Taiwania 65(4): 413–422, **2020**
DOI: 10.6165/tai.2020.65.413

Lowest copy nuclear genes in disentangling plant molecular systematics

Anjan HAZRA[1], Sriparna BHOWMICK[1], Chandan SENGUPTA[2], Sauren DAS[1,*]

1. Agricultural and Ecological Research Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata – 700108, West Bengal, India
2. Department of Botany, University of Kalyani, Nadia – 741235, West Bengal, India.
[*]Corresponding author's e-mail: sauren@isical.ac.in

(Manuscript received 17 March 2020; Accepted 28 July 2020; Online published 17 August 2020)

ABSTRACT: In recent years, low copy nuclear genes became a promising choice in plant phylogeny and systematic studies for being bi-parentally inherited and highly variable, thus possessing more phylogenetically informative sites in contrast to widely used organellar genes. Here, a set of nuclear genes has been fished out from the plant genome database targeting their single copy presence in whole genomes of most of the taxa. Low copy genes, that are yet to be included in molecular phylogenetic studies of plants, were selected. All group of green plants from algae to angiosperm has been considered for validating these markers towards determining both species level and deep lineage hierarchy. The reconstructed phylogeny with selected genes, in present work, exhibited good resolution up to family level with high statistical support. Moreover, *NAD*, *PS54*, *P4H*, *CDIPT,* and *GTF* could also serve well up to higher rank clustering. Concatenated species tree through best predicted substitution model with and without third codon position corroborated the prospects of nuclear gene-based phylogeny with some incongruences in the hierarchy. The study acclaimed fourteen low copy nuclear genes concerning the determination of their efficacy toward inferring the taxonomic relationship of green plants which might be used in further molecular systematics and population genetic studies.

KEY WORDS: Angiosperm, low copy nuclear genes, molecular systematics, phylogenomics, plants.

INTRODUCTION

The plants are supposed to be first appeared sometime during 400 Million years ago, thereby slowly colonized on earth and still is in the evolutionary process. Among all plant groups, flowering plants befitted the most successful land survivor and extremely diverse with around 3.0 lakhs extant species (http://www.theplantlist.org) (Christenhusz and Byng, 2016)). The progressive relationship among the evolutionary diverge taxa is the key resource in the modern trend of plant phylogeny and classification.

The angiosperm taxonomy, which was mainly based on morphological data so far, has now been reformed by the advancement of concurrent phylogenetic evidence (Li *et al.*, 2017; Zeng *et al.*, 2012; Zeng *et al.*, 2014; Zhang *et al.*, 2017). The widely accepted Angiosperm Phylogeny Group (APG) has recently released its fourth update of the plant classification system (Chase *et al.*, 2016). According to it, Ceratophyllaceae is sister to eudicots and both of them collectively are sister to monocots. The APG classification system also hypothesized relationships among and within the major angiosperm groups. Nevertheless, the relationships among different mesangiosperm groups are yet to resolved with robust statistical support (Jansen *et al.*, 2007; Moore *et al.*, 2007; Moore *et al.*, 2010; Moore *et al.*, 2011; Qiu *et al.*, 1999; Zhang *et al.*, 2012) and regarded as a major challenge in angiosperm phylogeny (Davis *et al.*, 2014). Furthermore, during the last decades, the analyses for resolving relationships among taxa were mostly depending on organellar genes and the outcome is still inadequate (Bell *et al.*, 2010). Due to occurrence of multiple copies of rDNA, these genes are involving rigorous evolution (Letsch and Kjer, 2011) and the sequence property differs among different loci of the same genome too. Therefore, the uncertainty in the phylogenetic relationship based on organellar genes is still vibrant (Buckler *et al.*, 1997).

Low copy nuclear genes have been introduced to overcome these limitations, as well as to infer relationships among formerly unresolved lineages (Duarte *et al.*, 2010; Salas-Leiva *et al.*, 2014; Yuan *et al.*, 2009; Zeng *et al.*, 2017). They possess contrasting sequence information like high conservation in most regions across species and uniqueness too in several sites that can serve as a phylogenetically informative marker (Hazra *et al.*, 2018). Being bi-parentally inherited, nuclear genes differ with organelle genes with uniparental inheritance, and serve as efficient markers to track evolution through Mendelian inheritance even during hybridization, speciation, or sorting of closely related species of incomplete lineage too (Duarte *et al.*, 2010; Zhang *et al.*, 2012). Utilizing several non-linked nuclear single-copy genes is more worthy and may decipher incongruences of organelle gene-based phylogeny (Lu *et al.*, 2014; Zeng *et al.*, 2014; Zhang *et al.*, 2012). Routine analysis of several nuclear genes were efficiently used towards molecular evolutionary studies of fungal and animal relationships due to the availability of many genomic and EST sequence datasets (Regier *et al.*, 2010; Rokas *et al.*, 2003; Smith *et al.*, 2011; Struck *et al.*, 2011). However, the involvement of nuclear genes for deciding the plant phylogeny is still inadequate (Morton, 2011; Zeng *et al.*, 2014; Zeng *et al.*, 2017; Zhang *et al.*, 2012). The recently sequenced assemblage of whole-genome sequences is ideal resources for the identification of single/low copy genes (De Smet *et al.*, 2013; Li *et al.*, 2016). There are reports of a total 422 sequenced genome of Angiosperm (last accessed 23rd July 2020) spanning 99 families (http://plabipd.de) (figure 1). Given the above, the present study aims to analyze some of the lowest copy genes from sequenced plant genomes and their efficacy towards constructing the phylogenetic pathway of the plant kingdom. The consequential outcome of the study would recommend the status of the genes as phylogenetic markers toward developing accurate phylogeny and systematic approach in green plant lineages using nuclear gene-based polymorphism.

MATERIAL AND METHODS

Taxon sampling and data retrieval: For identification of single-copy genes and their sequence information, the sequenced plant genome databases PLAZA (https://bioinformatics.psb.ugent.be/plaza/) (Proost *et al.*, 2009) and Phytozome v12.1 (https://phytozome.jgi.doe.gov/pz/portal.html) (Goodstein *et al.*, 2011) were utilized. The PLAZA database (Updated till 2017) (Van Bel *et al.*, 2017) comprises whole genome resource of 95 species from different plant groups (47 dicot species representing 24 families, 18 monocot species under 7 families, 11 gymnosperms, 1 pteridophyte, 2 bryophytes, and 16 algal members). Comparative information regarding the orthologs and paralogs among the evolutionary distant taxa can be retrieved for further downstream analyses. The gene family finder tool of this comparative genomics platform enables us to fish out the gene families with minimum copy numbers in each plant genome of the database. From the resultant table, the first 15 loci with minimum copy number were selected for this study. Although some of the retrieved loci existed more than one copy in a highly polyploid genome (such as wheat), they still mostly were single copy in the maximum number of taxa. Moreover, according to our knowledge, a phylogeny of these particular genes are yet not compared to the corresponding species tree for their molecular systematics implications. Gene annotations of the selected loci were checked in TAIR 10 (https://www.arabidopsis.org/index.jsp) (Lamesch *et al.*, 2011). Coding DNA sequences of different taxa were retrieved individually from dicot, monocot, gymno- and pico-PLAZA by consecutive BLAST searches. Gene family identities of the respective database with a mean length of the sequences and predicted annotations are summarised in Table 1.

Sequence alignment, taxon, and sequence filtering: DNA sequences of each individual were aligned using the MUSCLE program (Edgar, 2004) with default parameters implemented in Mesquite v3.51 (Maddison and Maddison, 2019). The alignment was further curated by Phylemon 2.0 (http://phylemon.bioinfo.cipf.es) (Sánchez *et al.*, 2011) webserver interface. Aligned sequences were manually examined to eliminate the gap only regions, partial sequences, and the sequences poorly aligned among distantly related taxa. Problematic sequences like too short or unusually diverged sequences (which might be due to poor sequence quality or annotation error) were eliminated because this may infer erratic phylogenetic signal. An amino-terminal protease (CAAX) was excluded from further analysis as it showed erroneous output for the current study due to its large heterogeneity, both in sequence length and composition which hindered the alignment process. Representative plant taxa with more than 15% missing data (eliminated as poor sequences) were also excluded from the final dataset. Finally selected 14 lowest copy nuclear genes and plastid rbcL sequences of the 74 taxa were undergone individual gene tree construction (Supplementary Table 1). Subsequently, all these genes were concatenated using FaBox (Villesen, 2007) for the generation of the final tree for species.

Phylogenetic analyses: The best suitable model for evolutionary inference was estimated by Maximum Likelihood fits analysis by 24 different nucleotide substitution models (Nei and Kumar, 2000). The lowest BIC (Bayesian Information Criterion) scores for each model which are deliberating to describe substitution pattern for the best has been evaluated in this method along with the determination of AICc value (Akaike Information Criterion), the number of parameters (including branch lengths) and Maximum Likelihood value (lnL). The phylogenetic tree was inferred through the PhyML tool (Guindon and Gascuel, 2003) using Gamma distribution model (+G) with 5 rate categories and by assuming, wherever applicable, that a certain fraction of sites is evolutionarily invariable (+I). Initially, phylogenetic analysis involved all codon positions (1, 2, and 3) and thereby without the 3rd codon position. Overall, sequences from different major classes of plant kingdom viz. Angiosperm, Gymnosperm, Pteridophyte, Bryophyte, and Algae were considered for phylogenetic reconstruction. Overall mean evolutionary distances were calculated in MEGA7 (Kumar *et al.*, 2016). A 1000 round bootstrap analysis included in each study and the percentage values of the same has been used for the representation of the nodes. The final species tree with considerable topological support has been envisaged, edited, and represented using FigTree 1.4.2 (Rambaut and Drummond, 2015) software.

Comparison with existing taxonomic hierarchy: The angiosperm phylogeny group classification (Chase

Table 1. Selected lowest copy gene loci with their putative annotations and source family identities at different PLAZA database.

Name (Abbreviation)	Biological process	Cellular component	Function	DNA sequence length	Amino acid length	Dicot Plaza gene family	Monocot Plaza gene family	Gymno Plaza gene family	Pico Plaza gene family
Peptidase S54 **(PS54)**		integral component of membrane	serine-type endopeptidase activity; protein binding	1157.30 ± 322.11	381.47 ± 104.83	HOM04D006357	HOM04M006151	HOM03D006628	HOM003081
NAD Dependent Epimerase-Dehydratase **(NAD)**			coenzyme binding; catalytic activity	1310.08 ± 436.41	436.69 ± 143.43	HOM04D005951	HOM04M005938	HOM03D003670	HOM002516
RNA methyltransferase **(RNAmt)**	RNA processing		RNA methyltransferase activity; RNA binding	1539.15 ± 395.71	509.84 ± 134.66	HOM04D005926	HOM04M006008	HOM03D004268	HOM000871
acyl-CoA-N acyltransferase **(NAT)**			N-acetyltransferase activity	569.95 ± 300.79	189.98 ± 100.27	HOM04D005862	HOM04M005736	HOM03D006091	HOM000622
N-terminal methylthiotransferase **(MTTase)**	tRNA modification		4 iron, 4 sulfur cluster binding; transferase activity	1600.80 ± 400.14	533.60 ± 133.31	HOM04D005668	HOM04M005156	HOM03D005388	HOM000450
Mur Ligase **(MurE/MurF)**	regulation of cell shape; cell division; biosynthetic process	cytoplasm	acid-amino acid ligase activity; ATP binding	2030.17 ± 607.05	676.72 ± 202.35	HOM04D005814	HOM04M003604	HOM03D004822	HOM003966
Prolyl 4-hydroxylase **(P4H)**	oxidation-reduction process		L-ascorbic acid binding; iron ion binding; oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	670.00 ± 167.82	221.90 ± 56.65	HOM04D005777	HOM04M006436	HOM03D006249	HOM003059
carotenoid oxygenase **(CDO)**	oxidation-reduction process		oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen	1845.26 ± 206.84	615.08 ± 68.93	HOM04D005745	HOM04M004262	HOM03D008510	HOM000831
Mog1/Psbp **(Mog 1)**	photosynthesis	photosystem II oxygen evolving complex; extrinsic component of membrane	calcium ion binding	736.37 ± 141.64	246.28 ± 46.30	HOM04D007049	HOM04M006675	HOM03D007040	HOM010199
Amino terminal protease **(CAAX)**		membrane	protease	2462.96 ± 2256.42	820.99 ± 752.14	HOM04D002859	HOM04M006624	HOM03D006200	HOM004942
Complex C subunit B **(CCB)**				777.81 ± 110.18	259.27 ± 36.73	HOM04D006047	HOM04M006504	HOM03D007284	HOM002872
Plastidial Cellular Component **(PCC)**	positive regulation of hydrogen peroxide biosynthetic process; defense response to oomycetes	plastid; chloroplast		671.50 ± 167.98	224.68 ± 55.34	HOM04D007059	HOM04M006703	HOM03D006508	HOM002825
Checkpoint protein RAD 17-24 **(Rad17/24)**	cell cycle;DNA repair	nucleus		1633.41 ± 548.95	545.37 ± 187.59	HOM04D005902	HOM04M005992	HOM03D006532	HOM003282
CDP-alcohol phosphatidyltransferase **(CDIPT)**	phospholipid biosynthetic process	membrane	phosphotransferase activity, for other substituted phosphate groups	983.24 ± 198.95	306.14 ± 90.25	HOM04D006360	HOM04M006224	HOM03D006774	HOM003242
Glycosyl transferase **(GTF)**		endoplasmic reticulum membrane	transferase activity, transferring hexosyl groups	1456.30 ± 298.84	423.62 ± 160.42	HOM04D006154	HOM04M006679	HOM03D004088	HOM003755

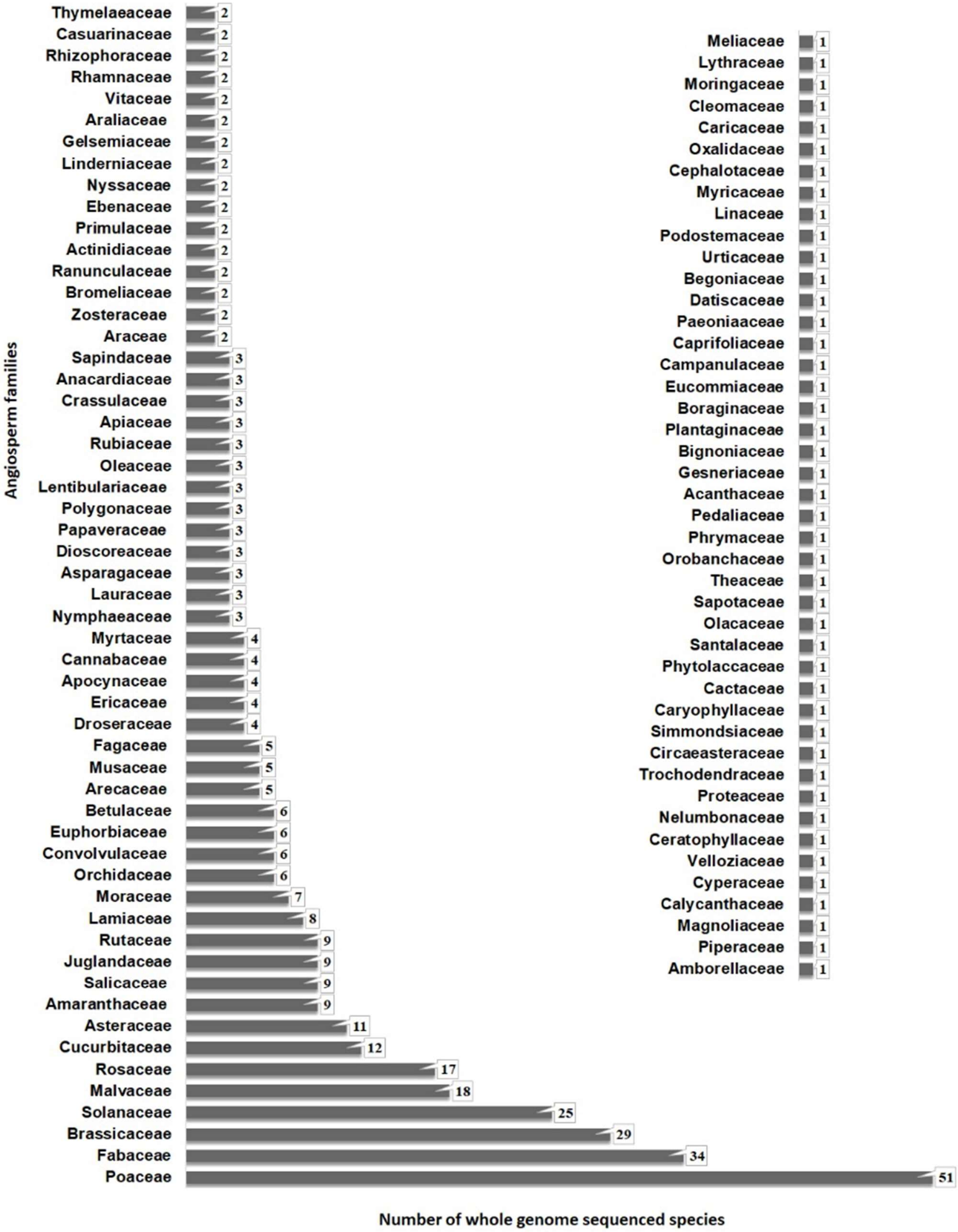

Fig. 1. Family wises statistics for availability of sequenced plant genomes (information retrieved from https://plabipd.de/; last accessed 23 July 2020).

et al., 2016) described with the detailed topology of the families in their official pages APweb (www.mobot.org/MOBOT/research/APweb) and here the reference tree based on the scheme was taken into consideration. Both the single gene trees and concatenated tree were undergone manual observation to compare with the established phylogenetic classification scheme. Congruence and dissimilarities of the same with the usual taxonomic hierarchy are discussed accordingly.

RESULTS

Here, minimum copy number genes have been identified from 101 species of whole-genome datasets in PLAZA database. After the manual screening, 15 families were selected with the lowest copy number in plant groups. Most of the species overall contained single gene copies (Figure 2) except *Glycine max*, *Chenopodium quinoa*, *Malus domestica,* and *Triticum aestivum*. The mean sequence length of the families varied from 569.95 ± 300.79 to 2462.96 ± 2256.42 base pairs (Table 1). After the elimination of problematic sequences from the individual alignments of each locus, 95 taxa remained in the list, which also came across a varied number of missing data. As mentioned in the method section, taxa with more than 15% missing data were excluded to threshold the minimum sequence coverage for concatenation phylogeny. In the final dataset of 74 taxa, most of the genes had only 0-3 % missing data except *CDO* (12.2 %) and *Mog1* (13.5 %). Aligned length of genes ranged from 1332 to 5297 sites long. The concatenated sequence alignment of 74 taxa contained a total 45025 positions which subsequently after the addition of rbcL alignment of 46782 nucleotides long. A maximum overall mean distance was found in the case of *CDIPT* whereas it was lowest in *NAT* (Table 2). The maximum likelihood fit test of the concatenated dataset resulted in GTR+G+I (GTR = General Time Reversible) for all three codon positions (Supplementary Table 2) and K2+G+I (K2 = Kimura 2-parameter) for first two codon positions excluding the third one as best by comparing all 24 substitution model.

The reconstructed gene trees of all 14 finally selected gene individuals exhibited excellent low-rank resolution with strong bootstrap support (>90%) specifically below the family level (Figure 3). Moreover, *NAD*, *PS54*, *P4H*, *CDIPT,* and *GTF* could also serve well up deep to lineage level clustering. The concatenated phylogeny of green plants considering the best-suggested substitution models clearly distinguished all seed plants as monophyletic groups, gymnosperms belonging to completely separate clades than that of angiosperms, *Amborella* as sister to all mesangiospermae, the monophyly of dicots and monocots with strong bootstrap support. Reconstructed maximum likelihood tree with and without codon position differed only in the level of statistical support at each node and consisted of similar hierarchical clustering. The incongruence of taxonomic positions observed when compared to APG IV classification was mostly resulted in weak bootstrap support. Concatenated multi-gene matrix when combined with plastidial rbcL loci, the subsequent tree ensured good statistical support of each node (mostly 100%). Throughout the study, a different set of phylogenetic trees with various combinations of loci often separated Brassicales and Malvaceae with a distinct point of origin.

Table 2. Characteristics of sequence dataset used in this study. Corresponding sequences from studied genomes are aligned using MUSCLE program and overall mean distance has been calculated from the alignment through MEGA7 tool.

Loci	Missing data %	Aligned length (bp)	Overall mean distance
PS54	1.4	3785	0.522
NAD	0.0	2434	0.418
RNAmt	2.7	4312	0.526
NAT	0.0	1332	0.349
MTTase	1.4	3636	0.5
MurE/MurF	2.7	5297	0.474
P4H	2.7	2269	0.501
CDO	12.2	4773	0.623
Mog 1	13.5	2123	0.546
CCB	2.7	1728	0.553
PCC	2.7	2106	0.553
Rad17/24	6.8	5080	0.605
CDIPT	0.0	2383	0.636
GTF	1.4	3767	0.556

DISCUSSION

Decoding accurate phylogenetic history of land plants integrates the understanding of the colonization, evolution, speciation, and diversification of plants on earth. Recent advancement of sequencing technology and the availability of a large set of potential molecular markers undoubtedly assists the relationship inferring process to a great extent. However, phylogenetic clustering among various plant lineages at different hierarchy levels remained contradictory and debatable (Li *et al.*, 2017). Therefore, generating more molecular datasets and utilization of the available ones in global analysis becomes the prime point of interest. Babineau *et al.* (2013) suggested that assessment of the phylogenetic information and level of taxonomic resolution of the existing LCNG sequences involving maximum possible taxonomic groups will be a cheaper and time-saving process instead of the extensive trial and error process in sequencing of the samples. In recent, the available plant genome resources have adequate information for conducting plant phylogeny and systematic study through a comparative genomic approach. According to the latest record, 99 angiosperm families represent with at least one genome sequenced member, among which Poaceae is the leading with the maximum number of sequenced genome plants (51)

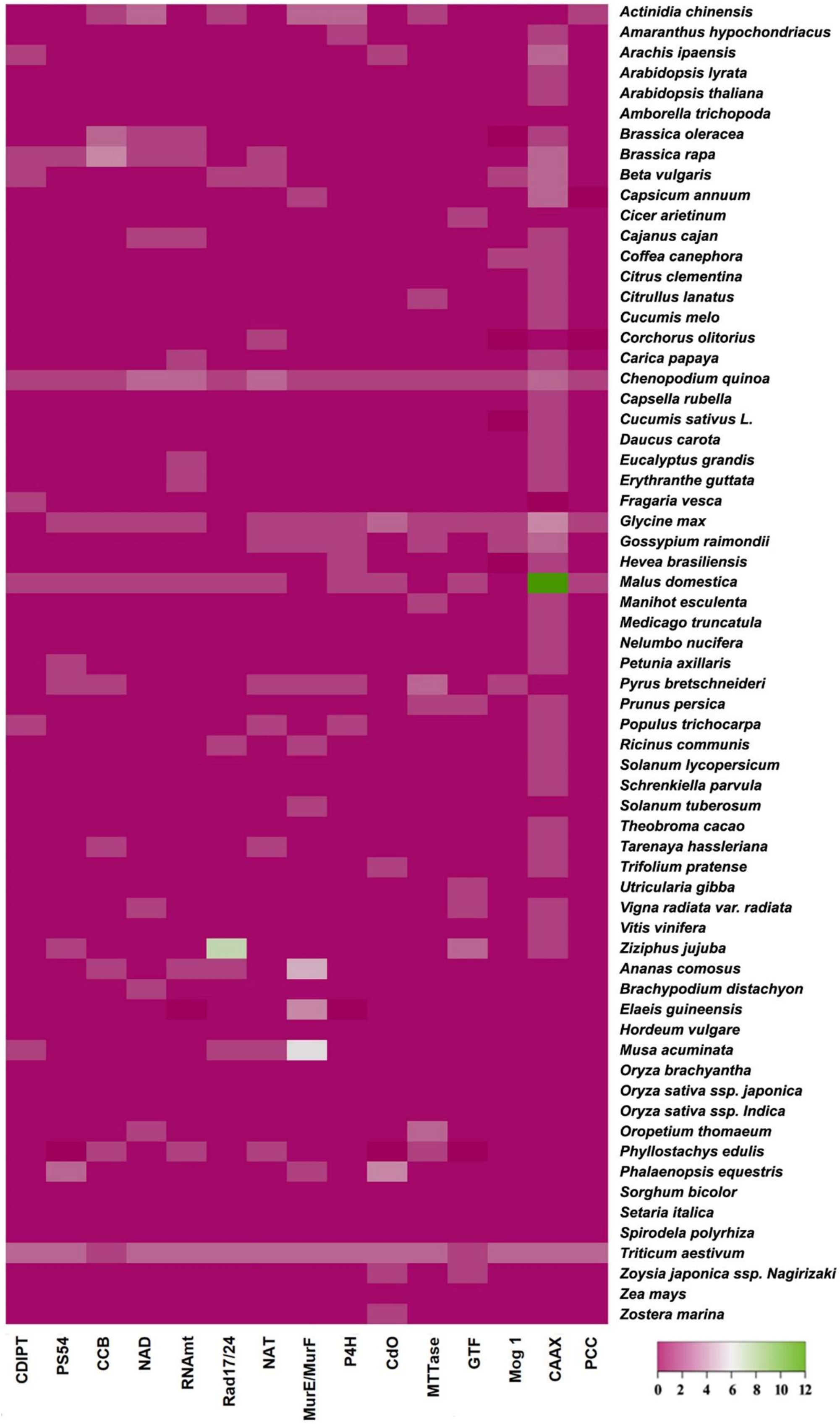

Fig. 2. Distribution of copy number of selected gene families in studied plant genomes.

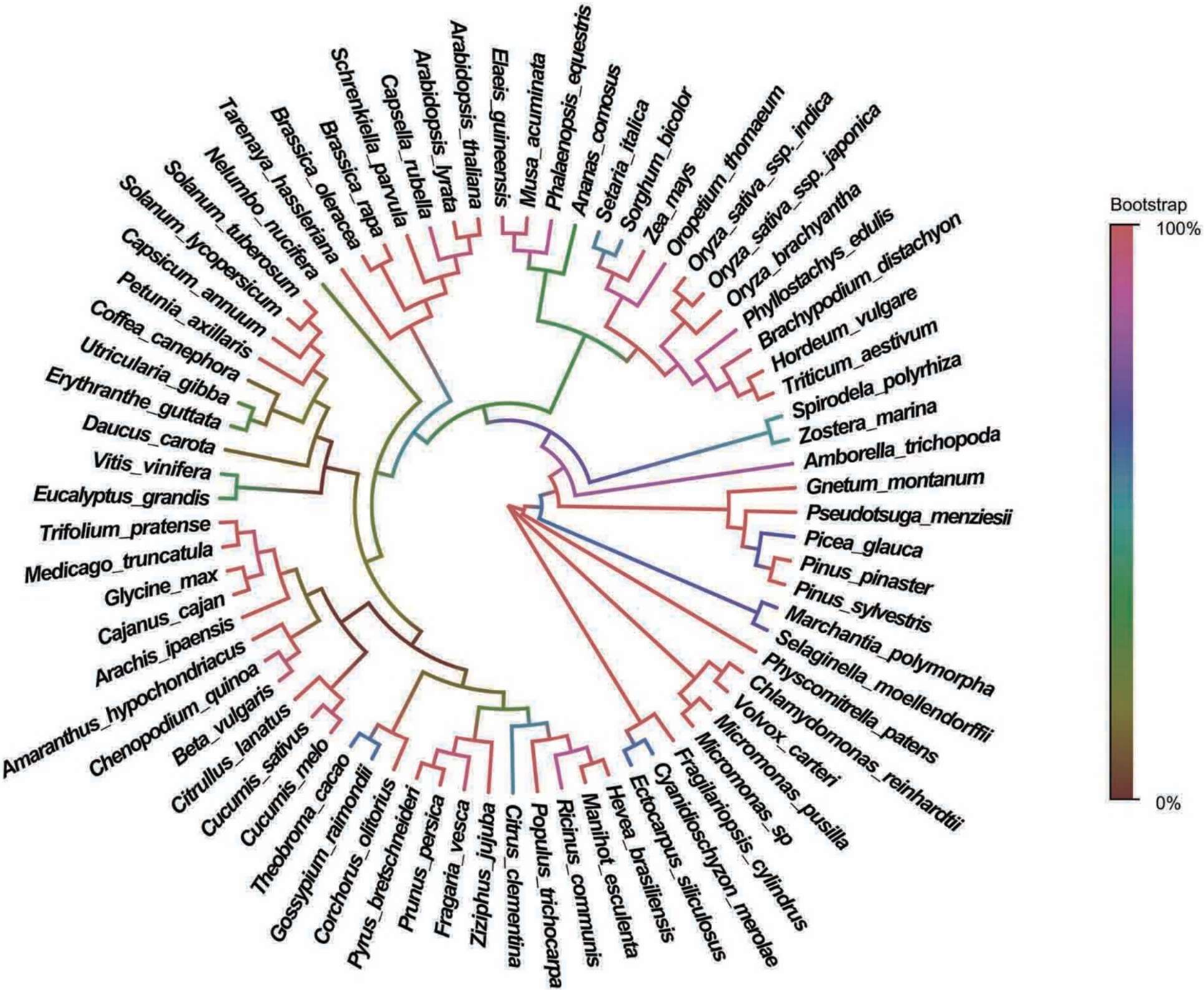

Fig. 3. Maximum likelihood tree inferred from the concatenated matrix of 14 low copy nuclear genes. Phylogenetic inference was drawn using Gamma distribution model (+G) with 5 rate categories and by assuming certain fraction of sites are evolutionarily invariable (+I), wherever applicable. Branch colours have been represented by percentages of corresponding bootstrap support among 1000 replicates.

followed by Fabaceae (34), Brassicaceae (29), Solanaceae (25), and others (figure 1). However, only a percentage of this included in multi-utility comparative genomics platforms like Phytozome, PLAZA, etc. The orthology information tool in the current approach enabled us to fish out the actual low copy genes with a mostly single copy in an organism's genome. This is because of molecular evolution and adaptive divergence signals during gene duplication of multiple copy gene families that might hinder the phylogenetic reconstruction exclusively targeted to taxonomy and systematics (Hazra *et al.*, 2019; Hilu *et al.*, 2014).

Sequences of low copy nuclear genes (LCNG) are being used as a useful resource for reconstructing plant phylogeny and systematics during recent years (Sang, 2002; Wu *et al.*, 2006). It is even advantageous over the organellar gene for resolving the relationship between middle and low-rank taxonomic groups. Cacho and Strauss (2013) reported a total of 11 primer sets based on single-copy nuclear genes, which can be useful in improving resolution at and above the species level across the Thelypodieae. Certain low copy nuclear Conserved Ortholog Set (COS) genes found to serve a higher proportion of parsimony informative sites in comparison with traditional phylogenetic markers like ITS and matK (Li *et al.*, 2008). In general, organellar genes are much conserved, therefore unable to provide adequate phylogenetic informative sites for taxonomic lower level resolution (Knoop, 2004). In contrast, comparatively higher variability in some regions of the nuclear gene assisted to recommend it as an ideal phylogenetic marker to reveal the complex history of angiosperms evolution (Cruz-Mazo *et al.*, 2009; Lu *et al.*, 2010). Moreover, the much smaller size of the organellar genome than that of nuclear genome and its uniparental inheritance can contribute to infer a partial evolutionary history of plants only (Jansen *et al.*, 2007; Moore *et al.*, 2010; Ness *et al.*, 2011; Zhang *et al.*, 2012). Thus, molecular systematics of angiosperms critically need LCNG markers to overcome the limitations of plastid marker-based relationships at lower taxonomic levels (eg. at species level) which remained debatable hitherto. Furthermore, the unsuitability of single copy plastid genes for resolving the position of Malpighiales, Cornales, and Ericales mandates the information of nuclear gene is necessary in this regard (Zhang *et al.*, 2012). Nevertheless, in this case, the methodological requirements such as orthology identification, copy number estimation of nuclear genes are much complex and therefore very few nuclear gene markers have already been implemented in phylogenetic reconstruction of plants (Babineau *et al.*, 2013; Li *et al.*, 2017; Ness *et al.*, 2011; Zeng *et al.*, 2014; Zhang *et al.*, 2012; Zhu and Ge, 2005).

The present study explored some LCNG markers, concatenated phylogeny of that can efficiently resolve systematic relationships well up to family level. This observation supports the earlier findings regarding the utility of the LCNG on phylogenetic reconstructions particularly at low taxonomic levels (Cacho and Strauss, 2013; Li *et al.*, 2008). However, this phylogeny did not strictly follow the superorder level resolution among selected taxa. These loci often separated Brassicales and Malvales clades which are supposed to be in the Malvid clade together according to the latest organellar gene-based phylogeny (Li *et al.*, 2019). The conflicting signal of some order/superorder level clades in the nuclear gene-based phylogeny with that of previously accepted organellar gene-based phylogeny came into the limelight earlier (Hilu *et al.*, 2014; Sun *et al.*, 2015). Nonetheless, it can be assumed that the incongruence in topology at the superorder level is most possibly due to a low sample size of the sequenced genome or missing taxa from representative groups that perform a vital role for inferring deep node phylogeny. Mostly, to depict such phylogeny, wide sampling from distant groups are considered that is evident in similar approaches conducted earlier. For example, in a recent breakthrough report, the tree of green plants has been enriched with robust phylogenomic analyses from the transcriptomes of 1,124 green plants (One Thousand Plant Transcriptomes Initiative, 2019). As, sufficient genome sequences spanning each green plant family is still not available in the public domain, so maximum possible representatives with clear and reliable sequences have been considered in this study. Finally, these lowest copy bearing genes from the whole genome dataset might serve as the marker of choice to ascertain the species level of phylogenetic reconstruction of the plant system and at least five among them perform well above family level too. Identification and analyses of the homologous regions from a broad range of taxa covering the relevant evolutionary groups would lead to their practical implications more precisely.

CONCLUSION

In the present study, the combination of precisely evaluated nuclear genes from available genomic resources has been examined toward their utilization in phylogenetic reconstruction at the various rank of taxa. Finally, the study provides important clues through the evaluation of genome-scale mining of several lowest copy nuclear genes concerning the accuracy of taxonomic relationships. The recommended LCNG from plant genomes might be used in further molecular systematics and population genetic study as well as the screening methodology would be useful in disentangling ideal low copy gene markers for such study.

ACKNOWLEDGMENTS

Authors are thankful to the anonymous reviewers for their helpful suggestions toward the improvement of the earlier version of the manuscript.

LITERATURE CITED

Babineau, M., E. Gagnon, A. Bruneau, 2013. Phylogenetic utility of 19 low copy nuclear genes in closely related genera and species of caesalpinioid legumes. S. Afr. J. Bot. **89:** 94–105.

Bell, C.D., D.E. Soltis, P.S. Soltis, 2010. The age and diversification of the angiosperms re-revisited. Am. J. Bot. **97**(**8**): 1296–1303.

Buckler, E.S., A. Ippolito, T.P. Holtsford. 1997. The evolution of ribosomal DNA divergent paralogues and phylogenetic implications. Genetics **145**(**3**): 821–832.

Cacho, N.I., S.Y. Strauss. 2013. Single-copy nuclear gene primers for Streptanthus and other Brassicaceae from genomic scans, published data, and ESTs. Appl. Plant Sci. **1**(**7**): 1200002.

Chase, M.W., M. Christenhusz, M. Fay, J. Byng, W.S. Judd, D. Soltis, D. Mabberley, A. Sennikov, P.S. Soltis, and P.F. Stevens. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. **181**(**1**): 1–20.

Christenhusz, M.J. and J.W. Byng. 2016. The number of known plants species in the world and its annual increase. Phytotaxa **261**(**3**): 201–217.

Cruz-Mazo, G., M. Buide, R. Samuel, and E. Narbona. 2009. Molecular phylogeny of Scorzoneroides (Asteraceae): Evolution of heterocarpy and annual habit in unpredictable environments. Mol. Phylogenetics Evol. **53**(**3**): 835–847.

Davis, C.C., Z. Xi, and S. Mathews. 2014. Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. BMC biology **12**(**1**): 11.

De Smet, R., K.L. Adams, K. Vandepoele, M.C. Van Montagu, S. Maere, and Y. Van De Peer. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. PNAS **110**(**8**): 2898–2903.

Duarte, J. M., Wall, P. K., Edger, P. P., Landherr, L. L., Ma, H., Pires, P. K., Leebens-Mack, J. and Claude, W.

D. 2010. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. BMC Evolutionary Biology **10**(**1**): 61.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(**5**): 1792–1797.

Goodstein, D.M., S. Shu, R. Howson, R. Neupane, R.D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, and N. Putnam. 2011. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. **40**(**D1**): D1178–D1186.

Guindon, S. and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**(**5**): 696–704.

Hazra, A., N. Dasgupta, C. Sengupta, and S. Das. 2019. MIPS: Functional dynamics in evolutionary pathways of plant kingdom. Genomics **111**(**6**): 1929–1945.

Hazra, A., P. Nandy, C. Sengupta, and S. Das. 2018. MIPS sequences: a promising molecular consideration in angiosperm phylogeny and systematics. BioTechnologia **99**(**1**): 5–12.

Hilu, K.W., C.M. Black, and D. Oza. 2014. Impact of gene molecular evolution on phylogenetic reconstruction: A case study in the Rosids (superorder Rosanae, Angiosperms). PLoS One **9**(**6**): e99725.

Jansen, R.K., Z. Cai, L.A. Raubeson, H. Daniell, C.W. Depamphilis, J. Leebens-Mack, K.F. Müller, M. Guisinger-Bellian, R.C. Haberle, and A.K. Hansen. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. PNAS **104**(**49**): 19369–19374.

Knoop, V. 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr. Genet. **46**(**3**): 123–139.

Kumar, S., G. Stecher, and K. Tamura. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. **33**(**7**): 1870–1874.

Lamesch, P., T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D.L. Alexander, and M. Garcia-Hernandez. 2011. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. **40**(**D1**): D1202–D1210.

Letsch, H.O. and K.M. Kjer, 2011. Potential pitfalls of modelling ribosomal RNA data in phylogenetic tree reconstruction: evidence from case studies in the Metazoa. BMC Evolutionary Biology **11**(**1**): 146.

Li, H.-T., T.-S. Yi, L.-M. Gao, P.-F. Ma, T. Zhang, J.-B. Yang, M.A. Gitzendanner, P.W. Fritsch, J. Cai, and Y. Luo. 2019. Origin of angiosperms and the puzzle of the Jurassic gap. Nat. Plants **5**(**5**): 461.

Li, M., J Wunder,. G. Bissoli, E. Scarponi, S. Gazzani, E. Barbaro, H. Saedler, and C. Varotto. 2008. Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. Cladistics **24**(**5**): 727–745.

Li, Z., A.R. De La Torre, L. Sterck, F.M. Cánovas, C. Avila, I. Merino, J.A. Cabezas, M.T. Cervera, P.K. Ingvarsson, and Y. Van De Peer. 2017. Single-copy genes as molecular markers for phylogenomic studies in seed plants. Genome Biol. Evol. **9**(**5**): 1130–1147.

Li, Z., J. Defoort, S. Tasdighian, S. Maere, Y. Van De Peer, and R. De Smet. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. The Plant Cell **28**(**2**): 326–344.

Lu, L., P.W. Fritsch, B.C. Cruz, H. Wang, and D.-Z. Li. 2010. Reticulate evolution, cryptic species, and character convergence in the core East Asian clade of Gaultheria (Ericaceae). Mol. Phylogenet. Evol. **57**(**1**): 364–379.

Lu, Y., J.-H. Ran, D.-M. Guo, Z.-Y. Yang, and X.-Q. Wang. 2014. Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. PLoS One **9**(**9**): e107679.

Maddison, W.P. and D.R. Maddison. 2019. Mesquite: a modular system for evolutionary analysis. Version 3.51 http://www.mesquiteproject.org

Moore, M.J., C.D. Bell, P.S. Soltis, and D.E. Soltis. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. PNAS **104**(**49**): 19363–19368.

Moore, M.J., N. Hassan, M.A. Gitzendanner, R.A. Bruenn, M. Croley, A. Vandeventer, J.W. Horn, A. Dhingra, S.F. Brockington, and M. Latvis. 2011. Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. Int. J. Plant Sci. **172**(**4**): 541–558.

Moore, M.J., P.S. Soltis, C.D. Bell, J.G. Burleigh, and D.E. Soltis. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. PNAS **107**(**10**): 4623–4628.

Morton, C.M. 2011. Newly sequenced nuclear gene (Xdh) for inferring angiosperm phylogeny1. Ann. Mo. Bot. Gard. **98**(**1**): 63–90.

Nei, M. and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford university press.

Ness, R.W., S.W. Graham, and S.C. Barrett. 2011. Reconciling gene and genome duplication events: using multiple nuclear gene families to infer the phylogeny of the aquatic plant family Pontederiaceae. Mol. Biol. Evol. **28**(**11**): 3009–3018.

One Thousand Plant Transcriptomes Initiative 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature **574**(**7780**): 679.

Proost, S., M. Van Bel, L. Sterck, K. Billiau, T. Van Parys, Y. Van De Peer, and K. Vandepoele. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. The Plant Cell **21**(**12**): 3718–3731.

Qiu, Y.-L., J. Lee, F. Bernasconi-Quadroni, D.E. Soltis, P.S. Soltis, M. Zanis, E.A. Zimme, Z. Chen, V. Savolainen, and M.W. Chase. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature **402**(**6760**): 404–407.

Rambaut, A. and A. Drummond. 2015. FigTree, ver. 1.4. 2. Available: http:/tree.bio.ed.ac.uk/software/figtree/.

Regier, J.C., J.W. Shultz, A. Zwick, A. Hussey, B. Ball, R. Wetzer, J.W. Martin, and C.W. Cunningham. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature **463**(**7284**): 1079–1083.

Rokas, A., B.L. Williams, N. King, and S.B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature **425**(**6960**): 798–804.

Salas-Leiva, D.E., A.W. Meerow, J. Francisco-Ortega, M. Calonje, M.P. Griffith, D.W. Stevenson, and K. Nakamura. 2014. Conserved genetic regions across angiosperms as tools to develop single-copy nuclear markers in gymnosperms: an example using cycads. Mol. Ecol. Resour. **14(4)**: 831–845.

Sánchez, R., F. Serra, J. Tárraga, I. Medina, J. Carbonell, L. Pulido, A. De María, S. Capella-Gutíerrez, J. Huerta-Cepas, and T. Gabaldón. 2011. Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. Nucleic Acids Res. **39(suppl_2)**: W470–W474.

Sang, T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. **37(3)**: 121–147.

Smith, S.A., N.G. Wilson, F.E. Goetz, C. Feehery, S.C. Andrade, G.W. Rouse, G. Giribet, and C.W. Dunn. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature **480(7377)**: 364–368.

Struck, T.H., C. Paul, N. Hill, S. Hartmann, C. Hösel, M. Kube, B. Lieb, A. Meyer, R. Tiedemann, and G. Purschke. 2011. Phylogenomic analyses unravel annelid evolution. Nature **471(7336)**: 95–98.

Sun, M., D.E. Soltis, P.S. Soltis, X. Zhu, J.G. Burleigh, and Z. Chen. 2015. Deep phylogenetic incongruence in the angiosperm clade Rosidae. Mol. Phylogenet. Evol. **83:** 156–166.

Van Bel, M., T. Diels, E. Vancaester, L. Kreft, A. Botzki, Y. Van De Peer, F. Coppens, and K. Vandepoele. 2017. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. **46(D1)**: D1190–D1196.

Villesen, P. 2007. FaBox: an online toolbox for fasta sequences. Mol. Ecol. Notes **7(6)**: 965–968.

Wu, F., L. A. Mueller, D. Crouzillat, V. Pétiard, and S.D. Tanksley. 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(3)**: 1407–1420.

Yuan, Y.W., C. Liu, H.E. Marx, and R.G. Olmstead. 2009. The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies. New Phytol. **182(1)**: 272–283.

Zeng, L., N. Zhang, Q. Zhang, P.K. Endress, J. Huang, and H. Ma. 2017. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. New Phytol. **214(3)**: 1338–1354.

Zeng, L., Q. Zhang, R. Sun, H. Kong, N. Zhang, and H. Ma. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat. Commun. **5(1)**: 1–12.

Zhang, N., L. Zeng, H. Shan, and H. Ma. 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. **195(4)**: 923–937.

Zhu, Q. and S. Ge. 2005. Phylogenetic relationships among A-genome species of the genus Oryza revealed by intron sequences of four nuclear genes. New Phytol. **167(1)**: 249–265.

Supplementary materials are available from Journal Website.