Identification of *Dioscorea* L. Species Based on Complete Chloroplast Genome

Xinlian Chen , Hui Yao , Shuangjiao Ma , Ying Li , Jianguo Zhou , Jingyuan Song , Shilin Chen
Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Abstract: Previous studies of DNA barcoding of *Dioscorea* species have shown that the identification efficiency was low. In this study, the complete chloroplast (CP) genomes of *D. opposite* and *D. collettii* were sequenced using Illumina HiSeq 2000. The complete CP genome lengths of *D. opposite* and *D. collettii* were 152,963 bp and 153,870 bp, respectively, which include four parts: two inverted repeats (IRs), one large single copy (LSC) and one small single copy (SSC). Both of the two CP genomes encoded 125 genes, including 87 protein coding genes, 30 tRNA genes and 8 rRNA genes. The GC contents of *D. opposite* and *D. collettii* were 37.04% and 37.17%, respectively. The result showed that the sequence variation of non-coding regions is higher than that of the protein coding region and the variations of in LSC and SSC are higher than that of the IRs. 10 molecular markers of the genus, which include 5 protein coding regions and 5 non-coding regions, were screened to be used as the potential DNA barcodes for authenticating *Dioscorea* species. The ML tree constructed by the complete CP genomes showed that the CP genome can be used as an ultra-barcode to distinguish this genus from each other. This study laid the foundation of super barcode utilization in this genus and provided us a molecular base for next investigation on this important medicinal species.

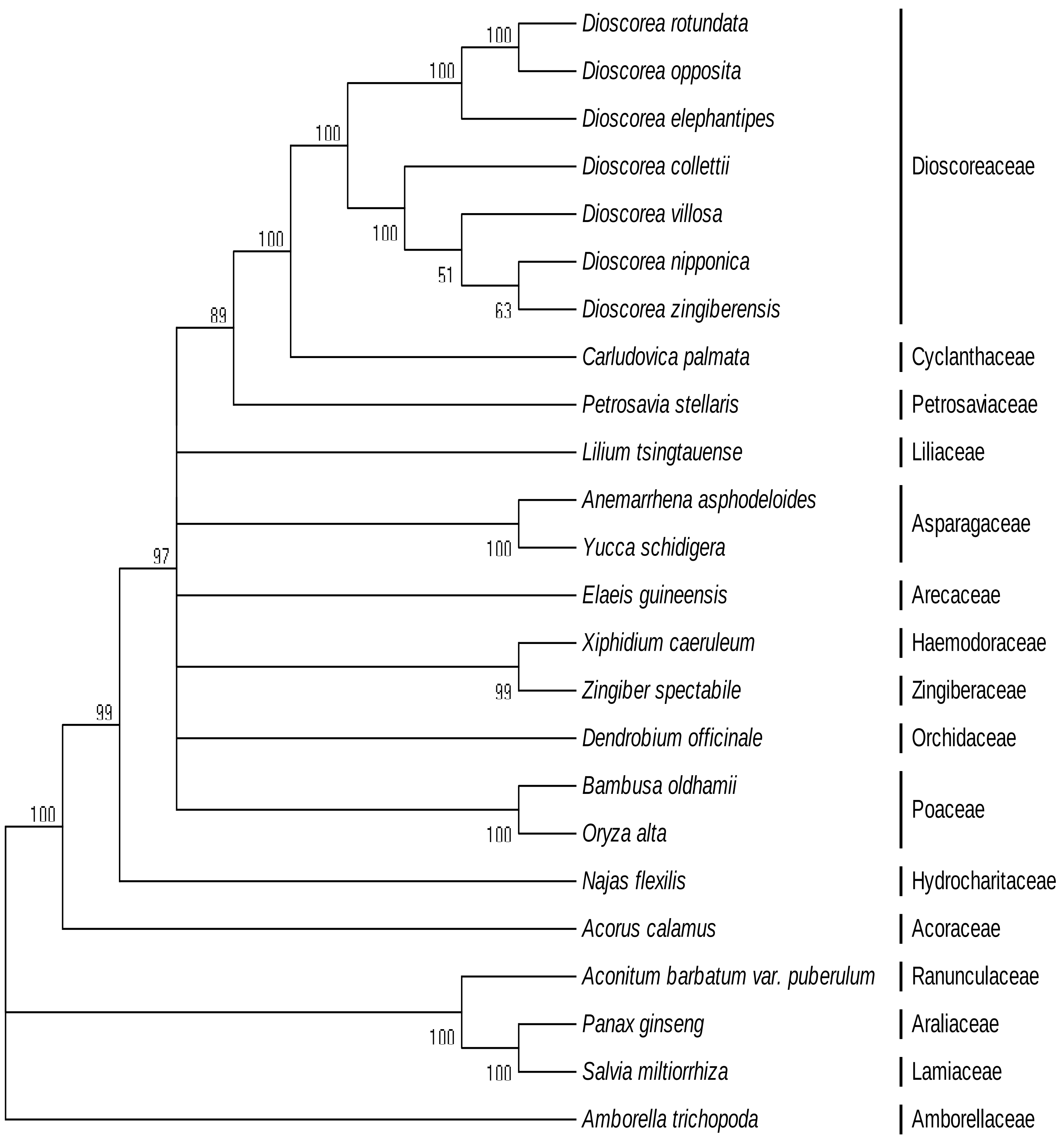
Background: *Dioscorea* L. has a long history in China as medicinal and edible plants, which have important economic values. However, the identification problem of *Dioscorea* L. is still unsolved because of the limitations of traditional methods. *Flora of China* records 52 plants, and some varieties of species are too complex to distinguish from morphology. Previous research showed that the universal DNA barcode sequences cannot identify *Dioscorea* L. effectively. Solving the identification problem of *Dioscorea* L. is the vital research content on the post-barcode era. In this study, the complete chloroplast (CP) genomes of *D. opposite* and *D. collettii* were sequenced using Illumina Hiseq X in order to screen specific DNA regions or explore the complete CP genome as a super barcode to identify this genus.

Results: The complete CP genome size of *D. opposite* is 152,960 bp and that of *D. collettii* is 153,869 bp. A pair of inverted repeats (IRs) of 50,986bp is separated by a large single-copy region (LSC, 83,152 bp) and a small single-copy region (SSC, 18,822 bp) in *D. opposite*. Moreover, a pair of IRs with a length of 51,182 bp is separated by LSC (83,824 bp) and SSC (18,863bp) in *D. collettii*. region (SSC, 18,822 bp) in *D. opposite*. Moreover, a pair of IRs with a length of 51,182 bp is separated by LSC (83,824 bp) and SSC (18,863 bp) in *D. collettii*.

Sequence variation of candidate sequences

Candidate sequence	sequence variation		
	Alignment length (bp)	Number of variation sites	Proportion of variation sites
rps19	330	29	8.79%
ndhF	2333	129	5.53%
ycf1	6435	2201	34.2%
rpl32	174	13	7.47%
rpl20	372	26	6.99%
matK-trnQ(TTG)	3521	376	10.68%
trnE(TTC)-trnT(GGT)	934	65	6.96%
atpI-rsp2	269	18	6.69%
trnT(TGT)-trnF(GAA)	1818	124	6.83%
ycf1-rps15	413	140	33.90%

Both species contain eight rRNAs and 30 tRNAs, whereas *D. opposite* has 89 protein-coding genes and that of *D. collettii* is 88. The specific DNA regions with high variation were screened as the potential DNA barcodes to identify *Dioscorea* species based on the complete CP genomes of *D. opposite* and *D. collettii* that are combined with four other complete CP genomes of *Dioscorea* L., *D. elephantipes* (EF380353), *D. rotundata* (KJ490011), *D. nipponica* (KP404629), and *D. zingiberensis* (KP899622).



The ML tree showed that the family Dioscoreaceae was sister taxa with respect to the family Cyclanthaceae, and these species were grouped with the family Petrosaviaceae. And the seven *Dioscorea* species could be distinguished from each other. This result Illustrated that complete CP genome sequence can be used as a super barcode to identify *Dioscorea* species.

References

Huang C H, Ku C Y, Jan T R. Diosgenin attenuates allergen-induced intestinal inflammation and IgE production in murine model of food allergy. Planta Med, 2009, 75(12): 1300.

Gao X, Zhu Y P, Wu B C, et al. Phylogeny of *Dioscorea* sect. Stenophora based on chloroplast *matK, rbcL* and *trnL-F* sequences. J Syst Evol, 2008, 46(3): 315-321.

Chen S L, Pang X H, Song J Y, et al. A renaissance in herbal medicine identification: From morphology to DNA. Biotechnol Adv, 2014, 32(15): 1237–1244.

Yao H, Song J Y, Liu C, et al. Use of ITS2 region as the universal DNA barcode for plants and animals. PloS ONE, 2010, 5(10): e13102.

Sun X Q, Zhu Y J, Guo J L, et al. DNA barcoding the *Dioscorea* in China, a vital group in the evolution of Monocotyledon: Use of *matK* gene for species discrimination. PloS ONE, 2012, 7(2): e32057.

Wu F H, Chan M T, Liao D C, et al. Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biol, 2010, 10(1): 68.

Funding Sources/Acknowledgements

Chinese Academy of Medical Sciences (CAMS) InnovationFund for Medical Sciences (CIFMS) (No. 2016-I2M-3-016) and The Key Projects in the National Science and Technology Pillar Program (No. 2011BAI07B08).