Syst. Biol. 55(4):623–636, 2006

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150600863109

Incorporating Allelic Variation for Reconstructing the Evolutionary History of Organisms from Multiple Genes: An Example from *Rosa* in North America

SIMON JOLY AND ANNE BRUNEAU

Institut de recherche en biologie végétale, Université de Montréal, 4101 Sherbrooke Est, Montréal (Québec), Canada H1X 2B2; E-mail: simon.joly@umontreal.ca (S.J.); anne.bruneau@umontreal.ca (A.B.)

Abstract.—Allelic variation within individuals holds information regarding the relationships of organisms, which is expected to be particularly important for reconstructing the evolutionary history of closely related taxa. However, little effort has been committed to incorporate such information for reconstructing the phylogeny of organisms. Haplotype trees represent a solution when one nonrecombinant marker is considered, but there is no satisfying method when multiple genes are to be combined. In this paper, we propose an algorithm that converts a distance matrix of alleles to a distance matrix among organisms. This algorithm allows the incorporation of allelic variation for reconstructing the phylogeny of organisms from one or more genes. The method is applied to reconstruct the phylogeny of the seven native diploid species of *Rosa* sect. *Cinnamomeae* in North America. The glyceralgehyde 3-phosphate dehydrogenase (*GAPDH*), the triose phosphate isomerase (*TPI*), and the malate synthase (*MS*) genes were sequenced for 40 individuals from these species. The three genes had little genetic variation, and most species showed incomplete lineage sorting, suggesting these species have a recent origin. Despite these difficulties, the networks (NeighborNet) of organisms reconstructed from the matrix obtained with the algorithm recovered groups that more closely match taxonomic boundaries than did the haplotype trees. The combined network of individuals shows that species west of the Rocky Mountains, *Rosa gymnocarpa* and *R. pisocarpa*, form exclusive groups and that together they are distinct from eastern species. In the east, three groups were found to be exclusive: *R. nitida–R. palustris*, *R. foliolosa*, and *R. blanda–R. woodsii*. These groups are congruent with the morphology and the ecology of species. The method is also useful for representing hybrid individuals when the relationships are reconstructed using a phylogenetic network. [Allelic variation; gene tree–species tree; haplotype trees; hybridization; incomplete lineage sorting; phylogenetic networks; *Rosa*; total evidence.]

Allelic variation at autosomal loci holds information regarding the relationships of organisms. Indeed, using two alleles instead of one can give better estimations of phylogenetic relationships because twice the amount of information is provided. This is especially true of closely related taxa for which incomplete lineage sorting is more likely (Rosenberg, 2002, 2003; Degnan and Salter, 2005). In addition, allelic variation allows the detection of hybrid individuals with a single marker, whereas at least two are required when only one allele per locus is sampled. But in spite of the amount of data contained in allelic variation, little effort has been directed to date at incorporating such information for reconstructing the phylogenetic relationships of organisms.

One solution when a single nonrecombinant marker is considered is to use haplotype trees, which frequently are used in evolutionary studies of closely related species (Schaal and Olsen, 2000). At present, however, there is no phylogenetic method that can easily incorporate allelic variation for more than one gene for reconstructing the evolutionary history of individuals. Yet the importance of investigating several markers for reconstructing the phylogeny of species is widely recognized as any single gene can be incongruent with the evolutionary history of species (Pamilo and Nei, 1988; Takahata, 1989; Wu, 1991; Doyle, 1992; Maddison, 1997; Nichols, 2001; Rosenberg, 2002, 2003; Degnan and Salter, 2005).

Most current approaches used for reconstructing phylogenies from multiple markers, either using a total evidence (e.g., Kluge, 1989; Yang, 1996; Seo et al., 2005) or a consensus approach (e.g., de Queiroz, 1993), cannot incorporate allelic variation for multiple genes because they use haplotypes as terminal units of the analysis. Because it makes no sense to concatenate alleles from different loci that segregate in natural populations, such methods are limited to using a single haplotype per individual. If the individuals, rather than the alleles, were the terminals of the analysis, it would be possible to combine information from different genes.

In this paper, we propose an algorithm that incorporates allelic variation for reconstructing the phylogeny of organisms. The proposed algorithm converts a distance matrix of alleles into a distance matrix of organisms so that individuals become the terminals of the analysis. The matrix of organisms for one marker can either be used alone or in combination with other matrices obtained from independently evolving markers to reconstruct a phylogeny of organisms.

The algorithm is applied to reconstruct the evolutionary history of the seven native diploid species of *Rosa* sect. *Cinnamomeae* in North America using allelic variation at three nuclear loci for 40 individuals. Very little is known of the phylogenetic relationships of these rose species, mostly because of the poor species sampling of previous phylogenetic studies (e.g., Millan et al., 1996; Matsumoto et al., 1998). Moreover, the little molecular variation found among North American species (Wissemann and Ritz, 2005; Joly et al., 2006) limits our understanding of their relationships and suggests that these species are of recent origin. Consequently, incomplete lineage sorting (or deep coalescence) could be an important issue in this group as it is expected to be most severe among recently diverged species (Rosenberg, 2002, 2003; Degnan and Salter, 2005). Hybridization also could be a confounding evolutionary process because of the propensity of these roses to hybridize (Erlanson, 1934; Ratsek et al., 1939, 1940; Lewis and Basye, 1961).

Therefore, this group represents a good case study to test the proposed algorithm because of the potentially important additional information that allelic variation can provide.

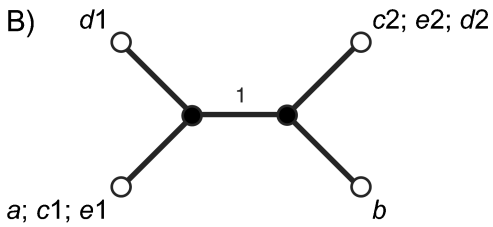
THE POFAD ALGORITHM

The POFAD (for Phylogeny of Organisms from Allelic Data) algorithm starts with a distance matrix of alleles for a given marker. The algorithm described below assumes that the organisms are diploids. The algorithm will be illustrated using a hypothetical example with five individuals (A to E) from which we have a haplotype distance matrix (Fig. 1A) that can be represented by a haplotype tree (Fig. 1B). In the example, letters are used to distinguish individuals: capital and lowercase letters represent individuals and alleles, respectively. Alleles within an individual are set apart by a number (1 or 2).

A)

	a	*b*	*c*1	*c*2	*d*1	*d*2	*e*1	*e*2
a	0							
b	3	0						
*c*1	0	3	0					
*c*2	3	2	3	0				
*d*1	2	3	2	3	0			
*d*2	3	2	3	0	3	0		
*e*1	0	3	0	3	2	3	0	
*e*2	3	2	3	0	3	0	3	0

B)



C)

	A	*B*	*C*	*D*	*E*
A	0				
B	3	0			
C	1.5	2.5	0		
D	2.5	2.5	1	0	
E	1.5	2.5	0	1	0

D)

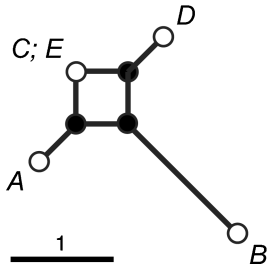


FIGURE 1. (A) Hypothetic haplotype distance matrix and (B) the unrooted haplotype tree obtained from it. (C) Matrix of distances between organism obtained from the haplotype distance matrix using the POFAD algorithm and (D) the NeighborNet network reconstructed from it. Letters distinguish individuals: capital and lowercase letters represent individuals and alleles, respectively. Alleles within an individual are distinguished by a number (1 or 2).

Calculating the Distance between Organisms

Let $d(A, B)$ be the distance between individuals A and B and $d(a, b)$ be the distance between alleles a and b. Moreover, let $\min[x; y]$ be the minimum of values of x and y. When evaluating the distance between two diploid individuals at a locus, three situations can be encountered:

(1) Both Individuals Have a Single Allele.—In this situation, the distance between individuals is equal to the allelic distance. If A and B are two individuals that both have 1 allele,

$$d(A, B) = d(a, b)$$

In the hypothetic example, $d(A, B) = 3$.

(2) One Individual Has One Allele and the Other Has Two Alleles.—If A is an individual with one allele (a) and C is an individual with two alleles ($c1, c2$), then

$$d(A, C) = \frac{d(a, c1) + d(a, c2)}{2}$$

If we apply this to the imaginary example, $d(A, C) = (0 + 3)/2 = 1.5$.

(3) Both Individuals Have Two Alleles.—Two individuals, D and E, both have two alleles ($d1, d2$ and $e1, e2$). There are two pairs of allelic distances possible among these individuals: $d(d1, e1)$ and $d(d2, e2)$ or $d(d1, e2)$ and $d(d2, e1)$. The distance between such organisms is the mean of the shortest pair of distances:

$$d(D, E) = \frac{\min[d(d1, e1) + d(d2, e2); d(d1, e2) + d(d2, e1)]}{2}$$

This minimizes the distance between the two pairs of alleles compared. It also compares an allele in one individual with the allele in another individual with which it shares a most recent common ancestor. Taking individuals D and E from the hypothetic example (Fig. 1b), allele $d1$ will be compared with allele $e1$ (that are distant by two mutations) and allele $d2$ with $e2$ (that are identical) because the mean distance for this pair of comparisons, one mutation, is less than the mean distance of three mutations obtained when $d1$ is compared to $e2$ and $d2$ to $e1$. Therefore, in the imaginary example, $d(D, E) = [d(d1, e1) + d(d2, e2)]/2 = [2 + 0]/2 = 1$. This distance is preferable to using the mean of the four different allelic distances, as this latter option can give a non-zero distance for two identical individuals. To illustrate this with the hypothetic example, take individuals C and E that have identical alleles. The sum for each pair of allelic distances is 0 for $[d(c1, e1) + d(c2, e2)]$ and 6 for $[d(c1, e2) + d(c2, e1)]$. Taking the mean of all four comparisons would give a distance of $(0 + 6)/4 = 1.5$ for the distance between C and E, which would not make sense because

they are genetically identical. In contrast, taking the mean of the pair with the shortest allelic distance gives a distance of 0.

Combining Information from Different Genes

The matrix of organisms obtained from one marker can either be used alone or be combined with matrices obtained from other markers. For the present paper, each gene matrix is reweighted so that each gene makes an equal contribution to the combined phylogeny. This is done by dividing each distance by the largest distance of the matrix, for each gene matrix. By attributing the same weight to each gene, every gene is considered to represent an independent estimation of the phylogeny. To fulfill this requirement, there needs to be no recombination within markers. In the presence of recombination, more than one evolutionary history is present in one marker and consequently the weight of the nonrecombining portions of a recombinant gene will be down-weighted. It is therefore recommended to test for recombination before combining different genes.

When combining multiple gene matrices, the final distance between two individuals is the mean of distances between these individuals in the individual matrices. If M and N are two individuals, then the mean distance between them will be:

$$d(M, N) = \frac{1}{n}\sum_{i=1}^{n} d_n(M, N)/d^n_{\max}$$

where n is the number of data sets and $d^n_{\max}$ is the maximum distance in matrix n. Once the final matrix is obtained, we can reconstruct the phylogeny of the organisms with any phylogenetic or network method that uses distances. The program POFAD, written in C++, implements these algorithms and is available at www.irbv.umontreal.ca/pofad.htm.

In our imaginary example, the relationship of individuals was reconstructed from the matrix of organisms (Fig. 1C) using the NeighborNet method (Bryant and Moulton, 2004; Fig. 1D).

MATERIAL AND METHODS

Plant Material

Forty individuals from all seven North American diploid species of *Rosa* sect. *Cinnamomeae* were investigated (Table 1). *Rosa gymnocarpa* Nutt. and *R. pisocarpa* Gray are found exclusively west of the Rocky Mountains; *R. blanda* Ait., *R. foliolosa* Nutt. ex Torr. & A. Gray, *R. nitida* Willd., and *R. palustris* Marsh. occur strictly east of the Rockies, and *R. woodsii* Lindl. can be found on both sides of these mountains. Two diploid species of section *Synstylae* found in North America, *R. setigera* Michx. (native) and *R. multiflora* Thunb. (introduced and now a noxious invasive [Meiners et al., 2001; Hunter and Mattice, 2002]), were included as outgroup taxa. DNA was extracted using the CTAB method of Doyle and Doyle (1987) modified as in Joly (2006).

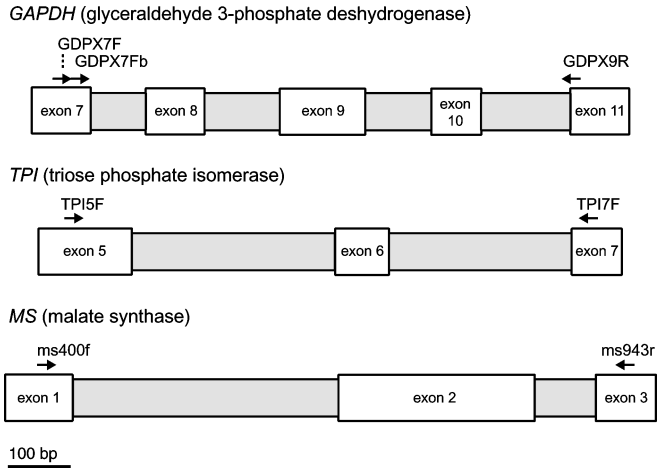


FIGURE 2. Scheme of the loci used in the study in North American *Rosa*. Primers are not to scale and their positions are approximate. Introns are in gray.

Gene Sequencing and Allele Sampling

Three nuclear genes were used in this study: glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*), triose phosphate isomerase (*TPI*), and malate synthase (*MS*). The *GAPDH* sequences are from Joly et al. (2006); GenBank DQ091014–027, 030–035, 038–057, 061–069, 072–086, 172–174). *TPI* was amplified and sequenced using forward primer TPI5F (5′-AAGGTGATCGCCTGTGTTGG-3′) and reverse primer TPI7R (Strand et al., 1997) located in the fifth and seventh exon of the gene, respectively (Fig. 2). The *MS* gene was amplified and sequenced using primers ms400f and ms943r (Lewis and Doyle, 2001); the amplified region covers the first two introns of the gene (Fig. 2). The PCR conditions were as in Joly et al. (2006) except that the annealing temperature was 52°C and 48°C for *TPI* and *MS*, respectively, and that a manual hotstart was used for *TPI* (i.e., the *Taq* was included after the sample reached 95°C). PCR purification and sequencing followed Joly et al. (2006). Allele recovery was achieved using the procedure described in Joly et al. (2006). In short, individuals with no polymorphic peaks in direct sequencing were considered to be homozygous. Alleles of individuals that showed a single polymorphic site were easily extrapolated, but individuals that showed more than one polymorphic site or that had indels among its alleles needed to be cloned. Three to four clones were sequenced per individual to allow the detection of PCR induced mutations and of in vitro recombinants. The cloning procedure is described in Joly et al. (2006).

Analyses

Recombination.—For each gene, recombination was tested using the homoplasy test (Maynard Smith and Smith, 1998), the neighbor similarity score (Jakobsen and Easteal, 1996), the Max chi-squared (χ^2; Maynard Smith, 1992), and the pairwise homoplasy index statistic (Φ; Bruen et al., 2006). These methods were selected because

TABLE 1. Individuals included in this study with their collectors and locality. The number of alleles found for the different genes is indicated and the number of clones sequenced for each species and for each gene is showed in brackets. A dash in brackets indicates that there were two alleles that differed by a single mutation and that cloning was not necessary.

Species	Accession	Collector	Province/ State [a]	Latitude, Longitude	*GAPDH*	*TPI*	*MS*
R. blanda	160	Joly and Starr 409	N.B.	45°57′43.7″N, 67°22′26.1″W	2 [3]	2 [4]	2 [5]
R. blanda	326	Joly and Starr 582	Ont.	42°15′29.7″N, 83°02′58.8″W	2 [3]	2 [4]	2 [5]
R. blanda	365	Joly and Starr 622	Wis.	42°39′07.5″N, 89°43′32.4″W	2 [4]	2 [4]	2 [5]
R. blanda	421	Joly and Starr 678	Minn.	48°06′36.3″N, 96°09′16.0″W	2 [4]	2 [2]	2 [5]
R. blanda	462	Joly and Starr 722	Man.	50°00′59.3″N, 96°55′35.2″W	2 [4]	2 [3]	2 [—]
R. blanda	528	Joly and Starr 788	Ont.	46°28′15.4″N, 80°29′27.2″W	1	2 [—]	2 [—]
R. blanda	567	Joly 921	N.Y.	—	2 [4]	1	2 [—]
R. blanda	652	Joly et al. 993	Que.	48°02′58.8″N, 65°28′43.6″W	2 [4]	1 [3]	2 [6]
R. blanda	1214	Bruneau et al. 1214	Que.	45°31′18″N, 73°50′02″W[b]	1	1	2 [5]
R. blanda	1219	Bruneau et al. 1219	Que.	45°30′18″N, 73°50′02″W[b]	1	2 [—]	2 [5]
R. blanda	1236	Bruneau et al. 1236	Que.	48°21′36″N, 68°45′36″W[b]	2 [4]	2 [4]	2 [1]
R. blanda	98016	Drouin 98-016	Que.	47°26′27″N, 70°30′18″W[b]	2 [—]	1	2 [5]
R. foliolosa	699	Lewis 15846-3	Okla.	34°24′N, 96°00′W	2 [3]	2 [4]	1
R. foliolosa	795	O'Kennon and McLemore 19069A	Tex.	33°24′32.2″N, 97°30′22.0″W	2 [—]	1	2 [5]
R. gymnocarpa	543	Ertter 18001	Idaho	—	1	2 [4]	1
R. gymnocarpa	751	Lewis 15852-1	B.C.	49°02′N, 118°13′W	2 [3]	2 [4]	1
R. gymnocarpa	767	Ertter 18293a	Idaho	—	1	1	2 [4]
R. multiflora	302	Joly and Starr 558	Pa.	42°08′48.4″N, 80°08′00.1″W	2 [4]	2 [3]	2 [5]
R. nitida	570	Meilleur s.n.	Que.	—	2 [4]	2 [2]	1
R. nitida	604	Joly et al. 941	N.B.	45°56′29.2″N, 64°52′07.3″W	2 [3]	2 [4]	2 [—]
R. nitida	675	Brouillet 03-55-1	Nfld.	—	2 [—]	2 [4]	2 [5]
R. nitida	812	Joly 1010-1	Que.	46°22′45.3″N, 75°00′20.6″W	2 [4]	2 [1]	1
R. palustris	168	Joly and Starr 417	N.B.	45°33′43.2″N, 67°25′31.2″W	2 [4]	1	1
R. palustris	304	Joly and Starr 560	Pa.	42°09′32.9″N, 80°07′10.7″W	2 [4]	2 [4]	1
R. palustris	317	Joly and Starr 573	Ont.	42°19′41.0″N, 82°18′49.0″W	2 [4]	2 [4]	1
R. palustris	331	Joly and Starr 587	Mich.	42°19′32.0″N, 84°29′51.2″W	1	2 [3]	1
R. palustris	386	Joly and Starr 644	Wis.	44°01′30.6″N, 89°43′13.1″W	1	1	2 [2]
R. palustris	581	Joly 912	N.Y.	—	1	2 [—]	2 [5]
R. pisocarpa	774	Ertter 18303a	Calif.	—	2 [4]	2 [4]	2 [4]
R. pisocarpa	847	Ertter 18428	Calif.	41°09.2′N, 123°49.2′W	2 [4]	2 [4]	2 [4]
R. setigera	298	Joly and Starr 554	Pa.	42°08′48.4″N, 80°08′00.1″W	1	1	2 [5]
R. woodsii	4	Spellenberg 12555	N.Mex.	—	1	2 [1]	2 [6]
R. woodsii	492	Joly and Starr 752	Sask.	49°12′35.3″N, 101°50′46.1″W	1	2 [—]	2 [4]
R. woodsii	498	Joly and Starr 758	N.Dak.	48°21′09.6″N, 99°47′07.5″W	2 [—]	2 [—]	1
R. woodsii	700	Saarela 266-1	Alta.	—	2 [—]	2 [4]	2 [5]
R. woodsii	733	Dickson 2017	Alta.	—	2 [—]	2 [—]	1
R. woodsii	741	Lewis 15848-1	B.C.	49°45′N, 120°50′W	2 [3]	2 [3]	2 [5]
R. woodsii	800	Joly 1005-1	Colo.	40°12′23.4″N, 104°49′54.0″W	1	2 [3]	1
R. woodsii	807	Joly 1008-1	Colo.	40°38′36.8″N, 104°20′32.0″W	1	2 [—]	2 [—]

[a] Abbreviations follow the nomenclature of Flora of North America (Flora of North America Editorial Committee, 1993).
[b] Approximate coordinates that were not determined by GPS.

they were demonstrated to perform well in datasets of low divergence (Posada and Crandall, 2001; Posada, 2002; Bruen et al., 2006). The homoplasy test was performed without an outgroup using Maynard Smith's program (1998) under conservative ($S_E = 0.6S$) and liberal ($S_E = S$) conditions, where S_E is the effective number of sites and S is the total number of sites in the dataset. The three other methods were implemented in a program written by T. Bruen (2005), testing the significance of the statistics using 1000 permutations. The χ^2 test used a sliding window of size corresponding to the number of polymorphic sites divided by 1.5 and the Φ test used a relative window size (w) of 100.

Phylogenetic Analyses.—For each gene, the gaps were recoded using the simple gap coding method (Simmons and Ochoterena, 2000) implemented in GapCoder (Young and Healy, 2003). Haplotype trees were obtained with PAUP* (ver. 4.10b; Swofford, 2002) by heuristic parsimony analysis with 10 random addition sequence replicates, each retaining a maximum of 1000 trees, TBR branch swapping, and saving all minimal trees during branch swapping.

Two methods were used for obtaining allelic distance matrices from sequences. The first used allelic distances corrected using the appropriate evolutionary model, according to the Akaike information criterion (AIC; Akaike, 1974) calculated in ModelTest (ver. 3.7, Posada and Crandall, 1998) from a neighbor-joinning tree using the matrices without the gaps recoded and treating gaps as missing data. The second used the uncorrected distance of PAUP* to recover allelic distances from the matrices with gaps coded as presence/absence characters.

The matrices of organisms were obtained from POFAD for each gene individually and for the three genes in combination. The phylogeny of organisms was reconstructed

using the NeighborNet algorithm .(Bryant and Moulton, 2004) implemented in SplitsTree (Huson and Bryant, 2006).

RESULTS

Sequences for the genes *TPI* and *MS* were deposited in GenBank (DQ200986 to DQ201120) and matrices used for the analyses are available from TreeBase (study accession number S1444). All gene regions have a greater proportion of intron than exon positions in the aligned matrix, with *TPI* having a greater proportion of intron positions than the other genes for the regions under study (Table 2). Of the three genes, *MS* is the most variable, particularly in the exons where it has a higher number of both synonymous and non-synonymous mutations (Table 2). Indeed, *GAPDH*, *TPI*, and *MS* have 1, 1, and 8 variable amino acid mutations, respectively. All data sets have several indels, which are all located in the intron except one that resulted in the removal of two amino acids in the *MS* gene.

Recombination

Of the four methods used for detecting recombination, only the homoplasy test showed evidence of recombination, returning a positive result for all three datasets (Table 3). This discrepancy between methods could be the consequence of the presence of rate variation among sites in the datasets (see Table 2) because the homoplasy test has been shown to give false evidence of recombination in presence of rate heterogeneity (Posada and Crandall, 2001; Posada, 2002). Therefore, it is more likely that there has been no recombination in the three datasets. Visual inspection of homoplasies on haplotype trees (Templeton et al., 1992) also did not reveal evidence of recombination, further supporting an absence of recombination in each of the three datasets.

Haplotype Trees

Because no recombination was detected in the datasets, it is appropriate to use haplotype trees to represent the genealogy of the haplotypes for each gene. The haplotype trees differ with respect to which taxa form a clade for the different genes (Figs. 3A, 4A, 5A). Haplotypes of *R. gymnocarpa* form a clade with *GAPDH* and *MS*, but not with *TPI*. Haplotypes of *R. pisocarpa* only group together with *GAPDH* and none of the other species have their alleles in a single clade, yet this is sometimes the consequence of one or few incongruent haplotypes. Although haplotypes are more often closer to haplotypes of its species than to those of other species, the overall pattern is a lack of differentiation of species for any single gene. Despite the little information available regarding species relationships, some species are found in different positions in the haplotype trees. For instance, *R. gymnocarpa* is sister to all remaining North American species of sect. *Cinnamomeae* for *GAPDH* but not according to the other genes.

TABLE 3. Recombination inference for the glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*), triose phosphate isomerase (*TPI*), and malate synthase (*MS*) of *Rosa* in North America. Methods used are the homoplasy test (Homo), the neighbor similarity score (NSS), the Max chi-squared (χ^2), and the Phi statistic (Φ) (see text). The probability for the null hypothesis of no recombination is shown for all methods.

Dataset	Mean diversity	S_E (= 0.6S)[a]	P(Homo)[b]	P(NSS)	P(χ^2)	P(Φ)
GAPDH	0.9%	325	0.000	0.637	0.973	0.922
TPI	1.1%	422	0.000	0.205	0.304	0.139
MS	1.2%	478	0.004	0.101	0.486	0.428

[a] The effective number of sites (S_E) is calculated from the total number of sites excluding the 1st and 2nd codon positions (S).
[b] Only the results with the conservative conditions are shown as these are all significant.

Organism Trees

The two ways of recovering allelic distances—the uncorrected distance using gap information and the corrected distance according to the appropriate evolutionary model—gave similar results although including gaps gave a slightly better resolution (data not shown). For this reason, only the results obtained with the uncorrected distance are shown. This choice is further motivated by the presence of several indels in the datasets. Indels are frequent among closely related species or individuals (Britten et al., 2003) and provide phylogenetic information (Kelchner, 2000) that should not be overlooked in phylogenetic studies. Moreover, because of the low divergence among species, it is less important to correct for multiple hits when calculating the distances.

The gene networks of organisms were more often congruent with the taxonomic boundaries than the haplotype trees (Figs. 3B, 4B, 5B). The haplotypes trees for the genes GAPDH, TPI, and MS resolved one, zero, and one species as monophyletic, respectively, whereas the

TABLE 2. Characteristics of the portions of the glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*), triose phosphate isomerase (*TPI*), and malate synthase (*MS*) genes used for inferring the phylogenetic relationships of *Rosa* sect. *Cinnamomeae* in North America. The best model of evolution and the gamma shape (α) is indicated for each gene. The mean pairwise divergence per site is also indicated for synonymous (dS), non-synonymous (dN), and intron (d(intron)) positions.

Dataset	Length	Ratio exon/intron[a]	Variable characters[b]	Informative characters[b]	Indels	Model of evolution[c]	α[c]	dS[b]	dN[b]	d(intron)[b]
GAPDH	739–755	0.75	64	37	10	TrN +Γ	0.222	0.0084	0.0009	0.0120
TPI	806–808	0.26	54	31	7	HKY +Γ	0.263	0.0015	0.0007	0.0110
MS	995–1045	0.74	60	27	17	HKY +Γ	0.536	0.0206	0.0037	0.0144

[a] Calculated from the aligned sequences.
[b] Excluding indels.
[c] Calculated with the outgroup.

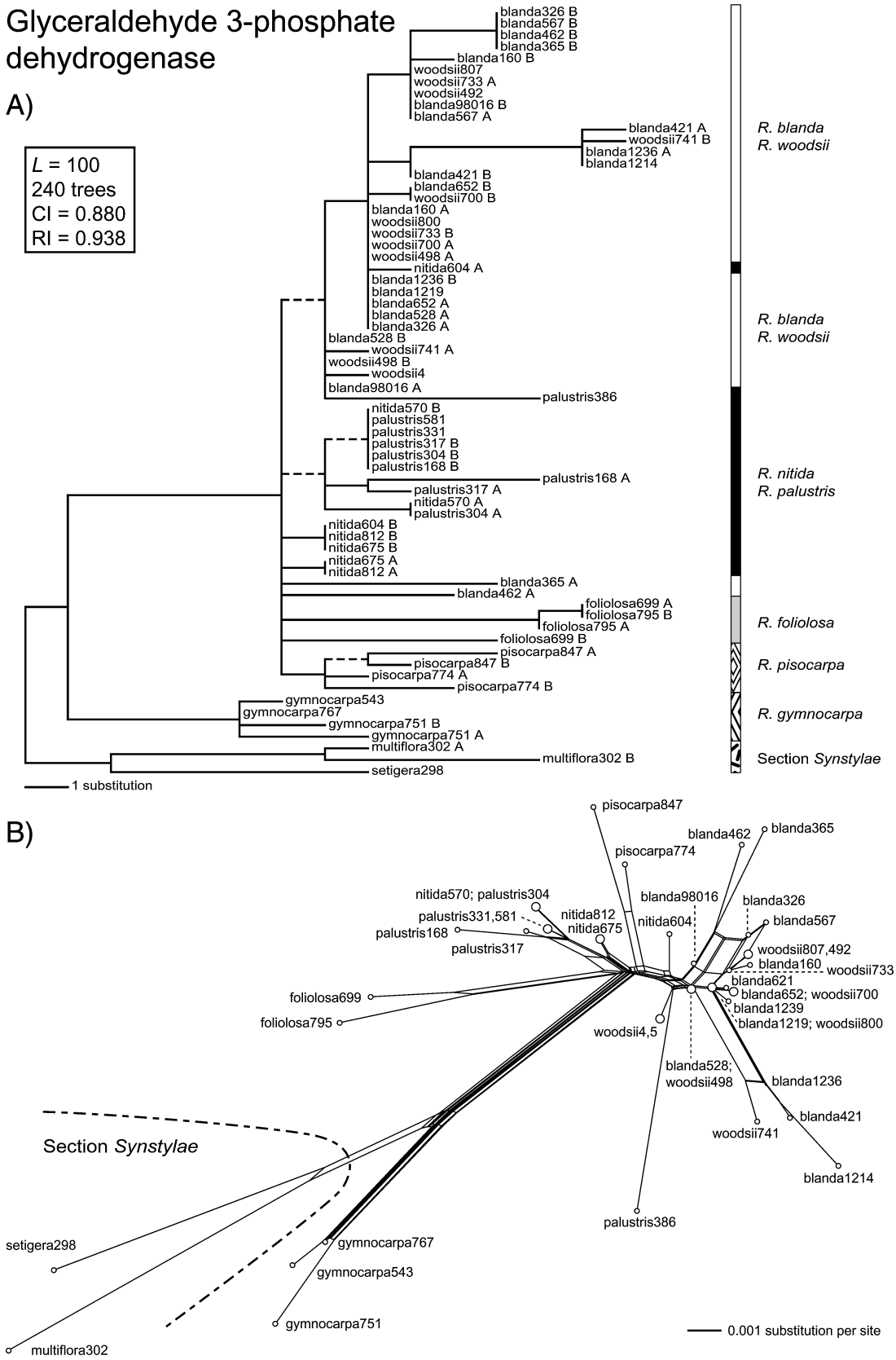


FIGURE 3. Analyses of the glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) dataset. (A) One of the 240 most parsimonious haplotype trees. Dashed lines indicate branches that are not found in the strict consensus tree. (B) Phylogenetic network (NeighborNet) of the organisms.

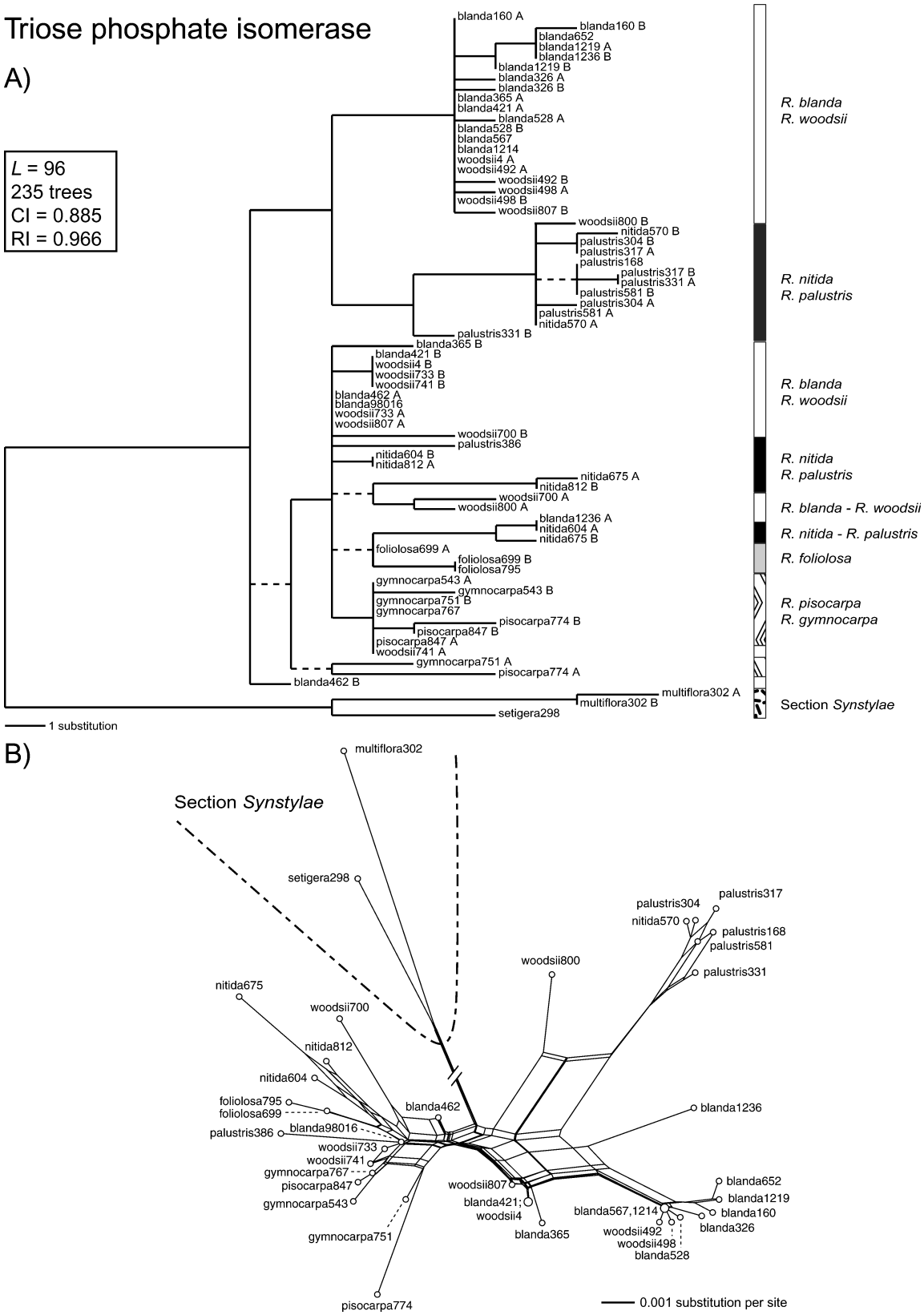


FIGURE 4. Analyses of the triose phosphate isomerase (*TPI*) dataset. (A) One of the 235 most parsimonious haplotype trees. Dashed lines indicate branches that are not found in the strict consensus tree. (B) Phylogenetic network (NeighborNet) of the organisms. The length of the branch connecting the outgroup to the ingroup is of 0.016.

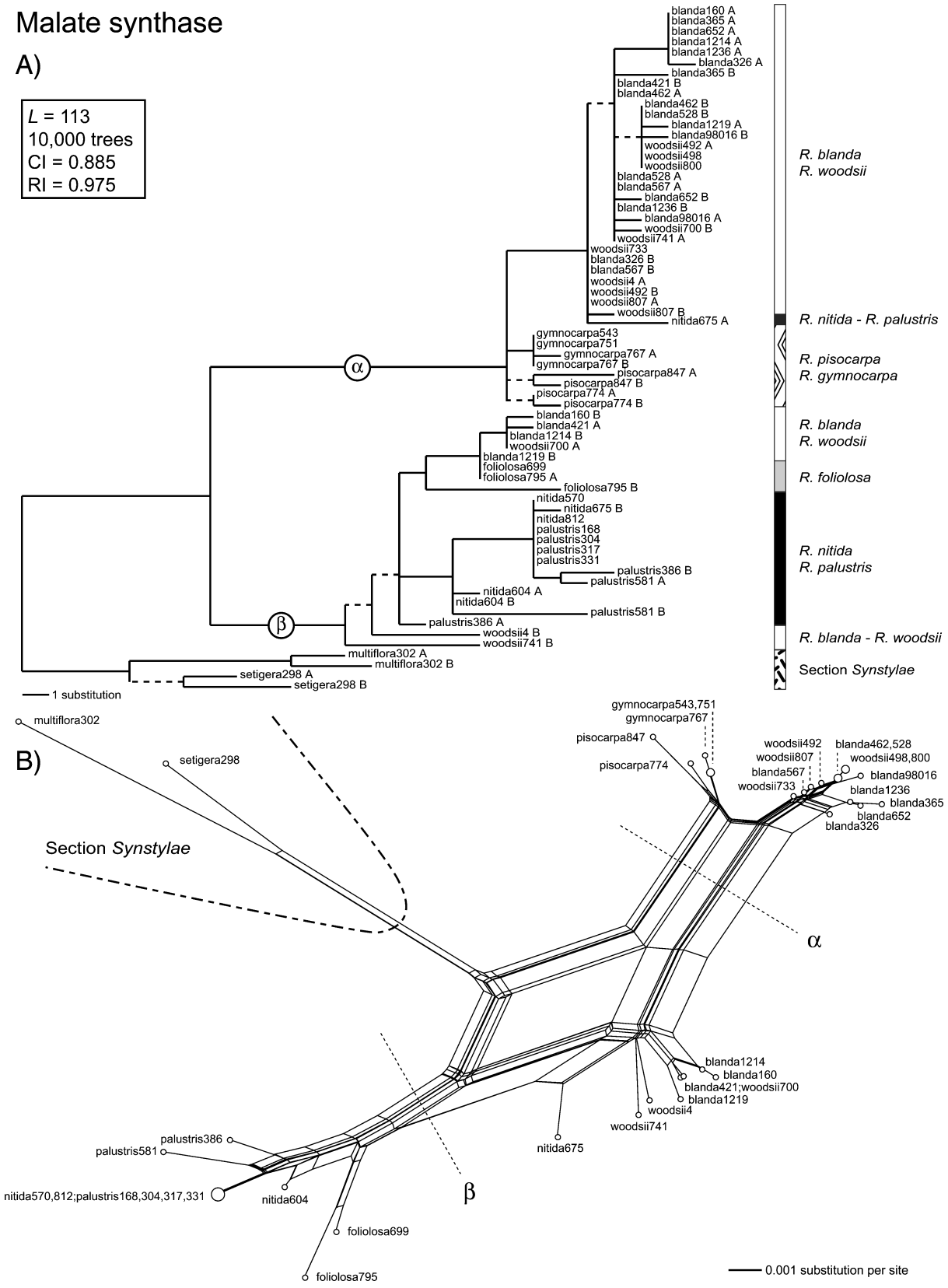


FIGURE 5. Analyses of the malate synthase (*MS*) dataset. (A) One of the 10,000 most parsimonious haplotype trees. Dashed lines indicate branches that are not found in the strict consensus tree. (B) Phylogenetic network (NeighborNet) of the organisms. α and β indicate two genetically distinct groups of alleles (A) or individuals (B) (see text).

network of organisms for the same genes had three, one, and three species resolved by splits. For example, *R. foliolosa* individuals are resolved by a split in all three genes and *R. pisocarpa* individuals group together with *GAPDH* and *MS*. Similarly, individuals of *R. nitida* and *R. palustris* together are resolved by splits with *GAPDH* and *MS*, with few exceptions. Finally, *R. blanda* and *R. woodsii* individuals together are resolved by a split with *GAPDH*, although this group also includes individual *palustris*386.

The networks of organisms appear to appropriately represent intermediate individuals. For example, many individuals (blanda[160, 421, 1214, 1219], woodsii[4, 700, 741], nitida675) have *MS* haplotypes that occur in each of the two major clades on the haplotype tree (α and β; Fig. 5A). Their intermediate status is clearly represented in the network of organisms as these individuals are positioned between the clusters corresponding to the two clades in the haplotype trees (α and β; Fig. 5B). Similar examples are found with the other genes.

The phylogenetic network obtained when the three nuclear genes are combined (Fig. 6) is more resolved and relationships are clearer than when genes are analyzed individually. The network clearly shows that individuals of *R. gymnocarpa* are supported by a split as are individuals of *R. pisocarpa*. However, the relationship of these western species with the eastern ones is not clear. For example, one split suggests that *R. gymnocarpa* is sister to all remaining North American species, whereas another

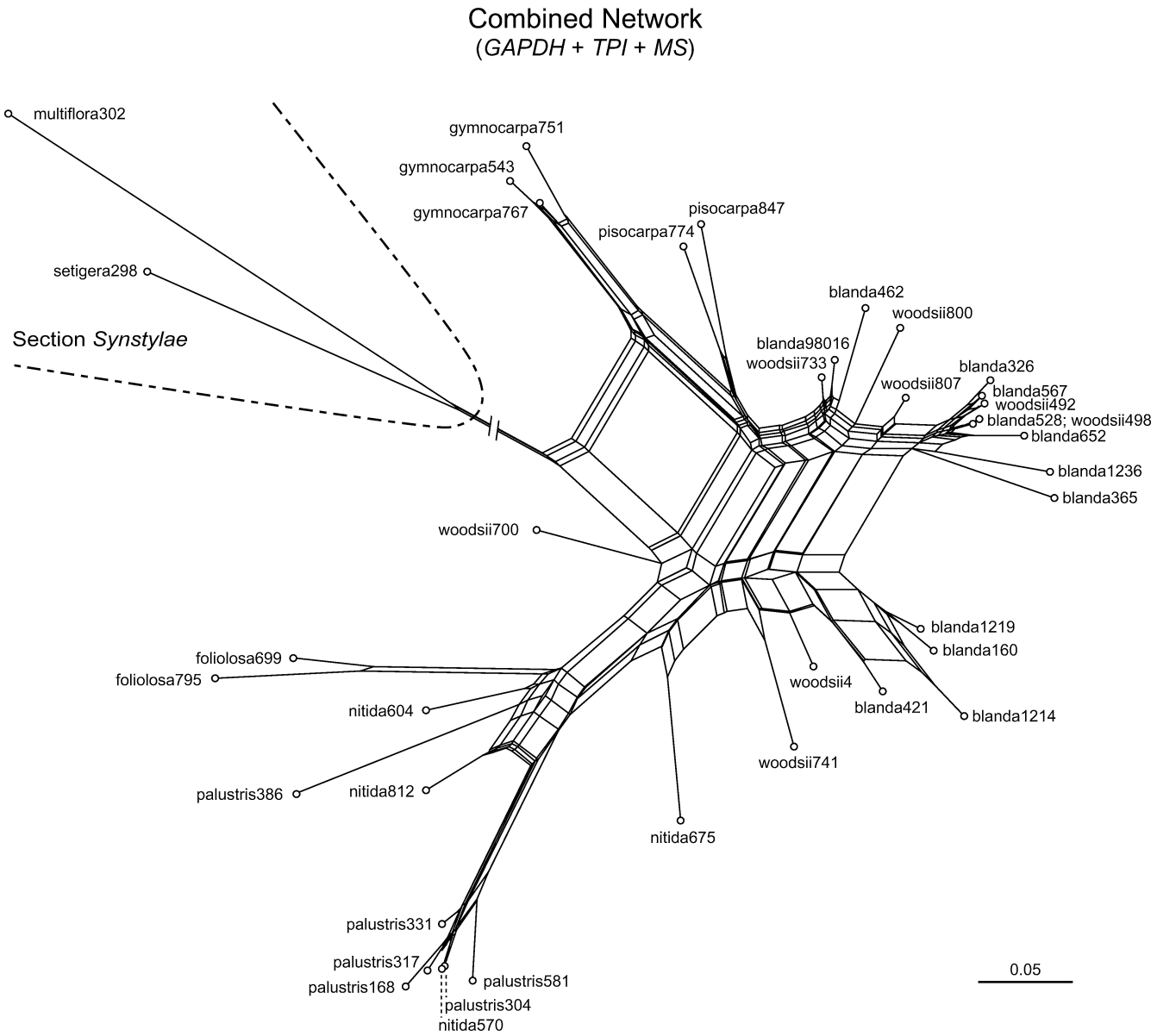


FIGURE 6. Phylogenetic network (NeighborNet) representing the relationships of the organisms obtained from the combined analysis of the glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*), triose phosphate isomerase (*TPI*), and malate synthase (*MS*) loci. The length of the branch connecting the outgroup to the ingroup is of 0.298. The scale only gives a relative indicator of distance because the matrices were standardized.

suggests that it is closer to *R. pisocarpa* and some individuals of *R. blanda* and *R. woodsii*. Neither *R. blanda* nor *R. woodsii* are exclusive in the combined analysis, but these two species together are resolved by a weak split (i.e., there is another contradictory split or bipartition of similar or greater length on the network), which groups all individuals except *woodsii*700. The species *R. nitida*, *R. foliolosa*, and *R. palustris* are resolved as a group on the network, being supported by a weakly contradicted split. Of these three species, *R. foliolosa* individuals are clearly distinct and are strongly resolved by a split. *Rosa nitida* and *R. palustris* are not distinguished from one another but they are grouped together by a weak split on the network (Fig. 5).

DISCUSSION

The POFAD Algorithm

The relationships obtained with the networks of organisms more closely match taxonomic boundaries than those obtained from the haplotype trees. This is probably because the proposed method increases the amount of information included per terminal by incorporating allelic variation for reconstructing the evolutionary history of organisms. For example, if an individual has an allele that is closer to alleles of another species because of deep coalescence, the individual could still group with its species depending on the other allele. This is indeed what happens with *R. foliolosa* that is resolved by a split in all networks of organisms but that is not monophyletic in any of the haplotype trees.

The incorporation of allelic data using the POFAD algorithm also potentially allows the detection and the representation of hybrid individuals if the phylogeny is reconstructed using a reticulate phylogenetic method. For instance, some individuals have malate synthase alleles that fall in two distinct clades in the haplotype tree and these individuals were represented as being intermediate between individuals belonging to these two clades in the network of organisms (see Results and Fig. 5). Using both alleles instead of one for autosomal loci allows the detection of hybrid individuals with a single marker, whereas a minimum of two markers is required when only one allele per individual is sampled. The power of detecting and representing hybrid individuals in phylogenies increases as more genes are investigated (Linder and Rieseberg, 2004), and the increased information contained in allelic variation should similarly improve our ability to reconstruct the evolution of hybrid individuals.

These examples demonstrate the importance of incorporating allelic variation whenever possible in phylogenetic analyses. Using allelic variation effectively doubles the number of lineages sampled. This increases the probability of sampling ancestral lineages within species that provide independent tests of the relationships among species (Rosenberg, 2002). With more ancestral lineages, there is an increased probability of sampling at least one lineage that will have a most recent interspecific coalescent event with its sister species, thereby improving chances of recovering the species phylogeny. This is particularly important for recently diverged species where haplotypes have had less time to coalesce within the species (Rosenberg, 2002).

Combining Multiple Genes.—The greatest interest of the POFAD algorithm certainly is its ability to incorporate allelic variation when reconstructing the phylogenetic history of organisms from multiple datasets. Because any single gene can be incongruent with the species tree, it is important to sample multiple independently evolving markers to be confident in the resulting phylogeny. When analyzing multiple markers, one approach is to combine the datasets first and then to proceed to an analysis of the concatenated dataset (Kluge, 1989; Yang, 1996; Seo et al., 2005). This approach suffers from the fact that alleles are the terminal units of the analysis, henceforth hindering the concatenation of alleles from different loci because they segregate in natural populations. One solution would be to use a consensus sequence of alleles for each individual (see Howarth, 2005), therefore making the organisms the terminals units of the analysis. However, this would result in a loss of information because ambiguities are optimized as to minimize the number of evolutionary changes in phylogenetic analyses. To illustrate this, consider a sequence that differs at a single site between two diploid individuals. Then suppose that an individual is coded as R (A or G) at the site (which means that it has one allele with an A and one with a G) and that the other individual has an A. These two individuals would then be treated as if they were identical even if the first individual has two alleles including one that is different from the alleles of the second individual.

The alternative to the total evidence approach is the "gene as character" approach that consists of combining the trees from each marker analysed independently, either by using consensus tree (e.g., de Queiroz, 1993), reconciled tree (Page and Charleston, 1997; Slowinski et al., 1997), or supertree (e.g., Doyle, 1992; Sanderson et al., 1998; Bininda-Emonds, 2004; Wilkinson et al., 2005) methods. As for the total evidence approach, these methods use haplotypes as terminal units and cannot incorporate allelic variation in phylogenetic analyses of multiple genes, with the exception of reconciled trees. Reconciled trees, however, differ from the POFAD method in that species, rather than individuals, are the terminal units of the analysis. Indeed, one assumption of this method is that gene transmission is strictly vertical among the terminal units of the analysis (Page and Charleston, 1997).

Because of these problems with existing methods, studies that have used allelic variation from multiple markers have either compared the topologies of the different haplotype trees (Hare and Avise, 1998), used allelic consensus sequences for individuals in a concatenated matrix (Howarth, 2005), found concordant signals among gene trees to identify nonrecombining groups of individuals (Koufopanou et al., 1997), or compared the demographic events that were found to have affected each genealogy (Templeton, 2002). The method proposed in this paper gives an alternative to these options by reconstructing a single phylogeny of organisms from multiple datasets that contain allelic information.

Applicability.—The POFAD method should be useful whenever haplotype trees are used, such as at the

intraspecific level or at the species interface among closely related species. At the intraspecific level, it could be useful for phylogeographic studies that wish to draw conclusions from more than one nuclear gene. The use of nuclear genes for phylogeographic studies is becoming frequent (e.g., Olsen and Schaal, 1999; Hare, 2001; Antunes et al., 2002; Joly and Bruneau, 2004) and some studies have already used multiple nuclear gene trees (Hare and Avise, 1998; Templeton, 2002). The proposed method could also be useful for studies at the species interface where it can help delimit species. Because alleles at nuclear loci segregate in natural populations due to sexual reproduction (gene segregation and recombination), relationships within species should be reticulate (tokogenetic), whereas they should be hierarchic (phylogenetic) among species. Tokogenetic relationships result in the sharing of alleles among individuals, which in turn tend to make individuals within species more similar to each other than to individuals of other species. This also implies that there should be no shared phylogenetic patterns among genes within species. In contrast, strong phylogenetic signals shared by a majority of genes should correspond to the speciation event (Koufopanou et al., 1997). These speciation events should therefore result in strong splits in the combined network of organisms if interspecific hybridization does not occur.

Phylogeny of North American Diploid Roses

Little is known of phylogenetic relationships among rose species in North America. Previous studies have provided little information because of the low resolution of molecular markers and poor species sampling (Millan et al., 1996; Matsumoto et al., 1998; Wissemann and Ritz, 2005). In contrast, the three nuclear genes sequenced for several individuals per species in this study allow an assessment of phylogenetic relationships among North American species but also provide information regarding species boundaries.

First of all, the diploid species of *Rosa* in North America appear to be of recent origin according to the low levels of genetic variation found in haplotype trees. Yet, it is also possible that the long generation time, which is typical for shrubs, could accentuate this trend. A rapid radiation is also supported by the lack of monophyly observed for most species. Indeed, recently diverged species are not expected to be reciprocally monophyletic and incomplete lineage sorting is expected to be frequent among such species (Rosenberg, 2002, 2003; Degnan and Salter, 2005). Nevertheless, polyphyletic species could also be the consequence of interspecific gene flow that is indicative of poorly defined species boundaries. Or course, the phenomenon responsible for nonmonophyletic species is likely to be different from one species to the other. But in spite of the low levels of genetic variation and of the absence of monophyly for most species for one or more of the genes studied, the combined analysis of individuals remains informative regarding the phylogenetic relationships of North American species.

Botanists generally have treated the western and eastern North American rose species as distinct entities (Lewis, 1957b; Erlanson MacFarlane, 1966). Yet, the hypothesis that western and eastern species form distinct phylogenetic groups has never been tested. The combined network suggests that a distinction between the west and the east may exist, although it is only supported by a weak split. Relative to the outgroup species of section *Synstylae*, one strong split suggests that *R. gymnocarpa* is sister to all remaining North American species, a signal mostly contributed by the *GAPDH* gene. The alternative solution, which is supported by a split of similar strength contributed mostly by the *MS* gene, groups *R. gymnocarpa* with *R. pisocarpa* and some individuals of *R. blanda* and *R. woodsii*. Congruent with this latter solution, a split on the network supports the monophyly of western species, but this split is rather weak. Because of the incongruence regarding the exact position of the western species among the genes studied, more genes will have to be investigated to determine the exact branching pattern and to confirm the distinction between western and eastern diploid species. Individually, however, both western species *R. gymnocarpa* and *R. pisocarpa* form exclusive groups of individuals, suggesting there is little or no genetic exchange between them. Thus, even if the sampling is limited for these species, the results suggest that these species are distinct.

In the east, the combined network shows that species are divided into two clear groups: one consist of *R. blanda* and *R. woodsii* and the other of *R. foliolosa*, *R. nitida*, and *R. palustris*. In the former group, individuals of *R. blanda* and *R. woodsii* cannot be distinguished from one another. However, both species together form a genetically variable group that is supported by a split in the combined analysis, with the exception of the *woodsii*700 individual. The high genetic diversity observed in this group may be explained in part by the widespread distribution of these species that could reduce the homogenizing effect of gene flow. *Rosa woodsii* ranges from California and British Columbia to the eastern Great Plains, whereas *R. blanda* is distributed from Manitoba and Minnesota in the west to New Brunswick and Maine in the east.

Several clues suggest that the lack of differentiation between *R. blanda* and *R. woodsii* is caused by ongoing gene flow. These species are indeed ecologically (they grow in mesic soils along woods and rivers) and morphologically similar and are difficult to tell apart (Lewis, 1962). Moreover, hybrids between these species have been shown to be highly fertile (Erlanson, 1934; Ratsek et al., 1939), and in the area where the two species overlap, Lewis (1962) described a hybrid zone. Clearly, the species status of these species needs to be reassessed.

The second eastern group revealed by the combined network consists of *R. foliolosa*, *R. nitida* and *R. palustris*. This group is congruent with morphological data because these species share many characters that distinguish them from other North American species. In fact, these species represent all the diploid species that were once placed in sect. *Carolinae* (Crépin, 1889).

Within this group, *R. foliolosa* distinguishes itself from the other species by having its two individuals clearly

resolved as a group on the network. Although only two individuals were investigated for *R. foliolosa*, the network suggests that it is genetically distinct from the other species. *Rosa foliolosa* is also distinct from the other species morphologically, being characterized by narrow leaflets and small pedicels (Lewis, 1957a, 1958). This species is also peculiar for having the smallest geographic distribution of all species of sect. *Cinnamomeae* in North America, as it occurs only in Oklahoma, Texas, and western Arkansas (Lewis, 1958).

Individuals of the last two species, *R. nitida* and *R. palustris*, cannot be distinguished from one another on the network but together are supported as a group, albeit by a weak split. If we consider that *R. foliolosa* individuals are clearly distinct from individuals of these species, then *R. nitida* and *R. palustris* together form a rather cohesive group. A close relationship between these species is not surprising as both have narrow stipules, hypanthium glands, and a preference for bogs and poorly drained soils. In contrast with *R. blanda* and *R. woodsii*, however, *R. nitida* and *R. palustris* are clearly morphologically distinct (Lewis, 1957b, 1957a). This suggests that the lack of genetic distinction between these species is the consequence of a recent origin rather than of poorly defined species boundaries. Although the prevalence of incomplete lineage sorting among species suggests that little time has occurred since the formation of species, the often small populations of these roses and the patchiness of populations over wide geographic areas can also contribute to the retention of ancient polymorphisms. For example, the *palustris*386 individual is from the western extremity of the distribution of *R. palustris*, where few populations are found. This could explain why this individual has retained alleles that are more closely related to *R. blanda* and *R. woodsii* haplotypes for the *GAPDH* and *TPI* genes.

Gene Trees and Species Tree and Individual Sampling within Species

In agreement with most phylogenetic studies investigating multiple markers, incongruence was observed among gene trees obtained from the three loci investigated (Chen and Li, 2001; Cronn and Wendel, 2003; Doyle et al., 2003; Rokas et al., 2003; Jennings and Edwards, 2005). Although some of the incongruence results from the relative position of species among gene phylogenies (i.e., *R. gymnocarpa*), most of the incongruence observed in this study was caused by the lack of monophyly of the species. Such incongruence could be the result of paralogy, incomplete lineage sorting, or hybridization. No signs of gene duplication were noted in this study so paralogy does not seem to be the cause of the lack of species monophyly. Incomplete lineage sorting is more likely to be the cause of incongruence when an incongruent allele is distant from alleles of other species and when their divergence is basal (Holder et al., 2001; Funk and Omland, 2003; Joly et al., 2006). This appears to be case for the allele *palustris*386 that falls in the group of *R. blanda* and *R. woodsii* individuals in the *GAPDH* haplotype tree. In contrast, hybridization should cause an incongruent haplotype to have diverged recently and to be similar to alleles of another species (Holder et al., 2001; Funk and Omland, 2003; Joly et al., 2006). For example, hybridization could explain the position of allele A of nitida604 in the *GAPDH* haplotype tree, which is located in an otherwise exclusively *R. blanda* and *R. woodsii* clade. It is not always obvious how to distinguish the two processes, however, and it may be often impossible to be confident of the process that caused incongruence (Holder et al., 2001; Joly et al., 2006).

Incongruence caused by nonmonophyletic species demonstrates the importance not only of sampling many genes but also of sampling many individuals per species when reconstructing the phylogenetic history of closely related species. Rosenberg (2002) indeed showed that enhanced haplotype sampling increases the probability that the gene tree is topologically concordant with the species tree, in particular for recent radiations as in North American diploid roses. Maddison and Knowles (2006) arrive at the same conclusion in a simulation study demonstrating that given limited resources, it is more advantageous to sample more individuals per species for a single gene than to sequence few individuals for more genes if the species have diverged recently. As discussed above in the context of allelic variation, sampling more individuals increases the probability of sampling ancestral lineages and gives a better chance of accurately reconstructing the phylogenetic history of species, particularly for recently diverged species (Rosenberg, 2002).

Studies that assess the gene tree vs. species tree problem often sample a single individual per species and highlight incompatibilities among the phylogenies obtained from different genes. In these studies, a gene can only be congruent or incongruent with the species tree. Yet, it is probably more frequent that for any particular gene there will be some haplotypes that agree with the species tree and some others that will be incongruent with it. As noted by Rosenberg (2003), without an appropriate sampling of individuals within species, one could conclude that a gene has coalesced within the species when it has not. Such incorrect inferences could result in biased conclusions concerning the evolutionary processes involved in speciation (Funk and Omland, 2003).

CONCLUSION

The algorithm described in this paper allows the incorporation of allelic variation in reconstructing the phylogenetic history of organisms of one or more genes. Allelic variation should provide important additional phylogenetic information when working with closely related species. It also gives the opportunity to reconstruct the phylogenetic history of hybrid individuals even with a single marker when a reticulate phylogenetic method is used. We hope that such a method will stimulate the incorporation of allelic data into phylogenetic analysis as it represents an important amount of information that too often is neglected.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the help of Julian Starr, Walter Lewis, Luc Brouillet, Elisabeth Dickson, Barbara Ertter, Alain Meilleur, Jeff Saarela, and Richard Spellenberg for providing plant material. Authors also thank François-Joseph Lapointe and Bernard Angers for their comments and Trevor Bruen for help and suggestions with recombination analyses. Rod Page, Allan Baker, David Bryant, and an anonymous reviewer gave helpful comments on a previous version of the manuscript. Financial help for this study came from research grants (AB) and fellowships (SJ) from the National Sciences and Engineering Research Council of Canada and from the Fonds québécois de la recherche sur la nature et les technologies.

REFERENCES

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr. 19:716–723.

Antunes, A., A. R. Templeton, R. Guyomard, and P. Alexandrino. 2002. The role of nuclear genes in intraspecific evolutionary inference: Genealogy of the *transferin* gene in the brown trout. Mol. Biol. Evol. 19:1272–1287.

Bininda-Emonds, O. R. 2004. The evolution of supertrees. Trends Ecol. Evol. 19:315-322.

Britten, R. J., L. Rowen, J. Williams, and R. A. Cameron. 2003. Majority of divergence between closely related DNA samples is due to indels. Proc. Nat. Acad. Sci. USA 100:4661–4665.

Bruen, T. C. 2005. PhiPack: PHI test and other tests of recombination. McGill University, Montréal, Québec, Canada. www.mcb.mcgill.ca/~trevor/.

Bruen, T. C., H. Philippe, and D. Bryant. 2006. A simple robust statistical test for detecting the presence of recombination. Genetics 172:2665–2681.

Bryant, D., and V. Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Chen, F. -C., and W. -H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. 68:444–456.

Crépin, F. 1889. Sketch of a new classification of roses. J. R. Hort. Soc. 11:217–228.

Cronn, R., and J. F. Wendel. 2003. Cryptic tryst, genomic mergers, and plant speciation. New Phytologist 161:133–142.

de Queiroz, A. 1993. For consensus (sometimes). Syst. Biol. 42:368–372.

Degnan, J. H., and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

Doyle, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot. 17:144–163.

Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19:11–15.

Doyle, J. J., J. L. Doyle, J. T. Rauscher, and A. H. D. Brown. 2003. Diploid and polyploid reticulate evolution throughout the history of the perennial soybeans (*Glycine* subgenus *Glycine*). New Phytologist 161:121–132.

Erlanson, E. W. 1934. Experimental data for a revision of the North American wild roses. Bot. Gazette 96:197–259.

Erlanson MacFarlane, E. W. 1966. The old problem of species in *Rosa* with special reference to North America. Am. Rose Annu. 51:150–160.

Flora of North America Editorial Committee. 1993. Flora of North America, vol. 1. Oxford University Press, New York.

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. Evol. Syst. 34:397–423.

Hare, M. P. 2001. Prospects for nuclear gene phylogeography. Trends Ecol. Evol. 16:700–706.

Hare, M. P., and J. C. Avise. 1998. Population structure in the american oyster as inferred by nuclear gene genealogies. Mol. Biol. Evol. 15:119–128.

Holder, M. T., J. A. Anderson, and A. K. Holloway. 2001. Difficulties in detecting hybridization. Syst. Biol. 50:978–982.

Howarth, D. G. 2005. Genealogical evidence of homoploid hybrid speciation in an adaptive radiation of *Scaevola* (Goodeniaceae) in the Hawaiian islands. Evolution 59:948–961.

Hunter, J. C., and J. A. Mattice. 2002. The spread of woody exotics into the forests of a northeastern landscape, 1938–1999. J. Torrey Bot. Soc. 129:220–227.

Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Jakobsen, I. B., and S. Easteal. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. CABIOS 12:291–295.

Jennings, W. B., and S. V. Edwards. 2005. Speciational history of Australian grass finches (*Poephilla*) inferred from thirty gene trees. Evolution 59:2033–2047.

Joly, S., and A. Bruneau. 2004. Evolution of triploidy in *Apios americana* (Leguminosae) revealed by the genealogical analysis of the histone H3-D gene. Evolution 58:284–295.

Joly, S., J. R. Starr, W. H. Lewis, and A. Bruneau. 2006. Polyploid and hybrid evolution in roses east of the Rocky Mountains. Am. J. Bot. 93:412–425.

Kelchner, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. Ann Miss. Bot. Garden 87:482–498.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38:7–25.

Koufopanou, V., A. Burt, and J. W. Taylor. 1997. Concordance of gene genealogies reveals reproductive isolation in the pathogenic fungus *Coccidioides immitis*. Proc. Nati. Acad. Sci. USA 94:5478–5482.

Lewis, C. E., and J. J. Doyle. 2001. Phylogenetic utility of the nuclear gene malate synthase in the palm family (Arecaceae). Mol. Phylogenet. Evol. 19:409–420.

Lewis, W. H. 1957a. A monograph of the genus *Rosa* in North America east of the Rocky Mountains. Ph.D. thesis, University of Virginia.

Lewis, W. H. 1957b. Revision of the genus *Rosa* in eastern North America: A review. Am. Rose Annu. 42:116–126.

Lewis, W. H. 1958. A monograph of the genus *Rosa* in North America. II. *R. foliolosa*. Southwestern Naturalist 3:145–153.

Lewis, W. H. 1962. Monograph of the genus *Rosa* in North America. IV. *R.* x *dulcissima*. Brittonia 14:65–71.

Lewis, W. H., and R. E. Basye. 1961. Analysis of nine crosses between diploid *Rosa* species. Proc. Am. Soc. Hort. Sci. 78:573–579.

Linder, C. R., and L. H. Rieseberg. 2004. Reconstructing patterns of reticulate evolution in plants. Am. J. Bot. 91:1700–1708.

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

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Matsumoto, S., M. Kouchi, J. Yabuki, M. Kusunoki, Y. Ueda, and H. Fukui. 1998. Phylogenetic analyses of the genus *Rosa* using the *mat*K sequence: Molecular evidence for the narrow genetic background of modern roses. Sci. Hort. 77:73–82.

Maynard Smith, J. 1992. Analysing the mosaic structure of genes. J. Mol. Evol. 34:126–129.

Maynard Smith, J. 1998. Homoplasy test: Datain and exph programs, version 3, University of Sussex, Brighton, UK. www.lifesci.sussex.ac.uk/home/John_Maynard_Smith/.

Maynard Smith, J., and N. H. Smith. 1998. Detecting recombination from gene trees. Mol. Biol. Evol. 15:590–599.

Meiners, S. J., S. T. A. Pickett, and M. L. Cadenasso. 2001. Effect of plant invasions on the species richness of abandoned agricultural land. Ecography 24:633–644.

Millan, T., F. Osuna, S. Bobos, A. M. Torres, and J. I. Cubero. 1996. Using RAPDs to study phylogenetic relationships in *Rosa*. Theor. Appl. Genet. 92:273–277.

Nichols, R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.

Olsen, K. M., and B. A. Schaal. 1999. Evidence on the origin of cassava: phylogeography of *Manihot esculenta*. Proc. Nat. Acad. Sci. USA 96:5586–5591.

Page, R. D. M., and M. A. Charleston. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. Mol. Phylogenet. Evol. 7:231–240.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
Posada, D. 2002. Evaluating methods for detecting recombination from DNA sequences: Empirical data. Mol. Biol. Evol. 19:708–717.
Posada, D., and K. A. Crandall. 1998. ModelTest: Testing the model of DNA substitution. Bioinformatics 14:817–818.
Posada, D., and K. A. Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Nat. Acad. Sci. USA 98:13757–13762.
Ratsek, J. C., W. S. Flory Jr., and S. H. Yarnell. 1940. Crossing relations of some diploid and polyploid species of roses. Proc. Am. Soc. Hort. Sci. 38:637–654.
Ratsek, J. C., S. H. Yarnell, and W. S. Flory, Jr. 1939. Crossing relations of some diploid species of roses. Proc. Am. Soc. Hort. Sci. 37:983–992.
Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–803.
Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61:225–247.
Rosenberg, N. A. 2003. The shape of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57:1465–1477.
Sanderson, M. J., A. Purvis, and C. Henze. 1998. Phylogenetic supertrees: Assembling the trees of life. Trends Ecol. Evol. 13:105–109.
Schaal, B. A., and K. M. Olsen. 2000. Gene genealogies and population variation in plants. Proc. Nat. Acad. Sci. USA 97:7024–7029.
Seo, T.-K., H. Kishino, and J. L. Thorne. 2005. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. Proc. Nat. Acad. Sci. USA 102:4436–4441.
Simmons, M. P., and H. Ochoterena. 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49:369–381.
Slowinski, J. B., A. Knight, and A. P. Rooney. 1997. Inferring species trees from gene trees: A phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. Mol. Phylogenet. Evol. 8:349–362.
Strand, A. E., J. Leebens-Mack, and B. G. Milligan. 1997. Nuclear DNA-based markers for plant evolutionary biology. Mol. Ecol. 6:113–118.
Swofford, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.
Takahata, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122:957–966.
Templeton, A. R. 2002. Out of Africa again and again. Nature 416:45-51.
Templeton, A. R., K. A. Crandall, and C. F. Sing. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics 132:619–633.
Wilkinson, M., J. A. Cotton, C. Creevey, O. Eulenstein, S. R. Harris, F.-J. Lapointe, C. Levasseur, J. O. Mcinerney, D. Pisani, and J. L. Thorley. 2005. The shape of supertrees to come: Tree shape related properties of fourteen supertree methods. Syst. Biol. 54:419–431.
Wissemann, V., and C. M. Ritz. 2005. The genus *Rosa* (Rosoideae, Rosaceae) revisited: molecular analysis of nrITS-1 and *atp*B-*rbc*L intergenic spacer (IGS) versus conventional taxonomy. Bot. J. Linn. Soc. 147:275–290.
Wu, C. -I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127:429–435.
Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.
Young, N. D., and J. Healy. 2003. GapCoder automates the use of indel characters in phylogenetic analysis. BMC Bioinformatics 4:6.

First submitted 15 September 2005; reviews returned 10 December 2005; final acceptance 6 January 2006
Associate Editor: Allen Baker