Molecular Ecology (2006) **15**, 3505–3513 doi: 10.1111/j.1365-294X.2006.03040.x

FAST TRACK

Blind population genetics survey of tropical rainforest trees

JÉRÔME DUMINIL,* HENRI CARON,* IVAN SCOTTI,† SAINT-OMER CAZAL† and RÉMY J. PETIT*
**INRA, UMR Biodiversity, Genes and Ecosystems, 69 route d'Arcachon, F-33612 Cestas Cedex, France, †INRA, UMR Ecologie des Forêts de Guyane, Campus agronomique BP 709, Avenue de France, 97387 Kourou Cedex, France*

Abstract

Rainforest tree species can be difficult to identify outside of their period of reproduction. Vascular tissues from *Carapa* spp. individuals were collected during a short field trip in French Guiana and analysed in the laboratory with nuclear and chloroplast markers. Using a Bayesian approach, > 90% of the samples could be assigned to one of two distinct clusters corresponding to previously described species, making it possible to estimate the genetic structure of each species and to identify cases of introgression. We argue that this blind procedure represents a first-choice rather than a fallback option whenever related taxa are investigated.

Keywords: Bayesian assignment tests, genetic diversity, geographical structure, South America, species delimitation

Received 19 March 2006; revision received 20 May 2006; accepted 2 June 2006

Introduction

Species are arguably the most fundamental units in ecology and evolution, hence the persistent interest in defining them conceptually and in delimitating them accurately and efficiently (e.g. Mallet 1995; Wiens & Servedio 2000; Hey *et al*. 2003; Sites & Marshall 2004; de Queiroz 2005; and references therein). In practice, species identification typically precedes more detailed investigations. For instance, in population genetic surveys, it comes before molecular typing and population data analysis. Yet, the type of data typically gathered by population geneticists (i.e. multilocus nuclear markers) is particularly well suited to delimitate species (Mallet 1995; Sites & Marshall 2004; Chase *et al*. 2005). Although cross-validation of species identification using different approaches is desirable, it can be costly and may be overkilling in routine studies when clear results are obtained with a single approach. Hence the rather unorthodox suggestion to proceed directly to molecular genetic investigations when studying a group of closely related species by skipping standard morphological identification for most samples.

Such a procedure is likely to be particularly advantageous in specific situations. One of them is the study of the population genetic structure of tropical forest trees. Tropical ecosystems contain over 70% of the ~60 000–80 000 estimated tree species in the world, but are submitted to high rates of deforestation. Exploring the genetic diversity of the rainforest woody flora represents therefore a pressing task. However, tropical forests can be difficult to access and the trees themselves can be difficult to identify because leaf morphology is often insufficient to precisely determine individuals. Hence, collecting well-identified plant material for subsequent molecular analyses in the laboratory presents real barriers for scientists.

Population genetics surveys aim at describing the genetic structure of a group of populations more or less connected by gene flow (i.e. belonging to the same biological species), as a starting point to interpret evolutionary processes. If closely related species coexist, it is of interest to confirm that the morphological divide that has motivated the designation of the different taxa also exists at the molecular level and to ensure that no other reproductive entity has remained unnoticed. Barcoding methods of species identification have attracted much attention recently (e.g. Hebert *et al*. 2004; Will & Rubinoff 2004; Hebert & Barrett 2005) but their performances clearly depend upon initial species delimitation by taxonomists (Meyer & Paulay 2005). Furthermore, they are not well suited to cases where hybridization and introgression are taking place. According to Chase *et al*. (2005), what is needed are 'more sophisticated barcoding tools, which would be multiple, low-copy nuclear markers with sufficient genetic variability and polymerase

Correspondence: Henri Caron, Fax: +33557122881;
E-mail: caron@pierroton.inra.fr

chain reaction (PCR)-reliability; these would permit the detection of hybrids and permit researchers to identify the "genetic gaps" that are useful in assessing species limits'.

Since the existence of low-copy nuclear markers combining broad taxonomic coverage and high level of polymorphism remains elusive, population geneticists interested by species complexes could instead rely on nuclear microsatellites. A major difference with barcoding approaches is that the degree of conservation of the microsatellite primers generally allows only the screening of related (e.g. congeneric) taxa and populations. Fortunately, techniques to develop microsatellites have continuously improved and are now broadly applicable (Zane *et al*. 2002; Selkoe & Toonen 2006). Nuclear microsatellites have a number of advantages: they do not depend on the environment, are highly variable, codominant and generally unlinked, all properties of particular interest for species delimitation. Hence, whenever such markers are available in a group of related species, blind sampling strategies can be implemented, followed by a posteriori delimitation of existing genotypic clusters using assignment methods developed to identify populations (e.g. Pritchard *et al.* 2000; reviewed in Manel *et al*. 2005). Eventually, if clear divides exist, the clusters identified in this first step can be analysed separately. For instance, the geographic distribution, population genetic structure and inbreeding coefficient of each cluster can be estimated, while cases of introgression might be identified using other type of DNA markers such as organelle DNA (Petit & Vendramin 2006). In addition, the inclusion of appropriate controls and/or the taxonomic identification of a subset of the samples can be used to check the correspondence of the genotypic clusters with previously described taxonomic species. This strategy is illustrated here with the example of a South American tropical forest tree genus, *Carapa*.

Carapa (mahogany family, Meliaceae) is a genus comprising medium-sized to large trees up to 30–35 m tall. Several species of *Carapa* have been described from tropical South America and Africa, but the genus is currently under taxonomic revision (David Kenfack, Missouri botanical garden, personal communication). *Carapa procera* D.C. occurs in the Guianas and Central Amazonia and in Western and Central Africa (Pennington *et al*. 1981; Ferraz *et al*. 2003). *Carapa guianensis* Aubl. has not been reported in Africa but is much more widespread in the American continent as it occurs from Central America to South America and has been identified in the Caribbean islands (Ferraz *et al*. 2002). Furthermore, *Carapa nicaraguensis* D.C. has been described from Nicaragua and Costa Rica but is considered as synonymous to *C. guianensis* by Pennington *et al*. (1981). A third species, *Carapa megistocarpa*, was described by Gentry (1988) in Ecuador. The two species present in the Guianas, *Carapa procera* and *C. guianensis*, are monoecious insect-pollinated diploid species characterized by high levels of outcrossing (Hall *et al*. 1994; Doligez & Joly 1997). However, a slightly lower multilocus outcrossing rate has been found in *C. procera* ($t_m \sim 0.83$; Doligez & Joly 1997) than in *C. guianensis* ($t_m \sim 1.00$; Hall *et al*. 1994).

Floral traits are generally used to differentiate these two species. *Carapa guianensis* flowers are described as being sessile or subsessile; they are predominantly tetramerous with eight anthers. In contrast, *C. procera* flowers always display a slender pedicel and are predominantly pentamerous with 10 anthers (Pennington *et al*. 1981). This difference between the flowers of the two species is also reflected in their fruits that have either a tetramerous or pentamerous structure. It has also been reported that the first leaves to emerge following germination are compound in *C. guianensis* and simple in *C. procera* (Ferraz *et al*. 2002). In any case, identification of adult *Carapa* individuals is tedious, since the canopy is hard to access and the required reproductive characters are not present most of the time.

In this study, we 'blindly' sampled individuals of *Carapa* in different populations of French Guiana, i.e. no attempt was made to systematically use morphological traits for identification purposes. Trees were sampled without accessing the canopy, by relying on vascular tissue material taken directly from the trunk, which was then used for DNA isolation (Cavers *et al*. 2006), thereby considerably simplifying the operations (Dick 2001). In addition, for a subset of populations, seeds were collected, germinated and included in the molecular analyses. Bayesian assignment analysis of multilocus nuclear genotypes was used to identify genotypic clusters. Correspondence between marker-based clusters and morphology-based taxa was then tested for the seedling material and for a subset of the adult material. Moreover, herbarium material was genotyped and used as further controls of the correspondence between taxonomic species and genotypic clusters. Finally, parameters of genetic diversity were estimated for each genotypic cluster and the correspondence between nuclear microsatellite genotypes and chloroplast DNA haplotypes was examined, to search for possible cases of introgression.

Materials and methods

Material was sampled from adult trees in 21 *Carapa* spp. populations in French Guiana. We also received additional plant material from Venezuela, Brazil and Costa Rica, resulting in a total of 155 adult trees from 26 populations (at least four individuals per population, except for one population in French Guiana; Table 1). In addition, 57 seeds were gathered from the ground in five of the French Guiana populations (population no. 8, 11, 13, 15 and 16, Table 2) and 19 samples were obtained from the Cayenne herbarium (French Guiana). Total DNA was isolated from

Table 1 Chloroplast DNA composition of the studied populations of *Carapa* (adult individuals) as a function of nuclear microsatellite assignment group

				procera					*guianensis*			*nicarag*	Unassigned individuals				
No.	Population name	Coordinates	Total	H2	H4	H5	H7	H8	H1	H2	H6	H3	H2	H3	H4	H5	H7
1	Bafog	5°29′N, 54°00′W	4	0	0	0	4	0	0	0	0	0	0	0	0	0	0
2	Crique Valentin	5°18′N, 53°12′W	4	0	0	0	4	0	0	0	0	0	0	0	0	0	0
3	Montagne de Fer	5°21′N, 53°32′W	6	0	0	0	6	0	0	0	0	0	0	0	0	0	0
4	Degrad Florian	5°28′N, 53°33′W	3	0	0	0	3	0	0	0	0	0	0	0	0	0	0
5	Counami	5°23′N, 53°12′W	5	0	0	0	5	0	0	0	0	0	0	0	0	0	0
6	Trou Poisson	5°20′N, 53°09′W	5	0	0	0	1	4	0	0	0	0	0	0	0	0	0
7	Saint Elie	5°17′N, 53°05′W	5	0	0	0	5	0	0	0	0	0	0	0	0	0	0
8	Paracou	5°18′N, 52°53′W	6	0	0	0	6	0	0	0	0	0	0	0	0	0	0
9	Petit Saut	5°06′N, 52°58′W	9	0	0	0	9	0	0	0	0	0	0	0	0	0	0
10	Montagne Plomb	5°01′N, 52°56′W	4	0	0	0	4	0	0	0	0	0	0	0	0	0	0
11	Montagne des Singes	5°03′N, 52°42′W	7	0	0	0	6	0	0	0	0	0	0	0	0	0	1
12	Risquetout	4°53′N, 52°33′W	5	0	0	0	5	0	0	0	0	0	0	0	0	0	0
13	Rorota	4°52′N, 52°16′W	6	0	0	0	0	0	4	2	0	0	0	0	0	0	0
14	Fourgassié	4°41′N, 52°19′W	9	0	0	0	0	0	0	9	0	0	0	0	0	0	0
15	Tibourou	4°26′N, 52°19′W	5	0	5	0	0	0	0	0	0	0	0	0	0	0	0
16	Massif de Belizon	4°13′N, 52°25′W	9	0	1	0	0	0	0	3	0	0	4	0	1	0	0
17	Kaw	4°34′N, 52°14′W	9	0	0	0	0	0	0	9	0	0	0	0	0	0	0
18	Approuague	4°16′N, 52°10′W	9	5	0	0	2	0	0	0	0	0	2	0	0	0	0
19	Parc Naturel de Régina	4°07′N, 52°11′W	5	0	0	0	0	0	0	5	0	0	0	0	0	0	0
20	Saut Maripa	3°52′N, 51°52′W	5	0	0	0	0	0	0	5	0	0	0	0	0	0	0
21	Saül	3°38′N, 53°12′W	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
22	Tapajos (Brazil)	2°51′S, 54°57′W	7	0	0	0	0	0	0	0	7	0	0	0	0	0	0
23	Manaus (Brazil)	2°30′S, 60°00′W	4	0	0	2	0	0	0	0	0	0	0	0	0	2	0
24	Upata (Venezuela)	7°50′N, 61°50′W	3	0	0	0	0	0	0	3	0	0	0	0	0	0	0
25	Ladrillera (Costa Rica)	10°27′N, 84°04′W	10	0	0	0	0	0	0	0	0	9	0	1	0	0	0
26	Corinto (Costa Rica)	10°12′N, 83°17′W	10	0	0	0	0	0	0	0	0	10	0	0	0	0	0

Table 2 *Carapa* populations used to examine correspondence between molecular-based assignments and seedlings morphology

			Number of individuals*		
No.	Population name	Number of seedlings	*C. procera*	*C. guianensis*	Unassigned
8	Paracou	8	8s	—	—
11	Montagne des Singes	8	8s	—	—
13	Rorota	22	1s/1c	7s/12c	1c
16	Massif de Belizon	5	—	1s/3c	1c
18	Approuague	14	11s	1s/2c	—

*'s' corresponds to simple leaves and 'c' to compound ones.

vascular tissues sampled underneath the bark (Cavers *et al*. 2006) (adults) or leaves (seedlings and herbarium material) with the QIAGEN DNeasy 96 plant kit.

Seven nuclear microsatellites developed from *Carapa guianensis* (Cg01, Cg11, Cg5, Cg6, Cg7, Cg16, Cg17) were analysed as described in Dayanandan *et al*. (1999) and Vinson *et al*. (2005). PCR products were separated by electrophoresis in denaturating polyacrylamide sequencing gels.

Chloroplast DNA (cpDNA) polymorphisms were revealed by PCR-restriction fragment length polymorphism (RFLP) for the 155 adult individuals using seven different combinations of PCR primer-restriction enzyme: *trn*H-*psb*A/*Taq*I, *trn*S-*trn*G/*Mse*I, *trn*S-*trn*G/*Hae*III, *psb*B-*psb*F/*Mse*I, *pet*Af-*psb*Er/*Mse*I, SR/*Mse*I and CD/*Mse*I. The six consensus primer pairs were selected among those listed by Grivet *et al*. (2001) and Hamilton (1999). PCRs were carried out in a PerkinElmer thermal cycler with annealing temperature and elongation time as described in Grivet *et al*. (2001)

or Hamilton (1999). The digestion step was carried out at 65 °C for 3 h for *Taq*I, and at 37 °C overnight for *Hae*III and *Mse*I in a total volume of 25 μL, containing 2.5 μL of buffer and 2 U of each restriction enzyme. Enzymes and buffers were obtained from Life Technologies, Gibco BRL. PCR-RFLP fragments were separated on 8% acrylamide gel and stained with ethidium bromide.

Deviations from Hardy–Weinberg equilibrium and linkage disequilibrium were checked using GENEPOP 3.4 (Raymond & Rousset 1995). To detect genetically homogenous groups of individuals, Bayesian model-based clustering was conducted on the microsatellite data as implemented with STRUCTURE 2.1 (Pritchard *et al*. 2000). Most of the running parameters were set to default values, as suggested in the user's manual (Pritchard & Wen 2004). Length of burn-in period and number of Markov chain Monte Carlo repetitions were set to 10 000 (Evanno *et al*. 2005). Analyses were performed on the total set of individuals. The number of clusters K was tested in the range from 1 to 26 (i.e. the total number of populations studied), with 15 iterations for each value of K. As highlighted by Evanno *et al*. (2005), $L(K)$, the posterior probability of the data for a given K does not always show a clear mode for the true K. They recommend the use of an ad hoc quantity based on the second order rate of change of the likelihood function with respect to $K(\Delta K)$; this approach was used to determine the number of clusters. We repeated the same analyses using only individuals from French Guiana populations to test the influence of the presence of extra French Guiana populations on assignment probabilities. Assignment probabilities thresholds of either 0.8 or 0.9 were used to build the clusters.

Three different tests were used to check the correspondence of the genotypic clusters with previously described taxonomic species:

1 Fruit morphology was observed in four populations (8, 13, 14 and 15, table 1).
2 In a subset of populations, seeds were collected and germinated and the first leaf morphology was recorded.
3 Samples from the Cayenne Herbarium that had been identified on the basis of floral morphology were genotyped using the same set of microsatellite markers (accessions 469, 539, 1027, 1093, 1160, 1173 and 2733 identified as *C. guianensis* and accessions 2683, 5029, 6209, 7189, 15123 and 21528 identified as *Carapa procera*). We also included a number of samples from the same herbarium putatively identified as *C. guianensis* but on the basis of leaf morphology only (accessions 4426, 4397, 4469, 4473, 4509 and 4511) (www.cayenne.ird.fr/aublet2/aublet2_uk.php3).

For individuals from French Guiana, and separately for each cluster, within-population nuclear gene diversity (H_S) and total gene diversity (H_T) were calculated using GENEPOP 3.4 (Raymond & Rousset 1995). *F*-statistics were estimated using the weighted analysis of variance method of Weir & Cockerham (1984).

CpDNA data were analysed for within-population (H_S) and total (H_T) diversity and for the level of population subdivision (G_{ST}) as well as the corresponding parameters when similarities between haplotypes are taken into account (v_S, v_T and N_{ST}) with the method described by Pons & Petit (1996) using PERMUT (available at www.pierroton.inra.fr/genetics/labo/Software/PermutCpSSR/).

Statistical parsimony was used to reconstruct phylogenetic relationships between haplotypes using TCS version 1.18 (Clement *et al*. 2000). Likelihood-ratio χ^2 tests were performed to test for the association between first leaf morphology of the seedlings and the group of assignment based on nuclear markers of the individuals.

Results

Cluster determination

On average, 6.5 loci were successfully amplified per individual for the fresh samples (212 individuals, corresponding to 155 adult trees and 57 seedlings). Bayesian assignment tests using a variable number of populations (K) indicate that the most likely number of populations is $K = 2$, as suggested by the distribution of ΔK values (results not shown). However, the individuals from two populations of Costa Rica (No. 25 and 26) were assigned to one cluster or to the other, depending on the iteration, while the other genotypes were always assigned to the same cluster (see Table S2, Supplementary material; the mean standard error of the assignment probability over all 15 iterations is only 0.004, excluding populations 25 and 26). Using a restricted but more balanced sample (60 individuals) including all Costa Rican individuals (20) as well as an equal number of individuals attributed to each of the two previously identified clusters, three clusters were identified. All individuals from the two Costa Rican populations were assigned to a first cluster, while the remaining 40 individuals were assigned to two other clusters that correspond to the previously identified ones (minimum assignment probability of 0.93). This supports the existence of a third species in Costa Rica, which might correspond to *Carapa nicaraguensis* (Ferraz *et al*. 2002). These results also illustrate the sensitivity of STRUCTURE to the proportion of each group in the total sample, as pointed out by other authors (e.g. Vähä & Primmer 2006).

Setting aside the individuals from Costa Rica, most of the 192 individuals were clustered in two groups (96% when using a threshold of 0.8 and 93% when using a threshold of 0.9). The difference between the two thresholds is negligible given the sharp transition in assignment probabilities between

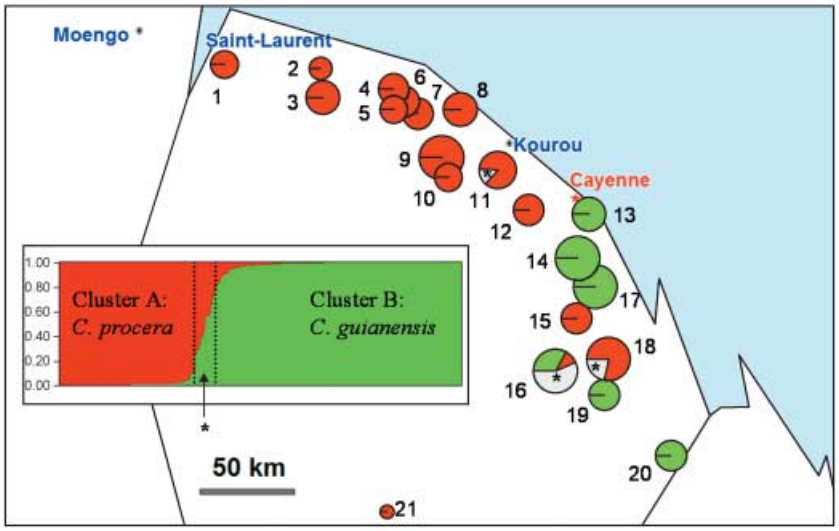

Fig. 1 Distribution of the two groups of *Carapa* sp. in French Guiana. Individuals assigned by STRUCTURE (threshold of 0.8) to *cluster A* are in red, those assigned to *cluster B* are in green, and those that could not be assigned to either cluster are in grey. Circles sizes are proportional to the number of individuals. The diagram (inset figure) represents assignment probabilities for each individual (for the complete sample of 192 adult trees and seedlings). The zone between the broken lines (*) corresponds to individuals that were not assigned to either cluster (0.8 > *probability of assignment* > 0.2).

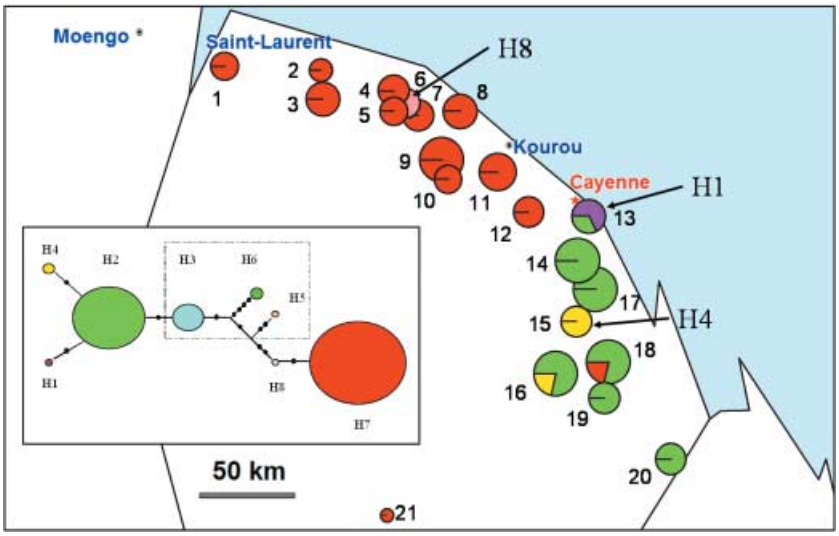

Fig. 2 Distribution of cpDNA haplotypes in the *Carapa* sp. complex in French Guiana. Circle size is proportional to the number of individuals. The haplotype network is shown in the inset; the circles are proportional to the haplotype overall frequency. Mutations detected by PCR-RFLP are represented by blots. The three haplotypes in the dotted rectangle (H3, H5 and H6) were found outside of French Guiana (see Table 1).

the two clusters (see inset in Fig. 2). The 0.8 threshold was selected as a compromise between efficiency and accuracy (Vähä & Primmer 2006).

Cluster A includes 78 adult individuals (red pie charts in Fig. 1), and cluster B, 47 (green pie charts), whereas 10 individuals remain unassigned at the 0.8 threshold (grey pie charts). Only one population includes individuals from both clusters A and B (population 16, Table 1 and Fig. 1). The 10 unassigned adult individuals belong to five different populations. However, five of them are found in a single population (16), the only one that has a mixed composition. This is a first indication that some genetic exchanges between the two clusters have taken place.

A short field survey in French Guiana after the flowering period in 2006 allowed us to examine fruit morphology in four of the previously investigated populations. In populations 13 and 14, fruits picked up on the ground present a tetrameric structure considered typical of *Carapa guianensis*, whereas in populations 8 and 15 they have a pentameric structure as in *Carapa procera*. All individuals from populations 13 and 14 had been assigned to cluster B, and all those from populations 8 and 15 to cluster A; this is a first suggestion that cluster A corresponds to *C. procera* and cluster B to *C. guianensis*.

In addition, in the five populations where seeds had been sampled and germinated, seedlings with either simple or compound leaves were observed and genotyped. Although not absolute, the association between seedling leaf morphology and molecular data is very strong ($\chi^2 = 27.3$, $P < 0.001$), with a predominance of seedlings with simple leaves assigned to cluster A and of seedlings with compound leaves assigned to cluster B, a further confirmation that cluster A might correspond to *C. procera* and cluster B to *C. guianensis* (Ferraz *et al*. 2002; Table 2).

Table 3 Genetic diversity of French Guiana *Carapa* populations

	Number of populations*	Nuclear data				Chloroplast data†					
		H_S	H_T	F_{ST}	F_{IS}	H_S	H_T	G_{ST}	v_S	v_T	N_{ST}
Carapa spp.	20	0.415	0.700	0.412	0.122	0.095 (0.044)	0.606 (0.083)	0.842 (0.063)	0.054 (0.036)	0.777 (0.083)	0.931‡ (0.045)
Carapa procera	14	0.363	0.514	0.288	0.194	0.063 (0.042)	0.336 (0.146)	0.814 (0.085)	0.063 (0.042)	0.360 (0.158)	0.826 (0.089)
Carapa guianensis	6	0.533	0.621	0.193	0.030	—	—	—	—	—	—

*with ≥ 3 individuals.
†'—' not computed. Standard errors in parentheses.
‡N_{ST} higher than G_{ST} at $P < 0.02$.

We also tested the correspondence between cluster assignation and species identification in the case of 19 previously identified herbarium samples. Multilocus genotypes (three to seven loci) were obtained for 12 samples out of 19 (mean of 5.2 loci amplified per individual among these 12 individuals). The remaining seven samples failed to amplify at all loci. Five individuals among the 12 (2683, 5029, 5737, 15123, 794) had been identified as *C. procera*. All were assigned to cluster A. Similarly, four samples (accessions 1027, 539, 1160, 469) had been identified as *C. guianensis*. All were assigned to cluster B. On the contrary, three samples identified as '*C. guianensis?*' on the basis of leaf morphology only (4469, 4473, 4511) were assigned to cluster A. Overall, these results confirm that cluster A corresponds to *C. procera* and cluster B to *C. guianensis* but indicate that leaf morphological traits alone do not allow safe species identification.

Species distribution

The individuals assigned to cluster A were mainly from western French Guiana and the Manaus population (Fig. 1). This broadly corresponds with the distribution of *C. procera* published by Pennington *et al.* (1981) in his monograph of the Meliaceae. However, some populations made of trees assigned to cluster A were also found in the eastern part of French Guiana (populations 15, 16 and 18, Table 1 and Fig. 1), although only one population included individuals from both clusters (population 16; Fig. 1). Cluster B comprises individuals from the eastern part of French Guiana as well as from the Brazilian Tapajos population and from the population from Venezuela. This is in agreement with the description of the geographic distribution of *C. guianensis* given by Pennington *et al.* (1981).

Population analysis of nuclear microsatellite variation

When measured over the whole *Carapa* complex, genetic differentiation among populations was high ($F_{ST} = 0.410$). This value contrasts with the results obtained for each cluster taken separately ($F_{ST} = 0.196$ for the *C. guianensis* cluster; $F_{ST} = 0.280$ for the *C. procera* cluster), a further indication that the two clusters correspond to distinct biological entities. Interestingly, these fairly high values of intraspecific differentiation, at least for trees (Hamrick & Godt 1989), did not compromise the assignment procedure, probably because the differentiation between the two clusters is quite strong ($F_{ST} = 0.381$). Hence, only few loci are needed to achieve good assignment accuracy (Vähä & Primmer 2006). This high value of interspecific differentiation also justifies our use of Evanno's ΔK method of inference of the number of clusters, as this method appears to perform better when differentiation among populations is strong (Waples & Gaggiotti 2006). The *C. procera* cluster presents a higher heterozygote deficit ($F_{IS} = 0.182$) than the *C. guianensis* cluster ($F_{IS} = 0.021$; Table 3). As mentioned earlier, a lower outcrossing rate had been measured in *C. procera* than in *C. guianensis* (Hall *et al.* 1994; Doligez & Joly 1997; see also Dayanandan *et al.* 1999), so this result also fits well with the expectations.

Analysis of chloroplast DNA variation

We detected 12 polymorphic restriction fragments (length variants) that combined into eight chloroplast haplotypes (H1 to H8, see Table S1, Supplementary material, and Fig. 2). Two were particularly abundant and had distinct geographic distributions, one being restricted to western French Guiana (H7) and the other (H2) to the eastern part (Table 1). The haplotypes from Costa Rica and Brazil (H3, H5, H6) were more closely related to H2 (Fig. 2). For the *Carapa* complex, the coefficient of cpDNA differentiation measured over the 20 French Guiana populations was very high ($G_{ST} = 0.842$), especially when haplotype similarities were taken into account ($N_{ST} = 0.917$; Table 3). The difference between the two parameters is significant ($P < 0.02$), pointing to the existence of a phylogeographic structure, as is often the case when genetic structure is well marked (Petit *et al.* 2005). The corresponding estimates for *C. procera* are provided in Table 3, whereas no estimate of genetic differentiation

could be obtained for *C. guianensis* due to its low cpDNA genetic diversity. In a previous study on cpDNA variation in *C. guianensis*, a strong genetic structure had been detected (G_{ST} = 0.96, Cloutier *et al*. 2005).

In French Guiana, five different haplotypes were associated with *C. procera* and only two with *C. guianensis* (Table 1). One haplotype (H2) is shared by the two species. Most *C. procera* individuals from the western part of French Guiana have either haplotype H2 or haplotype H4, which is closely related to H2 (Fig. 2). As H2 is the most common haplotype in *C. guianensis*, this suggests that pollen from *C. procera* might have displaced the nuclear genome of local populations of *C. guianensis* by hybridization and recurrent backcrosses, resulting in *C. procera* individuals with a cpDNA haplotype characteristic of *C. guianensis* (see Belahbib *et al*. 2001; Petit *et al*. 2004 for other examples in trees).

Discussion

The sampling of well-identified material for population genetics surveys can be a difficult task. The simple procedure used here proved very efficient; no identification was attempted in the field, which considerably facilitated sampling. Subsequently, most individuals were assigned to a single genotypic cluster, which was then tentatively attributed to previously described taxa using several independent lines of evidence. First, the morphology of a subset of individuals (adults and seedlings) was broadly consistent with their separation in two genotypic clusters, although, not surprisingly, one morphological character (fruit type) was more concordant than the other (leaf morphology of seedlings). Second, herbarium material identified by taxonomists was genotyped and assigned to the previously identified clusters. Again, complete correspondence was observed, except in the cases where the herbarium material had been identified on the basis of adult leaf material only. Third, the geographic distribution of the genotypic clusters largely fits with the geographic range of the species described previously by botanists. While these different lines of evidence provide a strong indication that the taxonomic species and genotypic clusters correspond well, we cannot rule out the possibility that more thorough surveys will detect individuals for which the assignment to genotypic clusters contradicts taxonomic determinations (based on reproductive material). This would not be surprising given the existence of introgression between the two genetic entities (inferred from the genetic composition of population 16 at nuclear microsatellites and from the comparison between nuclear and organelle variation).

In this study, and contrary to most previous surveys of genetic variation, the assignment of individuals to species was not based on morphological characters considered a priori to characterize each species but on the use of multiple highly variable codominant nuclear markers coupled with quantitative (i.e. Bayesian) assignment methods. Morphological data are needed only to establish the correspondence with previously described taxa on a subset of individuals. Such approaches have already been evaluated for the purpose of identifying parental populations and hybrids (Vähä & Primmer 2006 and references therein) but rarely so in the context of identifying species sampled during population genetic surveys (for an exception involving a pair of cryptic fish species investigated with allozymes see Hawkins *et al*. 2005).

While the 'myth of the molecule' has been sharply criticized by taxonomists (e.g. Will & Rubinoff 2004), we insist that the 'blind' approach used here to circumscribe clusters that closely correspond to taxonomic species is not based on a classical barcoding approach nor on phylogeographic analyses of a few gene trees (Templeton 2001) but on population genetic principles based on multilocus data. As stressed by Hey *et al*. (2003) and de Queiroz (2005), taxonomic species should merely be considered as hypotheses of evolutionary entities to be tested with more quantitative approaches. Recently, morphological species were shown to correspond fairly well to reproductively independent lineages, not only in animals but also in plants (Rieseberg *et al*. 2006). The correspondence between taxonomic species and phenotypic clusters was also investigated, using a meta-analysis of published data. However, the correspondence of taxonomic species and/or phenotypic clusters with genotypic clusters (as defined here, using polymorphic molecular markers) remains to be investigated for most taxa. Whenever the correspondence appears satisfactory, as in our example, a blind procedure such as the one exemplified here can be readily advocated. However, even if the correspondence among approaches is poor or cannot be easily assessed, we believe that this blind procedure remains a first-choice option. Our rationale is that the Bayesian assignment method aims at minimizing Hardy–Weinberg and linkage disequilibria within each of the genotypic clusters. This strikes us as appropriate (and quantitative) criteria to circumscribe 'biological species' as originally envisaged by Dobzhansky (1937). Mallet (1995) had explicitly proposed such an approach to delimitate species: 'we see two species rather than one if there are two identifiable genotypic clusters. These clusters are recognized by a deficit of intermediates, both at single loci (heterozygote deficits) and at multiple loci …'. He further pointed out that this definition applies best to populations in contact. Delimitation of species according to the genetic principles outlined by Mallet more than 10 years ago is now becoming easier, thanks to the development of new powerful statistical methods and the increased availability of highly polymorphic nuclear markers.

In conclusion, it appears that traditional population genetic surveys should facilitate the exploration of

biodiversity below but also above the species level in tropical forests and elsewhere. In the case of *Carapa* sp. in French Guiana, it proved possible to assign most samples to genotypic clusters corresponding to previously described species despite indications for hybridization and introgression. However, more studies are needed to investigate the feasibility of this approach in more situations, including in complex cases involving polyploidy, asexual reproduction, high selfing rates, larger numbers of congeneric species, low interspecific differentiation or complete allopatry.

Acknowledgements

We are grateful to Arndt Hampe and three anonymous referees for their critical comments on a previous version of the manuscript. We would like to thank Carlos Navarro, Elio Sanoja, Patrick Heuret, Stephen Cavers, Eric Bandou and Mimi Bertocchi for their valuable help in the sampling of populations. We also thank Cécile Aupic (herbarium of the Muséum National d'Histoire Naturelle in Paris) and Jean Jacques de Granville (herbarium of the IRD in Cayenne) for the access to herbarium collections, and Pierre-Michel Forget, Cyril Dutech, David Kenfack, Meriem Fournier, Christopher Baraloto, Lilian Blanc, Jean-Pierre Pascal, Laurent Maggia, Vincent Freycon and Pascal Petronelli for their help and for discussions. The study was supported by a grant from the EC, contract No. 003708 (project Seedsource) to HC and RJP. Additional funding was provided by a Bourse Dufrenoy from the Académie d'Agriculture de France to JD.

Supplementary material

The supplementary material is available from http://www.blackwellpublishing.com/products/journals/suppmat/MEC/MEC3040/MEC3040sm.htm

Table S1 Electrophoretic profiles and frequency of the eight haplotypes of *Carapa* spp. identified by PCR-RFLP

Table S2 Mean assignment probabilities across iterations as a function of the standard error of the probabilities. The dotted ellipse corresponds to the individuals from the Costa Rican populations, that are either associated to one cluster or to the other depending on the iteration (high standard error over the iterations). Other individuals do not present this pattern.

References

Belahbib N, Pemonge MH, Ouassou A *et al.* (2001) Frequent cytoplasmic exchanges between oak species that are not closely related: *Quercus suber* and *Q. ilex* in *Morocco*. *Molecular Ecology*, **10**, 2003–2012.

Cavers S, Bandou E, Caron H *et al.* (2006) A simple and effective technique for sampling tissue for DNA analysis of trees: collection and preservation of trunk cambium, an alternative to canopy leaves. *Silvae Genetica*, **54**, 265–269.

Chase MW, Salamin N, Wilkinson M *et al.* (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1889–1895.

Clement MD, Posada MD, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, **9**, 1657–1660.

Cloutier D, Povoa JSR, Procopio LC *et al.* (2005) Chloroplast DNA variation of *Carapa guianensis* in the Amazon basin. *Silvae Genetica*, **54**, 270–274.

Dayanandan S, Dole J, Bawa K, Kesseli R (1999) Population structure delineated with microsatellite markers in fragmented populations of a tropical tree, *Carapa guianensis* (Meliaceae). *Molecular Ecology*, **8**, 1585–1592.

Dick CW (2001) Genetic rescue of remnant tropical trees by an alien pollinator. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **268**, 2391–2396.

Dobzhansky T (1937) *Genetics and the Origin of Species*. Columbia University Press, New York.

Doligez A, Joly HI (1997) Mating system of *Carapa procera* (Meliaceae) in the French Guiana tropical forest. *American Journal of Botany*, **84**, 461–470.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Ferraz IDK, Camargo JLC, Sampaio PTB (2002) Sementes e plântulas de andiroba (*Carapa guianensis* Aubl. e *Carapa procera* D.C.): aspectos botânicos, ecologicos e tecnologicos. *Acta Amazonica*, **32**, 647–661.

Ferraz IDK, Camargo JLC, Sampaio PTB (2003) Andiroba *Carapa guianensis* Aubl. *Carapa procera* D.C. Meliaceae. *Manual de Sementes da Amazônia*, **1**, 1–6.

Gentry AH (1988) New species and a new combination for plants from trans-Andean South America. *Annals of the Missouri Botanical Garden*, **75**, 1429–1439.

Grivet D, Heinze B, Vendramin GG, Petit RJ (2001) Genome walking with consensus primers: application to the large single copy region of chloroplast DNA. *Molecular Ecology Notes*, **1**, 345–349.

Hall P, Orrell LC, Bawa KS (1994) Genetic diversity and mating system in a tropical tree, *Carapa guianensis* (Meliaceae). *American Journal of Botany*, **81**, 1104–1111.

Hamilton MB (1999) Four primer pairs for the amplification of chloroplast intergenic regions with intraspecific variation. *Molecular Ecology*, **8**, 521–523.

Hamrick JL, Godt MJW (1989) Allozyme diversity in plant species. In: *Plant Population Genetics, Breeding and Genetic Resources* (eds Brown AHD, Clegg MT, Kahler AL, Weir BS), pp. 43–63. Sinauer, Sunderland, Massachusetts.

Hawkins SL, Heifetz J, Kondzela CM *et al.* (2005) Genetic variation of rougheye rockfish (*Sebastes aleutianus*) and shortraker rockfish (*S. borealis*) inferred from allozymes. *Fishery Bulletin*, **103**, 524–535.

Hebert PDN, Barrett RDH (2005) Reply to the comment by L. Prendini on 'Identifying spiders through DNA barcodes'. *Canadian Journal of Zoology-Revue Canadienne de Zoologie*, **83**, 505–506.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, 1657–1663.

Hey J, Waples RS, Arnold ML, Butlin RK, Harrison RG (2003) Understanding and confronting species uncertainty in biology and conservation. *Trends in Ecology & Evolution*, **18**, 597–603.

Mallet J (1995) A species definition for the modern synthesis. *Trends in Ecology & Evolution*, **10**, 294–299.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, **20**, 136–142.

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, **3**, 1–10.

Pennington TD, Styles BT, Taylor DAH (1981) Meliaceae. *Flora Neotropica*, **28**, 406–419. Hafner, New York.

Petit RJ, Vendramin GG (2006) Plant phylogeography based on organelle genes: an introduction. In: *Phylogeography of Southern European Refugia* (eds Weiss S, Ferrand N). Springer, in press.

Petit RJ, Bodénès C, Ducousso A, Roussel G, Kremer A (2004) Hybridization as a mechanism of invasion in oaks. *New Phytologist*, **161**, 151–164.

Petit RJ, Duminil J, Fineschi S *et al.* (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology*, **14**, 689–711.

Pons O, Petit RJ (1996) Measuring and testing genetic differentiation with ordered versus unordered alleles. *Genetics*, **144**, 1237–1245.

Pritchard JK, Wen W (2004) Documentation for STRUCTURE software (version 2). Available from http://pritch.bsd.uchicago.edu/structure.html.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

de Queiroz K (2005) Different species problems and their resolutions. *BioEssays*, **27**, 1263–1269.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Rieseberg LH, Wood TE, Baack EJ (2006) The nature of plant species. *Nature*, **440**, 524–527.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.

Sites JW, Marshall JC (2004) Operational criteria for delimiting species. *Annual Review of Ecology, Evolution and Systematics*, **35**, 199–227.

Templeton AR (2001) Using phylogeographic analyses of gene trees to test species status and processes. *Molecular Ecology*, **10**, 779–791.

Vähä JP, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, **15**, 63–72.

Vinson CC, Azevedo VCR, Sampaio I, Ciampi AY (2005) Development of microsatellite markers for *Carapa guianensis* (Aublet), a tree species from the Amazon forest. *Molecular Ecology Notes*, **5**, 33–34.

Waples RS, Gaggiotti OE (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wiens JJ, Servedio MR (2000) Species delimitation in systematics: inferring diagnostic differences between species. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **267**, 631–636.

Will KW, Rubinoff D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*, **20**, 47–55.

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology*, **11**, 1–16.

Jerôme Duminil is working on comparative studies of genetic diversity in plants. He is particularly interested in ecological, life historical and chorological features of species that can explain the distribution of genetic diversity. Henri Caron is studying genetic variation and genetic processes in tropical tree species. Saintomer Cazal's main focus is in tropical botany and forest tree genetics. Ivan Scotti's main research interests are in ecological and evolutionary genetics of forest trees. Remy Petit is a population geneticist with broad interest in evolution, phylogeography and mating system of trees.
