International Journal of Biology; Vol. 7, No. 4; 2015
ISSN 1916-9671 E-ISSN 1916-968X
Published by Canadian Center of Science and Education

Discovery of Highly Accurate Plant DNA Barcodes via Novel Iterative Methodologies

Jaison Jain[1]

[1]Hunter College High School, New York, New York, United States of America

Correspondence: Jaison Jain, Hunter College High School, New York, New York, United States of America. Tel: 646-725-8320. Email: jaisonj708@gmail.com, jaisonjain@hunterschools.org

Received: July 13, 2015 Accepted: July 25, 2015 Online Published: August 23, 2015

doi:10.5539/ijb.v7n4p42 URL: http://dx.doi.org/10.5539/ijb.v7n4p42

Abstract

A DNA barcode is a short, variable segment of DNA used in species identification. A rapid and inexpensive molecular tool, DNA barcoding may allow for large-scale biodiversity assessments in the future, prompting logical and targeted conservation policies. However, a highly accurate DNA barcode for plants has not yet been found, hindering development of advanced DNA barcoding-based conservation paradigms and perpetuating loss of plant biodiversity.

Previous attempts to develop a plant barcode have examined only small fractions of the plant genome. In this study, a more rigorous methodology was devised: whole plant chloroplast genomes were iteratively analyzed in successive 500 base pair segments, such that *all* potential barcodes contained within the chloroplast genome were assessed for their ability to distinguish plants. Moreover, an algorithm was constructed to optimize this novel process, yielding a 17000% increase in efficiency. Both non-optimized and optimized methods uncovered two 500 bp regions of DNA (out of the 2950 tested) that had unprecedented, near-perfect identification accuracy in a comprehensive sample set of whole chloroplast sequences. These DNA barcodes may enable larger biodiversity assessments, aiding plant species preservation efforts. And, our algorithm may facilitate discovery of barcodes for other kingdoms of life, expanding the reach of our methodology.

Keywords: biodiversity, DNA barcoding, genomics, plants

1. Introduction

The current biodiversity crisis threatens the well-being of ecosystems and humankind. Baseline biodiversity levels and rates of species loss are largely unknown, resulting in unfocused conservation policies (Brooks, 2010; Krishnamurthy & Francis, 2012). In order to obtain these figures and thus better target conservation efforts, large-scale biodiversity assessments may be conducted (IUCN, 2000; Margules & Pressey, 2000; Myers et al., 2000). The relative expense and rarity of taxonomists, as well as the inherently time-consuming nature of taxonomy, has made this task unfeasible in previous years; however, the advent of DNA barcoding—the use of short, standardized fragments of DNA as a means of species identification—offers a potential solution (Smith et al., 2005; Alfonsi et al., 2013; Ardura et al., 2013; Nagy et al., 2013). Although controversial as a phylogenetic tool, DNA barcoding is widely used as a method of species identification, owing to its cost-effectiveness, efficiency, and functionality on a large scale (Hebert et al., 2003). Hence, DNA barcoding may likely be an integral tool that enables a global network of biodiversity monitoring in future conservation-oriented efforts (Smith et al., 2005; Krishnamurthy & Francis, 2012; Nagy et al., 2013; Cristescu, 2014).

However, there currently exists no universally accepted, highly accurate DNA barcode for the species identification of plants (Pennisi, 2007; Ledford, 2008; Hollingsworth et al., 2011; Mattia et al., 2011). The mitochondrial cytochrome c oxidase I gene (CO1 or cox1), a standard barcode for the animal kingdom (Hebert et al., 2003), is not sufficiently variable in plants to reliably distinguish between species (Chase et al., 2007). In fact, nuclear and mitochondrial DNA is generally unsuitable for plant barcoding (Pennisi, 2007). Plant barcodes currently in use are found in the chloroplast genome, which is more variable between plant species, but even these barcodes are associated with major limitations. For instance, the *rbcL* (RuBisCO large subunit) barcode region is 1400 base pairs long, whereas the ideal length is 300—800 bp (Chase et al., 2007; Ford et al., 2009). The *trnH-psbA* spacer is of inconsistent length and missing in many plant species, complicating sequence alignment (Chase et al., 2007). ITS (internal transcribed spacer), another commonly proposed barcode, frequently exists as multiple,

divergent copies within a single organism (Chase et al., 2007; Ford et al., 2009). Moreover, none of these segments has been determined to have species identification accuracy of greater than 70% (Chase et al., 2007; Ford et al., 2009). Although multi-segment barcodes resolve many of these issues, species identification accuracy is only marginally greater. The proposed *matK* + *rbcL* barcode, for instance, only has the ability to distinguish 72% of studied plant species (CBOL Plant Working Group, 2009), inadequate when compared to CO1, which averages upwards of 95% discrimination success in animal species (Thomas, 2009). Although there exist barcodes that work with great accuracy in specific plant clades (Pennisi, 2007), there is no highly accurate DNA barcode universally applicable to all plant species.

Numerous efforts in search of plant barcodes have been undertaken; however, they may not have employed sufficiently rigorous methods. For instance, the investigation of Ford et al. (2009) involved 41 coding regions of a single species' chloroplast genome (Ford et al., 2009). This study therefore did not account for all possible barcodes and did not recognize the potential utility of non-protein-encoding and uncharacterized "junk" regions in plant barcoding. The study of Kress et al. (2005) evaluated genomes of only two plant species, *Atropa belladonna* and *Nicotiana tabacumas*, far removed from a full-scale investigation (Kress et al., 2005). In addition, although introns, exons, and spacer regions were studied, other non-coding regions of the chloroplast were overlooked. Shaw et al. (2005) only examined 21 non-coding barcode regions (Shaw et al., 2005). Thus, a truly comprehensive screening of the plant chloroplast genome in search of a DNA barcode has not yet been done.

This present study, therefore, aimed to discover a highly accurate DNA barcode for plant identification using a more rigorous and thorough methodology, with the long-term prospect of combating plant biodiversity loss via more effective and focused conservation policies.

2. Materials and Methods

The goal of our methodology was to find a DNA barcode that could be practically implemented to identify plants. The methodology, therefore, tested for all essential attributes of an ideal DNA barcode, enumerated as follows:

1) universal presence among plant species;
2) demonstration of a "barcode gap," an indicator of perfect identification accuracy;
3) appropriately short length to enable DNA extraction (300-800 base pairs);
4) adjacently conserved primer sequences. (Kress et al., 2005; CBOL Plant Working Group, 2009; Board of Trustees, 2013).

Our methodology tested all 500 bp segments of a chloroplast genome alignment in an iterative fashion (successively repetitive) for the presence of a barcode gap. Segments that were not universally present were excluded. Afterwards, DNA regions flanking highly accurate barcodes were studied for their viability as primers.

2.1 Testing for the Barcode Gap

The barcode gap is present in a given set of pairwise genetic distances if and only if the maximum pairwise *intra*specific distance, a measure of variation within the same species, is less than the minimum pairwise *inter*specific distance, a measure of variation within different species. In other words, a DNA barcode with a barcode gap will always be more genetically similar to other barcodes of the same species than to barcodes of other species. Hence, the presence of a DNA barcode gap is indicative of an ideal DNA barcode with a hypothetical accuracy 100% in species identification. The barcode gap can be thought of as the 'gold-standard' of DNA barcodes. (Hebert et al., 2003; Meyer & Paulay, 2005).

The multiple sequence alignment of selected sequences (see *3.1. Sampling*) was performed by local MAFFT algorithms (Katoh et al., 2002). The alignment was then exported to MEGA6 (Tamura et al., 2011), which calculated pairwise interspecific and intraspecific distances.

A Bash script was written in the Unix shell of the Linux kernel to iteratively analyze 500 bp segments of the multiple sequence alignment for the barcode gap (maximum intraspecific distance < minimum interspecific distance). Successive brackets were instantiated every 125 base pairs in this "sliding window" of 500 bp. The 500 bp barcode length was applied in accordance with the ideal length of 300-800 bp (Kress et al., 2005), and the 125 bp shift of brackets used in order to test as many meaningfully distinct barcodes as possible within the practical limitations of MEGA6's computing capacity. In this manner, 2950 segments were evaluated for the presence of a barcode gap (bp 0-500, 125-625, 250-750, and so on for the entire length of the ~372,349 bp alignment).

Segments with barcode gaps were then tested for universality—that is, whether the barcode was present in >97% of sequences in the alignment (with "present" defined as presence of at least 475/500 bp).

Mann-Whitney U tests, corrected for multiple comparisons (Bonferroni correction), were performed in GraphPad Prism 6 to evaluate statistical significance values for each barcode that presented a barcode gap.

2.2 Additional Testing of Highest-Performing Barcodes

Further tests—namely, BLAST1, Nearest Distance, and Neighbor-Joining Tree Building tests—were performed using MEGA6's functions in order to confirm the universality and efficacy of barcodes experimentally found to have barcode gaps.

In BLAST1 tests, the species of known query barcode sequences (n = 866) inputted into BLAST (Basic Local Alignment Search Tool) ought to be consistent with the species of the top BLAST hit. Only hits having E values < 1 × 10^-5 were considered. Evidently, the query sequence itself cannot qualify as the top hit. In strict Nearest Distance tests, query sequences (n = 44) were aligned with the reference sample set (n = 822). The identity of a query sequence ought to be the same as the sequence having the smallest genetic distance from the query, given that the smallest distance value is below the 90th percentile of intraspecific distance values from the initial iterative analysis with that region. Otherwise, the identification is categorized as ambiguous or a failure. In strict Neighbor-Joining Tree-Building tests, a sequence's identity was determined by its placement on a MEGA6 neighbor-joining tree containing both the query sequences (n = 44) as well as the reference sequences (n = 822). The query sequence ought to be (a) contained within a monospecific clade and (b) of the same species as all other sequences within its monospecific clade (Ross, 2008).

2.3 Testing of the Algorithm

We devised an algorithm which, rather than searching for the barcode gap, calculated the percent effectiveness (in plant species identification) of each of the 2950 DNA segments, with the ultimate intent of pinpointing those segments with highest percent effectiveness. The algorithm's calculation of percent effectiveness of each segment was compared with the actual percent effectiveness of those same segments, each to the nearest percentage point. Actual values were determined using the BLAST1 method (see *3.3. Additional Testing*). BLAST tests were conducted for each 500 bp segment of the alignment, using each of the 866 original sequences. The number of accurate BLAST identifications out of 866 was equivalent to the actual percent effectiveness of that barcode. Percent error of the algorithm in predicting the accuracy of a given DNA barcode was then recorded.

In addition, the segments identified by the algorithm as having the highest percent effectiveness were compared to barcodes determined to have barcode gaps by the initial iterative methodology. If the algorithm identified the same regions as the iterative MEGA6 analysis (i.e, those segments with barcode gaps), the optimized algorithm would be considered an equally accurate alternative to the time-consuming iterative methodology.

2.4 Experimental Controls: Plant Barcodes Currently in Use

In order to insure that segments were not universally of disproportionately high accuracy in the restricted sample set of 866 sequences, plant barcodes currently in use (ex. *rbcL,* ITS*, matK*) underwent identical testing (initial iterative analysis, additional tests, algorithmic tests). These results were then compared to accepted identification accuracies in the scientific literature from their respective studies.

2.5 Primers

In order to screen for quality primers*,* a LAGAN alignment was generated (Brudno, 2007), which displayed the degree of similarity between sequences of the alignment. Potential primer sequences flanking the barcode ought to be highly conserved between different sequences. These flanking sequences were also assessed for GC-content, DNA repeats, and redundancy in the genome.

After assessment of potential primer sequences, NCBI Primer-BLAST was utilized in order to identify primers specific to the selected 500 bp barcode segments. The following primer sequences were identified: GAGAGTCGATACTCGGC and ATCCAATGCAACTACGC. Primer universality was assessed using the initial data set employed in the barcode exploration step, with a strict threshold of 19/20 base pairs in order to qualify for universality.

Following quality and universality assessments, primer success was analyzed in the initial sample set. Genomic DNA was extracted from fresh leaves obtained from the New York Botanical Gardens. The 50-μL PCR reactions contained PCR buffer, 0.5 mmol/L of each nucleotide, 0.5 μmol/L of each primer, 2.0 units Taq polymerase, and roughly 20–30 ng DNA (New England Biolabs, Ipswich, MA). After 5 min at 94°C, the PCR thermal cycling was organized as follows: 30 cycles of 30 s at 94°C, 40 s at 52°C, and 1 min at 72°C. After examination by electrophoresis on a 1% agarose gel, the PCR products were sequenced by Eurofins MWG Operon's DNA sequencing services. PCR/sequencing success for each sample was then evaluated with the following criteria:

Quality Value score >20, <1% low-quality bases, and <1% internal gaps and substitutions when aligning the forward and reverse reads.

3. Results

3.1 Sampling

Whole chloroplast genome sequences of various plants were obtained from GenBank. All specimens were vouchered and credibly identified. 822 samples were used for the main iterative analysis, encompassing 411 species (SI Table 1). 44 samples were used in additional testing, encompassing 22 species.

For primer universality tests, vouchers were obtained from the New York Botanical Gardens. Sample set encompassed 32 species of 32 distinct families (SI Table 2). Species were included from each of the following plant subdivisions: Bryophyta, Pteridophyta, Gymnospermae, Angiospermae.

3.2 Discovery of Two Novel Barcodes with Barcode Gaps

In a multiple sequence alignment of 372,349 base pairs, a total of 2950 segments (each 500 bp in length) were evaluated in the initial iterative analysis. Five (0.17%) of these segments presented with complete barcode gaps, indicating 100% species identification accuracy in the sample set (Figure 1). Two of these five barcodes were also universally present (98% of sequences) in the sequence alignment (SI Table 1, Figure 2). Universality rate was lowest in non-embryophytes and bryophyta.

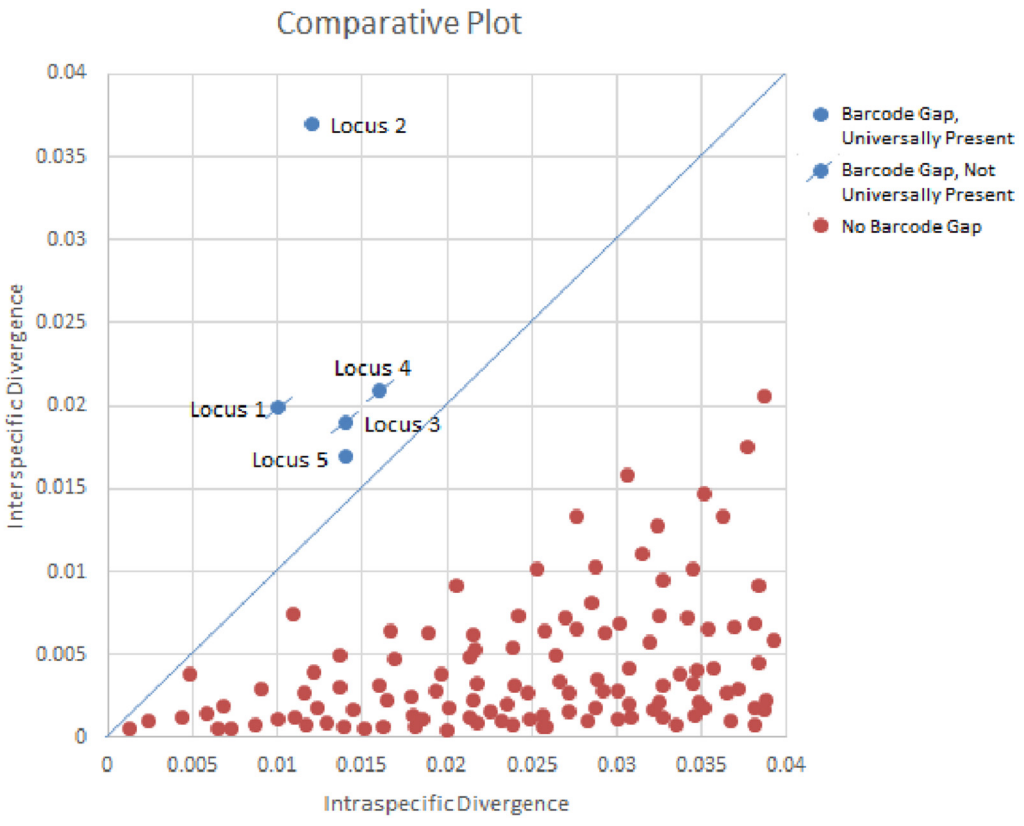

Figure 1. 1:1 Plot of Regions Found with Barcode Gaps

Note. Each data point comprises a cluster of approximately one-hundred 500 bp segments. At the 1:1 line shown in the graph, maximum intraspecific distance is equal to minimum interspecific distance. Any data points above the line have a barcode gap. Those below the 1:1 do not have a barcode gap.

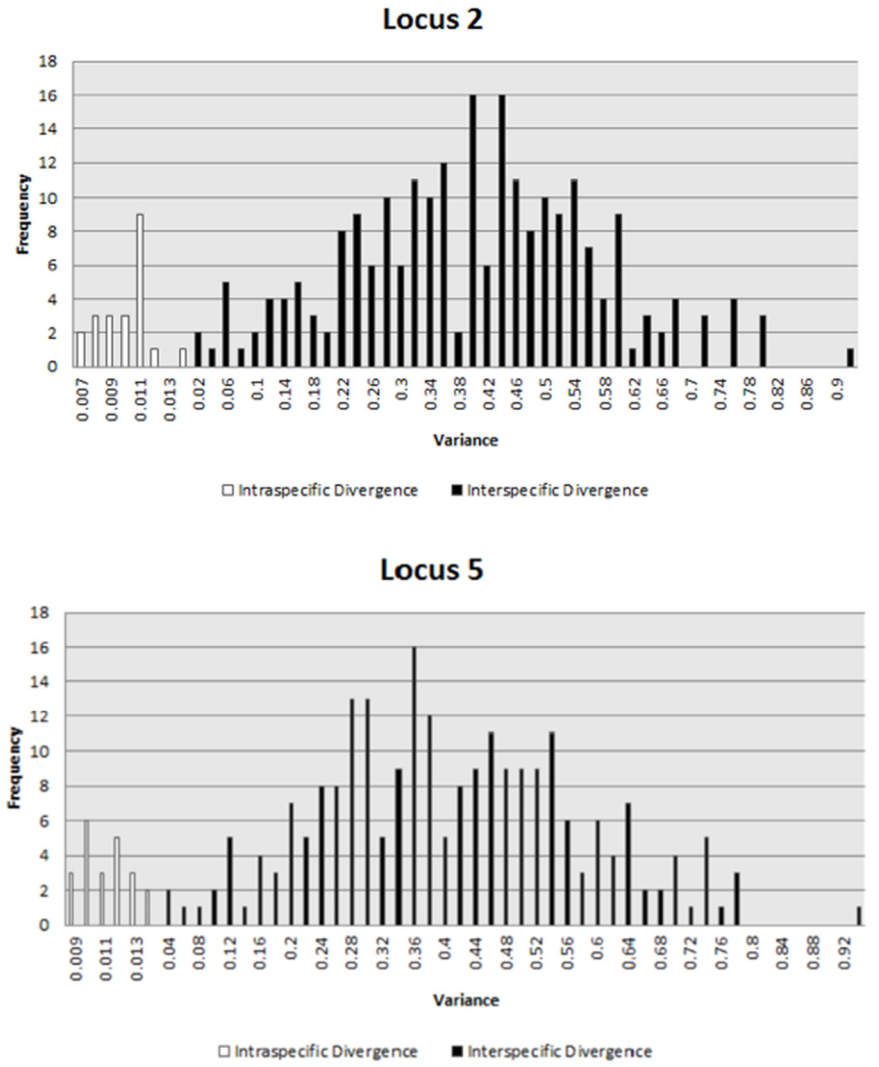

Figure 2. Histogram Representation of Barcode Gaps

Note. Both segments represented here have barcode gaps and universality. All pairwise interspecific distances are greater than interspecific distances.

All distance measures were values from 0 to 1, with 0 being representative of no evolutionary distance/variation and 1 being representative of the most variation possible (no mutually similar nucleotides between two sequences). Barcode gaps were, on average, 0.0096 p-distance units wide (gap = minimum interspecific distance - maximum intraspecific distance). All differences between interspecific and intraspecific distances were highly significant ($p < 0.001$).

These two particular barcode sequences were searched in the GenBank database. However, only whole plant chloroplast genomes resulted, indicating that the two regions are unnamed and not functionally characterized in any GenBank-associated study.

Both barcode segments were also flanked by conserved sequences, as indicated by the LAGAN alignment. The interspecific variation of the segments were seen as a “dips” in conservation levels (Figure 3). Immediately surrounding the dip, however, were highly conserved flanking sequences >96% similar, on average, in all examined plant species. In addition, manual analysis confirmed an optimal GC-content of 43%, and based on MEGA6’s local calculations, no repeats/redundancy existed for these two barcodes. Theoretically, therefore, quality primers can be generated for these barcodes. And, following primer construction with Primer-BLAST, DNA sequencing testing revealed PCR and DNA sequencing success rate of 94%, or 30/32 specimens (SI Table 2). Amplification and sequencing were least successful in bryophytes.

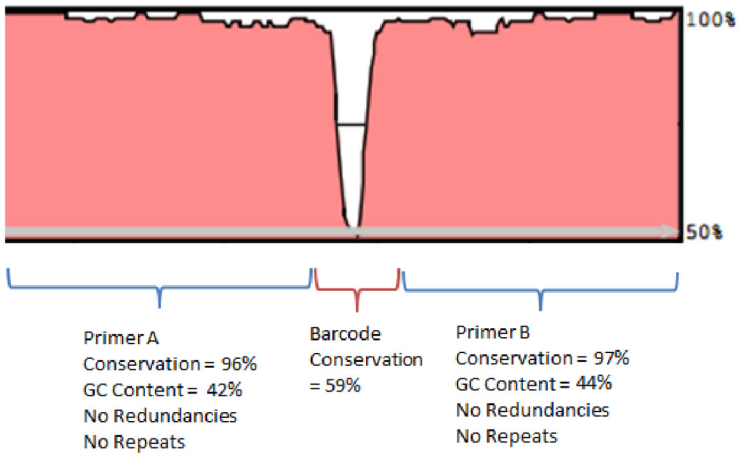

Figure 3. Tests for Viable Primers

Note. Sequences immediately flanking the two discovered barcode sequences were analyzed for GC-content, genetic conservation, redundancies within the genome, and DNA repeats. All tests indicated that both barcodes had viable primers. A representation of one of the barcode's primers is depicted.

Thus, our comprehensive, iterative methodology uncovered two novel DNA barcodes that were (1) of an optimal length of 500 bp, (2) universally present in all examined plant chloroplast sequences, (3) amplifiable with viable primers, and (4) high accuracy in species identification (indicated by the barcode gap), thereby satisfying all requirements of a functional plant DNA barcode.

3.3 Confirmation of Identification Accuracy via Additional Tests

BLAST1 tests confirmed both barcode segments to be 100% effective in plant species identification in the data set, consistent with the presence of a barcode gap. For both barcodes, all 866 query sequences matched their respective top BLAST hits. The strict Nearest Distance test demonstrated that one segment correctly identified query sequences in 100% of cases, and the second segment in 93% of cases (7% ambiguous, with a smallest genetic distance above the 90th percentile maximum required for non-ambiguous identifications). The Neighbor-Joining Tree test demonstrated 100% species identification for one barcode: all query sequences were placed in monospecific clades of the correct species, even in the presence of closely-related species. The second barcode exhibited 94% for this test.

3.4 Current Plant Barcodes Perform Poorly in Sample Set

In comparison to the two discovered segments, current plant barcodes (*rbcL+matK, trnH-psbA*, ITS, etc.) displayed no barcode gap: there was significant overlap between interspecific and intraspecific variation (ex. *rbcL* shown in Figure 4). Furthermore, they performed significantly poorer in BLAST1, Nearest Distance, and NJ-Tree-Building tests (Table 1). Current plant barcodes, in summary, do not perform as well in species discrimination as those segments with barcode gaps uncovered in the study.

Table 1. Performance of Discovered vs. Current Plant Barcodes.

Segment	Barcode Gap in Data Set	BLAST1	Nearest Distance	Neighbor Joining Tree
Barcode 1	Yes ($p < 0.001$)	100%	100%	100%
Barcode 2	Yes ($p < 0.001$)	100%	93% (7% ambiguous)	94%
rbcL	No	73%	73%	60%
matK	No	73%	67%	60%
rbcL* + *matK	No	67%	60%	53%
ITS	No	73%	67%	60%
trnH-psbA	No	67%	67%	67%

Note. In summary, two barcodes were found via the comprehensive iterative analysis that had barcode gaps and performed exceptionally well in all additional tests. In the same sample set of sequences, plant barcodes currently in use have overlap between intra- and inter-specific variation and perform poorly in additional testing.

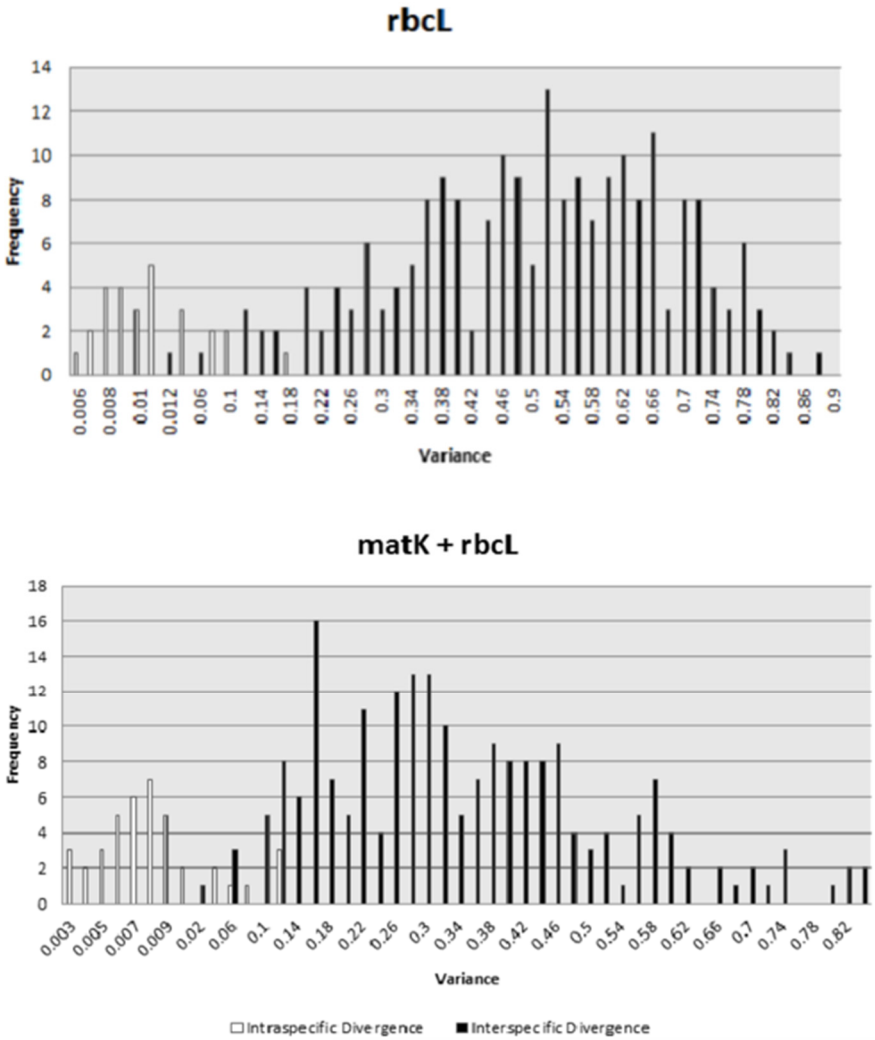

Figure 4. Histogram Representation of Current Plant Barcode Overlaps

Note. In contrast to those segments found to have barcode gaps (Figure 2), no plant barcodes currently in use have the barcode gap in the experimental sample set (ex. *rbcL* and *matK*+*rbcL,* as shown).

3.5 Development of a Faster Screening Process (Algorithm)

Because the original iterative analysis was very time-consuming, we sought to devise an alternative method of locating DNA barcodes that was faster and more efficient, without sacrificing accuracy of results. Thus, a modular algorithm was developed and, afterwards, the algorithm was tested.

We sought to devise an alternative method to identity universally-effective barcodes through screening of entire genomes that was faster and more efficient than the methods detailed above, without sacrificing accuracy of results. Thus, the following algorithm was tested:

$$E(p) = \frac{1}{\sum_{k=1}^{n}\binom{\sigma(\epsilon_k)}{2}}\sum_{k=1}^{n}\left[\sum_{j=1}^{\sigma(\epsilon_k)-1} P(\epsilon_{k,j},\epsilon_{k,j+1})\right] \times \frac{1}{\left[\left[\sum_{k=1}^{n}\sigma(\epsilon_k)\right]-1\right]^2 - \sum_{k-1}^{n}\binom{\sigma(\epsilon_k)}{2}}\sum_{k=1}^{n}\left[\sum_{j=1}^{\sigma(\epsilon_k)}\left[\sum_{i=k+1}^{n}\left[\sum_{h=1}^{\sigma(\epsilon_i)} P(\epsilon_{k,j},\epsilon_{i,h})\right]\right]\right],$$

where n is equivalent to the number of species present in the alignment, $\epsilon_{a,b}$, denotes instance number b of species number a, $\sigma(\epsilon_a)$ returns the number of instances of species ϵ_a (i.e. its maximum value of b), and $P(\epsilon_{a,b}, \epsilon_{c,d})$ returns the weighted conservativeness of the given genetic locus in the pairwise alignment of $\epsilon_{a,b}$ and $\epsilon_{c,d}$.

The algorithm is modular, namely each term in the global product carries a discrete meaning.

$$\sum_{k=1}^{n}\binom{\sigma(\epsilon_k)}{2}$$

denotes the number of possible intraspecific pairwise alignments.

$$\sum_{k=1}^{n}\left[\sum_{j=1}^{\sigma(\epsilon_k)-1} P(\epsilon_{k,j}, \epsilon_{k,j+1})\right]$$

represents the sum of the weighted averages of the conversation values on the target site in all possible intraspecific pairwise alignments,

$$\left[\left[\sum_{k=1}^{n}\sigma(\epsilon_k)\right]-1\right]^2-\sum_{k=1}^{n}\binom{\sigma(\epsilon_k)}{2}$$

denotes the number of possible interspecific pairwise alignments, and

$$\sum_{k=1}^{n}\left[\sum_{j=1}^{\sigma(\epsilon_k)}\left[\sum_{i=k+1}^{n}\left[\sum_{h=1}^{\sigma(\epsilon_i)} P(\epsilon_{k,j}, \epsilon_{i,h})\right]\right]\right],$$

represents the sum of the weighted averages of the conservation values on the target site in all possible interspecific pairwise alignments.

Though the algorithm provides an effective method by which the weighted conservativeness of a given locus in pairwise alignment can be resolved with many others, applying it to an entire 500 bp locus would prove inefficient and exhaust computing resources on the order of magnitude of the systematic iterative analysis.

However, by analyzing only a small fragment of the entire query genetic locus at a time, 2-3 bp segments, the entire analysis can be completed in a fraction of the previous time. This is done by integrating the calculated efficacy of a given DNA fragment with that of an adjacent one, and continuing the integration along the length of the barcode. This resolves the efficacy of hundreds of pairwise alignments entirely into that of an exponentially smaller data set. This process, when continued through what effectively becomes a binary tree, eventually culminates in the binary resolution of large DNA fragments into a globally calculated efficacy for the barcode itself.

Note that the algorithm must be scaled by a factor of the multiplicative inverse of the first global product minus 1 on each branch of the binary tree.

3.6 Optimization of Iterative Analysis via Novel Algorithm

Computer scripts for the original iterative analysis (which employed MEGA6's genetic distance calculations) ran for approximately 125 continuous hours. Furthermore, if only 5 more sequences were added to the alignment, all available RAM would have been depleted and computer scripts terminated. Thus, a method was needed to speed up processing and enable analysis of barcodes for larger sequence alignments in the future.

In total, only 43 minutes were necessary for the algorithmic process, comprising a 17000% increase in efficiency in comparison with the initial iterative analysis. The multiple sequence alignment, percent effectiveness calculations for each of the 2950 individual segments, and pinpointing of the most accurate DNA barcodes took the algorithm a total of 43 minutes to process completely.

3.7 Algorithm Retains Accuracy of Initial Analysis

Percent error of the algorithm in its percent effectiveness calculations was 4.87% (relative to the initial non-algorithmic method). However, it was observed that percent error of the algorithm's calculations was significantly lower for barcodes of higher percent effectiveness. The algorithm had only a 1.76% error rate when applied to barcodes with effectiveness 80-100% (Figure 5). This observation implies that most barcodes that researchers are concerned with (*i.e,* those that are highly effective) will have very accurately predicted percent efficacies. Furthermore, the algorithm identified the same barcodes as having *highest* percent effectiveness. Of 2950 segments, the algorithm pinpointed both regions with barcode gaps as having 100% effectiveness in species identification.

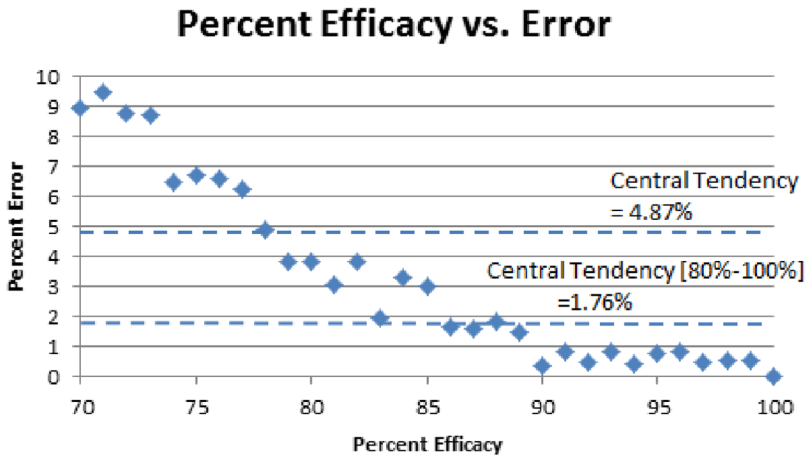

Figure 5. Accuracy of Algorithm vs. Main Iterative Analysis

Note. The algorithmic methodology is > 95% accurate in its calculations of DNA barcode effectiveness (error rate = 4.87%). Further, when applied to more effective barcodes (80-100% actual accuracy), the algorithm is > 98% accurate. Thus, it is a viable alternative to the initial iterative analysis.

4. Discussion

The present study comprised a comprehensive analysis in search of a plant DNA barcode. Unlike earlier investigations, multiple whole chloroplast genomes were studied, and analysis was done in a non-discriminatory fashion. All portions of the genome were screened. Two plant DNA barcodes with unprecedented, near-perfect accuracies in the data set were located, though the study must be expanded to a broader sample set before real-world use. These two 500 bp segments were found to have barcode gaps in the consulted data set. Further tests on additional sequences confirmed accuracy. Furthermore, these two short barcodes were universally present in all chloroplast sequences studied, and were flanked by highly conserved and viable primer sequences. Plant DNA barcodes currently in use today performed poorly in the sample set used, lacking barcode gaps and failing many of the additional tests.

We speculate that these two regions had not previously been recognized in plant barcoding because they are not protein-encoding, not functionally characterized, and not present as individual sequences in GenBank (only as part of whole chloroplast genome sequences). The likelihood of finding such DNA regions is therefore meager without a global sequential comparison of complete chloroplast genomes.

A comprehensive catalog of these two barcodes in a greater quantity of plant species must be constructed before the present findings may be of use. Given the growing quantity of whole genome sequences (from which these barcodes can easily be located) and facile amplification of these two DNA segments, the development of the aforementioned catalog of barcodes is feasible. And, the advent of next-generation sequencing methods that improve upon the traditional Sanger sequencing procedure will facilitate and simplify the rapid construction of large-scale DNA barcode libraries (Shokralla et al., 2014). Regardless, our algorithm may be utilized in evaluating and comparing DNA barcodes currently in use today in an efficient and exhaustive manner. Moreover, the algorithm may be useful in rapidly locating accurate DNA barcodes for small clades or taxons, perhaps to be employed by specialists of particular organisms.

Our long-term prospect is that upon further evaluation and cataloging, these two barcodes be used in aiding rapid biodiversity assessments, thus elucidating where rates of biodiversity loss are greatest and where conservation policies must be targeted. In the present day, such inexpensive and rapid mapping of plant biodiversity is not feasible (Brooks, 2010; Krishnamurthy & Francis, 2012). Furthermore, species extinction is currently occurring at ~1000x its natural rate (Pimm et al., 2014). And, with most plant life on Earth yet to be discovered, many plant species may vanish before we ever reap their potential medicinal, economic, and scientific value (Krishnamurthy & Francis, 2012; Scheffers et al., 2012). DNA barcoding, however, in addition to being a tool for species identification, can also be utilized for rapid species discovery with the help of taxonomists (Hebert et al., 2005). Thus, the DNA barcodes uncovered in this study may be utilized to identify new plant species before their disappearance. In addition to facilitating protection and description of plant biodiversity, these DNA barcodes may be used for biomedicine, pest control, medicinal plant authentication, and a bio-literacy tool for the public (Ajmal et al., 2014).

Furthermore, the devised algorithm yielded a 170-fold increase in efficiency in comparison with the earlier method, without sacrificing accuracy of the resultant barcode "percent effectiveness" calculations. Hence, our algorithm can be used in place of the original method, allowing for easy expansion of the present study into a larger sample set of more chloroplast sequences and species. And, unlike the original iterative analysis, restricted to 500 bp segments, our algorithmic methodology enables analysis of barcodes with a greater variability with regards to sequence length. In addition, the algorithm can be applied to other kingdoms of life—such as fungi and insects—for which more accurate barcodes may exist. Perhaps a universal, "barcode of life" can be found using our algorithm that encompasses all species on Earth. These applications of the algorithm will be explored in future studies.

Acknowledgements

I would like to acknowledge Julian Glowacz for his guidance on algorithm design and Ms. Gilana Reiss for her general research consultations and guidance.

There are no conflicts of interest.

References

Ajmal, A. M., Gyulai, G., Hidvégi, N., Kerti, B., Al Hemaid, F. M., Pandey, A. K., & Lee, J. (2014). The changing epitome of species identification – DNA barcoding. *Saudi Journal of Biological Sciences, 21*, 204-231. http://dx.doi.org/10.1016/j.sjbs.2014.03.003.

Alfonsi, E., Méheust, E., Fuchs, S., Carpentier, F. G., Quillivic, Y., Viricel, A., Hassani, S., & Jung, J. L. (2013). The use of DNA barcoding to monitor the marine mammal biodiversity along the French Atlantic coast. *ZooKeys, 365*, 25-48. http://dx.doi.org/10.3897/zookeys.365.5873

Ardura, A., Serge, P., & Garcia-Vazquez, E. (2013). Applications of DNA barcoding to fish landings: authentication and diversity assessment. *ZooKeys, 365*, 49-66. http://dx.doi.org/10.3897/zookeys.365.6409

Board of Trustees of the Royal Botanic Gardens, Kew. (2007). Establishing a standard DNA barcode for land plants. Retrieved from http://www.kew.org/science-conservation/research-data/science-directory/projects/establishing-standard-dna-barcode-land

Brooks, T. (2010). *Conservation Planning and Priorities.* In N. S. Sodhi, & P. R. Ehrlich (Eds.), *Conservation biology for all* (pp. 199-219). Oxford: Oxford University Press.

Brudno, M. (2007). An introduction to the Lagan alignment toolkit. *Methods Mol Biol., 395*, 205-220. http://dx.doi.org/10.1007/978-1-59745-514-5_13

CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, 106*, 12794-12797. http://dx.doi.org/10.1073/pnas.0905845106

Chase, M. W., & Fay, M. F. (2009). Barcoding of plants and fungi. *Science, 325*, 682-683. http://dx.doi.org/10.1126/science.1176906

Chase, M. W., Cowan, R., Hollingsworth, P., Berg, C., & Madrinan, S. (2007). A proposal for a standardised protocol to barcode all land plants. *Taxon., 56*, 295-299.

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities. *Trends in Ecology and Evolution, 29*, 566-571. http://dx.doi.org/10.1016/j.tree.2014.08.001.

Ford, C. S., Ayres, K. L., Toomey, N., Haider, N., Van Alphen Stahl, J., Kelly, L. J., & Wilkinson, M. J. (2009). Selection of candidate coding DNA barcoding regions for use on land plants. *Botanical Journal of the Linnean Society, 159*, 1-11. http,//dx.doi.org/10.1111/j.1095-8339.2008.00938.x

Hebert, P. D., & Gregory, R. T. (2005). The promise of DNA barcoding for taxonomy. *Systematic Biology, 54*, 852-859. http://dx.doi.org/10.1080/10635150500354886

Hebert, P. D., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B, 270*, 313-321. http://dx.doi.org/10.1098/rspb.2002.2218

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE, 6*, e19254. http://dx.doi.org/10.1371/journal.pone.0019255

International Union for Conservation of Nature. (2000). *A guide to the assessment of biological diversity*.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on Fast Fourier Transform. *Nucleic Acids Research, 30*, 3059-3066.

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., & Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences, 102*, 8369-8374. http://dx.doi.org/10.1073/pnas.0503123102

Krishnamurthy, K. P., & Francis, R. A. (2012). Critical review on the utility of DNA barcoding in biodiversity conservation. *Biodiversity and Conservation, 21*, 1901-1919. http://dx.doi.org/10.1007/s10531-012-0306-2

Ledford, H. (2008). Botanical identities. *Nature, 451*, 616. http://dx.doi.org/10.1038/451616b.

Margules, C. R., & Pressey, R. L. (2000). Systematic conservation planning. *Nature, 405*, 243-253. http://dx.doi.org/10.1038/35012251

Mattia, F.D., Bruni, I., Galimberti, A., Cattaneo, F., Casiraghi, M., & Labra, M. (2011). A comparative study of different DNA barcoding markers for the identification of some members of Lamiacaea. *Food Research International, 44*, 693-702. http://dx.doi.org/10.1016/j.foodres.2010.12.032

Meyer, C. P., & Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology, 3*, e422. http://dx.doi.org/10.1371/journal.pbio.0030422

Myers, N., Mittermeier R. A., Mittermeier C. G., da Fonseca G. A., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature, 403*, 853-858. http://dx.doi.org/10.1038/35002501

Nagy, Z. T., Sonet, G., Mortelmans, J., Vandewynkel, C., & Grootaert, P. (2013). Using DNA barcodes for assessing diversity in the family Hybotidae. *ZooKeys, 365*, 263-278. http://dx.doi.org/10.3897/zookeys.365.6070

Pennisi, E. (2007). Wanted: A barcode for plants. *Science, 318*, 190-191. http://dx.doi.org/10.1126/**science**.

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., Raven, P. H., Roberts, C. M., & Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science, 344*, 1246752. http://dx.doi.org/10.1126/science.1246752

Ross, H. A., Murugan, S., & Li, W. (2008). Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology, 57*, 216-230. http://dx.doi.org/10.1080/ 10635150802032990

Scheffers, B. R., Joppa, L. N., Pimm, S. L., & Laurance, W. F. (2012). What we know and don't know about Earth's missing biodiversity. *Trends in Ecology and Evolution, 27*, 501-510. http://dx.doi.org/10.1016/j.tree.2012.05.008

Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., Liu, W., Miller, J., Siripun, K. C., Winder, C. T., Schilling, E. E., & Small, R. L. (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. http://dx.doi.org/10.3732/ajb.92.1.142

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources, 14*, 892-901. http://dx.doi.org/10.1111/1755-0998.12236

Smith, M. A., Fisher, B. L., & Hebert, P. D. (2005). DNA barcoding for effective biodiversity assessment. *Philosophical Transactions of the Royal Society, 360*, 1825-1834. http://dx.doi.org/10.1098/rstb.2005.1714

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2011). Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution, 28*, 2731-2739. http://dx.doi.org/10.1093/molbev/msr121.

Thomas, C. (2009). Plant bar code soon to become reality. *Science, 325*, 526. http://dx.doi.org/10.1126/science.325_526