South African Journal of Botany xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

South African Journal of Botany

journal homepage: www.elsevier.com/locate/sajb

Phylogenetic utility of 19 low copy nuclear genes in closely related genera and species of caesalpinioid legumes

Marielle Babineau *, Edeline Gagnon, Anne Bruneau

Institut de Recherche en Biologie Végétale and Département de Sciences biologiques, Université de Montréal, 4101 rue Sherbrooke est, Montréal, Québec, Canada H1X 2B2

ARTICLE INFO

Available online xxxx

Edited by B-E Van Wyk

Keywords:
Caesalpineae clade
Caesalpinia s.l
Delonix
Duplication
Low copy nuclear genes
Phylogeny

ABSTRACT

Numerous studies have identified low copy nuclear genes (LCNG) with phylogenetic potential throughout angiosperms, several specifically focused on Leguminosae. However, phylogenetic resolution at the species- to subspecies-level is often inferred based only on a small subset of taxa scattered throughout the higher level study group. This study aims to reassess the phylogenetic resolution of 19 previously published nuclear regions in Leguminosae using 18 species from two clades of the Caesalpinieae representing both distantly related genera and closely related species. Nuclear regions were amplified and aligned throughout the sampled taxa. Sequences were cloned when polymorphism was noted. The plastid loci *matK*, *rps16*, *trnL*, *trnD–trnT* and the nuclear ribosomal ITS regions were also analyzed for comparison. Phylogenetic analyses using parsimony were performed on individual matrices. Three nuclear regions were eliminated due to the non-specificity of the primers (*RNAH*, *PTSB*, *GI*). Four regions showed lower resolution than predicted from previous studies (*MMK1*, *CYB6*, *RBPCO*, *EFGC*), while three revealed greater resolution than anticipated (*SQD1*, *AT103*, *EIF3E*). Three other markers indicated previously unidentified duplication events for the genus *Caesalpinia* s.l. (*ATCP* and *AROB*) and at the base of the *Caesalpinia* and *Peltophorum* clades (*CALTL*). Phylogenies reconstructed from the intron-spanning regions *AIGP*, *SHMT*, *AT103* and *EIF3E* are congruent with ITS and plastid data and show the best phylogenetic potential for studies of closely related species of caesalpinioid legumes. We present a screening strategy for the evaluation of LCNG for phylogenetic studies.

1. Introduction

One of the important criteria in choosing loci for molecular phylogenetic studies is having sufficient levels of variation to distinguish and group taxa, but not too much so as to cause problems of homology. Because of the inadequacy of the chloroplast genome to accurately resolve relationships among closely related plant genera and species in some groups, the last decade has been focused on generating, evaluating and testing low copy nuclear loci for species identification and evolutionary studies. The use of nuclear loci in phylogenetic studies is especially crucial in light of the large number of species thought to be directly implicated in, or evolved from, multiparental hybridization, ancient and recent polyploidy, and lateral gene transfer (Ishikawa et al., 2009; Rieseberg et al., 2000; Soltis et al., 2009). Low copy nuclear genes (LCNG) in plants are being developed rapidly (e.g., Armisen et al., 2008; Chapman et al., 2007; Choi et al., 2006; Doyle et al., 1999, 2002; Duarte et al., 2010; Ilut and Doyle, 2012; Li et al., 2008; Steele et al., 2008). However, most of these studies examine relationships based upon a few scattered species of the same family, or in a few studies, a diversity of species from distantly related angiosperm families: rarely have these LCNG been tested for their phylogenetic value outside of the group for which they were developed. Therefore the assessment of genus to subspecies level resolution is indirect and often speculative because of the lack of sampling of sister genera and species. In addition, these indirect estimations can be inaccurate due to specific properties of the nuclear genome such as hybridization, genome or gene duplication, levels of sequence variation of exons and introns, and incomplete lineage sorting. These phenomena can lead to an under or overestimation of the level of resolution. Hence, preliminary studies of nuclear genes are essential to evaluate the level of sequence variation of a specific locus for phylogenetic reconstruction and to estimate copy number and possible dynamics of gene duplication (Hughes et al., 2006; Soltis and Soltis, 1998).

* Corresponding author. Tel.: +1 514 343 6111x82356.
E-mail addresses: marielle.babineau@umontreal.ca (M. Babineau), edeline.gagnon@umontreal.ca (E. Gagnon), anne.bruneau@umontreal.ca (A. Bruneau).

0254-6299/$ – see front matter © 2013 SAAB. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sajb.2013.06.018

Here we assess the utility of 19 published LCNG for phylogenetic analysis, by comparing sequences across 18 species of caesalpinioid Leguminosae whose phylogenetic relationships are poorly resolved using chloroplast loci. Our test group includes two distantly related clades in the Caesalpineae clade, one for testing generic level variation, the other for examining species level variation. *Caesalpinia* s.l. consists of seven, possibly nine, distinct genera (Gagnon et al., in press; Lewis, 2005). The second group comprises a set of closely related species of the genus *Delonix*, and three related genera in the *Peltophorum* clade. The complementary structure of these two groups establishes a three-tier sampling strategy that allows us to investigate the level of variation at the species, the generic, and the clade level for several low copy nuclear genes and to determine their phylogenetic utility.

2. Materials and methods

2.1. Selection of taxa and markers

The first group sampled consists of ten species belonging to *Caesalpinia* s.l. that have been proposed to belong to different generic lineages (but several species have not been published under the new generic combination): *Caesalpinia pulcherrima* (genus *Caesalpinia* sensu stricto), *Guilandina bonduc*, *Coulteria pumila* (genus *Coulteria*), *Mezoneuron scortechinii*, *Libidibia paraguariensis*, *Caesalpinia mimosifolia* (genus *Erythrostemon*) and *Poincianella palmeri*, *Tara spinosa*, as well as *Caesalpinia decapetala* and *Caesalpinia madagascariensis*, whose generic affiliations are uncertain. The second group consists of five closely related species of *Delonix* (*Delonix pumila*, *Delonix floribunda*, *Delonix boiviniana*, *Delonix velutina* and *Delonix regia*), and *Colvillea racemosa*, which studies using chloroplast sequences have found to occur as nested within the *Delonix* clade (Haston et al., 2005; Hawkins et al., 2007; Simpson et al., 2003). Included within this second test group are the closely related *Conzattia multiflora* and the more distantly related *Parkinsonia aculeata*, see Supplementary Data for voucher information.

A total of 19 nuclear regions were chosen to be tested across our set of taxa. Thirteen of these (*RNAH*, *PTSB*, *MMK1*, *AIGP*, *tRALs*, *CTP*, *CALTL*, *ATCP*, *SHMT*, *EFGC*, *UDPGD*, *CYB6*, *RBPCO*) are from Choi et al. (2006) and were specifically designed for the family Leguminosae. Another six markers (*AGT1*, *GI*, *AROB*, *EIF3E*, *AT103*, *SQD1*) are COSII genes that were selected from Li et al. (2008) and which were developed to amplify across the angiosperms. Both intron-spanning regions and exon-only loci were selected.

Sequences for these nuclear regions also were retrieved from GenBank in order to evaluate the degree of resolution. *Glycine max*, *Glycine tomentella*, *Phaseolus vulgaris*, *Pisum sativum*, *Medicago truncatula*, *Vigna radiata*, *Lotus tenuis*, *Lotus corniculatus* var. *japonicus*, *Wisteria* sp. and *Wisteria sinensis* were selected from the Choi et al. (2006) study. Sequences for *P. aculeata* and *D. regia* from Choi et al. (2006) also were used in a subset of nuclear loci to verify that the correct sequences were indeed obtained in our own study. Sequences of *Prunus avium* and of *Ranunculus bulbosus* were selected from the study of Li et al. (2008). *G. max* or *G. tomentella* were used as outgroup taxa to root the trees for the nuclear loci developed by Choi et al. (2006), *P. avium* or *M. truncatula* were used to root the gene trees for markers from Li et al. (2008), and *P. aculeata* was used to root trees for the comparative matrices.

To assess the phylogenetic signal of these loci, results from the LCNG analyses were compared with results from a dataset of plastid markers (*trnL* intron, *matK* gene and *trnK*-3′ intron, *rps16*, *trnD–trnT* intergenic spacer) and a second dataset consisting of nuclear ribosomal internal transcribed spacer sequences (ITS). The complete dataset was assembled using 121 published sequences available from GenBank and 314 sequences generated in this study (Table A in Appendix).

2.2. Screening strategy for LCNG markers

An initial round of amplification was carried out on all taxa to screen for markers that produced a clear single band from the PCR reaction. These were sequenced directly. If the PCR was successful and free of polymorphisms, it was assumed that a single copy gene was amplified, with a single allele. If a single double-peaked nucleotide was detected on the sequencing chromatogram, the presence of two allelic sequences was deduced. If multiple double-peaked nucleotides were detected on direct sequencing chromatograms, this was interpreted as the presence either of two different alleles or of more than one gene copy. To distinguish these two possibilities, the samples were cloned, and three to six transformed colonies were sequenced.

2.3. Molecular methods

Extraction of DNA was carried out using a modified CTAB protocol, as described by Joly et al. (2006), or the QIAGEN DNeasy Plant Mini Kit (Mississauga, ON), following the manufacturer's instructions.

All PCR amplifications for nuclear and plastid loci were done in reaction volumes of 25 μl, with 1 × *Taq* DNA polymerase buffer without $MgCl_2$ (Roche Diagnostics, Indianapolis, IN), 3.0 mM of $MgCl_2$, 200 μM of each dNTP (Fermentas, Burlington, ON), 0.4 μM of each primer, 3 μg of bovine serum albumin (New England Biolabs, Ipswich, MA), 0.03% of tween-20, 3% of pure DMSO, one unit of *Taq* polymerase, and 50–300 ng of genomic DNA.

Plastid and nuclear ribosomal DNA regions were amplified according to previously published PCR cycling protocols: the *trnL* (UAA) intron was obtained following Bruneau et al. (2001) with primers trnL-C and trnL-D (Taberlet et al., 1991). The *matK* gene and 3′-*trnK* intron were amplified as described in Bruneau et al. (2008), with primers trnK685F (Hu et al., 2000), trnK4La (Wojciechowski et al., 2004), trnK2R*(Wojciechowski et al., 2004) and KC6 (Bruneau et al., 2008). The *trnD–trnT* region was amplified as described in Shaw et al. (2005), with primers trnD, trnT, trnE and trnY, while the *rps16* region was obtained using primers rps16F and rps2R from Oxelman et al. (1997). Primers AB101 and AB102 from Douzery et al. (1999) were used to amplify the ITS region.

For the LCNG markers designed by Li et al. (2008), we followed PCR cycling conditions similar to those used by the authors, but with a longer extension step (1 min 30 s). For markers designed by Choi et al. (2006), we used an initial denaturing step at 94 °C for 3 min, followed by 35 cycles of three steps at 94 °C for 30 s, 53 °C for 30 s, and 72 °C for 2 min, concluding with a final extension step at 72 °C for 5 min.

All PCR amplification products were subsequently purified using a 20% PEG protocol (Joly et al., 2006) and sequenced in both directions using Big Dye Terminator 3.1 chemistry (Applied Biosystems, Carlsbad, CA), as described in Bruneau et al. (2008). After an ethanol–sodium acetate precipitation followed with two 70% ethanol washes, sequencing products were run on an ABI 3100-Avant automated sequencer (Applied Biosystems). The chromatograms were assembled and visually inspected using Sequencher (versions 3.1–4.6, Gene Codes Corporation, Ann Arbor, MI).

After direct sequencing, sequences that needed cloning were amplified in triplicate reactions to help reduce PCR recombinants and *Taq* induced errors (Cronn et al., 2002; Joly et al., 2006; Judo et al., 1998). Cloning was performed using the CloneJET PCR cloning kit (Fermentas), following manufacturer instructions, but with lower concentrations of reagents for the sticky-end protocol and ligation reactions. All sequenced transformed colonies were visually inspected upon alignment; sequences that were suspected to be

Please cite this article as: Babineau, M., et al., Phylogenetic utility of 19 low copy nuclear genes in closely related genera and species of caesalpinioid legumes, South African Journal of Botany (2013), http://dx.doi.org/10.1016/j.sajb.2013.06.018