

DNA BARCODING

Testing candidate plant barcode regions in the Myristicaceae

S. G. NEWMAS^T,* A. J. FAZEKAS,* R. A. D. STEEVES* and J. JANOVEC†

*Floristic Diversity Research Group, Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada. N1G 2W1, †Botanical Research Institute of Texas (BRIT), 509 Pecan Street, Suite 101, Fort Worth, Texas 76102–4060, USA

Abstract

The concept and practice of DNA barcoding have been designed as a system to facilitate species identification and recognition. The primary challenge for barcoding plants has been to identify a suitable region on which to focus the effort. The slow relative nucleotide substitution rates of plant mitochondria and the technical issues with the use of nuclear regions have focused attention on several proposed regions in the plastid genome. One of the challenges for barcoding is to discriminate closely related or recently evolved species. The Myristicaceae, or nutmeg family, is an older group within the angiosperms that contains some recently evolved species providing a challenging test for barcoding plants. The goal of this study is to determine the relative utility of six coding (Universal Plastid Amplicon – UPA, *rpoB*, *rpoC1*, *accD*, *rbcL*, *matK*) and one noncoding (*trnH-psbA*) chloroplast loci for barcoding in the genus *Compsonuera* using both single region and multiregion approaches. Five of the regions we tested were predominantly invariant across species (UPA, *rpoB*, *rpoC1*, *accD*, *rbcL*). Two of the regions (*matK* and *trnH-psbA*) had significant variation and show promise for barcoding in nutmegs. We demonstrate that a two-gene approach utilizing a moderately variable region (*matK*) and a more variable region (*trnH-psbA*) provides resolution among all the *Compsonuera* species we sampled including the recently evolved *C. sprucei* and *C. mexicana*. Our classification analyses based on non-metric multidimensional scaling ordination, suggest that the use of two regions results in a decreased range of intraspecific variation relative to the distribution of interspecific divergence with 95% of the samples correctly identified in a sequence identification analysis.

Keywords: barcoding, biodiversity, *Compsonuera*, nutmegs, species discovery, taxonomy, tropical botany

Received 15 May 2007; revision accepted 29 August 2007

DNA barcoding is a method of species identification and recognition using DNA sequence data (Hebert *et al.* 2003, 2004a, b). Based on the mitochondrial gene, cytochrome *c* oxidase subunit 1 (COI or *cox1*), DNA barcoding is proceeding apace in several animal groups, which can be reviewed online via the Canadian Barcode of Life (<http://www.bolnet.ca>) and the Consortium for the Barcode of Life (CBOL) (<http://www.barcoding.si.edu>). In contrast to this rapid progress, efforts to enable such a system for plants have been slower for a number of reasons that have been previously discussed in detail (Chase *et al.* 2005; Kress *et al.* 2005; Cowan *et al.* 2006; Newmaster *et al.* 2006). The primary

challenge has been to identify a suitable DNA region (sequence) or regions, on which to focus the effort.

A suitable region or combination of regions must be able to encompass the natural variation found within and between species circumscriptions in plants. The slow relative substitution rates of plant mitochondrial DNA and the technical issues (e.g., Cowan *et al.* 2006) with the use of nuclear regions have focused attention on the plastid genome as the current best working option. A number of candidate regions have been suggested. Kress *et al.* (2005) proposed the *trnH-psbA* plastid spacer and suggested a multigene approach, which would include nuclear ribosomal ITS or a single-copy nuclear locus yet to be determined. Chase *et al.* (2005) supported the idea of a multigene approach and coined the term multilocus barcode (MBC). Newmaster

Correspondence: S. G. Newmaster, Fax: 519-767-1656; E-mail: snewmast@uoguelph.ca

et al. (2006) expanded the MBC to include the concept of a tiered or nested approach of analysis as a potential method of facilitating the use of noncoding regions in the bioinformatics processes that the Barcode of Life Database (BOLD) (<http://www.boldsystems.org>) currently uses. By utilizing a common, easily aligned gene as a first tier, a more variable (and therefore difficult to align across genera) locus can be incorporated in a nested analysis, which could be run in parallel. Schneider & Schuettpelz (2006) have also shown that DNA-based identification using *rbcL* has considerable potential for exploring the ecology of fern gametophytes, although this is not unexpected given the relatively high rates of plastid sequence evolution in this taxonomic group. We previously examined the utility of *rbcL* (Newmaster *et al.* 2006) as a first tier, noting the low interspecific divergence, and have progressed to test the utility of several other regions. Presting (2006) identified a plastid region conserved in all photosynthetic lineages, termed the universal plastid amplicon (UPA), and recently, Kress & Erickson (2007) have proposed a two-locus barcode based on *rbcL* and the *trnH-psbA* intergenic spacer. The Plant Working Group of the Consortium for the Barcode of Life (PWG-CBOL), led by principals at Royal Botanic Gardens, Kew, UK, identified a number of interim barcode loci candidates (see <http://www.rbgekew.org.uk/barcoding>), and recently proposed two sets of three plastid regions for plant barcoding (Chase *et al.* 2007).

One of the challenges for any barcoding region is its utility in discriminating closely related (i.e. sister-species) or recently evolved species. It is expected that a system based on any one, or small number of chloroplast genes will fail in certain taxonomic groups with extremely low amounts of plastid variation while performing well in other groups. We therefore wanted to focus on a group with low molecular divergence containing some recently evolved species that might be expected to be at the limit of resolution for the proposed regions. Our first objective here is to provide a concise analysis of the suitability of these regions for DNA barcoding in a group that exhibits these characteristics.

The Myristicaceae, or nutmeg family, is an older group within the angiosperms that contains recently evolved species that should provide a challenging test for barcoding plants. The nutmeg family is comprised of ~500 species of canopy to subcanopy trees native to tropical rainforest environments (Smith 1937; Janovec & Harrison 2002). *Compsonneura* is one of five Neotropical genera and is comprised of 21 described species of which some are reproductively isolated by vicariance in Central and South America. Since species of *Compsonneura* share similar leaf morphologies, identification of species relies heavily upon characteristics of the small flowers (1–4 mm) that are only present on adult trees for a few weeks every year (Armstrong 1997). Floral identification is largely dependant on characteristics of the androecium (Smith 1937; Armstrong

& Tucker 1986; Janovec & Harrison 2002), making identification of female members of these dioecious species exceedingly difficult. The Myristicaceae, and more specifically *Compsonneura*, are an ideal group for testing barcoding in plants as they present a taxonomic impediment and the family has been found to have low levels of molecular variation compared to other closely related families (Sauquet *et al.* 2003). *Compsonneura* contains some recently described taxa (Janovec & Neill 2002), a new species split (Janovec & Harrison 2002), and additional new species currently being described (Janovec *et al.* in preparation). The systematic relationships of this group are discussed in a monograph of the genus *Compsonneura* (Janovec *et al.* in preparation). The first part of this study determines the relative utility of seven plastid regions (UPA, *rpoB*, *rpoC1*, *accD*, *rbcL*, *matK* and *trnH-psbA*) for barcoding in this group. Given that a multigene barcode is being proposed for plants, we provide a cursory evaluation of the effect of combining loci on the overlap of intraspecific and interspecific variation.

Besides the issue of choosing an appropriate region for barcoding in plants, plant barcoding will suffer some of the same criticisms currently levied at animal barcoding projects. Of primary concern is the issue of choosing thresholds to delimit species (Ferguson 2002), particularly when intraspecific variation can be shown to be greater than interspecific variation. This is the basis of a recent critical evaluation of barcoding animals, which states that single-gene thresholds for species discovery can result in substantial error in detecting new species that have recent divergence times (Meyer & Paulay 2005; Hickerson *et al.* 2006) and misrepresent the correspondence between recently isolated populations and reproductively isolated species.

Additional logical problems can be encountered when using distance measures to circumscribe species. Intraspecific variation between individuals can exist such that one pairwise distance may exceed a threshold, resulting in an individual both included and excluded from a group. These issues have been more completely addressed previously with empirical and theoretical examples (Meier *et al.* 2006; Little & Stevenson 2007). Despite the known issues with distance or similarity metrics, a review of several sequence identification methods (Little & Stevenson 2007) suggests that algorithms employing similarity methods performed best when the data set is alignable. The second objective of our study is to test whether accurate species assignments can be made with this data set using a similarity method based on the region(s) that exhibit the greatest amount of variation.

Methods

Sampling

Forty individual trees were sampled representing eight species (40% of the known species) of *Compsonneura*, and

Table 1 List of Myristicaceae species with locations and distances among populations

Species	N	Countries	Minimum distance between samples	Maximum distance between samples
<i>Compsonneura atopa</i> (A.C.Sm.) A.C.Sm.	1	Ecuador	N/A	N/A
<i>Compsonneura capitellata</i> Warb.	5	Ecuador	0.5 km	590 km
<i>Compsonneura debilis</i> Warb.	4	Brazil, Venezuela	40 km	1750 km
<i>Compsonneura excelsa</i> A.C.Sm.	5	Costa Rica	0.5 km	50 km
<i>Compsonneura mexicana</i> (Hemsl.) Janovec	9	Belize, Costa Rica	0.5 km	840 km
<i>Compsonneura mutisii</i> A.C.Sm.	5	Ecuador	0.5 km	130 km
<i>Compsonneura sprucei</i> (A.DC.) Warb.	6	Ecuador, Peru	0.5 km	580 km
<i>Compsonneura ulei</i> Warb. ex Pilg.	3	Brazil	240 km	1150 km
<i>Iryanthera lancifolia</i> Ducke	1	Peru	N/A	N/A
<i>Virola sebifera</i> L.	1	Peru	N/A	N/A
Total	40			

N/A, not applicable.

one species each of *Virola* and *Iryanthera*. Since *Compsonneura* species in our analyses exhibit both widespread and endemic ranges in the Neotropics, specimens were selected from our collections of each respective species circumscription as to represent the largest geographical spread as possible (Table 1). Numerous species of *Compsonneura* have been described from only one or few specimens and most of these were not collected with molecular investigations in mind, were poorly preserved, and thus contain little to no amplifiable DNA, limiting the number of taxa in our investigation. Our taxonomic sampling reflects the realities of field sampling in the tropics and availability of specimens; rare species are less well represented than widespread ones, but we have tried to capture the geographical spread of most species. Specimen vouchers are deposited at the Botanical Research Institute of Texas (BRIT) and the Biodiversity Institute of Ontario (BIO) herbaria. Barcode DNA vouchers (leaf material in silica gel and DNA extractions) are filed at the BIO herbarium. Voucher numbers including specimen collection location coordinates and GenBank Accession numbers are provided in the Appendix.

Molecular analysis

Seven regions including portions of six coding regions (*UPA*, *rpoB*, *rpoC1*, *accD*, *rbcL*, *matK*) and one noncoding (*trnH-psbA* intergenic spacer) chloroplast region were selected to determine their suitability for barcoding. DNA extractions were performed using the DNeasy Plant Mini kit (QIAGEN). Manufacturer's protocols were followed except that the incubation step following homogenization was increased to 1 h. Polymerase chain reaction (PCR) amplification was performed on a PTC-100 thermocycler (Bio-Rad) using the published primers and thermal cycling conditions (Kress *et al.* 2005; <http://www.rbgekew.org.uk/barcoding>; Presting 2006). The PCR mix included 10 mM Tris-HCl pH 8.3, 50 mM

KCl, 2.5 mM MgCl₂, 0.2 mM each dntp, 0.1 µM each primer, 1 U AmpliTaq Gold Polymerase (Applied Biosystems), and up to 20 ng template DNA. Initial attempts at PCR failed for *matK*. New primers were designed for this region that differ slightly (one to two nucleotides) from those presented by the Plant Working Group (PWG-CBOL), *matK2.1-Myristicaceae* (5'-CCTATCCATCTGGATATCTTGG-3') and *matK5-Myristicaceae* (5'-GTTCTAGCACACGAAAATCG-3'). Amplification products were diluted and sequenced directly using the same primers used for amplification. Cleaned sequence products were run on an ABI 3730 sequencer (Applied Biosystems). Due to presumed degradation of some samples, not all samples yielded a PCR product for all regions. Sequences were aligned using CLUSTAL W (Higgins *et al.* 1994) and checked manually. Truncated sequence reads resulted in missing nucleotide data for two samples.

In order to obtain an estimate of variation in the seven regions examined, we calculated pairwise uncorrected p-distance for each region using MEGA 3.1 (Kumar *et al.* 2004) (Table 2). These distances were initially used to evaluate intraspecific and interspecific divergence in the samples for all seven regions. Five regions (*UPA*, *rpoB*, *rpoC1*, *accD*, *rbcL*) were largely invariant with a single haplotype predominating among all samples.

Two regions with considerable interspecific divergence (*matK* and *trnH-psbA*) were retained for a multivariate classification analysis. This analysis utilized the raw sequence data from each of the regions in a matrix with all 40 nutmeg specimens. Indels were coded as a fifth character state. Bray-Curtis average linkage was used to create three distance matrices of the 40 specimens using the informative sequence data (variable sites) from (i) *matK*, (ii) *trnH-psbA*, and (iii) both *matK* and *trnH-psbA* combined. Although coding indels as a fifth character state has the potential to overweight distance measures, some of the indels are associated with the

Table 2 Variability in six coding (*rpoB*, *rpoC1*, *accD*, *rbcL*, *matK*, UPA) and one noncoding (*trnH-psbA*) chloroplast regions, including a combination of two regions (*matK+trnH-psbA*) for barcoding ten Myristicaceae species (*resolution, number of species with haplotypes not found in any other species; †mean uncorrected p-distance)

Region	Resolution*	No. of variable sites	Indels (length)	Intraspecific p-distance†	Interspecific p-distance†	Aligned length	No. of samples
UPA	0	1	0	> 0.000	0.001	342	40
<i>rpoB</i>	0	4	0	0	0.001	486	38
<i>rpoC1</i>	0	5	0	> 0.000	0.002	403	40
<i>accD</i>	0	5	0	> 0.000	0.003	341	39
<i>rbcL</i>	0	2	0	0.001	0.002	668	36
<i>matK</i>	5	10	2 (6 bp, 9 bp)	0.005	0.042	761	39
<i>trnH-psbA</i>	10	40	5 (2–5 bp)	0.009	0.060	396	40
<i>matK+trnH-psbA</i>	10	50	7 (2–9 bp)	0.008	0.064	761 + 396	40

homopolymer runs and it is improbable that all indels of a given length in these regions are homologous. The relationship of classification structure in the species data to the molecular characters was analysed with nonmetric multidimensional scaling (NMS; Kruskal 1964; Primer 2002). In NMS, the Bray-Curtis distance measure was used because of its robustness for both large and small scales on the axes (Minchin 1987). Data were standardized by species maxima and two-dimensional solutions were appropriately chosen based on plotting a measure of fit ('stress') to the number of dimensions. Stress represents distortion in the data and a stress value over 0.15 is high enough that the results are invalidated (Primer 2002). One thousand iterations were used for each NMS run, using random start coordinates. The first two ordination axes were rotated to enhance interpretability with the different axes. As an independent check, detrended correspondence analysis (DCA; ter Braak 1998) was used to evaluate the NMS classification. A histogram of the intraspecific and interspecific p-distances is provided for each of the three classifications [(i) *matK*, (ii) *trnH-psbA*, and (iii) both *matK* and *trnH-psbA*] combined. A two-sample Kolmogorov-Smirnov nonparametric statistical test was used to distinguish between the levels of variation in the raw p-distance measures for (i) *matK* and *trnH-psbA*, and (ii) *trnH-psbA* and the combined *matK + trnH-psbA*.

In order to test whether accurate species assignments can be made among the samples in our data set, we used the 'best match' and 'best close match' functions of the program TAXONDNA (Meier *et al.* 2006). This program determines the closest match of a sequence from comparisons to all other sequences in an aligned data set. It establishes a similarity threshold based on the frequency distribution of the intraspecific pairwise distances. The threshold is set at a value below which 95% of all intraspecific pairwise distances are found (Meier *et al.* 2006). Unlike the ordinations we calculated, TAXONDNA ignores indels when calculating distance. These sequence identification methods were performed on the *matK*, *trnH-psbA*, and combined *matK + trnH-psbA* data

sets using uncorrected pairwise distances and a minimum sequence overlap of 300 bp. The inclusion of conspecific individuals is a key component of this type of analysis, as the query sequence is removed from the data set prior to determining its best or closest match. Initial program runs revealed that two species in the data set (*Virola sebifera* and *Iryanthera lancifolia*) were consistently outside the established threshold and failed to be identified correctly as they lacked conspecific individuals in the data set. These two individuals were therefore removed from the identification analysis. One other species (*C. atopa*) is also represented by only a single individual. This sample was retained as previous inspection of the data set indicated a close affinity between *C. atopa* and *C. capitellata*.

Results

Invariant regions

Five of the regions we tested exhibited marginal amounts of variation across all samples (Table 2). The maximum interspecific p-distances were low between all species for *rbcL* (< 0.003), *rpoB* (< 0.004), *rpoC1* (< 0.008), *accD* (< 0.013) and UPA (< 0.001), with many different species presenting identical sequences. The maximum intraspecific p-distance within species of *Compsonneura* was 0.001 in *rbcL*, 0.002 in *rpoC1* and 0.006 in *accD*, but was nil for *rpoB* and UPA. Clearly, none of these five regions have sufficient variation to be suitable for barcoding in *Compsonneura*.

Variant regions

Two of the regions (*matK* and *trnH-psbA*) had significant variation and show promise for barcoding in nutmegs. The level of variation in the raw p-distance measures for *trnH-psbA* was significantly (mean distance 0.060 ± 0.0037 ; $P < 0.001$) higher than those in *matK* (mean distance 0.042 ± 0.0030).

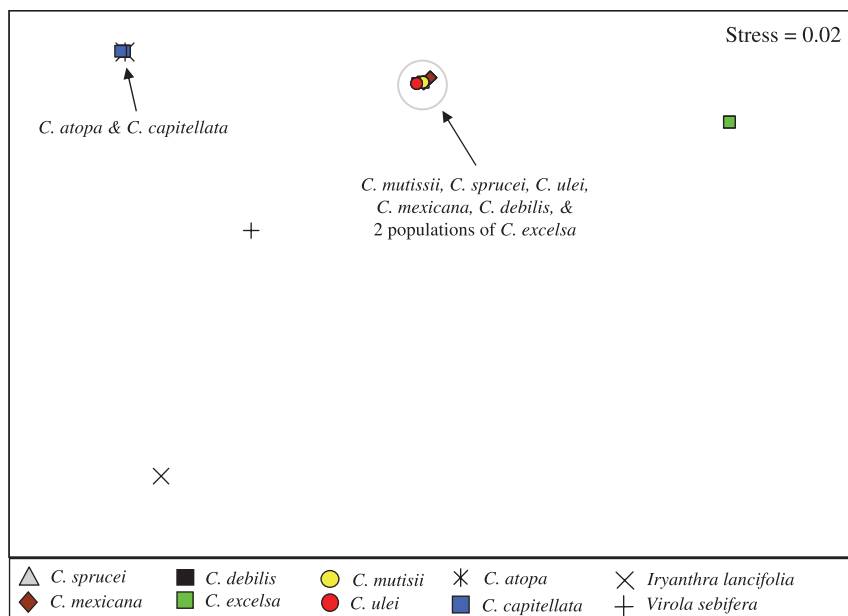


Fig. 1 NMS ordination of 39 individuals of nutmeg species using *matK* sequence data (25 variable sites). Grey circle represents species that exhibit relatively invariant sequences.

matK

The proposed barcoding portion of *matK* is approximately 760 base pairs in *Compsonneura*. One individual of *C. debilis* consistently failed to amplify. Across all samples, we observed 10 variable sites and two indels in *matK* (Table 2). The variation observed in *matK* does separate some species; however, there is a wide range of intraspecific and interspecific variation. The *matK* region presents a single common haplotype across all individuals of *C. mutissii*, *C. sprucei* and two individuals of *C. excelsa*. Within-species variation however, is observed in *C. excelsa*, *C. capitellata* and *C. mexicana*. The sequence divergence in *C. mexicana* appears to be geographically related, with samples from Belize and Costa Rica clustering separately. The sequence data for *matK* do differentiate *C. mexicana* and *C. sprucei* (a recently described species split, Janovec & Neill 2002) based on a consistent single nucleotide difference. The NMS classification used to ordinate the *matK* data (Fig. 1) and the histogram (Fig. 2) reflects the distribution in variation.

The results of the sequence identification analysis reflect the incidence of shared haplotypes between species with 19 allospecific pairwise matches. Under both the best match and best close match functions, less than 50% of the individuals in this data set could be identified using *matK* (Table 3).

trnH-psbA

The *trnH-psbA* intergenic spacer region, approximately 396 base pairs in length in this data set, was successfully sequenced for all 40 samples of nutmegs. Among all samples, there were three indels, 40 variable sites and two regions

of A:T homopolymer runs with length differences of up to 4 bp (Table 2). The variation in *trnH-psbA* as shown by the ordination (Fig. 3) was sufficient to separate all species, including the recently diverged *C. mexicana* and *C. sprucei*. However, the considerable range in intraspecific variation detected (pairwise distances > 0.06) does overlap with the low pairwise distances between some species pairs (0.01–0.04) (Fig. 2). For example, on the NMS ordination, two groups of *C. sprucei* individuals are widely spread on the X-axis (Fig. 3). Further inspection reveals that these two groups of individuals are geographically isolated in different forest ecosystems of the upper Amazon of Peru and Ecuador. Some intraspecific variation in *C. mexicana* was also observed, which is associated with a geographical distance of 840 km.

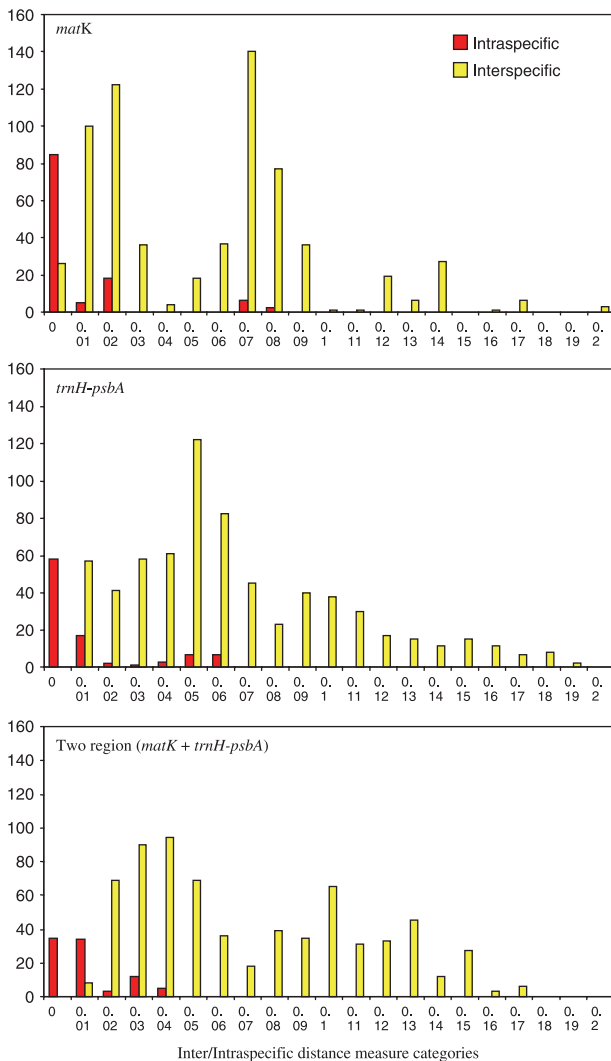
The increase in variation observed in the *trnH-psbA* region over *matK* is also apparent in the sequence identification analysis. Fewer allospecific pairwise matches with 0% distance are observed (Table 3) and the identification success approaches 70%. Although no allospecific sequences are actually identical, the exclusion of indels as informative characters results in reduced identification success.

Two-gene approach

The analysis of a two-gene approach revealed several considerations for barcoding plants. Although the *trnH-psbA* data set does indicate that it would be sufficient to distinguish the species used in this study if no new haplotypes were found (Fig. 3), the large amount of population variation observed would certainly make the placement of an unknown sample problematic. We were interested therefore to see the effect of combining the *matK* and *trnH-psbA* data sets.

Table 3 Identification success based on the 'best match' and 'best close match' functions of the program TAXONDNA (Meier *et al.* 2006)

Region	Best match			Best close match				
	Successfully identified	Ambiguous	Misidentified	Successfully identified	Ambiguous	Misidentified	No match	Threshold
<i>matK</i>	48.6%	48.6%	2.7%	40.5%	48.6%	2.7%	8.1%	0.26%
<i>trnH-psbA</i>	65.8%	28.9%	5.3%	65.8%	28.9%	5.3%	0%	2.52%
<i>matK+trnH-psbA</i>	94.7%	0%	5.3%	94.7%	0%	5.3%	0%	0.86%

**Fig. 2** Histograms of the number of pairwise (y-axes) intraspecific and interspecific divergence distances estimates (x-axes) among all 40 nutmeg samples for *matK*, *trnH-psbA* and a two-region matrix using variable sites from *matK* + *trnH-psbA*.

The level of variation in the raw p-distance measures from the combination of *matK* + *trnH-psbA* was not significantly higher (mean distance 0.064 ± 0.0031 ; $P < 0.116$) than those in *trnH-psbA* alone (mean distance 0.060 ± 0.0037).

However, the NMS classification of the combined data set (Fig. 4), demonstrates that these two regions are complementary, and that by combining data sets, the variation within species circumscriptions is reduced relative to the distance between species. Although inspection of the histogram for this combination does not show a clear gap between intraspecific variation and interspecific divergence, it does suggest a much more defined range where intraspecific variation is considerably lower than the distribution of interspecific divergence (Fig. 2). It is interesting to note that the samples within the sections *Hadrocarpa* and *Compsooneura* as delineated by Janovec (2000) and Janovec & Neill (2002), are identified, respectively, on the left and right side of the ordination (Fig. 4). This split among the *Compsooneura* species represents the sum of the greatest interspecific divergence distances in the histogram aligned with that of other intergeneric distances such those among *Iryanthera*, *Virola* and *Compsooneura*.

The combined data set yields the best result of the sequence identification analysis. No two individuals of different species share identical sequences and the percentage of correct identifications of all pairwise comparisons is 94.7% (Table 3). Two samples were incorrectly identified. The first, *C. atopa* had only a single individual represented, so it logically will be misidentified with its nearest match, since the algorithm removes the query sequence from the data set prior to determining the match. The second incorrect identification was an individual of *C. capitellata* whose nearest match was the single individual of *C. atopa*.

Discussion

The data presented here show that many of the suggested plastid regions for plant barcoding will not differentiate species in *Compsooneura*. As a member of the Myristicaceae, a group with a low amount of genetic divergence, they provide a useful test of the proposed regions. Of the seven regions examined, only the *trnH-psbA* intergenic spacer had unique sequences for each species. The portion of the coding region *matK* examined here was the only other region with considerable variation; however, only half of the species sampled had species-specific sequences. The data from *rbcL*, *rpoB*, *rpoC1*, *accD* and UPA indicated that these

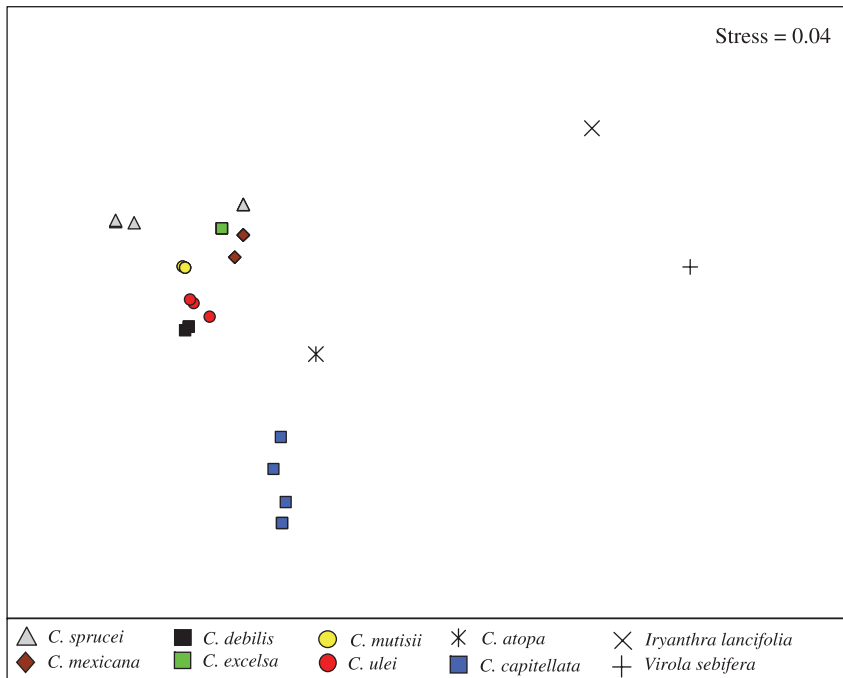


Fig. 3 NMS ordination of 40 individuals of nutmeg species using *trnH-psbA* sequence data (58 variable sites).

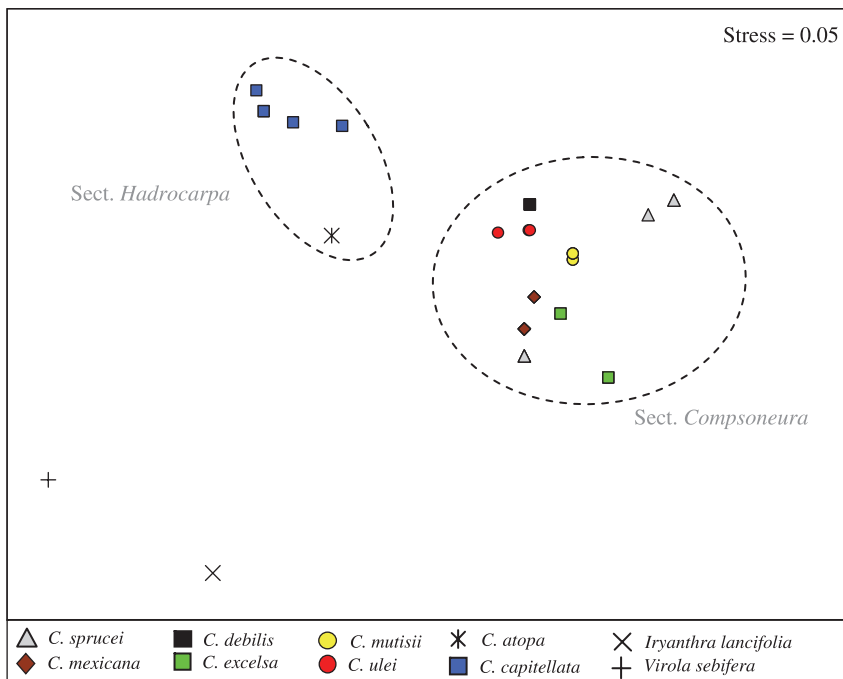


Fig. 4 NMS ordination of 40 individuals of nutmeg species using *matK* and *trnH-psbA* sequence data (83 variable sites). Dotted circles identify the species within the sections *Hadrocarpa* and *Compsoeura*.

regions would have little to no success in identifying species of *Compsoeura*. Although Kress & Erickson (2007) have recently proposed *rbcl* as part of a two-gene system for plant barcoding, inspection of the *rbcl* sequences in our data set shows that *rbcl* has only two polymorphic sites with three widely shared haplotypes. In this instance, the combination of *rbcl* with the *trnH-psbA* intergenic spacer would provide no more resolution than *trnH-psbA* alone.

The universal plastid amplicon identified by Presting (2006) was readily amplified using the conserved primers he identified, but it also had the lowest amount of variation of all regions, among the species and genera of Myristicaceae examined. This is not surprising because UPA is located in the highly conserved 23S rDNA. Although this region does not provide species-level resolution in this family, its utility for barcoding still needs to be addressed across a broader

survey of taxa. UPA may prove to be of significant utility if it is universal in all photosynthetic lineages (with conserved primers), particularly when dealing with environmental samples.

It is generally accepted that a multiregion approach to barcoding plants will be required (Chase *et al.* 2005, 2007; Kress *et al.* 2005; Cowan *et al.* 2006; Newmaster *et al.* 2006; Kress & Erickson 2007). Although a noncoding region such as the *trnH-psbA* spacer may provide good species resolution, it will also likely be so variable for some groups that every population will need to be sampled in order to capture a high proportion of haplotypes. Certainly, our data set reflects the need for widespread sampling. Rates of evolution in plants are highly variable across taxa and the approach to barcoding must consider this variation. An optimal combination of regions will therefore be one that encompasses a large range of variation. Using a combination of a coding region with a noncoding region achieves this, and facilitates new approaches to analysis. Our data set presented here demonstrates this point. When the *matK* and *trnH-psbA* data sets are combined, the largest values of intraspecific variation in the data set were reduced considerably. This also has the effect of reducing the amount of overlap in intraspecific and interspecific variation (Fig. 2) which is a concern when using barcoding to recognize new species (Meyer & Paulay 2005). Importantly, the pairwise distances that do overlap in Fig. 2 are not simply outliers, but represent specimens that warrant further attention.

Our data set also highlights the need for incorporating indels in the analysis. Although the sequence identification method had only a 65.8% success for the *trnH-psbA* spacer, the inclusion of indels would have likely increased the success rate as sequences from some individuals of *C. excelsa* and *C. mexicana* differ only in indel length.

The combination of *matK* plus *trnH-psbA* clearly has the most potential for barcoding in *Compsoeura*. Both the ordination and the sequence identification analysis indicate that by combining regions, a high proportion of individuals can successfully be identified in this group. One important question is whether our result here can be applied more broadly across other taxonomic groups. The higher amount of variation in the *matK* region compared to other plastid coding regions is well known, and it is unsurprising that the noncoding region would have the greatest amount of variation in this data set. However, it is important to characterize these regions more broadly to determine whether some of the less studied regions such as *rpoB* and *rpoC1* are consistently less variable than *matK* as observed here, and which regions are most complementary in a multiregion approach.

Although the *trnH-psbA* spacer and *matK* were the only regions with significant variation, both these regions have characteristics that may prove problematic for barcoding. The primers provided by the Plant Working Group for *matK* did not work well for our samples and required rede-

sign. It is likely that *matK* will require multiple sets of primers across many of the major clades of the plant kingdom. The trade-off between variability and primer universality will likely be an issue with any plastid coding region; a region with sufficient variation to separate species will be so variable that it may require multiple primers. As has been discussed previously (Kress *et al.* 2005; Cowan *et al.* 2006), difficulties with the *trnH-psbA* spacer include problems with alignment between species of different genera, low variability in some groups (e.g. palms, orchids), extreme length variation and absence from some lineages of land plants.

The relative importance of these issues is difficult to quantify. Although considerable effort is being spent on developing universal primers for the regions identified by the Plant Working Group, those involved in animal barcoding apparently view this issue as less important. The animal barcoding projects completed and currently underway, utilize many different sets of primers for *cox1*, with several sets for each taxonomic group. Previously (Chase *et al.* 2006; Newmaster *et al.* 2006), *rbcL* was evaluated as a possible region because of its universality, ease of amplification, ease of alignment, and because there is a significant body of data available for evaluation. It has been shown to differentiate a large percentage of congeneric plant species (Newmaster *et al.* 2006). These evaluations also provide a benchmark with which to compare other regions. *matK* may be a better candidate for a first region because it is more variable than *rbcL*. However, a minimal complement of primers will be optimal for applications of barcoding in ecology or applied projects. When presented with a completely unknown sample, it will be highly desirable to run it with the smallest number of primer sets as possible.

For the *trnH-psbA* spacer, there are likely few taxa with low variability or short fragment lengths, and there are approaches to analysis that have the potential to deal with the issue of alignment. Previously (Newmaster *et al.* 2006), we proposed a tiered or nested approach to analysis as one way to overcome the issue of alignment with noncoding regions such as the *trnH-psbA* spacer. The tiered approach is based on the use of a common, easily amplified, and aligned region (such as *rbcL* or *matK*) that can act as a scaffold on which to place data from a highly variable noncoding region. Searching algorithms can utilize the aligned portion to nest or localize a search, or even differentially weight the two regions. Other solutions incorporating combinations of clustering and similarity methods have also been proposed (Little & Stevenson 2007). For many taxonomic groups such as the Myristicaceae, a noncoding region such as the *trnH-psbA* spacer will likely be required for separating closely related species. Solutions for dealing with groups of plants with extremely low levels of plastid variation or with taxonomically complex groups (TCG's) such as hybrids and species that exhibit introgression, apomixis and backcrossing (Ennos *et al.* 2005; Cowan *et al.*

2006) will depend on technological advances that facilitate use of the nuclear genome.

Character-based approaches will become more attractive as new technologies and methods emerge in molecular biology and bioinformatics. Diagnostic methods for short barcodes based on informative characters have been proposed (DasGupta *et al.* 2005) and have the potential to be more efficient and cost-effective than sequencing long fragments (e.g. 600–800 bp). Hajibabaei *et al.* (2006) found that minimal (i.e. 100 bp) barcodes were effective in identifying animal specimens, confirming their utility in circumstances where full barcodes are unable to be obtained and the identification comparisons are within a confined taxonomic group.

We have identified population level differences in *C. sprucei* associated with ecotypic differences and in *C. mexicana* associated with vicariance. While we are not about to jump to the conclusion that we have recognized new species, the combination of genetic, ecological, and geographical data suggests a closer look is warranted. Janovec & Neill (2002) provide considerable morphological evidence to support differentiation within *C. mexicana* and *C. sprucei*. Individuals of these populations will now be examined in more detail to determine whether they warrant separate species designation. As with all new species discoveries, this would have to be validated by the taxonomic evidence including morphology and potentially additional molecular data. What will need to be investigated is the consistency of characters used for a barcode and whether these characters hold true with increasing sample size.

This barcoding study on nutmegs corroborates the recent description of *C. mexicana* (Janovec & Harrison 2002), and supports the generic split into sections *Hadrocarpa* and *Compsonera*, which was suggested previously based on a suite of morphological character evidence (Janovec 2000). Further sampling is needed to evaluate the population level variation observed in both *matK* and *trnH-psbA*. The considerable variation in *C. sprucei* may represent a widespread species complex undergoing what could be various lines of incipient speciation. We are exploring the variation in more populations, using additional regions to determine if this is best explained by geographical or ecotypic divergence, or an incipient speciation event.

One of the greatest utilities of barcoding is its use in overcoming taxonomic impediments (i.e. identifying unknown leaves, roots, etc.) in ecological studies. In this respect, it will be of enormous benefit in identifying species of *Compsonera* and Myristicaceae that are currently primarily separated by androecium characters in small, short-lived flowers. The future of barcoding holds great promise in invigorating taxonomy and providing other disciplines with a method to overcome the prominent taxonomic impediments and allow more in-depth ecological analysis of species groups and complexes.

Acknowledgements

This research was supported by Genome Canada through the Ontario genomics institute and the Canadian Foundation for Innovation. We thank John Gerrath, Heather Cole, Derek Pieper and Candice Newmaster for their assistance in our lab at the Biodiversity Institute Herbarium. We would also like to thank Robert Hanner, Kevin Burgess, Mehrdad Hajibabaei, Spencer Barrett, Sean Graham and Brian Husband for reviewing an earlier version of the manuscript.

References

- Armstrong JE (1997) Pollination by deceit in nutmeg (*Myristica insipida*, Myristicaceae): floral displays and beetle activity at male and female trees. *American Journal of Botany*, **84**, 1266–1274.
- Armstrong JE, Tucker SC (1986) Floral development in *Myristica* (Myristicaceae). *American Journal of Botany*, **73**, 1131–1143.
- ter Braak CJF (1998) *Canoco 4 Centre for Biometry*. Wageningen, The Netherlands.
- Chase MW, Cowan RS, Hollingsworth PM *et al.* (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**, 295–299.
- 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.
- Cowan RS, Chase MW, Kress WJ, Savolainen V (2006) 300 000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon*, **55**, 611–616.
- DasGupta B, Konwar KM, Mandoiu II, Shvartsman AA (2005) DNA-BAR: distinguisher selection for DNA barcoding. *Bioinformatics*, **21**, 3424–3426.
- Ennos RA, French GC, Hollingsworth PM (2005) Conserving taxonomic complexity. *Trends in Ecology & Evolution*, **20**, 164–168.
- Ferguson JWH (2002) On the use of genetic divergence for identifying species. *Biological Journal of the Linnean Society*, **75**, 509–516.
- Hajibabaei M, Smith MA, Janzen DH *et al.* (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, **6**, 959–964.
- Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identification through DNA barcodes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **270**, 313–321.
- Hebert PD, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a) Ten species in one: DNA barcoding reveals cryptic species in the Neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences, USA*, **101**, 14812–14817.
- Hebert PDN, Stoeckle MY, Zemplak TS, Francis CM (2004b) Identification of birds through DNA barcodes. *Public Library of Science, Biology*, **2**, e312 (www.plosbiology.org).
- Hickerson MJ, Meyer CP, Moritz C (2006) DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology*, **55**, 729–739.
- Higgins D, Thompson J, Gibson T *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.
- Janovec JP (2000) *A systematic monograph of Compsonera, a Neotropical genus of the Myristicaceae family*, PhD Dissertation, Texas A & M University, College Station, Texas.

- Janovec JP, Harrison JS (2002) A morphological analysis of the *Compsoeura sprucei* complex (Myristicaceae), with a new combination for the Central American species *Compsoeura mexicana*. *Systematic Botany*, **27**, 662–673.
- Janovec JP, Neill AK (2002) Studies of the Myristicaceae: an overview of the *Compsoeura atopa* complex, with descriptions of new species from Columbia. *Brittonia*, **54**, 251–261.
- Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *Public Library of Science ONE*, **2**, e508. doi: 10.1371/journal.pone.0000508.
- Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences, USA*, **102**, 8369–8374.
- Kruskal JB (1964) Non-metric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.
- Kumar S, Tamura K, Nei M (2004) MEGA 3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*, **5**, 150–163.
- Little DP, Stevenson WmD (2007) A comparison of algorithms for identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics*, **23**, 1–21.
- Meier R, Shiyang K, Vaidya G, Ng PKC (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.
- Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *Public Library of Science, Biology*, **3**, e422.
- Minchin P (1987) An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, **69**, 89–107.
- Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in the land plants: evaluation of *rbcL* in a multigene tiered approach. *Canadian Journal of Botany*, **84**, 335–341.
- Presting GG (2006) Identification of conserved regions in the plastid genome: implications for DNA barcoding and biological function. *Canadian Journal of Botany*, **84**, 1434–1443.
- Primer Software (2002) *Primer Multivariate Software Version 529*. PRIMER-E Ltd, Hedingham Gardens, Roborough Plymouth, UK.
- Sauquet H, Doyle JA, Scharaschkin T *et al.* (2003) Phylogenetic analysis of Magnoliales and Myristicaceae based on multiple data sets: implications for character evolution. *Botanical Journal of the Linnean Society*, **142**, 125–186.
- Schneider H, Schuettpelz E (2006) Identifying fern gametophytes using DNA sequences. *Molecular Ecology Notes*, **6**, 989–991.
- Smith AC (1937) The American species of the Myristicaceae. *Brittonia*, **2**, 393–510.

Appendix

Genbank Accession numbers, with isolate and collection coordinates (dd, decimal degrees)

Species	Isolate	UPA	<i>rpoB</i>	<i>rpoC1</i>	<i>accD</i>	<i>rbcL</i>	<i>matK</i>	<i>trnH-psbA</i> spacer	Latitude (dd)	Longitude (dd)
<i>Compsoeura atopa</i>	1374	EU090662	EU090544	EU090582	EU090430	EU090508	EU090469	EU090622	-0.78	-77.43
<i>Compsoeura capitellata</i>	835	EU090663	EU090545	EU090583	EU090431	EU090509	EU090470	EU090623	-6.15	-76.17
<i>Compsoeura capitellata</i>	855	EU090664	EU090546	EU090584	EU090432	EU090510	EU090471	EU090624	-3.52	-73.15
<i>Compsoeura capitellata</i>	872	EU090665	EU090547	EU090585	EU090433	EU090511	EU090472	EU090625	-3.52	-73.15
<i>Compsoeura capitellata</i>	875	EU090666	EU090548	EU090586	EU090434	EU090512	EU090473	EU090626	-3.52	-73.15
<i>Compsoeura capitellata</i>	889	EU090667	EU090549	EU090587	EU090435	EU090513	EU090474	EU090627	-1.07	-77.62
<i>Compsoeura debilis</i>	190	EU090668	EU090550	EU090588	EU090436	EU090514	EU090475	EU090628	-10.37	-58.37
<i>Compsoeura debilis</i>	6172	EU090669	—	EU090589	EU090437	—	—	EU090629	2.87	-67.22
<i>Compsoeura debilis</i>	7209	EU090670	EU090551	EU090590	EU090438	EU090515	EU090476	EU090630	2.51	-67.29
<i>Compsoeura debilis</i>	22972	EU090671	EU090552	EU090591	EU090439	—	EU090477	EU090631	1.85	-67.05
<i>Compsoeura excelsa</i>	636	EU090672	EU090553	EU090592	EU090440	EU090516	EU090478	EU090632	8.42	-83.12
<i>Compsoeura excelsa</i>	666	EU090673	EU090554	EU090593	EU090441	EU090517	EU090479	EU090633	8.44	-83.28
<i>Compsoeura excelsa</i>	668	EU090674	EU090555	EU090594	EU090442	EU090518	EU090480	EU090634	8.43	-83.30
<i>Compsoeura excelsa</i>	669	EU090675	EU090556	EU090595	EU090443	EU090519	EU090481	EU090635	8.26	-83.22
<i>Compsoeura excelsa</i>	671	EU090676	EU090557	EU090596	EU090444	EU090520	EU090482	EU090636	8.26	-83.22
<i>Compsoeura mexicana</i>	07	EU090677	—	EU090597	EU090445	EU090521	EU090483	EU090637	16.38	-88.82
<i>Compsoeura mexicana</i>	354	EU090678	EU090558	EU090598	EU090446	EU090522	EU090484	EU090638	10.42	-83.98
<i>Compsoeura mexicana</i>	362	EU090679	EU090559	EU090599	EU090447	EU090523	EU090485	EU090639	10.42	-83.98
<i>Compsoeura mexicana</i>	696	EU090680	EU090560	EU090600	EU090448	EU090524	EU090486	EU090640	16.23	-89.08
<i>Compsoeura mexicana</i>	701	EU090681	EU090561	EU090601	EU090449	EU090525	EU090487	EU090641	16.23	-89.08
<i>Compsoeura mexicana</i>	719	EU090682	EU090562	EU090602	EU090450	EU090526	EU090488	EU090642	16.12	-89.03
<i>Compsoeura mexicana</i>	720	EU090683	EU090563	EU090603	EU090451	EU090527	EU090489	EU090643	16.12	-89.03
<i>Compsoeura mexicana</i>	757	EU090684	EU090564	EU090604	—	EU090528	EU090490	EU090644	16.20	-89.10
<i>Compsoeura mexicana</i>	1283	EU090685	EU090565	EU090605	EU090452	EU090529	EU090491	EU090645	10.42	-83.98
<i>Compsoeura mutisii</i>	911	EU090686	EU090566	EU090606	EU090453	EU090530	EU090492	EU090646	1.05	-78.53
<i>Compsoeura mutisii</i>	913	EU090687	EU090567	EU090607	EU090454	EU090531	EU090493	EU090647	1.05	-78.53
<i>Compsoeura mutisii</i>	914	EU090688	EU090568	EU090608	EU090455	EU090532	EU090494	EU090648	1.05	-78.53
<i>Compsoeura mutisii</i>	1290	EU090689	EU090569	EU090609	EU090456	EU090533	EU090495	EU090649	0.10	-66.80
<i>Compsoeura mutisii</i>	1295	EU090690	EU090570	EU090610	EU090457	EU090534	EU090496	EU090650	0.10	-66.80
<i>Compsoeura sprucei</i>	812	EU090691	EU090571	EU090611	EU090458	EU090535	EU090497	EU090651	-6.06	-76.11
<i>Compsoeura sprucei</i>	817	EU090692	EU090572	EU090612	EU090459	EU090536	EU090498	EU090652	-6.06	-76.11
<i>Compsoeura sprucei</i>	821	EU090693	EU090573	EU090613	EU090460	EU090537	EU090499	EU090653	-6.06	-76.11
<i>Compsoeura sprucei</i>	884	EU090694	EU090574	EU090614	EU090461	EU090538	EU090500	EU090654	-1.04	-77.37
<i>Compsoeura sprucei</i>	887	EU090695	EU090575	EU090615	EU090462	EU090539	EU090501	EU090655	-1.04	-77.37
<i>Compsoeura sprucei</i>	903	EU090696	EU090576	EU090616	EU090463	EU090540	EU090502	EU090656	-1.04	-77.37
<i>Compsoeura ulei</i>	88	EU090697	EU090577	EU090617	EU090464	—	EU090503	EU090657	-3.83	-49.70
<i>Compsoeura ulei</i>	6192	EU090698	EU090578	EU090618	EU090465	EU090541	EU090504	EU090658	-5.82	-50.53
<i>Compsoeura ulei</i>	42644	EU090699	EU090579	EU090619	EU090466	EU090542	EU090505	EU090659	-2.42	-59.90
<i>Iryanthera lancifolia</i>	879	EU090700	EU090580	EU090620	EU090467	—	EU090506	EU090660	-3.52	-73.15
<i>Virola sebifera</i>	779	EU090701	EU090581	EU090621	EU090468	EU090543	EU090507	EU090661	-6.15	-76.17