MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2014) doi: 10.1111/1755-0998.12256

Automated DNA-based plant identification for large-scale biodiversity assessment

ANNA PAPADOPOULOU,* DOUGLAS CHESTERS,† INDIANA CORONADO,‡ GISSELA DE LA CADENA,* ANABELA CARDOSO,* JAZMINA C. REYES,‡ JEAN-MICHEL MAES,§ RICARDO M. RUEDA‡ and JESÚS GÓMEZ-ZURITA*

*Animal Biodiversity and Evolution, Institut de Biologia Evolutiva (CSIC-Univ. Pompeu Fabra), 08003 Barcelona, Spain, †Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China, ‡Herbario y Jardín Botánico Ambiental, Universidad Nacional Autónoma de Nicaragua, León, Nicaragua, §Museo Entomológico de León, León, Nicaragua

Abstract

Rapid degradation of tropical forests urges to improve our efficiency in large-scale biodiversity assessment. DNA barcoding can assist greatly in this task, but commonly used phenetic approaches for DNA-based identifications rely on the existence of comprehensive reference databases, which are infeasible for hyperdiverse tropical ecosystems. Alternatively, phylogenetic methods are more robust to sparse taxon sampling but time-consuming, while multiple alignment of species-diagnostic, typically length-variable, markers can be problematic across divergent taxa. We advocate the combination of phylogenetic and phenetic methods for taxonomic assignment of DNA-barcode sequences against incomplete reference databases such as GenBank, and we developed a pipeline to implement this approach on large-scale plant diversity projects. The pipeline workflow includes several steps: database construction and curation, query sequence clustering, sequence retrieval, distance calculation, multiple alignment and phylogenetic inference. We describe the strategies used to establish these steps and the optimization of parameters to fit the selected *psbA-trnH* marker. We tested the pipeline using infertile plant samples and herbivore diet sequences from the highly threatened Nicaraguan seasonally dry forest and exploiting a valuable purpose-built resource: a partial local reference database of plant *psbA-trnH*. The selected methodology proved efficient and reliable for high-throughput taxonomic assignment, and our results corroborate the advantage of applying 'strict' tree-based criteria to avoid false positives. The pipeline tools are distributed as the scripts suite 'BAGpipe' (pipeline for Biodiversity Assessment using GenBank data), which can be readily adjusted to the purposes of other projects and applied to sequence-based identification for any marker or taxon.

Keywords: BAGpipe script suite, DNA barcoding, dry tropical forest, Nicaragua, *psbA-trnH*, taxonomic assignment

Received 18 December 2013; revision received 17 February 2014; accepted 22 February 2014

Introduction

As tropical forests are rapidly decaying due to human activity and climate change, it is becoming increasingly critical to improve our efficiency in recording biodiversity and developing informed conservation strategies for these hyperdiverse ecosystems. The wide application of the DNA-barcoding approach (Hebert *et al.* 2003) has provided a valuable tool, which can greatly accelerate species identification and facilitate large-scale biodiversity assessment and environmental monitoring (Janzen *et al.* 2009; Yu *et al.* 2012). However, there are some limitations to this approach, including finding optimal DNA barcodes and robust methodologies for sound taxonomic identification based on these sequences. Especially in plants, DNA barcoding has been far more challenging than in animals (Fazekas *et al.* 2009; Cowan & Fay 2012). A range of chloroplast genes have been proposed by different authors as potential plant barcodes, but none of them have proved as successful as the *cox1* marker for animals, either in terms of sequence recovery and quality or in terms of species discrimination power (Hollingsworth *et al.* 2011). The CBOL Plant Working Group (2009) proposed *rbcL* and *matK* as the core two-loci plant barcode, with the *psbA-trnH* intergenic region being the next favoured option, as a supplementary barcoding marker. The *rbcL* barcode is easy to amplify and sequence across vascular plants but lacks variability

Correspondence: Jesús Gómez-Zurita, Fax: +34 93 221 1011; E-mail: j.gomez-zurita@ibe.upf-csic.es

to distinguish among closely related species, while the *matK* region provides high discriminatory power but unsatisfactory recovery rates (Kress *et al.* 2009; Hollingsworth *et al.* 2011). The *psbA-trnH* region is both universally amplifiable and one of the most variable chloroplastic intergenic spacers (Shaw *et al.* 2005) showing high discrimination success in most land plant groups (Kress *et al.* 2009; Pang *et al.* 2012). However, it raises concerns regarding its high length variation (Chase *et al.* 2007; Kress & Erickson 2007), the presence of intraspecific microinversions associated with palindromes (Whitlock *et al.* 2010; Jeanson *et al.* 2011) and sequencing problems related to mononucleotide repeats (Fazekas *et al.* 2008; Devey *et al.* 2009; but see Fazekas *et al.* 2010). Several authors discussed the performance of these markers in different taxonomic groups and geographical settings (Steven & Subramanyam 2009; Jeanson *et al.* 2011; Bruni *et al.* 2012). In the few studies that specifically targeted the tropical forest flora, *psbA-trnH* was consistently shown to outperform the other two markers when used on its own or in combination with other loci (González *et al.* 2009; Kress *et al.* 2009; Costion *et al.* 2011; Parmentier *et al.* 2013; Tripathi *et al.* 2013). We do not necessarily advocate *psbA-trnH* as a replacement for the standard plant binary DNA barcode. However, considering that database incompleteness is an issue for all loci, including the pair *rbcL* and *matK*, and that each locus is affected by different limitations, the potential of *psbA-trnH* or other length-variable markers as suitable candidates for taxonomic identification should not be neglected.

Most commonly applied methods for plant identification are based on a 'best-match' criterion (Meier *et al.* 2006), or similarly on top BLAST hits (Altschul *et al.* 1997), which heavily depend on the comprehensiveness of the database used (Koski & Golding 2001; Ross *et al.* 2008; Berger *et al.* 2011). The ensuing discussion about marker performance generally assumes nearly complete taxonomic coverage and a good representation of intraspecific diversity in the reference database (Parmentier *et al.* 2013). However, the level of completeness in the relevant public sequence databases (such as GenBank or BOLD) is still unsatisfactory, especially for the understudied and hyperdiverse tropical flora, and building exhaustive databases may take several decades (Parmentier *et al.* 2013). This is currently limiting the applicability of the method to very few well-studied tropical forest sites or research stations, such as the Barro Colorado Island in Panama where comprehensive DNA-barcode reference libraries have been constructed (Kress *et al.* 2009) and employed successfully for ecological studies (Jones *et al.* 2011). Otherwise, a purpose-built local reference database is required, which has only been attempted so far for relative small-scale tropical forest study sites (0.1–50 hectare plots; González *et al.* 2009; Costion *et al.* 2011; Parmentier *et al.* 2013).

In the interest of scaling-up DNA-based identification procedures for biodiversity assessment across broader tropical forest regions, it is necessary to enhance the representation of the relevant flora in sequence databases, but it might also be particularly important to select methods for sequence identification that are robust to sparse taxon sampling. Comparisons among different algorithms (BLAST, genetic distances, and tree-based methods) for species identification show that, given an incomplete reference database, only tree-based methods, coupled with a distance threshold, can protect against false positives (Ross *et al.* 2008). In large-scale tropical biodiversity assessment, where species-level identifications might often be infeasible, it is worth paying attention to methods that can efficiently assign sequences to higher taxonomic ranks, in the absence of any conspecific or even congeneric sequence in the reference set. This approach has been widely developed and applied in microbial metagenomics (von Mering *et al.* 2007; Alonso-Alemany *et al.* 2014), where the problem of sparse taxon sampling in the reference database can be particularly severe. In this context, phylogenetic methods have outperformed in accuracy algorithms based on pairwise sequence similarity (Munch *et al.* 2008; Berger *et al.* 2011).

These concerns have not been fully appreciated by the plant DNA-barcoding community, as exemplified by the BOLD Identification System (Ratnasingham & Hebert 2007), which currently employs BLAST searches and pairwise distances for taxonomic assignment. (Conversely, animal *cox1* identification combines sequence similarity methods with distance tree construction.) Similarly, Liu *et al.*'s (2011) PTIGS-IdIt web application, which was specifically developed to facilitate plant identification using the *psbA-trnH* intergenic region, is exclusively based on distances, and the resulting species identifications are deemed reliable only when an identical sequence is included in the reference database. To our knowledge, there is no publicly available tool to facilitate automated plant DNA sequence identification, while explicitly taking into account the incompleteness of public databases.

Based on these considerations, we advocate the incorporation of phylogenetic methods as a standard tool in routine analyses of plant DNA barcodes, especially when reference databases are incomplete, as expected in large-scale studies in tropical forests. However, we acknowledge that phylogenetic methods might not be always optimal for discriminating among recently separated taxa (Austerlitz *et al.* 2009), while they are labour intensive and computationally demanding for high-throughput species identification, and further complicated for markers exhibiting length variation and other problems related to multiple sequence alignment across distant taxa (e.g. *psbA-trnH*; Chase *et al.* 2007). Here, we develop a pipeline to facilitate and speed up the taxonomic assignment of large numbers of *psbA-trnH* sequences, by combining homology searches, pairwise distances and phylogenetic tools to circumvent alignment issues affecting this marker. Moreover, in contrast to others, our approach exploits indel information, as coding of indels in *psbA-trnH* sequences improves discrimination among closely related species (Costion *et al.* 2011; Liu *et al.* 2012). We exemplify our approach by focusing on the Mesoamerican seasonally dry tropical forest (SDTF), one of the world's most threatened biomes, suffering from extreme levels of fragmentation and anthropogenic disturbance, and requiring urgent conservation actions (Janzen 1988; Miles *et al.* 2006; Griscom & Ashton 2011). Neotropical dry forest patches have received considerably less attention from ecologists and conservationists than their neighbouring rain forests (Janzen 1988; Sánchez-Azofeifa *et al.* 2005), and their flora is massively under-represented in sequence databases. Their importance has been recently recognized, and there is an increased scientific interest to develop conservation initiatives for the remaining Neotropical SDTF fragments (Quesada *et al.* 2009; Linares-Palomino *et al.* 2010; Portillo-Quintero & Sánchez-Azofeifa 2010; Griscom & Ashton 2011). In this context, it is extremely important to establish a DNA-based identification framework for the Mesoamerican SDTF flora, which can greatly accelerate and improve large-scale inventories especially when reproductive organs are not available for accurate morphological identification, which is very common during field surveys (Dexter *et al.* 2010; Parmentier *et al.* 2013). Moreover, DNA-based plant identification is now permitting large-scale inventories of trophic interactions between herbivores and their host plants (Jurado-Rivera *et al.* 2009; Valentini *et al.* 2009; García-Robledo *et al.* 2013), incorporating knowledge about food-web structure in biodiversity assessment and conservation planning. With these ideas in mind, we developed the first purpose-built sequence reference database for the angiosperm flora of the Pacific side of Nicaragua, assembled to enhance the representation of this ecoregion in public sequence databases. Moreover, we have generated sequences from infertile plant samples and diet sequences amplified from herbivore beetles from the same forest fragments, to evaluate the efficiency and utility of the pipeline, compare the performance of phylogenetic vs. distance-based taxonomic assignment and assess the contribution of the local reference database to the efficiency of the identification.

Materials and methods

Sampling, DNA extraction and sequencing of plants

Angiosperm samples were collected from 15 sites focusing on SDTF fragments and surrounding transitional areas along the Pacific region of Nicaragua and high altitudes of the northern province of Estelí, between October 2011 and August 2012 (Table S1, Supporting information). For each sample, a small amount of leaf tissue was cut off and kept in silica gel for DNA work, and a voucher was deposited in the Herbarium of the Universidad Nacional Autónoma (UNAN, León, Nicaragua).

DNA was extracted from 624 plant samples. Small pieces of frozen leaf tissue (approximately 0.25 cm^2) were ground for 1–2 min at 50 Hz using 5-mm stainless steel beads in a TissueLyser LT (Qiagen, Heidelberg) and following the manufacturer's instructions. Total genomic DNA was extracted following the 'mini' protocol of the DNeasy Plant Mini Kit (Qiagen), but extending lysis incubation time at 65 °C to 2–3 h.

A fragment of the cpDNA *psbA-trnH* gene was PCR-amplified using primers psbAF (Sang *et al.* 1997) and trnH2 (Tate & Simpson 2003) from 1 μL of gDNA in a 25 μL reaction (at final concentrations: reaction buffer 1x, 3 mM $MgCl_2$, 0.2 mM dNTPs, 0.5 U *Taq* polymerase and 0.2 $\mu\mu$ for each primer) with an initial step of 3 min at 94 °C, 35 cycles of 30 s at 94 °C, 30 s at 55 °C and 1 min at 72 °C and a final step of 10 min at 72 °C. PCR products were purified using ammonium acetate and isopropanol and sequenced in both directions using the BigDye Terminator v3.1 Cycle sequencing kit (Applied Biosystems, Foster City, CA, USA). Sequences were assembled and edited using Geneious Pro 5.3.6 (Drummond *et al.* 2010) and submitted to the European Nucleotide Archive (EMBL-EBI) under Accession nos HG963487-HG964039.

Automated construction of a psbA-trnH database

For the purpose of this study, we established an efficient way (Fig. 1) to retrieve all sequences homologous to a marker of choice and associated taxonomic information from the latest release in public databases, giving them the same orientation and the length of the barcode region of interest (critical steps for successful multiple sequence alignment). Sequences were retrieved from GenBank either based on similarity searches under a range of conditions or according to gene annotation (entries including the strings 'psbA-trnH' or 'trnH-psbA' in the gene, product or definition fields), and results were compared using a custom Perl script. An important advantage of employing similarity searches for sequence retrieval is that the resulting pairwise alignments can be used to

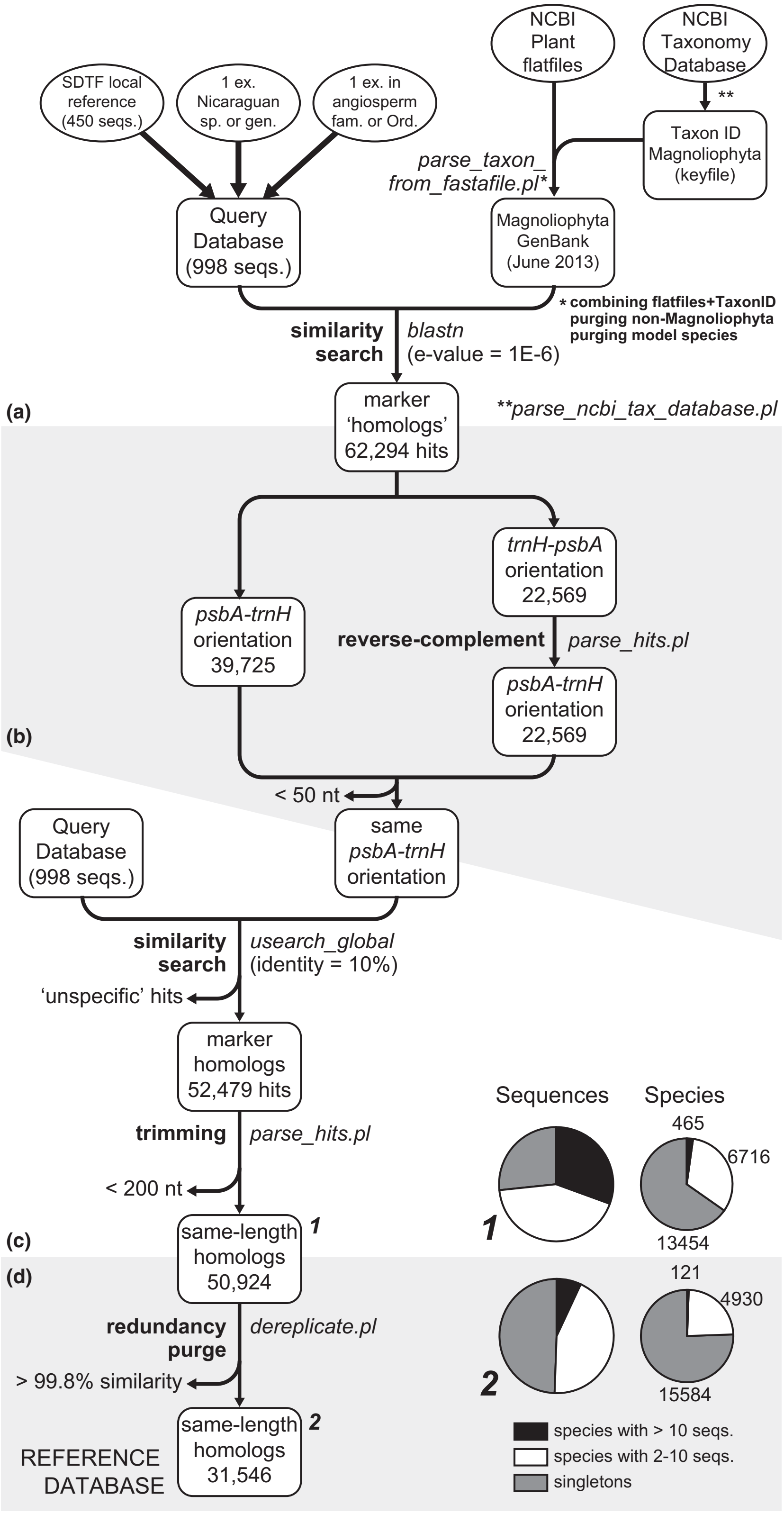
SDTF local reference (450 seqs.)
1 ex. Nicaraguan sp. or gen.
1 ex. in angiosperm fam. or Ord.
NCBI Plant flatfiles
NCBI Taxonomy Database
**
Taxon ID Magnoliophyta (keyfile)
parse_taxon_ from_fastafile.pl*
Query Database (998 seqs.)
Magnoliophyta GenBank (June 2013)
* combining flatfiles+TaxonID purging non-Magnoliophyta purging model species
similarity search
blastn (e-value = 1E-6)
(a)
marker 'homologs' 62,294 hits
**parse_ncbi_tax_database.pl
trnH-psbA orientation 22,569
psbA-trnH orientation 39,725
reverse-complement
parse_hits.pl
psbA-trnH orientation 22,569
(b)
< 50 nt
Query Database (998 seqs.)
same psbA-trnH orientation
similarity search
usearch_global (identity = 10%)
'unspecific' hits
marker homologs 52,479 hits
trimming
parse_hits.pl
< 200 nt
same-length homologs 50,924
1
(c)
(d)
redundancy purge
dereplicate.pl
> 99.8% similarity
same-length homologs 31,546
2
REFERENCE DATABASE
Sequences
Species
465
6716
13454
1
121
4930
15584
2
■ species with > 10 seqs.
□ species with 2-10 seqs.
▨ singletons

Fig. 1 Database construction pipeline workflow. The flowchart is presented with data obtained from its implementation to the construction of a *psbA-trnH* reference database for sclerophyll deciduous tropical forests in Nicaragua, and relevant scripts are indicated (see main text for details). The general procedure can be split in four successive stages: (a) fishing public DNA sequence databases for homologues based on similarity searches; (b) curating data for identical sequence orientation; (c) purging data based on global identity and clipping of sequence ends to similar length; and (d) removing of intraspecific redundancy. The inset pie charts show the purging effects of removing redundant data for overstudied taxa at the expense of increasing taxonomic singletons.

automatically check sequence orientation and overlap between query and database sequences, which allows trimming the retrieved sequences to the desired length. Moreover, it avoids errors and ambiguities in GenBank annotations. However, the parameters used for similarity searches can also affect significantly the efficiency of the process. We compared the efficiency of the *blastn* algorithm (Altschul *et al.* 1997; Zhang *et al.* 2000) under a range of E-value cut-offs (1E-15 to 1E-5), as implemented in the NCBI Blast toolkit (Camacho *et al.* 2009), against the *usearch* algorithm (Edgar 2010) with global alignment (*usearch_global*) and a range of identity threshold values (0.3–0.8), as implemented in USEARCH 4.2.66. (We used an older release of USEARCH, as the current 6.0.307 version had memory limitations and generated less satisfactory results.)

The above similarity searches were based on a query file which included as much taxonomically relevant information as possible. (i) All our Nicaraguan SDTF database sequences that covered the full length of the barcode region. (ii) One representative GenBank sequence per species or genus known to exist in the region, according to the flora of Nicaragua (Stevens *et al.* 2001) and the records of the UNAN Herbarium. GenBank sequences were retrieved based on organism name and gene annotation and were subsequently assessed for homology with the barcode region, checked for taxonomic ID errors and manually oriented and trimmed to the region of interest. And (iii) one sequence per angiosperm order and major family that was not represented by any of the above entries. Overall, the query file included 998 sequences. We also tested how recovery efficiency depended on comprehensiveness of the set of query sequences, using reduced sets.

After sequence retrieval from GenBank, reversing and trimming when necessary, the last step in the automated construction of the reference database involved removal of redundant sequences, that is, identical or nearly identical sequences of the same species. This step can save huge amounts of computing time in subsequent steps. This was achieved using a *blastn* all-against-all search among sequences with the same taxonomic ID, followed by single linkage clustering and sequence removal where similarity was above a certain threshold. A range of similarity threshold values (99–100%) were tested to select the value removing most sequences for taxa extensively studied at the population level, while retaining intraspecific variation, which is fundamental for sequence identification.

Optimization of automated taxonomic assignment of psbA-trnH sequences

We developed an automated strategy for taxonomic assignment of unidentified *psbA-trnH* sequences based on both phenetic and phylogenetic inference methods. Each step in the procedure was approached using alternative strategies, and a number of custom Perl scripts were developed to implement and combine the selected methods. To assess the reliability of each step and to optimize the parameter values for default *psbA-trnH* searches, we used our Nicaraguan SDTF plant sequence database as queries and contrasted the results with their known taxonomy. In general terms, the pipeline workflow considers four stages (Fig. 2): (i) clustering of query sequences, (ii) retrieval of related sequences from the database, (iii) multiple sequence alignment and phylogenetic inference and (iv) parsing taxonomic assignment from genetic distances and trees.

To maximize the efficiency of the identification procedure, we presumed that query sequences should not be processed individually or together as a single group, especially if they diverged substantially from each other. Instead, we considered an initial procedure where query sequences were clustered into groups based on similarity, which were processed individually. This clustering step speeds up the alignment and phylogenetic inference steps of the pipeline, especially in a multiprocessor system where the process can be easily parallelized, and it critically ensures that alignment problems of divergent length-variable *psbA-trnH* sequences will be minimized. Different sequence clustering strategies were tried on the SDTF control sequences to choose the best strategy considering: (i) *blastn* against *usearch_global* algorithms to calculate per cent identities among all pairs of sequences; (ii) clustering under 'nearest', 'average' or 'furthest neighbour' criteria (and a range of linkage fraction values between these, that is, 0.1–0.8); and (iii) using different clustering thresholds (60–90%). Several combinations of the above parameters were tested, and in each case, 20 of the resulting groups were picked at random and aligned using MAFFT 7.043b (Katoh *et al.* 2002; Katoh & Standley 2013) with the E-INS-i algorithm.

Sequences related to each query group were retrieved from the database using similarity searches with the same parameters and identity threshold as the ones selected in the previous step. Sequence length variation, very common for *psbA-trnH* even between closely related species, substantially affects sequence retrieval in some cases, because USEARCH and BLAST identity values are calculated based on the total number of alignment columns (including nonterminal gaps), that is, long indels greatly reduce the identity score between two otherwise very similar sequences of closely related taxa. Thus, we tried to lower the identity threshold and introduce an additional step after sequence retrieval involving pairwise distance calculation. We avoided the typically employed uncorrected p-distances, which do not take gaps into account, as indel information for this marker can be very important to distinguish among closely related species (Costion *et al.* 2011; Liu *et al.* 2012); instead, we calculated p-distances counting each string of nonterminal gaps as a single event. We used these distances for distance-based identification, as described

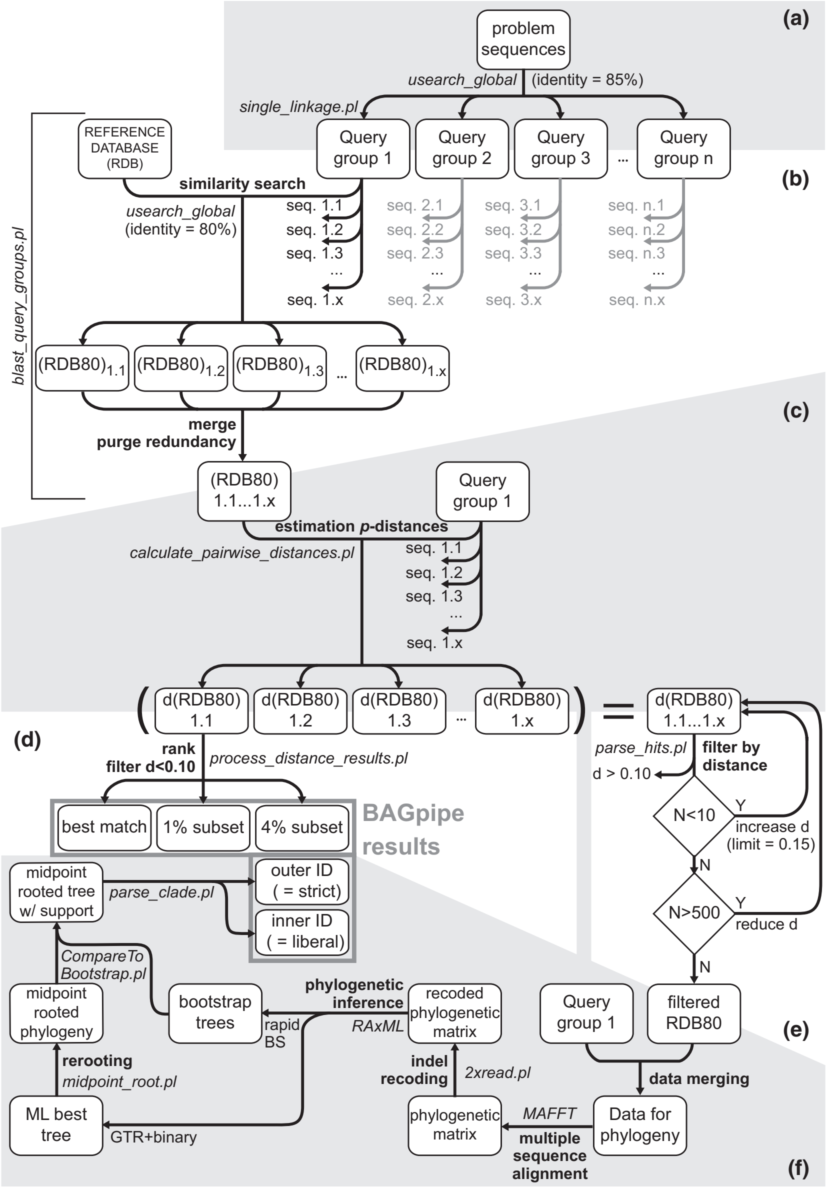

Fig. 2 Schematic workflow of the procedure implemented in the pipeline for Biodiversity Assessment using GenBank data (BAGpipe). The pipeline and parameters used are optimized for angiosperm *psbA-trnH* sequence data. The procedure includes six well-differentiated stages: (a) splitting data in query groups of sequences with high global similarity; (b) finding subsets of sequences from the reference database with global similarity to each of the sequences in the respective query group and merging data into a single subset for subsequent comparisons; (c) estimating genetic divergences for each of the sequences in the corresponding query group with these in the previous subset; (d) ordination of sequences in the previous subset to generate the phenetic results of BAGpipe by parsing relevant information from the ordinations; (e) using genetic distances to generate size-manageable subsets of sequences for efficient phylogenetic analyses; and (f) merging of query groups with their respective size-purged subset of genetically close sequences for standard maximum-likelihood phylogenetic inference (with assessment of bootstrap support).

below, but also as an additional criterion to decide which of the retrieved sequences would be included in the multiple alignments.

Once the final set of database sequences was selected, these were aligned against the query sequences of the corresponding query group using either MUSCLE (Edgar 2004) or MAFFT 7.043b with the E-INS-i, Q-INS-i or FFT-NS-2 algorithm. The resulting alignments were checked for obvious misaligned regions and were used for phylogenetic inference with RAxML 7.2.8 (Stamatakis 2006; Stamatakis *et al.* 2008), the leading programme for large-scale maximum-likelihood analysis, as it is faster and yields better likelihood scores than other comparable methods (Stamatakis 2006; Liu *et al.* 2012). ML inference relied on 20 independent searches starting from different stepwise addition parsimony trees, and clade support was assessed both by standard and rapid bootstrapping algorithms with 100 pseudoreplicates. We explored alternative indel coding strategies using the options available in RAxML: (i) gaps coded as fifth state and data analysis using a multistate GTR model; (ii) simple indel coding (SIC; Simmons & Ochoterena 2000) and analysing the resulting binary matrix as an additional partition under a binary model, known to perform well for modelling indel patterns (Berger & Stamatakis 2012); and (iii) treating gaps as missing data (default option). Our dynamic procedure to obtain input data prevents from identifying outgroups, and the resulting trees were subsequently rooted either using (i) midpoint rooting (i.e. rooting in the middle of the longest tip-to-tip path) or (ii) a recently added rerooting option in RAxML 7.7.1, which roots the tree at the branch that best balances the left and right subtree lengths (sum over branches in the subtree).

Distance-based species identification was assessed applying criteria that are commonly employed in the DNA-barcoding literature. For instance, we applied the 'best-match' criterion (Meier *et al.* 2006) and tested a range of predefined divergence thresholds (0.5–5%). The BOLD Identification System (Ratnasingham & Hebert 2007) uses 1% divergence for animal *cox1* sequences, but an analogous threshold has not been optimized for the *psbA-trnH* marker. We used the Nicaraguan SDTF sequences with conspecific sequences in GenBank to identify a suitable threshold for this data set.

Tree-based taxonomic assignment was based on a 'strict' criterion (*sensu* Ross *et al.* 2008), requiring the query sequence to subtend at least one node into a supported clade (bootstrap >70%) exclusively consisting of conspecifics (or congenerics, cotribal, confamilials, for higher taxonomic ranks). Additionally, we assessed the suitability of a 'liberal' tree-based criterion (*sensu* Ross *et al.* 2008), considering supported sister or unresolved relationships with conspecifics as useful information about the closest taxon to the query.

Pipeline development and availability

Each one of the pipeline steps and relevant parameters were optimized as described above. Custom Perl scripts were developed or modified to implement the selected methods in combination with previously available software for homology searches, clustering, alignment and phylogenetic inference. Further scripts were developed to parse phylogenetic trees automatically, process results linking them with GenBank taxonomy and produce user-friendly outputs. The pipeline is presented as a set of scripts and commands that can be replicated and customized by other users, allowing for the process to be run on any taxonomic and genetic data set. The full pipeline, including detailed descriptions of the scripts and instructions on how to use them, is freely available as a compact and structured package named BAGpipe (*pipeline for Biodiversity Assessment using GenBank data*: http://www.ibe.upf-csic.es/SOFT/Softwareanddata.html and http://sourceforge.net/users/dchesters).

Pipeline test: Evaluating the importance of the local reference database

When the pipeline was developed and test sequences of known taxonomy were run through successfully, we applied it to a set of unidentified plant sequences obtained from Nicaraguan samples from the same SDTF localities as the reference database, but with uncertain or unknown ID (mostly infertile samples). These sequences were run through the sequence identification part of the pipeline against: (i) the automatically compiled database from GenBank and (ii) a combined database including these sequences and our custom Nicaraguan SDTF database sequences. We compared the results from both analyses to assess the contribution of the local database to sequence identification.

Pipeline test: Inference of the diet of herbivore insects

We also tested the potential of the DNA-based taxonomic assignment tool to identify plant sequences from ingested plant tissue by herbivore insects. We sampled leaf beetles (Chrysomelidae) in 14 of the Nicaraguan SDTF localities by beating and sweeping vegetation. The specimens were immediately stored in 100% ethanol in the field. From these, we selected four abundant and easily recognizable species in the subfamily Cassidinae with known dietary habits (Table S2, Supporting information): (i) *Brachycoryna pumila* Guérin-Méneville, reported on *Malvastrum* and *Sida* (Malvaceae); (ii) *Heterispa vinula* (Erichson), reported from a wide range of plant taxa in the Malvaceae (*Apeiba* sp., *Guazuma ulmifolia, Sida* spp., *Triumfetta* sp.) as well as *Indigofera* sp. (Fabaceae); (iii) *Parorectis rugosa* (Boheman), recorded on *Physalis maxima* (Solanaceae); and (iv) *Physonota alutacea* Boheman, known to be a specialist of *Cordia* spp. (Boraginaceae).

In all, 79 leaf beetle specimens in these four species were subject to nondestructive whole-specimen DNA extractions using the DNeasy Blood and Tissue Kit (Qiagen), thus also obtaining DNA from ingested plant tissue. PCR amplification of the *psbA-trnH* fragment from insect DNA extractions used the same primers and mix composition as above with a touchdown protocol consisting of 3 min at 94 °C, 16 cycles of decreasing (60–43 °C) annealing temperature, 27 cycles with constant annealing at 42 °C (30 s) and a final 10 min step at 72 °C. When the amplification was weak, PCR products were reamplified using custom internal primers (psbA-Int2: CTCATAACTTCCCTCTAGAYYTAGC; trnH-Int1: GCCTTGATCCACTTGGCYAC) and fewer PCR cycles at higher annealing temperature: 3 min at 94 °C, 12 cycles annealing at 63 °C for 30 s (30 s at 94 °C, 1 min at 72 °C), and 10 min at 72 °C for final elongation. If multiple bands were obtained, coamplified PCR products were individually excised from the agarose gel and an aliquot used for reamplification using the same set of internal primers as above and 18 cycles of the previous PCR protocol.

Putative diet sequences obtained from the extracted leaf beetle specimens (ENA Accession nos HG964040-HG964098) were processed with the automated identification pipeline against the combined database including

GenBank sequences and our Nicaraguan SDTF reference database.

Results and discussion

Assemblage of a local psbA-trnH reference database: newly generated SDTF data

Of 624 DNA-extracted plant samples from Nicaraguan SDTF, only 554 were successfully amplified and sequenced. Poor sample preservation (46 samples), homopolymer issues (19 PCR products), apparent contamination problems (three samples) and heterologous amplifications (two samples) were the specific causes for failure. An additional 4% of the sequences were mildly affected by mononucleotide repeat regions, and we included them in the data sets after trimming the affected regions and replacing them with missing data for the analyses. Edited sequence length ranged from 148 to 876 nt (214–876 nt if only full-length sequences were considered). Of 554 samples sequenced, 450 voucher specimens were identified with confidence to belong to 437 species and were used to assemble our local reference database (Table S1, Supporting information). The other 104 samples remained unidentified as voucher specimens usually lacked reproductive organs and were identified secondarily using the automated procedure by comparisons with the reference database.

Of the 437 known plant species in our reference database, 75% were absent in public sequence databases for the *psbA-trnH* locus, while 22% were new taxa for GenBank (as for June 2013). The remaining 109 species were already available with at least one *psbA-trnH* sequence, but not from Nicaragua. The latter provided a chance to evaluate the importance of sampling local populations for species identification. For each of these 109 species (114 sequences in total), p-distances were calculated between the newly generated Nicaraguan sequence and its most similar conspecific GenBank sequence (using Needleman–Wunsch pairwise alignment, counting each string of gaps as a single event, excluding terminal gaps or ambiguities). Up to 60% of the Nicaraguan sequences were less than 1% divergent from their conspecific GenBank sequences (24% of them being identical); however, 17% were more than 3% divergent, including four cases that showed >10% divergence from their conspecifics (Fig. S1, Supporting information). It would be important to assess whether the estimated genetic divergences are correlated with geographical distance from Nicaragua; unfortunately most NCBI records do not include geographical information. We identified seven species (*Caesalpinia eriostachys, Croton niveus, Ficus insipida, F. obtusifolia, F. maxima, Guarea glabra*, and *Maclura tinctoria*) where increased intraspecific divergence was mostly owing to microinversions associated with palindromes, as inferred by *blastn* comparisons with their GenBank conspecifics and the EMBOSS *palindrome* algorithm (Rice *et al*. 2000). Pang *et al*. (2012) had reported already these inversions in *Ficus insipida* and some species of *Caesalpinia,* and they reoriented them for analysis. While reorienting inversions may be a common practice, we find it inappropriate in our case due to ambiguity in resolving their correct orientation prior to phylogenetic analyses, and realizing that these inversions may be fixed in populations and represent significant evolutionary events, we thus opted to keep them in the data. Finally, we cannot neglect the possibility that some high intraspecific distances are an artefact due to taxonomic misidentifications either in our samples or in the GenBank entries (see below).

The assembled reference database represents a valuable resource for our research programme, but it also critically enhances ecological and evolutionary research for the entire scientific community. So far, the Nicaraguan angiosperm flora had been represented in public sequence databases by a handful of barcode region sequences (three species each for *rbcL* and *psbA-trnH*, and one for *matK*), while 24% of the Nicaraguan SDTF angiosperm species (approximately 300 taxa) were represented by sequences from other parts of the world (or with unknown geographical origin). Our data more than double the representation of SDTF species in public sequence databases and incorporate local genetic variants for 29% of the available species.

Assemblage of a local psbA-trnH reference database: fishing GenBank

This stage of the procedure is schematized in Fig. 1a. A custom query file with SDTF data and selected sequences from GenBank was used to retrieve all available homologous and taxonomically relevant data from the public databases. The *blastn* algorithm with a relatively permissive E-value cut-off (1E-6) retrieved the maximum percentage of *psbA-trnH* annotated sequences from the NCBI database (98.85% of approximately 48 370 sequences; June 2013), while the *usearch* algorithm with the least stringent threshold tried left out about 7.7% of the annotated sequences. However, *blastn* searches also retrieved a large number of spurious (obviously nonhomologous) hits, such as 2500 full-chromosome sequences longer than 10 000 nt. There were 556 *psbA-trnH* annotated sequences which were not recovered by any of the similarity searches. A closer look at each of these nonretrieved sequences revealed that they corresponded mostly to cases that were not truly homologous to the fragment of interest or there was some other error in the GenBank submission, and we opted to exclude these sequences (Appendix S1, Supporting information). Repeating the homology searches with reduced query sets, for example not covering all angiosperm orders, resulted in missing many more of the *psbA-trnH* annotated sequences. On the other hand, removing some of the redundant sequences per genus or family from the query file did not alter the results significantly; however, we opted to keep all the Nicaraguan sequences in the query file to ensure that we retrieved all relevant taxa and, most critically, that they would be trimmed correctly.

Conversely, sequence retrieval based on annotation was a few thousand sequences short compared to similarity searches. A subset of 3200 sequences lost despite presenting high similarity with the query sequences (E-value cut-off: 1E-25) and an overlap of at least 300 nt with the region of interest were excluded owing to several causes: 1500 sequences containing 'psbA' and/or 'trnH' in their definition but not referring explicitly to the *psbA-trnH* intergenic region; 800 chloroplast sequences containing neither 'psbA' nor 'trnH' in their definition; 360 misspelling cases; 300 whole chloroplast genome sequences; and 260 entries using an alternative description of the *psbA-trnH* region ('psbA-tRNA-His', 'trnH(GUC)-psbA', etc.). Overall, these results highlight the importance of using homology searches for sequence retrieval, instead of relying on database annotations.

Assemblage of a local psbA-trnH reference database: polishing mined data

The *psbA-trnH* data obtained from GenBank were highly heterogeneous in orientation, sequence length, marker coverage and taxonomic redundancy, and thus required curation before being amenable to analysis. The corresponding steps in the global procedure are schematized in Fig. 1b–d. BLAST alignments are less informative when trimming database sequences to the extent of the queries. The local alignment procedure typically generated discontinuous tracts because of the hypervariable regions in the *psbA-trnH* intergenic spacer. Each homologue in the database was hit by a number of queries, although due to local alignment vagaries the particular start and end positions often differed between queries. Given multiple trimming possibilities for each sequence, two alternatives were assessed: trimming according to the longest hit or to the left-most and right-most positions over all hits. When BLAST results were used, the latter retained untrimmed some of the long nonhomologous chromosome sequences with multiple unspecific hits (more than 4% of the sequences were longer than 10 000 nt), while the former tended to overtrim even homologous sequences and to produce several very short fragments (approximately 20% of the retrieved sequences were shorter than 200 nt). However, when the USEARCH global alignments were used instead, the longest hit trimming option gave much more satisfactory results (only 2% of the retrieved sequences were shorter than 200 nt, while less than 0.4% were longer than 10 000 nt).

The *blastn* algorithm proved more efficient in sequence retrieval (Fig. 1a), while the *usearch_global* algorithm performed better the trimming step; therefore, we decided to combine both to get improved overall results. Running *blastn* under the selected conditions retrieved a total of 62 294 hits from the NCBI database (June 2013 release), 39 725 being in the *psbA-trnH* direction. All other sequences, in the *trnH-psbA* orientation according to the *blastn* results, were reversed and complemented (Fig. 1b). A second similarity search was run against the BLAST-retrieved sequences using the *usearch_global* algorithm with a very low identity threshold (0.1) and the same query file. This step removed 15% of the NCBI sequences, including most unspecific hits (only 3% of the long chromosome sequences found by the initial BLAST search remained in the database), but none of the *psbA-trnH* annotated sequences. The global alignments produced by USEARCH were used for trimming the sequences to the extent of the queries. After trimming, only sequences longer than 200 nt were retained in the database, which were in total 50 924 (Fig. 1c). Retaining shorter sequences appeared to cause problems in subsequent pipeline steps, as many of the short fragments corresponded exclusively to conserved coding regions of the *psbA* or the *trnH* genes, resulting in very low, misleading distances.

The sequences retained belonged to 20 635 species (inset Fig. 1), but nearly one-third (15 641) belonged to just 465 species, which were represented by more than ten sequences each, while there were 31 species with more than 100 sequences each. Conversely, 13 454 species were represented by a single sequence. Filtering redundancy with a 99.8% sequence similarity threshold, corresponding to approximately 1 nt difference between sequences, reduced significantly the size of the database to 31 546 sequences (June 2013 release). The number of over-represented taxa dropped to 121 species with more than ten sequences (7% of the total) and only five species with more than 50 sequences; the number of singleton species increased to 15 584. Decreasing the similarity threshold to 99% reduced the database to about half of its original size, with only 33 species being represented by more than ten sequences, but it also increased further the number of singletons to 17 125. Considering the importance of including multiple representatives per species and retaining intraspecific variation within the database for sequence identification, we kept the 99.8% threshold as default, which already resulted in a greatly reduced and overall more balanced database in terms of taxonomic coverage than the original, without producing too many extra singletons (Fig. 1d). However, this value (filter_threshold in *dereplicate.pl*) can be tuned depending on the marker and the taxonomic representation of the group of interest in public databases.

Clustering of query sequences for analysis: lessons from reference SDTF data

For a fast-evolving marker such as the *psbA-trnH* interspacer and the level of evolutionary divergence covered here, the angiosperms, query sequences had to be aggregated into alignable clusters. Applying *blastn* identity scores to cluster the test SDTF sequences of known taxonomy consistently produced a high percentage of very problematic, taxonomically incoherent groupings, which could not be aligned satisfactorily. This was partly because these identity scores are based on local pairwise alignments, thus a bad guide for clustering and especially tricky for our marker, containing both conserved and hypervariable regions. Alternatively, using *usearch_global* for calculating identities, followed by clustering under the same parameters, produced much more sensible groupings (sequences belonging to the same genus, family or order), which could be readily aligned in most cases. Consequently, the *usearch_global* algorithm was employed for this step, empirically establishing an average neighbour clustering (linkage fraction = 0.5) with an identity threshold = 85%, the lowest threshold that produced consistently satisfactory alignments. Increasing these values further produced an even greater number of clusters and singletons, which was not desirable either. Applying the selected settings, the 450 sequences of the Nicaraguan database were clustered into 198 query groups, including 84 clusters of 2–17 sequences and 114 singletons (Fig. 2a). All subsequent steps were performed individually for each of the query groups iteratively using a loop structure (Fig. 2 develops schematically the process for one such hypothetical group); nonetheless, we intended to optimize the threshold values across all groups.

A *usearch_global* search with default 85% identity threshold retrieved different numbers of sequences from the database for each of the 198 query groups, from nil to 2897. These large differences could be attributed mainly to variation of the *psbA-trnH* marker differing greatly across taxonomic groups, so that a uniform threshold unavoidably produced very imbalanced sets of sequences. Additionally, the representation in GenBank is uneven across taxa. Both extremes are problematic, as the failure to retrieve homologues impedes sequence identification, while the excess of similar sequences might considerably slow down multiple alignment and phylogenetic inference steps. These observations imply that different thresholds may be employed across taxa, but this requires some knowledge a priori on the taxonomy of the query sequences or some initial exploratory searches before running a large number of sequences through the entire pipeline. Three of the four query groups that retrieved >1000 homologue sequences belonged to Liliopsida, which is in accordance with *psbA-trnH* showing relatively lower variation in monocots compared to other angiosperms (e.g. Pang *et al.* 2012). The fourth such query group belonged to the dicot family Asteraceae, which also shows relatively low levels of variation for this marker, and is abundantly represented in GenBank. Conversely, there were 37 query groups belonging to several eudicot families, which only retrieved between 0 and 2 homologue sequences from our database, because of their poor representation in GenBank and an apparent high indel-length variation between the query and the most similar database sequences. To account for long indels, we lowered the identity threshold for sequence retrieval to 80% (Fig. 2b) and, as an additional step, we filtered the obtained sequences based on p-distances and with a relatively stringent threshold ($d < 0.10$) to improve multiple alignments (Fig. 2c). This strategy produced a better outcome compared to a single clustering step based on an identity threshold = 85–90%, as the latter removed some sequences with long indels even if they were closely related to the queries and were thus useful for their identification. In most dicot groups examined, all sequences retrieved within a 10% p-distance threshold belonged to the same family, while increasing the threshold to 12% or more retrieved other families too. Therefore, instead of using a uniform threshold across all groups, it seemed a sensible option to adjust the thresholds for data sets retrieving a particularly large or an extremely low number of homologues (Fig. 2e). To reduce significantly the seven biggest data sets (to <500 sequences), we gradually applied a more stringent threshold, which reached as low as 3% for the largest Liliopsida data sets, while we gradually relaxed it up to 15% for 60 of the query groups that retrieved < 10 sequences.

As a by-product of our careful examination of small taxonomic subsets of GenBank data, we could identify 53 sequences with presumably erroneous taxonomic annotation. They showed high similarity to orders or families different from the one suggested in their annotation; thus, they were removed from the reference database, as they affected the next steps of automated identification (Table S3, Supporting information). Nonetheless, this problem highlights an unavoidable limitation of any automated identification procedure that depends on taxonomic annotations from GenBank sequences (Nilsson *et al.* 2006).

Use of sequence clusters for distance-based taxonomic identification

Based on a 'best-match' criterion, 60% of the Nicaraguan SDTF test sequences of known taxonomy and having conspecifics in GenBank were correctly identified at the species level. An additional 33% hit only at the genus level (26% a different species of the same genus; 7% several species of the same genus, including the correct one). The remaining 7% did not match the correct genus as a first hit, but a confamilial species in a different genus, even though the correct genus appeared further down in the list of best hits, or in two cases both the correct species and a noncongeneric species produced the best match. The relatively high percentage of failure in identification at the species level correlates with high intraspecific variation in some of these species, that is, high divergence between Nicaraguan sequences and their conspecifics from other geographical sources (Fig. S1, Supporting information), and thus, it might indicate a poor performance of this marker when lacking a local reference database, but it could also be partly due to identification errors (in GenBank accessions and/or our samples). The SDTF sequences that were identified correctly at the species or genus levels using the 'best-match' criterion above were subsequently used to investigate the effect of different distance thresholds. Using a single threshold, the highest percentage of correct identifications at the species and genus levels was obtained for thresholds 1% and 2% (Figs 2d and 3a), but the rate of success did not exceed 30% at the species level (or 66% at the genus level). Indeed, uniform thresholds are not expected to be successful for species identification across angiosperms, as we showed that variation of *psbA-trnH* differs greatly across taxonomic groups. Therefore, we decided to use two thresholds, selecting as default the combination that maximized the percentage of correct identifications for our particular data set, which was 1% and 4% (Fig. 3a). However, these values do not have universal applicability and should be adapted for each individual data set.

Decisions for matrix assemblage and automated tree-based taxonomic assessment

Each group of closely related *psbA-trnH* sequences with correct size and excluding redundancy was the basis for tree-based taxonomic inference (Fig. 2f). Data were aligned using several strategies, and the Q-INS-i algorithm, which considers RNA secondary structure information (Katoh & Toh 2008), generally provided the most satisfactory alignments. However, it is only feasible for data sets with less than 200 sequences and had to be disregarded, together with other more exhaustive but slower algorithms, for the pipeline. Instead, the E-INS-i algorithm generally provided good multiple alignments for the *psbA-trnH* region, without being particularly slow for the selected size of data sets (< 500 sequences), representing a good trade-off between alignment accuracy and speed. Alternatively, MUSCLE and FFT-NS-2 were both much faster than E-INS-i, with MUSCLE providing less problematic alignments than FFT-NS-2, and it should be preferred when speed is of primary interest and/or bigger data sets are analysed.

Each data matrix was used for ML phylogenetic inference and evaluation of clade support using the rapid bootstrapping algorithm under a GTRCAT model, which proved more effective than standard bootstrapping in yielding very similar clade support values in considerably reduced time, especially for the larger data sets. In these analyses, treating indels as fifth state proved highly unsuitable for the *psbA-trnH* marker as indels caused long branches even between conspecific sequences that only differed by a single indel event. On the other hand, treating gaps as missing data failed to differentiate between closely related species differing mainly by indels. Simple indel coding increased resolution and nodal branch support (Simmons *et al.* 2007; Egan & Crandall 2008; Luan *et al.* 2013); therefore, an indel-recoding step was incorporated in the pipeline, and RAxML searches were performed using a partitioned model (GTR for the DNA partition and binary for the indel partition).

Rooting strategy (midpoint vs. RAxML rooting) only affected 17 cases of tree-based identification using the 450 SDTF test sequences, and in most (71%), the inference based on the midpoint-rooted tree was preferable (correct identification at a lower taxonomic rank) and was thus selected as default method for the pipeline.

By testing the tree-based procedure on the 450 taxonomically identified SDTF test sequences, we could quantify how many correct identifications as well as how many 'false positives' were inferred at each taxonomic rank. The 'strict' tree-based taxonomic assignment (Fig. 3b) allowed few identifications at the species level (1%). This was largely due to the very limited number of species with multiple sequences in the database, but the trade-off between computational demands and retaining genetic diversity recommended purging the latter, although as seen above, it can be tuned to increase intraspecific diversity and the number of strict identifications. The 'liberal' (i.e. closest taxon) approach and the two distance thresholds, neither depending on the existence of multiple conspecific sequences, yielded more correct candidate species identifications (7–11%), but the percentage of false positives at the species (6–18%) or even genus (2–8%) level also increased. 'Strict' tree-based assignment consistently yielded a higher proportion of correct inferences at higher taxonomic ranks compared

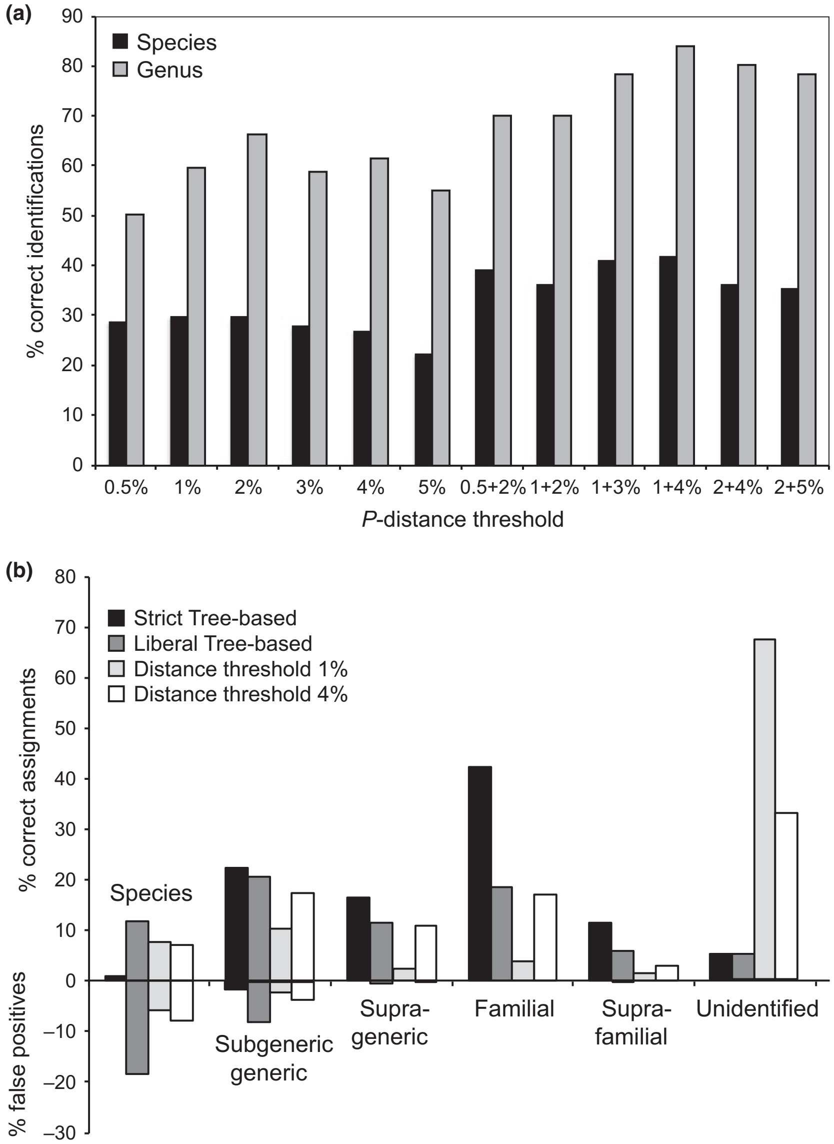

Fig. 3 (a) Percentage of correct assignment at species and genus level, when using a range of single distance thresholds or combinations of two. Comparisons are based on 107 SDTF test sequences, that is, sequences from known taxa which have conspecifics in GenBank and match the correct species or genus when the 'best-match' criterion is applied. (b) Percentage of correct or false assignment at different taxonomic ranks, using the 'strict' or 'liberal' tree-based criteria, and two distance thresholds. Comparisons are based on 450 SDTF test sequences.

to alternative approaches (22%, genus level; 16%, tribe or other suprageneric rank; 42%, family level), while keeping a low proportion of false positives (1.5% at the genus level). Some SDTF sequences (11%) were only assigned to order or higher taxonomic rank, and a few (5%) did not yield any results under default settings for sequence retrieval. However, the number of sequences that could not be assigned to any rank under the default settings using the 1% or 4% threshold approach was much higher (67% and 33%, respectively).

Overall, against a largely incomplete sequence database, the 'strict' tree-based taxonomic assignment approach proved to be reliable for taxonomic assignment at different ranks. Although a very conservative criterion, it appears to be the only safe and efficient approach to apply in the absence of a comprehensive reference database (Ross *et al.* 2008). Distance-based and 'liberal' approaches are still useful for providing candidate taxa, but not always safe for identification as they contribute false positives at different taxonomic ranks. Moreover, the threshold-based criteria can only be applied when there is something very similar in the database.

Structure, computing demands and novelty of the taxonomic identification pipeline

The structure and default settings of the pipeline were optimized to serve the taxonomic assignment of angiosperm *psbA-trnH* sequences (Figs 1 and 2; Appendix S2, Supporting information). However, the whole procedure can be readily modified and customized to serve the particular needs of other projects. Just by altering some of the default parameters, the pipeline can mine and build a relevant identification database for other loci or other taxa, with a number of options made available for users to query this database with new sequence data.

From the point of view of computational demands, perhaps the slowest step in the analysis affected the database construction part of the pipeline, specifically the initial BLAST search against the full plant sequence database. When the full 998 sequence query file was used, this step took about eight hours on our computer (two 6 core Dual Intel Xeon X5690 processors at 3.46 GHz, running Kernel Linux 2.6.33.3) and 11–13 h the whole database construction procedure. With a smaller query file, this time can be reduced, but as this procedure is only repeated approximately once every 2 months, when there is a new GenBank release, it is not considered of big concern. Most steps in the sequence identification part of the pipeline are relatively quick, apart from the phylogenetic inference step, which may slow down significantly the process, depending on the size of the aligned data sets (e.g. a typical 100-sequence query file took 5 h on average to run on our computer under default settings) and the possibility to parallelize processes. In summary, our pipeline allows to taxonomically identifying thousands of unknown plant sequences in few days, enormously helping to accelerate large-scale biodiversity assessment studies.

We are not aware of any tool available for automatic or semi-automatic identification of plant DNA sequences with the versatility and informative potential of our approach. The PTIGS-IdIt web application (Liu *et al.* 2011), specifically developed for the *psbA-trnH* locus, is exclusively based on distances; it does not provide assignment at higher taxonomic ranks, and it depends on a mostly complete reference database. Moreover, the existing implementation can only process one sequence at the time, making it unsuitable for large-scale projects. We intended to compare our results against the PTIGS-IdIt server, but this tool, both as web service and source code, was unavailable when we were preparing this manuscript (C. Liu, personal communication). Alternatively, phylogenetic approaches specifically developed for quick taxonomic assignment of short reads using reference alignments and reference trees (i.e. 'evolutionary placement algorithm'; Berger *et al.* 2011), would not be directly applicable to *psbA-trnH* sequences, because of alignment problems across distant taxa. As in our application, a clustering step and filtering of homologues would be required before multiple alignment. We consider combining the two methods in the future, which would be a necessary step for the implementation of our approach to NGS-data.

Example of automated identification: Importance of the local reference database

We ran 104 sequences obtained from infertile plants through the automated identification pipeline (Table S4, Supporting information). The distribution of 'best-match' distances, that is, pairwise distances of a query from its best-match database sequence, showed a clear shift towards lower values when the Nicaraguan SDTF reference sequences were included, with 46% of queries being identical to a database sequence (vs. 10% when only GenBank sequences were included) (Fig. 4a). The number of species-level identifications at the 1% p-distance threshold doubled when the local reference was considered (Fig. 4b). The results of the 'strict' tree-based

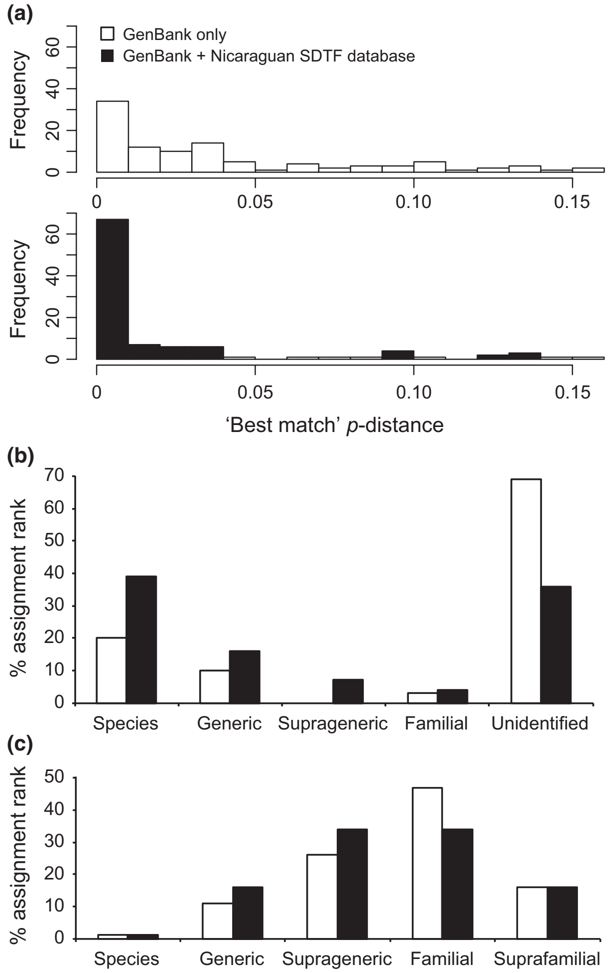

Fig. 4 Identification of infertile plant samples against GenBank or a combined GenBank and a local Nicaraguan SDTF reference database. Distribution of 'best-match' distances, that is, distances of queries from their closest database sequence (a). Percentage of sequences identified at different taxonomic ranks, when using a 1% distance threshold (b) or a 'strict' tree-based criterion (c).

identification at species level were not altered between treatments (1% in both cases), but taxonomic assignment at lower ranks was generally improved when the Nicaraguan sequences were included, with a significant increase in sequences identified at the genus or tribe levels (Fig. 4c). In summary, there was an important contribution of the local reference database for taxonomic assignment at different ranks, even if the improvement of species-level identifications would require including several individuals per plant species to adequately represent intraspecific variation. Nevertheless, taxonomic assignment at supraspecific ranks can still be very informative for biodiversity assessment and it can assist greatly in the morphological identification of problematic samples. For example, the reassessment of infertile voucher specimens based on our automated taxonomic assignment at higher ranks helped us to reach species- or genus-level identifications in 45 cases or nearly half of the unavailable identifications.

Example of automated identification: Assisting food-web ecology

Plant sequences, including fragments longer than 150 nt (a suggested maximum size for faecal DNA samples; Deagle *et al.* 2006), can be obtained from processed food of herbivore insects (Jurado-Rivera *et al.* 2009). Of 79 leaf beetle extractions, we amplified and sequenced successfully 59 putative diet sequences from 48 individuals in four species, ranging between 110 and 620 nt. 27 *psbA-trnH* were sequenced directly from single PCR products, while the rest required reamplification from excised gel bands. Seven and two individuals were sequenced for two and three coamplified products, respectively, while in twelve samples only one band was successfully sequenced.

Running the pipeline under the default settings assigned taxonomically 54 of the 59 diet sequences (Fig. 5; Table S2, Supporting information); the other five did not retrieve enough homologues from the database hampering meaningful identifications. For the cassid *Parorectis rugosa*, reported as feeding on groundcherries, our inferences showed that 16 (76%) of the retrieved sequences correctly fell within the *Physalis* clade (Fig. S2, Supporting information), one extra sequence was assigned to the closely related *Solanum*, and four sequences matched other families. In the case of the other cassid, *Physonota alutacea*, a manjack specialist, 70% of the sequences correctly fell within a *Cordia+Varronia* clade (two mutually paraphyletic borage genera, treated as synonyms by some authors; de Stapf 2010) (Fig. S3, Supporting information). The remaining three sequences matched other families. For the hispid *Brachycoryna pumila,* there were seven sequences matching the expected result at genus level (i.e. the mallow *Sida*) and two matching other families, while for the other hispid, *Heterispa vinula*, there were eleven sequences matching the expected genus or family and five matching other families. Overall, these results show that the framework developed here, that is, the automated identification pipeline together with a local reference database (even if incomplete), provides a great potential for inferring herbivore diet in the Mesoamerican SDTF, at least at generic level or above. This degree of resolution is already very

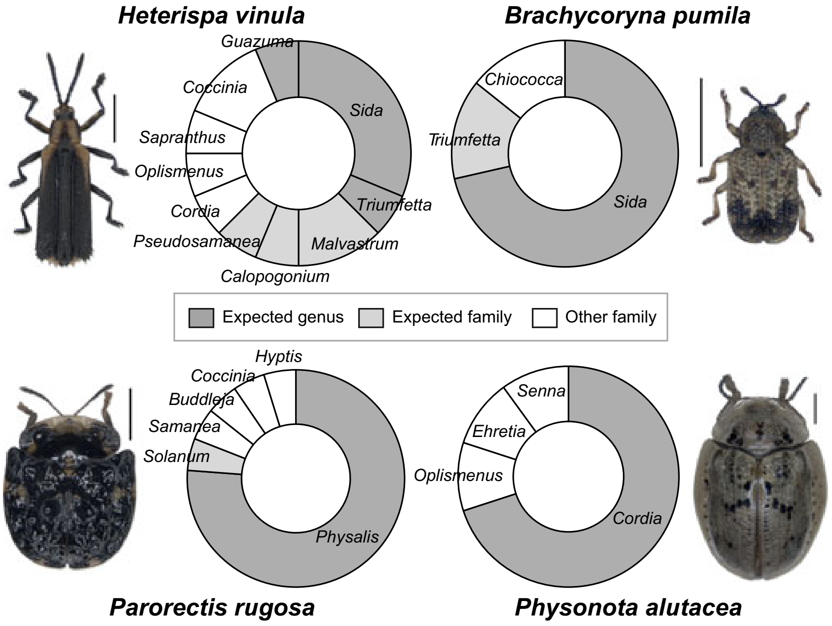

Fig. 5 Summary of diet inferences (at the plant generic level) after applying the automated taxonomic assignation BAGpipe procedure to identify *psbA-trnH* putative diet sequences from four species of leaf beetles of known dietary preference. (scale bars = 2.0 mm.)

important, as herbivore specialization is not necessarily realized at the species level, but possibly at higher ranks (Barrett & Heil 2012). Moreover, the results of the pipeline inform about which herbivores share the same diet, even in the absence of species-level plant identifications, which can still be useful to infer range of associations or food-web structure, among others. Unexpected family hits in inferred diets represent relevant ecological findings, maybe related to higher tolerance by adult insects to plant utilization compared to larvae, for instance, although we should not discard potential contamination or other artefacts.

Conclusions

We have developed a robust automated procedure for taxonomic identification based on DNA sequence data. Several highlights can be extracted from this study. (i) The phylogenetic framework is necessary for DNA-based taxonomic assignment of plant sequences against incomplete reference databases, a situation currently unavoidable in the tropics, given that the 'strict' tree-based criterion is the safest approach to avoid false positives. All identification procedures are sensitive to taxonomic coverage, to completeness of the reference system, and ours is not an exception. But we provide here with a flexible and fully automated approach which will allow us to investigate in a systematic way the impact of reference database composition on identification success rates. (ii) To compensate incompleteness of the reference database for our focal taxonomic and geographical scope, we have made a first step towards the construction of a sequence database for the Mesoamerican SDTF, which will greatly enhance the efforts for inventorying plant diversity and recording ecological interactions in this threatened ecosystem. (iii) The *psbA-trnH* marker, despite some problematic features, has overall performed well for taxonomic assignment across a wide range of tropical plant taxa. We provide a tool, specifically designed to overcome alignment difficulties of this marker and use it efficiently for species identification and taxonomic assignment within both phylogenetic and phenetic frameworks. This approach can be easily extended and applied to the two protein-coding plant barcode markers, *rbcL* and *matK*, to take advantage of the growing reference databases for these markers, respecting the standards set by CBOL Plant Working Group (2009). And (iv) we contribute the script suite 'BAGpipe' (pipeline for Biodiversity Assessment using GenBank data), specifically developed for large-scale plant biodiversity assessment in the tropics, but readily adjustable to the purposes of other sequence-based identification projects using any marker or taxon. We are currently working on a server-based implementation of 'BAGpipe' and intend to make it available for other commonly used markers in plant and animal biodiversity studies (Papadopoulou *et al*., unpublished).

Acknowledgements

This project was mainly funded by the 'Ecology and Conservation Biology' programme (BIOCON 08) of the Spanish private fund, 'Fundación BBVA' led by J.G.-Z. Additional funding was available for fieldwork thanks to Project C/032352/10 of the Interuniversity Cooperation and Research Programme (PCI-AECID 2010) of the Spanish Ministry of Foreign Affairs and Cooperation (MAEC). A.P. was funded by the postdoctoral programme 'Juan de la Cierva' of the Spanish Ministry of Science and Innovation (MICINN); G.D.C. by the predoctoral programme of the AECID (Spanish MAEC); and D.C. by the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. KSXC2-EW-B-02) and the Chinese National Science Foundation (Grants No. 30870268, 31172048, J1210002). The authors want to thank Inés Álvarez (Royal Botanical Garden, CSIC, Madrid) for her assessment in our initial steps to process plant material for genetic analyses, and we are sincerely indebted to Dr. Nicolas Puillandre and two anonymous reviewers for their encouragement and their help to improve this article.

References

Alonso-Alemany D, Barré A, Beretta S *et al*. (2014) Further steps in TANGO: improved taxonomic assignment in metagenomics. *Bioinformatics*, **30**, 17–23.

Altschul SF, Madden TL, Schaffer AA *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Austerlitz F, David O, Schaeffer B *et al*. (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics*, **10**(Suppl 14), S10.

Barrett LG, Heil M (2012) Unifying concepts and mechanisms in the specificity of plant-enemy interactions. *Trends in Plant Science*, **17**, 282–292.

Berger SA, Stamatakis A (2012) PaPaRa 2.0: a vectorized algorithm for probabilistic phylogeny-aware alignment extension. *Heidelberg Institute for Theoretical Studies*, http://sco.h-its.org/exelixis/publications.html. Exelixis-RRDR-2012-2015.

Berger SA, Krompass D, Stamatakis A (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under Maximum Likelihood. *Systematic Biology*, **60**, 291–302.

Bruni I, De Mattia F, Martellos S *et al*. (2012) DNA Barcoding as an effective tool in improving a digital plant identification system: a case study for the area of Mt. Valerio, Trieste (NE Italy). *PLoS ONE*, **7**, e43256.

Camacho C, Coulouris G, Avagyan V *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, **106**, 12794–12797.

Chase MW, Cowan RS, Hollingsworth PM *et al*. (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**, 295–299.

Costion C, Ford A, Cross H *et al*. (2011) Plant DNA barcodes can accurately estimate species richness in poorly known floras. *PLoS ONE*, **6**, e26841.

Cowan RS, Fay MF (2012) Challenges in the DNA barcoding of plant material. *Methods in Molecular Biology*, **862**, 23–33.

Deagle BE, Eveson JP, Jarman SN (2006) Quantification of damage in DNA recovered from highly degraded samples-a case study on DNA in faeces. *Frontiers in Zoology*, **3**, 11.

Devey DS, Chase MW, Clarkson JJ (2009) A stuttering start to plant DNA barcoding: microsatellites present a previously overlooked problem in non-coding plastid regions. *Taxon*, **58**, 7–15.

Dexter KG, Pennington TD, Cunningham CW (2010) Using DNA to assess errors in tropical tree identifications: how often are ecologists wrong and when does it matter? *Ecological Monographs*, **80**, 267–286.

Drummond AJ, Ashton B, Buxton S *et al.* (2010) *Geneious v5. 3*. Biomatters, Ltd, Auckland, New Zealand.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Egan AN, Crandall KA (2008) Incorporating gaps as phylogenetic characters across eight DNA regions: ramifications for North American Psoraleeae (Leguminosae). *Molecular Phylogenetics and Evolution*, **46**, 532–546.

Fazekas AJ, Burgess KS, Kesanakurti PR *et al.* (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE*, **3**, e2802.

Fazekas AJ, Kesanakurti PR, Burgess KS *et al.* (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources*, **9**, 130–139.

Fazekas AJ, Steeves R, Newmaster SG, Hollingsworth PM (2010) Stopping the stutter: improvements in sequence quality from regions with mononucleotide repeats can increase the usefulness of non-coding regions for DNA barcoding. *Taxon*, **59**, 694–697.

García-Robledo C, Erickson DL, Staines CL, Erwin TL, Kress WJ (2013) Tropical plant-herbivore networks: reconstructing species interactions using DNA barcodes. *PLoS ONE*, **8**, e52967.

González MA, Baraloto C, Engel J *et al.* (2009) Identification of Amazonian trees with DNA barcodes. *PLoS ONE*, **4**, e7483.

Griscom HP, Ashton MS (2011) Restoration of dry tropical forests in Central America: a review of pattern and process. *Forest Ecology and Management*, **261**, 1564–1579.

Hebert PDN, Ratnasingham S, DeWaard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B*, **270**(Supplement), S96–S99.

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

Janzen DH (1988) Tropical dry forests. The most endangered major tropical ecosystem. In: *Biodiversity*(ed Wilson EO), pp. 130–137. National Academy Press, Washington, DC.

Janzen DH, Hallwachs W, Blandin P *et al.* (2009) Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources*, **9**, 1–26.

Jeanson ML, Labat JN, Little DP (2011) DNA barcoding: a new tool for palm taxonomists? *Annals of Botany*, **108**, 1445–1451.

Jones FA, Erickson DL, Bernal MA *et al.* (2011) The roots of diversity: below ground species richness and rooting distributions in a tropical forest revealed by DNA barcodes and inverse modeling. *PLoS ONE*, **6**, e24506.

Jurado-Rivera JA, Vogler AP, Reid CAM, Petitpierre E, Gómez-Zurita J (2009) DNA barcoding insect-hostplant associations. *Proceedings of the Royal Society Series B*, **1657**, 639–648.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Katoh K, Toh H (2008) Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics*, **9**, 212.

Katoh K, Misawa K, Kuma KÄ, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution*, **52**, 540–542.

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS ONE*, **2**, e508.

Kress WJ, Erickson DL, Jones FA *et al.* (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences*, **106**, 18621–18626.

Linares-Palomino R, Kvist L, Aguirre-Mendoza Z, Gonzales-Inca C (2010) Diversity and endemism of woody plant species in the Equatorial Pacific seasonally dry forests. *Biodiversity and Conservation*, **19**, 169–185.

Liu C, Liang D, Gao T *et al.* (2011) PTIGS-IdIt, a system for species identification by DNA sequences of the psbA-trnH intergenic spacer region. *BMC Bioinformatics*, **12**, S4.

Liu K, Warnow TJ, Holder MT *et al.* (2012) SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology*, **61**, 90–106.

Luan P, Ryder OA, Davis H, Zhang Y, Yu L (2013) Incorporating indels as phylogenetic characters: impact for interfamilial relationships within Arctoidea (Mammalia: Carnivora). *Molecular Phylogenetics and Evolution*, **66**, 748–756.

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.

von Mering C, Hugenholtz P, Raes J *et al.* (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, **315**, 1126–1130.

Miles L, Newton A, Defries R *et al.* (2006) A global overview of the conservation status of tropical dry forests. *Journal of Biogeography*, **33**, 491–505.

Munch K, Boomsma W, Willerslev E, Nielsen R (2008) Fast phylogenetic DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 3997–4002.

Nilsson RH, Ryberg M, Kristiansson E *et al.* (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS ONE*, **1**, e59.

Pang X, Liu C, Shi L *et al.* (2012) Utility of the *trnH-psbA* intergenic spacer region and its combinations as plant DNA barcodes: a meta-analysis. *PLoS ONE*, **7**, e48833.

Parmentier I, Duminil J, Kuzmina M *et al.* (2013) How effective are DNA barcodes in the identification of African rainforest trees? *PLoS ONE*, **8**, e54921.

Portillo-Quintero CA, Sánchez-Azofeifa GA (2010) Extent and conservation of tropical dry forests in the Americas. *Biological Conservation*, **143**, 144–155.

Quesada M, Sanchez-Azofeifa GA, Alvarez-Añorve M *et al.* (2009) Succession and management of tropical dry forests in the Americas: review and new perspectives. *Forest Ecology and Management*, **258**, 1014–1024.

Ratnasingham S, Hebert PDN (2007) BOLD: the Barcode of Life Data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, **16**, 276–277.

Ross HA, Murugan S, Sibon Li WL (2008) Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology*, **57**, 216–230.

Sánchez-Azofeifa A, Quesada M, Rodríguez J *et al.* (2005) Research priorities for Neotropical dry forests. *Biotropica*, **37**, 477–485.

Sang T, Crawford D, Stuessy T (1997) Chloroplast DNA phylogeny, reticulate evolution, and biogeography of Paeonia (Paeoniaceae). *American Journal of Botany*, **84**, 1120.

Shaw J, Lickey EB, Beck JT *et al.* (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany*, **92**, 142–166.

Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology*, **49**, 369–381.

Simmons MP, Müller K, Norton AP (2007) The relative performance of indel-coding methods in simulations. *Molecular Phylogenetics and Evolution*, **44**, 724–740.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology*, **57**, 758–771.

de Stapf MNS (2010) Nomenclatural notes on *Varronia* (Boraginaceae *s.l.*) in Brazil. *Rodriguésia*, **61**, 133–135.

Steven NG, Subramanyam R (2009) Testing plant barcoding in a sister species complex of pantropical Acacia (Mimosoideae, Fabaceae). *Molecular Ecology Resources*, **9**, 172–180.

Stevens WD, Ulloa CU, Pool A *et al.* (2001) *Flora de Nicaragua*. Missouri botanical garden Press, St. Louis, MO.

Tate JA, Simpson BB (2003) Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. *Systematic Botany*, **28**, 723–737.

Tripathi AM, Tyagi A, Kumar A *et al.* (2013) The internal transcribed spacer (ITS) region and *trnH-psbA* are suitable candidate loci for DNA barcoding of tropical tree species of India. *PLoS ONE*, **8**, e57934.

Valentini A, Miquel C, Nawaz MA *et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Molecular Ecology Resources*, **9**, 51–60.

Whitlock BA, Hale AM, Groff PA (2010) Intraspecific inversions pose a challenge for the *trnH-psbA* plant DNA barcode. *PLoS ONE*, **5**, e11533.

Yu DW, Ji Y, Emerson BC *et al.* (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203–214.

J.G.-Z. and A.P. designed research; A.P., I.C., G.D.C., A.C., J.R., J.M.M., R.M.R. and J.G.-Z. performed research, including fieldwork and labwork; I.C. identified plant samples; A.P., D.C. and J.G.-Z. developed the pipeline; D.C. wrote the Perl scripts; A.P. and J.G.Z. wrote the paper; all authors discussed the results and approved the final manuscript.

Data Accessibility

DNA sequences: EBI Nucleotide Archive Accession nos HG963487-HG964098.

Sequence alignments: Dryad repository, doi:10.5061/dryad.j42c6

DNA sequences and other metadata: http://biologia-evolutiva.org/dryforest/

Computer Programs: The program, user manual and example data set are available from http://www.ibe.upf-csic.es/SOFT/Softwareanddata.html and http://sourceforge.net/users/dchesters.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Species of angiosperms from the Pacific side of Nicaragua and the northern province of Estelí used to construct a local reference *psbA-trnH* sequence database for automated identification of flora in the Mesoamerican seasonally deciduous tropical forest. Alongside the plant taxonomy, geographical sources, sample voucher numbers (UNAN Herbarium, IBE-JGZ DNA collection) and public sequence database Accession numbers are given

Table S2 Species of Cassidinae (Coleoptera: Chrysomelidae) of known diet used to test the performance of the automated DNA-based identification pipeline. Several individuals for each species were used to retrieve *psbA-trnH* sequences from whole-specimen DNA extractions, thus likely representing plant tissue remains in the insect gut. The number of specimens tested, of putative diet sequences retrieved and the results obtained with the pipeline are given

Table S3 GenBank *psbA-trnH* sequences removed a posteriori from the curated database for automated identification due to problematic taxonomic annotations

Table S4 Resuls of BAGpipe automated taxonomic assignment of 104 infertile plant samples based on *psbA-trnH* sequences and a reference database including GenBank data and a custom Nicaraguan sclerophyll deciduous tropical forest database

Fig. S1 Distribution of minimum p-distances between 114 Nicaraguan SDTF sequences and their conspecifics from GenBank.

Fig. S2 Maximum-likelihood tree with clade support values above 70% produced automatically by the BAGpipe pipeline for a group of 16 putative diet sequences of the cassid *Parorectis rugosa*. The pipeline isolated a group of closely related sequences from GenBank and our purpose-built SDTF local reference (NPL codes) and placed the diet sequences within a *Physalis* clade with high confidence.

Fig. S3 Maximum-likelihood tree with clade support values above 70% produced automatically by the BAGpipe pipeline for a group of seven putative diet sequences of the cassid *Physonota alutacea*. The pipeline isolated a group of closely related sequences from GenBank and our purpose-built SDTF local reference (NPL codes) and placed the diet sequences within a specific clade of *Cordia* (+*Varronia*) sequences and with high confidence.

Appendix S1 Examples of GenBank sequences, annotated as *psbA-trnH* but not retrieved by any of the similarity searches performed.

Appendix S2 BAGpipe manual. The manual for the whole procedure and specific instructions for scripts can be found at http://www.ibe.upf-csic.es/SOFT/Softwareanddata/ and http://sourceforge.net/users/dchesters.