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, dilluted 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 at 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-gene 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-m^2 subplot (n = 1,250 subplots). The analyses were conducted by using the software Phylocom 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 queens 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-m^2 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.

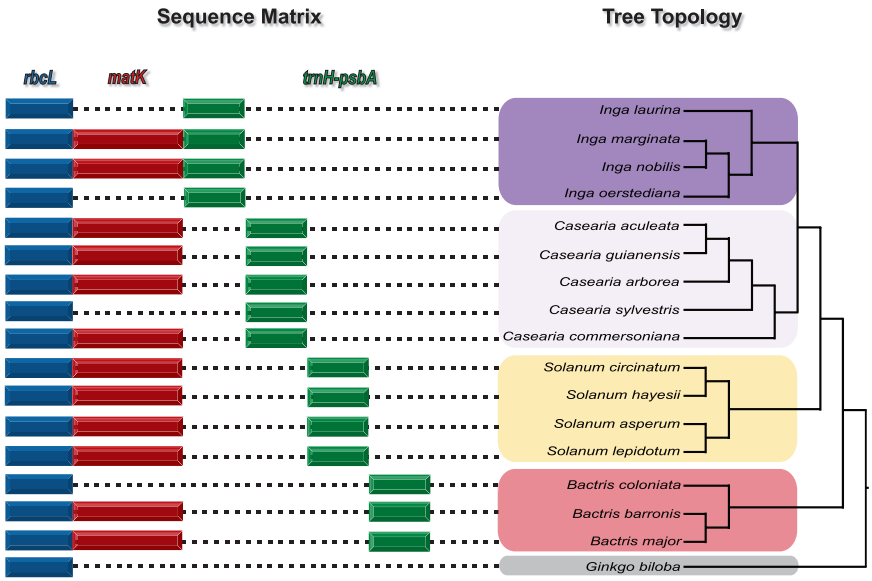

Fig. S1. Schematic representation of supermatrix approach in sequence alignment for phylogenetic analysis using exemplars from the BCI flora.

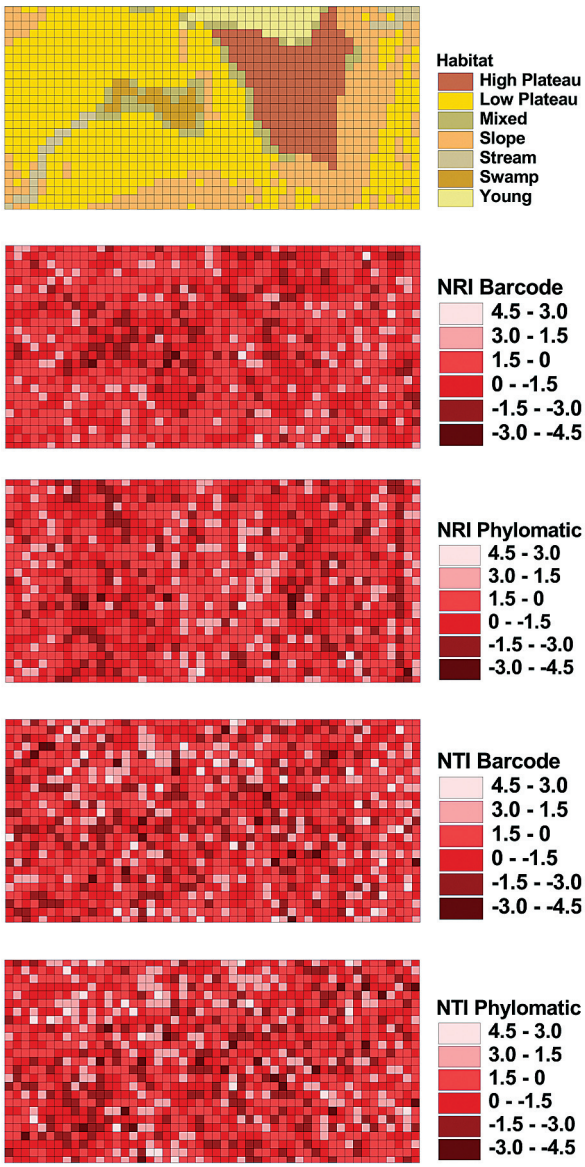

Fig. S2. The spatial distribution of the BCI habitats (41) and the NRI and NTI values in the 1,250 400-m^2 quadrants calculated by using the barcode phylogeny and the Phylomatic phylogeny. Negative NRI and NTI values indicate phylogenetic overdispersion and positive values indicate phylogenetic clustering. 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

	trnH-psbA		*rbcLa*		*matK*		*trnH-psbA* + *rbcL*		*matK* + *rbcL*		*matK* + *rbcL*+ *trnH-psbA*
	PCR	SEQ	PCR	SEQ	PCR	SEQ	Either SEQ	Both SEQ	Either SEQ	Both SEQ	All SEQ
No. of species (296)	290	280	277	277	252	205	291	279	279	205	204
%	0.97	0.94	0.93	0.93	0.85	0.69	0.98	0.94	0.94	0.69	0.69
No. of samples (1,035)	968	902	862	858	723	650	970	825	859	650	645
%	0.93	0.87	0.83	0.83	0.70	0.62	0.93	0.79	0.83	0.63	0.62
Order* (23)	—	23	—	22	—	22[†]	23	22	23	22	22
Family[‡] (57)	—	55	—	55	—	50	55	55	55	50	50
Genus[§] (181)	—	173	—	170	—	138	173	170	170	138	138

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

Taxonomic rank	Measure	Frequency of correct identification (BLAST), %		
		trnH-psbA	*rbcL*	*matK*
Species (296)	CI frequency	95	75	99
	Recovery × CI	90	70	69
Genus (181)	CI frequency	100	91	100
	Recovery × CI	95	85.5	69
Family (57)	CI frequency	100	100	100
	Recovery × CI	95	94	69

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

	Resolution							
	rbcLa + *matK*				*rbcLa* + *matK* + *trnH-psbA*			
Order (*n*)	<50%	>50% to 70%	>70% to 85%	>85%	<50%	>50% to70%	>70% to 85%	>85%
Fabales (35)	5	3	0	27	6	2	1	26
Rosales (25)	1	1	0	23	1	1	0	23
Malpighiales* (47)	0	10	0	37	1	6	1	39
Sapindales (26)	0	2	2	24	0	2	0	24
Malvales/Brassicales (14)	1	0	0	13	1	0	0	13
Myrtales† (27)	1	4	0	22	2	3	2	20
Gentianales (38)	1	2	0	35	1	1	0	36
Apiales/Solanales/ Boraginaceae (14)	2	2	0	10	2	2	0	10
Ericales/Carylphyllales (18)	0	1	1	16	0	1	1	16
Basal angiosperms‡ (37)	4	5	4	24	2	2	5	28
Total	15 (5.3%)	30 (10.7%)	5 (1.8%)	231 (82.2%)	16 (5.7%)	20 (7.1%)	10 (3.6%)	235 (83.6%)

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 Crossomatales.
‡Basal angiosperms comprised of Arecales, Laurales, Magnoliales, 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

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

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*

Marker	Primer	Sequence (5′ → 3′)	Size, bp	Ref.
rbcLa	SI_For	ATGTCACCACAAACAGAGACTAAAGC	554	1
	SI_Rev	GTAAAATCAAGTCCACCRCG		
matK	KIM 3F	CGTACAGTACTTTTGTGTTTACGAG	Avg	2
	KIM 1R	ACCCAGTCCATCTGGAAATCTTGGTTC	850	
	1329	TCTAGCACACGAAAGTCGAAGT	Avg	3
	320	CGATCTATTCATTCAATATTTC	≈880	
	5R	GTTCTAGCACAAGAAAGTCG	Avg	4
	XF	TAATTTACGATCAATTCATTC	≈900	5
trnH-psbA	psbA3′f	GTTATGCATGAACGTAATGCTC	Avg	6
	trnH	CGCGCATGGTGGATTCACAATCC	≈450	7

1. New primers developed at the Smithsonian.
2. Ki-Joong Kim, School of Life Sciences and Biotechnology, Korea University, Seoul, Korea, unpublished primers.
3. Cuenoud P, et al. (2002) Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. *Am J Bot* 89:132–144.
4. www.kew.org/barcoding/protocols.html.
5. Soltis DE, et al. (2001) Phylogenetic relationships and evolution in *Chrysosplenium* (Saxifragaceae) based on *matK* sequence data. *Am J Bot* 88:883–893.
6. Sang T, Crawford DJ, Stuessy TF (1997) Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *Am J Bot* 84:1120–1136.
7. Tate JA, Simpson BB (2003) Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. *Syst Bot* 28:723–737.

Other Supporting Information Files

Dataset S1 (PDF)