doi:10.1111/j.1558-5646.2007.00023.x

STATISTICAL TESTS FOR TAXONOMIC DISTINCTIVENESS FROM OBSERVATIONS OF MONOPHYLY

Noah A. Rosenberg[1,2]

[1] *Department of Human Genetics, Bioinformatics Program, and the Life Sciences Institute, University of Michigan, Ann Arbor, Michigan 48109-2218*

[2] *E-mail: rnoah@umich.edu*

Received January 30, 2006
Accepted October 17, 2006

The observation of monophyly for a specified set of genealogical lineages is often used to place the lineages into a distinctive taxonomic entity. However, it is sometimes possible that monophyly of the lineages can occur by chance as an outcome of the random branching of lineages within a single taxon. Thus, especially for small samples, an observation of monophyly for a set of lineages—even if strongly supported statistically—does not necessarily indicate that the lineages are from a distinctive group. Here I develop a test of the null hypothesis that monophyly is a chance outcome of random branching. I also compute the sample size required so that the probability of chance occurrence of monophyly of a specified set of lineages lies below a prescribed tolerance. Under the null model of random branching, the probability that monophyly of the lineages in an index group occurs by chance is substantial if the sample is highly asymmetric, that is, if only a few of the sampled lineages are from the index group, or if only a few lineages are external to the group. If sample sizes are similar inside and outside the group of interest, however, chance occurrence of monophyly can be rejected at stringent significance levels ($P < 10^{-5}$) even for quite small samples ($\approx$ 20 total lineages). For a fixed total sample size, rejection of the null hypothesis of random branching in a single taxon occurs at the most stringent level if samples of nearly equal size inside and outside the index group—with a slightly greater size within the index group—are used. Similar results apply, with smaller sample sizes needed, when reciprocal monophyly of two groups, rather than monophyly of a single group, is of interest. The results suggest minimal sample sizes required for inferences to be made about taxonomic distinctiveness from observations of monophyly.

KEY WORDS: **Coalescent, genealogical species concept, gene trees, phylogeography.**

A set of lineages of a common type—species, orthologous genomic regions in different individuals, or other taxonomic entities that have a tree-like genealogy—is monophyletic if no other lineages under consideration descend from the most recent common ancestor of the set. Monophyly is often used in assembling members of lower-level taxa into higher-level taxa. For example, monophyly of the copies of a genomic region in a set of individuals (or monophyly for a large fraction of the regions of their genomes) might be used to group the individuals into a distinctive population or species (Avise and Ball 1990; Moritz 1994; Baum and Shaw 1995; Hudson and Coyne 2002).

Studies of monophyly in a genomic region typically gather genetic sequences from individuals of a group whose monophyly status is of interest, as well as from individuals belonging to one or more additional groups. Trees constructed from these sequences are then used to make inferences about monophyly for the group of interest (for representative examples, see Johnson et al. (2005); Steiper (2006); Weisrock et al. (2006)). Suppose that data on a lineages from group A and b lineages from other groups are analyzed, and that the set of a lineages is seen to be monophyletic. If monophyly of the set of a lineages is treated as an important requirement for identifying group A as distinctive, the use of the observation of monophyly for this purpose requires at least two conditions to hold. First, evidence must exist in support of the view that the *a* lineages are indeed monophyletic. Second, it must be improbable that the chance branching of $a + b$ lineages within a single taxonomic group could lead to monophyly of the specified set of *a* lineages. In other words, even if the *a* lineages are known with certainty to be monophyletic, sample sizes must be large enough that chance can be eliminated as the source of the observed monophyly.

One way of assessing the first condition is with a likelihood ratio test (Huelsenbeck et al. 1996). With this method, the likelihood of a set of DNA sequences is maximized separately under a null hypothesis of monophyly of the set, and under an alternative hypothesis in which the set is not constrained to be monophyletic. The null hypothesis of monophyly is rejected if the ratio of the unconstrained and constrained maxima is sufficiently large.

This type of test, however, does not address the second condition. If *a* and *b* are small, under a null hypothesis of a single taxonomic entity in which lineages follow a random-branching model, the probability is sizeable that in a collection of $a + b$ total lineages, a specific set of *a* lineages is monophyletic (Brown 1994; Rosenberg 2003). Thus, even if it is strongly supported by such approaches as that of Huelsenbeck et al. (1996), the observation of monophyly might potentially be attributable to chance rather than to the distinctiveness of group *A*. Monophyly of the *a* lineages from group *A* must be unlikely under a suitable null model to assert that group *A* is indeed distinctive. For the special case that the *b* lineages all belong to the same group, scenarios of chance monophyly under a null model and of monophyly resulting from a barrier to gene flow are shown in Figure 1.

In this article, using a random-branching model and assuming that the monophyly status of a set of lineages can be known with certainty, I develop statistical tests of the null hypothesis that monophyly has occurred by chance. These new methods can be viewed as a type of genealogy-based test of a null hypothesis that lineages have been sampled from a single taxonomic group. For the new tests, as a function of the statistical significance level, I determine the sample sizes needed in genealogical studies for avoiding the conclusion that monophyly is a consequence of chance, so that appropriate sample sizes for an investigation of monophyly can be chosen prior to the study. The results are described in terms of genetic lineages at a locus for individuals from the taxonomic "groups" for which taxonomic distinctiveness is being tested. These groups can be viewed as different populations of the same species, or as populations that represent different putative species.

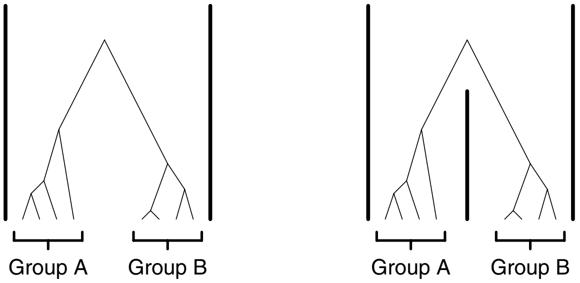

Figure 1. **Chance and isolation as causes of reciprocal monophyly. Lineages are classified as belonging to one of two groups, *A* and *B*. In the diagram on the left, the lineages of the two groups are reciprocally monophyletic by chance—both groups have small sample sizes, all lineages belong to the same unsubdivided population, and random branching within that population has by chance led to monophyletic groupings for *A* and *B*. In the diagram of the same genealogy on the right, the lineages of the two groups are reciprocally monophyletic, but instead as a result of genetic isolation (represented by the center line).**

TESTING FOR CHANCE OCCURRENCE OF MONOPHYLY

THE NULL MODEL

In the Yule model (Yule 1924; Harding 1971; Slowinski and Guyer 1989; Maddison and Slatkin 1991; Aldous 2001; Steel and McKenzie 2001; Rosenberg 2006), bifurcating genealogies are generated forward in time by a process in which each lineage has equal probability of being the next to branch into two; equivalently, they are generated backward in time in such a way that each pair of lineages has the same probability of being the next to coalesce. This model of random branching is a component of the coalescent model for the evolution of genealogical trees within populations (Nordborg 2003; Hein et al. 2005), and thus it provides a sensible null model for the branching of lineages within the taxonomic "groups" that we consider. Consequently, the observation of monophyly in a genealogical dataset can be viewed as a test statistic for the null hypothesis that the lineages in the dataset are drawn from a single taxonomic entity that evolves according to the Yule model. If the null hypothesis is rejected, then it is inferred that the random branching of the Yule model does not hold, perhaps because lineages were drawn from multiple distinctive groups separated by genetic barriers. If the null hypothesis is not rejected, then the observed monophyly is likely enough to have occurred by chance that it should not be regarded as evidence of a genealogical separation of groups.

Notice the distinction between these potential conclusions and those of the monophyly test of Huelsenbeck et al. (1996). In the test of Huelsenbeck et al. (1996), it is concluded that monophyly is either supported or not supported by the data. The test here, however, assumes that monophyly is supported and concludes that it either is or is not interesting—in other words, the new test examines whether monophyly has been produced by evolutionary processes or, of less interest, by insufficient sampling. I consider two versions of a test of the meaning of monophyly: one based on

Table 1. **Significance level at which the null hypothesis of random branching in a single taxon can be rejected, when monophyly of group *A* is observed. Given the sample sizes for each row and column, the probability of monophyly of group *A* is computed under the null hypothesis (eq. 1); a probability below 0.01 is assigned to the smallest integer power of 10 larger than or equal to it.**

Lineages from group *A* (*a*)	Lineages external to group *A* (*b*)								
	1	2	3	4	5	10	20	50	100
2	0.333	0.222	0.167	0.133	0.111	0.061	0.032	0.013	$<10^{-2}$
3	0.167	0.083	0.050	0.033	0.024	$<10^{-2}$	$<10^{-2}$	$<10^{-3}$	$<10^{-4}$
4	0.100	0.040	0.020	0.011	$<10^{-2}$	$<10^{-2}$	$<10^{-3}$	$<10^{-4}$	$<10^{-5}$
5	0.067	0.022	$<10^{-2}$	$<10^{-2}$	$<10^{-2}$	$<10^{-3}$	$<10^{-4}$	$<10^{-5}$	$<10^{-7}$
10	0.018	$<10^{-2}$	$<10^{-3}$	$<10^{-3}$	$<10^{-4}$	$<10^{-5}$	$<10^{-7}$	$<10^{-10}$	$<10^{-13}$
20	$<10^{-2}$	$<10^{-3}$	$<10^{-4}$	$<10^{-4}$	$<10^{-5}$	$<10^{-8}$	$<10^{-11}$	$<10^{-17}$	$<10^{-22}$
50	$<10^{-3}$	$<10^{-4}$	$<10^{-5}$	$<10^{-6}$	$<10^{-7}$	$<10^{-12}$	$<10^{-18}$	$<10^{-30}$	$<10^{-41}$
100	$<10^{-3}$	$<10^{-5}$	$<10^{-6}$	$<10^{-8}$	$<10^{-9}$	$<10^{-15}$	$<10^{-24}$	$<10^{-41}$	$<10^{-60}$

monophyly of the lineages from group *A* when the lineages not from *A* have arbitrary sources, and one based on the reciprocal monophyly of the lineages of groups *A* and *B* when all lineages derive from one of these two groups.

TESTING FOR CHANCE OCCURRENCE OF MONOPHYLY OF *A*

Probability of monophyly of A under the null model

Consider $c = a + b$ total lineages, where a of the lineages derive from group A, and the remaining b lineages are from groups other than A. We are interested in how to interpret an observation of monophyly for the a lineages from group A. Under the Yule model, the lineages of group A are monophyletic with probability (Rosenberg 2003, eq. 11)

$$P_A(a,\ b) \quad = \quad \frac{2}{\binom{a+b}{a}}\,\frac{a+b}{a(a+1)}. \qquad (1)$$

Using equation (1), Table 1 gives rejection probabilities for the null hypothesis that monophyly of the a specific lineages is due to random branching of the c total lineages. Provided that several lineages are considered both from group A and not from group A, the rejection probability rapidly becomes quite small: for example, $P_A(6, 6) \approx 6.18 \times 10^{-4}$.

Minimal sample sizes for rejecting the null model

We can use equation (1) to evaluate the sample sizes required for rejecting the null model at specified significance levels. It is straightforward to show that the function P_A (eq. 1) is monotonically decreasing in both a and b. Because of the decreasing nature of P_A as a increases, for any α between 0 and 1 and a fixed value of b, equation (1) can be used to obtain the minimal sample size a needed so that an observation of monophyly enables rejection of the null hypothesis at significance level α. If $b = 1$, that is, if the meaning of monophyly of group A is tested using a lineages from A and one lineage not from A, the minimal number of lineages needed is

$$a \quad = \quad \left\lceil \frac{\sqrt{\alpha^2 + 8\alpha}}{2\alpha} - \frac{1}{2} \right\rceil \qquad (2)$$

where $\lceil x \rceil$ is the smallest integer larger than or equal to x. More generally, by numerically solving equation (1) for a in terms of b and α, Table 2 gives the minimal sample size needed from group A to achieve $P_A \leq \alpha$.

Table 2. **For a fixed sample size external to group *A*, the minimal sample size needed from group *A* so that the probability of monophyly of group *A* is less than or equal to α ($P_A \leq \alpha$).**

Lineages external group *A* (*b*)	α					
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
1	4	14	45	141	447	1414
2	3	7	16	34	74	159
3	3	5	10	18	33	58
4	3	5	8	13	21	33
5	3	4	7	10	16	24
6	2	4	6	9	13	19
7	2	4	6	8	11	16
8	2	4	5	7	10	14
9	2	3	5	7	9	12
10	2	3	5	6	9	11
20	2	3	4	5	6	8
50	2	3	3	4	5	6
100	2	2	3	3	4	5
200	2	2	3	3	4	4
500	2	2	3	3	3	4
1000	2	2	2	3	3	3

Similarly, for a fixed value of a, the minimal value of b can be determined so that with a lineages from group A and b lineages not from group A, the probability under the null hypothesis that the lineages of A are monophyletic is no larger than α. This computation is only sensible for $a \geq 2$, because for $a = 1$, monophyly is guaranteed. For $a = 2$, this minimal value of b is

$$b = \left\lceil \frac{2}{3\alpha} \right\rceil - 1. \tag{3}$$

For $a = 3$, it is

$$b = \left\lceil \frac{\sqrt{\alpha^2 + 4\alpha}}{2\alpha} - \frac{3}{2} \right\rceil. \tag{4}$$

More generally, as a function of a, Table 3 gives the minimal sample size needed outside group A to achieve $P_A \leq \alpha$.

Finally, for a fixed value of $c = a + b$, the minimal sample sizes needed from group A and not from A to achieve $P_A \leq \alpha$ can be determined from equation (1) (Tables 4 and 5). At a fixed value of c, the null hypothesis of random branching is rejected at the most stringent significance level when $a \approx b$ (Fig. 2). By finding the smallest value of a for which the rejection probability with sample sizes $a + 1$ and $c - (a + 1)$ exceeds that for sample sizes a and $c - a$, it can be shown that for fixed $c \geq 3$, the rejection probability is smallest when $a = (c + 2)/2$ (even c) or $a = (c + 1)/2$ (odd c). Thus, use of nearly equal sample sizes, with a slightly greater sample size within group A than outside group A, leads to the most stringent rejection probability. More generally, a slightly greater effort in obtaining samples from A rather than external to A produces more stringent rejection probabilities. This result can be seen from the asymmetry in P_A, that is, from the fact that $a > b$ leads to $P_A(a, b) < P_A(b, a)$.

Table 3. **For a fixed sample size from group A, the minimal sample size needed external to group A so that the probability of monophyly of group A is less than or equal to α ($P_A \leq \alpha$).**

Lineages from group A (a)	α					
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
2	6	66	666	6666	66666	666666
3	2	9	31	99	315	999
4	1	5	12	27	61	132
5	1	3	8	15	28	51
6	1	3	6	10	18	30
7	1	2	5	8	13	21
8	1	2	4	7	11	16
9	1	2	4	6	9	13
10	1	2	3	5	8	12
20	1	1	2	3	5	6
50	1	1	1	2	3	4
100	1	1	1	2	2	3
200	1	1	1	1	2	2
500	1	1	1	1	1	2
1000	1	1	1	1	1	2

Table 4. **For a fixed total sample size (c=a+b), the minimal sample size needed from group A so that the probability of monophyly of group A is less than or equal to α ($P_A \leq \alpha$).**

Total lineages (c=a+b)	α					
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
10	2	4	–	–	–	–
20	2	3	4	5	7	–
30	2	3	4	5	6	7
40	2	3	3	4	5	6
50	2	3	3	4	5	6
60	2	3	3	4	5	5
70	2	2	3	4	4	5
80	2	2	3	4	4	5
90	2	2	3	4	4	5
100	2	2	3	4	4	5
200	2	2	3	3	4	4
500	2	2	3	3	3	4
1000	2	2	2	3	3	4

TESTING FOR CHANCE OCCURRENCE OF RECIPROCAL MONOPHYLY

Probability of reciprocal monophyly under the null model

We again consider $c = a + b$ total lineages, where a of the lineages derive from group A. Suppose now that the b lineages not from group A all derive from a second group B, that the lineages of

Table 5. **For a fixed total sample size (c=a+b), the minimal sample size needed external to group A so that the probability of monophyly of group A is less than or equal to α ($P_A \leq \alpha$).**

Total lineages (c=a+b)	α					
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
10	1	2	–	–	–	–
20	1	1	2	4	6	–
30	1	1	2	3	4	5
40	1	1	2	2	3	4
50	1	1	1	2	3	4
60	1	1	1	2	3	4
70	1	1	1	2	3	3
80	1	1	1	2	2	3
90	1	1	1	2	2	3
100	1	1	1	2	2	3
200	1	1	1	1	2	2
500	1	1	1	1	1	2
1000	1	1	1	1	1	2

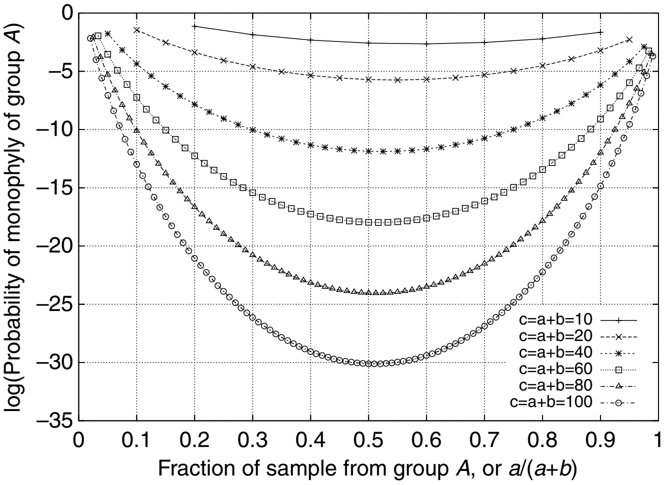

Figure 2. Logarithm (base 10) of the probability of monophyly of group *A* as a function of the fraction of a sample that derives from group *A*.

group A are monophyletic, and that the lineages of group B are separately monophyletic. We are now interested in how to interpret this observation of reciprocal monophyly. Under the Yule model, the lineages of the two groups are reciprocally monophyletic with probability (Brown 1994; Rosenberg 2003, eq. 9)

$$P_{AB}(a,\ b) \quad = \quad \frac{2}{\binom{a+b}{a}}\,\frac{1}{a+b-1}. \qquad (5)$$

Using equation (5), Table 6 gives rejection probabilities for the null hypothesis that reciprocal monophyly of the specified sets of lineages is due to random branching of the c total lineages. Reciprocal monophyly is less probable under the null hypothesis than monophyly of group A, so that with the same sample sizes, an observation of reciprocal monophyly of A and B leads to a smaller rejection probability than does an observation of monophyly of group A: for example, $P_{AB}(6, 6) = (7/22)P_A(6, 6) \approx 1.97 \times 10^{-4}$.

Minimal sample sizes for rejecting the null model

Analogously to the case in which monophyly of the lineages of group A is of interest, for fixed values of α and a, equation (5) can be used to calculate the minimal sample size b needed so that an observation of reciprocal monophyly enables rejection of the null hypothesis at significance level α (Table 7); because of the symmetry of the situation, the same results are obtained when b is fixed and a is allowed to vary. For a fixed value of c, equation (5) can be used to determine the minimal sample sizes needed from groups A and B to achieve $P_{AB} \leq \alpha$ (Table 8). At a fixed c, the null hypothesis is rejected most stringently when $a \approx b$ (Fig. 3); for fixed $c \geq 4$, it can be shown that the rejection probability is smallest when $a = c/2$ (even c) or $a = (c \pm 1)/2$ (odd c).

Discussion

The results here give rejection probabilities for the null hypothesis that sampled lineages follow a random-branching model (Tables 1 and 6), and minimal sample sizes necessary for rejecting this hypothesis from an observation of monophyly (Tables 2–5, 7, and 8). Thus, the methods presented can assist not only in evaluating the outcomes of genealogical studies of monophyly, but also in determining in advance the appropriate sample sizes that should be gathered for such studies.

It is observed that monophyly of a specified set of lineages is usually unlikely under the null hypothesis of random branching, that samples in which lineages from the index group constitute approximately half the total sample enable rejection of the null at the most stringent significance levels (Figs. 2 and 3), and that fairly small samples are sufficient for rejecting the null hypothesis.

Table 6. Significance level at which the null hypothesis of random branching in a single taxon can be rejected, when reciprocal monophyly is observed. Given the sample sizes for each row and column, the probability of reciprocal monophyly is computed under the null hypothesis (eq. 5); a probability below 0.01 is assigned to the smallest integer power of 10 larger than or equal to it.

Lineages from group	Lineages from group B (b)							
A (a)	2	3	4	5	10	20	50	100
2	0.111	0.050	0.027	0.016	$<10^{-2}$	$<10^{-3}$	$<10^{-4}$	$<10^{-5}$
3	0.050	0.020	$<10^{-2}$	$<10^{-2}$	$<10^{-3}$	$<10^{-4}$	$<10^{-5}$	$<10^{-6}$
4	0.027	$<10^{-2}$	$<10^{-2}$	$<10^{-2}$	$<10^{-3}$	$<10^{-5}$	$<10^{-6}$	$<10^{-8}$
5	0.016	$<10^{-2}$	$<10^{-2}$	$<10^{-3}$	$<10^{-4}$	$<10^{-5}$	$<10^{-7}$	$<10^{-9}$
10	$<10^{-2}$	$<10^{-3}$	$<10^{-3}$	$<10^{-4}$	$<10^{-6}$	$<10^{-8}$	$<10^{-12}$	$<10^{-15}$
20	$<10^{-3}$	$<10^{-4}$	$<10^{-5}$	$<10^{-5}$	$<10^{-8}$	$<10^{-12}$	$<10^{-18}$	$<10^{-24}$
50	$<10^{-4}$	$<10^{-5}$	$<10^{-6}$	$<10^{-7}$	$<10^{-12}$	$<10^{-18}$	$<10^{-30}$	$<10^{-42}$
100	$<10^{-5}$	$<10^{-6}$	$<10^{-8}$	$<10^{-9}$	$<10^{-15}$	$<10^{-24}$	$<10^{-42}$	$<10^{-60}$

Table 7. For a fixed sample size from group *A*, the minimal sample size needed from group *B* so that the probability of reciprocal monophyly is less than or equal to α ($P_{AB} \leq \alpha$).

Lineages from group *A* (*a*)	α					
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
2	3	7	15	33	73	158
3	2	4	9	17	32	57
4	2	3	7	12	20	32
5	2	3	5	9	14	22
6	2	3	5	7	12	17
7	2	2	4	6	10	14
8	2	2	4	6	9	12
9	2	2	3	5	8	11
10	2	2	3	5	7	10
20	2	2	2	3	4	6
50	2	2	2	2	3	4
100	2	2	2	2	2	3
200	2	2	2	2	2	2
500	2	2	2	2	2	2
1000	2	2	2	2	2	2

An important conclusion is that when monophyly of a particular set of lineages is of interest, incorporating several lineages outside that set into the analysis substantially decreases the total number of lineages that must be studied.

Reciprocal monophyly of two sets of lineages is less likely under random branching than monophyly of one set of lineages. Therefore, in the reciprocal monophyly case, smaller sample sizes are required for achieving corresponding levels of significance than in the case of monophyly of a single collection of lineages. Small sample sizes may be sufficient for very stringent rejection probabilities if for each of the two groups, at least several lineages are sampled (∼10).

Table 8. For a fixed total sample size ($c{=}a{+}b$), the minimal sample size needed from the group with smaller sample size so that the probability of reciprocal monophyly is less than or equal to α ($P_{AB} \leq \alpha$).

Total lineages ($c{=}a{+}b$)	α					
	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
10	2	2	5	–	–	–
20	2	2	2	3	5	8
30	2	2	2	3	4	5
40	2	2	2	2	3	4
50	2	2	2	2	3	4
60	2	2	2	2	3	3
70	2	2	2	2	3	3
80	2	2	2	2	2	3
90	2	2	2	2	2	3
100	2	2	2	2	2	3
200	2	2	2	2	2	2
500	2	2	2	2	2	2
1000	2	2	2	2	2	2

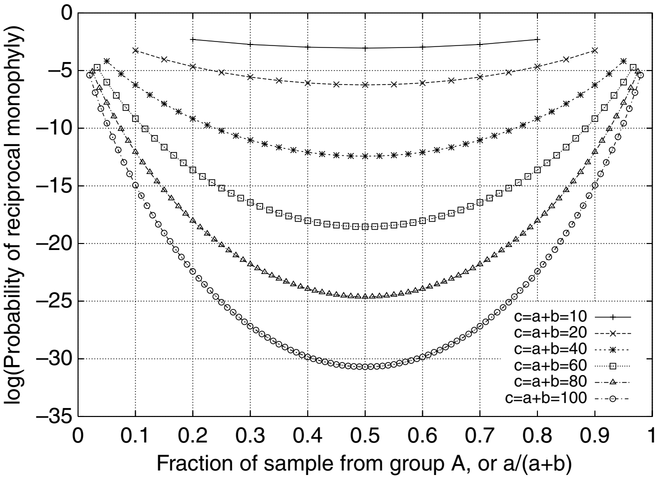

Figure 3. Logarithm (base 10) of the probability of reciprocal monophyly as a function of the fraction of a sample that derives from group *A*.

When one or both of the sample sizes is particularly small (<5), statements regarding monophyly should often be interpreted with caution. For small samples, however, the monophyly of lineages from the same set of individuals across a collection of multiple independent loci can assist considerably in rejecting a null hypothesis that there is no genealogical separation among groups. For *a* lineages from group *A* and *b* lineages outside group *A*, the probability under the null hypothesis of random branching that the corresponding *a* lineages across *L* loci are monophyletic for at least *k* of the loci is

$$\sum_{j=k}^{L} \binom{L}{j} P_A(a,\ b)^j \left[1 - P_A(a,\ b)\right]^{L-j}. \tag{6}$$

The analogous probability in the case of reciprocal monophyly, when the *b* lineages are all part of group *B*, is

$$\sum_{j=k}^{L} \binom{L}{j} P_{AB}(a,\ b)^j \left[1 - P_{AB}(a,\ b)\right]^{L-j}. \tag{7}$$

Even for small *a* and *b*, for *L* sufficiently large and *k* sufficiently close to *L*, these probabilities quickly become small, so that when multiple loci produce similar inferences, there is considerable potential for excluding chance as the cause of observations of monophyly.

A variant of the problem considered here was investigated by Nordborg (1998). A previous study (Krings et al. 1997) had examined mitochondrial DNA sequences from 986 modern humans and a Neanderthal, and had observed that the modern human sequences appeared to be monophyletic. To test if Neanderthals and modern humans could have mated randomly during the time of their coexistence, Nordborg (1998) commented that under plausible scenarios, the 986 sampled modern human sequences likely traced back to only a relatively small number of ancestral sequences extant at the time of the extinction of the Neanderthals. Thus, whereas a monophyletic sample of 986 modern human lineages concurrent with a single Neanderthal sequence would provide strong evidence that the Neanderthals and the modern humans were not a randomly mating population ($P_A(986, 1) \approx 2.06 \times 10^{-6}$), the observation of monophyly of a much smaller set of lineages ancestral to the 986 sampled lineages is considerably more equivocal (for example, $P_A(4, 1) = 0.1$). This difference in probabilities highlights the fact that if the samples from A and from external to A are taken from different points in time, computations of P_A and P_{AB}—following Nordborg (1998)—should make use of the distribution of the number of lineages ancestral to the more recent of the two samples, rather than the sample size in the present.

The results here are important for the application to genetic data of the genealogical species concept (Baum and Shaw 1995; Shaw 1998; Hudson and Coyne 2002), which delineates species by monophyly of the genetic lineages of their members, and in DNA barcoding (Hebert et al. 2003; Moritz and Cicero 2004; Meyer and Paulay 2005; Hickerson et al. 2006), which may also use monophyly in species demarcation. Such methods represent only a small subset of approaches to the identification of species; however, if these monophyly-based approaches are used, then sample sizes should be made sufficiently large that chance can be excluded as the cause of observed monophyly. The fact that the probability of monophyly under random branching decreases rapidly with sample size ensures that even with sample sizes as small as 10 inside and outside an index group, chance monophyly of the index group has probability below 2×10^{-6}. Thus, chance production of monophyly—which should not be disregarded for small samples—is not likely to cause misleading inferences about genealogical distinctiveness if reasonably large samples are used.

ACKNOWLEDGMENTS

I thank Michael Blum, Richard Harrison, Mark Tanaka, and three anonymous reviewers for suggestions. This work was supported by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences and by an Alfred P. Sloan Research Fellowship.

LITERATURE CITED

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat. Sci. 16:23–34.

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

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

Brown, J. K. M. 1994. Probabilities of evolutionary trees. Syst. Biol. 43:78–91.

Harding, E. F. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Prob. 3:44–77.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. B 270:313–321.

Hein, J., M. H. Schierup, and C. Wiuf. 2005. Gene genealogies, variation and evolution. Oxford Univ. Press, Oxford, U.K.

Hickerson, M. J., C. P. Meyer, and C. Moritz. 2006. DNA barcoding will often fail to discover new animal species over broad parameter space. Syst. Biol. 55:729–739.

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

Huelsenbeck, J. P., D. M. Hillis, and R. Nielsen. 1996. A likelihood-ratio test of monophyly. Syst. Biol. 45:546–558.

Johnson, J. A., R. T. Watson, and D. P. Mindell. 2005. Prioritizing species conservation: does the Cape Verde kite exist? Proc. R. Soc. Lond. B 272:1365–1371.

Krings, M., A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking, and S. Pääbo. 1997. Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30.

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.

Meyer, C. P., and G. Paulay. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. 3:2229–2238.

Moritz, C. 1994. Defining “evolutionary significant units” for conservation. Trends Ecol. Evol. 9:373–375.

Moritz, C., and C. Cicero. 2004. DNA barcoding: promise and pitfalls. PLoS Biol. 2:1529–1531.

Nordborg, M. 1998. On the probability of Neanderthal ancestry. Am. J. Hum. Genet. 63:1237–1240.

———. 2003. Coalescent theory. 2nd ed. Pp. 602–635 *in* D. J. Balding, M. Bishop, and C. Cannings, eds. Handbook of statistical genetics. Wiley, Chichester, UK.

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.

———. 2006. The mean and variance of the numbers of *r*-pronged nodes and *r*-caterpillars in Yule-generated genealogical trees. Ann. Comb. 10:129–146.

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

Slowinski, J. B., and C. Guyer. 1989. Testing the stochasticity of patterns of organismal diversity: an improved null model. Am. Nat. 134:907–921.

Steel, M., and A. McKenzie. 2001. Properties of phylogenetic trees generated by Yule-type speciation models. Math. Biosci. 170:91–112.

Steiper, M. E. 2006. Population history, biogeography, and taxonomy of orangutans (Genus: *Pongo*) based on a population genetic meta-analysis of multiple loci. J. Hum. Evol. 50:509–522.

Weisrock, D. W., H. B. Shaffer, B. L. Storz, S. R. Storz, and S. R. Voss. 2006. Multiple nuclear gene sequences identify phylogenetic species boundaries in the rapidly radiating clade of Mexican ambystomatid salamanders. Mol. Ecol. 15:2489–2503.

Yule, G. U. 1924. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. F. R. S. Phil. Trans. R. Soc. Lond. B 213:21–87.

Associate Editor: R. Harrison