EVALUATION OF THE CHLOROPLAST BARCODING MARKERS BY MEAN AND SMALLEST INTERSPECIFIC DISTANCES

DA-CHENG HAO[1*], PEI-GEN XIAO[2*], YONG PENG[2], JINGQUN DONG[1] AND WENXIA LIU[1]

[1]Biotechnology Institute, School of Environment, Dalian Jiaotong University, Dalian 116028, China
[2]Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Beijing 100193, China
Corresponding author's e-mail: hao@djtu.edu.cn

Abstract

It is more difficult in plants than in animals to distinguish between species using barcoding loci, thus it is vital to investigate the barcoding potential of various chloroplast (cp) regions besides the well-known plastid markers. It is also crucial to investigate whether the analytical metrics and depths of taxon sampling affect assessments of the barcoding utility. Here, we collected more than 9100 plant sequences, representing ten cp non-coding regions, by experiments and GenBank search. The inter- and intra-specific variations of the individual genus were calculated. The correlation between the number of sampled congeneric species and mean/smallest interspecific distance was quantitated. In addition, a selection of published barcoding data sets was reviewed to compare species discrimination of *trnH-psbA* in various plants. By comparing the barcoding gaps, we found that species pairs for different cp markers have variable size gaps between intra and interspecific genetic distances and the approach based on mean interspecific distances results in overblown estimates and may lead to misidentification of closely related species, which is problematic for all cp markers examined. It is also found that the smallest interspecific distances decrease with the number of species sampled in 6 out of 10 cp markers examined. The differences in the size of barcode gaps based on mean versus smallest interspecific distances may have major implications for the plant DNA barcoding. This study presents the first empirical evidence to advocate the simultaneous use of mean and smallest interspecific distances in assessing plant barcoding markers.

Introduction

The barcode of life is a short DNA sequence from a uniform locality on the genome, which is used for identifying species (Schori and Showalter, 2011). DNA barcoding is an emerging global standard for identifying species. Although efforts to identify a DNA barcode for discriminating among recognized species are less successful in plants than animals, in recent few years, researchers have reported many efficient cases for plant DNA barcoding (e.g., de Groot *et al*., 2011; Al-Qurainy *et al*., 2011; Kress *et al*., 2010). The non-coding chloroplast (cp) DNA regions, as well as the well-known nuclear ITS and the chloroplast protein-coding sequences, have been commonly used as the barcoding markers for a long time and for more species (e. g., Wang *et al*., 2010; Ma *et al*., 2010; Srirama *et al*., 2010; Hao *et al*., 2010a, b). However, studies comparing the barcoding utility of the non-coding cp DNA regions are lacking, and it is unknown whether the *trnH-psbA* intergenic spacer is superior to other non-coding regions. Since Shaw *et al*., (2007) suggested that nine newly explored cp non-coding regions, i.e., *petL-psbE* (M), *psbJ-petA* (N), *3'trnV-ndhC* (P), *psbD-trnT* (R), *atpI-atpH* (S), *trnQ-5'rps16* (T1), *3'rps16-5'trnK* (T2), *ndhF-rpl32* (V1),and *rpl32-trnL* (V2), offer better levels of variation at low taxonomic levels, the number of these sequences in NCBI GenBank has been rapidly increasing. In this study, we report both mean and the smallest interspecific distances for congeneric species and compare the barcoding utility of the classical *trnH-psbA* spacer and the new potential barcoding markers.

Materials and methods

DNA extraction, PCR amplification, and sequencing: Leaves were sampled from two individuals of *Ilex cornuta Lindl*, *I. latifolia Thumb*, *I. kaushue S. Y. Hu* (*I. kudingcha C.J. Tseng*), *I. pentagona S.K. Chen*, *I. paraguariensis*, and *Camellia sinensis*, respectively. Genomoc DNA was extracted from the leaves by Universal genomic DNA extraction kit 3.0 (Takara, Dalian, China). We used routine PCR, with no more than three attempts per sample to recover a PCR amplicon for all samples. The PCR cycling conditions were previously described (Shaw *et al*., 2007), with adjustment of lower annealing temperatures and more cycles when necessary. Primer pairs for each of the chloroplast regions are listed in Table 1 (Shaw *et al*., 2007). Successful PCR products were purified and transferred to a TA cloning vector pMD19-T (Takara). Sequencing was performed on an ABI 3730 Genetic Analyzer (Applied Biosystems). Sixty one cp sequences were deposited in GenBank under accession numbers FR849937–FR849997.

Sequence editing, alignment, and calculation of sequence divergence: Sequences of ten chloroplast noncoding regions (M, N, P, R, S, T1, T2, V1, V2, and *trnH-psbA*) were retrieved from NCBI GenBank and aligned with Clustal W2 (Larkin *et al*., 2007). Sequence variability was calculated with the software MEGA 5.04 (Tamura *et al*., 2011) under the following conditions: model, maximum composite likelihood (MCL); substitutions to include, transition + transversion; uniform rates among sites; pattern among lineages, different (heterogeneous); gaps/missing data treatment, pairwise deletion. MEGA only calculates the interspecific distance for the genus with at least three species, and the intraspecific distance for the species with at least three conspecific sequences. The sequences of subspecies, variety, hybrid, and cultured strain were excluded in determining the mean and the smallest interspecific distances for congeneric species. The genetic distance gap (barcoding gap) between inter- and intra-specific distance

(MCL) was determined for ten cp markers examined. The overlap between intra- and inter-specific variability after deleting the 5% largest intraspecific and the 5% smallest interspecific distances was calculated (Meier *et al*., 2008). The correlation between the number of sampled congeneric species and mean/smallest interspecific distance was quantitated and subject to the statistical tests. Values with $p<0.05$ are regarded as significant.

Results and discussion

Explorations of appropriate DNA barcoding locus for plants are the most important issue for practical uses of the technique in recent years, hence studies on evaluation/comparison of DNA barcodes are highly needed. In order to enumerate the interspecific distances we aligned 9,166 GenBank sequences for 5,829 species of land plants (Table 1 and data available upon request) based on nucleotide sequences using Clustal W2 . For each sequence for the 681 species with multiple sequences in the data set, the mean intraspecific MCL distance was calculated using MEGA5.04. MCL method, unlike K2P model or uncorrected distance, can estimate evolutionary distances between all pairs of sequences simultaneously, with substitution pattern heterogeneities among lineages. The mean and the smallest interspecific distances for congeneric species were also determined for all sequences. We then calculated the overlap between intra- and interspecific variability after deleting the 5% largest intraspecific and the 5% smallest interspecific distances.

Table 1. Barcoding gaps and sequence variability (sequence divergence calculated by model maximum composite likelihood) for land plants.

	No. of species/genera /sequences	No. of species with intraspecific sequences (≥3)	Mean intraspecific variability (%)	Mean interspecific variability (%)	Smallest interspecific variability (%)	Overlap[a] btw intra- and mean interspecific variability (%)	Overlap[a] btw intra- and lowest interspecific variability (%)
petL-psbE (M)	127/19/285	42	0.338 ± 0.851	3.22 ± 4.72	0.680 ± 1.01	None(0[b])	0-0.5(60.0[b])
psbJ-petA (N)	487/25/599	28	0.721 ± 1.18	2.88 ± 2.77	0.894 ± 2.03	0.2-3.6(57.8[b])	0-3.6(84.2[b])
3'trnV-ndhC (P)	138/18/218	15	0.093 ± 0.158	2.1 ± 2.81	0.711 ± 1.59	0.1-0.4(29.4[b])	0-0.4(76.4[b])
psbD-trnT (R)	209/19/283	15	0.793 ± 2.12	6.25 ± 13.43	0.422 ± 0.654	0.3-2.9(72.2[b])	0-2.6(100.0[b])
atpI-atpH (S)	270/40/398	29	0.100 ± 0.138	1.27 ± 1.79	0.331 ± 0.943	0-0.4(46.9[b])	0-0.4(87.5[b])
trnQ-5'rps16 (T1)	654/52/793	44	0.202 ± 0.48	1.97 ± 3.04	0.330 ± 0.866	0-0.6(44.8[b])	0-0.6(87.7[b])
3'rps16-5'trnK (T2)	353/40/622	36	0.236 ± 0.324	2.62 ± 4.08	0.554 ± 1.21	0.1-1.0(36.3[b])	0-1.0(75.7[b])
ndhF-rpl32 (V1)	321/26/583	50	0.422 ± 0.685	2.83 ± 2.87	0.641 ± 1.36	0.5-0.8(12.5[b])	0-0.8(75.0[b])
rpl32-trnL (V2)	350/37/599	60	0.263 ± 0.423	2.24 ± 3.13	0.436 ± 0.992	0.2-0.8(40.0[b])	0-0.8(84.0[b])
trnH-psbA	2920/328/4786	362	0.533 ± 1.16	4.65 ± 8.54	0.732 ± 1.70	0-2.6(60.4[b])	0-2.6(90.0[b])
Total	5829/604/9166	681					

[a] Overlap after deleting the 5% largest intra- and 5% smallest interspecific values
[b] Proportion of pairwise congeneric distances in overlap range
btw= between

An assessment of the different barcoding gaps reveals that the approach based on mean interspecific distances yields exaggerated estimates (Table 1). The differences are particularly conspicuous for V1, M, and P. For V1, the mean interspecific MCL distance is 2.83% ± 2.87%, whereas the lowest interspecific distance is 0.641% ± 1.36%. Correspondingly, the overlap between intra- and interspecific variability is also artificially small for mean values (12.5% of pairwise congeneric distances in overlap range). For M, the overlap zone is absent based on mean values, while it is zero to 0.5% for smallest interspecific distances and 60% of pairwise congeneric distances fall into the overlap range. For the commonly used barcoding marker *trnH-psbA*, based on mean values, 60.4% of all pairwise congeneric interspecific distances fall into the interval, whereas it is 90% based on smallest interspecific values, thus increasing the chance of misidentification. Among the ten cp markers examined, *trnH-psbA* ranked second in the mean interspecific distance (after R), second in the smallest interspecific distance (after N), and third in the mean intraspecific distance (after R and N). It seems that *trnH-psbA* is not superior to other cp markers examined, at least in the sense of the barcoding gap. Kress *et al*. (2005) found that *trnH-psbA* ranked first in divergence value in 6 of the 8 genera and in 11 of the 14 species pairs, compared with the other eight plastid regions, including T2. However, the sampling in this early study is limited and the barcoding gap was not calculated. Although the difference between mean inter- and intra-specific distances in *trnH-psbA* is much greater than that in T2 and other cp markers (Fig. 1), the difference between smallest interspecific and mean intraspecific distances in *trnH-psbA* is not greater than that in other cp markers examined. The mean interspecific distances are thus a poor estimator for the smallest interspecific distances.

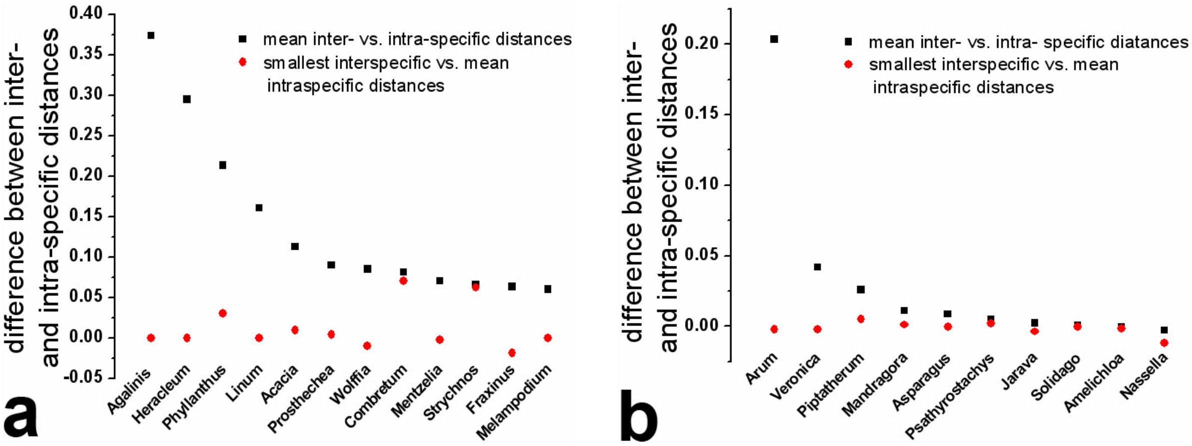

Fig. 1 Difference between inter- and intra-specific genetic distance (MCL) for species in the selected plant genera. a) *trnH-psbA*, b) T2. Each genus was represented by a minimum of three species. For *trnH-psbA*, 12 genera with the greatest difference between mean inter- and intra-specific distances are shown; for T2, all genera for which both metrics can be calculated are shown. Each species was represented by at least three sequences from different individuals.

A previously overlooked disadvantage of using the mean instead of the smallest interspecific, congeneric distances for quantifying the plant barcoding gap is that the latter decreases with taxon sampling. A denser taxon sample will generally decrease the smallest observed interspecific distance for a species. We tested these assumptions based on the 492 genera in our dataset that are represented by more than three species and in fact the smallest interspecific distances decrease with the number of species sampled (Fig. 2a) in six out of 10 cp markers, whereas the mean interspecific distances are not correlated with sampling intensity in all markers examined. Unlike COI sequences of animals (Meier *et al*., 2008), the difference between the smallest and the mean interspecific distances increases significantly with the number of congeneric species only in N (r = 0.567, P = 0.011; Fig. 2b), and in other cp markers, there is no statistically significant correlation between the 2 metrics, which might be due to the fact that the mean interspecific distances somewhat decrease with the number of species (Table 2, negative correlation between the 2 metrics in 8 out of 10 markers).

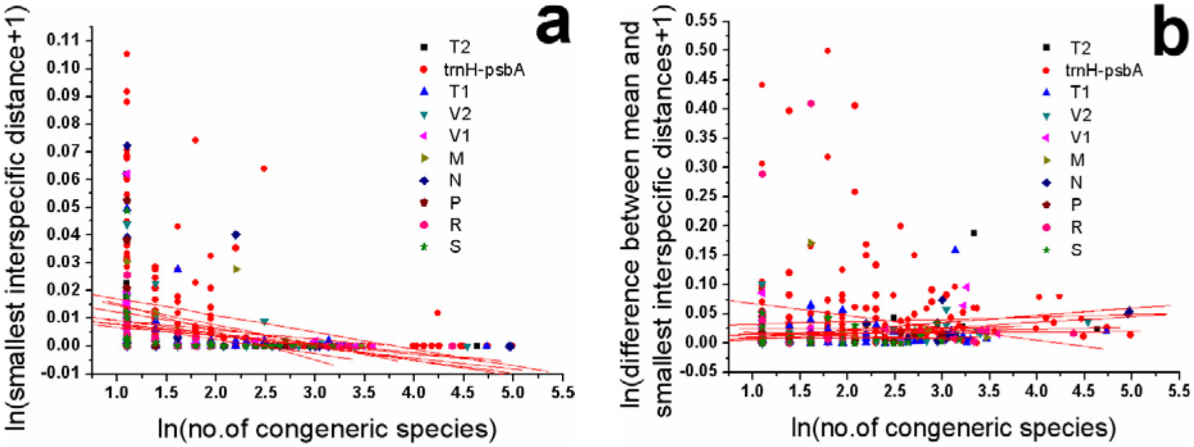

Fig. 2. The number of sampled congeneric species is significantly correlated with the smallest interspecific distance (a) but not the difference between mean and smallest interspecific distances (b). Correlation coefficients and P values are shown in Table 2.

In conclusion, this study uses plant barcoding data to explore the question of whether both mean and smallest interspecific distances should be utilized in the evaluation of the cp barcoding markers. We noted that relatively few cp non-coding data sets, except for *trnH-psbA*, exist for plants in which individuals from multiple populations from several species within a genus are sampled (Tables 1 & 2). Most typical studies in plant systematics usually consist of many species each with low (or no) population-level replication, and population genetic sampling, which often focuses on one or rarely a few species with more intensive population sampling. Only by more extensive species/population sampling for various cp non-coding markers can the hypothesis presented in this study be adequately tested. In plants, future DNA barcoding studies with denser species sampling and more intensive geographical sampling of species will be of great help. Differences in the size of barcode gaps based on mean versus smallest interspecific distances may have major implications for identifying query sequences using plant DNA barcoding. Proper measures of interspecific distances are not only important for distance-based identification techniques, but also important for recognizing the cryptic species. This study presents the first empirical evidence to advocate the simultaneous use of mean and smallest interspecific distances in assessing plant barcoding markers.

Table 2. Correlation of the number of sampled congeneric species, mean interspecific variability, and the smallest interspecific variability.

	No. of genera with sequences of ≥3 species	R btw ln (no. of congeneric species) and ln(smallest interspecific distance+1)	R btw ln (no. of congeneric species) and ln(mean interspecific distance+1)	R btw ln (no. of congeneric species) and ln(difference btw mean and smallest interspecific distances+1)
petL-psbE (M)	15	-0.252(P=0.362)	-0.055(P=0.844)	0.002(P=0.992)
psbJ-petA (N)	19	-0.382(P=0.106)	0.175(P=0.472)	0.567(P=**0.011**)
3'trnV-ndhC (P)	17	-0.473(P=0.055)	-0.422(P=0.091)	-0.288(P=0.260)
psbD-trnT (R)	18	-0.478(P=**0.044**)	-0.209(P=0.404)	-0.179(P=0.476)
atpI-atpH (S)	32	-0.473(P=**0.006**)	-0.194(P=0.285)	-0.097(P=0.595)
trnQ-5'rps16 (T1)	49	-0.384(P=**0.006**)	-0.067(P=0.644)	0.047(P=0.744)
3'rps16-5'trnK (T2)	33	-0.352(P=**0.044**)	0.179(P=0.316)	0.310(P=0.079)
ndhF-rpl32 (V1)	24	-0.521(P=**0.008**)	-0.093(P=0.662)	0.169(P=0.428)
rpl32-trnL (V2)	25	-0.335(P=0.101)	-0.001(P=0.993)	0.133(P=0.525)
trnH-psbA	260	-0.291(P<**0.0001**)	-0.018(P=0.766)	0.049(P=0.424)

Significant P values (<0.05) are in bold. R= correlation coefficient. Btw= between

References

Al-Qurainy, F., S. Khan, M.A. Ali, F.M. Al-Hemaid, M. Tarroum and M. Ashraf. 2011. Authentication of *Ruta graveolens* and its adulterant using internal transcribed spacer (ITS) sequences of nuclear ribosomal DNA. *Pak. J. Bot*., 43(3): 1613-1620.

de Groot, G.A., H.J. During, J.W. Maas, H. Schneider, J.C. Vogel and R.H. Erkens. 2011. Use of *rbcL* and *trnL-F* as a two-locus DNA barcode for identification of NW-European ferns: an ecological perspective. *PLoS One*, 6(1): e16371.

Hao, D.C., S.L. Chen and P.G. Xiao. 2010a. Sequence characteristics and divergent evolution of the chloroplast *psbA-trnH* noncoding region in gymnosperms. *J. Appl. Genet*., 51(3): 259-273.

Hao, D.C., S.L. Chen, P.G. Xiao and Y. Peng. 2010b. Authentication of medicinal plants by DNA-based markers and genomics. *Chin. Herb. Med*., 2(4): 250-261.

Kress, W.J., D.L. Erickson, N.G. Swenson, J. Thompson, M. Uriarte and J.K. Zimmerman. 2010. Advances in the use of DNA barcodes to build a community phylogeny for tropical trees in a Puerto Rican forest dynamics plot. *PLoS One*, 5(11): e15409.

Kress, W.J., K.J. Wurdack, E.A. Zimmer, L.A. Weigt and D.H. Janzen. 2005. Use of DNA barcodes to identify flowering plants. *Proc. Nati. Acad. Sci. U.S.A*., 102(23): 8369-8374.

Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson and D.G. Higgins. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21): 2947-2948.

Ma, X.Y., C.X. Xie, C. Liu, J.Y. Song, H. Yao, K. Luo, Y.J. Zhu, T. Gao, X.H. Pang, J. Qian and S.L. Chen. 2010. Species identification of medicinal pteridophytes by a DNA barcode marker, the chloroplast *psbA-trnH* intergenic region. *Biol. Pharm. Bull*., 33(11): 1919-1924.

Meier, R., G. Zhang and F. Ali. 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. *Syst. Biol*., 57(5): 809-813.

Schori, M. and A.M. Showalter. 2011. DNA barcoding as a means for identifying medicinal plants of Pakistan. *Pak. J. Bot*., 43(SI): 1-4.

Shaw, J., E.B. Lickey, E.E. Schilling and R.L. Small. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetics studies in angiosperms: the tortoise and the hare III. *Am. J. Bot*., 94(3): 275-288.

Srirama, R., U. Senthilkumar, N. Sreejayan, G. Ravikanth, B.R. Gurumurthy, M.B. Shivanna, M. Sanjappa, K.N. Ganeshaiah and R.U. Shaanker. 2010. Assessing species admixtures in raw drug trade of *Phyllanthus*, a hepato-protective plant using molecular tools. *J. Ethnopharmacol*., 130(2): 208-215.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol. Biol. Evol*., 28(10): 2731-2739.

Wang, W., Y. Wu, Y. Yan, M. Ermakova, R. Kerstetter and J. Messing. 2010. DNA barcoding of the Lemnaceae, a family of aquatic monocots. *BMC Plant Biol*., 10: 205.

(Received for publication 5 August 2011)