Commercial Teas Highlight Plant DNA Barcode Identification Successes and Obstacles

Mark Y. Stoeckle1, Catherine C. Gamble2, Rohan Kirpekar2, Grace Young2, Selena Ahmed3 & Damon P. Little4

1Program for the Human Environment, The Rockefeller University, New York, NY 10065, USA, 2Trinity School, New York, NY 10024, USA, 3Department of Biology, Tufts University, Medford, MA 02115, USA, 4Lewis B. and Dorothy Cullman Program for Molecular Systematics, The New York Botanical Garden, Bronx, New York, NY 10458, USA.

Appearance does not easily identify the dried plant fragments used to prepare teas to species. Here we test recovery of standard DNA barcodes for land plants from a large array of commercial tea products and analyze their performance in identifying tea constituents using existing databases. Most (90%) of 146 tea products yielded rbcL or matK barcodes using a standard protocol. Matching DNA identifications to listed ingredients was limited by incomplete databases for the two markers, shared or nearly identical barcodes among some species, and lack of standard common names for plant species. About 1/3 of herbal teas generated DNA identifications not found on labels. Broad scale adoption of plant DNA barcoding may require algorithms that place search results in context of standard plant names and character-based keys for distinguishing closely-related species. Demonstrating the importance of accessible plant barcoding, our findings indicate unlisted ingredients are common in herbal teas.

Aqueous infusions prepared from dried plants, broadly known as teas, are popular beverages with desirable physiologic activities and potential health benefits. Accurate labeling is important for consumers, marketers, and regulators, as tea constituents cannot be easily identified to species by visual appearance. Their taxonomic diversity and fragmentary nature present a ready and demanding test of DNA-based identification. Here we report the successes with and obstacles to identifying tea ingredients using a short DNA sequence from a uniform locality within the genome, DNA barcoding.

Tea properly refers to infusions prepared from leaves of the tea plant, Camellia sinensis (L.) Kuntze, an evergreen flowering tree in the family Theaceae, native to the mountainous regions of southwestern China and neighboring countries. The two main commercial varieties are small-leafed C. sinensis var. sinensis, adapted to cool weather and high altitude, and large-leafed C. sinensis var. assamica (J. W. Mast.) Kitam., which grows well in tropical and sub-tropical environments. Tea plant leaves contain a high concentration of phytochemicals including polyphenolic catechins and the methylxanthine caffeine. Tea drinking originated in southern China at least 2000 years ago, and today tea is the most widely consumed beverage in the world.

In 2009, the Plant Working Group of the Consortium for the Barcode of Life (CBOL) endorsed a proposal to use defined portions of the plastid genes rbcL (~550 bp segment) and matK (~790 bp segment) as standard barcodes for land plants. These and other candidate markers have been tested in various floristic and taxonomic settings. As compared to animals, plants generally have less barcode variation both within and among species. A relatively large proportion of plants (~15%-30%) share barcodes among multiple species. Plant barcodes generally do not exhibit the strong clustering pattern observed in most animal species (interspecific variation ≪ intraspecific variation). These observations apply even when longer sequences or additional markers are sampled.
which may reflect fundamental differences in plant and animal biology and evolution. Notwithstanding these limitations, standard plant barcodes are efficacious in a number of scientific and applied settings and have enormous potential for wider use.

In this study we explored a practical application of plant barcoding: matching commercial tea ingredients to product labels. We searched a public reference database for the closest match to each barcode sequence and compared the result to the listed ingredients. Because the tea specimens are morphologically unrecognizable, we cannot know with certainty if the source plants are represented in the reference database, a realistic and difficult test of barcode identification.

**Results**

**Barcode recovery, haplotypes, matches.** Using single sets of primers for each locus, readable *rbcL* or *matK* barcodes were recovered from 131 (90%) of 146 tea products, including 96% of CS and 84% of herbal teas. *rbcL* was recovered from 113/146 (77%), *matK* from 108/136 (79%), and both from 90/136 (66%). A total of 253 readable sequences were obtained, comprising 48 *rbcL* and 40 *matK* haplotypes (Figs. 1,2; additional details in Supplementary Tables S1,S2 online). There were no insertions or deletions in *rbcL* sequences; the *matK* alignment contained 14 different types of insertions or deletions. For each haplotype, BLAST searches of GenBank and Barcode of Life databases were performed. The closest match in each database was recorded. As compared to results with GenBank, BOLD matches were on average lower identity and fewer were label ingredients, indicating that at the time of the study BOLD was less well populated with barcodes of plants used in commercial tea products. As a result, subsequent analyses were performed using GenBank. The *rbcL* haplotypes matched 42 species in 24 families; the *matK* haplotypes matched 25 species in 16 families (Figs. 1,2).

Taking into account uncertainties arising from incomplete databases, shared barcodes, and ambiguous common names, of 48 *rbcL* haplotypes, 32 were assigned to species, 10 to genus, and 6 to family. Of 40 *matK* haplotypes, 27 were assigned to species, 8 to genus, and 5 to family (Figs. 1,2). In most cases (58%), barcodes recovered from commercial tea products matched listed ingredients. It should be noted that our study was designed to enable comparison between CS and herbal teas, and not among individual products or manufacturers. Given this and potential liability issues, we assigned arbitrary alphanumeric codes to each product to protect the manufacturer’s identity. Most of the barcodes that did not match listed ingredients reflected an incomplete reference database, lacking either a record for the relevant species or a record of an intraspecific variant. For example, an herbal tea labeled “Marshmallow (*Althaea officinalis*)” produced an *rbcL* sequence closest to *Anisodontea*

---

**Figure 1 | rbcL barcode identifications.** For each haplotype, alphanumeric code, number of isolates, identification, and graphic representation of match results are shown. Color bars depict percent identity of closest match, nearest neighbor (NN) in the same genus, and NN in a different genus, with scale at bottom. Haplotypes for which the second closest match was in a different genus have a blank in “NN same genus” column. (Note: *P. pentandrum* = *Pittosporum pentandrum*).
triloba (1 mismatch, 99.8% identity). However, at the time of the study there were no GenBank rbcL records for A. officinalis. Overall, at the time of the study about one-third of plant species listed on product labels lacked rbcL or matK records in GenBank. Reflecting incomplete representation of intraspecific variants, more than half of C. sinensis tea products yielded an rbcL barcode 100% identical to congeneric species C. oleifera and C. sasanqua but with one mismatch compared to the C. sinensis rbcL record.

Barcode identifications were incompatible with listed ingredients for some products, including 21/60 (35%) herbal and 3/70 (4%) CS teas (Table 1). Some of the non-label DNAs matched plants used in other tea products, some matched common weeds or other non-food plants, and some could not be identified. The most common non-label ingredient, found in seven products, was chamomile (Matricaria recutita). Four herbal teas yielded sequences identified as tea plant (C. sinensis), although none listed ingredients in the tea family (Theaceae). Regarding non-food plants, a product labeled “St. John’s wort (Hypericum perforatum),” a flowering plant, yielded an rbcL sequence identical to that of several fern species. A barcode from an herbal tea matched Poa annua, a widely cultivated meadow grass. Four products yielded barcodes closely matching plants in Apiaceae, the parsley family, although the particular species could not be determined. Apiaceae includes many food plants and ubiquitous wild relatives, but for the products in question none of the listed ingredients were in this family.

### Taxonomic resolution
For most rbcL haplotypes, the differences between closest match, nearest neighbor (NN) in the same genus, and NN in a different genus were modest or absent. Among the 48 haplotypes, the average percent identity was 99.9% for closest, 99.8% for congeneric NN, and 99.2% for NN in a different genus, or about 0.6, 1.1, and 4.6 nucleotide differences respectively (Fig. 1; additional details in Supplementary Table S1 online). Of 32 rbcL haplotypes with 100% match, 15 were also identical to one or more congeneric species and eight were identical to one or more species in a different genus.

For matK, the average identities were 99.5% for closest match, 99.5% for NN congeneric, and 98.1% for NN different genus, or about 3.8, 3.8, and 14.3 nucleotide differences (Fig. 2; additional details in Supplementary Table S2 online). Of 14 haplotypes with a 100% match, three were also identical to one or more congeneric species, and none were identical to species in a different genus.

### C. sinensis rbcL nucleotide sequence polymorphism
We observed nucleotide variation (A or C) in CS rbcL sequences at a site corresponding to position 68 of the coding region (gi 7525012:54958-56397 was used as a reference), with the predicted predicted amino acid being either asparagine (68A) or threonine (68C). The 68A sequence was identical to the C. sinensis rbcL GenBank record, whereas the 68C variant was identical to rbcL sequences of several congeneric species (C. albogigas, C. granthamiana, C. japonica, C. oleifera, C. sasanqua) and a related species Tutcheria hirta. Among tea products for which geographic or tea type information was available, the 68C variant was strongly associated with C. sinensis

---

**Figure 2 | matK barcode identifications.** For each haplotype, alphanumeric code, number of isolates, identification, and graphic representation of match are shown as described in Fig. 1 legend.

| Family | Haplotype | Closest Match | NN Same Genus | NN Diff. Genus |
|--------|-----------|---------------|---------------|---------------|
| Fabaceae | R154z-1 | Medicago sativa | Medicago sativa |  |
| Rosaceae | K26-1 | Medicago sativa | Medicago sativa |  |
| | TT221-1 | Senecio sp. indet. | Senecio sp. indet. |  |
| Myrtaceae | TT208-1 | Psidium guajava | Psidium guajava |  |
| Malvaceae | G13-1 | Althaea officinalis | Althaea officinalis |  |
| | TT219-1 | Tilia americana | Tilia americana |  |
| Caricaceae | TT216-1 | Carica papaya | Carica papaya |  |
| | G25-1 | Carissa papaya | Carissa papaya |  |
| Cannabisaceae | G27-1 | Carissa sinensis | Carissa sinensis |  |
| | R10-5 | Carissa sinensis | Carissa sinensis |  |
| | R1-1 | Carissa sinensis | Carissa sinensis |  |
| | R34-1 | Carissa sinensis | Carissa sinensis |  |
| | R36-2-59 | Carissa sinensis | Carissa sinensis |  |
| | K10-3 | Carissa sinensis | Carissa sinensis |  |
| | R25-1 | Carissa sinensis | Carissa sinensis |  |
| Ericaceae | TT216-1 | Vaccinium myrtillus | Vaccinium myrtillus |  |
| Oleaceae | R4A-1 | Jasminum sp. indet. | Jasminum sp. indet. |  |
| | TT219-1 | Verbenasum thapsus | Verbenasum thapsus |  |
| | TT205-1 | Lonicera sp. | Lonicera sp. |  |
| | R10A-1 | Scutellaria barbata | Scutellaria barbata |  |
| | TT207-1 | Ondorites venus | Ondorites venus |  |
| | G31-1 | Ilex sp. indet. | Ilex sp. indet. |  |
| | TT211-1 | Acalypha mellefolium | Acalypha mellefolium |  |
| | TT128-1 | Cinusam sp. indet. | Cinusam sp. indet. |  |
| | TT20A-1 | Echinacea sp. indet. | Echinacea sp. indet. |  |
| | G2-12 | Matricaria recutita | Matricaria recutita |  |
| | K18-2 | Matricaria recutita | Matricaria recutita |  |
| | R39-1 | Matricaria recutita | Matricaria recutita |  |
| | R41z-1 | Stellaria rebusculana | Stellaria rebusculana |  |
| Pittosporaceae | K28z-1 | P. peniandrum | P. peniandrum |  |
| Apiaceae | R8A-5 | Apiaceae | Apiaceae |  |
| | R33-1 | Apiaceae | Apiaceae |  |
| | R45-1 | Apiaceae | Apiaceae |  |
| | R46-1 | Foeniculum vulgare | Foeniculum vulgare |  |
Table 1 | DNA barcode identification of unlisted ingredients.

| Product | Label | Non-label DNA | Comment |
|---------|-------|---------------|---------|
| G1      | apple pieces, vitamin C, citric acid, natural flavor | chamomile (Matricaria recutita) | G1 matK 100% identity to *Camellia sinensis*, family Theaceae. No listed ingredients in Theaceae. |
| G2      | apple pieces, orange peel, rosehips, hibiscus, cornflower blossoms, clove, cinnamon, anise, pepper, natural flavor | chamomile (Matricaria recutita) | G2 rbcL 99.8% identity to *Pentzia incana*, G2 matK 98.9% match *Achillea millefolium*, both family Asteraceae. rbcL and matK sequences most likely represent chamomile (Matricaria recutita), based on 100% match to partial *M. recutita* sequences in GenBank and recovery of identical or nearly identical sequences from products listing chamomile as sole ingredient. Listed ingredient cornflower is not an approved name for chamomile. It refers to *Centaurea cyanus*, family Asteraceae. Compared to closest match, G2 sequence is more distant from *C. cyanus* rbcL (98.6%) [AB530955]; no *C. cyanus* matK sequences in GenBank. Other ingredients in different families. |
| G4      | raspberry pieces, apple pieces, orange peel, rosehips, hibiscus, lemongrass, vitamin C, natural raspberry flavor | chamomile (Matricaria recutita) | G4 matK 98.9% match *Achillea millefolium*, family Asteraceae, most likely represents chamomile (Matricaria recutita) (see Comment under G2). No listed ingredients in Asteraceae. |
| G6      | | Prunella vulgaris | G6 matK 98.9% match *Achillea millefolium*, family Asteraceae, most likely represents chamomile (Matricaria recutita) (see Comment under G2). No listed ingredients in Asteraceae. |
| G10     | | honeysuckle flower | G10 matK 98.9% identity *Achillea millefolium*, family Asteraceae, most likely represents chamomile (Matricaria recutita) (see Comment under G2). No listed ingredients in Asteraceae. |
| K21     | | opio d’arco inner bark | K21 rbcL and matK 100% match *Camellia sinensis*, family Theaceae. No listed ingredients in Theaceae. |
| K22     | | rosehips, orange peel, chamomile flowers, lemongrass, lemon myrtle, hibiscus flowers, nana mint, natural citrus flavors and other natural flavors | K22 rbcL 100% match to *P. annua* (annual bluegrass), family Poaceae. Listed ingredient lemongrass, *Citropogon citratus*, is also in Poaceae. However, compared to closest match, K24 sequence is more distant from *C. citratus* (94.0%) (GQ436383). No other ingredients in Poaceae. |
| K24     | | ginger root, natural flavors, Linden, lemon peel, blackberry leaves, lemongrass, citric acid | K24 rbcL 100% match to *P. annua* (annual bluegrass), family Poaceae. Listed ingredient lemongrass, *Citropogon citratus*, is also in Poaceae. However, compared to closest match, K24 sequence is more distant from *C. citratus* (94.0%) (GQ436383). No other ingredients in Poaceae. |
| K27     | | eleuthero, peppermint, cinnamon, ginger, chamomile, west indian lemongrass, licorice, catnip, tilia flowers, natural lemon flavor, hops, vitamins B6 and B12 | K27 matK 100% match to *Chenopodium album*, family Amaranthaceae. In BOLD, K27 matK is 100% match to *Chenopodium album*. No listed ingredients in Amaranthaceae. |
| R8      | | white goosfoot (*Chenopodium album*) | R8 rbcL 100% match to *Camelina oleifera*, family Thaceae. No listed ingredients in Theaceae. R8 rbcL 99.8% and matK 99.7% match *Pimpinella saxifraga*, family Apiaceae. No listed ingredients in Apiaceae. |
| R10     | | tea (*Camellia sinensis*), parsley family (Apiaceae) | R10 matK 99.6% match to *Pentzia incana*, R10 matK 98.9% match *Achillea millefolium*, both family Asteraceae, likely represents chamomile (Matricaria recutita) (see Comment under G2). No listed ingredients in Asteraceae. R10 matK 99.7% match *Scutellaria barbata*, family Lamiaceae. Listed ingredient spearmint (*Mentha spicata*) is also in Lamiaceae. However compared to closest match, R10 matK2 is relatively distant from *M. spicata* (91.8%) (GU381684). No other ingredients in Lamiaceae. |
| R15     | | black tea plus rooibos, black pepper, cardamom, cinnamon, ginger, organic cane sugar, natural flavors | R15 rbcL 99.6% match *Pentzia incana* and R15 matK 98.9% match *Achillea millefolium*, both family Asteraceae, likely represents chamomile (Matricaria recutita) (see Comment under G2). No listed ingredients in Asteraceae. R15 matK 99.7% match *Medicago sativa*, family Fabaceae. Listed ingredient rooibos (*Aspalathus linearis*) is also family Fabaceae. No *A. linearis* matK sequences in GenBank for comparison, but *A. linearis* and *M. sativa* rbcL sequences show limited identity (93.6%). Other ingredients in different families. |
Table 1 | (Continued)

| Product: | R41 |
| Label: | carob pod, Indian sarsaparilla root, ginger root, kava root, cinnamon bark, stevia leaf, cardamom seed, natural flavors, barley malt, essential oils |
| Non-label DNA: | tea (Camellia sinensis) |
| Comment: | R41 rbcL 99.8% identity to Camellia oleifera, family Theaceae. No listed ingredients in Theaceae. |

| Product: | R45 |
| Label: | lemongrass, blackberry leaves, citric acid, rose hips, spearmint, natural flavors, orange peel, safflowers, hibiscus flowers, rose petals, orange essence, ginger, licorice, natural flavors |
| Non-label DNA: | parsley family (Apiaceae) |
| Comment: | R45 rbcL 99.1% match Heteromorpha arborescens, matK 99.7% match Pimpinella saxifraga, both family Apiaceae. No listed ingredients in Apiaceae. |

| Product: | TT204 |
| Label: | St. John's wort (aerial part) (Hypericum perforatum) |
| Non-label DNA: | fern (Terpinschore sp. indet.) |
| Comment: | TT204 rbcL 100% match to several species in genus Terpinschore, family Polypodiaceae. No listed ingredients in Polypodiaceae. |

| Product: | TT207 |
| Label: | eyebright herb (Euphrasia officinalis) |
| Non-label DNA: | red bartsia (Odontites vernus) |
| Comment: | TT207 rbcL 100% and matK 99.5% match to Odontites vernus, family Orobanchaceae. Listed ingredient Euphrasia officinalis is also in family Orobanchaceae. There are no E. officinalis sequences in GenBank for direct comparison. However, the closest Euphrasia species with sequences in GenBank is relatively distant from recovered sequence: TT207 rbcL is 97.3% match to E. spectabilis AY849864, and TT207 matK is 93.7% match to E. spectabilis AY849603. |

| Product: | TT210 |
| Label: | rooibos (Aspalathus linearis), lemongrass (Cymbopogon citratus), stevia (Stevia rebaudiana) |
| Non-label DNA: | blackberry (Rubus sp. indet.) |
| Comment: | TT210 matK 99.9% match to Rubus discolor, family Rosaceae. No listed ingredients in Rosaceae. |

| Product: | TT213 |
| Label: | yellowlock root (Rumex crispus) |
| Non-label DNA: | papaya (Carica papaya) |
| Comment: | TT213 matK 100% match to Carica papaya, family Caricaeae. No listed ingredients in Caricaeae. |

| Product: | TT225 |
| Label: | hirpiricao (Hypericum perforatum) |
| Non-label DNA: | lemon balm (Melissa officinalis) |
| Comment: | TT225 rbcL 99.8% match to Melissa officinalis, family Lamiaceae. No listed ingredients in Lamiaceae. |

| Product: | MGm17 |
| Label: | orange, mango, cinnamon |
| Non-label DNA: | lantana (Lantana sp. indet.) |
| Comment: | MGm17 rbcL 99.6% identity Lantana camara, family Verbenaceae. No listed ingredients in Verbenaceae. |

| Product: | MRa6 |
| Label: | stevia (Stevia rebaudiana) |
| Non-label DNA: | MRa6 rbcL 100% identity Stevia rebaudiana, family Asteraceae, tribe Eupatorieae. Listed ingredient chicory (Cichorium intybus) is also in family Asteraceae, but in a different tribe, Cichoreae. MRa6 sequence has lower identity to C. intybus (97.4%) [L13652] than to S. rebaudiana sequence. No other ingredients in Asteraceae. |

| Product: | R23 |
| Label: | Formosa oolong tea |
| Non-label DNA: | heal all (Prunella vulgaris), chamomile (Matricaria recutita) |
| Comment: | R23 rbcL 100% match to Prunella vulgaris, family Lamiaceae. R23 matK 98.9% match to Achillea millefolium, family Asteraceae, likely represents chamomile (Matricaria recutita) (see Comment under G2). No listed ingredients in Lamiaceae or Asteraceae. |

| Product: | R33 |
| Label: | Sichuan tea |
| Non-label DNA: | parsley family (Apiaceae) |
| Comment: | R33 matK 99.7% match Pimpinella saxifraga, family Apiaceae. No listed ingredients in Apiaceae. |

| Product: | R36 |
| Label: | gunpowder tea |
| Non-label DNA: | parsley family (Apiaceae) |
| Comment: | R36 matK 99.7% match Pimpinella saxifraga, family Apiaceae. No listed ingredients in Apiaceae. |

var. assamica vs. C. sinensis var. sinensis (71% vs. 12%, p = 0.0002) (additional details in Supplementary Table S3 online).

**Discussion**

Reliable DNA identification of species requires recovery of a barcode sequence from the sample, representation of relevant species in the reference database, and sufficient nucleotide sequence variability to distinguish among closely-related species. Regarding the first requirement, we recovered rbcL or matK barcodes from 90% of commercial tea products using a single set of primers for each region. Success was less frequent with herbal as compared to CS teas (84% vs 96%), which may reflect primer mismatch, Taq inhibition, or DNA degradation in some of the diverse plant materials in herbal teas. In terms of markers, rbcL was recovered from a broader taxonomic range of plants than matK (42 species in 24 families vs. 25 species in 16 families; Figs. 1,2). These results are consistent with general observation that rbcL is more easily amplified from wide range of species than is matK.

The second condition for DNA identification of species is representation of relevant taxa in the reference database, in our case GenBank. As in most practical applications of barcoding, our specimens were morphologically unrecognizable, thus representation cannot be assessed directly. About one-third of the plant species listed on labels lacked GenBank records for rbcL, matK, or both at...
the time of the study. A more precise indicator of species representation is whether the recovered sequences are identical to any in the database. 62% of our barcode haplotypes did not have an identical match in GenBank (Figs. 1,2). This indicates that many plant species found in tea products are either not represented, have undocumented intraspecific variation, or that a sequencing error has occurred.

The third requirement for identifying species by barcode is biological: there must be sequence differences that discriminate among closely-related species. We can determine how well this condition is met for our specimens by comparing the best match and the congeneric nearest neighbor for each haplotype. For rbcL, these differed by only 1 site on average, and for matK these differed by only 2 sites on average (Figs. 1,2; see also Supplementary Tables S1,S2 online). Our results are consistent with the estimated 70%–85% species discrimination using rbcL + matK barcodes, and highlight the relatively small number of positions that distinguish many closely-related plant species8,23,24. Differences between congeneric species in this study are similar to those reported for intraspecific variation and are also the same magnitude as sequencing error. Thus a barcode that differs from its closest reference database sequence at just one or a few sites plausibly represents an unrecorded variant for that species, a closely-related species not in the reference database, or sequencing error.

Our results highlight a need for improved algorithms for assigning taxonomic names to plant barcode sequences, particularly if barcoding is to be applied by non-specialists, which is one of the goals of the effort1,12,25. Algorithms that place search results in the context of plant taxonomy and current database representation of related plants will be helpful. Character-based approaches may assist in distinguishing closely-related species, particularly if supported by expert annotation that flags diagnostic nucleotide positions27,28. In addition, although employing two markers adds precision to plant barcode identifications, it also generates a need for algorithms that integrate database search results. In our data, most extractions that yielded both markers gave discordant results, that is, the rbcL and matK barcodes matched different species in GenBank, largely reflecting differences in representation of species or intraspecific variants for the two markers.

A large fraction (35%) of herbal products yielded one or more barcodes that pointed to non-label ingredients. Possible explanations include database errors (e.g. sequences with incorrect species names), limitations of search algorithm (e.g. relevant sequences not recognized by BLAST), laboratory error (e.g. PCR contamination, sample mix-up), or presence of unlabeled ingredients. The disproportionate number of discordant sequences recovered from herbal specimens and the finding of species not listed on other products and not under study in the laboratory points to unnamed constituents. This could reflect inadvertent introduction, such as from harvested plant material mixed with unrecognized species, residual ingredients in processing machinery, or as part of unspecified flavorings listed on some products. The relative amount of such potential material in our samples is unknown and is beyond the scope of this study. The finding of unlabeled chamomile (M. recutita) or tea plant (C. sinensis) in multiple products suggests the possibility of addition or substitution to improve taste, appearance, or for economic reasons26.

To our knowledge, the polymorphism at rbcL position 68 is the first described plastid marker that differs among C. sinensis varieties, regions of cultivation, and tea processing types26,28. Our results are consistent with marketplace trends—India and Sri Lanka, largely devoted to cultivation of C. sinensis var. assamica, are the dominant global exporters of black tea, whereas China, largely cultivating C. sinensis var. sinensis, has become the dominant exporter of green tea, with 75% of world market29. Our findings may help inform future research on the geographic origin and diversity of wild and cultivated CS resources29,30.

In summary, plant DNA barcodes can be recovered from most commercial tea products using a standard protocol. At the same time, interpreting DNA barcode identifications in relation to product labels is challenging. New algorithms that place search results in the context of standard plant names and character-based keys for distinguishing closely-related species are needed. With appropriate software to guide non-experts, DNA barcoding can offer an effective method to help provide more accurate ingredient labels to consumers, thereby improving safety of food and botanicals25,32. This is particularly pertinent in an increasingly global economy where longer and more complex market chains distance suppliers from the source of products and where regulatory agencies are becoming more stringent with food and botanical labeling25,33.

Methods

Specimen collection. CS and herbal tea products from New York City stores, school dining halls, and homes of investigators were collected during October 2009–February 2010. 146 products were obtained from 25 locations, representing 33 manufacturers, 17 countries, and 82 plant common names. As this study was designed to enable comparison between CS and herbal teas and not among individual products or manufacturers, products were assigned an arbitrary alphanumerical code. 73 were C. sinensis, and 73 were herbal products prepared from other plant species.

Five herbal products contained C. sinensis together with other plants. 44 herbal teas (60.3%) listed a single ingredient; the remainder named 2–10 different plants. When not specified on the label, scientific and common name equivalents were determined from the reference used by the U.S. Food and Drug Administration29.

Reference samples. C. sinens var. assamica specimens (n = 17) were collected in Yunnan, China by SA during 2007–2009. C. sinensis var. sinensis specimens (n = 24) collected in China (7), Taiwan (7), Japan (7), and Argentina (3) were obtained from the Kunming Institute of Botany, Kunming, China. Reference sample rbcL sequences and additional collection information were deposited in GenBank under accession codes JN909623–JN909663. GenBank accesses used for comparison of C. sinensis rbcL haplotypes included C. alibogia (AF380033), C. grandithamiana (AF380034), C. japonica (AF380035), C. oleifera (GG436637), C. sasanqua (AF380036), C. sinensis (AF380037), and Tutcheria hirta (AF380067).

DNA extraction and sequencing. DNA was isolated from 5–15 mg dried tissue using a DNeasy96 Plant kit (Qiagen). The manufacturer’s protocol was modified as follows: tissue was disrupted and then incubated for 12–18 h with gentle mixing at 42 °C in 600 μL of the supplied AP1 buffer with 600 μg of protease K added (630 μL total volume). Polysaccharides were precipitated at 4 °C with 200 μL of AP2. The remaining steps followed the manufacturer’s protocol. For the 86% of specimens that appeared morphologically homogenous, a single extraction was performed. The remaining samples were divided into groups of morphologically homogenous material (average 3 range 2–8), and separate extractions were performed with the aim of recovering individual components.

Individual amplifications of matK and rbcL took place in a 15 μL volume containing: 1.5 μL buffer 200 mM Tris pH 8.8, 100 mM KCL 100 mM (NH4)2SO4, 20 mM MgSO4+7H2O, 1% (v/v) Triton X-100, 50% (w/v) sucrose, 0.25% (w/v) cresol red, 0.2 μM dNTPs, 0.025 μg/mL BSA, 0.5 (rbcL) or 1 (matK) μM of each primer, 1 unit of Taq, and 0.5 μL genomic DNA. For amplification and sequencing of matK, primers 3F (5’-CGT-ACA-GTA-CTT-TTG-TGT-CTA-GCA-G-3’) and 1R (5’-ACC-CAG-TCC-TAT-GGC-AAA-ATA-TGG-TGT-3’) were used with the following cycling conditions: 95 °C 2.5 min; 10 cycles: 95 °C 30 s, 56 °C 30 s, 72 °C 30 s; 25 cycles: 88 °C 30 s, 56 °C 30 s, 72 °C 30 s; 72 °C 2 min. For rbcL amplification and sequencing, primers F1 (5’-ATG-TCA-CCA-GAA-AGC-ACT-G-3’) and R634 (5’-GAA-ACG-TCT-GAT-CCA-GAT-3’) were used with the following cycling conditions: 95 °C 2.5 min; 35 cycles: 95 °C 30 s, 58 °C 30 s, 72 °C 30 s; 72 °C 2 min. PCR products were treated with ExoSAP-IT and bi-directionally sequenced with BigDye 3.1 chemistry on an ABI 3730 sequencer (High–Throughput Genomics Unit, University of Washington).

Portable laboratory. A subset of specimens (10) were analyzed in a portable laboratory. Equipment included a thermal cycler (Technne), microcentrifuge (Eppendorf minispin), vortex mixer, heating block, pipettes, and E-gel apparatus (Invitrogen), purchased used or reconditioned except for E-gel unit. DNA was isolated from DNeasy Plant Mini Kit (Qiagen) following manufacturer’s instructions. PCR was performed using rbcL primers as described above except that 25 μL reaction volume, 0.5 units TaKara Ex Taq, and buffer supplied by manufacturer were used. DNA and PCR yields were assessed on an E-gel EX 1% with a blue-light excitable nucleic acid stain. products were cleaned with QIA quick PCR purification kit (Qiagen), and unidirectional sequencing was performed at a commercial facility (Macrogen).

Sequence files and data analysis. Trace files were assembled in MacVector 11.0, and sequences with greater than 2% ambiguous bases were discarded, using QV of 40 for
bi-directional reads 20 for single reads. Sequences were aligned using ClustalW (rbcL) or MUSCLE v3.8.31 (matK). Sequence files are deposited in GenBank under accession codes HQ699082-HQ699129 (rbcL) and HQ699130-HQ699169 (matK). Fisher's exact test, two-tailed, was used for statistical comparisons.

Database searches. GenBank database was searched using megaBLAST during August-October 2010, with default parameters adjusted to retrieve 5000 sequences. To optimize correct identifications, the closest match for each rbcL and matK haplotype was defined as the target with highest percentage identity using an arbitrary cutoff of 90% or greater overlap with the query sequence. In most cases this corresponded to the sequence with the highest BLAST score. In other cases, the closest match was a shorter target with a higher percent identity. Ambiguous bases in query or target sequences were considered as matching. For queries that produced multiple identical matches, the target with a species name closest to a label ingredient was chosen when possible. A similar procedure was followed for BOLD searches, with the exception that the number of alignment results was 100, which is the maximum allowed. For consistency in reporting, the species of sequences deposited in GenBank and BOLD were used unaltered even though some may be in error or reflect outdated taxonomy.

1. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. Proc. Biol. Sci., 270, 313–321 (2003).
2. Chang, H. T. A taxonomy of the genus Camellia. Acta Sci. Nat. Univ. Sunyatseni, Monogr. Series 1, 1–180 (1981).
3. Chang, H. T. & Bartholomew, A. Camellias. London: Batsford, 211 p. (1984).
4. Ming, T. & Zhang, W. The evolution and distribution of genus Camellia. Acta Botanica Yunnanica 18, 1–13 (1996).
5. Balasaravanan, T., Pius, P. K., Kumar, R. R., Muraleedharan, N. & Shasany, A. K. Genetic diversity among south Indian tea germplasm (Camellia sinensis, C. assamica, and C. assamica spp. Lasiocalyx) using AFLP markers. Plant Sci. 165, 365–372 (2003).
6. Ni, S., Yao, M., Chen, L., Zhao, L. & Wang, X. Germplasm and breeding research of tea plant based on DNA marker approaches. Front. Agric. China 2, 200–207 (2008).
7. Katoh, Y., Katoh, M., Takeda, Y. & Omori, M. (2003) Genetic diversity within cultivated teas based on nucleotide sequence comparison of ribosomal RNA maturease in plastid DNA. Euphytica 134, 287–295 (2003).
8. Wachira, F. R., Powell, W. & Waugh, R. An assessment of genetic diversity of Camellia sinensis (L.) cultivated tea and its wild relatives based on randomly amplified polymorphic DNA and organell-speciﬁc ITS. Heredity 78, 603–611 (1997).
9. Chen, J., Wang, P., Xia, Y., Xu, M. & Pei, S. Genetic diversity and differentiation of Camellia sinensis L. (cultivated tea) and its wild relatives in Yunnan province of China, revealed by morphology, biochemistry, and alloenzyme studies. Genet. Resources Crop. Eval. 52, 41–52 (2005).
10. Singh, D. & Ahuja, P. S. SS rDNA gene diversity in tea (Camellia sinensis (L.) O. Kuntze) and its use for variety identification. Genome 49, 91–96 (2006).
11. Lai, J.-A., Yang, W.-C. & Hsia, J.-Y. An assessment of genetic relationships in cultivated tea clones and native wild tea in Taiwan using RAPD and ISSR markers. Bot. Bull. Acad. Sin. 42, 93–100 (2001).
12. Li, H. L. The domestication of plants in China: ecogeographical considerations. In: Keightley, D. N., editor. The Origins of Chinese Civilization. Berkeley: University of California Press, pp. 21–64 (1982).
13. Ceresa, M. (1996) Diffusion of tea-drinking habit in pre-Tang and early Tang period. Asiatica Venetiana 1, 19–25 (1996).
14. Magoma, G. N., Wachira, F. N., Obanda, M., Imbuga, M. & Agong, S. G. The use of catechins as biochemical markers in diversity studies of tea (Camellia sinensis). Genet. Resources Crops. Eval. 47, 107–114 (2000).
15. Ahmed, S., et al. Pu-erh tea tasting in Yunnan, China: correlation of drinkers' perceptions to phytochemistry. J. Ethnopharmacol. 132, 176–185 (2010).
16. Centers for Disease Control. Anticholinergic poisoning associated with an herbal tea—New York City, 1994. Morb. Mortal. Wkly. Rep. 44, 193–195 (1995).
17. Kumana, C. R. et al. Herbal tea induced veno-occlusive disease: quantification of toxic alkaloid exposure in adults. Gut 26, 101–104 (1985).
18. Toxicology and Clinical Pharmacology of Herbal Products, editor Cupp, M. J. Totowa: Humana Press, 325 p. (2000).