American Journal of Botany

American Journal of Botany 99(2): 320–329. 2012.

Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA[1]

Nolan Kane[2,5], Saemundur Sveinsson[2], Hannes Dempewolf[2], Ji Yong Yang[2], Dapeng Zhang[3], Johannes M. M. Engels[4], and Quentin Cronk[2]

[2]Department of Botany, University of British Columbia, Vancouver BC, Canada V6T 1Z4;
[3]SPCL, USDA-ARS 10300 Baltimore Avenue, Bldg. 001 Barc-West, Beltsville, Maryland 20705 USA;
and [4]Bioversity International 00057 Maccarese, Rome, Italy

- *Premise of study:* To reliably identify lineages below the species level such as subspecies or varieties, we propose an extension to DNA-barcoding using next-generation sequencing to produce whole organellar genomes and substantial nuclear ribosomal sequence. Because this method uses much longer versions of the traditional DNA-barcoding loci in the plastid and ribosomal DNA, we call our approach ultra-barcoding (UBC).
- *Methods:* We used high-throughput next-generation sequencing to scan the genome and generate reliable sequence of high copy number regions. Using this method, we examined whole plastid genomes as well as nearly 6000 bases of nuclear ribosomal DNA sequences for nine genotypes of *Theobroma cacao* and an individual of the related species *T. grandiflorum*, as well as an additional publicly available whole plastid genome of *T. cacao*.
- *Key results:* All individuals of *T. cacao* examined were uniquely distinguished, and evidence of reticulation and gene flow was observed. Sequence variation was observed in some of the canonical barcoding regions between species, but other regions of the chloroplast were more variable both within species and between species, as were ribosomal spacers. Furthermore, no single region provides the level of data available using the complete plastid genome and rDNA.
- *Conclusions:* Our data demonstrate that UBC is a viable, increasingly cost-effective approach for reliably distinguishing varieties and even individual genotypes of *T. cacao.* This approach shows great promise for applications where very closely related or interbreeding taxa must be distinguished.

Key words: cacao; chloroplast; DNA barcoding; evolution; genomics; *Gossypium*; Malvaceae; plastid; *Theobroma*; ultra-barcoding.

Reliable identification of variation below the species level would provide valuable insight into subspecies ranges, habitat differentiation, and patterns of gene flow and migration. Additionally, it would aid in conservation of important variation within species. For example, the conservation status of 2160 subspecies, varieties, or subpopulations of plants and animals has been evaluated by the International Union for Conservation of Nature (IUCN, 2006), with many high profile subspecies listed as endangered, threatened, or extinct. Reliable identification of taxa below the species level would also be important for fisheries and wildlife management, and in breeding efforts to improve crop and lifestock species. The use of DNA markers to identify subspecies has a long history (e.g., Morin et al., 1992; Ross et al., 1992; Bardakci and Skibinski, 1994), but is beyond the scope of current widely applied DNA-barcoding efforts by the Consortium for the Barcode of Life (CBOL, 2009).

To systematically examine variation below the species level, we propose to sequence and assemble whole organellar genomes and large (greater than 5 kb) contiguous portions of the nuclear genome. We call this ultra-barcoding (UBC; Kane and Cronk, 2008) for the simple reason that the approach expands the traditional barcoding regions to their full, many-kilobase length. This approach yields a tremendous amount of data for each locus, making it is far more sensitive than traditional DNA-barcoding (Parks et al., 2009; Nock et al., 2011; Steele and Pires, 2011) and thus may provide the necessary information to examine variation below the species level. The plastid is generally uniparentally inherited (Birky, 2001) and behaves as a single nonrecombining locus, providing a strong signal of population and phylogenetic history (Petit and Vendramin, 2007). As such, it is an ideal locus for DNA-barcoding in most taxa. However, evolution is generally so slow in the plastid genome that there is little variation per nucleotide (Palmer, 1985; Wolfe et al., 1987; Zurawski and Clegg, 1987). Thus, the amount of variation present over short regions may be too low to distinguish recently diverged taxa (e.g., Piredda et al., 2011). Because the plastid is present at many copies per nuclear

[1]Manuscript received 6 December 2011; revision accepted 19 December 2011.

This work was funded by a World Bank Development Marketplace grant awarded to Q.C. and J.E. We thank Chris Grassa, Lambert Motilal, and Loren Rieseberg for helpful discussions and advice; Brian Irish (USDA ARS) and Kyle Wallick (USBG) for providing leaf samples; Stephen Pinney (USDA ARS) for assistance with DNA extractions; and Jon Armstrong and Jarret Glasscock (Cofactor Genomics, St. Louis) for sequencing assistance. Q.C. also acknowledges laboratory support from NSERC.

[5]Author for correspondence (e-mail: nkane@biodiversity.ubc.ca)

doi:10.3732/ajb.1100570

genome, plastid genomes can be sequenced relatively inexpensively using low-coverage whole genome shotgun sequence generated by next-generation sequencing approaches (Dempewolf et al., 2010; Meyers and Liston, 2010; Nock et al., 2011; Straub et al., 2011, 2012). Such a low-pass scan of the entire genome focuses entirely on the high-copy number fraction of the genome (Steele and Pires, 2011; Straub et al., 2012), unlike whole genome sequencing and assembly using traditional approaches, which typically covers mainly the single-copy fraction of the genome.

The UBC approach requires far more sequencing than traditional barcoding, but the cost to generate this additional sequence has become increasingly affordable with the rise of next-generation sequencing, particularly short-read technologies such as the Illumina HiSeq. Whole chloroplast genomes and other ultra-barcodes can be inexpensively and accurately sequenced using next-generation approaches (Cronn et al., 2008; Parks et al., 2009; Dempewolf et al., 2010; Whittall et al., 2010; Doorduin et al., 2011; Steele and Pires, 2011; Straub et al., 2011; 2012 Zhang et al., 2011).

The main focus of this UBC approach in plants is the plastid, but nuclear sequence is generated as well and provides additional useful information. As with the chloroplast, the ribosomal DNA (rDNA) has the advantage of being multicopy, so requires far less coverage than single-copy nuclear genes. Additionally, although rDNA is multicopy, it does not have the same problems of paralogy as other multicopy nuclear loci such as transposons because the copies do not evolve independently, apparently due to biased gene conversion and unequal crossing over (Arnheim et al., 1980; Coen et al., 1982; Hillis et al., 1991; Wendel and Albert, 1992; Elder and Turner, 1995). Termed concerted evolution, because the paralogous copies are evolving in concert, this property means that rDNA gives strong phylogenetic and even population-level signals (Hillis and Davis, 1988). The downside is that if concerted evolution is incomplete or slow, there may be infraindividual allelic variation at rDNA loci (Schaal and Learn, 1988). However, within-individual variation can be dealt with to some extent by generating consensus sequences for each individual, as next-generation sequencing approaches are quantitative, providing sequences and copy-numbers for all multicopy rDNA alleles per individual (Ganley and Kobayashi, 2007).

We recognize that DNA barcoding is conventionally understood to mean delimitation at the species level. However, we choose to widen the scope to include varietal delimitation, and our use of the word hereafter should be understood as such, particularly when using the term ultra-barcoding or UBC. As an example of how UBC can be applied, we will here examine whole plastid genomes and ribosomal DNA in *Theobroma cacao* L. (Malvaceae), a tropical tree crop of great economic importance (Wood and Lass, 2001). Cocoa beans are an internationally traded commodity used in the manufacture of chocolate. The enormous range of genetic variation within the species (e.g., Motamayor et al., 2008) can be used to produce chocolate with distinctive flavor profiles. At present, relatively little chocolate is marketed with specific varietal information, although the use of varieties is increasing due to connoisseur interest in single variety and known provenance chocolates. However, a rise in interest in cacao varieties is likely to have positive impacts on the genetic diversity in the production systems. First, it is likely to enhance the product value chain, especially to the economic benefit of small producers. It is also likely to have positive impacts for conservation under the paradigm of "genetic resource conservation through use", because local varieties are more likely to be retained by farmers, rather than be replaced by standard varieties. To realize these benefits, however, reliable means of characterizing and identifying genetic variation in cacao are important.

Variation within cultivated cacao has traditionally been divided into three main varietal groups (Cheesman, 1944): Criollo, Forastero, and Trinitario. Criollo varieties are considered to give the best quality product, while Forastero varieties are considered to be generally more disease resistant and robust (International Cocoa Organization, 2011). Hybridization between imported Criollo and Forastero varieties in Trinidad and Tobago is the hypothesized origin of the so-called Trinitario varieties that combine the best characteristics of the parental lineages (Cheesman, 1944). Diverse cacao germplasm is very important for the plant breeders (Bartley, 2005), and there is a need for markers to adequately distinguish accessions (for instance in field germplasm banks), for which microsatellites have been generally used (Motilal et al., 2009; Irish et al., 2010).

Genomic resources for cacao are developing fast. As well as complete nuclear genome sequencing (Argout et al., 2010), there are also two publicly available complete cpDNA genome sequences, including our Scavina-6 genotype and a genotype of unknown provenance (Jansen et al., 2011). It is therefore now timely to investigate the feasibility of sequence-based approaches in varietal identification, using low-pass whole genome scans.

For these reasons, *T. cacao* provides an ideal test case for the ultra-barcoding concept. While previous researchers have very successfully sequenced individual whole chloroplast genomes (Meyers and Liston, 2010; Straub et al., 2011) or even chloroplasts and ribosomal DNA (Nock et al., 2011), or used pooled chloroplast DNA samples to assess variation within species (Doorduin et al., 2011), the utility of generating this kind of data for multiple nonpooled individuals within a species has not been tested. In this paper, we use a large data set of whole genome scans of 10 individuals, in addition to publicly available resources to assess whole-organellar and whole-ribosomal variation both within *T. cacao* accessions and between *T. cacao* and a closely related congener.

MATERIALS AND METHODS

***Sample selection and DNA preparation*—**Samples for sequencing were selected to encompass a wide range of cacao variation (Table 1, Appendix S1; see Supplemental Data with the online version of this article) but focusing on the Trinitario varietal group, which has a complex hybrid origin involving multiple parental sources of Criollo and Forastero. Leaf material was obtained from accessions held in USDA collections at Beltsville, Maryland, and at the Tropical Agricultural Research Station (TARS) in Puerto Rico (Irish et al., 2010).

***Illumina sequencing*—**Total genomic DNA was extracted from fresh leaf tissue using the DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA) according to the manufacturer's protocol. Illumina sequencing libraries were prepared using standard chemistry and protocols and one lane of paired-end 60- or 80-bp sequence was generated for each sample on Illumina GAII machines by Cofactor Genomics of St. Louis (http://www.cofactorgenomics.com/).

***De novo assembly of chloroplast and rDNA*—**De novo assembly of reference *T. cacao* chloroplast and rDNA was accomplished using approaches described in Dempewolf et al. (2010), for the Scavina-6 genotype. Briefly, reads were trimmed and cleaned, and assembled using the ABySS (Simpson et al., 2009) and SOAPdenovo (http://soap.genomics.org.cn/soapdenovo.html) assemblers. Chloroplast-encoded contigs were identified using NCBI program blastn (Camacho et al., 2009), with a minimum e-value of 10^{-20} and minimum sequence identity of 85% to *Gossypium hirsutum* chloroplast genome (Lee et al., 2006),

TABLE 1. Provenances of *Theobroma* material sequenced.

Name	Source of material	Notes	Accession
EET-64 (*T. cacao*)	USDA (TARS), Puerto Rico	Hybrid between Upper Amazon Forastero (Nacional from Ecuador) and Trinitario (from Venezuela).	PI 275669
Criollo-22	USDA (SPCL), Beltsville, MD	Pure Criollo variety	Criollo-22
Stahel (*T. cacao*)	USDA (TARS), Puerto Rico	Trinitario with similarities to lower Amazon Forastero	MIA 27956
Pentagonum (*T. cacao*)	USDA (TARS), Puerto Rico	Trinitario (Criollo-type)	TARS 12044
Scavina-6 (*T. cacao*)	USDA (SPCL), Beltsville, MD	Upper Amazon Forastero, Peru	MIA 29885
Amelonado (*T. cacao*)	USDA (TARS), Puerto Rico	Lower Upper Amazon Forastero	TARS 16542
ICS39 (*T. cacao*)	USDA (TARS), Puerto Rico	Trinitario	TARS 16664
ICS06 (*T. cacao*)	USDA (TARS), Puerto Rico	Trinitario	TARS 16658
ICS01 (*T. cacao*)	USDA (TARS), Puerto Rico	Trinitario	TARS 16656
*T. grandiflorum (*Cupuaçu)	USDA (TARS), Puerto Rico	Species related to *T. cacao*. Wild and cultivated in Amazon Basin	04-0254

Notes: MD: Maryland, USA; TARS, Tropical Agriculture Research Station; SPCL, Sustainable Perennial Crops Laboratory; USDA, U. S. Department of Agriculture

which was the most closely related published full chloroplast genome at the time of the analysis. The most complete chloroplast assemblies resulted from SOAPdenovo, but both assemblies were aligned to the published chloroplast genome of *G. hirsutum* (Lee et al., 2006). Adjacent contigs were merged when overlap was greater than 15 bp, and shorter overlaps and small gaps were filled using custom perl scripts with trimmed Illumina read data as in Dempewolf et al. (2010). Quality of the final assembly was confirmed and error-corrected by aligning the quality-trimmed Illumina reads using the program MOSAIK (see below) for each genotype. Similarly, to generate a reference rDNA, assemblies were blasted against the existing rDNA sequence for *T. cacao* (GI:7595560, GI:27447189, GI:133854413) and related species (GI:19919576–19919578; Soltis et al., 2003). Adjacent contigs were again merged when overlap was greater than 15 bp, and gap filling and extension was performed where possible using the unassembled Illumina reads.

The full-length chloroplast sequence was annotated using DOGMA (Dual Organellar GenoMe Annotator; Wyman et al., 2004), with additional information about splice sites and open reading frames provided from comparisons with another completely sequenced *T. cacao* plastid genome (Jansen et al., 2011) as well as the fully annotated *Gossypium* chloroplast genomes (Ibrahim et al., 2006; Lee et al., 2006). The completed annotation was illustrated using OGDraw (Organellar Genome Draw; Lohse et al., 2007).

***SNP genotyping for each sample*—**Trimmed, cleaned paired-end reads were mapped to the reference *T. cacao* chloroplast and rDNA sequence using MOSAIK (Hillier et al., 2008), with a hash size of 12, 12 mismatches allowed, and an ACT score of 35. Sorted alignment files were converted to BAM format and single nucleotide polymorphisms (SNPs) were called for each sample against the reference using the program SAMTOOLS (Li et al., 2009), with a minimum quality of 20 for SNPs and 50 for insertions and deletions. Additionally, alignments were manually checked and confirmed with the raw reads for all significant differences from the reference.

***Sequence accession numbers*—**All of our WGS Illumina/Solexa reads are available from NCBI's Short Read Archive (accession SRA048198.1), with our annotated whole plastid genome assemblies also uploaded (GenBank accession HQ244500 for the de novo assembly, and GenBank accessions JQ228379–JQ228389 for the reference-guided assemblies). The rDNA sequences are GenBank accession JQ228369–JQ228378.

***Phylogenetic analysis*—**An alignment of the rDNA and the plastid sequences (our 10 samples as well as an additional sample of *T. cacao,* GI 328924764), which included all SNPs observed in the data nset was made using the program MUSCLE (Edgar, 2004) and used for phylogenetic analyses. A phylogeny for the plastid data was analyzed under maximum likelihood (ML; Felsenstein, 1973) using the program Garli 2.0 (Zwickl, 2006). The TIM base substitution model was chosen for the ML analysis, based on model testing with the program jModeltest 0.1.1 (Guindon and Gascuel, 2003; Posada, 2008). One thousand bootstrap replicates were performed to obtain support values for the phylogeny. The majority rule consensus tree was generated and tree drawn and rooted using the program Figtree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/). The rDNA data set was analyzed with a network-based approach, as rDNA undergoes recombination and is therefore likely to violate the assumption of evolving in a bifurcating manner. The program SplitsTree4 (Huson and Bryant, 2006) was used to visualize the relationships among the *T. cacao* varieties and to *T. grandiflorum* and were displayed as a phylogenetic network. Support values at each node in the network were estimated by running 1000 bootstrap replicates.

***Comparisons with Gossypium*—**Overall structural similarities between the full-length plastid genomes of *Theobroma* and *Gossypium* (Lee et al., 2006; Ibrahim et al., 2006) were examined aligning the two genomes with the blastz algorithm (Schwartz et al., 2003) and illustrated using the program zPicture (Ovcharenko et al., 2004).

RESULTS

***Illumina sequencing*—**We obtained between 1.7- and 4.6-Gbp sequence per sample after removing low-quality reads,

TABLE 2. Illumina sequence summary statistics and observed average coverage of the nuclear and chloroplast genome for *Theobroma* based on Burrows-Wheeler Aligner (BWA) alignments (see text).

Name	Read length (bp)	No. of pairs of reads	No. of pairs after filtering	Total bp	Nuclear genome coverage	Chloroplast genome coverage	rDNA coverage
EET-64 (*T. cacao*)	60	3.3E+07	3.2E+07	3.9E+09	9.36685	871	2669.9
Criollo-22 (*T. cacao*)	60	2.3E+07	2.1E+07	2.5E+09	6.08258	850.1	900.1
Stahel (*T. cacao*)	60	3.0E+07	2.8E+07	3.4E+09	8.17719	1229.1	2377.6
Pentagonum (*T. cacao*)	80	2.9E+07	2.6E+07	4.1E+09	9.87637	1539.8	4437.4
Scavina-6 (*T. cacao*)	60	1.9E+07	1.4E+07	1.7E+09	4.18122	186.9	783.9
Amelonado (*T. cacao*)	60	3.5E+07	3.4E+07	4.1E+09	9.89147	1291.4	3945.4
ICS39 (*T. cacao*)	80	3.2E+07	2.8E+07	4.5E+09	10.7130	1117.6	2735
ICS06 (*T. cacao*)	80	3.2E+07	2.8E+07	4.6E+09	10.9555	1324.7	3269.8
ICS01 (*T. cacao*)	60	2.5E+07	2.4E+07	2.9E+09	7.01850	944.6	3043.6
T. grandiflorum (Cupuaçu)	60	3.6E+07	3.4E+07	4.1E+09	9.86942	772.5	2662.2

or 4.2–11× coverage of the nuclear genome per sample based on an estimated genome size of 416 Mbp (Figueira et al., 1992). This was more than adequate coverage to assemble both plastid and rDNA: 187–1540× average coverage of the plastid and 784–3946× average coverage of the rDNA (Table 2).

De novo assembly of chloroplast and rDNA—Our de novo chloroplast assembly covered the entire 160 546-bp circular plastid genome (NC_014676, Fig. 1). Sanger sequence of 2162 bp of the plastid genome confirmed the assembly in nine regions (GI: 338190271–338190279), and also confirmed the length of repeats in microsatellites (Yang et al., 2011). The de novo rDNA assembly spanned the entire expressed portion including ETS, 18S, ITS1, 5.8S, ITS2, and 28S (Fig. 2), for a total of 5826 bp.

SNP genotyping for each sample—Numerous SNPs were found both within and between *Theobroma* species (Figs. 1, 2; Tables 3, 4), with far more variation between than within species for most regions. Three SNPs were confirmed in 95 individuals using Sanger sequence. Each of the examined canonical barcoding regions showed substantial variation between *T. cacao* and *T. grandiflorum* (Table 3), but none had enough variation to distinguish any of the major varieties of *T. cacao*, with only *rbcL* showing any variation and that only at one nucleotide. The most variable 500-bp region of the chloroplast within *T. cacao* was in the 3′ end of the *ccsA* gene, with four SNPs within *T. cacao* and two SNPs between *T. cacao* and *T. grandiflorum*. When using the entire plastid genome, however, we observed orders of magnitude more vatiation: 78 SNPs segregating within *T. cacao* and 415 SNPs observed between species. Indeed, there were multiple high-quality SNPs uniquely distinguishing each sample sequenced with the exception of EET and Criollo-22, which were identical. The rDNA showed similarly high variation, particularly within ITS1, ITS2 and the ETS (Table 4). Again, however, the variation when including the entire rDNA sequence was far greater than any one region, enabling unique identification of each individual sequenced, this time including EET and Criollo-22, which differed at several sites.

Phylogenetic analysis—The phylogenetic analyses revealed significant divergence between the major clades of *T. cacao*. The ML tree inferred by Garli (Fig. 3) showed two strongly supported clades within *T. cacao* corresponding to the origin of the Forastero and Criollo varieties. Furthermore, the tree shows that maternal lineages of Trinitario samples come from both

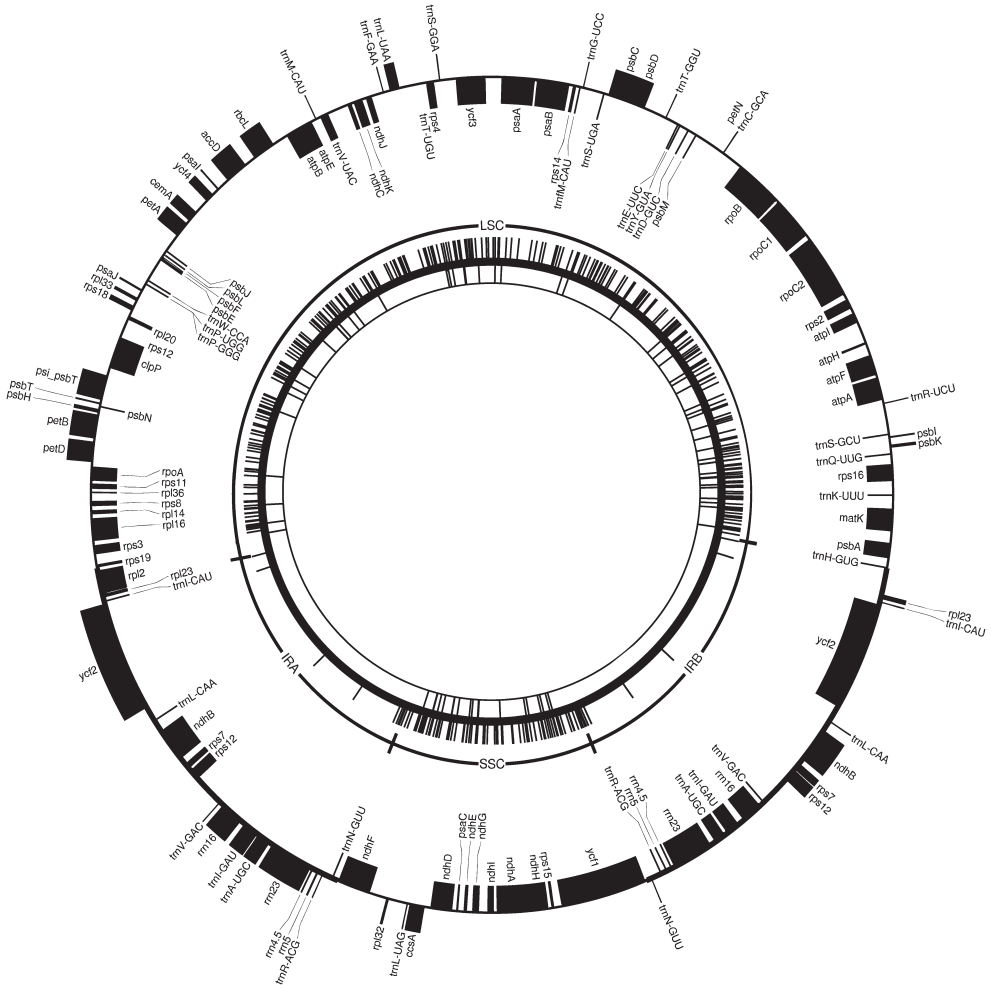

Fig. 1. Map of the annotated circular chloroplast for *Theobroma cacao* (outer circle). Inner circle: SNPs segregating within *T. cacao*. Middle circle: SNPs fixed between *T. cacao* and *T. grandiflorum*.

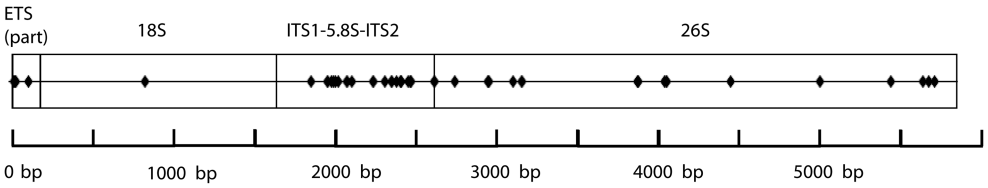

Fig. 2. Schematic diagram of the *Theobroma cacao* ribosomal DNA, including ETS, 18S, ITS1, 5.8S, ITS2, and 28S, with SNPs indicated by diamonds.

Forastero and Criollo backgrounds. Additional clades are also well supported within the two main branches of *T. cacao* chloroplast genomes. The *T. cacao* sample sequenced by Jansen et al. (2011), which has no publicly available provenance information, groups strongly with the Forastero cpDNA types, so is most likely Forastero or Trinitario.

The rDNA SplitsTree4 network showed similar patterns to the phylogenetic tree, with Forastero and Criollo clustering into distinct clades, and Trinitario intermediate (Fig. 4). Evidence of substantial gene flow can be observed between the clades as well, particularly in the Trinitario samples.

***Comparisons with Gossypium*—**Over 98% of the *Gossypium* and *Theobroma* genomes aligned well, with 94% of the bases identical overall. The only major differences between the genomes is a large insertion between *ycf3* and *trnF*. The other large regions of nonalignment include the intergenic region between *ndhC* and *trnV*, the longest at nearly 1 kb, and numerous smaller, mainly intergenic regions. In contrast, the inverted repeat regions are much more similar (99% identical) over almost 24 kb.

DISCUSSION

***Advantages of ultra-barcoding*—**Taxon identification by means of DNA sequencing (commonly called DNA barcoding) uses hundreds of base pairs of sequence from an organellar genome (mitochondria in animals, plastids in plants) or nuclear locus ITS (widely used in fungi; Bellemain et al., 2010). The standard approach in animals involves sequencing a 650-bp portion of the mitochondrial cytochrome oxidase I (*COI*), a region that contains substantial variation but highly conserved primer sequences (Hebert et al., 2003). This makes DNA barcoding relatively straightforward in most animal taxa, but progress was initially stymied in plants because of the difficulty of identifying a high-information region with conserved flanking sequences appropriate for primer design (CBOL, 2009).

Several alternative chloroplast-encoded barcoding regions were proposed for plants (Pennisi, 2007), including portions of several protein-coding genes (*matK*, *rbcL*, *rpoB*, and *rpoC1*) and intergenic spacers (*atpF-atpH*, *trnH-psbA*, and *psbK-psbI*). Recently, a two-locus combination of *rbcL* and *matK* was chosen by the Consortium for the Barcode of Life based on a number of factors including ease of amplification and sequencing and the ability to discriminate between closely related species (CBOL, 2009). The reliance on coding regions rather than introns or intergenic spacers helps to avoid problems with sequence quality and alignment. However, these coding regions have lower variation (and thus lower ability to discriminate species) than some intergenic regions. Moreover, difficulties remain even when using these established regions, for instance due to difficulties in amplifying *matK* (Kress and Erickson, 2008; CBOL, 2009) likely due to secondary structure, and this has so far only been partially resolved by optimizing PCR primers and protocols (CBOL, 2009).

Despite these difficulties, DNA barcoding has been quite successful for a number of applications. Barcoding has been shown to be a useful tool for identifying species in biodiversity hot-spots (Lahaye, 2008), identifying material from species listed under the Convention on International Trade of Endangered Species (CITES) appendixes for conservation or regulatory purposes (Lahaye, 2008; Dexter, et al., 2010; Yesson, et al., 2011), revealing cryptic species (Herbert et al., 2003; Lahaye, 2008; Ragupathy et al., 2009; Newmaster and Ragupathy, 2010), vouchering specimens (Dugan et al., 2007), delimiting species (Muellner et al., 2009), confirming endangered species identification for reintroduction programs (Rowntree et al., 2010), and identifying invasive species (Bleeker et al., 2007). Ethnobotanical research can be greatly aided by this approach to species identification as well (Ragupathy et al., 2009; Newmaster and Ragupathy, 2010). Important applications in forensic botany (Ferri et al., 2009; Ward et al., 2009), medicine (Howard, 2010; Lou et al., 2010), and consumer protection (Kress and Erickson, 2008) have also been proposed. These and other potential applications make it clear that DNA barcoding is a useful tool.

Here we assess a related approach using much larger versions of the traditional DNA-barcoding loci to examine within-species variation. For the identification of haplotype groups within a species, such as for the characterization of crop plant diversity (e.g., Coart et al., 2006), a large amount of sequence may be advantageous. The total length of the plastid genome provides an upper limit for the amount of sequence that can be interrogated from this genomic component, and thus the maximum amount of data that can be obtained for that locus.

The results reported here demonstrate the benefits of ultra-barcoding for application below the species level. While traditional barcoding often struggles to reliably differentiate species, our plastid and rDNA data distinguish every individual cacao plant sequenced, each with its own unique SNP pattern. Furthermore, there was much higher divergence between species

TABLE 3. Variation in single nucleotide polymorphisms (SNPs) within *Thoebroma cacao* (column 2) and fixed SNP differences between *T. cacao* and *T. grandiflorum* (column 3) in four plastid regions.

Region	Within *T. cacao*	Between *Theobroma* species
psbA-trnH	0	7
rbcL	1	4
matK	0	1
ccsA	4	2
Entire plastid genome	78	415

TABLE 4. Variation within *Theobroma cacao* and fixed differences between *T. cacao* and *T. grandiflorum* in six rDNA regions.

Region	Within species	Between species
ETS	3	2
18S	1	0
ITS1	5	6
5.8S	1	1
ITS2	6	5
28S	3	14
Entire rDNA	19	27

of *Theobroma*, with several hundred bases differentiating species across the entire plastid genome for any interspecific comparison, while there were none or a handful of differences in the smaller barcoding regions. Similarly, sequencing of the majority of the chloroplast genome in 17 individuals of *Jacobaea vulgaris* showed that even with low levels of chloroplast sequence variation per base, the entire chloroplast provides the potential for a very substantial data set even using pooled samples (Doorduin et al., 2011). That study identified somewhat fewer segregating SNPs (32 in total) within the species, although concluded that this may be an underestimate due to low coverage of many regions of the genome. UBC also reveals a considerable number of plastid microsatellite loci (described in Yang et al., 2011), which may be a useful tool to examine large numbers of individuals.

Plastid haplotype data, due to their nonrecombining nature, are particularly powerful in distinguishing phylogeographic patterns (Soltis et al., 1997; Cavers et al., 2003), such as those evident in crop wild ancestors and traditional landraces. In our samples, for example, the fact that Criollo-22 and EET-64 have 100% identical plastid genomes suggests a very recent common ancestor, presumably a maternal grandmother or great-grandmother. Records show that one parent of EET-64 is a Venezuelan Trinitario, 'Venezuelan Amarillo', which is a hybrid between Criollo and Lower Amazon Forastero. Our results characterize the maternal lineage of that hybridization event and confirm that some of the Ecuadorian Arriba cacao have a Criollo maternal pedigree, which was inherited through the Venezuelan Amarillo.

Broader phylogeographic studies in plants may be limited by the lack of variation in standard plastid loci (Schaal et al., 1998). The short sequences used as standard barcoding loci mean that phylogeography is generally considered to be beyond the scope of traditional DNA barcoding. This impediment is largely removed by ultra-barcoding. Moreover, where hybridization is rife, such as between crop lineages or between crops and their wild relatives, a haplotype approach only tells part of the story, so the ability of ultra-barcoding to provide high-resolution markers for tracking maternal inheritance as well as nuclear loci subject to biparental inheritance can be of considerable assistance in plant breeding and diversity studies. As sequencing costs fall, it is likely that Sanger-based sequencing of a few plastid loci will be increasingly replaced by routine whole-plastid sequencing of multiple individuals.

Another advantage of ultra-barcoding is the generation of both nuclear as well as plastid data. The low-depth total genomic DNA scans that generate whole plastid data also reliably sequence high copy-number regions of the nuclear genome, such as the ribosomal repeat region. Two subregions of the ribosomal repeat, the internal transcribed spacers (ITS1 and ITS2), have been a reliable workhorse for species identification and phylogenetics (Yao et al., 2010). However, these spacers are often limited by their short length and low variation. Ultra-barcoding readily generates much longer ribosomal repeat region sequences, including both highly conserved and highly variable regions, considerably expanding the number of nuclear polymorphisms detected in our analysis. Unfortunately the

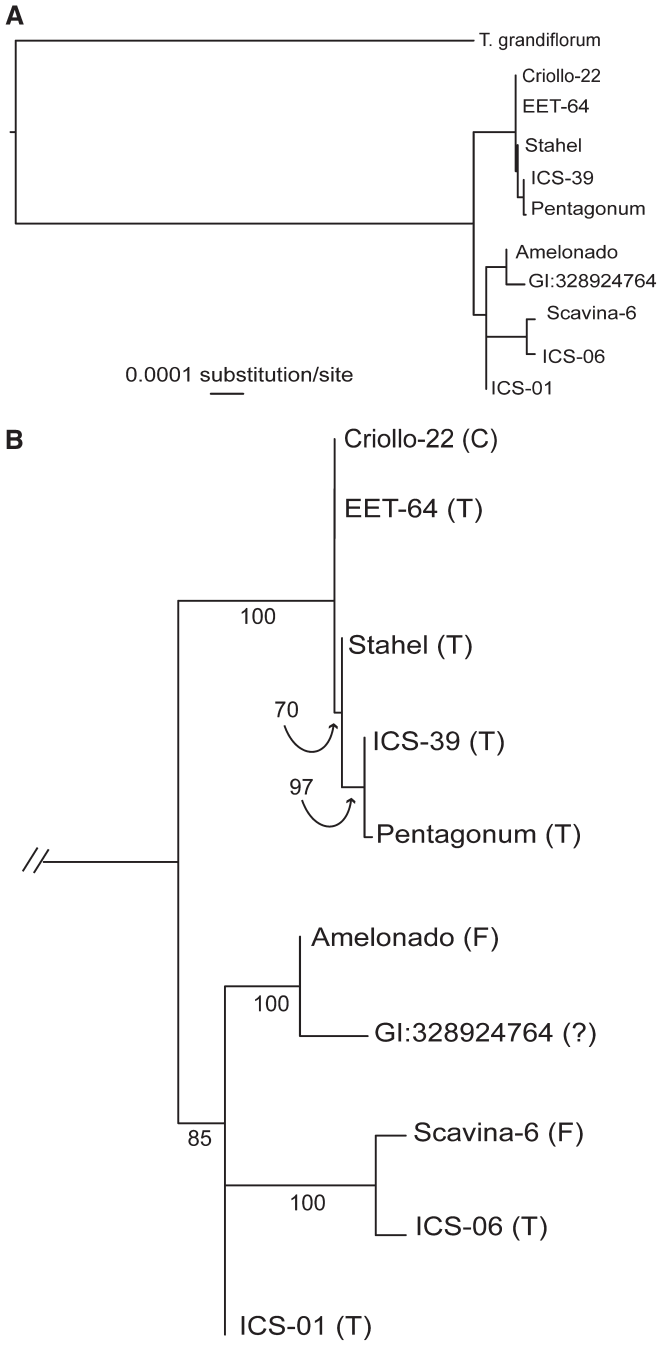

Fig. 3. Phylogenetic tree representing relationships of sequenced *Theobroma cacao* genotypes and *T. grandiflorum,* based on complete plastid genomes for each individual (−lnL = 221 170.271). Sample names are appended with variety information: (C) Criollo, (F) Forastero, (T) Trinitario.

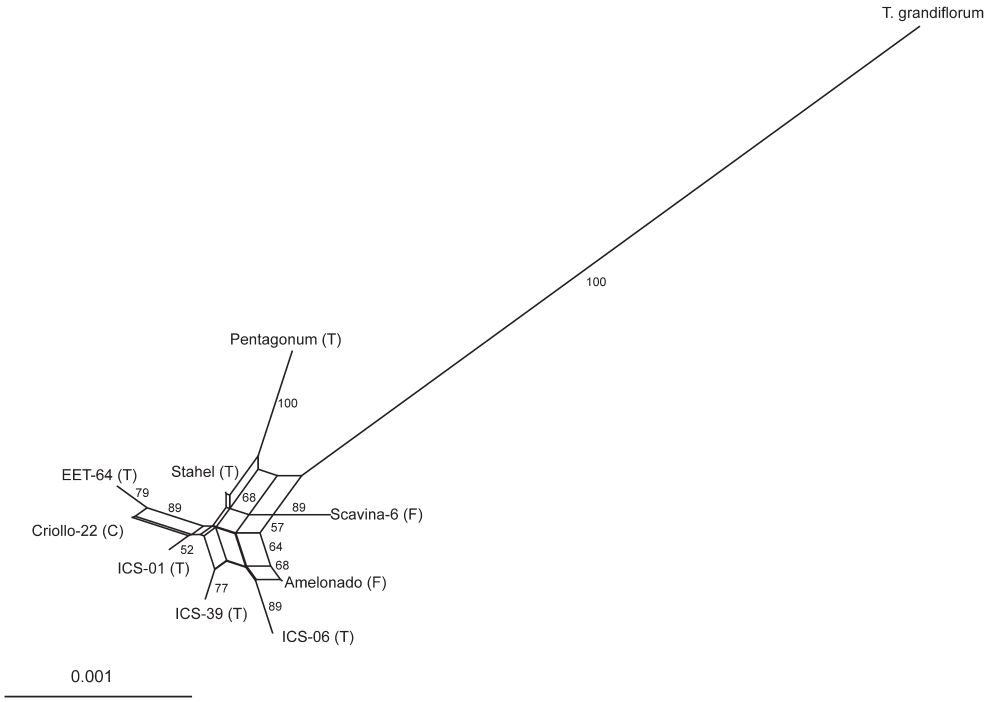

Fig. 4. SplitsTree4 split decomposition network of the rDNA consensus sequence for each individual sequenced. Only bootstrap values above 50 are shown. Sample names are appended with variety information: (C) Criollo, (F) Forastero, (T) Trinitario.

most variable region of the ribosomal repeat unit, the intergenic spacer (IGS), is presently proving difficult to work with. This is likely to be due to the rapid evolutionary rate of this region and the existence of multiple variants, particularly length polymorphisms (Ambrose and Crease, 2011). The variation within individuals is to be expected in rapidly evolving regions as there will be insufficient time for concerted evolution to homogenize copies after they arise. However, even excluding this region, there is ample SNP polymorphism to examine the comparative evolutionary history of nucleus and plastid.

***Practical barcoding applications in cacao*—**Given the results reported here, it is possible to ask how well a whole-genome scan technique works for practical applications and also to assess the prospects for future work. There is considerable need to identify and characterize cacao germplasm for numerous purposes including breeding work, industrial applications and efficient genetic resource conservation. Accurate varietal identification can help to support the "value chain", from consumer to grower, with the aim that local growers obtain increased economic benefits. Similar considerations are also arising from increasing niche marketing of other crops, such as coffee (Fitter and Kaplinsky, 2001). The use of genetic technologies to support the value chain is important in other bioproduct industries such as the meat industry (Sosnicki and Newman, 2010).

There are also expected conservation benefits from increased growing of local varieties when supported by accurate varietal identification. This "conservation-through-use" paradigm (Coomes, 2004; Newton, 2008) is an important aspect of genetic resource conservation for plants of economic value, besides the value of higher genetic diversity in the production system to increase resilience, e.g., in times of change.

Our data indicate that low-pass genome scans can identify varieties even when a recent ancestor is shared. The plastid SNP variation is sufficient to uniquely identify nearly every individual examined, as well as showing clear distinctions between the major varieties. The chloroplast genomes showed a clear phylogenetic signal as would be expected of a nonrecombining region (Fig. 3). Interestingly, there is good support for a separation of the Criollo and Forastero varieties, with the Trinitario varieties being scattered through the tree, again, as would be expected from a mixed group of hybrids (Fig. 3).

The data from the rDNA is sufficient to uniquely identify each accession. With these data, however, a less strong hierarchical structure is recovered as is to be expected from a recombining nuclear region. We therefore chose to represent this by a network graph (Fig. 4). Remarkably, though, there is also a suggestion of a separation between Criollo and Forastero varieties. In Fig. 4, Criollo and Forastero varieties are widely separated on the graph with Trinitario varieties in between, consistent with their hybrid origin.

Previously, nuclear microsatellites have been used with success for germplasm identification in cacao (Irish et al., 2010; Motilal et al., 2009). Sequence data offer some advantages as an alternative. The lower mutation rate of SNPs means that they are less prone to recurrent mutation, back-mutation and consequent homoplasy. SNPs are also preferable for measuring population genetic parameters (Kronholm et al., 2010; Whitlock 2011). Furthermore, the ability to simultaneously gather information on variation in two genomes (plastid and nuclear) offers great advantages in interpreting patterns of ancestry and parentage for breeders.

***The ribosomal DNA information gap*—**Repetitive sequences in general are given short shrift in full-length genomes, with

full-length rDNA repeats not present in most publically available "fully sequenced" genomes. For instance, neither of the two released *T. cacao* genomes (GenBank CACC01000001–CACC01025912, Argout et al., 2010, and http://www.cacaogenomedb.org/) has even a single full rDNA repeat, but rather only very small portions on several chromosomes. This is likely due to difficulty of assembly due to polymorphism among copies (Straub et al., 2011) as well as a focus on the single-copy portion of the genome. The question remains as to whether the IGS is a treasure trove of information waiting to be tapped or whether the amount of infra-individual variation is so high (e.g., Chou et al., 1999) that the problems of assembly, alignment, and analysis are intractable (Rogers and Bendich, 1987). These problems with the IGS are all potentially solvable, with better algorithms, better technology, and better approaches. For the IGS, longer reads such as 454's new chemistry may resolve some of these problems. For the moment, however, the variation present at the IGS is tantalizingly out of reach for most taxa. Similarly, while whole plastid genomes are available for 126 angiosperm species and subspecies as of June 2011, with more available nearly weekly, only 22 angiosperm mitochondria are available, most being clustered within a handful of families with nearly half (45%) being from the Poaceae alone. As more genome projects involve major evolutionary components, we hope that the gaps in rDNA and mitochondrial sequence will be filled.

Technology and cost—Cost per megabase of data is dropping rapidly. This is particularly true for the new HiSeq chemistry, where reagent costs are less than US$0.04/Mb, although many other technologies are close to this value (Glenn, 2011). With expected yield of close to 40 Gb/lane (Glenn, 2011), however, the standard 12 Illumina barcodes (http://www.illumina.com) are far too few, yielding much more coverage than necessary for an ultra-barcoding approach in most species. However, with the use of 48 barcodes, a 0.5-1Gb genome would have ~1× coverage, more than enough for UBC (Straub et al., 2012). Our hope is that major work is given to low-cost chemistry and protocols for library preparation. A combination of improved and more affordable robotics and low-cost reliable chemistry for library preparation are the key to making these techniques cost-effective for the widest-possible range of applications. Manufacturers are working in this direction, with 24-plex indexing Illumina DNA sample prep kits available soon (http://www.illumina.com/support/faqs.ilmn), and 48-level indexing validated primers already available from other vendors (e.g., http://www.biooscientific.com). With current commercial costs of 48-plex indexing and library preparation reagents below US$42/sample (e.g., http://www.biooscientific.com), and roughly US$1500 in sequencing costs, a lane of 48 individuals could be run for ~US$3500 (US$73/sample) in reagent costs. While still more expensive than a single Sanger sequence per individual (typically only a few dollars per sequence, or less for some at-cost facilities), the vast increase in data quantity and ability to distinguish lineages within species may make the extra expense of UBC worthwhile for many applications even at today's prices. As next-generation sequencing prices continue to decline, it is clear that the UBC approach will become increasingly attractive.

The future of DNA barcoding—UBC is a promising approach, resolving many of the difficulties with current widely applied DNA-barcoding approaches. This approach has numerous advantages, including higher resolution than smaller regions, and universality, as the need for taxon-specific primers is avoided. Nuclear loci such as ribosomal DNA provide substantial additional information that makes this approach even more powerful. Additional loci such as the mitochondrial genome, nuclear microsatellites, LTRs, TEs, and other repeats could also be assayed with this approach. For numerous breeding, medicinal, consumer protection, and forensic botany applications, it would be highly useful to identify samples to below the species level. Additionally, numerous questions related to phylogeography, population genetics, phylogenetics, and molecular evolution would benefit greatly by this type of data set. Finally, we should emphasize again that UBC does not remove the need for continued use of traditional barcoding methods currently in use, but rather provides data below the species level. The main stumbling blocks for UBC are cost and the somewhat higher quality and quantity of DNA required, as well as the bioinformatic and computational resources needed to deal with large amounts of next-generation sequence data. Nevertheless, as technology and methodologies continue to improve rapidly, we believe that it will soon be practical to sequence and assemble whole plastid genomes for a broad range of applications.

LITERATURE CITED

AMBROSE, C. D., AND T. J. CREASE. 2011. Evolution of the nuclear ribosomal DNA intergenic spacer in four species of the Daphnia pulex complex. *BMC Genetics* 12: 13.

ARGOUT, X., J. SALSE, J. M. AURY, M. J. GUILTINAN, G. DROC, J. GOUZY, ET AL. 2010. The genome of Theobroma cacao. *Nature Genetics* 43: 101–108.

ARNHEIM, N., M. KRYSTAL, R. SCHMICKEL, G. WILSON, O. RYDER, AND E. ZIMMER. 1980. Molecular evidence for genetic exchanges among ribosomal genes on non-homologous chromosomes in man and apes. *Proceedings of the National Academy of Sciences, USA* 77: 7323–7327.

BARDAKCI, F., AND D. O. F. SKIBINSKI. 1994. Application of the RAPD technique in tilapia fish: Species and subspecies identification. *Heredity* 73: 117–123.

BARTLEY, B. G. D. 2005. The genetic diversity of cacao and its utilization. CABI Publishing, Wallingford, UK.

BELLEMAIN, E., T. CARLSEN, C. BROCHMANN, E. COISSAC, P. TABERLET, AND H. KAUSERUD. 2010. ITS as an environmental DNA barcode for fungi: An in silico approach reveals potential PCR biases. *BMC Microbiology* 10: 189.

BIRKY, W. C. 2001. The inheritance of genes in mitochondria and chloroplasts: Laws, mechanisms and models. *Annual Review of Genetics* 35: 125–148.

BLEEKER, W., S. KLAUSMEYER, M. PEINTINGER, AND M. DIENST. 2007. Chloroplast DNA variations of cultivated radish and its wild relatives. *Plant Science* 168: 627–634.

CAMACHO, C., G. COULOURIS, V. AVAGYAN, N. MA, J. PAPADOPOULOS, K. BEALER, AND T. L. MADDEN. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10: 421.

CAVERS, S., C. NAVARRO, AND A. J. LOWE. 2003. Chloroplast DNA phylogeography reveals colonization history of a Neotropical tree, Cedrela odorata L., in Mesoamerica. *Molecular Ecology* 12: 1451–1460.

CBOL [Consortion of Barcode of Life]. 2009. A DNA barcode for land plants. Proceedings of the National Academy of Sciences, USA 106: 12794–12797.

CHEESMAN, E. E. 1944. Notes on the nomenclature, classification and possible relationships of cocoa populations. *Tropical Agricure* 21: 144–159.

CHOU, C. H., Y. C. CHIANG, AND T. Y. CHIANG. 1999. Within- and between-individual length heterogeneity of the rDNA-IGS in Miscanthus sinensis var. glaber (Poaceae): Phylogenetic analyses. *Genome* 42: 1088–1093.

COART, E., S. VAN GLABEKE, M. DE LOOSE, A. S. LARSEN, AND I. ROLDAN-RUIZ. 2006. Chloroplast diversity in the genus Malus: New insights into the relationship between the European wild apple (Malus sylvestris (L.) Mill.) and the domesticated apple (Malus domestica Borkh). *Molecular Ecology* 15: 2171–2182.

COEN, E., T. STRACHAN, AND G. DOVER. 1982. Dynamics of concerted evolution of ribosomal DNA and histone gene families in the melanogaster species subgroup of Drosophila. *Journal of Molecular Biology* 158: 17–35.

COOMES, O. T. 2004. Rain forest 'conservation-through-use'? Chambira palm fibre extraction and handicraft production in a land-constrained community, Peruvian Amazon. *Biodiversity and Conservation* 13: 351–360.

CRONN, R., A. LISTON, M. PARKS, D. S. GERNANDT, R. SHEN, AND T. MOCKLER. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* 36: e122.

DEMPEWOLF, H., N. C. KANE, K. L. OSTEVIK, M. GELETA, M. S. BARKER, Z. LAI, M. L. STEWART, E. BEKELE, J. M. ENGELS, Q. C. B. CRONK, AND L. H. RIESEBERG. 2010. Establishing genomic tools and resources for Guizotia abyssinica (L.f.) Cass.—The development of a library of expressed sequence tags, microsatellite loci and the sequencing of its chloroplast genome. *Molecular Ecology Resources* 10: 1048–1058.

DEXTER, K. G., T. D. PENNINGTON, AND C. W. CUNNINGHAM. 2010. Using DNA to assess errors in tropical tree identifications: How often are ecologists wrong and when does it matter? *Ecological Monographs* 80: 267–286.

DOORDUIN, L., B. GRAVENDEEL, Y. LAMMERS, Y. ARIYUREK, T. CHIN-A-WOENG, AND K. VRIELING. 2011. The complete chloroplast genome of 17 individuals of pest species Jacobaea vulgaris: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. *DNA Research* 18: 93–105.

DUGAN, L. E., M. F. WOJCIECHOWSKI, AND L. R. LANDRUM. 2007. A large scale plant survey: Efficient vouchering with identification through morphology and DNA analysis. *Taxon* 56: 1238–1244.

EDGAR, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

ELDER, J. F. JR., AND B. J. TURNER. 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly Review of Biology* 70: 297–320.

FELSENSTEIN, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* 22: 240–249.

FERRI, G., M. ALU, B. CORRADINI, AND G. BEDUSCHI. 2009. Forensic botany: Species identification of botanical trace evidence using a multigene barcoding approach. *International Journal of Legal Medicine* 123: 395–401.

FIGUEIRA, A., J. JANICK, AND P. GOLDSBROUGH. 1992. Genome size and DNA polymorphism in Theobroma cacao. *Journal of the American Society for Horticultural Science* 117: 673–677.

FITTER, R., AND R. KAPLINSKY. 2001. Who gains from product rents as the coffee market becomes more differentiated? A value chain analysis. *IDS Bulletin* 32: 69–82.

GANLEY, A. R. D., AND T. KOBAYASHI. 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research* 17: 184–191.

GLENN, T. C. 2011. Field guide to next-generation DNA sequencers. Molecular Ecology Resources.

GUINDON, S., AND O. GASCUEL. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52: 696–704.

HEBERT, P. D. N., A. CYWINSKI, S. L. BALL, AND J. R. DEWAARD. 2003. Biological identifications through DNA barcodes. *Proceedings. Biological Sciences* 270: 313–321.

HILLIER, L. W., G. T. MARTH, A. R. QUINLAN, D. DOOLING, G. FEWELL, D. BARNETT, P. FOX, ET AL. 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nature Methods* 5: 183–188.

HILLIS, D. M., AND S. K. DAVIS. 1988. Ribosomal DNA: Intraspecific polymorphism, concerted evolution, and phylogeny reconstruction. *Systematic Zoology* 37: 63–66.

HILLIS, D. M., C. MORITZ, C. A. PORTER, AND R. J. BAKER. 1991. Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science* 251: 308–310.

HOWARD, C. 2010. The development of deoxyribonucleic acid (DNA) based methods for the identification and authentication of medicinal plant material. Ph.D. dissertation, De Montfort University, Leicester, UK. Website http://hdl.handle.net/2086/3972.

HUSON, D. H., AND D. BRYANT. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267.

IBRAHIM, R. I., J. AZUMA, AND M. SAKAMOTO. 2006. Complete nucleotide sequence of the cotton (Gossypium barbadense L.) chloroplast genome with a comparative analysis of sequences among 9 dicot plants. *Genes & Genetic Systems* 81: 311–321.

INTERNATIONAL COCOA ORGANIZATION. 2011. About cocoa [online]. International Cocoa Organization, London, UK. Website http://www.icco.org/about/growing.aspx [accessed 20 June 2011].

IRISH, B. I., R. GOENAGA, D. ZHANG, R. SCHNELL, S. BROWN, AND J. C. MOTAMAYOR. 2010. Microsatellite fingerprinting of the USDA-ARS tropical agriculture research station cacao (Theobroma cacao L.) germplasm collection. *Crop Science* 50: 656–667.

IUCN [International Union for Conservation of Nature]. 2006. 2006 IUCN Red List. Website http://www.iucn.org [accessed December 16, 2011].

JANSEN, R. K., C. SASKI, S. B. LEE, A. K. HANSEN, AND H. DANIELL. 2011. Complete plastid genome sequences of three rosids (Castanea, Prunus, Theobroma): Evidence for at least two independent transfers of rpl22 to the nucleus. *Molecular Biology and Evolution* 28: 835–847.

KANE, N. C., AND Q. CRONK. 2008. Botany without borders, barcoding in focus. *Molecular Ecology* 17: 5175–5176.

KRESS, W. J., AND D. L. ERICKSON. 2008. DNA-barcoding—A windfall for tropical biology? *Biotropica* 40: 405–408.

KRONHOLM, I., O. LOUDET, AND J. DE MEAUX. 2010. Influence of mutation rate on estimators of genetic differentiation—Lessons from Arabidopsis thaliana. *BMC Genetics* 11: 33.

LAHAYE, R. M. van der Bank, D. Bogarin, J. Warner, F. Pupulin, G. Gigot, O. Maurin, S. Duthoit, T. G. Barraclough, and V. Savalainen. 2008. DNA barcoding the floras of biodiversity hotspot. Proceedings of the National Academy of Sciences, USA 105: 2923–2928.

LEE, S.-B., C. KAITTANIS, R. K. JANSEN, J. B. HOSTETLER, L. J. TALLON, C. D. TOWN, AND H. DANIELL. 2006. The complete chloroplast genome sequence of Gossypium hirsutum: Organization and phylogenetic relationships to other angiosperms. *BMC Genomics* 7: 61.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map (SAM) format and SAMtools. *Bioinformatics (Oxford, England)* 25: 2078–2079.

LOHSE, M., O. DRECHSEL, AND R. BOCK. 2007. OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* 52: 267–274.

LOU, S. K., K. L. WONG, M. LI, P. P. H. BUT, S. K. TSUI, AND P. C. SHAW. 2010. An integrated web medicinal materials DNA database: MMDBD (Medicinal Materials DNA Barcode Database). *BMC Genomics* 11: 402.

MEYERS, S., AND A. LISTON. 2010. Characterizing the genome of wild relatives of Limnanthes alba (meadowfoam) using massively parallel sequencing. *Acta Horticulturae* 859: 309–314.

MORIN, P. A., J. J. MOORE, AND D. S. WOODRUFF. 1992. Identification of chimpanzee subspecies with DNA from air and allele specific probes. *Proceedings. Biological Sciences* 249: 293–297 .

MOTAMAYOR, J. C., P. LACHNEAUD, J. W. DA SILVA E MOTA, R. LOOR, D. N. KUHN, J. S. BROWN, AND R. J. SCHNELL. 2008. Geographic and genetic population differentiation of the Amazonian chocolate tree (Theobroma cacao L). *PLoS ONE* 3: e3311.

MOTILAL, L. A., D. ZHANG, P. UMAHARAN, S. MISCHKE, M. BOCCARA, AND S. PINNEY. 2009. Increasing accuracy and throughput in large-scale microsatellite fingerprinting of cacao field germplasm collections. *Tropical Plant Biology* 2: 23–37.

MUELLNER, A. N., H. GREGER, AND C. M. PANNELL. 2009. Genetic diversity and geographic structure in Aglaia elaeagnoidea (Meliaceae, Sapindales), a morphologically complex tree species, near the two extremes of its distribution. Blumea—Biodiversity. *Evolution and Biogeography of Plants* 54: 207–216.

NEWMASTER, S. G., AND S. RAGUPATHY. 2010. Ethnobotany genomics—Discovery and innovation in a new era of exploratory research. *Journal of Ethnobiology and Ethnomedicine* 6: 2.

NEWTON, A. C. 2008. Conservation of tree species through sustainable use: How can it be achieved in practice? *Oryx* 42: 195–205.

NOCK, C. J., D. L. WATERS, M. A. EDWARDS, S. G. BOWEN, N. RICE, G. M. CORDEIRO, AND R. J. HENRY. 2011. Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal* 9: 328–333.

OVCHARENKO, I., G. G. LOOTS, R. C. HARDISON, W. MILLER, AND L. STUBBS. 2004. zPicture: Dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Research* 14: 472–477.

PALMER, J. D. 1985. Evolution of chloroplast and mitochondrial DNA in plants and algae. In R. J. MacIntyre [ed.], Monographs in evolutionary biology: Molecular evolutionary genetics, 131–240. Plenum, New York, New York, USA.

PARKS, M., R. CRONN, AND A. LISTON. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* 7: 84.

PENNISI, E. 2007. Wanted: A barcode for plants. *Science* 318: 190–191.

PETIT, R. J., AND G. G. VENDRAMIN. 2007. Plant phylogeography based on organelle genes: An introduction. In S. Weiss and N. Ferrand [eds.], Phylogeography of southern European refugia, 23–97. Springer, Dordrecht, Netherlands.

PIREDDA, R., C. S. MARCO, M. ATTIMONELLI, R. BELLAROSA, AND B. SCHIRONE. 2011. Prospects of barcoding the Italian wild dendroflora: Oaks reveal severe limitations to tracking species identity. *Molecular Ecology Resources* 11: 72–83.

POSADA, D. 2008. jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution* 25: 1253–1256.

RAGUPATHY, S., S. G. NEWMASTER, M. MURUGESAN, AND V. BALASUBRAMANIAM. 2009. DNA barcoding discriminates a new cryptic grass species revealed in an ethnobotany study by the hill tribes of the Western Ghats in southern India. *Molecular Ecology Resources* 9: 164–171.

ROGERS, S. O., AND A. J. BENDICH. 1987. Ribosomal RNA genes in plants: Variability in copy number and in the intergenic spacer. *Plant Molecular Biology* 9: 509–520.

ROSS, B. C., K. RAIOS, K. JACKSON, AND B. DWYER. 1992. Molecular cloning of a highly repeated DNA element from Mycobacterium tuberculosis and its use as an epidemiological tool. *Journal of Clinical Microbiology* 30: 942–946.

ROWNTREE, J. K., R. S. COWAN, M. LEGGETT, M. M. RAMSAY, AND M. F. FAY. 2010. Which moss is which? Identification of the threatened moss Orthodontium gracile using molecular and morphological techniques. Conservation Genetics 11: 1033–1042. Schaal, B. A., D. A. Hayworth, K. M. Olsen, J. T. Rauscher, and W. A. Smith. 1998. Phylogeographic studies in plants: Problems and prospects. *Molecular Ecology* 7: 465–474.

SCHAAL, B. A., AND G. H. LEARN JR. 1988. Ribosomal DNA variation within and among plant populations. *Annals of the Missouri Botanical Garden* 75: 1207–1216.

SCHWARTZ, S., W. J. KENT, A. SMIT, Z. ZHANG, R. BAERTSCH, R. C. HARDISON, D. HAUSSLER, AND W. MILLER. 2003. Human-mouse alignments with BLASTZ. *Genome Research* 13: 103–107.

SIMPSON, J. T., K. WONG, S. D. JACKMAN, J. E. SCHEIN, S. J. JONES, AND I. BIROL. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research* 19: 1117–1123.

SOLTIS, D. E., M. A. GITZENDANNER, D. D. STRENGE, AND P. S. SOLTIS. 1997. Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. *Plant Systematics and Evolution* 206: 353–373.

SOSNICKI, A. A., AND S. NEWMAN. 2010. The support of meat value chains by genetic technologies. *Meat Science* 86: 129–137.

SOLTIS, D. E., A. E. SENTERS, M. J. ZANIS, S. KIM, J. D. THOMPSON, P. S. SOLTIS, L. P. RONSE DE CRAENE, P. K. ENDRESS, AND J. S. FARRIS. 2003. Gunnerales are sister to other core eudicots: Implications for the evolution of pentamery. *American Journal of Botany* 90: 461–470.

STEELE, P. R., AND J. C. PIRES. 2011. Biodiversity assessment: State-of-the-art techniques in phylogenomics and species identification. *American Journal of Botany* 98: 415–425.

STRAUB, S. C. K., M. FISHBEIN, T. LIVSHULTZ, Z. FOSTER, M. PARKS, K. WEITEMIER, R. C. CRONN, ET AL. 2011. Building a model: Developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. *BMC Genomics* 12: 211.

STRAUB, S. C. K., M. PARKS, K. WEITEMIER, M. FISHBEIN, R. C. CRONN, AND A. LISTON. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* 99: ~~350–365~~.

WARD, J., S. R. GILMORE, J. ROBERTSON, AND R. PEAKALL. 2009. A grass molecular identification system for forensic botany: A critical evaluation of the strengths and limitations. *Journal of Forensic Sciences* 54: 1254–1260.

WENDEL, J. F., AND V. A. ALBERT. 1992. Phylogenetics of the cotton genus (Gossypium): Character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. *Systematic Botany* 17: 115–143.

Whitlock, M. 2011. G'_{ST} and D do not replace FST. *Molecular Ecology* 20: 1083–1091.

WHITTALL, J. B., J. SYRING, M. PARKS, J. BUENROSTRO, C. DICK, A. LISTON, AND R. CRONN. 2010. Finding a (pine) needle in a haystack: Chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* 19 (supplement 1): 100–114.

WOLFE, K. H., W. H. LI, AND P. M. SHARP. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA* 84: 9054–9058.

WOOD, G. A. R., AND R. A. LASS. 2001. Cocoa, 4th ed. Longman Group, Blackwell, UK.

WYMAN, S. K., R. K. JANSEN, AND J. L. BOORE. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics (Oxford, England)* 20: 3252–3255.

YANG, J. Y., A. LAMBERT, L. A. MOTILAL, H. DEMPEWOLF, K. MAHARAJ, AND Q. C. B. CRONK. 2011. Chloroplast microsatellite primers for cacao (Theobroma cacao). *American Journal of Botany* 98: e372–e374.

YAO, H., J. SONG, C. LIU, K. LUO, J. HAN, Y. LI, X. PANG, H. XU, Y. ZHU, P. XIAO, AND S. CHEN. 2010. Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS ONE* 5: e13102.

YESSON, C., R. T. BÁRCENAS, H. M. HERNÁNDEZ, M. DE LA LUZ RUÍZ-MAQUEDA, A. PRADO, V. M. RODRÍGUEZ, AND J. A. HAWKINS. 2011. DNA barcodes for Mexican Cactaceae, plants under pressure from wild collecting. *Molecular Ecology Resources* 11:

ZHANG, Y.-J., P.-F. MA, AND D.-Z. LI. 2011. High-throughput sequencing of six bamboo chloroplast genomes: Phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). *PLoS ONE* 6: e20596.

ZURAWSKI, G., AND M. T. CLEGG. 1987. Evolution of higher-plant chloroplast DNA-encoded genes: Implications for structure–function and phylogenetic studies. *Annual Review of Plant Physiology* 38: 391–418.

ZWICKL, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, University of Texas at Austin, Austin, Texas, USA.