doi:10.1111/j.1558-5646.2008.00442.x

A GENEALOGICAL APPROACH TO QUANTIFYING LINEAGE DIVERGENCE

Michael P. Cummings,[1,2] Maile C. Neel,[3,4] and Kerry L. Shaw[5,6,7]

[1]*Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742*

[2]*E-mail: mike@umiacs.umd.edu*

[3]*Department of Plant Science and Landscape Architecture, and Department of Entomology, University of Maryland, College Park, Maryland 20742*

[4]*E-mail: mneel@umd.edu*

[5]*Department of Biology, University of Maryland, College Park, Maryland 20742*

[6]*E-mail: kls4@cornell.edu*

Received November 3, 2007
Accepted May 22, 2008

We introduce a statistic, the genealogical sorting index (*gsi*), for quantifying the degree of exclusive ancestry of labeled groups on a rooted genealogy and demonstrate its application. The statistic is simple, intuitive, and easily calculated. It has a normalized range to facilitate comparisons among different groups, trees, or studies and it provides information on individual groups rather than a composite measure for all groups. It naturally handles polytomies and accommodates measures of uncertainty in phylogenetic relationships. We use coalescent simulations to explore the behavior of the *gsi* across a range of divergence times, with the mean value increasing to 1, the maximum value when exclusivity within a group reached monophyly. Simulations also demonstrate that the power to reject the null hypothesis of mixed genealogical ancestry increased markedly as sample size increased, and that the *gsi* provides a statistically more powerful measure of divergence than F_{ST}. Applications to data from published studies demonstrated that the *gsi* provides a useful way to detect significant exclusivity even when groups are not monophyletic. Although we describe this statistic in the context of divergence, it is more broadly applicable to quantify and assess the significance of clustering of observations in labeled groups on any tree.

KEY WORDS: **Ancestral polymorphism, congruence, exclusivity, genealogy, lineage sorting, monophyly, paraphyly, phylogeny, polyphyly, speciation, species.**

A primary goal of systematic biology is to identify monophyletic groups of organisms (e.g., Sites and Marshall 2003). Major efforts are focused at the species level because species are the fundamental units of the taxonomic hierarchy and the process of speciation underlies the evolution of biological diversity (Templeton 2001). An explicitly phylogenetic view of species became possible when the intraspecific history of genes could be understood through changes in the underlying DNA sequence as pioneered by Avise et al. (1979). This historical perspective enables a unified approach to the intraspecific study of populations and the phylogenetic study of closely related species. The power of this unification is that it ultimately facilitates understanding of the intraspecific processes that produce character evolution directly involved in the origin of new species. This potential has inspired investigations of genealogical relationships at a variety of scales, and has revealed the complex nature of species boundaries (e.g., Sites and Marshall 2004). This complexity arises in part from the variety of possible mechanisms that can give rise to new species.

[7]Present address: Department of Neurobiology and Behavior, Cornell University, Ithaca, New York 14853

Despite the many speciation mechanisms and the equally diverse range of species concepts, the divergence of species should be reflected in the genealogy of most loci. Ultimately gene copies at a given locus within species should share a common evolutionary history to the exclusion of gene copies within other species, that is, they will form monophyletic groups (Avise and Ball 1990; Baum and Shaw 1995; Liu and Pearl 2007). Studies of closely related species, however, are increasingly demonstrating that gene tree–species tree mismatches are quite common (Templeton 2001; Shaw 2002; Funk and Omland 2003; Machado and Hey 2003; Edwards et al. 2005; Pollard et al. 2006) and these mismatches have presented problems for inferring species relationships from allelic variation (Takahata 1989; Knowles and Maddison 2002; Slatkin and Pollack 2006; Knowles and Carstens 2007; Liu and Pearl 2007). As Templeton (2001) points out, the groups that are actively diverging or that have diverged recently are most likely to have gene tree–species tree mismatches. Although they present challenges, these groups are likely to be most informative about speciation processes.

Mismatches between a gene tree and a species tree result chiefly from shared ancestral polymorphism or introgressive hybridization. Shared ancestral polymorphism is an expected stage in the transition from polyphyly to paraphyly to monophyly among diverging populations (Tajima 1983; Avise and Ball 1990; Baum and Shaw 1995; Maddison 1997). Initially, when two interbreeding populations begin to diverge from a single population the gene copies within both descendent populations for any particular locus will share many ancestors in common. Thus, gene copies within either of the two diverging populations will exhibit polyphyletic ancestry. In absence of genetic exchange between these populations, over time genetic drift will lead to sorting of the gene lineages. As some gene lineages proliferate and others go extinct, patterns of exclusive ancestry within each population evolve. Individual populations will become monophyletic at some loci whereas others remain paraphyletic. Eventually, however, gene copies at all loci within each interbreeding population will evolve to a state of reciprocal monophyly if the process of genetic drift is unopposed. From coalescent theory we know the timeframe of this transition will vary for neutral loci due to the stochastic nature of the evolutionary sampling process and as a function of the inbreeding effective size of the locus considered (Hudson 1990; Moore 1995; Degnan and Rosenberg 2006). Specifically, under standard neutral coalescent theory assumptions, a single locus has a 0.95 probability of achieving reciprocal monophyly in 8.70 N_e (inbreeding effective population size) generations (Hudson and Coyne 2002).

The standard categorical phylogenetic concepts of polyphyly, paraphyly, and monophyly describe qualitative relationships among organismal groups (such as species) and fail to quantify the degree of genealogical divergence along a continuum. Monophyly represents an endpoint in the divergence process detectable through the topology of a phylogenetic tree, and as a binary condition (i.e., a group is either monophyletic or not monophyletic) it is ill-suited for quantifying divergence at other points along the continuum of genealogical divergence. Quantitative assessments of the extent to which genealogical relationships depart from monophyly or conversely the degree to which they depart from random polyphyly have been lacking. To further our understanding of the relationships between microevolutionary processes and phylogeny, statistical tools to objectively quantify the degree of genealogical sorting are needed. In this article, we develop the genealogical sorting index (*gsi*), a simple statistic that estimates the degree of exclusive ancestry of individuals in labeled groups on a rooted tree. We assess the behavior of the *gsi* across a range of divergence times through simulation, assess its statistical significance through permutation tests, and demonstrate its usefulness through application to several published phylogenetic trees. The *gsi* enables a departure from the typological view of relationships being either monophyletic or not by quantifying the degree of exclusive ancestry short of the end point of monophyly.

THE GENEALOGICAL SORTING INDEX

To achieve our objectives we need a measure that quantifies the relative degree of exclusive ancestry of a group on a rooted tree topology, where a group is defined as a set of commonly labeled branch tips and exclusivity is the amount of ancestry for a group that is common to only members of the group. Typically, branch tips in a tree will represent gene copies from a given locus, sampled from the individuals of the group. Thus, the value of the *gsi* may be taken to estimate the degree of genealogical exclusivity among the sampled gene copies as well as a quantification of the relationships among individuals from which the gene copies were sampled, depending on the intended inference. Throughout this article, we refer to the unit sampled as "individual" while acknowledging this duality. We start by calculating a measure, denoted *gs* for genealogical sorting, for any group on any tree of interest as the minimum number of nodes on a fully resolved tree required to unite a group, divided by the number of nodes actually uniting the group. Thus, the numerator represents the fully exclusive case (i.e., monophyly), the denominator represents the observed amount of exclusivity, and the quotient of these terms is a measure of relative exclusivity. Formally, *gs* is defined as

$$gs = \frac{n}{\sum_{u=1}^{U}(d_u - 2)}, \qquad (1)$$

where d is the degree of node u of U total nodes uniting a group (estimated coalescent events) through a most recent common ancestor, and n is the minimum number of nodes (coalescent events) required to unite a group of size $n + 1$ through a most recent common ancestor on a given genealogy. In graph theory, the degree is the number of edge (branch) ends at a vertex (node).

The maximum possible *gs* value for any group is

$$\max(gs) = 1. \tag{2}$$

The max(gs) is reached when a group is monophyletic.

The minimum possible *gs* value for any group on a tree can be calculated as

$$\min(gs) = \frac{n}{\sum_{i=1}^{I}(d_i - 2)}, \tag{3}$$

where *i* is one of I total nodes on the tree. Thus the minimum value would result if all nodes on a tree were required to unite a group (i.e., $U = I$). For trees with groups of equal size min(gs) → $1/k$ as $n \to \infty$, where k is the number of groups. Disparity in group size leads to smaller possible min(*gs*) values for smaller groups.

To provide a normalized statistic to quantify the degree of genealogical sorting along the unit interval, [0, 1], we calculate the *gsi* as follows:

$$gsi = \frac{\text{observed}gs - \min(gs)}{\max(gs) - \min(gs)}. \tag{4}$$

Some examples help illustrate the calculation. In a tree with two groups, *a* and *b*, both of size four (Fig. 1), the numerator from equation (1) is $n = 3$. To determine the value of the denominator one needs to identify the smallest subtree containing all individuals in the group of interest and sum the degree of each node minus 2, including the root node of the subtree. In the current example there are three nodes uniting all *a* individuals, enclosed with solid lines. The denominator is calculated by summing the degree of each node minus 2 over all nodes uniting all *a* individuals: $(3 - 2) + (3 - 2) + (3 - 2) = 3$. So the *gs* value for group *a* is $3/3 = 1$, which is also the value of *gsi*. The comparable calculation for group *b* (nodes enclosed with dashed lines) has the same numerator as for group *a*. The denominator is calculated by summing the degree of each node minus 2 over all nodes uniting all *b* individuals: $(3 - 2) + (3 - 2) + (3 - 2) + (3 - 2) = 4$. Thus the *gs* value for group *b* in Figure 1 is 3/4. Normalization of this value using equation (4) yields the *gsi* value $((3/4) – (3/7))/(1 – (3/7)) = 0.563$. Note that the monophyletic group *a*, by definition, has $gsi = 1$, the maximum value, whereas $gsi < 1$ for the paraphyletic group *b*.

ACCOMMODATING UNRESOLVED RELATIONSHIPS

The *gsi* naturally accommodates polytomies. Figure 2 illustrates a tree of two labeled groups *a* and *b*, each of size 4. Thus the *gs* numerator is $n = 3$ for both groups. The single node uniting *a* is enclosed within an ellipse with a solid line. The degree of this node is 5, making the denominator calculation $5 - 2 = 3$, and so the *gs* value for group *a* is again $3/3 = 1$, as is the value of *gsi*. We can make a similar calculation for group *b* in Figure 2 using the nodes enclosed with dashed lines. Summing the degree of each node minus 2: $(3-2)+(3-2)+(4-2)=4$. So the *gs* value for group *b* in this case is $3/4 = 0.75$, and subsequent normalization yields the same *gsi* value as for group b in Figure 1, 0.563. This example illustrates that the *gsi* of a group is affected by polytomies only if the polytomies affect exclusive ancestry of the group with respect to other groups. This property is desirable, because resolutions of polytomies involving only one group have no influence on the degree to which individuals of that group share ancestry with

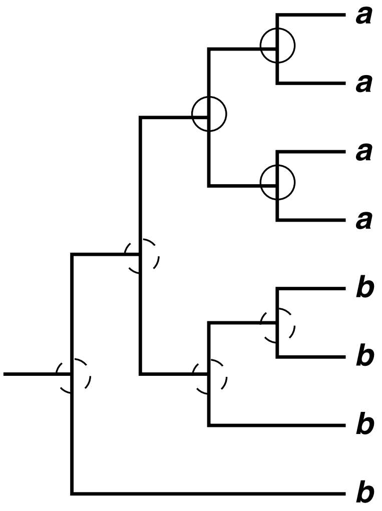

Figure 1. Hypothetical fully resolved phylogenetic tree showing the relationships of representatives from two labeled groups, *a* (uniting nodes enclosed by solid lines) and *b* (uniting nodes enclosed by dashed lines).

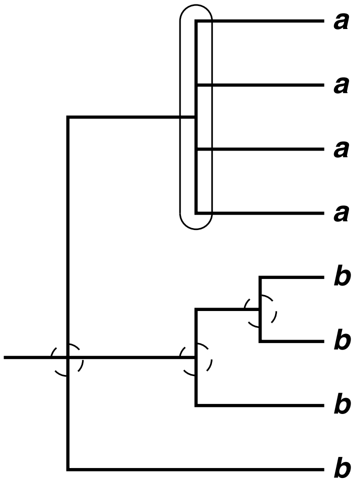

Figure 2. Hypothetical partially resolved phylogenetic tree showing the relationships of representatives from two labeled groups, *a* (uniting nodes enclosed by solid lines) and *b* (uniting nodes enclosed by dashed lines).

other groups on a tree. In contrast, polytomies involving more than one group increase the number of ancestors over the minimum possible and will lower *gsi* values from their maximum.

ASSESSING THE STATISTICAL SIGNIFICANCE OF THE GENEALOGICAL SORTING INDEX VALUES

The *gsi* quantifies the exclusive ancestry of a labeled group on a tree. The null hypothesis we wish to test is that the degree of exclusive ancestry of branch tips observed is that which might be observed at random. In other words the null hypothesis is that labeled groups form a single group of mixed genealogical ancestry for a given tree topology. An appropriate test of this hypothesis is a permutation test that holds the tree constant and permutes the group labels assigned to the tips of the tree (Maddison and Slatkin 1991), thus randomizing the common ancestry of members of the groups. We then calculate a *gsi* value for a random labeling of the tree tips. By generating many such permuted labelings and determining a *gsi* value for each, we can obtain a distribution of values from which we can determine the frequency of *gsi* values equal to or greater than that which we observed from the original labeled tree. This frequency is our *P*-value: the probability of observing by chance alone *gsi* values equal to or greater than the observed *gsi* value under the null hypothesis.

There is a wide range of applications for permutation testing because it is based on the empirical observations at hand, and thus subsumes any idiosyncrasies embodied in the original data (e.g., the number of groups, size of each group, and resolution of the topology). Permutation tests require no assumptions about underlying distributions of either the data or the statistics calculated, and the statistical power is usually equal to the most powerful parametric alternatives where these can be applied (Bickel and Van Zwet 1978). As is typical for measures on trees, no appropriate parametric alternative to assess the significance of the *gsi* is immediately obvious. Permutation tests are also exact in that estimated *P*-values are accurate (unbiased) with precision determined by the number of permutations evaluated.

We performed a permutation test as described above for groups *a* and *b* on the tree in Figure 1 using software developed for this purpose (A. L. Bazinet, M. C. Neel, K. L. Shaw, and M. P. Cummings, unpubl. ms.). With 10^4 replicates (the original observation plus 9999 permutations) $P = 0.0121$ for group *a*, and $P = 0.0130$ for group *b*. Thus the results show that values of the *gsi* equal to or greater than the observed values are unlikely to be observed by chance alone.

INTEGRATING ACROSS MULTIPLE TREES AND ACCOUNTING FOR UNCERTAINTY IN INFERRED RELATIONSHIPS

There are numerous instances in which multiple gene trees are estimated for a given set of groups. For example, comparisons within and among closely related populations or species often include estimates from multiple unlinked loci (where a locus is defined as a nonrecombining genomic region) to capture different patterns of lineage sorting in the transition from polyphyly to paraphyly to monophyly (Shaw 1998). Given sufficient variation, a composite measure across gene trees from multiple unlinked loci will provide the best picture of the degree to which a genome for a group has become exclusive in relation to the genome for another group.

Multiple trees will also result from bootstrap analysis (Felsenstein 1985) or from the posterior probability distribution determined by Markov chain Monte Carlo sampling in Bayesian analysis (Rannala and Yang 1996; Yang and Rannala 1997; Larget and Simon 1999; Mau et al. 1999). Uncertainty in inferred relationships can be incorporated into the *gsi* by calculating an ensemble statistic including all tree topologies resulting from a bootstrap analysis or trees from the posterior probability distribution determined by Markov chain Monte Carlo sampling in Bayesian analysis. Including uncertainty in the *gsi* calculation by integrating over tree topologies from bootstrap or Bayesian analysis of phylogenetic relationships (e.g., the 0.95 credible set of topologies) embraces the philosophy and analytical power of these approaches by allowing one to calculate a *gsi* value that is weighted by the probability of each constituent tree topology.

For any ensemble of trees we can calculate a single statistic, gsi_T, defined as the weighted sum of the *gsi* values from each tree topology:

$$gsi_T = \sum_{t=1}^{T} gsi_t P_t, \tag{5}$$

where T is the total number of topologies in the ensemble, gsi_t is the *gsi* value based on topology t, and P_t is the probability of the gene tree topology t. The probability of each gene tree topology is equal to its proportional representation in the sample of gene trees.

As with the *gsi* for a single tree, we can assess the significance of gsi_T through permutation. There are two ways that permutation might be applied to an ensemble of trees: permuting the class labels for each tree independently, or applying the same class labels to the same individuals across all trees in the ensemble for each replication. This latter permutation scheme preserves any correlated structure within groups across loci if such correlation is present, whereas the former permutation scheme does not. Thus the second approach might be appropriate if consideration of correlation among loci is germane to the problem at hand. However, under standard neutral coalescent assumptions (see below) the genealogical relationships among gene copies are uncorrelated across loci within groups for the null hypothesis of mixed genealogical ancestry (i.e., no divergence). Therefore, for the general case both permutation methods yield the same results with precision determined by the number of replicates (this has been confirmed by simulation, results not shown). The gsi_T and the associated *P*-value provide an objective criterion for assessing evolutionary distinctiveness, and these ensemble statistics have the advantage of variance reduction obtained by averaging across multiple loci.

SIMULATIONS TO EVALUATE THE BEHAVIOR OF THE GENEALOGICAL SORTING INDEX

Following other studies (e.g., Slatkin and Maddison 1990; Hudson and Coyne 2002; Rosenberg 2003) we used the coalescent (Kingman 1982) as the basis for simulating the divergence between groups, and we evaluated the behavior and power of the *gsi* as a function of divergence time. We simulated coalescent genealogies using the program ms (Hudson 2002) with some simplifying assumptions: the locus analyzed is subject to neutral evolutionary processes only (i.e., no selection), population sizes are constant, there is no migration between populations/species, and there is no recombination. We calculated the *gsi* for both groups at a range of divergence times between 0 and 16 N_e generations. Each *gsi* calculation was performed with a sample of $n = 5$, 10, or 20 gene copies for each group. Power of the *gsi* as a test statistic to reject the null hypothesis of mixed genealogical ancestry was evaluated by determining the proportion of *P*-values ≤ 0.05. For each divergence time point and sample size we calculated $\bar{x}(gsi)$ for 10^3 replicate genealogies, and the probability for each genealogy using permutation with 10^4 replicates. We then compared the power of the *gsi* statistic to reject the null hypothesis with the proportion of monophyletic genealogies, which is the power of monophyly to reject the null hypothesis of a single (mixed) group.

To provide a reference to a standard statistic, we also compare the *gsi* to F_{ST} (Wright 1951). Values of F_{ST} were estimated from the simulated haplotypes output from the program ms using an appropriate calculation (Hudson et al. 1992). Despite differences in their calculation and interpretation, comparison of the *gsi* to F_{ST} provides reference to a well-established method of measuring divergence. Among the fundamental differences between the statistics is that F_{ST} yields a single measure of differentiation among subpopulations whereas the *gsi* measures the degree of exclusivity for each group in a genealogy. To limit the consequences of the differences in these measures, our simulations were conducted with fully symmetric groups (e.g., equal samples sizes, and equal divergence time since a common ancestral group), thus making the expected *gsi* values identical for all groups at the same divergence times in the simulation. This design allows us to compare the *gsi* value for a single group to F_{ST}.

To complete the large number of analyses required for this study we used Grid computing (Cummings and Huskamp 2005) through The Lattice Project (Bazinet and Cummings 2008). A Grid service calculating *gsi* and associated *P*-values was developed using a special programming library and associated tools (Bazinet et al. 2007). Following the model of a previous study (Cummings et al. 2003), which used an earlier Grid system (Myers and Cummings 2003), we used Grid computing to distribute required files among many computers in which the analyses were conducted asynchronously in parallel.

With increasing divergence time, $\bar{x}(gsi)$ increased to 1, the maximum value when exclusivity of gene copies within a group reached monophyly (Fig. 3, left panel). Depending on the number of gene copies in the sample, $\bar{x}(gsi) > 0.95$ was reached by 3.2–4.8 N_e generations and $\bar{x}(gsi) > 0.99$ by 6.4–8.0 N_e generations. Power of the *gsi* to reject the null hypothesis increased

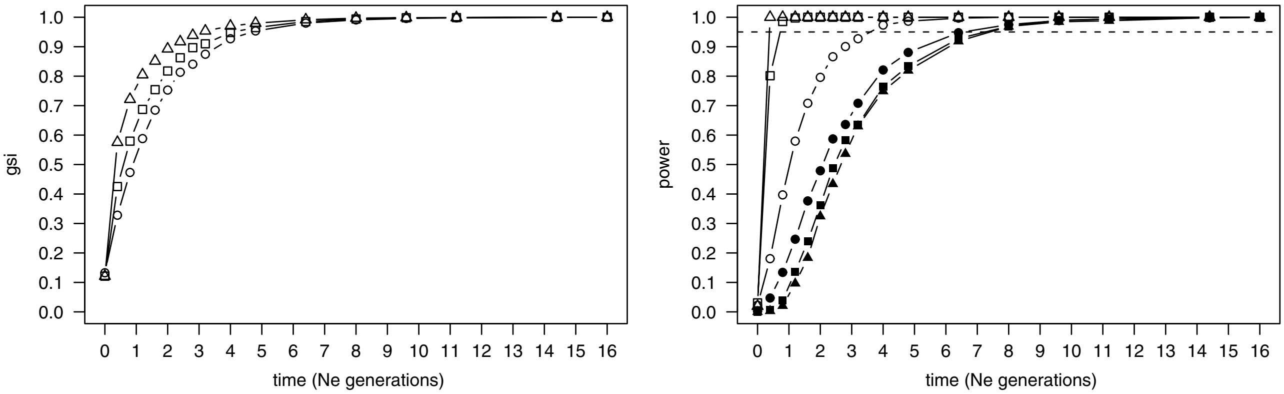

Figure 3. Relationships between population divergence in N_e generations (abscissa) and the genealogical sorting index (*gsi*, ordinate, left panel), and statistical power as a function of sample size (ordinate, right panel) based on coalescent simulations of two groups. Power of the *gsi* (open symbols) is expressed as the proportion of permutation tests with $P < 0.05$; power (probability) of monophyly (filled symbols) is the proportion of replicate simulations in which at least one of the two groups is monophyletic. Each point represents the mean of 10^3 replicate simulations: circles, groups of size 5; squares, groups of size 10; and triangles, groups of size 20. Dashed line designates power of 0.95.

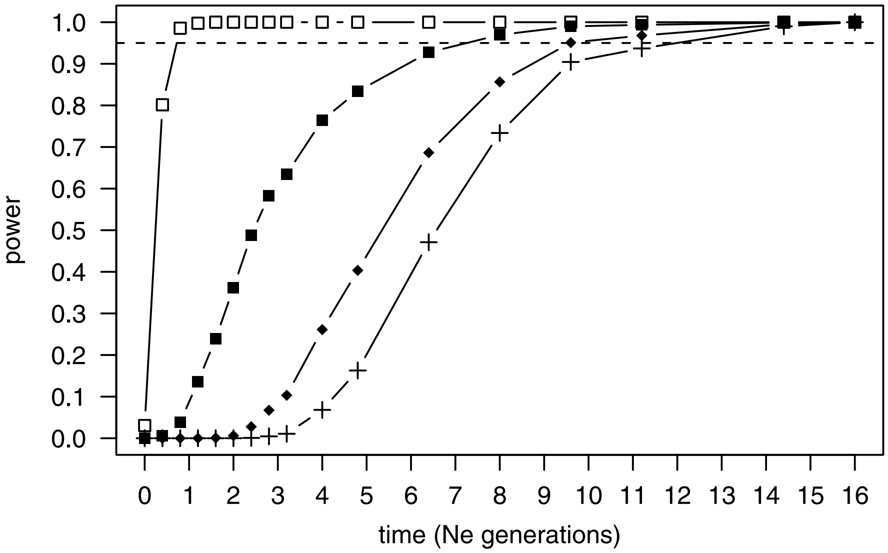

Figure 4. Statistical power as a function of the number of loci examined and population divergence in N_e generations based on 10^3 coalescent simulations of two groups of size 10. Power of the genealogical sorting index (open squares) is expressed as the proportion of permutation tests with $P < 0.05$, and is the same for all numbers of loci; power (probability) of monophyly (filled symbols) is the proportion of replicate simulations in which at least one of the two groups is monophyletic. Each point represents the mean of 10^3 replicate simulations: squares, 1 locus; diamonds, 5 loci; and pluses, 10 loci. Dashed line designates power of 0.95.

markedly, and much more rapidly than the *gsi* value itself, at early divergence times (Fig. 3, right panel). The power to detect significant genealogical structuring using the *gsi* was substantially higher than the probability of monophyly of lineages for any given time and number of gene copies (Fig. 3, right panel). Furthermore, increasing the number of gene copies increased the power of the *gsi* to reject the null hypothesis for a given divergence time. In contrast, the probability of monophyly at any given divergence time decreases as the number of gene copies increases (Fig. 3, right panel). With increasing numbers of loci sampled the expected $\overline{x}(gsi_T)$ remains constant (Fig. 4), and the variance declines (results not shown). In contrast, the expected times to monophyly increase with the number of loci sampled (Fig. 4).

The mean *gsi* values increase much more rapidly than $\overline{x}(F_{st})$ and reach maximum values at much earlier divergence times (Fig. 5, left panel). The power for the *gsi* is much greater than F_{ST} in the simulations, rejecting the null hypothesis of a single mixed population with a probability of 0.95 approximately 4 N_e generations earlier (Fig. 5, right panel). These results suggest that the *gsi* may have more power to detect divergence among subpopulations than F_{ST} in other cases in which where both statistics might be appropriately applied. The *gsi* has the additional benefit of providing a separate measure for each group on a tree.

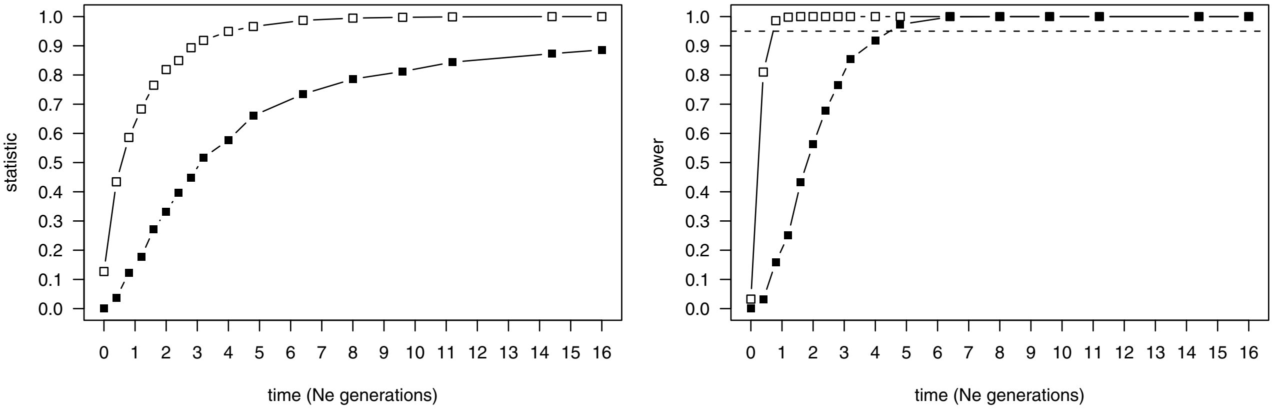

Figure 5. Relationships between population divergence in N_e generations (abscissa) and the genealogical sorting index (*gsi*, open squares) and F_{ST} (filled squares). The statistical measures (ordinate, left panel) and statistical power (ordinate, right panel) are based on coalescent simulations of two groups of size 10. Each point represents the mean of 10^3 replicate simulations of two groups of size 10. Dashed line designates power of 0.95.

APPLICATIONS TO EMPIRICAL DATA

To evaluate the *gsi* with real data, we used nuclear gene genealogies from a study of historical, demographic and selective factors associated with speciation in field crickets (*Gryllus firmus*, *Gryllus pennsylvanicus*, and *Gryllus ovisopis*) (Broughton and Harrison 2003) and in dolphins (*Lagenorhynchus obliquidens* and *Lagenorhynchus obscurus*) (Hare et al. 2002).

Field crickets.—The cricket genealogies were based on DNA sequence data from introns of four nuclear protein coding genes: *Cam*, calmodulin; *Cyt-c*, cytochrome c; *Ef1*α, elongation factor 1α; and *Pgi*, phosphoglucose isomerase. These data are particularly interesting because the degree of exclusive ancestry varies among groups across genes and the species/gene combinations exhibit a broad range of both the *gsi* and *P*-values allowing us to examine the performance of the statistic. Each tip in a genealogy was assigned to a group representing one of three recognized species (Fig. 6).

We redrew the cricket nuclear gene trees of Broughton and Harrison (2003) using MacClade (Maddison and Maddison 1989). We then calculated the *gsi* for each species under the null hypothesis that *G. firmus*, *G. pennsylvanicus*, and *G. ovisopis* formed a single group of mixed genealogical ancestry. The *gsi* values across the trees for the four sequence regions for *G. firmus* ranged from 0.0655 to 0.6533 (Table 1). Only *Cam* was not significant ($P = 0.3162$). Values of the *gsi* for *G. pennsylvanicus* ranged from 0.0455 to 0.7292, and only the value for *Cyt-c* was significant ($P = 0.0001$) (Table 1). For all trees *G. ovisopis* has *gsi* values of 1, with the exception of *Ef1*α, which had a value of 0.0118. Correspondingly, all *gsi* values for *G. ovisopis* were significant with the exception of that for *Ef1*α ($P = 0.5364$). The value of gsi_T, the value for the ensemble of the four gene trees, was significant for each of the labeled groups (Table 1), thus indicating the potential power of gsi_T to detect divergent genealogical structure by integrating across genes even when there is little obvious grouping evident in individual gene genealogies.

This empirical example raises an important issue regarding the effect of the relative group sizes being compared and the size of a group relative to the size of the whole tree. As the number of groups increases, a wider range of *gsi* values is often possible because the minimum *gs* value is a function of the total size of all

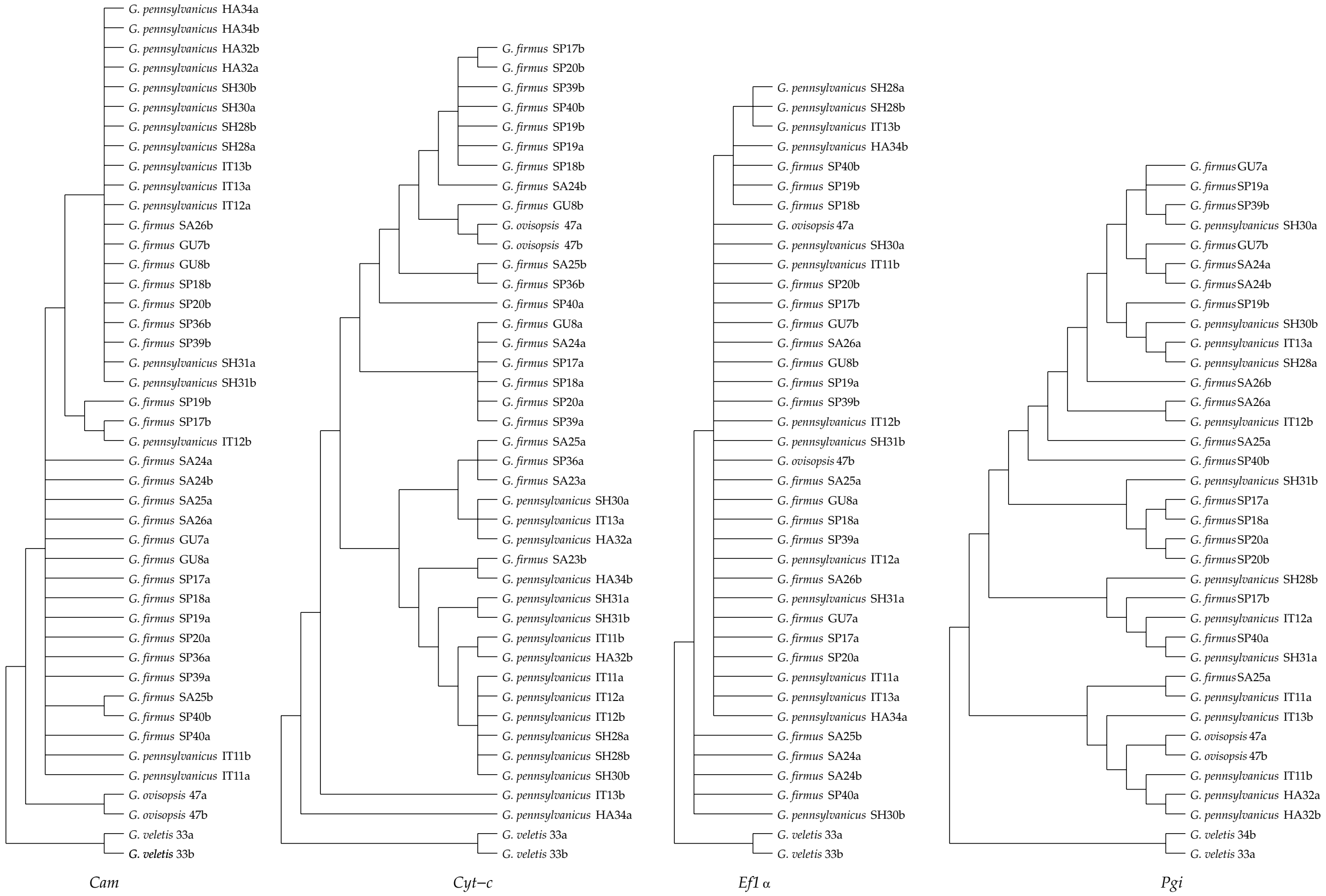

Figure 6. Gene genealogies for field crickets based on intron sequence data for the *Cam*, *Cyt-c*, *Ef1*α, and *Pgi* from Broughton and Harrison (2003). Gene copies are labeled by recognized species designations. Note these trees may differ slightly in topology from those presented in Broughton and Harrison (2003) due to difficulty in interpreting some short internal branches in the original figures. The results presented in Tables 1 and 2 are consistent with trees in this figure.

Table 1. Genealogical sorting index (*gsi*) and probability values for gene trees of Broughton and Harrison (2003) under the null hypothesis that gene copies labeled *Gryllus firmus*, *Gryllus pennsylvanicus,* and *Gryllus ovisopis* are a single mixed group.

	Gryllus firmus		*Gryllus pennsylvanicus*		*Gryllus ovisopis*	
Gene	*gsi*	*P*	*gsi*	*P*	*gsi*	*P*
Cam	0.0655	0.3162	0.0455	0.2226	1	0.0048
Cyt-c	0.6533	0.0001	0.7292	0.0001	1	0.0082
Ef1-a	0.0750	0.0228	0	1	0.0118	0.5364
Pgi	0.2861	0.0271	0.1444	0.3859	1	0.0193
all (gsi_T)	0.2700	0.0001	0.2298	0.0051	0.7529	0.0001

Table 2. Genealogical sorting index (*gsi*) and probability values for gene trees of Broughton and Harrison (2003) under the null hypothesis that gene copies labeled *Gryllus firmus* and *Gryllus pennsylvanicus* are a single mixed group.

	Gryllus firmus		*Gryllus pennsylvanicus*	
Gene	*gsi*	*P*	*gsi*	*P*
Cam	0	1	0.0164	0.5532
Cyt-c	0.7109	0.0001	0.7197	0.0001
Ef1-a	0.0909	0.0120	0	1
Pgi	0.2335	0.1541	0.1389	0.3546
all (gsi_T)	0.2588	0.0001	0.2188	0.0001

groups combined (eq. 3). Further, disparity in group size leads to a wider range of possible *gsi* values in small groups, and a narrower range of possible *gsi* values in larger groups. Correspondingly, opportunities for observing significant values decrease for larger groups compared to smaller groups. Thus in strongly unbalanced sampling designs, larger groups may have high *gsi* values, but the probability of observing such values will be increased and thus may not be significant. At the same time it is possible that a small group of only moderate apparent exclusivity could have a highly significant *gsi* value. This pattern is not a function of the *gsi* per se, but a consequence of constraints on the possible distributions of the group labels on a tree. For large groups that comprise the majority of a tree with unbalanced group representation, constraints in the array of possible label distributions results in decreased power to detect significant exclusive relationships. Thus, it is possible to make the *gsi* value for a group of moderate exclusivity significant because the group comprises a small proportion of a tree. Clearly it is essential that the null hypothesis and groups for comparison be carefully considered and that the sample design be appropriate for the hypothesis to be tested. As with most group comparisons in any context, ideally, sample sizes among groups should be similar and the groups of interest should comprise a substantial proportion of the sample. In the example here, the sample size for *G. ovisopis* is substantially smaller than the sample sizes for *G. firmus* and *G. pennsylvanicus*, and generally demonstrates a sampling design to be avoided when applying *gsi* if hypotheses regarding *G. ovisopis* were of interest. Similar consequences of small or unequal sampling among groups were noted by Rosenberg (Rosenberg 2007) in an examination of the chance occurrence of monophyly.

To examine only the groups with relatively similar sample sizes we tested a second null hypothesis that gene copies labeled *G. firmus* and *G. pennsylvanicus* formed a single intermixed group. To test this hypothesis we removed the data for *G. ovisopis* because this group had a much smaller sample size. Compared to the results for the first hypothesis, the *gsi* values only differed in those cases in which gene copies labeled *G. ovisopis* were previously nested within lineages composed of other labeled groups (Fig. 6, Tables 1 and 2). The values of the *gsi* for *Cyt−c* were significant for *G. firmus* and *G. pennsylvanicus* ($P = 0.0001$); additionally *Ef1*α was significant for *G. firmus* ($P = 0.0120$). As in the previous comparison that included *G. ovisopis*, gsi_T was significant for *G. firmus* and *G. pennsylvanicus* (Table 2).

Dolphins.—In addition to showing the utility of the *gsi* in detecting genealogical structure despite absence of reciprocal monophyly, our second empirical example demonstrates how uncertainty in phylogenetic relationships is incorporated into calculation of the *gsi*. We reanalyzed sequence data from introns of four nuclear protein coding genes: ACT, actin; BTM, Butyrophilin; CAMK, calcium calmodulin-dependent kinase; and HEXB, lysosomal beta-hexosaminidase for two dolphin species, *L. obliquidens* (Pacific white-sided dolphin) and *L. obscurus* (dusky dolphin), using alignments from Hare et al. (2002). We used PAUP* (Swofford 2003) to estimate a maximum likelihood tree and parameters for a general time reversible substitution model (Tavare 1986) with among-site rate variation modeled with invariable sites and a gamma distribution (Yang 1994; Gu et al. 1995) using a successive approximation estimation procedure (Swofford and Sullivan 2003). In a bootstrap analysis substantial computational savings can be realized without significant differences in the results by using the estimates of model parameters from the original data, rather than reestimating the parameter values for each bootstrap replicate (Cummings et al. 2003). Hence, we used the model parameters estimated for the original samples and performed bootstrap analyses with 2000 replicates and we saved all trees. We also generated a 50% majority-rule bootstrap consensus tree for each set of bootstrap replicates (Fig. 7).

To incorporate uncertainty in the inferred gene genealogies into our estimation of exclusivity we calculated gsi_T across all bootstrap trees (Table 3). The consequence of using all weighted topologies from a bootstrap analysis is illustrated in the data for dolphins in which there is no evidence for genealogical

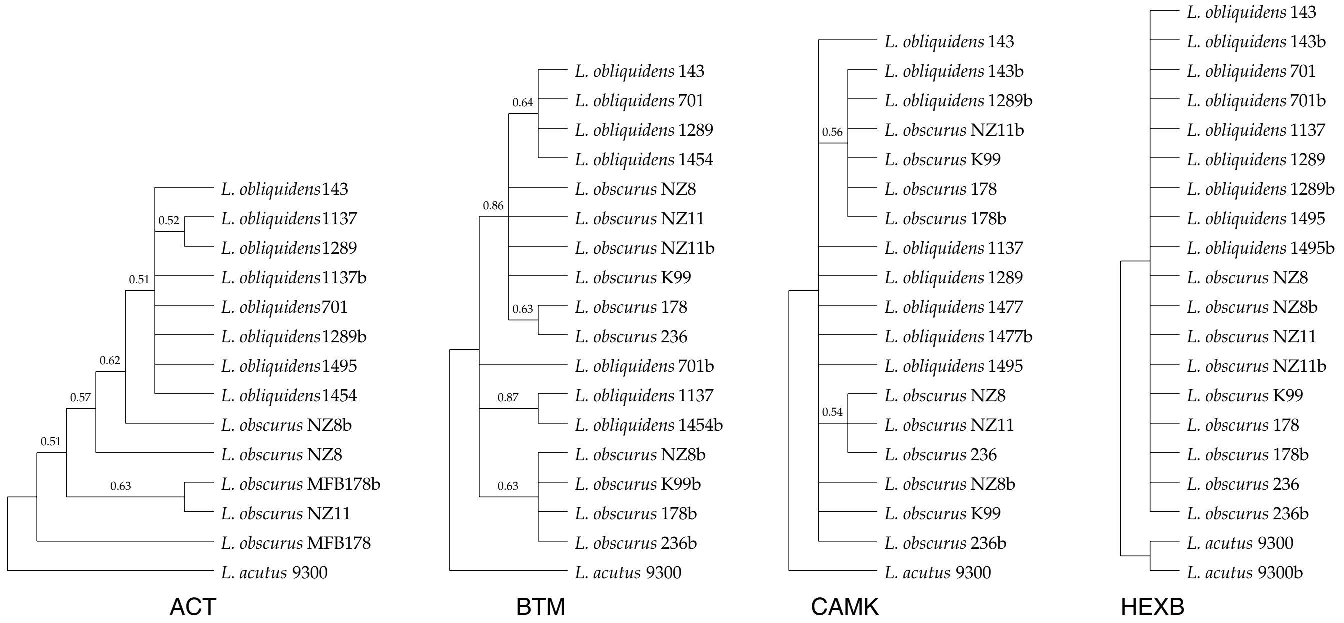

Figure 7. **Bootstrap consensus trees of gene genealogies for *Lagenorhynchus obliquidens* (Pacific white-sided dolphin) and *Lagenorhynchus obscurus* (dusky dolphin) based on intron sequence data for the ACT, BTM, CAMK, and HEXB reanalyzed from Hare et al. (2002). Gene copies are labeled by recognized species designations.**

sorting in the consensus tree (Fig. 7), but there is such evidence when bootstrap replicate topologies are used to calculate the *gsi* (Table 3). Calculating the ensemble statistic, gsi_T, on the full array of bootstrap replicates or most credible trees provides a robust means of detecting divergence that is not possible if uncertainty is not appropriately considered.

Discussion

As the number of gene-tree investigations has accumulated in the study of evolutionary genetics and speciation, the number of species exhibiting paraphyletic and polyphyletic gene trees has also increased (e.g., Funk and Omland 2003). The widely accepted explanation for these observations is that the time required to evolve reciprocal monophyly between diverging lineages is longer than the time since initial isolation (Tajima 1983; Hudson and Coyne 2002; Rosenberg 2003). From a phylogenetic standpoint, the interpretation of paraphyly allows one to assert little more than an absence of phylogenetic support for the historical integrity of a grouping (e.g., monophyly of a species). However, because the transition from polyphyly to monophyly is a continuous evolutionary process, a nonrandom distribution of allelic ancestry begins to accumulate within daughter lineages long before they have reached a state of reciprocal monophyly (Fig. 3, right panel). Until now, there has been no metric available to measure this accumulation of common genetic ancestry prior to monophyly. Together with permutation, the *gsi* enables one to test the hypothesis of significant genealogical divergence at a given locus well before monophyly is achieved.

Table 3. **Ensemble genealogical sorting index (gsi_T) from bootstrap replicates (calculated using eq. 5), and probability values for each gene intron region based on data from Hare et al. (2002) under the null hypothesis that gene copies labeled *Lagenorhynchus obscurus* and *Lagenorhynchus obliquidens* form a single, genealogically mixed, group.**

	Lagenorhynchus obscurus		*Lagenorhynchus obliquidens*	
Gene	gsi_T	*P*	gsi_T	*P*
ACT	0.3802	0.0089	0.7640	0.0006
BTM	0.2885	0.0001	0.1213	0.0173
CAMK	0.3531	0.0002	0.2474	0.0015
HEXB	0.0353	0.0112	0.1085	0.0002
all (gsi_T)	0.2642	0.0001	0.3103	0.0005

As such, the *gsi* meets an important need in evolutionary biology. The *gsi* is also (1) intuitive; (2) simple and easily calculated; (3) normalized to facilitate comparisons among groups and trees; (4) applicable to each individual group separately rather than providing a composite measure for all groups; (5) applicable to trees with polytomies; (6) accommodating to uncertainty in phylogenetic relationships as measured by bootstrap or posterior probability values; and (7) quantifiable on any tree regardless of how it was generated. Potential limitations are not a function of *gsi* itself but are inherent in the use and interpretation of many statistics, particularly those related to trees. For example, multiple topological arrangements can result in identical levels of genealogical exclusivity. Hence, *gsi* values and their significance must be interpreted in context of the biology of the groups of interest. Additionally, it is important to keep in mind that a *gsi* value for a single locus provides but one estimate of the genealogical patterns in a vast genome. More robust determinations regarding genealogical divergence can be derived from gsi_T and associated *P*-values determined by integrating across loci. Furthermore, as discussed above, the standard pitfalls associated with unequal sample sizes among groups will affect the power of *gsi* with the smaller group having inflated power and the larger group having decreased power relative to a balanced sampling design.

The results of our simulations demonstrate that the *gsi* tracks the expected gene-tree transition from polyphyly to paraphyly to monophyly. In the initial stages of divergence, *gsi* values are at or near 0, thus appropriately reflecting absence of exclusive ancestry. At the final stages of lineage sorting, *gsi* values reach 1 (Fig. 3, left panel), representing completely exclusive ancestry (i.e., monophyly). In between these two extremes, $\overline{x}(gsi)$ increases relatively rapidly. For example, depending on the number of gene copies sampled $\overline{x}(gsi)$ is > 0.95 at 3.2 to 4.8 N_e generations and >0.99 at 6.4 to 8.0 N_e generations, suggesting that exclusive relationship, or the amount of gene ancestry common to members of a lineage, accumulates rapidly once gene lineages begin to diverge. Evolution of reciprocal monophyly is expected to proceed much more slowly (Hudson and Coyne 2002) with a large variance in the actual time due to the stochastic nature of the coalescent process (Hudson and Turelli 2003). With respect to the ability to reject the hypothesis of a single genetically mixed group the power of *gsi* is much greater than monophyly at any particular time point (Fig. 3, right panel). Additionally, with increasing numbers of loci sampled for the calculation of gsi_T, although there is no change in expected mean value, the variance in *gsi* declines through averaging (results not shown). In contrast, expected times to monophyly increase with the number of loci sampled due to the multiplicative effects on the probability of multiple independent events (Fig 4; Hudson and Coyne 2002). Furthermore, in comparison to F_{ST}, estimates of the *gsi* increase more rapidly and reach the maximum possible value at shorter divergence times (Fig. 5, left panel). Correspondingly the *gsi* is a much more powerful measure of divergence, rejecting the null hypothesis of a single mixed population with 0.95 probability approximately 4 N_e generations before F_{ST} (Fig. 5, right panel).

Applying the *gsi* to real data, we found evidence of significant genealogical divergence for the field cricket *G. firmus* at three of the four protein-coding loci sampled by Broughton and Harrison (2003) and for *G. pennsylvanicus* at one of these loci. Moreover we showed significant genealogical divergence for both of these two species when *gsi* values from the four separate gene trees were integrated into an ensemble statistic, gsi_T. This quantification provides a powerful tool. In absence of such a statistic Broughton and Harrison (2003) concluded that only a small fraction of the genomes need to differentiate for speciation to occur. Whether this is the case, we document here that in fact substantial genealogical differentiation is evident among these species despite the lack of monophyly for any single gene tree (Tables 1 and 2). In the other empirical example we were also able to refute the hypothesis that the dolphin species *L. obscurus* and *L. obliquidens* form a single genealogical group, although monophyly is not observed at any individual locus for either species. The evidence in dolphins is more powerful than for crickets in that all *gsi* values at individual loci are significant when the ensemble *gsi* is calculated using all bootstrap replicates.

The ability to detect significant genealogical divergence before the condition of monophyly has evolved may prove useful for species delimitation. If phylogenetic differentiation is a criterion for species distinctions (e.g., Shaw 1998), the *gsi* provides a tree-based measure based on genealogical exclusivity to document boundaries between very young species. On the other hand, if the species criterion is not one of phylogenetic differentiation, but some other criterion such as reproductive incompatibility (Mayr 1963), then the *gsi* provides a robust, quantitative tool for evaluating the statement that speciation can occur prior to significant genomic differentiation (e.g., such as was claimed by Broughton and Harrison (2003), but refuted above based on analysis using the *gsi*).

Although we describe the *gsi* in the context of gene genealogies, the statistic has broad application to any situation in which it is desirable to quantify the degree of exclusive association among group representatives on a tree. Group labels are not limited to operational taxonomic units; rather they can represent any classification of interest such as one based on ecological characteristics, geographic locations, gene expression patterns, or biochemical pathways. For example, a tree might be built from sequence data for phytophagous insects and the tips labeled with the host plant with which individual insects were associated. The *gsi* and associated *P*-value could then be used to quantify the exclusivity of groups and to test hypotheses regarding group associations defined by host.

There is no underlying requirement as to how a tree is constructed. In a phylogenetic context the tree could be based on distances or discrete characters and it could be generated by distance, parsimony, maximum likelihood, or Bayesian methods. More broadly, a tree could be the result of any cluster analysis. However, when considering the transition from polyphyly to paraphyly to monophyly the *gsi* is only appropriately applied to hierarchical tree-like histories within species such as estimated gene-genealogies.

In summary, the *gsi* is a unique and powerful statistic that measures the exclusivity of lineages. It presents a novel and objective way to quantify genealogical structure by assessing the amount of exclusive ancestry of a group on a tree and determining the probability of observing that amount of exclusive ancestry at random in a single lineage. The *gsi* quantifies the historical relationships among groups, independent of any specific topology or distribution of coalescent times, to enable novel insight into the evolutionary process. The *gsi* thereby transcends the categorical view of monophyly and nonmonophyly characteristic of phylogenetic systematics. Likewise, *gsi* captures historical information about diverging populations from its quantification of exclusive relationship, independent of a reliance on estimates of coalescent times characteristic of historical population genetics. By focusing on this property of diverging lineages, evidence of differentiation is apparent much earlier than monophyly or topological congruence among genealogies. Thus, the *gsi* is a powerful statistic that may help bridge evolutionary insights gained via phylogenetic systematics and historical population genetics.

ACKNOWLEDGMENTS

We thank A. L. Bazinet and D. S. Myers for assistance with programming; A. L. Bazinet, M. Conte, and C. Milliron for assistance with simulations; M. Hare, R. Page, S. P. Otto, J. Sullivan, and two anonymous reviewers for helpful comments on an earlier version of the manuscript. MPC is supported in part by grants from the US National Science Foundation, Department of Energy, and a gift from Fujitsu Laboratories of America, Inc. KLS is supported in part by a grant from the US National Science Foundation. MCN is supported in part by funding from the US Fish and Wildlife Service and the Department of Defense.

LITERATURE CITED

Avise, J. C., and R. M. Ball, Jr. 1990. Principles of genealogical concordance in species concepts and biological taxonomy. Oxf. Surv. Evol. Biol. 7:45–67.

Avise, J. C., R. A. Lansman, and R. O. Shade. 1979. Use of restriction endonucleases to measure mitochondrial-DNA sequence relatedness in natural populations .1. Population structure and evolution in the genus *Peromyscus*. Genetics 92:279–295.

Baum, D. A., and K. L. Shaw. 1995. Genealogical perspectives on the species problem. Pp. 289–303 *in* P. C. Hock, and A. G. Stevenson, ed., Experimental and molecular approaches to plant biosystematics. Missouri Botanical Garden, St. Louis.

Bazinet, A. L., and M. P. Cummings. 2008. The Lattice Project: a grid research and production environment combining multiple grid computing models In press *in* M. H. W. Weber, ed. Distributed & Grid Computing—Science Made Transparent for Everyone. Rechenkraft.net, Marburg.

Bazinet, A. L., D. S. Myers, J. Fuetsch, and M. P. Cummings. 2007. Grid services base library: a high-level, procedural application program interface for writing Globus-based grid services. Future Gener. Comp. Sy. 23:517–522.

Bickel, P. J., and Van Zwet, W. R. 1978. Asymptotic expansion for the power of distribution free tests in the two-sample problem. Ann. Stat. 6:987–1007.

Broughton, R. E., and R. G. Harrison. 2003. Nuclear gene genealogies reveal historical, demographic and selective factors associated with speciation in field crickets. Genetics 163:1389–1401.

Cummings, M. P., and J. C. Huskamp. 2005. Grid computing. Educause Review 40:116–117.

Cummings, M. P., S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas, and K. Winka. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52:477–487.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. Plos Genet. 2:762–768.

Edwards, S. V., S. B. Kingan, J. D. Calkins, C. N. Balakrishnan, W. B. Jennings, W. J. Swanson, and M. D. Sorenson. 2005. Speciation in birds: genes, geography, and sexual selection. Proc. Natl. Acad. Sci. USA 102:6550–6557.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

Funk, D. J., and K. E. Omland. 2003. Species-level paraphyly and polyphyly: frequency, causes and consequences with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Syst. 34:397–423.

Gu, X., Y.-X. Fu, and W.-H. Li. 1995. Maximum likelihood estimation of the heterogeneity of substitution rates among nucleotide sites. Mol. Biol. Evol. 12:546–557.

Hare, M. P., F. Cipriano, and S. R. Palumbi. 2002. Genetic evidence on the demography of speciation in allopatric dolphin species. Evolution 56:804–816.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7:1–44.

———. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hudson, R. R., and J. A. Coyne. 2002. Mathematical consequences of the genealogical species concept. Evolution 56:1557–1565.

Hudson, R. R., and M. Turelli. 2003. Stochasticity overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57:182–190.

Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. Estimation of levels of gene flow from DNA-sequence data. Genetics 132:583–589.

Kingman, J. F. C. 1982. The coalescent. Stochas. Proc. Appl. 13:235–248.

Knowles, L. L., and B. C. Carstens. 2007. Delimiting species without monophyletic gene trees. Syst Biol 56:887–895.

Knowles, L. L., and W. P. Maddison. 2002. Statistical phylogeography. Mol. Ecol. 11:2623–2635.

Larget, B., and D. L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

Liu, L., and D. K. Pearl. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Machado, C. A., and J. Hey. 2003. The causes of phylogenetic conflict in a classic *Drosophila* species group. Proc. R Soc. Lond. B. 270:1193–1202.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Maddison, W. P., and D. R. Maddison. 1989. Interactive analysis of phylogeny and character evolution using the computer program MacClade. Folia Primatol. 53:190–202.

Maddison, W. P., and M. Slatkin. 1991. Null models for the number of evolutionary steps in a character on a phylogenetic tree. Evolution 45:1184–1197.

Mau, B., M. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.

Mayr, E. 1963. Animal species and evolution. Harvard Univ. Press, Cambridge, MA.

Moore, W. S. 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees. Evolution 49:718–726.

Myers, D. S., and M. P. Cummings. 2003. Necessity is the mother of invention: a simple grid computing system using commodity tools. J. Parallel Distrib. Comput. 63:578–589.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Rosenberg, N. A. 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57:1465–1477.

———. 2007. Statistical tests for taxonomic distinctiveness from observations of monophyly. Evolution 61:317–323.

Shaw, K. L. 1998. Species and the diversity of natural groups. Pp. 44–56 *in* D. J. Howard, and S. H. Berlocher, eds. Endless forms: species and speciation. Oxford Univ. Press, Oxford, England.

———. 2002. Conflict between mitochondrial and nuclear DNA phylogenies of a recent species radiation: what mitochondrial DNA reveals and conceals about modes of speciation in Hawaiian crickets. Proc. Natl. Acad. Sci. USA 99:16122–16127.

Sites, J. W., and J. C. Marshall. 2003. Delimiting species: a Renaissance issue in systematic biology. Trends Ecol. Evol. 18:462–470.

Sites, J. W., Jr., and J. C. Marshall. 2004. Empirical criteria for delimiting species. Annu. Rev. Ecol. Evol. Syst. 35:199–229.

Slatkin, M., and W. P. Maddison. 1990. Detecting isolation by distance using phylogenies of genes. Genetics 126:249–260.

Slatkin, M., and J. L. Pollack. 2006. The concordance of gene trees and species trees at two linked loci. Genetics 172:1979–1984.

Swofford, D. L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and Other Methods). Sinauer Associates, Sunderland, MA.

Swofford, D. L., and J. Sullivan. 2003. Phylogeny inference based on parsimony and other methods using PAUP*, Practice. Pp. 182–206 *in* M. Salemi, and A.-M. Vandamme, eds. The Phylogenetic handbook, a practical approach to DNA and protein phylogeny. Cambridge Univ. Press, Cambridge, England.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Takahata, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957–966.

Tavare, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math Life. Sci. 17:57–86.

Templeton, A. R. 2001. Using phylogeographic analyses of gene trees to test species status and processes. Mol. Ecol. 10:779–791.

Wright, S. 1951. The genetical structure of populations. Annals Eugen. 15:323–354.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

Associate Editor: S. Otto