ARTICLE IN PRESS

Molecular Phylogenetics and Evolution xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

The utility of nuclear conserved ortholog set II (COSII) genomic regions for species-level phylogenetic inference in *Lycium* (Solanaceae)

Rachel A. Levin *, Andrew Whelan, Jill S. Miller

Department of Biology, McGuire Life Sciences Building, Amherst College, Amherst, MA 01002, USA

ARTICLE INFO

Article history:
Received 9 December 2008
Revised 24 June 2009
Accepted 13 August 2009
Available online xxxx

Keywords:
Bayesian estimation of species trees (BEST)
Concatenation
COSII
Fine-scale relationships
GBSSI
Gene trees
Hybridization
Incongruence
ITS
Lycium
Multi-locus
Phylogeny
Species trees
Solanaceae

ABSTRACT

The identification of genomic regions with sufficient variation to elucidate fine-scale relationships among closely related species is a major goal of phylogenetic systematics. However, the accumulation of such multi-locus data sets brings its own challenges, given that gene trees do not necessarily represent the true species tree. Using genomic tools developed for *Solanum* (Solanaceae), we have evaluated the utility of nuclear conserved ortholog set II (COSII) regions for phylogenetic inference in tribe Lycieae (Solanaceae). Five COSII regions, with intronic contents ranging from 68% to 91%, were sequenced in 10 species. Their phylogenetic utility was assessed and compared with data from more commonly used nuclear (GBSSI, nrITS) and cpDNA spacer data. We compared the effectiveness of a traditional total evidence concatenation approach versus the recently developed Bayesian estimation of species trees (BEST) method to infer species trees given multiple independent gene trees. All of the sampled COSII regions had high numbers of parsimony-informative (PI) characters, and two of the COSII regions had more PI characters than the GBSSI, ITS, and cpDNA spacer data sets combined. COSII markers are a promising new tool for phylogenetic inference in Solanaceae, and should be explored in related groups. Both the concatenation and BEST approaches yielded similar topologies; however, when multiple individuals with polyphyletic alleles were included, BEST was clearly the more robust approach for inferring species trees in the presence of gene tree incongruence.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Although DNA sequence data have allowed for major advances in plant systematics, there remains a considerable challenge to identify regions with sufficient variation to differentiate closely related species. Recently there has been much progress in the identification of rapidly evolving cpDNA regions (Shaw et al., 2005, 2007), with many of these regions advocated for use in DNA barcoding (Kress and Erickson, 2007; Lahaye et al., 2008). In addition, considerable efforts have been made in the identification of low copy nuclear regions that are useful for phylogenetic inference (reviewed by Sang, 2002; Small et al., 2004; see also Cronn et al., 2002; Howarth and Baum, 2002; Wu et al., 2006; Chapman et al., 2007; Álvarez et al., 2008; Li et al., 2008; Steele et al., 2008; Yuan and Olmstead, 2008). Despite recent advances in the development of phylogenetic markers for plant systematics, it has become clear that no single gene region is the answer; rather, studies have emphasized the importance of the combined use of large numbers of orthologous sequences for robust phylogenetic inference (Cronn et al., 2002; Sang, 2002; Rokas et al., 2003; Small et al., 2004).

Although sequence data from multiple loci may result in robust topologies, each locus has its own genetic history, which may result in incongruence between topologies inferred from individual loci. The various evolutionary processes that may contribute to this incongruence include incomplete lineage sorting (or deep coalescence), introgression, and hybridization (reviewed in Small et al., 2004). Given that the inference of a species tree is a primary goal for systematics and for addressing evolutionary questions of character evolution, there has been much recent interest in exploring methods for inference of species trees from multi-locus data (Degnan and Salter, 2005; Maddison and Knowles, 2006; Ané et al., 2007; Carstens and Knowles, 2007; Belfiore et al., 2008; Brumfield et al., 2008; Degnan et al., 2008; Eckert and Carstens, 2008; Liu et al., 2008). Traditionally, systematists have concatenated data from multiple loci, with the idea that the most likely species topology will emerge from a combination of data, or total evidence (Kluge, 1989; Barrett et al., 1991; see also Gatesy and Baker, 2005), even if there is incongruence among individual gene topologies. Although incongruence between loci may result in

* Corresponding author. Fax: +1 413 542 7955.
E-mail address: rlevin@amherst.edu (R.A. Levin).

1055-7903/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2009.08.016

Please cite this article in press as: Levin, R.A., et al. The utility of nuclear conserved ortholog set II (COSII) genomic regions for species-level phylogenetic inference in *Lycium* (Solanaceae). Mol. Phylogenet. Evol. (2009), doi:10.1016/j.ympev.2009.08.016

low resolution of inferred species trees when data are concatenated, incongruence between data sets may also result in clades resolved in the concatenated species trees that are not observed in any of the individual gene trees (Gatesy and Baker, 2005; Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007; Rosenberg and Tao, 2008). More recently, there have been considerable advances in coalescent-based methods, which attempt to minimize problems of genealogical incongruence due to incomplete lineage sorting or deep coalescence, to infer species trees from multi-locus datasets (Maddison and Knowles, 2006; COAL: Carstens and Knowles, 2007; BEST: Edwards et al., 2007; Liu and Pearl, 2007; Liu et al., 2008; reviewed in Degnan and Rosenberg, 2009). Belfiore et al. (2008) and Eckert and Carstens (2008) have suggested that such methods are more robust to differences among gene trees than traditional concatenation approaches.

The current explosion in the availability of complete genomes and large EST databases has permitted the identification of putative single-copy conserved orthologous genes and development of universal primers for Euasterids I (including Solanaceae, Wu et al., 2006) and Asteraceae (Chapman et al., 2007; Álvarez et al., 2008). Although few studies have yet been published using these nuclear regions, they are emerging as promising tools for molecular phylogenetic systematics, given the magnitude of sequence data that they provide. Recently, researchers have reported COSII regions (Wu et al., 2006) to be useful for phylogenetic inference among wild tomato (*Solanum* sect. *Lycopersicon*; Rodríguez et al., 2006; Rodríguez and Spooner, 2007, 2008) and potato species (*Solanum* sect. *Petota*; Ames and Spooner, 2008; Fajardo and Spooner, 2008).

The development and demonstrated phylogenetic utility of COSII regions among closely related *Solanum* species has suggested a potential new set of genomic regions for use in fine-scale systematic studies of related taxonomic groups. In this paper we investigate their utility in a second group of Solanaceae, specifically species in tribe Lycieae. The tribe includes ca. 87 species across three genera (*Lycium*, *Grabowskia*, *Phrodus*), with the vast majority of species in *Lycium* (ca. 83 spp.). Whereas *Phrodus* and *Grabowskia* have limited distributions in the Americas, the genus *Lycium* is widespread and occurs worldwide. Although the majority of species diversity (ca. 85% of *Lycium* species) is concentrated in the Americas and southern Africa, *Lycium* is documented on all continents (except Antarctica) and many oceanic islands (e.g., Hawaiian, Canary, and Easter islands).

Previous studies (Levin and Miller, 2005; Levin et al., 2007) have defined backbone relationships in Lycieae; however, the markers used to date have relatively low levels of genetic variation, and unresolved lineages remain (note the short branch lengths in Fig. 1, Levin et al., 2007). Within *Lycium* there are ten species with gender dimorphism, and dimorphism has clearly evolved multiple times in the New and Old World species (Levin and Miller, 2005; Levin et al., 2007). Gender dimorphism is strongly associated with increases in ploidy levels, and all documented cases of separate sexes are in species with tetraploid (or higher) chromosome counts (Miller and Venable, 2000). Interestingly, in North American *Lycium californicum* there are both hermaphroditic and gender dimorphic populations (Yeung et al., 2005); these tend to be separated geographically, but both types of populations occur in relatively close proximity on the Baja California peninsula. Remarkably, the association between gender dimorphism and polyploidy within this species mirrors that observed in the genus as a whole; hermaphroditic populations are diploid, whereas dimorphic populations are polyploid (Yeung et al., 2005; Miller and Levin, unpubl. data). Previous phylogenetic studies including *L. californicum* have resulted in incongruence among loci (Levin and Miller, 2005; Levin et al., 2009), suggesting incomplete lineage sorting or hybridization. Coupled with reports of hybrids elsewhere in the genus (Bernardello, 1986; Venter, 2000), it is likely that hybridization may be important in the history of *Lycium*. However, lack of resolution (due to markers with insufficient variation) and limited sampling of multiple individuals within species have restricted confident interpretation of evolutionary patterns.

Given its cosmopolitan distribution, natural variation in sexual systems (both across and within species), and the potential role of hybridization, *Lycium* has emerged as an excellent candidate genus for fine-scale phylogenetic studies. Species-level knowledge of evolutionary relationships can be applied to a range of questions including detection of the frequency and direction of hybrid species formation, interpretation of historical biogeographic patterns of dispersal, and the evolution of combined versus separate sexes, including studies of the molecular evolutionary genetics of mating system genes (Savage and Miller, 2006; Miller et al., 2008).

Specifically, our goals for this study were to: (1) establish whether COSII regions can be used for phylogenetic inference in Lycieae and compare these COSII data with data from the more commonly used nuclear granule-bound starch synthase I (GBSSI), nuclear ribosomal internal transcribed spacer (ITS), and cpDNA spacer regions, (2) determine levels of within versus among species variation, (3) test allelic monophyly and interpret in light of possible hybrid histories (specifically for *L. californicum*), and (4) examine topological congruence of gene trees and compare species trees inferred using a concatenation approach with a coalescent-based approach [Bayesian estimation of species trees (BEST): Liu, 2008; Liu et al., 2008] for combining gene trees to infer species trees.

2. Materials and methods

2.1. Taxon sampling

We sampled nine species across tribe Lycieae (see Fig. 1 in Levin et al., 2007), including closely related species as well as more distantly related species, and geographically disparate species including those from North America (*L. carolinianum*, *L. andersonii*, *L. puberulum*, and *L. californicum*), South America (*Phrodus microphyllus*, *L. chilense*), and southern Africa (*L. ferocissimum*, *L. oxycarpum*, and *L. tenue*). To assess variation within species and possible hybrid ancestry, individuals from eight populations of *Lycium californicum* were also included for a subset of the nuclear loci. A tenth species, *Nolana werdermannii*, was included as an outgroup. All taxa and GenBank Accession numbers are in Table 1.

2.2. COSII regions

We initially amplified 44 COSII regions (available from http://www.sgn.cornell.edu/markers/cosii_markers.pl; Wu et al., 2006) for three distantly related *Lycium* species (*L. andersonii*, *L. oxycarpum*, and *L. puberulum*). These 44 regions were selected based on the previously documented high intronic content and length (<1500 bp) in *Solanum*. From these, five COSII regions were identified that amplified well, had >60% intronic content (in *Solanum*), and yielded single-banded PCR products. These five regions are listed in Table 2 with primer sequences. All five of these regions are located on different chromosomes and include a single large intron, with flanking regions from two exons (within which the primers are nested). Putative functions for many of these regions have been inferred (Sol Genomics network, www.sgn.cornell.edu; Mueller et al., 2005). For example, the unigene for COS14 has similarities to dehydrogenases, COS16 has similarities to acyltransferases, COS27 has similarities to aminotransferases, COS30 has similarities to acid phosphatase/vanadium-dependent haloperoxidases, and COS40 has similarities to tubulin alpha, although this unigene awaits further annotation.

Please cite this article in press as: Levin, R.A., et al. The utility of nuclear conserved ortholog set II (COSII) genomic regions for species-level phylogenetic inference in *Lycium* (Solanaceae). Mol. Phylogenet. Evol. (2009), doi:10.1016/j.ympev.2009.08.016

Table 1

Taxa, collection localities, voucher information, and GenBank Accession numbers for all sequences included in this study. GenBank Accession numbers are listed in the following order: GBSSI, ITS, cpDNA spacers ($trnH^{GUG}$–*psbA*, *rpl32*–$trnL^{UAG}$, *ndhF*–*rpl32*, $trnL^{UAA}$–$trnF^{GAA}$, $trnT^{UGU}$–$trnL^{UAA}$, $trnD^{GUC}$–$trnT^{GGU}$), COS14, COS16, COS27, COS30, and COS40. Many of the COSII regions yielded multiple alleles (usually two) per individual; thus, GenBank Accession numbers for all retrieved alleles are listed. Multiple *L. californicum* individuals included in the COSII analyses (Fig. 1) are noted in parentheses. Voucher specimens are deposited in the following herbaria: ARIZ, University of Arizona; CORD, Universidad Nacional de Córdoba; MASS, University of Massachusetts; TAIC, Texas A&M University, Kingsville; WTU, University of Washington.

Tribe Lycieae Hunz.

Lycium L.

L. andersonii A. Gray—Baja California, Mexico, *Miller 97-12* (ARIZ); DQ124503, DQ124620, FJ189607, FJ189736, FJ189669, DQ124562, DQ124439, FJ189631, N/A, N/A, N/A, N/A, N/A

L. andersonii A. Gray—Pinal Co., Arizona, USA, *Miller and Levin 06-01* (MASS); N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301177, GQ301119–GQ301120, GQ301085–GQ301086, GQ268879–GQ268880, GQ301146

L. californicum Nutt. ex A. Gray—Pinal Co., Arizona, USA, *Miller and Levin 04-15* (MASS); DQ124509, DQ124647, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A

L. californicum Nutt. ex A. Gray—Pima Co., Arizona, USA, *Miller 01-2* (ARIZ); N/A, N/A, FJ189608, FJ189737, FJ189670, DQ124572, DQ124449, FJ189632, N/A, N/A, N/A, N/A, N/A.

L. californicum Nutt. ex A. Gray—Zacatecas, Mexico, *Miller and Levin 05-56* (MASS); N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301121–GQ301122, GQ301087–GQ301088, GQ268881–GQ268882, GQ301147–GQ301148. (*L. californicum* PR)

L. californicum Nutt. ex A. Gray—Coahuila, Mexico, *Miller and Levin 05-43* (MASS); N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301123–GQ301124, GQ301089–GQ301090, GQ268883–GQ268884, GQ301149–GQ301150. (*L. californicum* LL)

L. californicum Nutt. ex A. Gray—Sonora, Mexico, *Miller and Levin 05-68* (MASS); N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301091–GQ301092, GQ268885, GQ301151–GQ301152. (*L. californicum* LB)

L. californicum Nutt. ex A. Gray—Sonora, Mexico, *Miller and Levin 05-80* (MASS); N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301093–GQ301094, GQ268886–GQ268887, GQ301153–GQ301154. (*L. californicum* KB)

L. californicum Nutt. ex A. Gray—Sonora, Mexico, *Miller and Levin 05-81* (MASS); N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301125–GQ301126, GQ301095–GQ301096, GQ268888–GQ268889, GQ301155–GQ301156. (*L. californicum* PL)

L. californicum Nutt. ex A. Gray—Maricopa Co., Arizona, USA, *Miller and Levin 05-85* (MASS); N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301097–GQ301099, GQ268890–GQ268891, GQ301157–GQ301159. (*L. californicum* PX)

L. californicum Nutt. ex A. Gray—San Diego Co., California, USA, no voucher; N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301100–GQ301101, GQ268892–GQ268893, GQ301160. (*L. californicum* KN)

L. californicum Nutt. ex A. Gray—Baja California, Mexico, no voucher; N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301102, GQ268894–GQ268895, GQ301161–GQ301162. (*L. californicum* SSC)

L. carolinianum Walt.—Texas, USA, *Hempel 843* (TAIC); DQ124512, DQ124622, GQ301192, GQ301193, GQ301194, DQ124573, DQ124450, FJ444883, GQ301178–GQ301179, GQ301127–GQ301128, GQ301103, GQ268896–GQ268897, GQ301163–GQ301164

L. chilense Bertero var. *chilense*—Argentina, *Miller et al. 04-101* (MASS); EF137765, N/A, FJ189611, N/A, FJ189673, FJ189719, FJ189727, FJ189635, GQ301180–GQ301182, GQ301129–GQ301130, GQ301104–GQ301105, GQ268898–GQ268899, GQ301165–GQ301166

L. chilense Bertero—Argentina, *Bernardello 877* (CORD); N/A, GQ301196, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301114–GQ301115, N/A, N/A.

L. chilense var. *vergarae* (Phil.) Bernardello—Argentina, *Miller et al. 04-109* (MASS); N/A, N/A, N/A, FJ189740, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A

L. ciliatum Schltdl.—Argentina, *Bernardello 876* (CORD); N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, GQ301113/GQ301116, N/A, N/A.

L. ferocissimum Miers—South Africa, *Miller et al. 05-199* (MASS); GQ301195, N/A, FJ189615, FJ189744, FJ189677, FJ189720, FJ189728, FJ189639, GQ301183, GQ301131–GQ301132, GQ301106–GQ301107, GQ268900–GQ268901, GQ301167

L. ferocissimum Miers—South Africa, cult. Strybing Arboretum and Botanical Gardens 98-0143; N/A, GQ301197, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A

L. oxycarpum Dunal—South Africa, *Miller et al. 05-206* (MASS); EF137784, N/A, FJ189618, FJ189747, FJ189680, FJ189721, FJ189729, FJ189642, GQ301184–GQ301185, GQ301133–GQ301134, GQ301108, GQ268902–GQ268903, GQ301168–GQ301169

L. puberulum A. Gray—Texas, USA, *Levin 97-6* (ARIZ); DQ124537, AF238985, FJ189620, FJ189749, FJ189682, DQ124599, DQ124476, FJ189644, GQ301188–GQ301189, GQ301137–GQ301138, GQ301110–GQ301111, GQ268906–GQ268907, GQ301171–GQ301172

L. tenue Willd.—South Africa, *Miller et al. 05-220* (MASS); EF137794, N/A, FJ189625, FJ189754, FJ189687, FJ189724, FJ189732, FJ189649, GQ301190–GQ301191, GQ301139–GQ301140, GQ301112, GQ268908–GQ268909, GQ301173–GQ301174

L. tenue Willd.—South Africa, *Olmstead 99-13* (WTU); N/A, GQ301198, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A, N/A.

Phrodus Miers

Phrodus microphyllus (Miers) Miers—Chile, *Miller et al. 04-92* (MASS); EF137801, FJ439765, FJ189627, FJ189756, FJ189689, FJ189725, DQ124495, FJ189651, GQ301186–GQ301187, GQ301135–GQ301136, GQ301109, GQ268904–GQ268905, GQ301170

Outgroup

Nolana werdermannii I.M. Johnst.—Chile, *Miller et al. 04-77* (MASS); EF137799, FJ439764, FJ189604, FJ189733, FJ189666, DQ124616, DQ124493, FJ189628, GQ301175–GQ301176, GQ301117–GQ301118, GQ301084, GQ268878, GQ301144–GQ301145

Please cite this article in press as: Levin, R.A., et al. The utility of nuclear conserved ortholog set II (COSII) genomic regions for species-level phylogenetic inference in *Lycium* (Solanaceae). Mol. Phylogenet. Evol. (2009), doi:10.1016/j.ympev.2009.08.016

Table 2
The COSII regions, official region names as assigned by the Sol Genomics Network (SGN; Mueller et al., 2005), unigene available on the SGN, and primers used in this study (see also Wu et al., 2006).

Region	SGN number	Unigene	Forward primer	Reverse primer
COS14	C2_At1g24360	SGN-U315110	5′-TCCGGTTGTTATTGTCACTGGAGC-3′	5′-TGGAAACTTCTTCTGCCTCCTTTG-3′
COS16	C2_At1g78690	SGN-U321585	5′-TCCAGAAGGGAAGGTCTGTCAAGAAG-3′	5′-AGTCATGTACAGACATTTTTGTGCTGC-3′
COS27	C2_At1g80360	SGN-U320089	5′-ATGGTTACTGCCGGTGCAAATCAG-3′	5′-TCGGTAACACCTGTCATCTGGAATG-3′
COS30	C2_At3g21610	SGN-U316177	5′-ATGGGATTCAAAAAGGATGCTTAGC-3′	5′-AGCCTAACACCAGTAGCATCATACATTAC-3′
COS40	C2_At1g50020	SGN-U326704	5′-TTGCTTACTCTTGGTGGAACATTC-3′	5′-TGTCTGTGATATCCTCTCTTCTTC-3′

Fifty microliter reactions contained 1× buffer, 2.5 mM $MgCl_2$, 0.20 mM dNTPs, 0.40 μM of each primer, 1.25 U of Taq polymerase, and 1–2 μL DNA. Thermal cycler conditions included initial denaturation at 94 °C for 3 min; 35 cycles at 94 °C for 30 s, 55 °C for 30 s, 72 °C for 1 min; ending with an extension at 72 °C for 7 min. Alternatively, a touchdown procedure was used with an initial denaturation at 94 °C for 3 min; 20 cycles at 94 °C for 45 s, 58–49 °C (decreasing one degree every two cycles) for 30 s, 72 °C for 1 min 30 s; 20 cycles at 94 °C for 45 s, 48 °C for 30 s, 72 °C for 1 min 30 s; with a final extension at 72 °C for 7 min. PCR products were cleaned using either the QIAQuick PCR Purification kit or the MinElute PCR Purification kit (Qiagen, Inc., Valencia, CA) or PEG precipitation (see Paithankar and Prasad, 1991). A few products were directly sequenced on an ABI automated sequencer by the DNA Sequencing Facility of the Biotechnology Resource Center at Cornell University (Ithaca, NY, USA); as these yielded unreadable sequences due to allelic length polymorphisms, all products were cloned prior to sequencing.

Cloning used either the Novagen Perfectly Blunt Cloning Kit or the Novagen pSTBlue-1 AccepTor Vector Giga Kit (Novagen, EMD Chemicals, Inc., Madison, WI). Colonies were PCR amplified in 25 μl reactions using 5 μl of each colony prep and the vector primers R20 and U19 at final concentrations of 0.125 mM. Reactions contained 1 U Taq polymerase, 1× buffer, 0.25 mM dNTPs, and 1.5 mM $MgCl_2$. The thermal cycler program had an initial denaturation at 94 °C for 5 min; 6 cycles at 94 °C for 1 min, 55–53 °C (decreasing 1 degree every 2 cycles) for 1 min, 72 °C for 2 min; 30 cycles at 94 °C for 1 min, 52 °C for 1 min, 72 °C for 2 min; ending with an extension at 72 °C for 5 min.

PCR products were cleaned as above or using ExoSAP with 5 μl PCR product, 1 μl shrimp alkaline phosphatase (1 U/μl, SAP), 0.5 μl Exonuclease I (10 U/μl), and 1.5 μl 10× PCR buffer. Six to eight colonies per accession were sequenced in a single direction using the vector primers. Sequencing was conducted by the DNA Sequencing Facility of the Biotechnology Resource Center at Cornell University (Ithaca, NY, USA) or the Pennsylvania State University Nucleic Acid Facility (University Park, PA, USA). As alleles were identified, one colony per allele was sequenced in the opposite direction in order to obtain a complete sequence.

2.3. Sequence alignment and phylogenetic analyses

Sequences were edited and assembled using Sequencher v. 4.8 (Gene Codes Corp., 1991–2007), and alleles were identified. The allele sequences (usually two; occasionally one or three) for each genomic accession were manually aligned across species using SeAl v. 2.0a11 (Rambaut, 2002). In rare cases, a consensus of multiple colony sequences for the same allele, rather than a single colony consensus, was included in the multi-species alignment. Locations of introns and exons were determined by alignment to unigenes (Table 2), which are available through the Sol Genomics Network (Mueller et al., 2005).

The sequences of the same ten species for three additional loci were compared to the COSII data. These three additional loci included previously published data from the nuclear ribosomal ITS region (Miller, 2002; Yeung et al., 2005; Levin et al., 2009), the nuclear granule-bound starch synthase gene (GBSSI: Levin and Miller, 2005; Levin et al., 2007), and cpDNA data that combines sequence data from six spacer regions (*ndhF–rpl32*, *rpl32–trnL*, *trnD–trnT*, *trnH–psbA*, *trnT–trnL*, *trnL–trnF*: Levin and Miller, 2005; Levin et al., 2009; Miller et al., 2009).

For each data set, model parameters were determined using the Akaike information criterion in Modeltest v. 3.7 (Posada and Crandall, 1998). The best-fit model for each data set was used in separate maximum likelihood (ML) analyses in PAUP* (Swofford, 2002). We used an ML algorithm to analyze each of the eight data sets, each of which included 10 species, except for COS14 and nrITS, which included nine species. ML settings in PAUP* included the heuristic search option, all most-parsimonious trees from a parsimony analysis of the data set (branch and bound) as the starting trees (note that not all of the most-parsimonious starting trees are actually used by PAUP*, depending on their ML scores), tree bisection reconnection (TBR) branch-swapping, and the MulTrees option in effect. *Nolana werdermannii* was defined as the outgroup, as *Nolana* has been shown to be sister to Lycieae in previous analyses (Levin et al., 2007; Olmstead et al., 2008). In most analyses there was only a single ML tree; when there was more than one ML tree, all most-likely trees were combined in a strict consensus tree. An ML non-parametric bootstrap (BS) analysis was conducted for each data set (PAUP*), using the same model parameters as in the original ML analysis, and 100 full heuristic bootstrap replicates, each with 10 random-addition sequence replicates, TBR branch-swapping, and the MulTrees option in effect.

In addition to separate analyses of each region, analyses were also conducted with combined data sets. For most combined analyses, only one allele per species was included. Alleles within species were generally chosen randomly; however, where there were clearly different alleles (i.e., were in different well supported clades in ML analyses of individual data sets), we included the allele with the most probable phylogenetic affinity (see Results for more detail).

Combined analyses employed either a concatenation approach or the recently developed Bayesian estimation of species trees (BEST v. 2.2; Liu, 2008; Liu et al., 2008). Concatenated data sets for five nuclear loci (COSII only) and all 8 loci (7 nuclear loci plus cpDNA) were analyzed using a partitioned Bayesian analysis with MrBayes v. 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Whereas these data sets included only a single accession of *L. californicum*, we also used the same analysis method for a data set including two accessions of *L. californicum* and the four COSII loci (COS16, COS27, COS30, and COS40) for which there were data for all 10 species. In all three analyses, each locus was a single partition, with its own substitution model (as determined by Modeltest, above); branch lengths were estimated separately for each partition. Analyses were run for 2 million generations using 4 Markov chains; trees were saved every 100 generations. Fifty percent majority-rule consensus trees with posterior probabilities were retrieved using the sumt command and burnin = 25 (5 loci), burnin = 35 (4 loci), or burnin = 55 (8 loci). The burnin was determined by graphically examining the tree likelihoods versus

Please cite this article in press as: Levin, R.A., et al. The utility of nuclear conserved ortholog set II (COSII) genomic regions for species-level phylogenetic inference in *Lycium* (Solanaceae). Mol. Phylogenet. Evol. (2009), doi:10.1016/j.ympev.2009.08.016

ARTICLE IN PRESS

generation time, and excluding those trees saved prior to the convergence of likelihood values.

In contrast to the concatenation approach, BEST (Liu, 2008; Liu et al., 2008) determines a Bayesian posterior distribution of species trees based on the distribution of gene trees. A majority-rule consensus tree is then constructed, in order to yield a species tree with posterior probabilities. Using the same data sets used in the concatenated analyses, three BEST analyses were conducted: one accession of *L. californicum* was included for analyses of five (COSII only) and eight loci (COSII, nrITS, GBSSI, and cpDNA), whereas two accessions of *L. californicum* were included for an analysis of the four COSII loci (see above) for which there were data for all 10 species. Each locus had its own substitution model (the same locus-specific models as discussed above). Analyses were run for 100–200 million generations, with 4 chains and a temperature of 0.15; trees were saved every 1000 generations. Tracer v. 1.4 (Rambaut and Drummond, 2007) was used to determine the appropriate burnin value. The majority-rule consensus for the estimated distribution of species trees was constructed in BEST using the sumt command and burnin = 50,000 (5 loci), burnin = 30,000 (8 loci), and burnin = 10,000 (4 loci).

3. Results

3.1. COSII sequence data

The five COSII regions vary in average length from 741 to 1332 bp (Table 3), with an average of three informative indels per region. COS40 is the longest region, and also contains considerable microsatellite repeats, resulting in incomplete sequences for two taxa (*L. andersonii* and *L. tenue*). As previously mentioned, these five COSII regions all include one intron, with flanking exons. For all five regions, the majority of their length is comprised of a single large intron (average intronic content of 82%), with an average intronic content per region ranging from 68% for COS16 to 91% for COS27 (Table 3). Those regions with the highest intronic content do not necessarily have the most parsimony-informative (PI) characters (Table 3). For example, COS14 has one of the highest intronic contents (89%) and also has, by far, the highest number of PI characters; however, COS27, which has the highest intronic content (91%), has the fewest PI characters, even though this region is ca. 250 bp longer than COS14.

Maximum likelihood analyses of each individual COSII marker resulted in single ML trees except for the analysis of COS27, which yielded two ML trees (Fig. 1). In general, species were monophyletic with respect to alleles, with the most notable exception being *L. californicum*, the one species for which multiple individuals were included. For COS27 and COS40 (COS14 did not amplify for *L. californicum*), *L. californicum* alleles were monophyletic and closely related to *L. carolinianum* or *L. andersonii* (Fig. 1C and E). In contrast, alleles of COS16 and COS30 were resolved in two distinct clades (Fig. 1B and D); one lineage of alleles is closely related to North American *L. carolinianum* and *L. andersonii*, whereas the other group is sister to a clade including the Old World species *L. ferocissimum*, *L. tenue*, and *L. oxycarpum*. For all but one accession (*L. californicum* PR), alleles within an individual grouped together in the same clade.

3.2. GBSSI, ITS, cpDNA sequence data

These three regions ranged in length from an average of 4953 bp for cpDNA (six spacer regions combined) to 1699 bp for GBSSI and 677 for ITS (Table 3). Average intronic content of this region of GBSSI (from exons 2 through 10) was 43%, markedly lower than the COSII regions. The number of PI characters was also considerably lower for these three regions, ranging from 22 PI characters for ITS to 29 PI characters for GBSSI (Table 3). Single ML trees resulted from separate analyses of these three regions. With ITS and cpDNA data, *L. californicum* is closely related to *L. carolinianum* and *L. andersonii* (Fig. 1G and H), whereas GBSSI suggests a close relationship with the Old World species and South American *L. chilense* (Fig. 1F).

3.3. Data sets combined: concatenation

For the combined analyses, we randomly chose alleles for inclusion [all alleles were not included due to the unequal numbers of alleles (generally 1 or 2) retrieved across species and gene regions], given the almost uniform monophyly within an accession (Fig. 1). However, for *L. californicum* we chose an allele for both COS16 and COS30 that reflected this species placement by the majority of loci (i.e., an allele of *L. californicum* PL, related to *L. andersonii* and *L. carolinianum*). The concatenated analyses including all eight loci (topology not shown) and just the five COSII loci (Fig. 2A) yielded well resolved topologies (posterior probabilities >99% for all nodes within Lycieae), with generally concordant results. In both analyses, *L. californicum* was in a clade with *L. andersonii* and *L. carolinianum*. The main difference between these two analyses was in the placement of *L. chilense*, which was sister to this clade of *L. californicum* + *L. andersonii* + *L. carolinianum* when all 8 loci were included, but outside both this clade and the Old World clade when only the COSII data were analyzed (Fig. 2A).

Given that inclusion of a single *L. californicum* allele strongly favors a particular topology regarding relationships of *L. californicum*, we also analyzed data including four COSII loci for two accessions of this species (*L. californicum* accessions LL and PL; Fig. 1B and D). This concatenated analysis yielded the well supported placement of *L. californicum* in two distinct clades (Fig. 3A), reflecting the same relationships suggested by the individual COS16 and COS30 data sets (Fig. 1B and D).

3.4. Data sets combined: BEST

The BEST analyses including all eight loci (topology not shown) and just the five COSII loci (Fig. 2B) yielded well resolved

Table 3
Comparison of the ten species (* nine species) data sets, with a single allele per species across all COSII regions. PI, parsimony-informative characters; CI, consistency index; RC, rescaled consistency index; π, nucleotide diversity.

Name	Aligned length	Mean length (range) bp	Mean % intronic content	Number variable sites (introns only)	Proportion variable sites	PI	CI (RC)	π (std. dev.)
COS14*	1135	1086 (1043–1104)	89	287 (281)	0.25	126	0.95 (0.87)	0.08652 (0.01954)
COS16	1390	1034 (980–1259)	68	215 (185)	0.21	65	0.96 (0.86)	0.05863 (0.01146)
COS27	1538	1332 (1203–1389)	91	282 (270)	0.18	41	0.94 (0.64)	0.05356 (0.01562)
COS30	984	741 (567–903)	78	185 (170)	0.19	81	0.92 (0.84)	0.07748 (0.00835)
COS40	1842	1326 (886–1660)	87	305 (296)	0.17	51	0.94 (0.70)	0.04929 (0.00910)
GBSSI	1780	1699 (1570–1716)	43	150 (95)	0.08	29	0.96 (0.83)	0.03380 (0.000053)
ITS*	681	677 (676–678)	N/A	71 (N/A)	0.10	22	0.82 (0.49)	0.02142 (0.00425)
cpDNA	5061	4953 (4919–5020)	N/A	119 (N/A)	0.02	25	0.96 (0.88)	0.00628 (0.000004)

Please cite this article in press as: Levin, R.A., et al. The utility of nuclear conserved ortholog set II (COSII) genomic regions for species-level phylogenetic inference in *Lycium* (Solanaceae). Mol. Phylogenet. Evol. (2009), doi:10.1016/j.ympev.2009.08.016

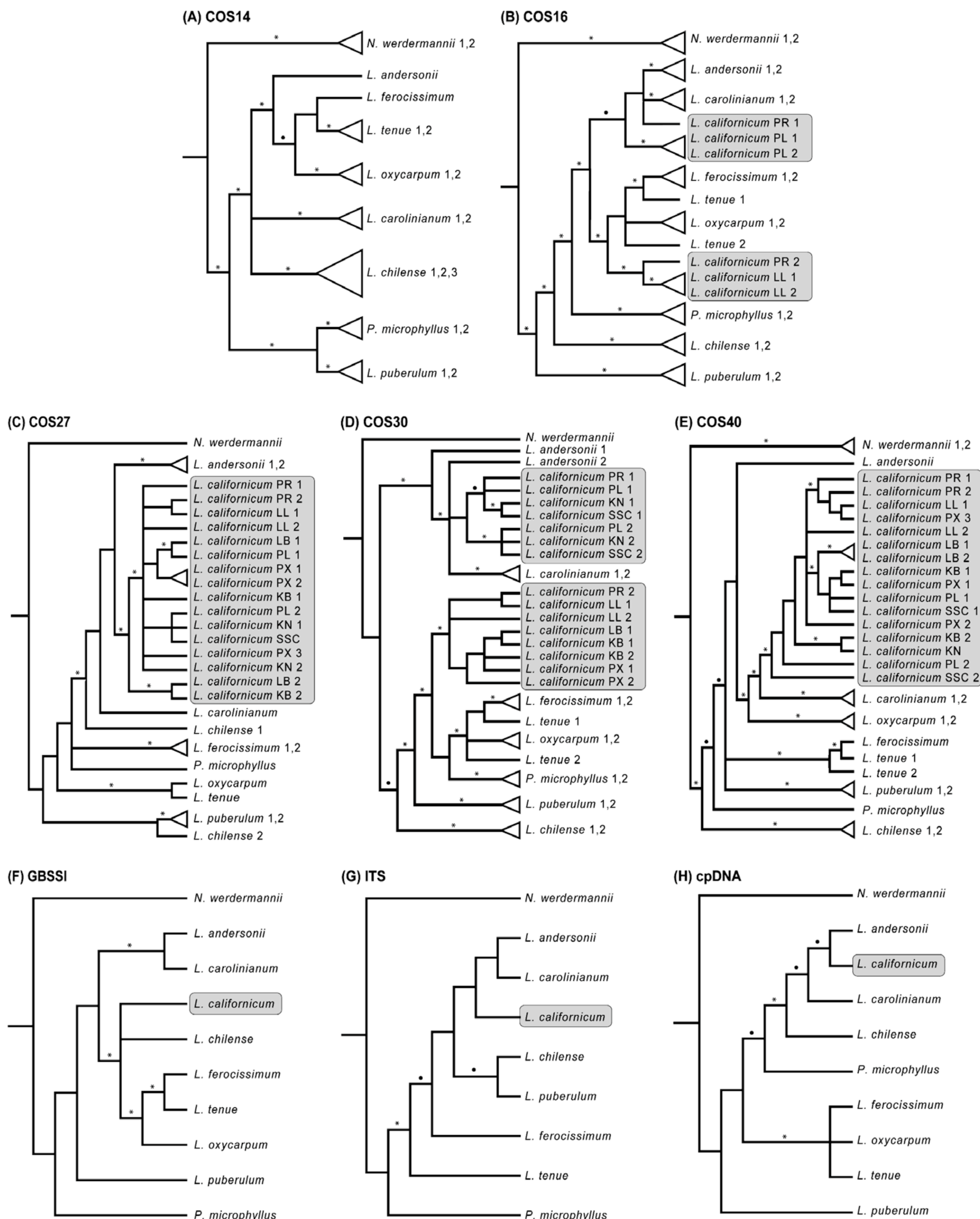

Fig. 1. (A–H) Gene trees from ML individual locus analyses. All topologies are single ML trees (analyses yielded a single ML tree except for COS27, for which a random tree out of the two ML trees was chosen). Bullets above branches indicate BS ⩾ 70%; asterisks indicate BS ⩾ 90%. *N. werdermannii*, *Nolana werdermannii*; *P. microphyllus*, *Phrodus microphyllus*. Letter abbreviations (PR, PL, KB, LL, PX, LB, KN, and SSC) indicate different *L. californicum* individuals (see Table 1). Numbers after species names indicate a different allele of the same accession; there is no number if only one allele was retrieved. Shading highlights the different individuals and alleles of *L. californicum*.

topologies that were generally congruent. There were, however, considerably higher posterior probabilities for the five COSII locus analysis (100% of clades resolved with posterior probabilities >75% versus only 63% of clades with all eight loci), likely due to the low

ARTICLE IN PRESS

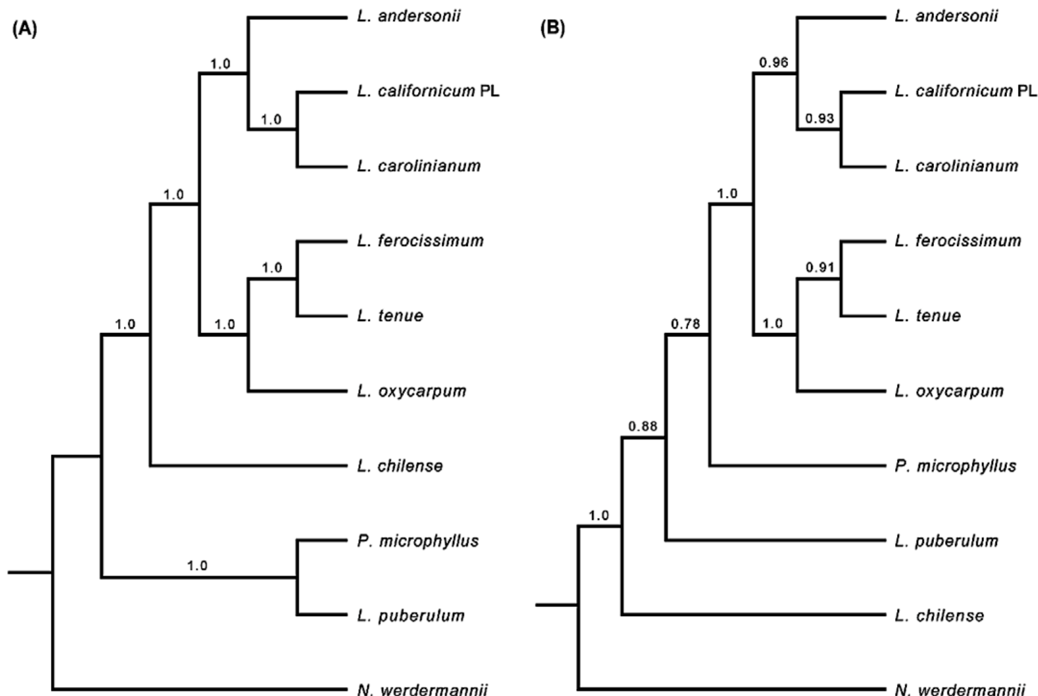

Fig. 2. (A) The majority-rule Bayesian consensus tree of the concatenated five COSII locus data set. (B) The majority-rule consensus tree inferred from the five COSII locus data set using BEST. Bayesian posterior probabilities are shown above the nodes.

percentage of parsimony-informative characters in the GBSSI, ITS, and cpDNA data sets and the high homoplasy in ITS (Table 3). The BEST analyses yielded topologies that were quite similar to that inferred using the concatenated approach (Fig. 2A). Specifically, in both BEST analyses *L. californicum* was in a clade with *L. andersonii* and *L. carolinianum*. Further, as with the concatenated analyses, *L. chilense* was sister to *L. californicum* + *L. andersonii* + *L. carolinianum* when all 8 loci were included, but outside this clade when only the COSII data were analyzed (Fig. 2). However, unlike the five locus concatenated analysis, in the BEST analysis of the same data set *L. chilense* was resolved at the base of all Lycieae species (Fig. 2B).

The four locus BEST analysis including two *L. californicum* accessions (Fig. 3B) yielded a well supported, monophyletic *L. californicum*. This is in contrast to the placement of *L. californicum* by the concatenated analysis (Fig. 3A).

4. Discussion

4.1. Gene trees and congruence

Across the various single locus analyses, there were a number of topological similarities, as well as instances of incongruence due to

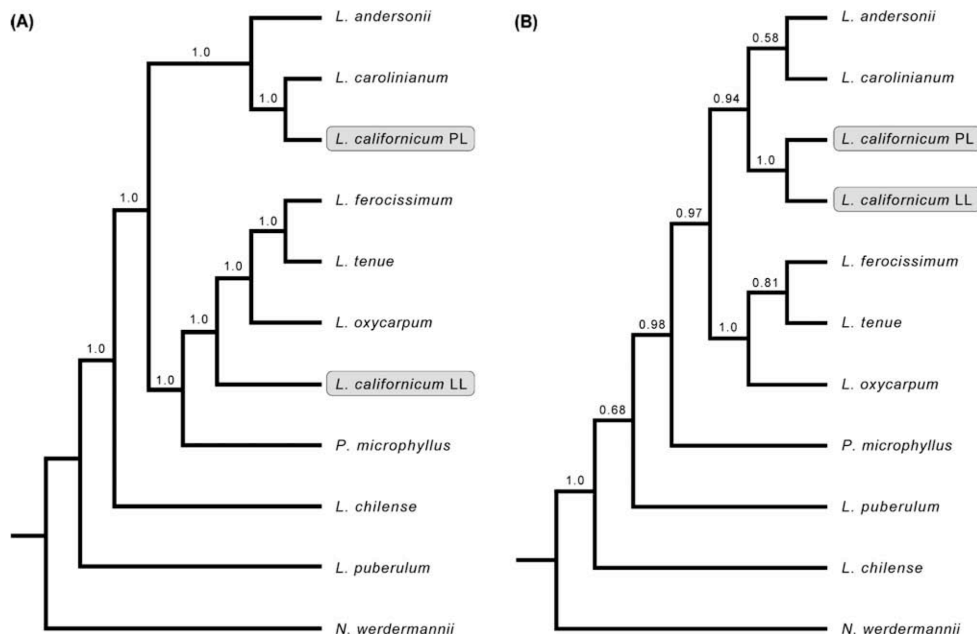

Fig. 3. (A) The majority-rule Bayesian consensus tree of the concatenated four COSII locus data set. (B) The majority-rule consensus tree inferred from the four COSII locus data set using BEST. Bayesian posterior probabilities are shown above the nodes. Shading highlights the two individuals of *L. californicum* included in these analyses.

Please cite this article in press as: Levin, R.A., et al. The utility of nuclear conserved ortholog set II (COSII) genomic regions for species-level phylogenetic inference in *Lycium* (Solanaceae). Mol. Phylogenet. Evol. (2009), doi:10.1016/j.ympev.2009.08.016

incomplete lineage sorting, introgression, and/or hybridization. Many of the individual gene trees showed remarkably strong support for relationships, especially COS14, COS16, and COS30 (Fig. 1). For resolved nodes supported by bootstrap percentages greater than 70%, 75–81% of these nodes had 100% BS support with COS14 and COS16, and 87% of these nodes had >90% BS support with COS30. A lack of support for deep relationships in some of the topologies is likely due to taxon sampling, given that sampling within *Lycium* was limited.

A main cause of incongruence across the eight gene trees (Fig. 1) is the polyphyly of alleles within species. In *Lycium californicum*, the taxon for which multiple individuals were sampled, all *L. californicum* alleles from multiple accessions generally comprise a monophyletic group regardless of locus (Fig. 1C and E; also for GBSSI and ITS in Yeung et al., 2005; Levin et al., 2007). The two exceptions to this pattern are the topologies inferred from COS16 and COS30 (Fig. 1B and D). Gene trees inferred from these two loci show the same pattern of *L. californicum* alleles; one clade of alleles is more closely related to the North American species *L. andersonii* and *L. carolinianum*, whereas the other clade is closely related to the Old World *Lycium* species. Topologies inferred from the other genomic regions reflected one of these two relationships (Fig. 1). Specifically, cpDNA, ITS, COS40, and COS27 suggest that *L. californicum* is closely related to the geographically proximate species *L. andersonii* and *L. carolinianum*, whereas GBSSI suggests a close relationship with the Old World species and South American *L. chilense*. The same phylogenetic affinities for *L. californicum* are also indicated when additional taxa are included (Miller, 2002; Levin and Miller, 2005; Levin et al., 2007, 2009). These results suggest either incomplete lineage sorting (supported by a species tree estimated using BEST, see below) or an ancient hybridization event, involving a maternal species in the *andersonii* clade and a paternal species allied with the Old World clade. Analyses with greater taxon sampling (Levin and Miller, 2005; Levin et al., 2007) suggest that such a hybridization event would have likely involved not an Old World species per se (which would be geographically unlikely), but rather a South American species in the ancestral clade that originally dispersed to the Old World.

The other notable instance of allelic polyphyly occurs in the South American species *Lycium chilense*. Across all of the COSII topologies, *L. chilense* alleles are monophyletic except for COS27; the gene tree for this region shows strong support for the two alleles in disjunct locations of the topology (Fig. 1). To test the robustness of this result, we sequenced COS27 for a second accession of *L. chilense* and also *L. ciliatum*, a species that previous studies (Levin and Miller, 2005; Levin et al., 2007) have shown to be sister or possibly conspecific with *L. chilense*. Remarkably, both additional accessions yielded alleles that were resolved in the same two disjunct areas of the gene tree (topology not shown) as the two *L. chilense* alleles in Fig. 1C. These results strongly suggest incomplete lineage sorting, introgression and/or hybridization. Possible hybrids between *L. ciliatum* and two other South American species (*L. cestroides* and *L. elongatum*), as well as polyploid individuals of *L. ciliatum* and *L. chilense*, have been reported (Bernardello, 1982, 1986). In fact, three alleles were retrieved from *L. chilense* in the present study (COS14; Fig. 1A). Bernardello (1986) also suggests that *L. ciliatum* and *L. chilense* are often difficult to distinguish morphologically, with some specimens appearing intermediate between the two species. The similarity of alleles (in the same two disjunct clades) between *L. chilense* and *L. ciliatum* provides additional evidence that these species may, in fact, be conspecific. Further sampling of species within South America, as well as additional accessions of *L. chilense* and *L. ciliatum*, are necessary to fully understand the evolutionary history of these species.

4.2. Species trees: concatenation versus BEST

When data were concatenated, inferred species relationships usually reflected those that occurred in a subset of the individual gene trees. However, among the individual COSII gene trees (Fig. 1), approximately half of the non-conspecific, well supported clades (BS $\geqslant$ 70%) within Lycieae were in conflict with the well supported clades from the five COSII locus concatenated analysis (Fig. 2A), although this number is somewhat inflated due to inclusion of only one *L. californicum* allele. The species tree inferred from the five COSII locus BEST analysis is generally concordant with that inferred using the concatenated approach (Fig. 2). The only notable difference in topology between the concatenated and BEST analyses is in the sister relationship between *Phrodus microphyllus* and *L. puberulum*. None of the BEST analyses (including data sets with all 8 loci, 5 COSII loci, and 4 COSII loci) show this relationship, which was only suggested by one individual gene topology (COS14; Fig. 1A). This observation is consistent with a documented disadvantage of concatenation, namely that relationships may be supported that are not suggested by the individual gene loci (Gatesy and Baker, 2005; Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007; Belfiore et al., 2008; Rosenberg and Tao, 2008).

Interestingly, the species tree inferred from the concatenated data set had higher clade posterior probabilities than the species tree estimated by BEST. In the concatenated species tree, all nodes within Lycieae were well resolved with a posterior probability >90%, whereas in the BEST topology only 75% of resolved nodes had posterior probabilities >90% (Fig. 2). Belfiore et al. (2008) and Liu et al. (2008) also recovered similar species tree topologies with Bayesian concatenation and BEST approaches, but with higher posterior probabilities in the species tree from the concatenated analysis. It is perhaps not surprising that clade support values for the species tree inferred using BEST are lower than the support values from the concatenated analysis. With the concatenated approach there may be hidden emergent support that results from combining data, as demonstrated by Gatesy and Baker (2005). Such support would not appear when using the BEST approach, given that the species tree posterior probabilities are a direct result of the number of genes and the signal in the individual gene trees (Liu et al., 2008).

When two individuals of *L. californicum* (LL and PL; from separate clades in Fig. 1B and D) were included in analyses of 4 COSII loci, there was a notable difference between the concatenation and BEST topologies (Fig. 3). In the concatenated analysis *L. californicum* was polyphyletic (Fig. 3A), reflecting the incongruent placement of *L. californicum* alleles across loci (Fig. 1). By contrast, the BEST topology resolved the two *L. californicum* individuals as a monophyletic group, with >99% posterior probability (Fig. 3B). These results are strong evidence that BEST does indeed reduce the effects of deep coalescence on species tree inference, as suggested by Liu et al. (2008). As such, these results suggest that incomplete lineage sorting, rather than hybridization, may explain the allelic polyphyly of *L. californicum*.

4.3. Phylogenetic and phylogeographic relationships

Given that our study was focused on determining the variability and potential utility of COSII regions and, as a result, included a small sample of taxa, limited conclusions can be made regarding phylogenetic relationships. However, across almost all loci there is consistent support for the monophyly of Old World species (included in this study: *L. ferocissimum*, *L. oxycarpum*, and *L. tenue*), a relationship that has been strongly suggested by both GBSSI and cpDNA data and five to eight times greater taxonomic sampling (Levin and Miller, 2005; Levin et al., 2007). There is also a close relationship between the three North American species

L. andersonii, *L. californicum*, *L. carolinianum*; the latter species also occurs on various pacific islands including Hawaii and Easter Island. A close relationship between at least two of these three species has been previously suggested by cpDNA and nuclear GBSSI, ITS, and NIA data (Miller, 2002; Levin and Miller, 2005; Levin et al., 2007, 2009). Thus, phylogenetic relationships using COSII regions generally reflect those observed in other studies, with much greater taxon sampling, which suggests that COSII regions will be useful for understanding species-level relationships in *Lycium*.

Inclusion of multiple individuals of *Lycium californicum* indicates that some of the COSII regions may also be useful for phylogeographic studies. It is notable that in *Lycium californicum*, alleles within individuals for COS16 and COS30 occurred in only one of two clades (i.e., either the *L. californicum* clade sister to other North American species, *L. andersonii* and *L. carolinianum*, or the clade sister to the Old World species, *L. ferocissimum*, *L. oxycarpum*, and *L. tenue*), and this pattern may suggest population-level fixation of COS16 and COS30 alleles. The only exception is the *L. californicum* PR individual from Zacatecas, Mexico, which had alleles in both clades for both COS16 and COS30 (Fig. 1B and D). To fully understand possible allele fixation and the utility of COS16 and COS30 for population-level studies, additional sampling within populations is warranted, given that only one individual per population was included in the present study.

The distribution of *Lycium californicum* alleles between the two clades in the COS30 gene tree further suggests that there may be a geographic pattern in the allele affinities (Fig. 1D). Excluding the above mentioned *L. californicum* PR individual, the alleles in the clade with *L. andersonii* and *L. carolinianum* are from individuals in the more northern parts of the species range (San Diego, California, USA; the northern part of Baja California and northern Sonora, Mexico). In contrast, alleles included in the clade with the Old World species plus *Phrodus microphyllus* are from individuals in more southern and eastern parts of the species range (central to southern Sonora and Coahuila, Mexico), with an exception being *L. californicum* PX, collected from one of the more northern parts of the species range outside of Phoenix, Arizona. These results indicate that COS30 may be useful for assessing relationships among populations. Addition of individuals from more southern populations (e.g., from Baja California Sur, Mexico) and more northern populations (e.g., from the Channel Islands, California, USA) would provide a test of this hypothesis of geographic structuring of alleles.

4.4. Phylogenetic utility of COSII regions

Although the COSII sequences are highly variable, inclusion of multiple individuals of *Lycium californicum* suggests that the variation is generally much lower within species than among species. When the amount of phylogenetic signal (as measured with parsimony) is directly compared across ten species for the same eight loci (Table 3), three of the five COSII regions (COS14, COS16, and COS30) have low homoplasy (rescaled consistency index >0.80), which is comparable to that of GBSSI and cpDNA. This low homoplasy is perhaps surprising, given that these COSII regions are over twice as variable as GBSSI and cpDNA data and have very high intronic content (Table 3). COS14, COS16, and COS30 have the highest numbers of parsimony-informative (PI) characters (65–126) and a considerably higher number of PI characters than GBSSI, ITS, or cpDNA (22–29). Furthermore, these three COSII regions range in length from 984 to 1390 bp and can be sequenced in two reactions with considerable overlap. As mentioned above, the topologies inferred from the COSII regions generally agree with those relationships suggested by the more commonly used loci, and the conflicts can be attributed to either taxon sampling or hybridization, introgression, or incomplete lineage sorting.

Our data for Lycieae suggest that the highly variable COSII regions investigated here will be useful for phylogenetic inference among closely related species and genera. This has also been suggested for *Solanum* species (Rodríguez et al., 2006; Rodríguez and Spooner, 2007, 2008; Ames and Spooner, 2008; Fajardo and Spooner, 2008); however, the identity of the COSII markers employed in these studies is not yet clear. It will be interesting to determine if those regions most useful for Lycieae (COS14, COS16, and COS30) were similarly useful in *Solanum*. If the same COSII regions are indeed informative for fine-scale phylogenetic studies in both *Lycium* and *Solanum*, then this would provide strong evidence for their prospective utility within other Solanaceae genera.

COSII regions are also potentially useful in other Asterids. To-date a few of these regions have been amplified in *Coffea* (Wu et al., 2006) and Acanthaceae (Levin et al., unpubl. data), but detailed studies have yet to be conducted. In addition, it is likely that other COSII regions, perhaps with lower intronic contents than those sequenced in the present study, would be useful for phylogenetic studies at deeper phylogenetic levels. There are ca. 1650 available COSII primer sets (http://www.sgn.cornell.edu/markers/cosii_markers.pl; Wu et al., 2006), which offer a vast resource for addressing phylogenetic questions at a range of taxonomic levels.

5. Conclusions

The level of variation demonstrated by the COSII regions suggests that they should be effective for inference of fine-scale relationships, particularly among the closely related and species rich (ca. 35 species) Old World clade of *Lycium*. Within this group there are both diploid, hermaphroditic and polyploid, dimorphic species. With ongoing and increased sampling in the Old World, we should be able to better understand whether hybridization has contributed to this pattern and, if so, elucidate possible parental lineages. One drawback of the COSII regions employed in the present study is that allelic length polymorphisms are common, requiring cloning. Thus, it is important to determine whether more sequence data or more taxa are best for addressing a specific question of interest. As to the universality of such markers, they will definitely be useful for researchers across Solanaceae. However, they may also be useful for other Asterids and possibly other lineages, given that the markers were originally designed using both *Solanum* and *Capsicum* (Solanaceae) and *Coffea* (Rubiaceae) and have known homologies with *Arabidopsis* (Wu et al., 2006). Other similar markers from the pentatricopeptide repeat gene family are currently being explored in *Arabidopsis*, various Poaceae, and Verbenaceae (Yuan et al., 2009); the intronless nature of these regions offer a benefit over COSII, given that one of the major drawbacks of COSII is the considerable allelic length polymorphisms in the large intronic regions.

Finally, we agree with recent studies (Belfiore et al., 2008; Brumfield et al., 2008; Liu et al., 2008) that BEST is a powerful new approach to addressing the challenge of inferring a robust species phylogeny from multiple (and sometimes incongruent) gene trees, especially where alleles within species are polyphyletic. A main constraint at present is that BEST analyses are time consuming, and with larger numbers of taxa (as in most phylogenetic studies) such an approach will be untenable. Development of this method, as well as other similar approaches, are an active field of research and provide a valuable (and much needed) set of tools for understanding relationships among closely related groups of species or populations.

Acknowledgments

The authors thank an anonymous reviewer for helpful comments; N. Feliciano, S. Barmettler, and A. Kamath for laboratory

assistance; Y. Wang for guidance navigating the available COSII markers; and L. Liu for assistance with BEST. This work was supported by National Science Foundation Grant DEB-0343735 to R.A.L. and J.S.M. and an REU supplement for A.W. Additional support was provided by the Department of Biology at Amherst College.

References

Álvarez, I., Costa, A., Nieto Feliner, G., 2008. Selecting single-copy nuclear genes for plant phylogenetics: a preliminary analysis for the Senecioneae (Asteraceae). J. Mol. Evol. 66, 276–291.
Ames, M., Spooner, D.M., 2008. Phylogeny of *Solanum* Series *Piurana* and Putatively Related Series Based on Morphological and Molecular Data. Abstract, Botany 2008, Vancouver, BC, Canada.
Ané, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24, 412–426.
Barrett, M., Donoghue, M.J., Sober, E., 1991. Against consensus. Syst. Zool. 40, 486–493.
Belfiore, N.M., Liu, L., Moritz, C., 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). Syst. Biol. 57, 294–310.
Bernardello, L.M., 1982. Estudios en *Lycium* (*Solanaceae*) II. Recuentos cromosomicos en entidades Argentinas. Hickenia 1, 321–328.
Bernardello, L.M., 1986. Revisión taxonómica de las especies sudamericanas de *Lycium* (Solanaceae). Bol. Acad. Nac. Cienc. Córdoba 57, 173–356.
Brumfield, R.T., Liu, L., Lum, D.E., Edwards, S.V., 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae, *Manacus*) from multilocus sequence data. Syst. Biol. 57, 719–731.
Carstens, B.C., Knowles, L.L., 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst. Biol. 56, 400–411.
Chapman, M.A., Change, J., Weisman, D., Kesseli, R.V., Burke, J.M., 2007. Universal markers for comparative mapping and phylogenetic analysis in the Asteraceae. Theor. Appl. Genet. 115, 747–755.
Cronn, R.C., Small, R.L., Haselkorn, T., Wendel, J.F., 2002. Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Am. J. Bot. 89, 707–725.
Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2, e68. doi:10.1371/journal.pgen.0020068.
Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340.
Degnan, J.H., Salter, L.A., 2005. Gene tree distributions under the coalescent process. Evolution 59, 24–37.
Degnan, J.H., DeGiorgio, M., Bryant, D., Rosenberg, N.A., 2008. Properties of consensus methods for inferring species trees from gene trees. arxiv:/0802.2355v1 [q-bio.PE].
Eckert, A.J., Carstens, B.C., 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Mol. Phylogenet. Evol. 49, 832–842.
Edwards, S.V., Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104, 5936–5941.
Fajardo, D., Spooner, D.M., 2008. Preliminary Phylogeny of *Solanum* Series *Conicibaccata* Based on Morphological and Molecular Data. Abstract, Botany 2008, Vancouver, BC, Canada.
Gatesy, J., Baker, R.H., 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst. Biol. 54, 483–492.
Howarth, D.G., Baum, D.A., 2002. Phylogenetic utility of a nuclear intron from nitrate reductase for the study of closely related plant species. Mol. Phylogenet. Evol. 23, 525–528.
Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17, 754–755.
Kluge, A.G., 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst. Zool. 38, 7–25.
Kress, W.J., Erickson, D.L., 2007. A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. PLoS ONE 2, e508. doi:10.1371/journal.pone.0000508.
Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56, 17–24.
Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T.G., Savolainen, V., 2008. DNA barcoding the floras of biodiversity hotspots. Proc. Natl. Acad. Sci. USA 105, 2923–2928.
Levin, R.A., Miller, J.S., 2005. Relationships within tribe Lycieae (Solanaceae): paraphyly of *Lycium* and multiple origins of gender dimorphism. Am. J. Bot. 92, 2044–2053.
Levin, R.A., Bernardello, G., Venter, A.M., Shak, J.R., Miller, J.S., 2007. Evolutionary relationships in tribe Lycieae (Solanaceae). Acta Hort. 745, 225–239.
Levin, R.A., Blanton, J., Miller, J.S., 2009. Phylogenetic utility of nuclear nitrate reductase: a multi-locus comparison of nuclear and chloroplast sequence data for inference of relationships among American Lycieae (Solanaceae). Mol. Phylogenet. Evol. 50, 608–617.
Li, M., Wunder, J., Bissoli, G., Scarponi, E., Gazzani, S., Barbaro, E., Saedler, H., Varotto, C., 2008. Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. Cladistics 24, 727–745.
Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24, 2542–2543.
Liu, L., Pearl, D.K., 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56, 504–514.
Liu, L., Pearl, D.K., Brumfield, R.T., Edwards, S.V., 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution 62, 2080–2091.
Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.
Miller, J.S., 2002. Phylogenetic relationships and the evolution of gender dimorphism in *Lycium* (Solanaceae). Syst. Bot. 27, 416–428.
Miller, J.S., Venable, D.L., 2000. Polyploidy and the evolution of gender dimorphism in plants. Science 289, 2335–2338.
Miller, J.S., Kamath, A., Levin, R.A., 2009. Do multiple tortoises equal a hare? The utility of nine noncoding plastid regions for species-level phylogenetics in tribe Lycieae (Solanaceae). Syst. Bot. 34, in press.
Miller, J.S., Levin, R.A., Feliciano, N.M., 2008. A tale of two continents: Baker's rule and the maintenance of self-incompatibility in *Lycium* (Solanaceae). Evolution 62, 1052–1065.
Mueller, L.A., Solow, T.H., Taylor, N., Skwarecki, B., Buels, R., Binns, J., Lin, C., Wright, M.H., Ahrens, R., Wang, Y., Herbst, E.V., Keyder, E.R., Menda, N., Zamir, D., Tanksley, S.D., 2005. The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond. Plant Physiol. 138, 1310–1317.
Olmstead, R.G., Bohs, L., Abdel Migid, H., Santiago-Valentin, E., Garcia, V.F., Collier, S.M., 2008. A molecular phylogeny of the Solanaceae. Taxon 57, 1159–1181.
Paithankar, K.R., Prasad, K.S.N., 1991. Precipitation of DNA by polyethylene glycol and ethanol. Nucleic Acids Res. 19, 1346.
Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.
Rambaut, A., 2002. Se-Al: sequence alignment editor, version 2.0. Available from: <http://evolve.zoo.ox.ac.uk/>.
Rambaut, A., Drummond, A.J., 2007. Tracer v 1.4. Available from: <http://beast.bio.ed.ac.uk/Tracer/>.
Rodríguez, F., Spooner, D.M., 2007. Multi Nuclear Gene Phylogenetic Analyses of Potatoes and Tomatoes. Abstract, Botany & Plant Biology 2007, Chicago, IL, USA.
Rodríguez, F., Spooner, D.M., 2008. Do Potatoes and Tomatoes have a Single Evolutionary History? What Proportion of the Genome Supports this History? Abstract, Botany 2008, Vancouver, BC, Canada.
Rodríguez, F., Wu, F., Tanksley, S., Spooner, D., 2006. A Multiple Single-Copy Gene Phylogenetic Analysis of Wild Tomatoes (*Solanum* L. section *Lycopersicon* (Mill.) Wettst.) and their Outgroup Relatives. Abstract, Solanaceae 2006 Meeting, Madison, WI, USA.
Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.
Ronquist, F., Huelsenbeck, J.P., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Rosenberg, N.A., Tao, R., 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. Syst. Biol. 57, 131–140.
Sang, T., 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. 37, 121–147.
Savage, A.E., Miller, J.S., 2006. Gametophytic self-incompatibility in *Lycium parishii* (Solanaceae): allelic diversity, genealogical structure, and patterns of molecular evolution. Heredity 96, 434–444.
Shaw, J., Lickey, E.B., Beck, J.T., Farmer, S.B., Liu, W., Miller, J., Sirpun, 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. Am. J. Bot. 92, 142–166.
Shaw, J., Lickey, E.B., Schilling, E.E., Small, R.L., 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am. J. Bot. 94, 275–288.
Small, R.L., Cronn, R.C., Wendel, J.F., 2004. Use of nuclear genes for phylogeny reconstruction in plants. Aust. Syst. Bot. 17, 145–170.
Steele, P.R., Guisinger-Bellian, M., Linder, C.R., Jansen, R.K., 2008. Phylogenetic utility of 141 low-copy nuclear regions in taxa at different taxonomic levels in two distantly related families of rosids. Mol. Phylogenet. Evol. 48, 1013–1026.
Swofford, D.L., 2002. PAUP*: phylogenetic analysis using parsimony (* and other methods), version 4. Sinauer, Sunderland, Massachusetts, USA.
Venter, A.M., 2000. Taxonomy of the genus *Lycium* L. (Solanaceae) in Africa., Ph.D. dissertation, University of the Free State, Bloemfontein, South Africa.
Wu, F., Mueller, L.A., Crouzillat, D., Pétiard, V., Tanksley, S.D., 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics 174, 1407–1420.
Yeung, K., Miller, J.S., Savage, A.E., Husband, B.C., Igic, B., Kohn, J.R., 2005. Association of ploidy and sexual system in *Lycium californicum* (Solanaceae). Evolution 59, 2048–2055.
Yuan, Y.W., Olmstead, R.G., 2008. Evolution and phylogenetic utility of the *PHOT* gene duplicates in the *Verbena* complex (Verbenaceae): dramatic intron size variation and footprint of ancestral recombination. Am. J. Bot. 95, 1166–1176.
Yuan, Y.W., Liu, C., Marx, H.E., Olmstead, R.G., 2009. The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies. New Phytol. 182, 272–283.

Please cite this article in press as: Levin, R.A., et al. The utility of nuclear conserved ortholog set II (COSII) genomic regions for species-level phylogenetic inference in *Lycium* (Solanaceae). Mol. Phylogenet. Evol. (2009), doi:10.1016/j.ympev.2009.08.016