Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama

W. John Kress1, David L. Erickson, F. Andrew Jones, Nathan G. Swenson, Rolando Perez, Oris Sanjur, and Eldredge Bunting

1Department of Botany, MRC-166, National Museum of Natural History, Smithsonian Institution, P.O. Box 37012, Washington, DC 20013-7012; 2Smithsonian Tropical Research Institute, P.O. Box 0843-03092, Balboa Ancón, Republic of Panama; 3Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot, Berkshire SL5 7PY, United Kingdom; and 4Center for Tropical Forest Science - Asia Program, Harvard University Herbaria, 22 Divinity Avenue, Cambridge, MA 02138

Communicated by Daniel H. Janzen, University of Pennsylvania, Philadelphia, PA, September 3, 2009 (received for review May 13, 2009)

The assembly of DNA barcode libraries is particularly relevant within species-rich natural communities for which accurate species identifications will enable detailed ecological forensic studies. In addition, well-resolved molecular phylogenies derived from these DNA barcode sequences have the potential to improve investigations of the mechanisms underlying community assembly and functional trait evolution. To date, no studies have effectively applied DNA barcodes sensu strictu in this manner. In this report, we demonstrate that a three-locus DNA barcode when applied to 296 species of woody trees, shrubs, and palms found within the 50-ha Forest Dynamics Plot on Barro Colorado Island (BCI), Panama, resulted in >98% correct identifications. These DNA barcode sequences are also used to reconstruct a robust community phylogeny employing a supermatrix method for 281 of the 296 plant species in the plot. The three-locus barcode data were sufficient to reliably reconstruct evolutionary relationships among the plant taxa in the plot that are congruent with the broadly accepted phylogeny of flowering plants (APG II). Earlier work on the phylogenetic structure of the BCI forest dynamics plot employing less resolved phylogenies reveals significant differences in evolutionary and ecological inferences compared with our data and suggests that unresolved community phylogenies may have increased type I and type II errors. These results illustrate how highly resolved phylogenies based on DNA barcode sequence data will enhance research focused on the interface between community ecology and evolution.

The most difficult challenge for DNA barcoding in plants is discriminating among taxa of highly speciose genera where rates of species identification by using a variety of putative barcodes rarely exceed 70% (1). In some cases of complex recently evolved species groups DNA barcoding may simply be inappropriate as an identification tool (2). This difficulty is especially acute in cases where certain life history traits have affected the rates of molecular evolution in a lineage, which in turn may affect rates of species assignment by DNA barcodes [e.g., generation times (3) and age-of-crown group diversification (4)].

In the absence of a universal barcode region capable of discriminating among all species in all groups of plants, it is clear that DNA barcodes will be most effectively applied in the identification of a circumscribed set of species that occur together in a floriistic region or ecological community, rather than in distinguishing among an exhaustive sample of taxonomically closely related species. In these cases only a limited number of closely related species occur in the same region, so identification to genus is all that is required.

It is now generally agreed that a plant barcode will combine more than one locus (5–7) and will include a phylogenetically conservative coding locus (rbcL) with one or more rapidly evolving regions (part of the matK gene and the intergenic spacer trnH-psbA). Although more laborious than a single-locus barcode, multilocus DNA barcodes can also be advantageous in phylogenetic applications owing to increased nucleotide sampling: the conserved coding locus will easily align over all taxa in a community sample to establish deep phylogenetic branches whereas the hypervariable region of the DNA barcode will align more easily within nested subsets of closely related species and permit relationships to be inferred among the terminal branches of the tree.

In this respect a supermatrix design (8, 9) is ideal for using a mixture of coding genes and intergenic spacers for phylogenetic reconstruction across the broadest evolutionary distances, as in the construction of community phylogenies (10). We define a supermatrix as a phylogenetic matrix that may contain a high incidence of missing data and the data content for any one taxon is stochastic (11) (Fig. S1). Confidence of correct sequence alignment is critical in building such complex matrices and testing the robustness and application of these complex data structures in phylogeny reconstruction is a major endeavor (9).

In the field of community ecology, recent investigations have focused on the factors responsible for species assemblages within a specific ecological community (12). In most cases estimates of the phylogenetic relationships among species in an assemblage are the least robust (10). Ideally characters used to generate a phylogeny of a community, e.g., DNA sequence data, should be independent of the functional characters under investigation and reflect true evolutionary relationships among the species actually present in the assemblage. Unfortunately, most community-based studies lack DNA sequence data for all of the taxa and rely on previously published taxon-specific phylogenies (13) organized with programs, such as Phylomatic (14) to construct community phylogenies usually only resolved to the generic and family levels. Despite these drawbacks, such a phylogenetic framework allows one to test the hypotheses that co-occurring species are 1) more closely related than by chance (phylogenetic clustering), 2) more distantly related than by chance (phylogenetic overdispersion) or 3) randomly distributed (10, 15).

In a community analysis of this type, Kenkel and Hubbell (16) constructed a community phylogeny of the 312 co-occurring tree species (>1 cm in diameter) in the forest dynamics plot on Barro Colorado Island in Panama, based on previously published phylogenies assembled with Phylomatic. They found that tree species in the younger forest and drought-stressed plateaus were phylogenetically clustered, whereas coexisting species in the swamp and slope habitats were more distantly related (over-


The authors declare no conflict of interest.

Ffreely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. GQ981925–GQ982412).

1To whom correspondence should be addressed. E-mail: kress@si.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0909820106DCSupplemental.
Results

Sequence Recoverability and Quality. PCR and sequencing success were extremely high for the rbcL region (95%; Table S1). matK had the lowest overall rate of recovery (69%), but the highest sequence quality for recovered samples. For trnH-psbA, we obtained a very high rate of recovery (94%) although some problems were encountered in sequencing (see below). On average the fraction of bases with quality scores >20 were: matK = 95%, rbcL = 93%, and trnH-psbA = 88%. At least one of the gene regions was generated for 98% of the 296 species. For combinations of two regions, recovery rate declined to the rate of the least successful marker (rbcL and matK = 69%; rbcL and trnH-psbA = 93%; trnH-psbA and matK = 69%); all three regions together were recovered 69% of the time (Table S1).

Sequence quality for the rbcL marker was high for all taxa: 85% of the successfully amplified taxa had >50% contig overlap and <9% of sequences were partial. For matK, 50% of the 205 recovered samples had >50% contig overlap, which was largely because of the excessive length of the amplicon (~850 bp using KIM3F.KIM1R primers). Another 20% of matK sequences had sequence quality only in one direction. For trnH-psbA 74% of the 280 samples with sequence data had contigs with >50% contig overlap between sequence reads; the remaining 26% of sequences were interrupted by mononucleotide repeats and had either low overlap between reads or resulted in only partial sequences. Fully 92% of trnH-psbA contigs had full-length sequence when all contigs were included.

Species Assignment and Identification. The highest success of species-level assignment using BLAST with individual barcodes (Table S2) was obtained with matK (99%), followed by trnH-psbA (95%), and then rbcL (75%). However, when rates of assignment are incorporated with rates of sequence recovery (i.e., the product of PCR recovery rate times the correct assignment rate), matK is less successful, resolving 69% of the 286 species compared with 70% for rbcL and 90% for trnH-psbA (Table S2). At the species-level for both matK and trnH-psbA, all failures to assign to the correct species occurred in genera that had numerous species in the plot (Inga, Ficus, and Piper). Assignment to the level of genus ranged from 91% to 100% and all markers provided 100% correct assignment to family (Table S2). The combinations of matK and rbcL correctly identified 92% of all species and trnH-psbA and rbcL correctly identified 95% of all species whereas the three-locus barcode identified 98% of the 286 taxa for which sequence data were available.

Phylogenetic Reconstruction. Each of the trees that were generated from the three different combinations of markers (i.e., trnH-psbA + rbcL, matK + rbcL, and all three regions combined) differed in topology and the degree of resolution. The three-locus combination provided the most fully resolved and well-supported tree (Fig. 1 and Table S3) thus we focus on the rbcL + matK + trnH-psbA tree in the following discussion. The three-locus tree closely matched the topology of the APG II ordinal-level phylogeny (Fig. 2) with no significant discrepancies among the 23 orders. In some cases where the APG II tree does not resolve ordinal relationships (e.g., within the asterids and the rosids), the three-locus barcode phylogeny does (e.g., Lamiales, Solanales, and Boraginaceae; Oxalidales, Celastrales, Malpighiales). In several lineages, the barcode phylogeny slightly contradicted the APG II topology (e.g., Piperales and Arecales; Sapindales, Malvales, and Brassicales). Assignment of families within orders was 100% for all marker combinations (Figs. 1 and 2); the topology of families within the major groups of angiosperms as defined on the barcode tree was also concordant with the APG II classification (e.g., asterids; Fig. 2).

To assess and quantify branch support we defined four separate categories of parsimony ratchet values for each node of the tree: 85% or greater = strong; 70–85% = moderate; 50–70% = weak; 50% or less = poor or no support. The most strongly supported tree was the three-locus supermatrix tree constructed by using MP (Table S3). All 23 orders were supported as monophyletic (ratchet values >70%) with 19 of these orders strongly supported. We were able to obtain good sequence data in 56 of the total 57 families in the plot (with poor quality sequence for one sample in the Rhamnaceae). Of these 56 families 53 either had only one genus represented on BCI or were supported as being monophyletic with 46 (89%) demonstrating strong support (medium support level = 98%). The mononophriletic families were the Bignoniaceae, Euphorbiaceae, and Salicaceae. Of the 47 genera with more than one species on BCI 35 (74%) were monophyletic (nonmonophyletic genera were in the Laurales, Arecales, Ericaceae, Myrtales, Sapindales, Malpighiales, and Fabales); 91% of these monophyletic genera were strongly supported (median support of 97%). The total fraction of nodes throughout the tree that were supported at each level was: strong = 85.7%; moderate = 1.8%; weak = 10.2%; poor = 2.2%. In contrast, in the most fully resolved of the ML bootstrap analyses (the three-locus supermatrix) the fraction of nodes that were supported at each level was: strong = 54.3%; medium = 11.3%; weak = 16%; poor = 21%. In a comparison of the three-locus barcode tree with that of the most up-to-date Phylomatic tree for all taxa on BCI, the barcode tree has 97.5% of the nodes resolved, whereas only 48.4% of the nodes of the Phylomatic tree are resolved. Within the large plant families present on BCI, this disparity between the barcode tree and the phylomatic tree in resolved nodes ranges from 100% versus 19.35% in the Rubiaceae (32 taxa) to 82.35% versus 50.0% in the Fabaceae (35 taxa).

Community Phylogenetic Structure Analyses. Our pruned version of the Kembel and Hubbell (16) Phylomatic tree yielded the same set of inferences as they concluded in their original work (table 5 in ref. 16; Table S4). Specifically, we found phylogenetic clustering in the “high plateau” and “young” habitats by using dispersal (10, 19–22).
just the ordinal relationships among the species in the BCI flora (see Fig. 2). Nodes with strong ratchet support (50–70%) are indicated by an open triangle. Nodes with weak ratchet support (70–85%) or weak (50–70%) are indicated by an open triangle.

Fig. 1. Maximum-parsimony tree of 281 species of woody plants in the Forest Dynamics Plot on Barro Colorado Island (BCI) based on a supermatrix analysis of rbcl, matK, and trnH-psbA sequence data. Color highlights indicate orders represented on BCI. The small tree at the bottom of the central column shows the ordinal relationships among the species in the BCI flora (see Fig. 2). Nodes with strong ratchet support (>85%) are indicated by an asterisk and nodes with moderate (>70–85%) or weak (<50–70%) support by an open triangle.
Using the barcode phylogeny, we found significant phylogenetic structuring for eight of the 14 inferences made (Table S4). When comparing our barcode phylogeny to the less resolved Phylomatic tree, nine of 14 inferences were dissimilar. Analyses based on the barcode phylogeny identified significant phylogenetic structure in five cases for which the pruned Kembel and Hubbell phylogeny did not. In two cases (NRI values on the “slope” and “young” habitats) significant phylogenetic structure was revealed in both analyses, but the opposite inference was made. In the remaining two cases, the pruned Kembel and Hubbell phylogeny demonstrated significant phylogenetic structuring and the barcode phylogeny did not. Only two of the five cases of significant phylogenetic structuring reported in Kembel and Hubbell and confirmed with our pruned Phylomatic tree were supported by the barcode phylogeny. When directly comparing the NRI and NTI values from the 1,250 individual subplots within the seven habitat types by using paired t tests, eight of the 14 comparisons were significantly different between the pruned Kembel and Hubbell phylogeny and the barcode tree (Table S4).

Discussion

Species Identification. We have demonstrated here that a three-locus DNA barcode provides accurate species identifications for 98% of the 281 taxa on BCI for which we had data. The three loci, *rbcL*, *matK*, and *trnH-psbA*, correctly identified species at 75%, 99%, and 95% success, respectively (Table S2). As expected *rbcL* demonstrated insufficient sequence variation to distinguish among closely related species (5, 23). *matK* was principally handicapped in our study by poor PCR recovery. Other investigations have shown higher rates of recovery for *matK* (7, 24), which suggests that the success of this locus may be improved in the future.

The *trnH-psbA* spacer alone showed the highest rates of species identification (90%) of the three loci when both sequence recovery and correct assignment were taken into account. By combining all three loci correct species assignment was improved to an impressive 98%. The taxonomic distribution of the 296 species on BCI, which includes a high proportion of single-species families and genera (Dataset S1), almost certainly yields higher rates of correct assignment than would be the case for communities comprised of more taxonomically dense samples of species.

In summary, our results show that DNA barcodes can be very effective in the context of a clearly circumscribed floristic sample or plant community, and that additional data, such as geography and morphology may be required to obtain higher rates of species identification in other contexts. The importance of a solid taxonomic foundation (17) for such applications of DNA barcoding cannot be overemphasized.

Community Phylogenies. Our application of an aligned supermatrix of the three barcode loci permitted the construction of a well-resolved phylogeny of the woody plant community in the forest dynamics plot on BCI. The ability to build such well-resolved phylogenies adds extra value to the development of DNA barcode libraries in general and supports future endeavors to collect DNA barcodes in ecologically important environments, especially in the tropics.

The high level of correspondence of the topology of the BCI barcode phylogeny to the APG II classification is a clear reflection of the central role that the plastid gene *rbcL* has played in our overall understanding of the evolution of the angiosperms (13, 25). In this respect the selection of *rbcL* as a component of the plant DNA barcode was not simply fortuitous (5). However, the ability of a 540-bp portion, which was selected for use as a plant barcode, of the complete 1,428-bp gene to resolve higher
level relationships across a broad section of the angiosperms was surprising. The successful alignment of matK through back-translation across the entire taxonomically diverse sample set, as well as, the use of a supermatrix alignment for the hypervariable trnH-psbA spacer also account for the success of our phylogenetic reconstruction. Supermatrix methodology allows for broad taxonomic sampling, where alignment across all phylogenetic markers and for all taxa is problematic. Thus, despite the sparseness of our alignment matrix (94% of the cells were missing data) the topological arrangements closely approximated expected relationships according to APG II. The importance of including in a supermatrix one gene that is present for all taxa and readily aligns across the full matrix, e.g., rbcL in our analysis, has been shown in other studies using supermatrices for building phylogenies (8, 9).

The most advantageous aspects of generating community phylogenies from DNA barcode sequences are the refinements and improvements provided in the resolution of the terminal branches of the tree. Trees produced by Phylomatic may in principle result in equally resolved terminal branches if the species have been included in previously published molecular phylogenetic investigations (14). This situation is unlikely for a majority of communities and it is particularly unlikely for tropical communities where phylogenetic information at the species-level is rare. The ability of the barcode phylogenies both to resolve the terminal branches of phylogenies and to provide refined measures of branch lengths is because the phylogenies are built from sequence data of individuals representing the species in the community.

Highly variable intergenic spacers, which present some disadvantages when used as a DNA barcode because of problems with alignment and length variation (26), provide a significant amount of sequence variation when reconstructing phylogenies within a family or a genus (27, 28). In such taxonomically narrow phylogenies, intergenic spacers contain few indels and are likely to have many more point mutations than a coding region, thereby providing more phylogenetic information. Our application of a supermatrix approach to the analysis of community phylogenies has allowed us to avoid the alignment problems of intergenic spacers while taking advantage of the high levels of sequence variation found within them. The combination of the slower evolving coding gene rbcL with the faster evolving spacer trnH-psbA and coding matK gene in a supermatrix provides the appropriate complement of sequence evolution for constructing phylogenies of the set of disparate taxa which occur in tropical plant communities, such as BCI. If greater phylogenetic resolution is required for a particular group of taxa, the analysis can always be enhanced with more extensive and targeted sequence data from additional genes, such as the nuclear internal transcribed spacer (29).

Community Structure. The primary use of phylogenies in plant community ecology has been to determine whether coexisting species are more or less closely related than expected by chance, which would be evidence for the overriding influence of habitat filtering as a determinant of community membership. Conversely, the coexistence of distantly related species would be evidence for the overriding influence of biotic interactions in determining community membership.

Of the seven habitat types in the BCI forest dynamics plot, our analyses demonstrated that five contained nonrandom phylogenetic assemblages. Thus, in general we can conclude that the coexistence of species in this forest is structured by abiotic and biotic interactions. In particular, both the low plateau and slope habitats had assemblages where the coexisting species were more closely related than expected by chance. This result suggests that habitat filtering plays a dominant role in structuring these communities. The high plateau and the mixed and young habitats were phylogenetically overdispersed and thus, in these habitats biotic interactions are likely to play a dominant role in determining species coexistence. Along streams and in swamps species assemblages were no different from those expected by chance, which suggests that phylogenetic relatedness may not play a role in determining the structure of communities in these habitats.

Our reexamination of the phylogenetic structure of subplots in the different BCI forest habitats revealed the power of a more finely resolved barcode phylogeny, which increased the statistical power to reject the null expectation and showed that the less resolved Phylomatic trees are prone to falsely accepting a null hypothesis of no effect (Type II error). For example, in five cases the barcode phylogeny found significant phylogenetic structure whereas the Phylomatic phylogeny found none. Furthermore, of the five cases of significant phylogenetic structuring detected by the Phylomatic tree, only one was supported by the barcode analyses and most of the barcode results tended to find more phylogenetic overdispersion.

The popularity and success of the Phylomatic approach to building community phylogenies has been based upon the ability of the program to produce rapid estimates of phylogenetic relatedness of complex lists of species within a community. It is now clear that studies of community evolution will be substantially improved by the production of more finely resolved phylogenetic trees and that DNA barcodes stand poised to serve as an efficient and effective approach to building community phylogenies. Such studies will have a bearing on the factors that govern not only local plant community assembly and dynamics, but also on understanding niche conservatism and the dynamics of species composition at landscape and global scales (30, 31).

Materials and Methods

Tissue Sampling and DNA Extraction, Amplification, and Sequencing. The 1,000 × 500 m Forest Dynamics Plot on Barro Colorado Island (BCI) in Panama was established in 1982 and all tree stems >1 cm diameter at breast height (dbh) have been mapped and identified to species in repeated censuses (32). DNA barcodes were generated for one to four tagged individuals for each of the 296 species (1,035 total samples) recorded from the plot in the last census in 2006 representing 23 orders, 57 families, and 181 genera. The standing plant serves as a living voucher for the life of each tagged individual. In addition, an herbarium voucher for each species sampled for this study is deposited at STRI and US (see Dataset S1).

Tissue samples were field collected at BCI and preserved through either frozen storage or silica gel desiccation. Approximately 50 mg of each sample of leaf material was placed within a well of a 2-μl polypropylene 96-well matrix screen-mate plate (Matrix Technologies) and transferred to the Laboratories of Analytical Biology at the Smithsonian Institution for DNA extraction and sequencing. Details on extraction, PCR, and robotic sequencing are provided in SI Materials and Methods and Table S5.

Sequence Editing, Alignment, and Assembly into a Supermatrix. Recovered trace files for each of the three markers were imported into Sequencher 4.8, trimmed, and assembled into contigs. Each of the three markers was handled differently in alignment. Alignment of rbcL was unambiguous because of the absence of indel variation and all sequences were readily assembled. For matK we used transAlign (33) to perform alignment via back-translation. The matK sequences were then aligned with each other and concatenated onto the rbcL alignment using MacClade (34) to produce a two-gene alignment for all taxa.

The trnH-psbA sequences were partitioned by family or order to build eighteen separate taxonomically structured sequence files, which were each then aligned by using Muscle (35). In cases where only one species per family was present in the plot, that individual was aligned with another family in the same order. When only a single species was represented for an order the sequence was not included in the phylogenetic analysis because the rbcL sequence would accurately place the taxon on the tree. The individual sets of aligned trnH-psbA sequences were then assembled into a supermatrix by sequentially concatenating them with the rbcL + matK alignment by using MacClade (Fig. S1). The resulting matrix had >94% of the cells consisting of
Methods and Materials.

Rates of Sequence Assignment to Species. The rate at which each barcode marker could be assigned to the correct species was determined by using BLASTn (short nearly exact search; 36). All recovered sequences at each of the three markers were formatted as both database and query; all query sequences were compared against the entire database library of sequences. For barcode sequences that varied within a species, all variant haplotypes were included within the database and queried in the BLASTn searches. For species lacking intraspecific variation only a single individual was included. A sequence was counted as being correctly assigned when that species had the highest score among all candidates; a sequence was not counted as correctly assigned when the correct species was either tied with another species, or received a lower score. All barcodes were tested singly and in combination.

Phylogenetic Reconstruction. We reconstructed a community phylogeny for the woody species of the BCI plot using maximum likelihood (ML) and maximum parsimony (MP) algorithms with three different marker combinations: rbcL + matK, rbcL + trnH-psaA, and rbcL + matK + trnH-psaA. For each combination either 281 or 277 of the 296 species were included depending on the available sequences. ML analyses were conducted using RAxML (37) and the MP analysis using PAUP v.4.0 (38), both run through the CIPRES supercomputer cluster (www.phylo.org). Implementation of the MP ratchet (39) was used to assess support for trees. To assess phylogenetic structure the trees were compared with the topology of APO II (13) using Mesquite (34) by directly projecting ordinal-level trees and family-level trees (within the asteroid clade) in opposition. For more details on phylogeny construction see SI Materials and Methods (S1).

Community Phylogenetic Structure Analyses. One of the 121 equally parsimonious trees of the three-locus MP analysis of 281 BCI taxa was selected to quantify the phylogenetic structure of tree assemblages in different habitats in the forest dynamics plot. Proportional branch lengths from the MP barcode phylogeny were applied to the community analyses, which were designed to emulate the original phylogenetic structuring-habitat analyses of this plot performed by Kemel and Hubbell (16). To facilitate a direct comparison between the two approaches, we pruned their Phylomatic phylogeny to include only the taxa found in our MP barcode phylogeny and retained their original branch lengths produced by using the Phylocom algorithm blad (40).

For both the barcode and Phylomatic phylogenies, we quantified by using Phylocom (40) the Net Relatedness Index (NRI: 10) and the Nearest Taxon Index (NTI: 10) for each 400-m² subplot (n = 1,250 subplots). Positive NRI and NTI values indicate phylogenetic clustering; negative NRI and NTI values indicate phylogenetic overdispersion.

Because the NRI and NTI values in the 1,250 subplots were spatially autocorrelated, we estimated the mean NRI and NTI values within habitats by using simultaneous spatial autoregression analyses. Next, we used the habitat defined for each 400 m² subplot (41) to determine, using t-tests, whether each habitat tended to contain subplots that were on average phylogenetically clustered, overdispersed, or random. Finally, for each subplot, we compared the barcode NRI and NTI values to those values calculated from the Phylomatic phylogeny by using paired t-tests. For more details on the community structure analyses see SI Materials and Methods (S1).

ACKNOWLEDGMENTS. We thank Ida Lopez for assistance with the figures, Lee Weigt and Ken Wurdack for advice in the lab, Jamie Whittaker for sequence data management, Kyle Harms for providing the habitat maps for Barro Colorado Island, and Steve Hubbell for the insights and energy that he has put into the Forest Dynamics Plot. This work supported by National Science Foundation Grant DEB 043665, the Smithsonian Institution, and a Tupper Postdoctoral Fellowship in Tropical Biology.

Supporting Information

Kress et al. 10.1073/pnas.0909820106

SI Materials and Methods

DNA Extraction, Amplification, and Sequencing. Field collected tissues preserved through flash freezing or silica gel desiccation were derived entirely from photosynthetic material. Leaf material was disrupted in a Tissuelyzer (Qiagen after which tissues were incubated overnight at 55°C using a CTAB-based extraction buffer from by AutoGen. Following incubation the supernatant was removed and placed in a clean 2-mL 96-well plate for submission to an AutoGen 960 DNA extraction robot. DNA extractions were transferred to Matrix barcode tubes (Matrix Technologies) and stored at −80°C. Working stocks of DNA were transferred to a microtiter plate, diluted 5× with 100 μM Tris-HCl (pH 8.0) and then taken to the PCR laboratory.

PCR and Sequencing. We used routine PCR, with no more than three attempts per sample to recover a PCR amplicon for all 1,035 samples. The PCR cycling conditions were exactly the same for rbcLa and trnH-psbA [95°C 3 min (94°C 30 s, 55°C 30 s, 72°C 1 min) × 33 cycles, 72°C 10 min] following procedures outlined in Kress and Erickson (5), with matK requiring lower annealing temperatures and more cycles [95°C 3 min (94°C 30 s, 49°C 30 s, 72°C 1 min) × 40 cycles, 72°C 10 min] following Fazekas et al. (6) and always included DMSO to a final concentration of 5%. Primer pairs for each of the gene regions are listed in Table S5. Successful PCRs were purified by using a 5× diluted mixture of ExoSap (USB). For sequencing, 2–4 μL of the purified PCR was used in a 12-μL reaction [0.8 μL of BigDye terminator sequencing mixture (V3.1; ABI), 2.0 μL of a 5× buffer (400 μM Tris-HCl, pH 8.0), 1 μL of 1 μM primer, and distilled water to volume]. Sequencing of matK PCR products included DMSO to a final concentration of 4% in the reaction mixture. Cycling sequencing protocols were the same for all markers, [95°C 15 s (94°C 15 s, 50°C 15 s, 60°C 4 min) × 30]. Following cycle sequencing, products were purified on a column of sephadex and sequence reactions were read on an ABI 3730.

Sequence Editing, Alignment, and Assembly into a Supermatrix. Recovered trace files for each of the three markers were imported into Sequencher 4.8 (GeneCodes Corp.), trimmed, and assembled into contigs. Each of the three markers was handled differently in alignment. The rbcLa marker was aligned in Sequencher 4.8. Alignment was unambiguous because of the absence of indel variation and all rbcLa sequences were readily aligned with each other in a global alignment. The global rbcLa alignment was then exported from Sequencher as an aligned nexus file.

For alignment of matK, sequences were exported individually (i.e., unaligned) in FASTA file format from Sequencher. We then used transAlign (33) to perform alignment via back-translation. The matK sequences (with one per species as available) were aligned simultaneously with each other in this manner and saved as an aligned FASTA file. That aligned matK FASTA file was then concatenated onto the rbcLa alignment by using MacClade (34) to produce a two-genus alignment for all taxa.

For trnH-psbA, contigs were exported from Sequencher as unaligned FASTA files. FASTA sequences were then partitioned taxonomically, primarily by family. In cases where only one species per family was present in the plot, that individual was aligned with another family in the same order (e.g., Cassipourea elliptica in the Rhizophoraceae of the Malpighiales). When only a single species was represented for an order (e.g., Turpinia occidentalis in the Staphyleaceae, which was one of only six cases in the analysis, Dataset S1), the trnH-psbA sequence of that species was not included in the phylogenetic alignment. Each set of taxonomically structured sequences was then aligned by using Muscle (35) with default parameters. A total of eighteen separate taxonomically structured files were generated in this way. Assembly of the different sets of aligned trnH-psbA sequences into a supermatrix was achieved by sequentially concatenating them with the rbcLa + matK alignment in a supermatrix format as described below (e.g., Fig. S1) again by using MacClade. The resulting matrix was very sparse, with >94% of the matrix consisting of missing data or gaps.

Phylogenetic Reconstruction. We reconstructed a community phylogeny for BCI by using maximum likelihood (ML) and maximum parsimony (MP) algorithms. Three different marker combinations were examined for performance in phylogenetic reconstruction: rbcL + matK, rbcL + trnH-psbA, and rbcL + matK + trnH-psbA. For the combinations of rbcL + matK and rbcL + matK + trnH-psbA, 281 of the 296 species were included with four taxa missing rbcL data; for the rbcL + trnH-psbA matrix 277 of the 296 species were included and all taxa had rbcL data. ML analyses were conducted using RAxML (37) via the CIPRES supercomputer cluster (www.phylo.org). The different locus combinations were partitioned for independent model assessment at each marker. For all combinations of markers a single most likely tree was estimated in addition to running 200–250 bootstrap replicates depending on the marker set. The same gene combinations were used in a MP analysis using PAUP* v.4.0 (38) run through the CIPRES cluster. Implementation of the MP ratchet (a rapid approximation to bootstrap values; ref. 39) was used to assess support for trees, with 200 ratchet iterations for each marker combination used. For both ML and MP trees, a 50% majority tree was constructed and used to quantify overall levels of support for each node within the trees, the rates of well-supported monophyly for taxonomic hierarchies (order, family, genus) and concordance with expected topologies.

Trees were then compared with expectations from the topology of the APG II tree (13) directly using Mesquite (34) by projecting ordinal- and family-level trees in opposition and comparing topologies. Hence for each marker combination, a comparison with expectations of ordinal relationship derived from APG II was performed. Family-level relationships within the asterid clade were also compared with APG II.

Community Phylogenetic Structure Analyses. One of the 121 equally parsimonious trees of the three-locus MP analysis of 281 BCI taxa was selected to quantify the phylogenetic structure of tree assemblages in different habitats in the forest dynamics plot. Proportional branch lengths from the MP barcode phylogeny were applied to the community analyses, which were designed to emulate the original phylogenetic structuring-habitat analyses of this plot performed by Kembel and Hubbell (16). To facilitate a direct comparison between the two approaches, we downloaded the Phylomatic phylogeny from Kembel and Hubbell (16) and pruned it to include only the taxa found in the three gene MP barcode phylogeny. We retained the branch lengths of the Kembel and Hubbell tree that they originally produced by using the Phylocom algorithm bladj (40). The bladj algorithm could, in principle, be applied to our barcode phylogeny, as well. However, this process would require all branch lengths in the barcode
phylogeny to be first set to one thereby losing valuable information on genetic distances resulting from the three loci used to construct the tree. Future sensitivity analyses will be needed to explore the relative advantages of using molecular branch lengths compared with those estimated by using algorithms such as bladj.

For both the barcode and Phylomatic phylogenies, we quantified the Net Relatedness Index (NRI: 10) and the Nearest Taxon Index (NTI: 10) for each 400-m² subplot (n = 1,250 subplots). The analyses were conducted by using the software Phylcom using the independent swap null model (40). Positive NRI and NTI values indicate that co-occurring species are more closely related than expected by chance (i.e., phylogenetically clustered). Negative NRI and NTI values indicate that co-occurring species are more distantly related than expected by chance (i.e., phylogenetically overdispersed).

Because the NRI and NTI values in the 1,250 subplots were spatially autocorrelated, we estimated the mean NRI and NTI values within habitats by using simultaneous spatial autoregression analyses. Specifically, we transformed all NRI and NTI values on the basis of a first order queen case spatial connectivity matrix by using the R package spdep. These estimates were calculated to simulate the spatial autoregression estimates that were quantified by Kembel and Hubbell (16). Next, as in Kembel and Hubbell (16), we used the habitat defined for each 400-m² subplot (41) to determine whether each habitat tended to contain subplots that were on average phylogenetically clustered, phylogenetically overdispersed, or phylogenetically random by using $t$ tests. Finally, for each of the 1,250 subplots, we compared the NRI and NTI values quantified from the barcode phylogeny to those calculated from the Phylomatic phylogeny using a paired $t$ test.
Fig. S1. Schematic representation of supermatrix approach in sequence alignment for phylogenetic analysis using exemplars from the BCI flora.
Fig. S2. The spatial distribution of the BCI habitats (41) and the NRI and NTI values in the 1,250 400-m² quadrants calculated by using the barcode phylogeny and the Phylomatic phylogeny. Negative NRI and NTI values indicate phylogenetic overdispersion and positive values indicate phylogenetic clustering. The color scales across all NRI and NTI maps were made equivalent to allow for direct visual comparisons between the four maps.
Table S1. PCR and sequencing (SEQ) results for 296 species, genera, and families of trees and shrubs in the Forest Dynamics Plot on Barro Colorado Island

<table>
<thead>
<tr>
<th></th>
<th>trnH-psbA PCR</th>
<th>trnH-psbA SEQ</th>
<th>rbcL PCR</th>
<th>rbcL SEQ</th>
<th>matK PCR</th>
<th>matK SEQ</th>
<th>trnH-psbA + rbcL Either SEQ</th>
<th>trnH-psbA + rbcL Both SEQ</th>
<th>matK + rbcL Either SEQ</th>
<th>matK + rbcL Both SEQ</th>
<th>matK + rbcL + trnH-psbA All SEQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of species (296)</td>
<td>290</td>
<td>280</td>
<td>277</td>
<td>277</td>
<td>252</td>
<td>205</td>
<td>291</td>
<td>279</td>
<td>279</td>
<td>205</td>
<td>204</td>
</tr>
<tr>
<td>%</td>
<td>0.97</td>
<td>0.94</td>
<td>0.93</td>
<td>0.93</td>
<td>0.85</td>
<td>0.69</td>
<td>0.98</td>
<td>0.94</td>
<td>0.94</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>No. of samples (1,035)</td>
<td>968</td>
<td>902</td>
<td>862</td>
<td>858</td>
<td>723</td>
<td>650</td>
<td>970</td>
<td>825</td>
<td>859</td>
<td>650</td>
<td>645</td>
</tr>
<tr>
<td>%</td>
<td>0.93</td>
<td>0.87</td>
<td>0.83</td>
<td>0.83</td>
<td>0.70</td>
<td>0.62</td>
<td>0.93</td>
<td>0.79</td>
<td>0.83</td>
<td>0.63</td>
<td>0.62</td>
</tr>
<tr>
<td>Order* (23)</td>
<td>—</td>
<td>—</td>
<td>23</td>
<td>—</td>
<td>22</td>
<td>—</td>
<td>23</td>
<td>22</td>
<td>23</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Family† (57)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>22</td>
<td>—</td>
<td>22</td>
<td>23</td>
<td>22</td>
<td>23</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Genus‡ (181)</td>
<td>—</td>
<td>173</td>
<td>—</td>
<td>170</td>
<td>—</td>
<td>138</td>
<td>173</td>
<td>170</td>
<td>170</td>
<td>138</td>
<td>138</td>
</tr>
</tbody>
</table>

For the number of species and number of samples the count for each category is given with the percentage relative to the total possible listed above.

*Six orders have one species only.
†matK missing for Picramia which is an order containing a single species on BCI.
‡16 families have one species only.
§127 genera have one species only.
Table S2. BLAST results for frequency of correct identification (CI) for all species, genera, and families

<table>
<thead>
<tr>
<th>Taxonomic rank</th>
<th>Measure</th>
<th>trnH-psbA</th>
<th>rbcL</th>
<th>matK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species (296)</td>
<td>CI frequency</td>
<td>95</td>
<td>75</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Recovery × CI</td>
<td>90</td>
<td>70</td>
<td>69</td>
</tr>
<tr>
<td>Genus (181)</td>
<td>CI frequency</td>
<td>100</td>
<td>91</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Recovery × CI</td>
<td>95</td>
<td>85.5</td>
<td>69</td>
</tr>
<tr>
<td>Family (57)</td>
<td>CI frequency</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Recovery × CI</td>
<td>95</td>
<td>94</td>
<td>69</td>
</tr>
</tbody>
</table>

The number of taxa for each taxonomic level is given in parentheses. For each level of the taxonomic hierarchy the frequency of correct identification for each marker and the product of sequence recovery and CI are provided.
Table S3. Parsimony ratchet support values for nodes of phylogenies using two- and three-gene regions

<table>
<thead>
<tr>
<th>Order (n)</th>
<th>Resolution</th>
<th>( rbcLa + matK )</th>
<th>( rbcLa + matK + trnH-psbA )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;50%</td>
<td>&gt;50% to 70%</td>
<td>&gt;70% to 85%</td>
</tr>
<tr>
<td>Fabales (35)</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Rosales (25)</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Malpighiales* (47)</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Sapindales (26)</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Malvales/Brassicales (14)</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Myrtales† (27)</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Gentianales (38)</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Apiales/Solanales/ Boraginaceae (14)</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Ericales/Caryophyllales (18)</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Basal angiosperms‡ (37)</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>15 (5.3%)</td>
<td>30 (10.7%)</td>
<td>5 (1.8%)</td>
</tr>
</tbody>
</table>

Results are partitioned taxonomically by order, with the number of nodes per order given in parentheses after the ordinal name.

*Malpighiales plus Oxalidales and Celastrales.
†Myrtales plus Picramiaceae and Crossopterales.
‡Basal angiosperms comprised of Arecales, Laurales, Magnoliidae, and Piperales.
Table S4. Estimated mean and standard error (SE) of the NRI and NTI values in the BCI habitats estimated by using first-order simultaneous spatial autoregression for the barcode and the Phylomatic phylogenies

<table>
<thead>
<tr>
<th>Habitat</th>
<th>Barcode</th>
<th></th>
<th></th>
<th>Phylomatic</th>
<th></th>
<th></th>
<th></th>
<th>Barcode</th>
<th></th>
<th></th>
<th>Phylomatic</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean (SE)</td>
<td>P</td>
<td>Mean (SE)</td>
<td>P</td>
<td>Mean (SE)</td>
<td>P</td>
<td>Mean (SE)</td>
<td>P</td>
<td>Mean (SE)</td>
<td>P</td>
<td>Mean (SE)</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>High plateau</td>
<td>170</td>
<td>-0.179 (0.046)</td>
<td>&lt;0.05</td>
<td>0.173 (0.069)</td>
<td>&lt;0.05</td>
<td>-0.352 (0.058)</td>
<td>&lt;0.05</td>
<td>-0.199 (0.072)</td>
<td>&lt;0.05</td>
<td>0.070 (0.089)</td>
<td>0.43</td>
<td>-0.269 (0.091)</td>
<td>&lt;0.05</td>
<td></td>
</tr>
<tr>
<td>Low plateau</td>
<td>620</td>
<td>0.171 (0.049)</td>
<td>&lt;0.05</td>
<td>0.097 (0.052)</td>
<td>0.06</td>
<td>0.074 (0.049)</td>
<td>0.13</td>
<td>0.275 (0.089)</td>
<td>&lt;0.05</td>
<td>-0.061 (0.048)</td>
<td>0.2</td>
<td>0.336 (0.07)</td>
<td>&lt;0.05</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>66</td>
<td>-0.353 (0.120)</td>
<td>&lt;0.05</td>
<td>0.125 (0.135)</td>
<td>0.36</td>
<td>-0.478 (0.138)</td>
<td>&lt;0.05</td>
<td>-0.022 (0.092)</td>
<td>0.81</td>
<td>0.090 (0.101)</td>
<td>0.37</td>
<td>-0.112 (0.099)</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>284</td>
<td>0.309 (0.071)</td>
<td>&lt;0.05</td>
<td>-0.198 (0.081)</td>
<td>&lt;0.05</td>
<td>0.507 (0.074)</td>
<td>&lt;0.05</td>
<td>0.259 (0.053)</td>
<td>&lt;0.05</td>
<td>0.154 (0.089)</td>
<td>0.08</td>
<td>0.105 (0.081)</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Stream</td>
<td>32</td>
<td>-0.179 (0.164)</td>
<td>0.2844</td>
<td>0.127 (0.167)</td>
<td>0.45</td>
<td>-0.306 (0.172)</td>
<td>0.08</td>
<td>-0.172 (0.221)</td>
<td>0.44</td>
<td>-0.267 (0.137)</td>
<td>0.06</td>
<td>0.095 (0.155)</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>Swamp</td>
<td>30</td>
<td>-0.539 (0.061)</td>
<td>&lt;0.05</td>
<td>-0.624 (0.185)</td>
<td>&lt;0.05</td>
<td>0.085 (0.144)</td>
<td>0.56</td>
<td>-0.149 (0.118)</td>
<td>0.22</td>
<td>-0.059 (0.111)</td>
<td>0.6</td>
<td>-0.09 (0.142)</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>Young</td>
<td>48</td>
<td>-0.609 (0.109)</td>
<td>&lt;0.05</td>
<td>0.473 (0.204)</td>
<td>&lt;0.05</td>
<td>-1.082 (0.122)</td>
<td>&lt;0.05</td>
<td>0.143 (0.104)</td>
<td>0.17</td>
<td>0.387 (0.131)</td>
<td>&lt;0.05</td>
<td>-0.244 (0.128)</td>
<td>0.06</td>
<td></td>
</tr>
</tbody>
</table>

The P values in the “Barcode” and “Phylomatic” columns were calculated using two-tailed t tests. Negative values indicate that the observed average NRI or NTI was phylogenetically overdispersed; positive values indicate that the observed average NRI or NTI score was phylogenetically clustered. The columns labeled “Difference” provide the mean of the difference between the Barcode and Phylomatic NRI and NTI values in each habitat; the P values were calculated using two-tailed paired t tests. All results presented in boldface in the table indicate that results based on the Phylomatic tree were significantly different than results based on the barcode tree.
Table S5. Primer pairs for barcode loci *rbcLa*, *matK*, and *trnH-psbA*

<table>
<thead>
<tr>
<th>Marker</th>
<th>Primer</th>
<th>Sequence (5′ → 3′)</th>
<th>Size, bp</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>rbcLa</em></td>
<td>SI_For</td>
<td>ATGTCACCACAAACACAGACTAAAGC</td>
<td>554</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SI_Rev</td>
<td>GTAAATCAAGTCCACCRCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>matK</em></td>
<td>KIM 3F</td>
<td>CGTACAGTACTTTTGTTTACGAG</td>
<td>Avg</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>KIM 1R</td>
<td>ACCCAGTCATCTGGAAATCTTGGTTC</td>
<td>850</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1329</td>
<td>TCTAGCAACAGAAAAGTGAAGT</td>
<td>Avg</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>320</td>
<td>CGATCTATTCAATCTAATTTCC</td>
<td>~880</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5R</td>
<td>GTTCAAGCAAAAGAAGTCG</td>
<td>Avg</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>XF</td>
<td>TAATTTAGATCAATTTCC</td>
<td>~900</td>
<td>5</td>
</tr>
<tr>
<td><em>trnH-psbA</em></td>
<td>psbA3′f</td>
<td>GTTATGCATGAACGTAAATGCTC</td>
<td>Avg</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>trnH</td>
<td>CGCGCATGTTGGATTCCAATCC</td>
<td>~450</td>
<td>7</td>
</tr>
</tbody>
</table>

1. New primers developed at the Smithsonian.
2. Ki-Joong Kim, School of Life Sciences and Biotechnology, Korea University, Seoul, Korea, unpublished primers.

Other Supporting Information Files

Dataset S1 (DOCX)