Published in final edited form as:
Taxon. 2008 November 1; 57(4): 1304–1316.

DNA BARCODING IN LAND PLANTS: DEVELOPING STANDARDS TO QUANTIFY AND MAXIMIZE SUCCESS

David L. Erickson[1,*], **John Spouge**[2], **Alissa Resch**[2], **Lee A. Weigt**[3], and **W. John Kress**[1]

[1]Department of Botany, MRC-166, National Museum of Natural History, Smithsonian Institution, P.O. Box 37012, Washington, DC 20013-7012, U.S.A.

[2]National Center for Biotechnology Information, Computational Biology Branch, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, U.S.A.

[3]Laboratories of Analytical Biology, National Museum of Natural History, Smithsonian Institution, 4210 Silver Hill Rd, Suitland, MD 20746, U.S.A.

Abstract

The selection of a DNA barcode in plants has been impeded in part due to the relatively low rates of nucleotide substitution observed at the most accessible plastid markers. However, the absence of consensus also reflects a lack of standards for comparing potential barcode markers. While many publications have suggested a host of plant DNA barcodes, the studies cannot be readily compared with each other through any quantitative or statistical parameter, partly because they put forward no single compelling rationale relevant to the adoption of a DNA barcode in plants. Here, we argue that the efficacy of any particular plant DNA barcode selection should reflect the anticipated performance of the resulting barcode database in assignment of a query sequence to species. While legitimate scientific disagreement exists over the criteria relevant to "database performance", the notion gives a unifying rationale for prioritizing selection criteria. Accordingly, we suggest a measure of barcode efficacy based on the rationale of database performance, "the probability of correct identification" (PCI). Moreover, the definition of PCI is left flexible enough to handle most of the scientific disagreement over how to best evaluate DNA barcodes. Finally, we consider how different types of barcodes might require different methods of analysis and database design and indicate how the analysis might affect the selection of the most broadly effective barcode for land plants.

Keywords

Probability; Identification; PCI; Barcoding; Land Plants; Evolution

Introduction

DNA barcoding has rapidly achieved recognition as an important tool with the power to aid many basic research and applied endeavors in taxonomy and species identification (Hebert & al., 2003; Savolainen & al., 2005; Hajibabaei & al., 2007). In animals, adoption of a DNA barcoding locus (cytochrome c oxidase 1 or CO1) was largely synonymous with adoption of DNA barcoding at large, as early studies employing CO1 demonstrated that DNA barcoding was feasible (Herbert & al., 2003). The significant levels of nucleotide substitution between species of animals (>10% divergence among species in many animal lineages) combined with the relative ease of acquiring sequence data, readily distinguished CO1 as a viable DNA

*To whom correspondence should be addressed. ericksond@si.edu.

barcode in animals. However, the selection of a DNA barcode for plants has been far less straightforward (Chase & al., 2005; Cowan & al., 2006; Kress & Erickson, 2008). This is in part due to the intrinsic differences in observed nucleotide substitution rates at virtually all of the barcode loci proposed for plants relative to CO1 in animals, but it is also due to ambiguity in what criteria should be used to select a barcode marker. The criteria employed in selecting a barcode should be driven by how the resulting database fulfills the objectives of DNA barcoding, which although diverse in practice, should be narrowly defined when establishing criteria for selection. The two main applications of DNA barcoding have been (1) the identification of recognized species and (2) the discovery of novel genotypes that may form the basis of subsequent species discovery (Hebert & al., 2004a, b). We consider identification of recognized species to be central, with discovery of novel genotypes desirable, but not defining. To this end, we outline the most relevant criteria for evaluating putative plant DNA barcode markers. We rank criteria and give examples of why those criteria are more or less important, all within the context of what will maximize our power to identify recognized species correctly. We note that the selection of a barcode marker necessarily involves a balance of trade-offs among several options. By elucidating these options and the trade-offs we can establish priorities from which a rational decision can be made regarding the selection of a plant barcode. Throughout this paper, we use the term "marker" to mean any genic region (coding or non-coding) that may be used as part of a barcode, and the term "DNA barcode" or "barcode" to mean any one or more markers that are used as the genetic identifiers. Moreover, we distinguish (protein) coding loci from "intergenic spacers", which are markers consisting of a non-coding region flanked by two coding loci (which provide conserved PCR primers).

The use of standards established by the Consortium for the Barcode of Life (CBOL) for adopting barcodes in plants has thus far not been sufficiently stringent to discriminate among a host of proposed candidates (see http://www.barcoding.si.edu/). The CBOL standards focus on contrasting proposed markers with the performance of CO1. In principle, a single uniform DNA barcode (and barcode database) for all organisms based on CO1 is desirable, but the limitations of mitochondrial genes, including CO1, have excluded CO1 as a candidate plant DNA barcode (Table 1; Cowan & al., 2006) and rendered the comparison of CO1 with other proposed nuclear or chloroplast markers an insufficient standard. Nearly all chloroplast and nuclear markers tested have proven more diagnostic than CO1 in plants (e.g. Table 1); yet additional criteria for discriminating among putative chloroplast and nuclear barcode markers tested so far remain lacking. We propose additional standards that a plant DNA barcode should meet to recommend for its adoption. In addition, the relative priority of different criteria used to evaluate barcode markers needs to be established. These criteria include the recovery rate (universality) of PCR and sequence data, the proportion of species that can be correctly identified (resolution) and the magnitude of that differentiation, the expected taxonomic breadth covered by a DNA barcode under a narrow range of laboratory conditions, and complementation of markers used in a multi-locus DNA barcode. In addition, we suggest that ultimately, a single shared statistic is required to quantitatively analyze all putative barcodes simultaneously with regard to recovery rates and divergence among congeneric species. Only through direct comparison of a single standard statistic can a uniform, rational decision be made among putative plant DNA barcodes

Probability of Correct Identification in Plant Barcode Markers

Numerous markers, singly and in combination, have been suggested for the plant DNA barcode (Chase & al., 2007; Kress & Erickson, 2007; Newmaster & al., 2006; Lahaye & al., 2008a). However, the rationale for selecting individual markers or their use in combination has not been analyzed in a quantitative context (Chase & al., 2007; Taberlet & al., 2006; Kress & al., 2005) with a handful of exceptions (Kress & Erickson 2007; Lahaye & al. 2008a, b; Fazekas & al., 2008). One emerging statistic has been the use of monophyly to infer efficacy of a barcode marker(s) (Lahaye & al., 2008a, b; Fazekas & al., 2008). The goal of a DNA barcode is not to infer monophyly, however, and we note that unambiguous assignment of sequences to a species is still possible, even with ambiguities in estimating homology, and hence monophyly. Here, we propose several closely related statistics, called generically the "probability of correct identification" (PCI), to evaluate the efficacy of a putative barcode for species identification. The ambiguity in the definition of PCI encompasses most (if not all) of the legitimate disagreements over barcode criteria, but all versions of PCI have a unifying rationale, which should become clear shortly. Most practical algorithms for species assignment start by comparing two DNA sequences to produce a distance between the sequences; they then use a nearest neighbor algorithm to assign an unknown sample organism to a species by finding the closest database sequence to the sample sequence. Examples of sequence distances include: Kimura-2-Paramater Distance, Needleman-Wunsch Algorithm for Global Alignment Distance, and the Smith-Waterman Algorithm (similar to BLAST) for Local Alignment Similarity. PCI implicitly depends on the species assignment algorithm; thus, a poor algorithm makes the PCI correspondingly poor. In addition, there are many of factors affecting the PCI, including: (1) the inclusion of PCR success, (2) taxonomic weighting, (3) scaling the analysis by species or individual, and (4) probabilistic versus discrete assignment within species. The Appendix examines each of these issues in greater detail and shows how each of the four issues can affect the PCI. Different scientists might wish to quantify different aspects of species identification by assessing PCI for, e.g., datasets restricted to individuals with data from all marker regions (e.g. where PCR success is 100%), or datasets including only angiosperms or cryptogams, etc. Thus, divergent views on which features are most important in a barcode can be addressed by applying PCI to different datasets, which reflect diverse but legitimate views of which types of species identification are most important to a DNA barcode . We consider PCI with a simple example below and in the Appendix, as well as with a re-analysis of published data.

Even with all its simplicity, however, the definition of PCI as a "probability of correct identification" contains some subtleties, which we now explore with a basic example (outlined in Figure 1). Consider a sample of 8 species, labeled 1 to 8, and the barcode based on the single marker "A". Within marker "A", only one base varies: in species 1, 2, and 3, it is always A; in species 4 and 5, it is always C; in species 6 and 7, it is always G; and in species 8, it is always T. As mentioned above, the first subtlety is that species identification requires a procedure, i.e., a bioinformatics algorithm. Fortunately, with only a single varying base, and no intra-specific variation, the algorithm is obvious: use the varying base in marker "A" to identify species. The three species 1, 2, and 3, share one barcode sequence; the two species 4 and 5 share a second, the two species 6 and 7 share a third; and only species 8 has a unique barcode sequence. In the "probabilistic species assignment", if one has no reason to treat the species differently and is forced to assign to a single species, the chance of correct identification is 1/3 for species 1, 2, and 3; 1/2 for species 4, 5, 6, and 7; and 1 for species 8. In the example, therefore, the PCI is

$$((1/3)\times 3+(1/2)\times 4+1)\ /\ 8=4/8=0.5.$$

Some scientists might consider any ambiguity at all as a complete failure of species identification. Under "discrete species assignment", where species are either correctly or incorrectly assigned, with no middle ground, because only species 8 can be identified unambiguously, the PCI is

$$(1)\ /\ 8=1/8=0.125.$$

Although discrete species assignment might be a realistic model for certain types of situations (e.g., ones not tolerating any assignment error at all), the remainder of the paper permits shades of gray and considers probabilistic species assignment only.

The versions of PCI above do not allow for PCR failure, which can be accounted for, as follows. Assume that PCR fails to produce the sequence for marker "A" species 3, 5, and 7, effectively making their chance of correct identification 0. The chance of correct identification becomes 1/2 for species 1 and 2; 1 for species 4, 6, and 8; and 0 for species 3, 5, and 7. Thus, when the effects of PCR failure assert themselves (as they will in practice), the PCI becomes

$$((1/2) \times 2+1 \times 3) \ / \ 8=4/8=0.5.$$

Although it might seem counterintuitive at first, the phenomenon is reasonable (and in fact general): if a failed PCR would only have produced ambiguous identification, the failure does not influence the PCI.

If, however, PCR succeeds for all species except species 8, the PCI becomes

$$((1/3) \times 3+ (1/2) \times 4) \ / \ 8=3/8=0.375,$$

so if a failed PCR would have produced unambiguous identification, the failure lowers the PCI. In essence, PCR failure is noticeably deleterious if the barcode is effective; but less so, if the barcode is not.

Note that data associated with a barcode record, such as geography or morphology, might discriminate among species with identical DNA barcode sequences. Thus, a species assignment algorithm can be extended beyond pure sequence recognition, and we anticipate that this type of meta-data will make important contributions to species identification, further emphasizing the importance of PCR recovery so that a barcode record with both sequence and ancillary meta-data can be constructed. Although PCI can be extended to included algorithms using meta-data, PCI as described above represents the most complete available method available to compare how well different genetic markers will function as DNA barcodes.

The PCI can be applied to any sample and across any taxonomic range relevant to the efficacy of a barcode. Any quantitative comparison must be considered in context, however, e.g., the number of attempts made to recover the PCR amplicon and the number of different primers and reaction conditions employed for each locus must also be considered. If the same sets of taxa were analyzed for different putative plant DNA barcodes under a designated set of reaction conditions, however, the PCI would readily contrast their suitability for broad scale DNA barcoding. (Bioinformatics software tools specifically relevant to calculating the PCI can be found at the URL http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/Software/.)

To further demonstrate PCI, we conducted a re-analysis of data from Lahaye & al. (2008b) using data from two putative barcode markers, *matK* and *trnH-psbA* (Fig. 2). The datasets for each marker were restricted to specimens with 100% PCR success rate, i.e., only individuals with sequence data at the corresponding marker were considered. Furthermore, only species represented by more than one individual were considered. We considered eight different distance algorithms, each using either an alignment or evolutionary distance. Alignment methods were varied to include both global and local alignment, along with other alignment variants (semi-global and overlap alignment). We assessed taxonomic assignment at the level of genus as well as species. Finally, we partitioned taxonomic assignment into categories of "correct assignment". At the bottom of each graph in Figure 2, the green bars quantify the fraction of species where all samples were unambiguously correctly assigned; at the top, the red bars quantify the fraction of species where all sample were unambiguously incorrectly assigned; and in the middle, the yellow bars quantify the remaining fraction of species ("ambiguously identified species"). The yellow bars are subdivided further into the two fractions: at the bottom surrounded by green lines, the fraction of nearest neighbors to each sample, which were from the same species, averaged over all species; at the top surrounded by red lines, the fraction of nearest neighbors to each sample, which were from different species, averaged over all species. (Qualitatively, the yellow bars represent species where identification can be ambiguous. The more the yellow bar is surrounded by green line, the more likely a particular sample will be correctly identified. Discrete species assignment counts the ambiguous assignments generating the yellow bars as incorrect assignments, thereby essentially adding them to the red bars.)

Generally, the *trnH-psbA* marker produced higher PCIs than *matK*, reflecting the higher rates of sequence evolution reported by Lahaye & al. (2008a). Note that the choice of alignment algorithm can affect the PCI. Figure 2 therefore demonstrates how PCI can be applied to data to evaluate DNA barcodes and algorithms objectively. It also shows how the algorithms can affect conclusions about the suitability of barcode markers, and that different markers can be affected differently (e.g., global alignment distance is favorable for *matK* and unfavorable for *trnH-psbA*).

Other Considerations

Beyond having a single statistic that is applied to barcode loci, other factors must be considered in the general endeavor of barcoding. We propose five factors that should be weighted in terms of their importance in selecting a plant DNA barcode:

1. Universal PCR amplification;
2. Power of species differentiation;
3. Complementation among loci;
4. Breadth of taxonomic application;
5. Bioinformatic analysis

1. Universal PCR amplification

Universal PCR amplification must be the primary criterion for selecting a DNA barcode. Here, the term universal PCR amplification means 'the highest rates of recovery of a barcode region by PCR.' Universal PCR amplification does not mean 100% recovery, but instead means the highest relative rate of recovery among putative DNA barcodes. For plants the most challenging trade-off exists between the universal PCR amplification criterion and rates of sequence divergence. This trade-off is less problematic for non-coding regions, as universal primers may anneal in highly conserved genes that flank the hypervariable intergenic spacer. It is, however, true for all markers that are built from coding loci. Ultimately, the barcode user community must decide what constitutes a sufficient level of universal amplification and under what set of conditions. DNA barcoding at its core represents a diagnosis of organismal diversity rather than genic diversity (as opposed to genomics which captures maximum genic diversity at the expense of organismal diversity). As we discuss in subsequent sections, both the efficiency of barcoding, and its application depends on our ability to recover sequence data at the DNA barcode marker(s). Thus, while the absolute number for rates of recovery is arbitrary, some rationale should be considered for selecting that number. We suggest that a minimum of 90% amplification and sequence recovery should be established for an acceptable plant DNA barcode. Hajibabaei & al. (2005) suggested that 95% recovery should be a standard, and we agree that this would be desirable, but are uncertain if it can be attained, particularly with a single marker. If multiple genetic markers are employed for the barcode, we urge that each marker should meet the 90% minimum and the number of non-recovered samples should be lowered by at least one-half with the addition of each new barcode marker. Our rationale for selecting such high values derives from two considerations. The first consideration is that end users of DNA barcodes will include a much wider range of practitioners than the group who will decide which markers are employed. If DNA barcoding is to be practical, the methodology must be accessible and easily carried out. A very narrow, standard range of PCR conditions along with a very limited set of PCR primers per locus will provide a robust barcode marker. Secondly, the power of DNA barcoding is directly proportional to the database of barcodes available in the library; building a very complete DNA barcode database will greatly increase the power of DNA barcoding (Ekrem & al., 2007).

Sequencing 350,000 or more plant species, each with multiple replicates represents a massive undertaking where even modest decreases in recovery generate significant asymmetries in the barcode database. Let us make two assumptions: (1) every marker is recovered independently of every other marker; and (2) for every marker, the recovery rate is 90% = 0.9. Consider a barcode composed of two markers, each is recovered 90% of the time, and thus not recovered 10% of the time (=0.1); then the frequency that neither marker is recovered is the product of the failure rate for each marker: 0.1^2=0.01= 1% of the time, making the recovery rate for at least one marker 99%. Both markers are recovered together 81% of the time (the product of the success rate for both markers (=0.9^2=0.81). For a barcode composed of three markers (same 90% recovery rate) at least one of the three markers are recovered 99.9% of the time (1-failure rate = 1-0.1^3=0.999), hence a gain of only 0.9%. All three markers are recovered together only 72.9% (0.9^3=0.729≈73%) of the time, at 50% more effort and cost. The recovery of two versus three markers therefore noticeably favors two, unless the three markers compensate by diagnosing an increased fraction of species. This, however, is unlikely to be the case (Kress & Erickson 2007; Lahaye & al., 2008a; Fazekas & al., 2008) as empirical studies have demonstrated a rapidly declining rate of improved discrimination with more than two markers. Given the same level of ability to identify exact species, a large barcode database containing relatively unvarying sequences is more useful than a small database containing very variable sequences. As noted above, in a well populated database, meta-data such as geographic location or morphology could partition the database barcodes and help to identify species. Currently, GenBank does not permit association of meta-data to records, but its inclusion could greatly improve species identification with barcodes. Such ancillary meta-data lessens the emphasis on selecting a barcode locus with maximal rates of divergence, toward selecting one with universal recovery. Our barcode survey of the plants on Plummers Island in the Potomac River near Washington, DC (Kress & Erickson, unpublished data) included 239 species in 72 genera and 51 families. We found the *rbcL* gene alone could differentiate nearly every species, even without the inclusion of the highly variable *trnH-psbA* spacer. Together, the two barcodes correctly identified all species.

Obviously, it is desirable to have the complete sequence from the entire set of barcode markers for all taxa. Moreover, as extra markers enter a barcode database to compensate for poor sequence recovery, computer programs for species identification and phylogeny become more complex. Thus, although many factors affecting the utility of DNA barcodes (rapid rates of speciation versus molecular evolution; spurious species assignments; hybridization and introgression) are beyond our control, the rate of sequence recovery is not, because we can select loci with PCR primers that are maximally universal. We therefore consider the ability to recover sequence from samples as preeminent among our selection criteria, and all subsequent criteria as subordinate.

2. Degree of species differentiation

The prior argument regarding the primacy of universal recovery is relevant only if the markers employed exhibit a reasonable degree of divergence between species. The minimum of sequence divergence between species or the proportion of congeneric species that must be distinguished by a DNA barcode are again arbitrary values and conditional upon how the DNA barcodes are employed. Unlike recovery success, in a bioinformatic context the degree of species differentiation should be more flexible. This is in large part because we have no control over the rates of divergence among species, whereas we do have some control over the recovery rates through the selection and use of different genetic markers. That is, there will be a fraction of species groups that cannot be resolved with any suggested DNA barcode marker, but whose recovery can be improved through marker selection. We suggest that the average level of divergence between congeneric species is not the best statistic for defining the power of a barcode marker, and that instead the probability of correct species identification is a more valuable statistic. Thus while the divergent applications of barcode data invite a diversity of criteria for evaluating barcode markers (e.g. reconstruction of monophyletic groups with a marker (Lahaye & al., 2008a; Fazekas & al., 2008)), the essential enterprise of DNA barcoding is resolving species identities, which can be narrowly defined as correct assignment of a barcode sequence to a species represented in the barcode database. Because DNA barcoding does not seek to estimate the homology of mutations, low levels of divergence that would not be sufficient to estimate phylogenetic relationships may be sufficient to distinguish among taxa. The establishment of a minimum divergence requirement for barcode markers may also drive DNA barcoding toward DNA taxonomy, which is a central concern for those expressing skepticism of DNA barcoding (Seberg & al., 2003).

Empirical results employing diverse plant markers as barcodes suggest that rates of correct identification (combining recovery and discrimination) of species by a plant DNA barcode may be below that observed for most animal species, regardless of what markers are chosen for a plant barcode. The consistent differences in rates of nucleotide substitution between plant and plant chloroplast and animal mitochondrial genes suggest an intrinsic difference in evolutionary history between these lineages. Whether this is due to increased reinforcement and lineage sorting in animal clades is debatable, but what is clear is that an estimate of unambiguous species assignment with plant barcodes is likely to be in the range of 60-70%. Indeed one of the criticisms of DNA barcoding in general arises from highly divergent expectations of how well DNA barcodes should perform at species identification (Elias & al., 2007). Work by Kress & Erickson (2007) observed that no more than 87% of species could be unambiguously resolved when using pairs of markers applied to 48 congeneric species pairs, and further addition of markers was unable to improve resolution above that level. A reassessment of the data from Lahaye & al. (2008a) by Hollingsworth (2008) also point to a similar level of unambiguous assignment at nearly 60% using *matK* only, and Fazekas & al. (2008) suggest similarly low levels of resolution for all land plants. Whether these values reflect an underlying taxonomy that needs to be revised or whether it reflects the limitations of DNA barcodes to resolve true species is irrelevant. What matters is that a realistic assessment of the power of DNA barcodes exists given existing taxonomy, and recognition that the power of a plant DNA barcode to unambiguously and correctly assign species is less than 100%. A realistic assessment of expected resolution should help guide marker selection, and what will be needed is a way to select among different combinations of putative barcode markers that give similar results.

Lahaye & al. (2008a) observed that the *trnH-psbA* intergenic spacer was significantly more informative at species resolution than was the plastid gene *matK*, yet rejected *trnH-psbA* because it could not be readily aligned among all species for a global phylogenetic analysis. We feel that such selection criteria should be more thoroughly justified, as the appearance of *ad-hoc* decision making may cripple the core endeavor of plant DNA barcoding given the expected low rates of resolution among all land plants. We assert that anything improving PCI should be favored, with bioinformatic and analytic tools adjusted accordingly. We not address issues of alignment here, but not that if a coding locus were combined with a variable length intergenic spacer as a plant barcode, then use of a supermatrix would ameliorate problems with global sequence alignment (Driskell & al., 2005). Lastly, the incidence of discrimination should be an average across all land plants, and the inclination to advance or discount markers due to performance within one lineage (e.g. Cycads [Sass & al., 2007], ferns, etc.) must be avoided.

The ability to identify a sample to genus or family may also be sufficient in many barcoding contexts. In addition, if the barcode database can be partitioned such that a submitted barcode sequence is not compared against all sequences within the database (e.g. where geographic or morphological metadata can be utilized in barcode database searches), then our context may change for how variable a marker must be to be diagnostic at the species level. One issue that consistently arises is selecting a marker with a high nucleotide substitution rate but decreased PCR recovery, with the hope that better, more universal PCR primers will be developed in the future. While it is true that in some cases new and better PCR primers are continually designed for specific taxa, we recommend that the adoption of a barcode locus today should be based upon results from today, particularly in light of adopting markers that have been employed for decades in phylogenetic studies, but have still not yet been demonstrated to be broadly universal.

3. Complementation among loci

Driven mostly by the low intrinsic rates of sequence evolution observed at most plant chloroplast loci, broad agreement seems to support a plant barcode consisting of two or more marker loci (Kress & al., 2005; Kress & Erickson, 2007; Chase & al., 2007; Rubinoff & al., 2005). The criteria for selecting loci to combine deserve special attention. They include: (1) the maximization of PCI; (2) the complementation in phylogenetic applications (i.e., one marker works well where the other does not); and (3) the use of unlinked loci, to decrease correlations among markers. Given the very real increase in time and money required to process additional barcode markers, we strongly suggest that consistent measures and criteria be applied to access the improvement gained by adding additional loci, PCI provides a framework to do just that. Similarly, the difficulty in managing data in a database of all land plants greatly favors employing the smallest number of markers possible. Two examples of measures that would be useful in delimiting the number of markers are: (1) each additional locus must reduce the number of non-recovered PCR by 50% (hereafter called the “50% rule”), and (2) the composite PCI must be 10% higher after adding an additional locus. The reasoning behind the 50% rule follows from the 50% increase in effort in going from a two to three marker barcode (as outlined earlier). Similarly, empirical results from Kress & Erickson (2007) demonstrate that the greatest increase in PCI when markers were combined also reduced the number of non-recovered samples by at least 50%. Given that the resolution of DNA barcodes appears to plateau rapidly (all published papers suggest 2 markers work nearly as well as 3 or more; Kress & Erickson 2007; Fazekas & al., 2008; Lahaye & al., 2008a,b) a high threshold of improvement that is proportional to time and money employed seems to make sense. Similarly, an approximate 10% increase in PCI was routinely the greatest increase observed when going from a single to a two-marker barcode (see Tables 1 & 2; Kress & Erickson 2007; Fazekas & al., 2008; Newmaster & al., 2008), with further additions rapidly decreasing in their contribution to resolution. We note that PCI can be readily scaled to multiple markers, providing a platform for assessing the efficacy of different marker combinations in DNA barcoding.

The improvement of a barcode marker may come from the addition of a less universal, but rapidly evolving marker or from a highly universal, but more slowly evolving marker. For example, in Figure 1, the complementation of a rapidly evolving marker with a much more slowly evolving marker results in a substantial increase in PCI. The PCI and the 50% rule can provide a measure of comparison between different strategies. In our investigations (Kress & Erickson, 2007; unpublished results), much of the improvement in species identification came from enhanced universal recovery (Tables 1 & 2) as where combining *trnH-psbA* with *rbcLa* exhibited greater power to identity than did combining *trnH-psbA* with *matK* due to its reduced universality. Similarly, improvement in identification with addition of extra markers quickly declined as barcodes composed of more than two markers failed to improve identification; a result that was mirrored in the results of Lahaye & al. (2008a). The complementation of loci that assured that at least one sequence is obtained for the maximum number of taxa showed a greatest improvement in our results (Kress & Erickson 2007), which is also predicted by the PCI. A simple and portable quantitative measure (PCI plus 50% rule) can allow us to test how complementation may improve the plant DNA barcode and guide selection with maximum benefit.

4. Breadth of taxonomic application

Another of the intrinsic trade-offs of candidate barcode markers is the successful application in different lineages of land plants (Small & al., 2005; Schneider & Schuettpelz, 2006). Many putative barcode markers appear to be limited in their utility when applied to non-angiosperms, such as monilophytes and mosses. Should different lineages of plants be weighted by ecological importance and difficulty of identification or strictly by their proportional representation in plants as a whole? Given that a large majority of plants are angiosperms in today's biomes, should markers be chosen that work best for them, or should cryptograms and gymnosperms, where identification may be more problematic, be given special attention? Additionally, should a multi-locus barcode be structured such that different marker loci work best for different lineages? If we accept that a DNA barcode is to facilitate identification, then it seems reasonable that the barcode should work well on groups were identification based on morphology is most difficult. Because species divergences among non-angiosperms appear higher than for many angiosperms, more slowly evolving markers where universality is emphasized may be more appropriate. At least one study that examined success in non-angiosperms for different barcodes suggested that the most universal barcode markers were also successful at species discrimination in the sample set employed (Kress & Erickson 2007).

5. Bioinformatics

Bioinformatic considerations influence the choice of a plant barcode, because barcode analysis places requirements on feasible database design and feasible algorithms. Regardless of the plant barcode selected, either accepted bioinformatics tools or reasonable alternatives must be in place. Fortunately, in most respects, the basic design of BOLD (the Barcode of Life Database; which handles all existing barcode data) extends to plant barcodes, because plant barcodes share many attributes with the animal barcodes already in BOLD (http://www.barcodinglife.org). Presently, however, the animal barcode is restricted to a single coding locus (CO1) so a barcode including an intergenic marker might require BOLD to modify the types of algorithms employed. Below, we discuss how bioinformatics should influence selection of a plant DNA barcode. In particular, we consider the ability of existing search algorithms to deal with variable length intergenic markers, the importance and ease of assessing the confidence of species assignments, and application of barcode data to phylogenetic reconstruction.

i. Sequence alignment algorithms—Sequence alignment is the starting point for consideration of all the bioinformatics issues. Small barcode studies lack rich bioinformatics resources, so they align sequences two at a time, with pairwise sequence alignment. In contrast, BOLD aligns many barcode sequences at once with Hidden Markov Models (Eddy, 1995, Durbin & al., 1998). Coding barcode sequences are (for the most part) easier to align than intergenic barcode sequences, because of the many insertions and deletions in the latter. In the initial step for an intergenic barcode, if multiple alignment of barcode sequences proved unfeasible, an alignment of every sequence pair could be substituted (Steinke & al., 2005). Multiple alignment does scale better with database size than pairwise alignment, however. Thus if the database contains N sequences, a multiple alignment can add a new sequence in constant time, whereas pairwise alignment is much slower and requires time proportional to N.

A coding barcode contains relatively few insertions and deletions, so most of its mutations are point mutations. The multiple alignment of a coding barcode therefore contains few gaps and has a length commensurate with its longest sequence, both desirable features when maintaining a barcode database. An intergenic barcode, on the other hand, contains many insertions and deletions, so the corresponding multiple alignment contains many gaps of variable length. In addition, gaps slow multiple sequence alignment and might become problematic for intergenic barcodes (Steinke & al., 2005). Ambiguity in the resulting multiple alignment might lead to incorrect species identification (Little & Stevenson 2007). Experience with the multiple alignment of intergenic barcodes is limited, however, leaving the magnitude of the potential difficulties uncertain.

Moreover, while the use of intergenic markers complicates analysis and design of a barcode database, several methods exist that might surmount the challenges. For example, to match an intergenic barcode from a specimen to the barcode database, a preliminary heuristic step could speed the database search. First, a heuristic search could find sequences matching the approximate length of the specimen's barcode. Comprehensive pairwise alignments could then locate matches to the specimen among database sequences of comparable length. Alternatively, in a barcode combining an intergenic marker with a coding marker, the coding marker could serve in a multiple alignment to localize assignment, after which pairwise alignment of the intergenic marker could identify to species.

Other possible alignment problems, e.g., partial sequences or sequences with many missing or ambiguous bases, already occur with animal barcodes, so other than the multiple alignment computation itself, an intergenic plant barcode presents few novel problems to the initial multiple alignment step in the present taxonomic algorithms.

ii. Confidence of species assignment—Ideally, any statement about recognition and identification should include a probability or confidence interval. By their nature, probabilistic methods for recognition and identification like the likelihood-ratio method (Matz & Nielsen 2005), Bayesian estimation (Nielsen & Matz 2006), and identification by coalescent models (Abdo & Golding 2007) usually have this desirable property. The likelihood-ratio method (Matz & Nielsen 2005) is similar to methods for determining paternity in genealogical reconstruction (Marshall & al., 1998), and like most probabilistic methods, through the calculus of probabilities it readily combines marker information from multiple barcode loci. Typically, probabilistic methods require an alignment of many local similarities between a specimen sequence and database sequences, and so they remain compatible with the use of an intergenic barcode.

Probabilistic methods are most efficient if data are plentiful. In particular, if copious data can provide an accurate picture of within-species variation, species identification improves noticeably (Matz and Nielsen 2005; Nielsen and Matz 2006). Probabilistic methods also work best when a barcode database contains sequences from all possible alternative species of interest. In particular, Bayesian methods require a complete enumeration of the alternatives to calculate the posterior probability of identifying a specimen as a particular species and excluding other closely related species. In general, probabilistic methods show that identification improves with plentiful data, confirming the importance of choosing barcodes with high universality. Although Nielsen and Matz (2006) urge exclusive use of barcodes with the greatest mutation rate, the competing demands of mutation rate and PCR universality suggests a mixed approach towards the selection of a plant barcode, one mutating rapidly but still susceptible to PCR. By this criterion, the intergenic regions between PCR primer sites in tightly conserved genes make ideal barcodes.

Unfortunately, probabilistic methods are usually impractically slow for the high throughput required in a public barcode database, so we do not consider them further. For reasons of computational speed, BOLD uses only (non- probabilistic) nearest neighbor algorithms based on the Kimura 2-Parameter (Kimura 1980), Jukes-Cantor (Jukes and Cantor 1969), or Pairwise distances to recognize, identify, and classify specimen sequences. As a partial substitute for a probability or confidence interval concerning matches between the query sequence and the database sequences, BOLD offers “% Specimen Similarity”.

In BOLD, algorithms for recognition and identification align barcodes from a new specimen to the existing multiple alignment to calculate “distances” from the new specimen to database specimens. The new specimen can have one or several “nearest neighbors” in the database, because the new specimen might be at the nearest distance to several database specimens. Based on the nearest neighbors in the database, the algorithms then yield: (1) a list of possible species to which the specimen belongs; and (2) the nearest-neighbor distance, which indicates how probable it is that the specimen belongs to one of the species. The distance has the smallest value of 0, achieved for identical sequences (and possibly others); and its increase suggests a decreasing probability of correct identification. Most algorithms have similar output as nearest-neighbor algorithms, substituting only another type of number, e.g., a probability for a distance. No algorithm seems to improve noticeably on the identifications from a nearest neighbor identification algorithm (Austerlitz 2007).

To recognize a known species, BOLD applies a threshold to the nearest-neighbor distance. If the distance is less than the threshold, the specimen is recognized as a known species. Whereas a probabilistic algorithm can automatically adjust parameters as species and sequences enter a barcode database, a nearest-neighbor recognition algorithm must make *ad hoc* adjustments to its species-specific thresholds. For sparsely represented species, a recognition threshold of 0.02 point mutations per site for an evolutionary distance like Kimura 2-Parameter distance is popular, though somewhat arbitrary. Indeed, the naïve use of barcoding thresholds to discover novel species has drawn much criticism, and we do not recommend it.

The algorithms for recognition and identification are the same for coding and intergenic barcodes, with one important caveat. Alignments between intergenic barcode sequences generally produce few sites displaying point mutations. At present, evolutionary distances examine only point mutations: they discard the insertions and deletions that occur in intergenic barcode alignments. Thus, at least in theory, nearest-neighbor algorithms for intergenic barcodes should probably use alignment distances (which account for gaps), and not evolutionary distances (which do not). Preliminary results indicate that even with evolutionary distances, the identification algorithms perform well with intergenic barcodes, however, because of their enhanced variability (Fig. 2).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Conclusions

We have outlined a number of issues relating to the selection of a DNA barcode for plants. Ultimately, the selection of a DNA barcode represents a challenge in navigating a series of trade-offs. The relatively low levels of nucleotide substitution at most coding loci in green plants have eliminated the hope of a “silver bullet” locus that would address all concerns, and has pushed researchers to consider a host of options, notably consideration of multiple loci including non-coding loci. Our central goal in this paper is to raise critical questions that must be addressed in order to evaluate candidate barcode loci and suggest some criteria by which those questions can be answered. We stress that employing a standard quantitative parameter, the PCI, for evaluating and comparing single and multi-locus barcodes that is portable across experiments remains central. The universal acquisition of sequence data that can populate the database should be a top criterion for selection, with discretionary power second and taxonomic range and complementation among loci facilitating these two priorities. We also assert that use of either coding or non-coding marker regions would be suitable as barcodes from a bioinformatic perspective, that their use in combination may represent a good compromise, and that while combining information from multiple loci represents a challenge to database design and analysis, these challenges can be overcome.

Our review focuses on the technologies of today. But we readily recognize the pace at which improvements in technology may dramatically alter what we think of when we discuss DNA barcodes. Whole genome sequencing of mitochondria and chloroplasts may render obsolete discussions of what marker regions to employ. At that stage, the debate will shift toward how to best manage and analyze the avalanche of data that will be generated. What will not change is the difficulty in capturing the process of speciation in a static set of DNA sequences. However, the assembly of a database of even one or two sequences for all organisms has the potential to shed new light on this dynamic evolutionary process.

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, National Library of Medicine.

Appendix

APPENDIX

Given any particular level for weighting (e.g., species or individuals), and given any particular dataset, PCI is the probability of correct species identification at the level. To explain the notion of “level of weighting”, for the level of individuals, e.g., PCI is the probability of correct species identification for a random individual; for the level of species, PCI is the probability of correct species identification for a random species. Thus, a particular PCI is completely defined, once the dataset, the level of division, and a definition for the phrase “correct species identification”, have all been specified. Note that given all the specifications, that PCI can be applied to all barcode markers, singly and in combination, and that a definition for the phrase “correct species identification” implicitly assumes an algorithm for identifying species, given a barcode (an obvious requirement for a functioning barcode database).

Let us consider the definition's ramifications. First, we must deal with issues related to statistical sampling. Implicitly and ideally, the definition of PCI refers to performance of a given barcode database (e.g., a future BOLD containing a plant barcode). In practice, the ideal PCI must be estimated from preliminary samples much smaller than a final database (and possibly of lower standardization or quality). Sampling issues are always present, and we can only suggest that any preliminary sample mimic the composition of the intended database as much as possible. Thus, as promised, our measure of barcode efficacy, PCI, is based on the rationale of database performance. Many other measures of barcode efficacy, e.g., the “barcode gap”, plots of type I and type II error in all pairwise comparisons, etc., have at most a tangential relevance to barcode database performance.

Second, if species identification in certain taxa is of great concern, PCI may be restricted to those taxa, e.g., the PCI for angiosperms may be determined separately from the PCI for ferns, simply by restricting the relevant dataset to angiosperms. To illustrate some of the remaining ramifications, consider Table A1 below, which represents a hypothetical dataset. The first column labels 10 species #1 to #10. The second column gives the number of individuals sampled from each species. Species #10 was heavily sampled and constitutes a full 80 of the 100 individuals sampled. The third and fourth columns represent two markers, Marker a and b. For simplicity, assume that each species corresponds to a unique genomic sequence. Correct species identification therefore is unambiguous and does not depend on which individuals were sampled from the species. Assume for diagrammatic simplicity that the sequences for Marker a differ only at a single distinguishing position in the genomic alignment; likewise for Marker b. Let the algorithm for species identification be to examine the distinguishing position for a (or b, or a+b together). If the letter(s) in the distinguishing position(s) is unique to the species, the species (and all its individuals) are correctly identified, indicated by “*” in Table A1. In practical situations, the choice of algorithm can be more obscure and can greatly influence the PCI. We defer a discussion of algorithm selection to barcode bioinformatics section below.

Third, the “level for weighting” relevant to PCI corresponds to a decision about weighting probabilities by, e.g., individuals or species. Possibly, e.g., a barcode database might be likely to receive query sequences from particular species, so its performance might need to be weighted by query (i.e., by individuals) rather than species. We prefer to weight by species, to avoid drawing incorrect conclusions about barcode database performance from biased datasets, ones with excessively many samples from a few species. Thus, in the absence of a specific justification to the contrary, our recommendation is to use PCI with species as the level for weighting.

Table A1 illustrates the third ramification with Species #10, which contains 80 of the 100 individuals sampled. (Marker a has 100% PCR success, so restriction of the dataset to PCR success is immaterial.) In Table A2, if the level for weighting is “Individual”, the PCI for Marker a is 0.1; if “Species” 0.8. Obviously, Species #10 has primary influence on the PCI if the level for weighting is “Individual”, probably undesirably so.

Fourth, there is legitimate scientific disagreement over whether PCR success should be included in a measure of barcode efficacy. We argue as follows: PCR success influences database performance, so a legitimate measure of barcode efficacy like PCI should include the effects of PCR success. If, however, the reader disagrees with the straightforward logic of this statement, the dataset in definition of PCI can be restricted to the individuals displaying a given type of PCR success (e.g., to individuals where at least one marker displayed successful PCR).

Table A1 illustrates the fourth ramification with Marker b, where PCR failed on Species #6-#10. Because the same number of individuals was sampled from the species with PCR success, Table A2 shows that if PCI is restricted to the dataset showing PCR success, the level of weighting does not matter: the PCI is 0.6. Now, consider the dataset consisting of all individuals. If the level of weighting is “Species”, the PCI falls to 0.3; if “Individual”, to 0.06 (again reflecting primarily the result for Species #10). For Marker a+b, there was no PCR failure on both Marker a and b, so PCI always considers the full dataset. If the level of weighting is “Species”, the PCI is 0.5; if “Individual”, 0.88. Note that at the level of weighting “Species”, the PCI for the dataset restricted to PCR success drops from 0.6 to 0.5 when Marker a is introduced into the barcode, because the requirement for PCR success includes a larger dataset for Marker a+b than for Marker b. No such anomaly occurs if the dataset includes all sampled individuals and the PCI incorporates the effect of PCR failure.

Fifth, the definition of “correct species identification” can encompass many views about the barcode database performance. On one hand (and not displayed in the examples in Table A1 and Table A2), e.g., if only 1% = 0.01 of individuals from a particular species are wrongly assigned, one might consider identification of that species as incorrect, making the relevant probability of correct species identification 0. Such an “all-or-nothing” evaluation might be appropriate, e.g., if the incorrect identification then requires human examination of every individual from the species. On the other hand, if one is willing to accept occasional incorrect identifications, a “probabilistic” evaluation might consider the probability of correct identification for the species 99% = 0.99. PCI can thus accommodate legitimate scientific differences over what constitutes “correct species identification”. Because the all-or-nothing evaluation permits a tiny subset within the sample (the 1% of incorrect identifications within the species) a large influence on the PCI, we recommend the probabilistic evaluation.

Literature Cited

Abdo Z, Golding GB. A step toward barcoding life: A model-based, decision-theoretic method to assign genes to preexisting species groups. Syst. Biol 2007;56:44–-56. [PubMed: 17366136](doi: 10.1080/10635150601167005)

Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V. Land plants and DNA barcodes: short-term and long-term goals. Philos. Trans. Ser. B 2005;360:1889–-1895. (doi:10.1098/rstb.2005.1720)

Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madriñán S, Petersen G, Seberg O, Cameron KM, Kress WJ, Hedderson TAJ, Conrad F, Salazar G, Richardson JE, Hollingsworth M, Jørgsensen T, Kelly L, Wilkinson M. A proposal for a standardised protocol to barcode all land plants. Taxon 2007;56:295–-299.

Cowan RS, Chase MW, Kress WJ, Savolainen V. 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. Taxon 2006;55:611–-616.

Driskell AC, Ane C, Burleigh JG, McMahon MM, O'Meara BC, Sanderson MJ. Prospects for building the tree of life from large sequence databases. Science 2005;306:1172–1174. [PubMed: 15539599] (doi: 10.1126/science.1102036)

Durbin, R.; Eddy, SR.; Krogh, A.; Mitcheson, G. Biological Sequence Analysis. Cambridge University Press; Cambridge: 1998.

Eddy SR. Multiple alignment using hidden Markov models. Proc. Int. Conf. Intell. Syst. Mol. Biol 1995;3:114–-20. [PubMed: 7584426]

Ekrem T, Willassen E, Stur E. A comprehensive DNA sequence library is essential for identification with DNA barcodes. Mol. Phylo. Evol 2007;43:530–-542.(doi.org/10.1016/j.ympev.2006.11.021)

Elias M, Hill RI, Willmott K, Dasmahapatra K, Brower AVZ, Mallet J, Jiggins CD. Limited performance of DNA barcoding in a diverse community of tropical. Proc Roy Soc B-Biol Sci 2007;274:2881–-2889. (doi:10.1098/rspb.2007.1035)

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG. Multiple Multilocus DNA Barcodes from the Plastid Genome Discriminate Plant Species Equally Well. PLoS ONE 2008;3:e2802. [PubMed: 18665273]al(doi:10.1371/journal.pone.0002802)

Hajibabaei M, deWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, Mackie PM, Hebert PDN. Critical factors for assembling a high volume of DNA barcodes. Phil. Trans. R. Soc. B 2005;360:1959–1967. [PubMed: 16214753](doi:10.1098/rstb.2005.1727)

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet 2007;23:167–-172. [PubMed: 17316886](doi:10.1016/j.tig.2007.02.001)

Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proc. R. Soc. B 2003;270:313–-321.(doi:10.1098/rspb.2002.2218)

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. Identification of birds through DNA barcodes. PLoS Biol 2004a;2:e312. [PubMed: 15455034](doi:10.1371/journal.pbio.0020312)

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc. Natl Acad. Sci. USA 2004b;101:14812–-14817. [PubMed: 15465915](doi:10.1073/pnas.0406166101)

Hollingsworth P. DNA barcoding plants in biodiversity hotspots: progress and outstanding questions. Heredity 2008;101:1–2. [PubMed: 18398425](doi:10.1038/hdy.2008.16)

Jukes, TH.; Cantor, CR. Evolution of protein molecules. In: Munro, HN., editor. Mammalian Protein Metabolism. New York; New York: 1969. p. 21--123.

Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol 1980;16:111–-20. [PubMed: 7463489]

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. Use of DNA barcodes to identify flowering plants. Proc. Natl Acad. Sci. USA 2005;102:8369–-8374. [PubMed: 15928076](doi:10.1073/pnas.0503123102)

Kress WJ, Erickson DL. A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the noncoding *trnH-psbA* spacer region. PLoS ONE 2007;2:e508. [PubMed: 17551588](doi:10.1371/journal.pone.0000508)

Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V. DNA barcoding the floras of biodiversity hotspots. Proc. Natl Acad. Sci. USA 2008a;105:2923–-2928. [PubMed: 18258745](doi:10.1073/pnas.0709936105)

Lahaye R, Savolainen V, Maurin O, Duthoit S, van der Bank M. A test of psbK-psbI and atpF-atpH as potential plant DNA barcodes using the flora of the Kruger National Park as a model system (South Africa). Nature Precedings. 2008bhdl:10101/npre.2008.1896.1

Little D, Stevenson DW. A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. Cladistics 2006;22:1–-21.(doi:10.1111/j.1096-0031.2006.00126.x)

Marshall TC, Slate J, Kruuk LEB, Pemberton JM. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol 1998;7:639–-655. [PubMed: 9633105](doi:10.1046/j.1365-294x.1998.00374.x)

Matz MV, Nielsen R. A likelihood ratio test for species membership based on DNA sequence data. Phil. Trans. R. Soc. B 2005;360:1969–-1974. [PubMed: 16214754](doi:10.1098/rstb.2005.1728)

Newmaster SG, Fazekas A, Ragupathy S. DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach Can. J. Bot 2006;84:335–-441.

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J. Testing candidate plant barcode regions in the Myristicaceae. Molecular Ecology Resources 2008;8:480–490.(doi: 10.1111/j.1471-8286.2007.02002.x)

Nielsen R, Matz M. Statistical approaches for DNA barcoding. Syst. Biol 2006;55:162–-169. [PubMed: 16507534](doi:10.1080/10635150500431239)

Ratnasingham, S.; Hebert, PDN. BOLD: The Barcode of Life Data System; Mol. Ecol. Notes. 2007. p. 355--364.(www.barcodinglife.org)(doi:10.1111/j.1471-8286.2006.01678.x.)

Rubinoff D, Cameron S, Will K. Are plant DNA barcodes a search for the Holy Grail? Trends Ecol. Evol 2005;21:1–-2. [PubMed: 16701459](doi:10.1016/j.tree.2005.10.019)

Sass C, Little DP, Stevenson DW, Specht CD. DNA barcoding in the Cycadales: testing the potential of proposed barcoding markers for species identification of cycads. PLoS ONE 2007;2:e1154. [PubMed: 17987130](doi:10.1371/journal.pone.0001154)

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R. Towards writing the encyclopedia of life: an introduction to DNA barcoding. Philos. Trans., Ser. B 2005;360:1850–-1811.(doi:10.1098/rstb.2005.1730)

Schneider H, Schuettpelz E. Identifying fern gametophytes using DNA sequences. Mol. Ecol. Notes 2006;6:989–-991.(doi:10.1111/j.1471-8286.2006.01424.x)

Seberg OC, Humphries J, Knapp S, Stevenson DW, Petersen G, Scharff N, Andersen NM. Shortcuts in systematics? A commentary on DNA-based taxonomy. Trends Ecol. Evol 2003;18:63–-65.(doi:10.1016/S0169-5347(02)00059-9)

Small RL, Lickey EB, Shaw J, Hauk WD. Amplification of noncoding chloroplast DNA for phylogenetic studies in lycophytes and monilophytes with a comparative example of relative phylogenetic utility from Ophioglossaceae. Mol. Phylo. Evol 2005;36:509–-522.(doi:10.1016/j.ympev.2005.04.018)

Steel MA, Szekely LA. Inverting random functions II: Explicit bounds for discrete maximum likelihood estimation, with applications. Siam J. Discrete Math 2002;15:562–-575.(doi:10.1137/S089548010138790X)

Taberlet P, Eric Coissac E, Pompanon F, Gielly L, Mique C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E. Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. Nucleic Acids Res 2006;35(3):e14. [PubMed: 17169982](doi:10.1093/nar/gkl938)

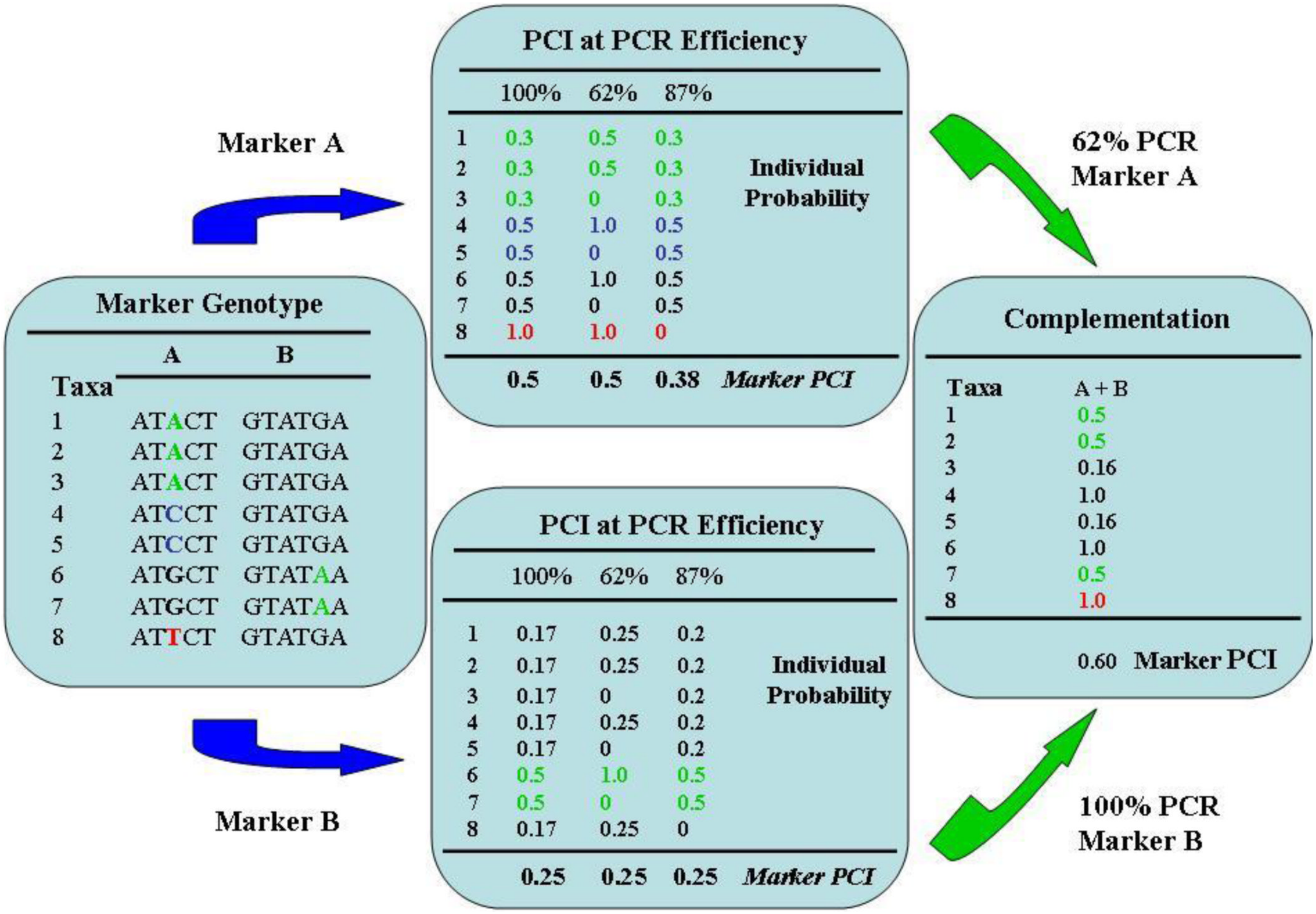

Figure 1.

Following the example in the text, the probability of correct identification (PCI) is outlined with two markers for 8 taxa. Panel I shows genotypes at two markers for 8 species – nucleotide changes are color coded as in the text. The panels II and III outline individual probability of assignment under different PCR recovery rates for each species, as well as a combined PCI for the marker at each of the three PCR recovery rates. Panel IV outlines the PCI for when marker “A” and “B” are used as a multi-locus barcode at a PCR recovery rate of 100% for marker “B” and 62% for marker “B”. Note that PCR failure affects PCI only when a marker is informative, such that PCI for marker “B“ does not decline with PCR failure, but marker “A” can when PCR failure affects individual identification sufficiently.

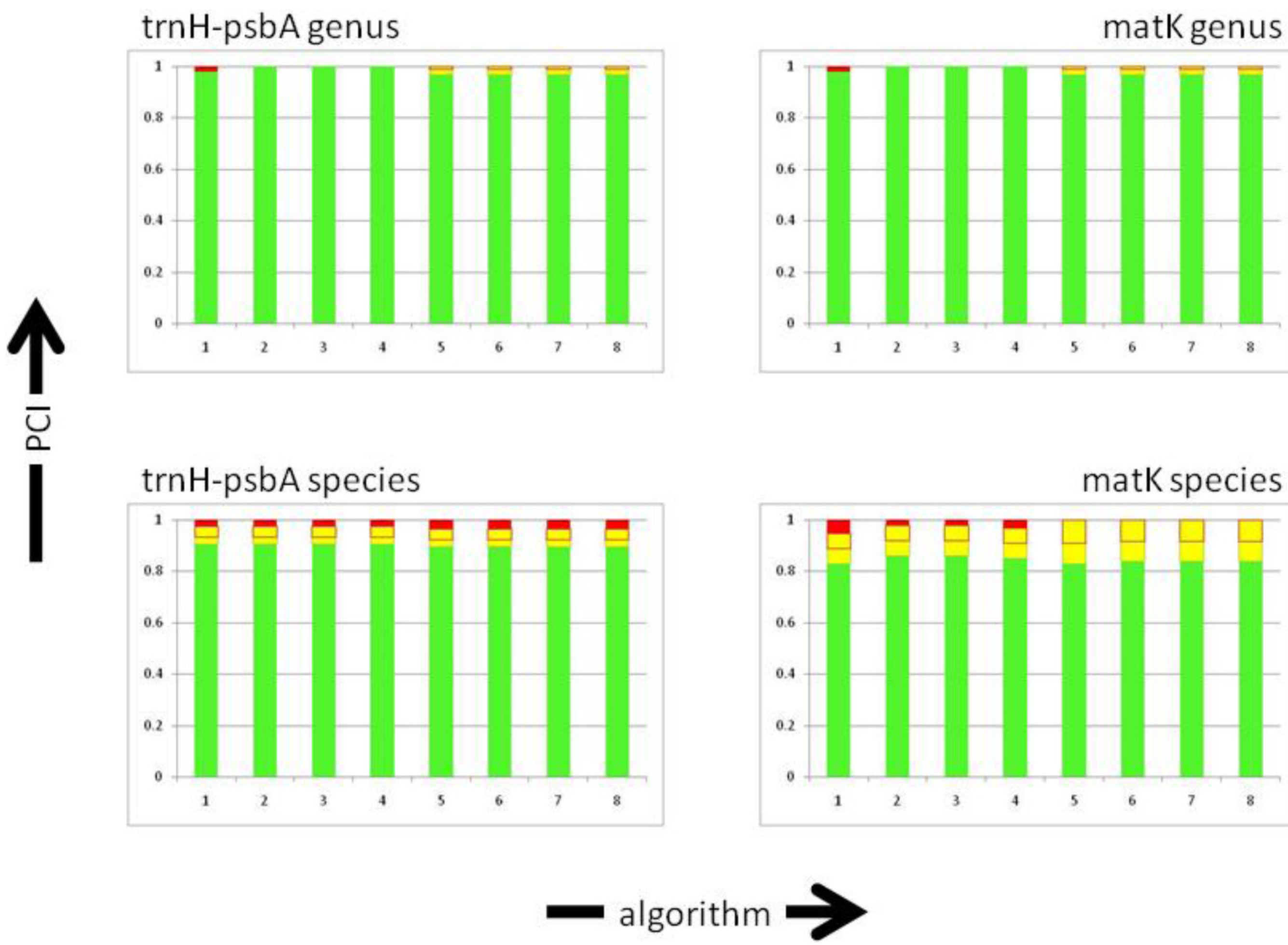

Figure 2.
PCI values for *trnH-psbA* and *matK* are plotted, comparing results for 8 different distance and alignment algorithms. For each marker, Figure 2 plots the PCI for assignment to the correct Genus and Species on the Y-axis against the algorithms, numbered on the X-axis. From the bottom up, the taxon assignments (genus or species) are subdivided by "correctness" into correct (in green), sometimes correct (either correct or incorrect), and never correct as described in the text. Data used was that presented in Lahaye & al. (2008b) for *trnH-psbA* and *matK*, using only the set of samples with PCR success at the corresponding marker. The algorithms examined were: 1 - Global Distance, 2 – Local Distance, 3 – Overlap Distance, 4 – Semi-Global Distance, 5 – Jukes-Cantor Distance, 6 – K2P Distance, 7 – Jin 1.0 Distance, 8 – Tamura Distance. Global alignment finds the best alignment of two complete sequences against each other; local alignment, of two subsequences; semi-global alignment, of one complete sequence against a subsequence of the other (or vice versa); and overlap alignment, of the left of one sequence against the right of the other (or vice versa). The distances normalized the alignment score by dividing by the length of the alignment.

Table 1

Summary of amplification success and divergence rates for six putative plant DNA barcodes. Each locus was tested on the same set of congeneric pairs of species. PCR amplifications used primers from Kress & Erickson (2007) for *trnH-psbA* and *rbcL*-a, from the RBG Kew website for *matK*, rpoC and rpoB, and from Hajibabaei, M.,(pers. Comm.) for CO1.

Region	*trnH-psbA*	*rbcL*-a	*rpoC*	*CO1*	*rpoB*	*matK*
Species pairs tested	48	48	48	48	48	46
Mean locus length (bp; standard deviation)	373 (147)	530 (27.5)	531 (31.9)	485 (n/a)	485 (15.5)	501 (18.4)
Percent PCR success	95.8%	94.8%	89.5%	79.2%	77.1%	39.3%
Mean percent sequence divergence (n; Std Dev)*	2.69% (43; 3.54)	1.29% (43; 2.07)	1.38% (40; 4.14)	0.34% (37)	2.05% (34; 3.65)	1.13% (14; 3.76)
Proportion of genera in which species were differentiated (n/n)**	82.6% (38/46)	69.8% (30/43)	60% (24/40)	35% (13/37)	61.8% (21/34)	64.3% (9/14)
Probability of Correct Assignment	0.791	0.662	0.537	0.277	0.476	0.253

* Mean percent sequence divergence between species pairs across genera that were successfully amplified (n = # of species pairs)

** Proportion of genera in which both species were successfully amplified and exhibited sequence divergence between species (n/n = # of genera in which species of a pair were differentiated/total # of pairs amplified)

Table 2

Selected comparisons of pairs of two loci combining *trnH-psbA* with *rpoB*, *rpoC*, *rbcL*-a, *matK*, and CO1 tested on 48 species pairs of land plants.

Region	*trnH-psbA + rbcL*-a	*trnH-psbA + rpoB*	*trnH-psbA + rpoC*	*trnH-psbA + CO1*	*trnH-psbA + matK*
Percent PCR success*	100% (96/96)	100% (96/96)	100% (96/96)	99% (95/96)	96% (92/96)
Proportion of genera in which species were differentiated (n/n)**	87.5% (42/48)	87.5% (42/48)	87.5% (42/48)	87.2% (41/47)	78.3% (36/46)
Probability of Correct Identification	0.875	0.875	0.875	0.872	0.75

*PCR amplification of either locus for both members of a generic pair is regarded as successful amplification for that generic pair

**Proportion of genera in which both species were successfully amplified and exhibited sequence divergence between species (n/n = # of genera in which species of a pair were differentiated/total # of pairs amplified)

Table A1

Species	Individuals In Species	Marker a	Marker b	Marker a+b
#1	2	A	A	AA*
#2	2	A	C*	AC*
#3	2	A	G*	AG*
#4	2	A	T*	AT*
#5	2	C	A	CA*
#6	2	C	?	C?
#7	2	C	?	C?
#8	3	G	?	G?
#9	3	G	?	G?
#10	80	T*	?	T?*

Table A2

Marker	Taxonomic Level	PCR Success	All Individuals
a	Species	0.1	0.1
a	Individual	0.8	0.8
b	Species	0.6	0.3
b	Individual	0.6	0.06
a+b	Species	0.5	0.5
a+b	Individual	0.88	0.88