JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12201

Research Article

Global versus Chinese perspectives on the phylogeny of the N-fixing clade

Hong-Lei Li[1,2,3], Wei Wang[1], Rui-Qi Li[1], Jing-Bo Zhang[1], Miao Sun[1], Rehan Naeem[1], Jun-Xia Su[4], Xiao-Guo Xiang[1], Peter E. Mortimer[5,6], De-Zhu Li[6,7], Kevin D. Hyde[5,6,8], Jian-Chu Xu[5,6], Douglas E. Soltis[9,10,11], Pamela S. Soltis[10,11], Jianhua Li[12], Shou-Zhou Zhang[2], Hong Wu[3], Zhi-Duan Chen[1*], and An-Ming Lu[1*]

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[2]Shenzhen Key Laboratory of Southern Subtropical Plant Diversity, Fairylake Botanical Garden, Shenzhen and Chinese Academy of Sciences, Shenzhen 518004, Guangdong, China
[3]College of Life Sciences, South China Agricultural University, Guangzhou 510642, China
[4]College of Life Sciences, Shanxi Normal University, Linfen 041004, Shanxi, China
[5]World Agroforestry Centre, East and Central Asia, Kunming 650201, China
[6]Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China
[7]Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China
[8]School of Science, Mae Fah Luang University, Chiang Rai 57100, Thailand
[9]Department of Biology, University of Florida, Gainesville, FL 32611, USA
[10]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA
[11]Genetics Institute, University of Florida, Gainesville, FL 32610, USA
[12]Biology Department, Hope College, Holland, MI 49423, USA
*Authors for correspondence. Z.-D. Chen. E-mail: zhiduan@ibcas.ac.cn. Tel.: 86-10-62836434. Fax: 86-10-62836095.
A.-M. Lu. E-mail: anmin@ibcas.ac.cn. Tel.: 86-10-62836448. Fax: 86-10-62836095.
Received 9 November 2015; Accepted 3 March 2016; Article first published online 22 July 2016

Abstract There has been increasing interest in integrating a regional tree of life with community assembly rules in the ecological research. This raises questions regarding the impacts of taxon sampling strategies at the regional versus global scales on the topology. To address this concern, we constructed two trees for the nitrogen-fixing clade: (i) a genus-level global tree including 1023 genera; and (ii) a regional tree comprising 303 genera, with taxon sampling limited to China. We used the supermatrix approach and performed maximum likelihood analyses on combined *matK*, *rbcL*, and *trnL-F* plastid sequences. We found that the topology of the global and the regional tree of the N-fixing clade were generally congruent. However, whereas relationships among the four orders obtained with the global tree agreed with the accepted topology obtained in focused analyses with more genes, the regional topology obtained different relationships, albeit weakly supported. At a finer scale, the phylogenetic position of the family Myricaceae was found to be sensitive to sampling density. We expect that internal support throughout the phylogeny could be improved with denser taxon sampling. The taxon sampling approach (global vs. regional) did not have a major impact on fine-level branching patterns of the N-fixing clade. Thus, a well-resolved phylogeny with relatively dense taxon sampling strategy at the regional scale appears, in this case, to be a good representation of the overall phylogenetic pattern and could be used in ecological research. Otherwise, the regional tree should be adjusted according to the correspondingly reliable global tree.

Key words: global tree of life, N-fixing clade, phylogeny, regional tree of life, supermatrix.

The tree of life has been widely applied as a useful tool in different areas of ecological research (Ma, 2013; Lu et al., 2014). An upsurge in ecological studies incorporating phylogenetic information with community dynamics has been seen in the last decade (Webb et al., 2002; Kraft et al., 2007; Cavender-Bares et al., 2009; Asner & Martin, 2011; Gregory et al., 2014; Robert, 2015). Nitrogen-fixing (N-fixing) plants are an important component of biological communities. The ability to fix and use atmospheric N through a process known as biological N-fixation complements the limited bioavailability of N and has received a lot of attention. In particular, there is hope that N-fixing genes can be transferred to non-N-fixing crop species (Soltis et al., 1995; Ferguson et al., 2010; Santi et al., 2013; Venkateshwaran et al., 2013).

The N-fixing clade was first recovered and proposed by Soltis et al. (1995) based on initial molecular phylogenetic analyses of angiosperms. The clade as now recognized contains 28 families (of which, 10 families contain N-fixing species), over 1300 genera, and approximately 30 000 species. Many species of the N-fixing clade have great economic importance as crop plants including legumes (Fabaceae), fruit crops (Rosaceae), and vegetables (Cucurbitaceae). Many other species play a crucial role in biological communities because of their ability to fix atmospheric N through a symbiosis with N-fixing bacteria in root nodules (Li et al., 2015). Additionally, the clade possesses many woody species that are distributed in temperate and tropical forests (e.g., Betulaceae, Fagaceae, Fabaceae, and Ulmaceae) (Croat, 1978; Elias, 1980; Lopez et al., 1987), forests of extreme importance to biological communities (Wang et al., 2009; Fang et al., 2012; Huang et al., 2012).

The N-fixing clade has long been recognized as monophyletic and inclusive of the four orders Cucurbitales, Fabales, Fagales, and Rosales (Soltis et al., 2000, 2008, 2011; Zhu et al., 2007; APG III, 2009; Wang et al., 2009; Moore et al., 2010, 2011). Relationships among these orders, as well as within these orders, are generally well resolved. For example, the relationships among families in Fagales are well resolved except for the position of Myricaceae (Li et al., 2004; Herbert et al., 2006; Bell et al., 2010; Soltis et al., 2011; Xiang et al., 2014). Within Cucurbitales, the relationships of Begoniaceae, Datiscaceae, and Tetramelaceae remain unclear (Swensen et al., 1994, 1998; Zhang et al., 2006; Schaefer & Renner, 2011). The four families within Fabales were strongly supported as monophyletic, but branching orders of the families have not yet been clarified (Bello et al., 2009; Wang et al., 2009; Soltis et al., 2011), and the main clades of Fabaceae are generally resolved, however, the relationships among some of them are still unclear (Wojciechowski et al., 2004; Cardoso et al., 2012a, 2012b; The Legume Phylogeny Work Group, 2013). The interfamilial relationships within Rosales were resolved with Rosaceae sister to the rest of the order (e.g., Savolainen et al., 2000a, 2000b; Hilu et al., 2003; Soltis et al., 2007, 2011), with the remaining families forming two distinct clades: one clade of Ulmaceae and relatives, and a second comprising Rhamnaceae and relatives (see Richardson et al., 2000; Savolainen et al., 2000a; Sytsma et al., 2002; Zhang et al., 2011). However, phylogenetic relationships among Barbeyaceae, Dirachmaceae, Elaeagnaceae, and Rhamnaceae within Rosales remain unclear. Moreover, within Rosaceae, three subfamilies are monophyletic with strong support, however, the position of Dryadoideae remains uncertain (Potter, 2003; Potter et al., 2007; Chin et al., 2014). Informally, Rosaceae may be composed of clades of rhamnoids, ziziphoids, and ampeloziziphoids; however, the relationships among these are unresolved (Richardson et al., 2000).

The combination of many genes and whole plastid genome sequence data has led to an improved understanding of the deep-level phylogeny of the N-fixing clade within the framework of all angiosperms (Leebens-Mack et al., 2005; Jansen et al., 2007; Moore et al., 2010, 2011; Soltis et al., 2011). However, fewer than 40 taxa of the N-fixing clade were included in the most taxonomically robust of these studies. Some studies recovered the basic phylogenetic framework of the orders within the N-fixing clade using broad taxonomic coverage, but the support values for some crown clades was low (Bello et al., 2009; Zhang et al., 2011). Crown clades may be resolved with denser taxonomic sampling, but denser sampling typically requires a prohibitive volume of sequence data for a many-gene phylogeny. As an alternative, deep-level phylogenies of families or orders may be resolved with a smaller number of markers from spacer regions (Richardson et al., 2000; Li et al., 2004; Zhang et al., 2006). Using spacer regions is a possible way to construct a well-resolved phylogeny with dense sampling.

A well-supported and resolved topology is crucial to carry out reliable, downstream ecological analysis (Cavender-Bares et al., 2009; Roquet et al., 2013). A regional tree of life is useful for studying phylogenetic diversity, community assembly rules, conservation biology, and niche evolution in a distinct area (Whitney et al., 2009; Asner & Martin, 2011; Schaefer et al., 2011). However, the impact of the taxon sampling strategy on the topology at the regional compared to a global scale has not been rigorously studied.

In this study, we selected three plastid regions, *matK*, *rbcL*, and *trnL-F* spacer and reconstructed the most comprehensive phylogeny (global tree) of the N-fixing clade to date, comprising 1023 species at the generic level using the supermatrix approach. We also reconstructed a regional tree for the N-fixing clade with 303 genera native to China. Another two global trees that include the remaining 726 operational taxonomic units (OTUs) and 303 randomly selected OTUs from this remaining group (including 6 outgroups) were also reconstructed. Our aims are to establish a tree of life of the N-fixing clade at the generic level and to test the impact of the taxon sampling density at the regional or global scale by comparing the two topologies.

Material and Methods

Taxon sampling

Through our sampling approach we tried to maximize the taxonomic coverage of each of the previously recognized genera (Stevens, 2001 onwards and references therein) within the N-fixing clade. DNA samples for some species used here were extracted from dried materials in silica gel. Sequences of most species were obtained from GenBank. We constructed a three-marker data matrix for 1023 species. Species names and GenBank accession numbers are presented in Table S1.

For newly generated sequences, we isolated genomic DNA from silica gel-dried materials using a Plant Genomic DNA Kit (Beijing Biomed, Beijing, China) or from herbarium samples following a modified CTAB procedure (Doyle & Doyle, 1987). DNA regions were amplified with polymerase chain reaction (PCR). We carried out PCR amplifications using the primers in Li et al. (2013) and 2× *Taq* PCR MasterMix (Beijing Biomed) in 25-μL reactions with the following thermocycler program: 2 min at 95 °C for denaturation, then 35 cycles of 30 s at 95 °C, 30–60 s at 53–57 °C for annealing, 2 min 30 s at 72 °C for primer extension, and a 10-min incubation at 72 °C following the cycles. The PCR products were purified using a GFX PCR DNA and Gel Band Purification Kit (Amersham Pharmacia Biotech, Piscataway, NJ, USA) and then directly sequenced them. Sequencing reactions were carried out using an ABI Prism BigDye Terminator Cycle Sequencing Kit (Applied Biosystems,

Beijing, China). We then processed the sequences using ABI 3730xl DNA Analysis Systems and following the manufacturer's protocols.

For sequences from GenBank, all available nucleotide sequences were selected from the three plastid regions (*matK*, *rbcL*, and *trnL-F*) representing the N-fixing clade. For each taxon, we tried to use the same species and DNA sample across the three plastid markers, but some composite accessions were necessary to represent genera. The longest sequence was selected when multiple sequences were available, and randomly selected one sequence when there were multiples of the same length. Most of the DNA sequences have been used in previously published studies (e.g., Li et al., 2004; Wojciechowski et al., 2004; Zhang et al., 2011).

DNA alignment

The *rbcL* sequences were aligned directly in the program MUSCLE using the default settings at the high accuracy parameter (Edgar, 2004), and the resulting alignment was manually adjusted by eye, using BioEdit version 5.0.9 (Hall, 1999). A two-step strategy was used to align the fast-evolving *matK* and *trnL-F* regions. First, we divided the sequences into clusters according to sequence length and taxonomic unit. Each cluster was aligned in MUSCLE under default high accuracy parameters, and then manually adjusted the alignment. Then we aligned the clusters with the profile–profile alignment algorithm in MUSCLE. Final adjustments were made to the alignments for these two genes using the MUSCLE refinement algorithm and then manually, especially to trim for quality and maximum coverage. The aligned global matrix contains 1023 OTUs. To compare with the global tree of the N-fixing clade, we constructed a regional matrix with 303 OTUs representing the N-fixing clade. In the regional matrix, all the genera have representatives distributed in China. The remaining 726 OTUs and 303 randomly selected OTUs from this remaining group (including 6 outgroups) were also prepared for the maximum likelihood (ML) analyses.

Phylogenetic analyses

The program RAxML version 7.6.6 (Stamatakis, 2006) was used to carry out the initial phylogenetic analysis under the ML criterion for each marker. No significant bootstrap (BS) support for conflicting nodes was evident (taken here as exceeding 70%), so the data from different markers for subsequent analyses were combined. Phylogenetic analyses of the combined dataset of three DNA regions using ML methods were carried out. The ML analysis was performed using RAxML with the following options: three data partitions (*rbcL*, *matK*, and *trnL-F*), GTR + I + Γ nucleotide substitution model, and 1000 non-parametric BS replicates. The gaps were treated as missing data. The program was run on the CIPRES network (Miller et al., 2010).

Results

For most nodes, the three-marker global tree showed higher BS support than the individual marker and regional trees. Thus, only the global tree is described below (Fig. S1; and interconnected subtrees in Fig. S2 for clearer visualization). We examine the regional tree (Figs. S3, S4) in the Discussion section under "Comparison of global and regional trees of the N-fixing clade" (below). The topologies with 726 and 303 OTUs are shown in Figs. S5 and S6.

Based on the combined three-marker dataset, we generated a well-resolved phylogeny of the N-fixing clade. Each of the four orders is strongly supported as monophyletic with BS value >80% (Fig. S1). Fabales are sister to the other three orders (BS = 100%), and Rosales are sister to Fagales and Cucurbitales (BS = 84%). Relationships within the four orders are summarized as follows.

Within Fabales, the monophyly of the four families are well supported (BS ≥ 99%). Within Fabaceae, subfamily Caesalpinioideae are paraphyletic and at the base of the family, whereas subfamilies Mimosoideae and Papilionoideae are well supported as monophyletic. In subfamily Caesalpinioideae, eight monophyletic clades were recovered: Cercideae, Deterieae s.l., Dialiinae s.l., Umtiza clade, Cassia clade, Caesalpinia clade, Tachigali clade, and Peltophorum clade. Within Mimosoideae, resolution of the large, higher-level mimosoid clades (e.g., tribal or generic level) is problematic. In subfamily Papilionoideae, 15 monophyletic clades were recovered: Swartzioid, Dipterygeae clade, Amburana clade, Cladrastis clade, Andira clade, Lecointeoid clade, Vataireoid, Dalbergioid s.l., Genistoid, Baphioid, Mirbelioids, Robinioid clade, inverted repeat-lacking clade (IRLC), Indigofereae, and Millettioid clade (inverted-repeat-lacking clade).

Within Rosales, Rosaceae were resolved as sister to other members of Rosales. The remaining families comprise a well-supported clade (BS = 99%). Within Rosaceae, three subfamilies Spiraeoideae, Dryadoideae, and Rosoideae were retrieved. Informally, we identified three well-supported clades in Rhamnaceae: Ampeloziziphoids (BS = 100%), Rhamnoids (BS = 99%), and Ziziphoids (BS = 100%). Ulmaceae comprise two well supported clades, each with BS = 99%: *Ampelocera* + *Holoptelea* and *Hemiptelea* + (*Zelkova* + *Ulmus*). Within Cannabaceae, *Aphananthe* was well supported as sister to the rest of the family. *Gironniera* + *Lozanella* was sister to the remainder. Within Moraceae, well-supported monophyletic clades of Castilleae (BS = 99%) and Dorstenieae s.l. (Clement & Weiblen, 2009) (BS = 99%) were detected. Two additional well-supported clades were Moreae (minus *Streblus*) (BS = 100%) and Artocarpeae (excluding *Hullettia* and *Parartocarpus*) (BS = 90%). In Urticaceae, four strongly supported clades (clade I–IV, Fig. S2k) were recognized and the relationships among them were well resolved.

In Cucurbitales, strong support (BS = 100%) was found for a clade of Corynocarpaceae + Coriariaceae as sister to a moderately supported (BS = 63%) clade consisting of the remaining Cucurbitales. Cucurbitaceae were well represented at the genus level in the current study. We recovered most tribes, including Fevilleeae, Actinostemmateae, Telfairieae, Bryonieae, Sicyeae, Schizopeponeae, Coniandreae, Cucurbiteae, and Benincaseae, *sensu* Schaefer & Renner (2011) based on the analyses of 14 DNA regions from the three plant genomes.

All families within Fagales had BS = 100%. Nothofagaceae were sister to the remaining Fagales (BS = 100%), followed by Fagaceae, which are sister to the remainder of the Fagales, with strong support (BS = 100%). The rest of Fagales formed

two clades: Casuarinaceae + (Ticodendraceae + Betulaceae) (BS = 100%) were sister to Myricaceae + (Rhoipteleaceae + Juglandaceae) (BS = 59%). In Betulaceae, *Alnus* was the sister to the remainder of Betulaceae (BS = 100%), with subsequent divergence order of *Betula* as sister to two clades: *Corylus* + *Ostryopsis* (BS = 99%) and *Ostrya* + *Carpinus* (BS = 100%). Within the Myricaceae, *Canacomyrica* was resolved as sister to *Myrica* + *Comptonia*. In Juglandaceae, two major clades were recovered: (i) *Alfaroa* + (*Engelhardia* + *Alfaropsis*) with BS = 99%; and (ii) *Annamocarya*, subsequently followed by *Platycarya*, *Cyclocarya*, and *Pterocarya* as sister to *Juglans* + *Carya* with BS = 98%.

Discussion

New interfamilial and intrafamilial relationships

Within the Fabales, defining the relationships among the four families has been particularly problematic in the past (Wojciechowski et al., 2004; Bruneau et al., 2008; Bello et al., 2009, 2012; Bell et al., 2010; Soltis et al., 2011). The topology resolved here is Quillajaceae + Surianaceae as sister to a weakly supported clade of Polygalaceae + Fabaceae. Persson (2001) suggested the relationships Polygalaceae + (Surianaceae + (Quillajaceae + Fabaceae)), but there was little support. In Doyle et al. (2000), Quillajaceae are sister to the other three families. Qiu et al. (2010) supported the relationships of Quillajaceae + (Fabaceae + (Surianaceae + Polygalaceae)) with weak support. In other analyses, the topology Polygalaceae + (Leguminosae + (Quillajaceae + Surianaceae)) is considered as the most likely hypothesis of interfamilial relationships of the order (Wojciechowski et al., 2004; Bruneau et al., 2008; Bello et al., 2009, 2012). Soltis et al. (2011) recovered a topology (Polygalaceae + Quillajaceae) + (Leguminosae + Surianaceae) upon the analyses of 17 genes, however, the support was weak and the taxon sampling in Fabales was low.

In the Rosales, the monophyly of Rhamnaceae has not been resolved by our work. Nevertheless, we identified three well-supported clades in Rhamnaceae: Ampeloziziphoids, Rhamnoids, and Ziziphoids. *Ventilago* was sister to Rhamnoids *sensu* Richardson et al. (2000) with strong support and should be included in Rhamnoids. Within Rosaceae, in agreement with Chin et al. (2014), Spiraeoideae are sister to Dryadoideae + Rosoideae. This result differs from a prior study focused on the family (Potter et al., 2007). However, the sister relationship of Dryadoideae and Rosoideae was supported by the result of the independent gene trees of *rbcL* and *matK* in Potter et al. (2007).

Relationships in Cucurbitales are similar to other recent analyses (e.g., Zhang et al., 2006; Soltis et al., 2007, 2011; Schaefer & Renner, 2011). However, we found that there was strong support (BS = 100%) for a clade of Corynocarpaceae + Coriariaceae as sister to the remaining Cucurbitales. Begoniaceae were resolved as sister to a well-supported (BS = 86%) clade of Datiscaceae + Tetramelaceae. However, Begoniaceae are resolved as sister to Datiscaceae with only moderate support in some analyses (Zhang et al., 2006; Schaefer et al., 2009; Schaefer & Renner, 2011).

Within the Fagales, the position of Myricaceae we present here is in agreement with the results of previous analyses (Li et al., 2004; Bell et al., 2010; Soltis et al., 2011). In contrast, Xiang et al. (2014) shows a close relationship between Myricaceae and clade Casuarinaceae + (Ticodendraceae + Betulaceae). In Betulaceae, *Alnus* is the sister to the remainder of Betulaceae (BS = 100%), followed by *Betula* as sister to two clades: *Corylus* + *Ostryopsis* (BS = 99%) and *Ostrya* + *Carpinus* (BS = 100%). These results agree well with Li et al. (2004). In some prior analyses (e.g., Forest et al., 2005; Grimm & Renner, 2013) *Betula* was resolved as sister to *Alnus*, but the support value was low.

Comparison of global and regional trees of the N-fixing clade

To test whether regional taxon sampling results in a tree with different branching patterns compared to a global tree, we compared the differences in the phylogenetic relationships of the N-fixing clade among our global and Chinese regional trees. The global and regional trees showed congruence in general, but the regional tree showed weaker support for some relationships (Figs. S1–S4). Within the Fabales, Polygalaceae were sister to Fabaceae in the regional tree, as in the global tree, although the support was lower (BS = 39% regional; BS = 45% global). In Surianaceae of the Fabales, our global tree showed strong support for *Recchia* + *Lundellia* as sister to *Suriana* + (*Cadellia* + *Stylobasium*), as described in Crayn et al. (1995), Forest et al. (2007), and Bello et al. (2009). In Polygalaceae, we found four monophyletic tribes with Xanthophylleae sister to the remaining Polygalaceae and Moutabeae sister to Carpolobieae + Polygaleae. These results are in agreement with previous molecular studies, especially Forest et al. (2007) and Bello et al. (2012). We recovered major clades in Fabaceae that were in accordance with previous studies (Doyle et al., 1997; Bruneau et al., 2001, 2008; Wojciechowski et al., 2004; Cardoso et al., 2012a, 2012b; Manzanilla & Bruneau, 2012). Similar topologies were also recovered in the regional tree. However, in a small number of relationships within Fabaceae, the regional tree showed higher support than the global tree. For example, Cassia clade was sister to Caesalpinia clade with BS = 66% in the regional tree, but support was <50% in the global tree. The sister relationship between Millettioid and Indigofereae got higher BS support in the regional tree (BS = 90%) than in the global tree (BS = 77%). In Rosales, both the global and regional trees agree with other analyses in providing strong support (BS = 99% global; BS = 98% regional) for the placement of Rosaceae as sister to other members of Rosales (Wang et al. 2009; Soltis et al. 2011; Zhang et al., 2011). The clade Ulmaceae + (Cannabaceae + (Moraceae + Urticaceae)) was recognized with strong support (BS = 99% global; BS = 100% regional) and the relationships among these four families were well resolved as in Soltis et al. (2011) and Zhang et al. (2011). Within Urticaceae, the global and regional trees resolved clade I as sister to clade IV (BS = 96% global; BS = 70% regional), and clade II as sister to clade III (BS = 98% global; BS = 68% regional). Similarly, the topology of Cucurbitales in the regional tree was comparable to the global tree, but the support value of the clade Cucurbitaceae + Tetramelaceae + Begoniaceae was lower (BS = 63% global; BS < 50% regional). In Fagales, our global tree showed that Nothofagaceae were sister to the remaining Fagales, followed by Fagaceae as sister to the remainder of Fagales. These findings

are congruent with other published phylogenies (Li et al., 2004; Soltis et al., 2007, 2011; Bell et al., 2010; Xiang et al., 2014).

In some cases, a number of deep level relationships were sensitive to taxon sampling, but the support values of these internal nodes were lower than 80%. The relationships among the four orders in the global tree are congruent with those recovered using large numbers of genes, but with a lower density of taxon sampling (Wang et al., 2009; Moore et al., 2010; Soltis et al., 2011). However, in the regional tree the relationships among the four orders were different or poorly resolved. In the regional tree, Cucurbitales are sister to the other three orders. Fabales were resolved as sister to a moderately supported Fagales + Rosales (BS = 65%) with BS < 50%, which is different from our global tree (Fig. S1). The placement of Fabales within the 726 global tree (Fig. S5) and 303 global tree (Fig. S6) are congruent with that in the global tree. Within Rosaceae, subfamily Dryadoideae were sister to Spiraeoideae in the regional tree with BS < 50%, but different in the global tree with Dryadoideae as sister to Rosoideae (BS = 76%). The relationships among the three subfamilies of Rosaceae in both the 726 global tree and 303 global tree are the same as that in the global tree, other than that in the regional tree. This indicates taxon sampling strategy at the regional scale could lead to a different topology in some cases compared with the sampling strategy at a global scale, albeit with BS < 80%. In Rhamnaceae, the sister relationship of Rhamnoids and Ziziphoids was found in our global tree, and in a number of studies (Richardson et al., 2000; Bell et al., 2010; Soltis et al., 2011; Zhang et al., 2011), although our BS support for the relationship was only 55%. The regional tree recovered a similar relationship, however, *Ventilago* of Rhamnoids is sister to Ziziphoids with low support (BS < 50%). In particular, Myricaceae were sister to Casuarinaceae + Betulaceae with BS = 72% in the regional tree, rather than sister to Juglandaceae in the global tree with BS = 59%. The position of Myricaceae in the 726 global tree agrees with that in the global tree, whereas the position of Myricaceae in the 303 global tree is the same as that in the regional tree. This indicates that the different placement of Myricaceae is caused by the density of taxa sampling other than the regional scale sampling strategy.

Conclusions and Perspectives

The N-fixing clade, *sensu* APG III, contains 28 families (of which 10 are N-fixing), over 1300 genera, and approximately 30 000 species. We present the most comprehensive genus-level phylogenetic hypothesis to date for the N-fixing clade, developed after analysis of three plastid loci, *matK*, *rbcL*, and *trnL-F*. Furthermore, we tested the impacts of taxon sampling strategy at the regional or global scale on the topology by comparing the global and regional trees.

Based on the combined three-marker dataset, we generated a well-resolved phylogeny of the N-fixing clade composed of four plant orders. Each of the four orders was strongly supported as monophyletic (Fig. S1). The deep and crown clades of the global tree recovered in our analyses are largely congruent with those in previous studies, highlighting the utility of spacer regions with sufficient taxon coverage for phylogenetic resolution. Generally, no strong conflicts (BS > 80%) are found among the major clades of global and regional trees of life. Internal support throughout the phylogeny could be improved with denser taxon sampling. A well-resolved phylogeny with relatively dense taxon sampling strategy at the regional scale does not have a negative impact on deep-level branching patterns of the N-fixing clade. Thus, a well-resolved phylogeny (internal nodes with BS > 80%) with a taxon sampling strategy at the regional scale could be used in ecological research. Otherwise, the regional tree should be adjusted according to the correspondingly reliable global tree before being used in ecological research.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant Nos. 31590822, 31500175, 31270268 and 40830209), the National Basic Research Program of China (Grant No. 2014CB954100), Chinese Academy of Sciences International Institution Development Program (No. SAJC201315), Chinese Academy of Sciences External Cooperation Program of the Bureau of International Co-operation (No. GJHZ201321), Chinese Academy of Sciences Visiting Professorship to Douglas E. Soltis as a senior international scientist (Grant No. 2011T1S24), and Shenzhen Key Laboratory of Southern Subtropical Plant Diversity (Grant No. SSTLAB-2014-04).

References

APG III (Angiosperm Phylogeny Group III). 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* 161: 105–121.

Asner GP, Martin RE. 2011. Canopy phylogenetic, chemical and spectra1 assembly in a lowland Amazonian forest. *New Phytologist* 189: 999–1012.

Bell CD, Soltis DE, Soltis PS. 2010. The age and diversification of the angiosperms re-revisited. *American Journal of Botany* 97: 1296–1303.

Bello MA, Bruneau A, Forest F, Hawkins JA. 2009. Elusive relationships within order Fabales: Phylogenetic analyses using *matK* and *rbcL* sequence data. *Systematic Botany* 34: 102–114.

Bello MA, Rudall PJ, Hawkins JA. 2012. Combined phylogenetic analyses reveal interfamilial relationships and patterns of floral evolution in the eudicot order Fabales. *Cladistics* 28: 393–421.

Bruneau A, Forest F, Herendeen PS, Klitgaard BB, Lewis GP. 2001. Phylogenetic relationships in the Caesalpinioideae (Leguminosae) as inferred from chloroplast *trnL* intron sequences. *Systematic Botany* 26: 487–514.

Bruneau A, Mercure M, Lewis GP, Herendeen PS. 2008. Phylogenetic patterns and diversification in the caesalpinioid legumes. *Botany* 86: 697–718.

Cardoso D, de Lima HC, Rodrigues RS, de Queiroz LP, Pennington RT, Lavin M. 2012a. The realignment of *Acosmium* sensu stricto with the Dalbergioid clade (Leguminosae: Papilionoideae) reveals a proneness for independent evolution of radial floral symmetry among early-branching papilionoid legumes. *Taxon* 61: 1057–1073.

Cardoso D, de Queiroz LP, Pennington RT, de Lima HC, Fonty É, Wojciechowski MF, Lavin M. 2012b. Revisiting the phylogeny of papilionoid legumes: New insights from comprehensively

sampled early-branching lineages. *American Journal of Botany* 99: 1991–2013.

Cavender-Bares J, Kozak KH, Fine PVA, Kembel SW. 2009. The merging of community ecology and phylogenetic biology. *Ecology Letters* 12: 693–715.

Chin S-W, Shaw J, Haberle R, Wen J, Potter D. 2014. Diversification of almonds, peaches, plums and cherries – Molecular systematics and biogeographic history of *Prunus* (Rosaceae). *Molecular Phylogenetics and Evolution* 76: 34–48.

Clement WL, Weiblen GD. 2009. Morphological evolution in the mulberry family (Moraceae). *Systematic Botany* 34: 530–552.

Crayn D, Fernando E, Gadek P, Quinn C. 1995. A reassessment of the familial affinity of the Mexican genus *Recchia* Moçiño & Sessé ex DC. *Brittonia* 47: 397–402.

Croat TB. 1978. *Flora of Barro Colorado Island*. Palo Alto: Stanford University Press.

Doyle JJ, Chappill JA, Bailey DC, Kajita T. 2000. Towards a comprehensive phylogeny of legumes: Evidence from *rbcL* sequences and non-molecular data. In: Herendeen PS, Bruneau A eds. *Advances in legume systematics*. Richmond: Royal Botanic Gardens, Kew. 1–20.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.

Doyle JJ, Doyle JL, Ballenger JA, Dickson EE, Kajita T, Ohashi H. 1997. A phylogeny of the chloroplast gene *rbcL* in the Leguminosae: Taxonomic correlations and insights into the evolution of nodulation. *American Journal of Botany* 84: 541–554.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Elias T. 1980. *The complete trees of North America: Field guide and natural history*. New York: van Nostrand Reinhold.

Fang J, Shen Z, Tang Z, Wang X, Wang Z, Feng J, Liu Y, Qiao X, Wu X, Zheng C. 2012. Forest community survey and the structural characteristics of forests in China. *Ecography* 35: 1059–1071.

Ferguson BJ, Indrasumunar A, Hayashi S, Lin M-H, Lin Y-H, Reid DE, Gresshoff PM. 2010. Molecular analysis of legume nodule development and autoregulation. *Journal of Integrative Plant Biology* 52: 61–76.

Forest F, Chase MW, Persson C, Crane PR, Hawkins JA. 2007. The role of biotic and abiotic factors in evolution of ant dispersal in the milkwort family (Polygalaceae). *Evolution* 61: 1675–1694.

Forest F, Savolainen V, Chase MW, Lupia R, Bruneau A, Crane PR. 2005. Teasing apart molecular- versus fossil-based error estimates when dating phylogenetic trees: A case study in the birch family (Betulaceae). *Systematic Botany* 30: 118–133.

Gregory PA, Roberta EM, Raul T, Christopher BA, Felipe S, Loreli C-J, Paola M. 2014. Amazonian functional diversity from forest canopy chemical assembly. *Proceedings of the National Academy of Sciences USA* 111: 5604–5609.

Grimm GW, Renner SS. 2013. Harvesting Betulaceae sequences from GenBank to generate a new chronogram for the family. *Botanical Journal of the Linnean Society* 172: 465–477.

Hall TA. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41: 95–98.

Herbert J, Chase MW, Möller M, Abbott RJ. 2006. Nuclear and plastid DNA sequences confirm the placement of the enigmatic *Canacomyrica monticola* in Myricaceae. *Taxon* 55: 349–357.

Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R. 2003. Angiosperm phylogeny based on *matK* sequence information. *American Journal of Botany* 90: 1758–1766.

Huang J, Chen B, Liu C, Lai J, Zhang J, Ma K. 2012. Identifying hotspots of endemic woody seed plant diversity in China. *Diversity and Distributions* 18: 673–688.

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences USA* 104: 19369–19374.

Kraft NJB, Cornwell WK, Webb CO, Ackerly DD. 2007. Trait evolution, community assembly, and the phylogenetic structure of ecological communities. *The American Naturalist* 170: 271–283.

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, de Pamphilis CW. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. *Molecular Biology and Evolution* 22: 1948–1963.

Li H-L, Wang W, Lin L, Zhu XY, Li JH, Zhu XY, Chen ZD. 2013. Diversification of the phaseoloid legumes: Effects of climate change, range expansion and habit shift. *Frontiers in Plant Science* 4: 386.

Li H-L, Wang W, Mortimer PE, Li R-Q, Li D-Z, Hyde KD, Xu J-C, Soltis D-E, Chen Z-D. 2015. Large-scale phylogenetic analyses reveal multiple gains of actinorhizal nitrogen-fixing symbioses in angiosperms associated with climate change. *Scientific Reports* 5: 14023.

Li RQ, Chen ZD, Lu AM, Soltis DE, Soltis PS, Manos PS. 2004. Phylogenetic relationships in Fagales based on DNA sequences from three genomes. *International Journal of Plant Sciences* 165: 311–324.

Lopez AJ, Little EL, Ritz GF, Rombold JS, Hahn WJ. 1987. *Arboles comunes del Paraguay: Ñande yvyra mata kuera*. Washington: Cuerpo de Paz. (in Spanish)

Lu LM, Sun M, Zhang JB, Li HL, Lin L, Yang T, Chen M, Chen ZD. 2014. Tree of life and its applications. *Biodiversity Science* 22: 3–20.

Ma KP. 2013. A mini review on the advancement of biodiversity research in China in 2012. *Biodiversity Science* 21: 1–2.

Manzanilla V, Bruneau A. 2012. Phylogeny reconstruction in the Caesalpinieae grade (Leguminosae) based on duplicated copies of the sucrose synthase gene and plastid markers. *Molecular Phylogenetics and Evolution* 65: 149–162.

Miller MA, Holder MT, Vos R, Midford PE, Liebowitz T, Chan L, Hoover P, Warnow T. 2010. The CIPRES Portals [online]. Available from www.phylo.org [accessed 5 January 2015].

Moore MJ, Hassan N, Gitzendanner MA, Bruenn RA, Croley M, Vandeventer A, Horn JW, Dhingra A, Brockington SF, Latvis M, Ramdial J, Alexandre R, Piedrahita A, Xi Z, Davis CC, Soltis PS, Soltis DE. 2011. Phylogenetic analysis of the plastid inverted repeat for 244 species: Insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. *International Journal of Plant Sciences* 172: 541–558.

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences USA* 107: 4623–4628.

Persson C. 2001. Phylogenetic relationships in Polygalaceae based on plastid DNA sequences from the *trnL-F* region. *Taxon* 50: 763–779.

Potter D. 2003. Molecular phylogenetic studies in Rosaceae. In: Sharma AK, Sharma A eds. *Plant genome: Biodiversity and evolution, Vol. 1, Part A: Phanerogams*. Enfield: Science Publishers. 319–351.

Potter D, Eriksson T, Evans RC, Oh S, Smedmark JEE, Morgan DR, Kerr M, Robertson KR, Arsenault M, Dickinson TA, Campbell CS. 2007. Phylogeny and classification of Rosaceae. *Plant Systematics and Evolution* 266: 5–43.

Qiu YL, Li L, Wang B, Xue JY, Hendry TA, Li RQ, Brown JW, Liu Y, Hudson GT, Chen ZD. 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *Journal of Systematics and Evolution* 48: 391–425.

Richardson JE, Fay MF, Cronk QCB, Bowman D, Chase MW. 2000. A phylogenetic analysis of Rhamnaceae using *rbcL* and *trnL-F* plastid DNA sequences. *American Journal of Botany* 87: 1309–1324.

Robert ER. 2015. Intrinsic dynamics of the regional community. *Ecology Letters* 18: 497–503.

Roquet C, Thuiller W, Lavergne S. 2013. Building megaphylogenies for macroecology: Taking up the challenge. *Ecography* 36: 13–26.

Santi C, Bogusz D, Franche C. 2013. Biological nitrogen fixation in non-legume plants. *Annals of Botany* 111: 743–767.

Savolainen V, Chase MW, Hoot SB, Morton CM, Soltis DE, Bayer C, Fay MF, De Bruijn AY, Sullivan S, Qiu YL. 2000a. Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. *Systematic Biology* 49: 306–362.

Savolainen V, Fay MF, Albach DC, Backlund A, van der Bank M, Cameron KM, Johnson S, Lledó MD, Pintaud JC, Powell M. 2000b. Phylogeny of the eudicots: A nearly complete familial analysis based on *rbcL* gene sequences. *Kew Bulletin* 55: 257–309.

Schaefer H, Hardy OJ, Silva L, Barraclough TG, Savolainen V. 2011. Testing Darwin's naturalization hypothesis in the Azores. *Ecology Letters* 14: 389–396.

Schaefer H, Heibl C, Renner SS. 2009. Gourds afloat: A dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proceedings of the Royal Society B: Biological Sciences* 276: 843–851.

Schaefer H, Renner SS. 2011. Phylogenetic relationships in the order Cucurbitales and a new classification of the gourd family (Cucurbitaceae). *Taxon* 60: 122–138.

Soltis DE, Bell CD, Kim S, Soltis PS. 2008. Origin and early evolution of Angiosperms. *Annals of the New York Academy of Sciences* 1133: 3–25.

Soltis DE, Gitzendanner MA, Soltis PS. 2007. A 567-taxon data set for angiosperms: The challenges posed by Bayesian analyses of large data sets. *International Journal of Plant Sciences* 168: 137–157.

Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS, Bell CD, Latvis M, Crawley S, Black C, Diouf D, Xi Z, Rushworth CA, Gitzendanner MA, Sytsma KJ, Qiu YL, Hilu KW, Davis CC, Sanderson MJ, Beaman RS, Olmstead RG, Judd WS, Donoghue MJ, Soltis PS. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704–730.

Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoot SB, Fay MF, Axtell M, Swensen SM, Prince LM, Kress WJ, Nixon KC, Farris JS. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of the Linnean Society* 133: 381–461.

Soltis DE, Soltis PS, Morgan DR, Swensen SM, Mullin BC, Dowd JM, Martin PG. 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *Proceedings of the National Academy of Sciences USA* 92: 2647–2651.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688–2690.

Stevens PF. 2001 onwards. Angiosperm Phylogeny Website. Version 12, July 2012 (and more or less continuously updated since) [online]. Available from www.mobot.org/MOBOT/research/APweb [accessed 5 January 2015].

Swensen SM, Luthi JN, Rieseberg LH. 1998. Datiscaceae revisited: Monophyly and the sequence of breeding system evolution. *Systematic Botany* 23: 157–169.

Swensen SM, Mullin BC, Chase MW. 1994. Phylogenetic affinities of Datiscaceae based on an analysis of nucleotide sequences from the plastid *rbcL* gene. *Systematic Botany* 19: 157–168.

Sytsma KJ, Morawetz J, Pires JC, Nepokroeff M, Conti E, Zjhra M, Hall JC, Chase MW. 2002. Urticalean rosids: Circumscription, rosid ancestry, and phylogenetics based on *rbcL*, *trnL–trnF*, and *ndhF* sequences. *American Journal of Botany* 89: 1531–1546.

The Legume Phylogeny Work Group. 2013. Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species-rich clades. *Taxon* 62: 217–248.

Venkateshwaran M, Volkening JD, Sussman MR, Ané J-M. 2013. Symbiosis and the social network of higher plants. *Current Opinion in Plant Biology* 16: 118–127.

Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences USA* 106: 3853–3858.

Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics* 33: 475–505.

Whitney KD, Ahern JR, Campbell LG. 2009. Hybridization-prone plant families do not generate more invasive species. *Biological Invasions* 11: 1205–1215.

Wojciechowski MF, Lavin M, Sanderson MJ. 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *American Journal of Botany* 91: 1846–1862.

Xiang X-G, Wang W, Li R-Q, Lin L, Liu Y, Zhou Z-K, Li ZY, Chen ZD. 2014. Large-scale phylogenetic analyses reveal fagalean diversification promoted by the interplay of diaspores and environments in the Paleogene. *Perspectives in Plant Ecology, Evolution and Systematics* 16: 101–110.

Zhang LB, Simmons MP, Kocyan A, Renner SS. 2006. Phylogeny of the Cucurbitales based on DNA sequences of nine loci from three genomes: Implications for morphological and sexual system evolution. *Molecular Phylogenetics and Evolution* 39: 305–322.

Zhang SD, Soltis DE, Yang Y, Li DZ, Yi TS. 2011. Multi-gene analysis provides a well-supported phylogeny of Rosales. *Molecular Phylogenetics and Evolution* 60: 21–28.

Zhu XY, Chase M, Qiu YL, Kong HZ, Dilcher D, Li JH, Chen ZD. 2007. Mitochondrial *matR* sequences help to resolve deep phylogenetic relationships in rosids. *BMC Evolutionary Biology* 7: 217.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12201/suppinfo:

Fig. S1. Summary tree resulting from maximum likelihood analysis of three genes (5564 bp; *matK*, *rbcL*, and *trnL-F*) for 1023 genera in the N-fixing clade, with tips representing families and major clades based on APG III (2009) and references therein.

Fig. S2. Majority-rule consensus of maximum likelihood trees containing 1023 genera in the N-fixing clade.

Fig. S3. Summary tree of the 303-species tree from maximum likelihood analysis of the N-fixing clade, with tips representing families and major clades based on APG III (2009) and references therein.

Fig. S4. Large-scale maximum likelihood majority-rule consensus containing 303 species of the N-fixing clade.

Fig. S5. Maximum likelihood majority-rule consensus containing 726 species that represent genera of the N-fixing clade with distributions outside China.

Fig. S6. Maximum likelihood majority-rule consensus containing 303 species randomly selected from the 726-species matrix of the N-fixing clade.

Table S1. Taxa used in this study of the N-fixing clade with GenBank accession numbers.