Brief heading: Chemocoding: identification tool for species delimitation in species-rich genera in the tropics.

Title: Chemocoding as an identification tool where morphological- and DNA-based methods fall short: *Inga* as a case study

María-José Endara[1,2], Phyllis D. Coley[1,3], Natasha L. Wiggins[4], Dale L. Forrister[1], Gordon C. Younkin[1], James A. Nicholls[5], R. Toby Pennington[6], Kyle G. Dexter[6,7], Catherine A. Kidner[5,6], Graham N. Stone[5], and Thomas A. Kursar[1,3]

[1]Department of Biology, University of Utah, Salt Lake City, UT 84112-0840, USA
[2]Centro de Investigación de la Biodiversidad y Cambio Climático (BioCamb) e Ingeniería en Biodiversidad y Recursos Genéticos. Facultad de Ciencias de Medio Ambiente, Universidad Tecnológica Indoamérica, Quito EC170103, Ecuador
[3]Smithsonian Tropical Research Institute Box 0843-03092, Balboa, Ancón, Republic of Panamá
[4]School of Biological Sciences, University of Tasmania, Sandy Bay TAS 7001, Australia
[5]Ashworth Labs, Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JY, UK
[6]Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR, UK
[7]School of GeoSciences, University of Edinburgh, Edinburgh EH9 3FF, UK

Correspondence: María-José Endara; Department of Biology, University of Utah, Salt Lake City, UT 84112-0840, USA; majo.endara@utah.edu, tel. +01 (801) 581-7086

Word count: Main body: 5,704; Introduction: 897; Materials & Methods: 1,480; Results: 1,245; Discussion: 1,945; Acknowledgments: 82.
Tables: 1
Figures: 5 (all color)
Supplementary information: Tables: 3; Figures: 9

Summary

- The need for species identification and taxonomic discovery has led to the development of innovative technologies for large-scale plant identification. DNA barcoding has been useful, but fails to distinguish among many species in species-rich plant genera, particularly in tropical regions. Here, we show that chemical fingerprinting, or "chemocoding", has great potential for plant identification in challenging tropical biomes.
- Using untargeted metabolomics in combination with multivariate analysis, we constructed species-level fingerprints, which we define as chemocoding. We evaluated the utility of chemocoding with species that were defined morphologically and subject to next-generation DNA sequencing in the diverse and recently radiated neotropical genus, *Inga* (Leguminosae), both at single study sites and across broad geographic scales.
- Our results show that chemocoding is a robust method for distinguishing morphologically similar species at a single site and for identifying widespread species across continental-scale ranges.
- Given that species are the fundamental unit of analysis for conservation and biodiversity research, the development of accurate identification methods is essential. We suggest that chemocoding will be a valuable additional source of data for a quick identification of plants, especially for groups where other methods fall short.

Key-words: Chemocoding, *Inga,* metabolomics, species identification, tropical forests

Introduction

Cataloguing the world's plant diversity has been a challenge for centuries, and because of accelerated anthropogenic extinctions, the rapid documentation of biodiversity is more critical than ever (Mace, 2004; Valentini *et al.,* 2009; Vernooy *et al.,* 2010; Cristescu, 2014). This is particularly true in tropical rainforests, where the high diversity and co-occurrence of morphologically and ecologically similar congeneric species have presented significant challenges for identification (Gonzalez *et al.,* 2009; Dexter *et al.,* 2010; Kress *et al.,* 2015; Liu *et al.,* 2015). Plant species have typically been identified by botanical experts based on morphological characteristics, or more recently, by sequencing several chloroplast DNA regions or the internal transcribed spacer of nuclear ribosomal DNA (ITS), referred to as "DNA barcoding" (Gemeinholzer *et al.,* 2006; Hollingsworth *et al.,* 2009; Chen *et al.,* 2010; Kress *et al.,* 2010; China Plant BOL Group *et al.,* 2011; Hollingsworth *et al.,* 2011). Here we present evidence that chemical fingerprinting or "chemocoding" can be another tool for species identification in confusing and/or closely related tree species in challenging, hyperdiverse biomes.

No single identification method is without drawbacks. Morphological methods are often labor-intensive, rely upon taxonomic expertise and are the most prone to subjective errors, particularly where phenotypic plasticity and cryptic taxa are prevalent (de Carvalho *et al.,* 2005; Costion *et al.,* 2011). Although DNA barcoding is rapid and straightforward, it can fail to distinguish closely related plant species because of insufficient sequence divergence in standard barcode markers (Kress *et al.,* 2009; Dexter *et al.,* 2010, Liu *et al.,* 2015) and may lead to faulty identifications in genera where species are of recent origin (Razafimandimbison *et al*., 2004, Naciri & Linder, 2015; Pennington & Lavin, 2015). These limitations are often exacerbated in highly diverse tropical systems, as the proportion of species belonging to young, species-rich genera is high (Richardson *et al.,* 2001). For example, in a subtropical Chinese forest, for the 44% of species belonging to genera with more than two species, >50% shared barcoding sequences and could not be distinguished (Liu *et al*., 2015). This led to an overall species resolution of only 67%. A similar problem is encountered in New World tropical rainforests, where DNA barcoding cannot reliably discriminate species within

ecologically important, species-rich genera such as *Inga, Ficus* and *Piper* (Gonzalez *et al.,* 2009; Kress *et al.,* 2009). While standard DNA barcoding for problematic groups can be improved by adding data for additional loci, recently diverged species will always be hard to distinguish using sequence data and no standardized marker sets for such extended DNA barcoding exist.

Many studies assessing diversity in surveys or plots for conservation or basic science rely on identifying all individuals in a plot to species-level, including the large majority of individuals that are without flowers or fruits (Dexter *et al.,* 2010). Plots in the tropics represent a daunting task, and are still faced with problematic identifications despite extremely well-trained botanists. A related problem is the difficulty of achieving uniform species identifications across multiple sites. For example, in a recent, extensive analysis of the identifications for eight genera, the three genera with the highest error rates were *Andira* and *Tachigali* (approximately 50%) and *Inga* (about 40%; Baker et al., 2017). And yet another substantial challenge is correlating the identities to species level for saplings, small trees and adult trees, because ontogenetic changes in leaf morphology may be considerable and juveniles do not bear flowers or fruits. For example, a consortium to understand forest dynamics has established 63 plots around the world where all woody plants >1 cm DBH are mapped and identified, the majority of which are juveniles (www.forestgeo.si.edu). Thus, the issues and errors associated with morphological species identifications of thousands of trees in the tropical forests are serious.

Given that species are a fundamental unit of analysis for conservation, for quantifying biodiversity, and for understanding ecological and evolutionary processes, the development of accurate methods for identifying them is essential. In this paper, we suggest that chemical fingerprinting (here termed chemocoding) can provide an additional identification tool for species identification in a species-rich tropical tree genus, particularly for morphologically confusing or cryptic species. We examine its utility both for distinguishing species within a single site, and for characterizing within-species variation over wider geographic scales. Moreover, chemocoding may be inexpensive enough to allow for every individual tree to be tested.

We test the potential of chemocoding for species identification within *Inga* Mill. (Leguminosae, Mimosoideae) because species in this genus are difficult to distinguish morphologically and show insufficient variation in barcoding sequences (Richardson *et al.,* 2001; Kress *et al.,* 2009; Dexter *et al.,* 2010; Dick & Webb, 2012). *Inga* is one of the most abundant and diverse Neotropical genera in lowland forest communities (Valencia *et al.,* 1994; ter Steege *et al.,* 2013), is widely distributed, and has undergone recent, rapid diversification (Richardson *et al.,* 2001). For *Inga*, genetic and morphological differentiation of closely related species is low and the identification of a species can therefore be difficult.

We propose the use of small, defense-related chemical markers characterized via untargeted metabolomics in combination with multivariate analysis for the construction of a phytochemical, species-level fingerprint, which we define as chemocoding. We evaluate the units defined by chemocoding with those defined morphologically in a recent taxonomic monograph (Pennington, 1997) as well as with those defined using next-generation DNA sequencing data of many hundreds of nuclear genes (Nicholls *et al.,* 2015 & unpubl. data).

Material and Methods

Study sites

Samples were collected at five sites that include a wide range of soils but very similar climates throughout the Amazon and Panama (Fig. 1). Barro Colorado Island is a field station administered by the Smithsonian Research Tropical Institute located in the Panama Canal (9°N 80°W). It is a lowland moist forest with 2649 mm of precipitation a year and 4-month dry season with mean monthly temperatures of 27°C (Leigh, 1999). The other four sites do not have a pronounced dry season. The Nouragues Ecological Research Station, French Guiana (4°N 53°W) is located inside the Nouragues National Reserve on the Guiana Shield. The mean annual precipitation is 2990 mm and the mean annual temperature is 26.3°C (Grimaldi & Riera, 2001). Tiputini Biodiversity Station is located in the eastern lowland Ecuadorian Amazon (0°S 75W), inside the Yasuni Biosphere Reserve. The climate is humid and aseasonal, with an annual precipitation of 3320 mm and an average annual temperature of 26° C (Valencia *et al.,* 2004). Kilometer

41 (KM41, 2°S 59°W) is a field station of the Biological Dynamics of Forest Fragments project located near Manaus, Brazil. The mean annual temperature is 26°C and average annual precipitation is 2651 mm (Radtke *et al.,* 2007). Los Amigos Biological Station is located in the southeastern lowland Peruvian Amazon, in the Madre de Dios Department (13°S 70°W). The mean annual rainfall is between 2700 to 3000 mm. Due to winter cold spells, the daily minimum temperature can drop to less than 10°C; the mean monthly temperature range is from 21 to 26°C (Pitman, 2007). For simplicity in the text, each site will be referred to by the country only.

Study species

We examined saplings of *Inga* because this size class is the most frequently censused for ecological research and also is the size class that can be very difficult to identify. Morphological identifications were based on the most recent taxonomic *Inga* monograph (Pennington, 1997), and made by four researchers (MJE, TAK, PDC and KGD) who have worked in the field identifying *Inga* for about five decades collectively. They consulted with the botanists working in the 50-ha CTFS (Center for Tropical Forest Science) plots in Panama and Ecuador and the plots in Nouragues, French Guiana.

Sampling

We determined the power of chemocoding for discriminating among species and among geographically disjunct populations within species. We included different taxa that are similar in vegetative morphology and are very difficult to identify in the absence of reproductive structures. Our examples include cases in which these coexist at the same site or in two well separated sites. We also included cases of one species that is morphologically uniform in collections from up to four sites (Table 1).

In addition, we tested the ability of chemocoding to correctly distinguish simultaneously between populations of several dozens of species at a regional scale (see Random Forest Analysis, below). More examples can be found in the supplementary information.

Collections

For each species, samples of expanding leaves were collected from five saplings, 0.5 – 4 m in height, in the shaded understory. We focused on expanding leaves as part of a study of plant-herbivore interactions and also because secondary metabolites are at greater concentration during the expansion stage than in leaves that have matured and toughened (Wiggins *et al.,* 2016). For each sapling, we collected leaves that were between 20% and 80% of the average maximum size. Fresh leaves were dried at room temperature with fans and silica gel during 24-48h, transported to the University of Utah, and stored at -20°C. For DNA analysis we typically included one sample per species per site.

Metabolomic analysis

Metabolites were extracted and analyzed following the protocol of Wiggins *et al.,* (2016), specifically designed for secondary metabolites having intermediate polarity. In *Inga*, these are mainly phenolics and saponins. Briefly, 100 mg of ground leaves were extracted in 1.0 ml of extraction buffer (44.4 mM ammonium acetate (pH 4.8): acetonitrile, 60:40, v/v). After extraction for 5 min and centrifugation (13 793 x g) for 5 min, the supernatant was transferred to a glass vial and the extraction repeated. The extracts were diluted fivefold by combining 200 µL of crude extract with 790 µL of acetonitrile:water (60/40, v/v) plus 10 µL of internal standard (1mg mL^{-1} biochanin A in acetonitrile:water, 50/50). Soluble metabolites were analyzed by ultra-performance liquid chromatography coupled to mass spectrometry (UPLC-MS) using an Acquity UPLC® *I-Class* system and a Xevo® G2 Q-ToF MS equipped with LockSpray and an electrospray ionization source (Waters, Milford, MA). Data were collected in negative ionization mode.

Raw data from the UPLC-MS were processed for peak detection, peak alignment and peak filtering using MassLynx (Waters) and the R package XCMS (Smith *et al*., 2006; Tautenhahn *et al.,* 2008; Benton *et al*., 2010). The parameters used were: peak detection method *"centWave"* (ppm=15, peakwidth=c(5,12), snthresh=5); peak grouping method *"density"* (bw=2); retention time correction method *"obiwarp"*; integrate areas of missing peaks method *"chrom"*. XCMS processing was performed for each species independently, with five leaf samples included as replicates. The results obtained by XCMS were post-processed in the R package CAMERA to assign the various ions derived from one compound (termed "features") to that compound (Kuhl *et al.,* 2012).

This uses a defined set of rules for linking the precursor ion with adducts and neutral losses (Table S1 for a list of these). The parameters used were: peak grouping after retention time *"groupFWHM"* (perfwhm*:* 0.8); verify grouping "*groupCorr";* annotate isotopes "*findIsotopes";* annotate adducts *"findAdducts"* (polarity= "negative"). For each case study, the resulting peak tables for each species were combined into a single peak table (*m/z* and retention time for each peak) using the R package metaXCMS (Tautenhahn *et al.,* 2011) with the following parameters: peak filtering: none; *m/z* and retention time tolerance: 0.05 and 12 seconds respectively. Peak tables are stored in MetaboLights (https://www.ebi.ac.uk/metabolights/), a publicly available database.

Statistical Analysis

The variation in metabolites across samples was quantified using unsupervised multivariate methods (no prior classification of samples), which is suitable for metabolomics data. For the first four case studies, we chose methods for data reduction and pattern recognition that group and visualize samples according to their similarities without prior assignment of samples to classes (Bartel *et al.,* 2013).

In order to visualize grouping patterns across samples, we used hierarchical clustering with multiple agglomerative algorithms because this method works well for a limited number of species (classes) and provides statistical power (Embrechts *et al.,* 2013). Peak intensities, or the total ion current (TIC), were normalized by dividing by the sum of the TIC for all features in the chromatogram of a sample. Subsequently, we fitted a hierarchical clustering model to the normalized data with 10000 permutations using the R package pvclust (Suzuki & Shimodaira, 2014). The hierarchical clustering was performed using the Pearson's correlation similarity measure, a routine method adopted for "omics" data (Reeb *et al.,* 2015). The clustering algorithm selection for each analysis was based on the correlation between the original distance matrix and the patristic distance in the hierarchical cluster diagram. Clusters with AU (Approximately Unbiased) p-values of ≥95 % are considered to be strongly supported by the data. For more details see Wiggins *et al.,* (2016).

In addition, to test the overall accuracy of chemocoding to identify samples when presented with a very large number of species (classes) simultaneously, we used

supervised statistical learning methods. Specifically, we chose Random Forest Analysis, which is a powerful classification method for multivariate datasets with many weak predictor variables along with a large number of species (cases). This method has been widely adopted in remote sensing, high dimensional biological data (various "omics), and ecology (Breiman, 2001; Lawrence *et al.,* 2006; Cutler *et al.,* 2007; Cutler *et al.,* 2009). Based on models that we constructed with the metabolomics data, we used Random Forest in order to predict how well samples can be classified to species. For this, a single sample-by-compound matrix was generated using XCMS as described above. A total of 1000 trees was generated for each Random Forest Model and 100 variables were used at each split, which was sufficient to arrive at a model with minimal prediction error (Breiman, 2011). We performed the analyses for 82 species with samples selected from a single site, as well as for 26 species found at two to four sites. Analyses were performed using the randomForest R package (Liaw & Weiner, 2002). R code for all of the analyses is provided in the Supplementary Section (Methods S1-S2).

Next-generation DNA sequence data

To determine the accuracy of our approach, we compared delimitations based on chemocoding with the first resolved phylogeny of *Inga,* accomplished through targeted enrichment and sequencing of 194 loci (259,313 bases; Nicholls *et al.,* 2015 & unpubl. data). Due to *Inga*'s recent, rapid radiation (Richardson *et al.*, 2001), a previous phylogeny with over 6kb of plastid and nuclear DNA sequence did not resolve species-level relationships fully (Kursar *et al.,* 2009).

Results

Case study 1: Two morphologically confusing or cryptic species that are present at the same site

Inga alata Benoist and *Inga pezizifera* Benth are sister species (Nicholls *et al.,* 2015, & unpubl. data; Fig. 2c). The saplings can only be successfully differentiated by expert field workers using subtle differences in the shape of the extrafloral nectaries, the number of leaflets, the number of primary lateral leaf veins, and the color of the expanding leaves (Fig. S1). In addition, they differ in the morphology of their inflorescences (Pennington,

1997), but this feature is not available in the saplings studied by many ecologists. We investigated how chemocoding might be useful to separate these two confusing species in Nouragues, French Guiana, where one species is often found meters away from the other.

Consistent with DNA sequence differences (Fig. 2c), chemocoding accurately determined species limits for five saplings each of the two species. The profiles of secondary metabolites showed visually evident differences between species (Fig. 2a). Hierarchical clustering of UPLC-MS metabolomics data (98% AUP value, Fig. 2b) clustered the samples into two distinct groups, one for each species.

We evaluated four further groups of species that are hard to separate morphologically, and coexist at a single site (available in the supplementary section). For three of these, chemocoding-based separation agreed with DNA sequence differences (Fig. S2: *I. coruscans* & *I. laurina* in Peru, Fig. S3: *I. umbellifera* T50, T72 and T73 in Ecuador (where T numbers correspond to codes for morphotypes in Tiputini, Ecuador), and Fig. S4: *I. chartaceae* & *I. sapindoides* in Ecuador). For the fourth supplemental case, chemocoding found substantial differences in secondary metabolites between two morphotypes of *I. leiocalycina* (T65 and T86, Fig. S5). T86 occurs in terra firme forests and T65 in floodplains in the Ecuadorian Amazon, whereas DNA data placed these two morphotypes into a monophyletic group (Fig. S5d). This may be the result of plasticity in response to differences in habitat (although we found no intermediate morpho- or chemotypes). Alternatively, these may reflect strong selection in the face of gene flow across this environmental gradient or these may be distinct species.

Case study 2: A single species that shows morphological variation within a site

Inga acreana Harms is a widely-distributed species across South America, from the Guyanas to the Amazon Basin in Colombia, Ecuador, Peru and Bolivia (Pennington, 1997; Pennington & Revelo, 1997). In the Tiputini Biological Station in Ecuador, it comprises two distinct types that have the same morphology, but that differ subtly in the color of the expanding leaves (T28 and T56; T numbers are codes for morphotypes; Fig. S1), and co-occur in floodplain habitats. We used chemocoding of five saplings of each of these two leaf variants to determine if they were the same or different chemotypes.

The two morphotypes showed no consistent metabolomic differences (Fig. 3a). Hierarchical clustering models fitted to the UPLC-MS data reveal no separate clusters (Fig. 3b). These results agree with DNA data; the phylogeny from targeted enrichment data placed these accessions representing these two morphotypes into the same, otherwise unstructured, monophyletic group (Fig. 3c). Most likely, the color difference between the two morphotypes is due to a difference in anthocyanin production, a form of intraspecific variation sometimes seen in expanding leaves.

Case study 3: Distinguishing among two morphologically similar species across three sites

Cross-checking identifications amongst morphologically similar tree species that occur at different sites is particularly challenging in tropical forests (Baker *et al.,* 2017). *Inga brachystachys* Ducke and *Inga obidensis* Ducke are closely related (Fig. 4d), with overlapping distributions across the Amazon (Pennington, 1997; Pennington & Revelo, 1997). They are morphologically very similar in the vegetative state, making it problematic to assign accurate species names (Fig. S1).

Chemocoding and hierarchical clustering of five saplings from each of the three sites delimited the samples into two groups, separating the French Guiana and Brazil samples from those collected in Peru (100% AUP value, Fig. 4c). Together, DNA sequence data (Fig. 4c), chemocoding and morphology suggest that samples collected in Brazil and French Guiana are a single species, *I. obidensis,* and that the chemically distinct Peruvian samples may represent a different, as yet unidentified species that, in its vegetative morphology, is similar to *I. brachystachys.*

Case study 4: Identification of a widespread species across its range

In order to assess the variation of chemocoding profiles across geographic space in a widespread species, we collected data on five saplings per population of *Inga auristellae* Harms, a species that occurs across northern South America. These came from four geographically separated populations: French Guiana, Ecuador, Brazil, and Peru (Fig. 1).

The different populations showed consistency in their chemistry across their wide geographic range (Fig. 5a). The hierarchical clustering model shows no significant

differences between *I. auristellae* populations from different geographic areas (94% AUP value, Fig. 5b). The DNA sequences analyses show that although the four populations belong to the same species, there is a strong geographic structure, with an initial split into east vs west Amazonia, and then within each of these groups the samples cluster by population (Fig. 5c). Thus, chemocoding shows consistency in species characterization, despite some evidence of genetic population structure without corresponding structuring of chemistry. A similar pattern was observed for other widespread species: *I. pezizifera* (Fig. S6). A second widespread species, *I. alba*, showed no geographic structure either in chemistry or DNA (Fig. S7).

Case 5: Identification of a large number of species, including comparisons among populations of widespread species

The detailed case studies presented above allow a small number of samples to be grouped by chemical similarity and statistically validated without prior classification of the samples (unbiased). Additionally, we sought to evaluate how often chemocoding could correctly identify samples to species when we include a large number of classes (species). To this end, we use statistical learning techniques in a supervised strategy (prior classification of the samples) to build a model based on metabolomics data for identification of samples to species. Specifically, we used Random Forest to assess the overall accuracy of chemocoding to distinguish between the 82 species of *Inga* that we sampled at the five study sites (five saplings per species per site). Random Forest works by creating many classification trees each trained using random bootstrapped samples from the original metabolomics data. A consensus classification is then chosen based on the majority vote from all trees (Breiman *et al.,* 1984). Out of the five samples per species we iteratively dropped one sample to train the model with four samples and test the predictive accuracy with the fifth sample, such that each sample was used once for testing and four times for training.

The resulting Random Forest model accurately classifies 94% of the 410 individuals representing the 82 *Inga* species (each with representatives from a single site, Table S2). Twenty-six of these species occurred at two to four of the study sites, so we also examined the model's classification accuracy when regional variation across sites

was included. In this case, our analyses correctly classified samples to species 96% of the time (Table S3). In addition, 90% of the time the classification model identified the correct species and site, indicating that there were regional differences in secondary metabolites within a species (Table S3). Overall, these results demonstrate that even in the face of cross-site intraspecific variation, there is sufficiently high interspecific variation to allow unknown samples to be efficiently classified into units that correspond to species as defined by morphology and DNA (Table S2).

Discussion

The need for new tools. The urgent need to catalogue, manage and understand the ecology and evolution of plant diversity has led to the development of innovative tools to improve the discrimination and identification of plant species. Traditional morphological- and molecular-based taxonomic identification methods have proved problematic for species-rich regions and highly species-rich genera (Dick & Webb, 2012; Seberg & Petersen, 2009). And it is precisely these diverse situations where accurate species identifications are most crucial. Alternative new technologies for plant identification in tropical trees include the use of near-infrared (NIR) leaf spectroscopy. However, its potential as a taxonomic tool has not been assessed across broad geographic scales for widespread species (Dugarte *et al.,* 2013; Lang *et al.,* 2013; Baker *et al.,* 2017).

Here, we add chemocoding to the tool box, and show that small, defense-related chemical markers characterized via untargeted metabolomics have great potential for species identification and in providing additional evidence for species delimitation and taxonomic discovery. Chemocoding was a robust method for species identification at a single site and across broad geographic scales, even for *Inga*, where levels of interspecific morphological variation are low and DNA barcoding is ineffective (Kress *et al.,* 2009; Gonzalez *et al.,* 2009; Dexter *et al.,* 2010).

Accuracy of identifications. An effective species identification method must be diagnostic of species (recognizing all populations rather than only sub-specific units or populations), and involve traits that are always present (rather than inducible or otherwise phenotypically plastic). Our analysis of *Inga* shows that entities identified as discrete morphospecies or phylogenetic taxa can indeed show constitutive differences in chemical

401 defenses that result in distinct chemocodes. For example, our results show that
402 chemocoding correctly identified species at a single site and, most importantly, across
403 their ranges (Figs. 2-5, Figs. S2-S9, Table S3). Even with 410 individuals from 82
404 species of *Inga,* with a given species occurring at multiple sites, the Random Forest
405 Analysis based on chemocodes accurately classified 96% of the individuals (Table S2). It
406 also appears to be a robust method for identifying widespread species where intraspecific
407 geographical variation might be problematic. For example, chemocoding of the widely
408 scattered populations of *I. auristellae* resolves them as a single group, which is also
409 resolved as monophyletic in our DNA sequence-based phylogeny (Fig. 5). Nevertheless,
410 it is important to consider that depending on levels of migration, polymorphism, and
411 selection, other outcomes are possible. In particular, some species that are widespread
412 may show significant divergence in chemistry (e.g. *I. alata* (Fig. S8), and *I. marginata*
413 (Fig. S9)). Hence, we caution that the efficacy of chemocode-based identification should
414 be explored in each candidate taxon. However, we conclude that given the abundance and
415 diversity of *Inga* species in neotropical forests, and the difficulty of identifying them
416 using morphological characters (particularly in sterile material), chemocodes provide a
417 valuable taxonomic tool.
418 Chemocoding is unlikely to identify species reliably where major components of
419 plant chemistry show phenotypic plasticity. Rather than being constitutive, chemistry
420 could be age-dependent (ontogeny or tissue age) or inducible by herbivores, pathogens,
421 light, etc. Since this could generate significant within-site variation, we recognize that
422 chemocoding may not work for species where important chemical markers vary.
423 However, studies with several species-rich genera in the tropics have found that inter-
424 specific differences in the defensive metabolome are large relative to intra-specific
425 variation, even considering factors that are recognized as generators of plastic variation
426 such as leaf ontogeny (expanding vs. mature leaves; Sedio *et al.,* 2017, Wiggins *et al.,*
427 2016), light environment (sun vs. shade; Sinimbu *et al.,* 2012; Bixenmann *et al.,* 2016;
428 Sedio *et al.,* 2017), season (dry vs. wet; Sedio *et al.,* 2017), and induction by herbivory
429 (Bixenmann *et al.,* 2016). Even though plasticity may be an issue in some taxa, it does
430 not rule out useful application of chemocoding, but highlights the need to separate
431 diagnostic from phenotypically plastic characters.

Practicality. An identification method also needs to be practical. In other words, it needs to be accurate, rapid and inexpensive, as is the case with DNA barcoding. However, in groups where barcoding using standard markers cannot discriminate among species, sequencing of many genes may be necessary, which can be time-consuming and expensive. In the case of *Inga,* obtaining the resolved phylogeny took several years and hundreds of thousands of dollars (Nicholls *et al.,* 2015). Furthermore, only a few individuals of each species were sequenced. However, as per-base pair prices for sequencing are dropping in next-generation sequencing approaches, discriminating among species based on DNA sequences from many hundreds of loci may become more feasible.

For widespread use of chemocoding, we envision the creation of a public library of reference "chemocodes", analogous to iBOL (ibol.org). In principal, chemocodes could be similar to barcodes in that they employ a limited set of compounds that are both constitutively produced and diagnostic of species. Instead, our method relies on the entire chemical fingerprint (typically a suite of more than 100 compounds) to identify species. Our choice is based on the variation observed single species within and across sites (e.g. *I. auristellae*, Fig. 5), taking into account variation caused by ontogeny.

We have successfully tested chemocoding on taxa from other groups, such as species from the families Euphorbiaceae, Malvaceae, Moraceae, Rubiaceae, Violaceae, among others (data not shown), suggesting that our approach works with groups other than *Inga.* Our data are publicly available (see Material and Methods) allowing others to attempt to develop diagnostic compounds for species identification. Nevertheless, the most challenging issue to address before chemocoding can be widely used, will be the application of our approach across different laboratories. For this, one must ensure that the same metabolic traits are used in all laboratories. In contrast to DNA barcoding, which uses standardized markers across taxa, each species is scored using different traits. At present, we do not know the extent to which chemical fingerprints will differ, depending for example on the exact column used for liquid chromatography or the exact model of mass spectrometer used. To address this issue such that chemocoding can be applied generally may require a more rigorous approach, in particular, the application of tandem MS or MS/MS (instead of simplifying the analysis as suggested below). Another

issue is that, while we used compounds with intermediate polarity for chemocoding of *Inga*, non-polar compounds such as terpenes or highly polar compounds such as non-protein amino acids may work best with other clades.

Because ecological studies based on long-term monitoring plots require the accurate identification of thousands of juvenile and adult trees, we consider here whether chemocoding could be applied to large numbers of samples. Currently, we have used chemocoding rather than DNA-based analyses for classifying over one thousand samples of *Inga* and have found chemocoding to be effective and convenient.

In terms of scaling up, one consideration is that these analyses can be carried out more expediently. The methods used in the present study range from easily accomplished to more complex methods. For example, sample preparation (e.g. drying) in the field and sample extraction in the laboratory are both straightforward. In addition, chemical analysis is largely free of contamination issues, which are a serious concern in DNA barcoding analyses, where specificity of primers may be low (Hollingsworth *et al.,* 2011). Other components could be simplified to streamline the analysis. These include collecting mature instead of expanding leaves. Most often, rainforest plants do not have expanding leaves, restricting chemocoding to a minority of saplings at any given point in time and reducing the utility of chemocoding. We found that mature leaves have most, but not all, of the chemical signals found in expanding leaves (Lokvam *et al.,* 2007; Wiggins *et al.,* 2016), so the use of mature leaves should be feasible. Additionally, we used UPLC with a 150mm column, followed by detection with a high-resolution time-of-flight mass spectrometer, an expensive analysis. To simplify this, we recommend using a shorter, 50mm column, saving solvent and instrument time, followed by detection with a quadrupole mass spectrometer. Single quadrupole, triple quadrupole or ion trap detectors are much less expensive than a time-of-flight. While these provide lower mass resolution, both negative and positive mode data can be obtained in a single run, something that generally is not possible with a time-of-flight spectrometer. Based on our extensive work on *Inga*, only some of which is presented here, most pairs of similar species should be distinguishable using the proposed simplifications. There are some cases where two species have similar chemistry and morphology and the simpler chemical analyses may lump these into a single chemotype. But our experience suggests that these would show

high "within-chemotype" variation, indicating the need for more sophisticated chemical methods that can effectively answer the question at hand. In our hands, chemocoding gave clear results and was practical in terms of time and cost. Using the simplifications suggested above would decrease cost and time. We estimate that manual extraction and automated chromatography and data analysis could take 20 min and $20.00 per sample at the time of writing. In summary, because chemocoding may work in circumstances where barcoding does not, we propose that it presents a novel and practical approach for surveying large number of individuals, possibly thousands of samples.

Conclusions

Our aspiration is not to claim that chemocoding will replace DNA barcoding or morphology-based identification methods, nor that chemocoding can determine species boundaries. Instead we suggest that it is a tool that provides valuable additional source of data to facilitate identification of plant species, especially for groups where traditional methods fall short. Chemocoding may be especially valuable in identifying species in recent radiations where morphological distinctions between species are slight and standard barcode markers do not provide sufficient resolution. However, it could also be used more broadly for species identification in cases where hundreds or thousands of samples need analysis, with the added benefit of providing information on defensive metabolites. In general, it can help by standardizing species names across multiple sites, and even in pin-pointing entities that may be species new to science. Such is the case for our third example (case study 3: *I. obidensis* and *I.*cf *brachystachys* in Brazil, French Guiana and Peru), where chemocoding and DNA suggest that the samples collected in Peru might represent a new species. Our approach is especially amenable for field biologists who work in networks of forest inventory plots since it can consistently distinguish amongst multiple species across geographic space (Figs. 4, 5; Tables S2, S3), and hence, can help in taxonomic integration across plots.

Experience with some taxa may show that chemocodes could help to distinguish groups of individuals that show similar plastic responses to shared abiotic environments or natural enemies (regardless of whether these correspond to species, e.g. Fig. S5). If so, chemocoding could be a very valuable way of dividing individual plants into ecologically

significant sets and improving our understanding of the plant-herbivore adaptive landscape. Given the key role of plant chemistry in many aspects of plant-herbivore-enemy interactions, it may be tremendously valuable to see plant communities in terms of chemotypes of hosts experienced by herbivores (Endara *et al.,* 2017).

The use of metabolites as tools in systematics has a long history and many antecedents (Gibbs, 1974; Smith, 1976; Harborne & Turner, 1984). While this has traditionally been used to investigate evolutionary relationships, based on the presence, absence and distinctive structures of specific classes of secondary metabolites in different groups at all taxonomic levels (Singh, 2016), chemocoding differs in that it is designed to *quantitatively* discriminate species across samples. This uses an unsupervised statistical approach that will likely yield consistent results independently of a priori ideas of species classifications. We see great potential in chemocoding to assist in the inventory of species-rich forests and potentially in the discovery of new species.

Acknowledgments

We thank the Ministry of Environment of Ecuador and the Ministry of Agriculture of Peru for granting the research and exportation permits. Valuable field assistance was provided by Zachary Benavidez, Allison Thompson, Yamara Serrano and Mayra Ninazunta. This work was supported by grants from the National Science Foundation (DEB-0640630 and DIMENSIONS of Biodiversity DEB-1135733), and Nouragues Travel Grants Program, CNRS, France to T.A.K. and P.D.C., and the Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación del Ecuador (SENESCYT) to M.J.E.

Author contributions

M.J.E., P.D.C., N.L.W., D.L.F and T.A.K. designed and conducted the research. M.J.E., N.L.W. and D.L.F. designed and performed the data analysis. G.C.Y. contributed to the metabolomic analysis. J.A.N., R.T.P., K.G.D., C.A.K. and G.N.S. contributed the next-generation DNA sequence data. M.J.E., P.D.C., N.L.W., D.L.F., J.A.N., R.T.P., K.G.D., C.A.K., G.N.S. and T.A.K. wrote the manuscript.

References

Baker TR, Pennington RT, Dexter KG, Fine PVA, Fortune-Hopkins H, Honorio EN, Huamatunpa-Chuquimaco I, Klitgård BB, Lewis GP, de Lima HC, *et al.* 2017. Maximizing synergy among tropical plant systematists, ecologists, and evolutionary biologists. *Trends in Ecology & Evolution* **32:** 258-267.

Bartel J, Krumsiek J, Theis FJ. 2013. Statistical methods for the analysis of high-throughput metabolomics data. *Computational and Structural Biotechnology Journal* **4**: e201301009.

Benton P, Want EJ, Ebbels TMD. 2010. Correction of mass calibration gaps in liquid chromatography–mass spectrometry metabolomics data. *Bioinformatics* **26**:2488–2489.

Bixenmann RJ, Coley PD, Weinhold A, Kursar TA. 2016. Higher herbivore pressure favors constitutive over induced defense. *Ecology and Evolution* **6:** 6037-6049.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and Regression Trees*. Monterey, CA, USA: Wadsworth.

Breiman L. 2001. Random forests. *Machine Learning* **45**: 5–32.

Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, *et al*. 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE* **5**: e8613.

China Plant iBOL Group, Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, Chen ZD, Zhou SL, Chen SL, *et al.* 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences USA* **108**: 19641–19646.

587 Costion C, Ford A, Cross H, Crayn D, Harrington M, Lowe A. 2011. Plant DNA
588 barcodes can accurately estimate species richness in poorly known floras. *PLoS ONE* **6**:
589 e26841.
590
591 Cristescu ME. 2014. From barcoding single individuals to metabarcoding biological
592 communities: towards an integrative approach to the study of global biodiversity. *Trends*
593 *in Ecology & Evolution* **29**: 566-571.
594
595 Cutler DR, Edwards KH, Beard AC, Kyle TH, Jacob G, Joshua JL. 2007. Random forests
596 for classification in ecology. *Ecology* **88**: 2783–2792.
597
598 Cutler A, Cutler DR, Stevens JR. 2009. Tree-based methods. In: Li X, Xu R, eds. *High*
599 *dimensional data analysis in cancer research.* New York, USA: Springer-Verlag, 83-101.
600
601 de Carvalho MR, Bockmann FA, Amorim DS, de Vivo M, de Toledo-Piza M, Menezes
602 NA, de Figueiredo JL, Castro R, Gill AC, McEachran JD. 2005. Revisiting the taxonomic
603 impediment. *Science* **307**: 353.
604
605 Dexter KG, Pennington TR, Cunningham CW. 2010. Using DNA to assess errors in
606 tropical tree identifications: How often are ecologists wrong and when does it matter?
607 *Ecological Monographs* **80**: 267-286.
608
609 Dick CW, Webb CO. 2012. Plant DNA barcodes, taxonomic management and species
610 discovery in tropical forests. In: Kress WJ, Erickson DL, eds. *DNA barcodes: Methods*
611 *and protocols.* Totowa, NJ: Humana Press, 379-383.
612
613 Dugarte FM, Higuchi N, Almeida A, Vicentina A. 2013. Species spectral signature:
614 Discriminating closely related species in the Amazon with Near-Infrared Leaf-
615 Spectroscopy. *Forest Ecology and Management* **291:** 240-248.
616

Endara MJ, Coley PD, Ghabash G, Nicholls JA, Dexter KG, Donoso DA, Stone GN, Pennington RT, Kursar TA. 2017. Coevolutionary arms race versus host defense chase in a tropical-herbivore plant system. *Proceedings of the National Academy of Sciences USA*, **114**: E7499–E7505.

Embrechts MJ, Gatti CJ, Linton J, Roysam B. 2013. Hierarchical Clustering for large data sets. In: Georgieva P, Mihaylova L, Jain LC, eds. *Advances in intelligent signal processing and data mining: Theory and applications.* New York, USA: Springer, 197-233.

Gemeinholzer B, Oberprieler C, Bachmann K. 2006. Using GenBank data for plant identification: possibilities and limitations using the ITS 1 of Asteraceae species belonging to the tribes Lactuceae and Anthemideae. *Taxon* **55**: 173-187.

Gibbs RD. 1974. *Chemotaxonomy of flowering plants.* Montreal and London: McGill's Queen's University Press.

Gonzalez MA, Baraloto C, Engel J, Mori SA, Pétronelli P, Riéra B, Roger A, Thébaud C, Chave J. 2009. Identification of Amazonian trees with DNA barcodes. *PLoS ONE* **4**: e7483.

Grimaldi C, Riéra B. 2001. Geography and climate. In: Bongers F, Charles-Dominique, P, Forget PM, Théry, M, eds. *Nouragues: Dynamics and plant-animal interactions in a Neotropical rain forest*. New York, USA: Springer-Science + Business Media.

Harborne JB, Turner BL. 1984. *Plant chemosystematics.* London, UK: Academic Press.

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* **270**: 313-321.

Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, *et al.* 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences USA* **106**: 12794–12797.

Hollingsworth PM, Graham SW, Little DP. 2011. Choosing and using a plant DNA barcode. *PLoS ONE* **6**: e19254.

Kress WJ, Erickson DL, Jones FA, Swenson NG, Perez R, Sanjur O, Bermingham E. 2009. Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences USA* **106**: 18621-18626.

Kress WJ, Erickson DL, Swenson NG, Thompson J, Uriarte M, Zimmerman JK. 2010. Advances in the use of DNA barcodes to build a community phylogeny for tropical trees in a Puerto Rican forest dynamics plot. *PLoS ONE* **5**: e15409.

Kress WJ, García-Robledo C, Uriarte M, Erickson DL. 2015. DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution* **30**: 25-35.

Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S. 2012. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry* **84:** 283-289.

Kursar TA, Dexter KG, Lokvam J, Pennington RT, Richardson JE, Weber MG, Murakami ET, Drake C, McGregor R, Coley PD. 2009. The evolution of antiherbivore defenses and their contribution to species coexistence in the tropical tree genus *Inga*. *Proceedings of the National Academy of Science USA* **106**:18073–18078.

Lang C, Costa FRC, Camargo JLC, Durgante M, Vicentini A. 2015. Near infrared spectroscopy facilitates rapid identification of both young and mature Amazonian tree species. *PLoS ONE* **10**: e0134521.

678 Lawrence RL, Shana DW, Roger LS. 2006. Mapping invasive plants using hyperspectral
679 imagery and Breiman Cutler Classifications (randomForest). *Remote Sensing of*
680 *Environment* **100**: 356–362.
681
682 Leigh EG. 1999. *Tropical forest ecology: a view from Barro Colorado Island*. New York,
683 USA: Oxford University Press.
684
685 Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* 2: 18-
686 22.
687
688 Liu J, Yan HF, Newmaster SG, Pei N, Ragupathy S, Ge XJ. 2015. The use of DNA
689 barcoding as a tool for the conservation biogeography of subtropical forests in China.
690 *Diversity of Distributions* **21**: 188-199.
691
692 Lokvam J, Clausen TP, Grapov D, Kursar TA. 2007. Galloyl depsides of tyrosine from
693 young leaves of *Inga laurina. Journal of Natural Products* **70:** 134-136.
694
695 Mace GM. 2004. The role of taxonomy in species conservation. *Philosophical*
696 *Transactions of the Royal Society of London Series B, Biological Sciences* **359**: 711–719.
697
698 Martucci MEP, De Vos RCH, Carollo CA, Gobbo-Neto L. 2014. Metabolomics as a
699 potential chemotaxonomical tool: application in the genus *Vemonia* Schreb. *PLoS ONE* **9**:
700 e93149.
701
702 Naciri Y, Linder HP. 2015. Species delimitation and relationships: the dance of the seven
703 veils. *Taxon* **64**: 3–16.
704
705 Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, Dexter
706 KG, Stone GN, Kidner CA. 2015. Using targeted enrichment of nuclear genes to increase
707 phylogenetic resolution in the neotropical rainforest genus *Inga* (Leguminosae:
708 Mimosoideae). *Frontiers in Plant Science* **6:** 1-20.

Pennington TD. 1997. *The Genus Inga: Botany.* UK: The Royal Botanic Gardens, Kew.

Pennington TD, Revelo N. 1997. *El género Inga en el Ecuador.* UK: The Royal Botanic Gardens, Kew.

Pennington RT, Lavin M. 2015. The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *New Phytologist* **210:** 25-37.

Pitman N. 2007. An overview of the Los Amigos watershed, Madre de Dios, southeastern Peru. October 2007 version of an unpublished report available from the author at npitman@amazonconservation.org

Radtke MG, Da Fonseca CRV, Williamson GB. 2007. The old and young Amazon: Dung beetle biomass, abundance and species diversity. *Biotropica* **39**: 725–730.

Razafimandimbison SG, Kellogg EA, Bremer B. 2004. Recent origin and phylogenetic utility of divergent ITS putative pseudogenes: a case study from Naucleeae (Rubiaceae). *Systematic Biology* **53**: 177-192.

Reeb PD, Bramardi SJ, Steibel JP. 2015. Assessing dissimilarity measures for sample-based hierarchical clustering of RNA sequencing data using plasmode datasets. *PloS ONE* **10**: e0132310.

Richardson JE, Pennington RT, Pennington TD, Hollingsworth PM. 2001. Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science* **293**: 2242–2245.

Seberg O, Petersen G. 2009. How many loci does it take to DNA barcode a *Crocus*? *PloS ONE* **4**: e4598.

Sedio BE, Rojas Echeverri JC, Boya CA, Wright JS. 2017. Sources of variation in foliar secondary chemistry in a tropical forest tree community. *Ecology* **98:** 616-623.

Singh R. 2016. Chemotaxonomy: a tool for plant classification. *Journal of Medicinal Plant Studies* **4:** 90-93.

Sinimbu G, Coley PD, Lemes MR, Lokvam J, Kursar TA. 2012. Do the antiherbivore traits of developing leaves in the Neotropical tree *Inga paraensis* (Fabaceae) vary with light availability? *Oecologia* **170**: 669-676.

Smith PM. 1976. *The chemotaxonomy of plants.* London, UK: Edward Arnold.

Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry* **78**: 779–787.

Suzuki R, Shimodaira H. 2014. pvclust: hierarchical clustering with p-values via multiscale bootstrap resampling. *Bioinformatics* **22**:1540–1542.

Tautenhahn R, Böttcher C, Neumann S. 2008. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9**:504.

Tautenhahn R, Patti GJ, Kalisiak E, Miyamoto T, Schmidt M, Lo FY, McBee J, Baliga NS, Siuzdak G. 2011. MetaXCMS: second-order analysis of untargeted metabolomics data. *Analytical Chemistry* **83**: 696-700.

ter Steege H, Pitman NCA, Sabatier D, Baraloto C, Salomão RP, Guevara JE, Phillips OL, Castilho CV, Magnusson WE, Moline JF, *et al.* 2013. Hyperdominance in the Amazonian tree flora. *Science* **342**: 1243092.

771 Valencia R, Balslev H, Miño GP. 1994. High tree α diversity in Amazonian
772 Ecuador. *Biodiversity and Conservation* **3**: 21–28.
773
774 Valencia R, Foster RB, Villa G, Condit R, Svenning J-C, Hernández C, Romoleroux K,
775 Losos E, Magård E, Balslev H. 2004. Tree species distributions and local habitat
776 variation in the Amazon: large forest plot in eastern Ecuador. *Journal of Ecology* **92:**
777 214-229.
778
779 Valentini A, Pompanon F, Taberlet P. 2009. DNA barcoding for ecologists. *Trends in*
780 *Ecology & Evolution* **24**: 110-117.
781
782 Vernooy R, Haribabu E, Muller MR, Vogel JH, Hebert PDN, Schindel DE, Shimura J,
783 Singer GAC. 2010. Barcoding life to conserve biological diversity: beyond the taxonomic
784 imperative. *PLoS Biology* **8**: e1000417.
785
786 Wiggins NL, Forrister DL, Endara MJ, Coley PD, Kursar TA. 2016. Quantitative and
787 qualitative shifts in defensive metabolites define chemical defense investment during leaf
788 development in *Inga,* a genus of tropical trees. *Ecology and Evolution* **6**: 478-492.
789
790 **Figure legends**
791
792 **Figure 1.** Study sites: 1) Barro Colorado, Panamá, 2) Nouragues, French Guiana, 3)
793 Tiputini, Ecuador, 4) KM41 near Manaus, Brazil and, 5) Los Amigos, Perú.
794
795 **Figure 2.** Case study 1: Two morphologically confusing or cryptic species that are
796 present at the same site: *I. alata* and *I. pezizifera* in French Guiana. a) Total ion
797 chromatograms showing relative intensities of peaks from the LC-QToF-MS in negative
798 mode, b) Hierarchical cluster dendrograms based on relative abundances of UPLC-MS
799 metabolites. The numbers in red above each branch point are the Approximately
800 Unbiased confidence levels, these indicate the probability that the samples below that
801 point are a cluster. Clusters with values of 95 signify P=0.05, indicating that these

clusters are strongly supported by the data, c) Clade containing *I. alata* and *I. pezizifera* adapted from a resolved phylogeny based on next-generation DNA sequence data (Nicholls *et al.,* 2015 & unpubl. data). Numbers in black represent bootstrap support values. Values greater than 95 indicate that the clade is strongly supported by the data.

Figure 3. Case study 2: A single species that shows morphological variation within a site: *I. acreana* T28 and *I. acreana* T56 in Ecuador. a) Total ion chromatograms showing relative intensities of peaks from the LC-QToF-MS in negative mode, b) Hierarchical clustering based on relative abundances of UPLC-MS metabolites. The numbers in red above each branch point are the Approximately Unbiased confidence levels, these indicate the probability that the samples below that point are a cluster. Clusters with values of 95 signify P=0.05, indicating that these clusters are strongly supported by the data, c) Clade containing *I. acreana* T28 and *I. acreana* T56 adapted from a resolved phylogeny based on next generation DNA sequence data (Nicholls *et al.,* 2015 & unpubl. data).

Figure 4. Case study 3: Distinguishing among two morphologically similar species across three sites: *I. obidensis* from Brazil and French Guiana and *I. brachystachys* from Peru. a) Total ion chromatograms showing relative intensities of peaks from the LC-QToF-MS in negative mode, b) Hierarchical clustering based on relative abundances of UPLC-MS metabolites. The numbers in red above each branch point are the Approximately Unbiased confidence levels, these indicate the probability that the samples below that point are a cluster. Clusters with values of 95 signify P=0.05, indicating that these clusters are strongly supported by the data, c) Clade containing *I. obidensis* and *I.* cf *brachystachys* adapted from a resolved phylogeny based on next-generation DNA sequence data (Nicholls *et al.,* 2015 & unpubl. data).

Figure 5. Case study 4: Identification of a widespread species across its range: *I. auristellae.* a) Total ion chromatograms showing relative intensities of peaks from the LC-QToF-MS in negative mode, b) Hierarchical clustering based on relative abundances of UPLC-MS metabolites. The numbers in red above each branch point are the

Approximately Unbiased confidence levels, these indicate the probability that the samples below that point are a cluster. Clusters with values of 95 signify P=0.05, indicating that these clusters are strongly supported by the data, c) Clade containing *I. auristellae* adapted from a resolved phylogeny based on next-generation DNA sequence data (Nicholls *et al.,* 2015 & unpubl. data).

Supporting Information

Fig. S1. Photographs of the *Inga* species for each study case.

Fig. S2. Two morphologically confusing or cryptic species that are present at the same site: *Inga coruscans* and *I. laurina* in Peru.

Fig. S3. Two morphologically confusing or cryptic species that are present at the same site: *Inga umbellifera* and *I. microcoma* in Ecuador.

Fig. S4. Two morphologically confusing or cryptic species that are present at the same site: *Inga chartacea* and *I. sapindoides* in Ecuador.

Fig. S5. Morphological and chemical variation within a site: *Inga leiocalycina* T65 and *I. leiocalycina* T86 in Ecuador.

Fig. S6. Identification of a widespread species across its range: *Inga pezizifera* in Panama, French Guiana and Brazil.

Fig. S7. Identification of a widespread species across its range: *Inga alba* in French Guiana, Brazil and Peru.

Fig. S8. Identification of a widespread species across its range: *Inga alata* in French Guiana, Ecuador and Peru.

Fig. S9. Identification of a widespread species across its range: *Inga marginata* in Panama, French Guiana, Ecuador and Peru.

Table S1. Mass difference rules used for peak annotation in negative mode with CAMERA.

Table S2. Species accuracy classification from Random Forest Analyses for *Inga* species with representatives from a single site.

Table S3. Species accuracy classification from Random Forest Analyses for *Inga* species that occurred at two to four of the study sites.

Methods S1. R code for LC/MS raw data pre-processing in XCMS

Methods S2. R code for Random Forest Analysis

895

Table 1. Study species

Figures	Case study	Species and sites
Figure 2	Two morphologically confusing species within a site	*Inga alata* Benoist, and *Inga pezizifera* Benth, French Guiana
Figure 3	Morphological variation within one species at a site	*Inga acreana*[1] Harms, Ecuador
Figure 4	Two morphologically similar species across sites	*Inga* cf. *brachystachys* Ducke, and *Inga obidensis* Ducke, French Guiana, Brazil and Peru
Figure 5	Identification of a widespread species across its range	*Inga auristellae* Harms, Ecuador, Brazil and Peru
Figure S1	Two morphologically confusing species within a site	*Inga coruscans* Humb. & Bonpl. ex Willd., and *Inga laurina* (SW.) Willd., Peru
Figure S2	Two morphologically confusing species within a site	*Inga microcoma* Harms and *Inga umbellifera*[2] (Vahl) Steud., Ecuador
Figure S3	Two morphologically confusing species within a site	*Inga chartacea* Poepp. and *Inga sapindoides* Willd., Ecuador
Figure S4	Morphological and chemical variation within on species at a site	*Inga leiocalycina*[3] Benth., Ecuador

Table 1. Continued…

Figure	Case study	Species and site
Figure S5	Identification of a widespread species across its range	*Inga alata* Benoist, French Guiana, Ecuador and Peru
Figure S6	Identification of a widespread species across its range	*Inga pezizifera* Benth, Panama, French Guiana and Brazil
Figure S7	Identification of a widespread species across its range	*Inga alba* (SW.) Willd., French Guiana, Brazil and Peru
Figure S8	Identification of a widespread species across its range	*Inga marginata* Willd. Panama, French Guiana, Ecuador and Peru

[1]*I. acreana* includes two morphotypes in Ecuador: T28 and T56.

[2] *I. umbellifera* includes two morphotypes in Ecuador: T50 and T73.

[3]*I. leiocalycina* includes two morphotypes in Ecuador: T65 and T86.