeographical hypotheses. In: Nelson G, Rosen DE, editors. *Vicariance biogeography: a critique*. New York: Columbia University Press. p. 40–63.
- Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22:2688–2690.
- Steel MA. 1988. Distribution of the symmetric difference metric in phylogenetic trees. *SIAM J Discr Math*. 1:541–555.
- Steel MA, Penny D. 1993. Distributions of tree comparison metrics: some new results. *Syst Biol*. 42:126–141.
- Storey JD. 2002. A direct approach to false discovery rates. *J R Stat Soc Ser B Biol Sci*. 64:479–498.
- Suchard MA. 2005. Stochastic models for horizontal gene transfer: taking a random walk through tree space. *Genetics* 170: 419–431.
- Susko E, Leigh J, Doolittle WF, Baptiste E. 2006. Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria. *Mol Biol Evol*. 23:1019–1030.
- Swofford DL. 1991. When are phylogeny estimates from molecular and morphological data incongruent? In: Miyamoto MM, Cracraft JJ, editors. *Phylogenetic analysis of DNA sequences*. Oxford: Oxford University Press. p. 295–333.
- Swofford DL. 2003. PAUP\* 4.0 b10. Phylogenetic analysis using parsimony (\* and other methods). Sunderland (MA): Sinauer Associates.



- Virgilio M, De Meyer M, White IM, Backeljau T. 2009. African *Dacus* (Diptera: Tephritidae): molecular data and host plant associations do not corroborate morphology-based classifications. *Mol Phylogenet Evol.* 51:531–539.
- Waddell PJ, Kishino H, Ota R. 2000. Rapid evaluation of the phylogenetic congruence of sequence data using likelihood ratio tests. *Mol Biol Evol.* 17:1988–1992.
- Walsby AE. 1994. Gas vesicles. *Microbiol Rev.* 58:94.
- Watanabe RL, Morett E, Vallejo EE. 2008. Inferring modules of functionally interacting proteins using the Bond Energy Algorithm. *BMC Bioinformatics.* 9:285.
- Waterman MS, Smith TF. 1978. On the similarity of dendrograms. *J Theor Biol.* 73:784–900.
- Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol.* 18(5):691–699.
- Woese CR. 1987. Bacterial evolution. *Microbiol Rev.* 51:221–271.
- Wu B, et al. 2008. Assessment of codivergence of mastreviruses with their plant hosts. *BMC Evol Biol.* 8:335.
- Wu Y. 2009. A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics* 25:190.
- Yellaboina S, Ranjan S, Chakhaiyar P, Hasnain SE, Ranjan A. 2004. Prediction of DtxR regulon: identification of binding sites and operons controlled by Diphtheria toxin repressor in *Corynebacterium diphtheriae*. *BMC Microbiol.* 4:38.
- Zelwer M, Daubin V. 2004. Detecting phylogenetic incongruence using BIONJ: an improvement of the ILD test. *Mol Phylogenet Evol.* 33:687–693.
- Zuckerandl E, Pauling L. 1965. Molecules as documents of evolutionary history. *J Theor Biol.* 8:357–366.
- Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [PhD thesis]. Austin (TX): University of Texas at Austin.

**Associate editor:** Bill Martin