e-ISSN:2320-0189
p-ISSN:2347-2308

Research & Reviews: Journal of Botanical Sciences

The Complete Chloroplast Genomes of *Asteraceae Species*

Ying Zhang[1], Wei Guan[2], Xiaonan Zhang[1] and Lei Li[1]*

[1]Key Laboratory of Tropical Animal and Plant Ecology of Hainan Province, College of Life Science Hainan Normal University, Hainan, Haikou 571158, PR China

[2]Research Institutes of Tropical Forest, CAF, Guangzhou 510520, PR China

Research Article

Received date: 21/12/2015

Accepted date: 15/02/2016

Published date: 19/02/2016

For Correspondence: Lei Li, College of Life Science Hainan Normal University, Haikou 571158, Hainan, China, Tel : 86-898-65883521

E- mail: lei-li @126.com

Keywords: Chloroplast Genome, Asteraceae, Pseudogenes, DNA barcodes, Phylogenetic tree.

ABSTRACT

Until now, twenty-seven Asteraceae complete chloroplast genomes were uncovered in the Gene bank. The highly conservative nature and slow evolutionary rate of the chloroplast genome demonstrated that it was uniform enough to perform comparative studies across different species but divergent sufficiently to capture evolutionary events, which makes it a suitable and invaluable tool or molecular phylogeny and molecular ecology studies. The researches about the size, genome content, LSC, SSC, IR- LSC/SSC borders, pseudogenes and DNA barcodes of these twenty-seven complete chloroplast genomes of Asteraceae were reviewed here. Based on the above information, the complete chloroplast genome of each species provides a more accurate relationship in Asteraceae and can be used as a more suitable marker for species identification.

INTRODUCTION

The family Asteraceae is a complex species belonging to the second largest family of plants in the world and consisting of 2,400 species distributed in 170 genera [1]. With the exception of Antarctica, the Asteraceae are distributed on all continents. The extremely various expressions in secondary chemistry, inflorescence morphology and chromosome number were found in the research of Asteraceae plants [2]. Furthermore, this family includes members of economically important food crops, herbal species, ornamentals for the cut-flower industry, weedy with the economic and ecological impact and some invasive species [3-7].

Chloroplasts (cp), which originate from ancient eubacteria invasions [8], are multifunctional organelles possessing their own genetic material. As the essential organelle in plant cell, it conducts photosynthesis in the presence of sunlight. The highly conservative nature and slow evolutionary rate of the chloroplast genome demonstrated that it was uniform enough to perform comparative studies across different species but divergent sufficiently to capture evolutionary events, which makes it a suitable and invaluable tool or molecular phylogeny and molecular ecology studies [7].

Since the publication of the first cp genome, the number of complete cp genomes available (http:// www.ncbi.nlm.nih.gov/genome) has increased rapidly thanks to the development of high-throughput technologies [3,6,9,10]. Today, there are 792 complete cp genome that were deposited in the Genebank organelle Genome Resource, while were 329 in 2014 and about 200 in 2011 [6,7]. In the meantime, from the 2012, the first complete cp genome of *Lactuca sativa* belonging to the family Asteraceae was published, until now, 26 other Asteraceae plants are reported in the Genebank. Among them, 12 subfamilies were found. *Cynara naetica*, *Cynara cardunculus*, *Cynara cornigera* and *Cynara humilis* [11] belonging to *Carduoideae*; *Leontopodium leiolepis* belonging to *Leontopodium*; *Parthenium argentatum* belonging to *Parthenium*; *Silybum marianum* belonging to *Silybum* subfamily; *Artemisia frigida* and *Artemisia montana* belonging to *Artemisia*; *Aster spathulifolius* and *Jacobaea vulgaris* belonging to *Aster* [2-5,11-13]. *Centaurea diffusa* is *Centaurea* species; *Chrysanthemum indicum* and *Chrysanthemum x morifolium* are *Chysanthemum* species; *Guizotia abyssinica* is *Guizotia* plant; *Heloanthus* subfamilies have 8 species were found with the whole cp genome sequences: *Helianthus annuus*, *Helianthus decaetalus*, *Helianthus divaricatus*, *Helianthus grosseserratus*, *Helianthus hirsutus*, *Helianthus masimiliani*, *Helianthus strumosus and Helianthus tuberisus* [5,14,15]. *Praxelis clematidea* [6] and *Ageratinn adenophora* are *Eupatorium* subfamily [6,7]. In this article, we describe the size, genome content, LSC, SSC, IR-LSC/SSC borders, Pseudogenes

and DNA barcodes of the Asteraceae cp genomes. Based on the above information, the complete chloroplast genome of each species provides a more accurate relationship in Asteraceae and can be used as a more suitable marker for species identification.

Size and Genome Content

From the information of all sequenced cp genomes, most to them range from 120 to 160 kb in length and have GC contents of 30 to 40% [3,6]. The cp genomes of Asteraceae species are from 149.51 bp (As. *spathulifolius*) to 153.202 bp *(S. marianum)* and differ slightly in length **(Table 1).** These are the larger cp genomes of *Asteraceae* compared with other plants. Multiple complete *Asteraceae* cp genomes available provide an opportunity to compare the sequence variation within the family at the genome-level. The sequence identity of all the twenty-seven *Asteraceae* cp genome was plotted using VISTA program with the annotation of *A. adenophora* as reference **(Figure 1A-1G)**, percent identity plot as summarized in **(Table S1).** The genomes comprise more than eighty protein- coding genes from 83 (Ch. indicum) to 90 (*C. diffusa*) except one species: *P. argentatum*, it's cp genome only contains 55 proteins-coding genes annotation in NCBI, but the number is 85 in Kumar's paper [12]. The number of rRNA is from seven to nine. Four genes: rrn 23, rrn 16, rrn 5 and rrn 4.5 are double for locating in the two copies of inverted repeats (IRs) can be found in majority species [6,11]. The differences are the disappear of rrn 5 in *L. sativa* and the join of rps19 in the rRNA in Helianthus subfamily except *H. annus*. The number of gene is from 106 (*P. argentatum*) to 138 (*H. annuus*) [5,12]. For the tRNA, there is also the least 17 in *P. argentatum* and the maximum 43 in *H. annuus* **(Table 1).** The whole aligned sequences indicate that the Asteraceae cp genomes are rather conservative, although some divergent regions are found between these genomes. Similar to other angiosperms, the coding region is more conservative than the non-coding counterpart. Of all genes, ycf1, ycf68 and rps19 gene is the most divergent [3,7]. rpoC1 gene contains two introns same with *A. adenophora* also shows high sequence divergence [7]. Furthermore, a number of regions are found to show high divergence, including trnk-psbK, aptL-aptF, trnS-trnG, ndhC-trnM, psbL-petG rpl14-rpl16, and accD-psaI [6] **(Table S1).**

Table 1. Size and genes of 27 Asteraceae cp genomes.

Species	Accession number	Size (Kb)	Protein	rRNA	tRNA	Gene	Pseudogene
Lactuca sativa	NC_007578.1	152.765	84	7	37	128	-
Parthenium argentatum	NC_013553.1	152.803	55	8	17	106	16
Chrysanthemum indicum	NC_020320.1	150.972	83	8	34	125	-
Praxelis clematidea	NC_023833.1	151.41	84	8	32	131	7
Chrysanthemum x morifolium	NC_020092.1	151.033	85	8	35	128	-
Helianthus giganteus	NC_023107.1	151.066	85	8	36	131	2
Leontopodium leiolepis	NC_027835.1	151.072	85	8	37	132	2
Helianthus annuus	NC_007977.1	151.104	85	8	43	138	2
Guizotia abyssinica	NC_010601.1	151.762	85	8	37	132	2
Ageratina adenophora	NC_015621.1	150.698	86	8	37	136	5
Artemisia montana	NC_025910.1	151.13	86	8	37	133	2
Cynara cardunculus	KM035764	152.529	86	8	37	131	6
Aster spathulifolius	NC_027434.1	149.51	87	8	37	132	-
Jacobaea vulgaris	NC_015543.1	150.689	87	8	37	132	-
Artemisia frigida	NC_020607.1	151.076	87	8	37	134	2
Cynara baetica	NC_028005.1	152.548	87	8	37	136	4
Cynara cornigera	NC_028006.1	152.55	87	8	37	136	4
Cynara humilis	NC_027113.1	152.585	87	8	36	135	4
Silybum marianum	NC_028027.1	153.202	87	8	37	136	4
Centaurea diffusa	NC_024286.1	152.559	90	8	36	135	1
Helianthus maximiliani	NC_023114.1	151.007	85	9	36	131	1
Helianthus grosseserratus	NC_023108.1	151.017	85	9	36	131	1
Helianthus strumosus	NC_023113.1	151.044	85	9	36	131	1
Helianthus divaricatus	NC_023109.1	151.045	85	9	36	131	1
Helianthus hirsutus	NC_023111.1	151.045	85	9	36	131	1
Helianthus tuberosus	NC_023112.1	151.047	85	9	36	131	1
Helianthus decapetalus	NC_023110.1	151.048	85	9	36	131	1

LSC, SSC and IR-LSC/SSC Borders

The cp genome forms a double stranded, circular molecule, which is highly conserved in size, structure and gene content [7]. The quadripartite organization is shared by almost all cp genomes, consisting of a large-single-copy region (LSC; 80-90 kb) and a small-single-copy region (SSC; 16-27 kb), as well as two copies of inverted repeats (IRs) of ~20 to 28 kb in size [9,10]. The gene content and structure of angiosperm cp genome is highly conserved [11,12]. In 27 *Asteraceae* species, *G. abyssinica* cp genome contains one of the largest LSCs. *C. diffusa* has the smallest LSCs and the largest SSCs. *Ar. frigida* has the smallest SSC region **(Figure 2).** Expansion and contraction of the IR as well as gene and intron losses have been documented in a wide range of angiosperms [13,14]. Chloroplast gene order is also highly conserved among land plants, but in most instances when changes do

occur, they involve one or few inversions [16]. There are several groups of land plants that have experienced substantial numbers of cpDNA rearrangements, including conifers, the angiosperm families Campanulaceae, Fabaceae, Geraniaceae and Lobeliaceae [17,18]. Two cpDNA inversions of a large about 23kb and a smaller about 3.3 kb are shared by all major clades of Asteraceae, except members of Barnadesioideae, indicating that the two inversions may be a key future of the Asteraceae cp genomes [5,6,12,18]. The possible existence of an inverted SSC in Asteraceae cp genomes is still to be conformed but cannot be exclude given the nature of the flip-flop mechanism of the inverted repeats [19]. In *Ar. frigida*, a totally inversion SSC were observed compared with other angiosperm species, such as Arabidopsis [6]. However, the specific primers were used to validate the presumed inversion event would amplify the SSC no matter its orientation [3].

At the two SSC boundaries in cp genomes, the general structure was revealed in dicots (i.e., tobacco, Panax and *Arabidopsis*), and includes *ycf*1 spans and a *ycf*1 peseudogene adjacent to JSB in IRb [20]. The locations of the genes: *rps*19, *ycf*1, *ndh*F, *ycf*1* and *rps*19* except *trn*H are un-conservative in Asteraceae cp genomes **(Figure 2).** The *ycf*1 gene is distributed in the SSC region or IRb/SSC region, but only locates in the IRb region in *C. indicum*. In *Ar. Montana* the *rps*19* gene is in the IRa region, but others in the LSC region except being disappear in *As. spathulifolius*, *C. diffusa*, *Ch. indicum*, *Ch. x morilolium*, *J. vulgaris* and *L. sativa*. The *ndh*F varied in distance from the IRa/SSC border, and was entirely located in the SSC region in all *Asteraceae* species except *H. decapetalus* in IRa region and *S. marianim* in SSC/IRa border. In both *L. sativa* and *Ar. frigida*, *ndh*F located only 1 bp and 75 bp near the IRb/SSC border, and both the two species are invasive plants [6]. Compared with other monocot and dicot species, the position of the *trn*H gene in the cp genome is quite conserved. In general, the *trn*H gene is located in the IR region in the monocots, compared with its location in the LSC region in the dicots [21,22]. Same with all the dicots, in all Asteraceae species, the *trn*H gene is located in the LSC region [6].

Pseudogenes

Pseudogenes are functionless relatives of genes that have lost their gene expression in the cell or their ability to code protein [23]. Pseudogenes often result from the accumulation of multiple mutations within a gene, whose product is not required for the survival of the organism. Although not protein-coding, the DNA of pseudogenes may be functional, similar to other kinds of non-coding DNA which can have a regulatory role [24]. Twenty-two cp genomes were found pseudogenes among the twenty-seven Asteraceae plants **(Table 1)** and the different pseudogenes can be found in each cp genomes. In *C. cardunculus* three pseudogenes were identified: *ycf*68, in the IR, contains a premature stop codon in its coding sequence; the remaining two pseudogenes, *ycf*1 and *rps*19, are located in the boundary regions between IRb/SSC and Ira/SSC, respectively. The lack of their protein-coding ability is due to partial gene duplication [3]. The same three pseudogenes can also be found in *A. adenophora*, *Ar. Frigida* and *Praxelis clematiea* [6,7,13]. The difference is *ycf*68 in the IR become pseudogene due to several premature stop codons present in its coding sequence in *Ar. frigida* [25]. The *atp*B gene in relation to coding genes in *As. spathulifolius* [13], contained a start codon and formed a pseudogene due to deletion. The *atp*B gene is related to ATP synthase, and much more closely related to the *rbc*L gene with respect to its genetic structure. The *atp*B gene has often been used in evaluations of the upper family level And it also considered to be beneficial to phylogenetic research of the genus *Aster* and closely related groups [13]. But in *As. spathulifolius* it is not registered in the Genebank. In a major invasive species, *P. argentatum*, twelve pseudogenes were found: *atp*F, *ycf*3, *ycf*4, *rps*12, *clp*P, *rpl*16, *rps*3, *rpl*2, *rps*12, *ycf*1, *ndh*A, *ndh*B [7]. However, in *Helianthus* species, no more than two pseudogenes were found as *ycf*1 and *rps*19 in *H. annuus* and *ycf*1 in *H. decapetalus*. The gene *ycf*1 encodes a protein of unknown function that is essential, which appears to be a multi-pass trans-membrane protein, with no clear association to known functional domains [5,26].

DNA Barcodes

Several studies have analyzed the phylogenetic relationships in Asteraceae family based on cp sequences. One of the most comprehensive analyses included 108 taxa [27]. But until now, there were still no some special gene or combined genes can be the suitable DNA barcodes to discriminate all Asteraceae plants at the species level and below. For Asteraceae, the *ycf*1 and *ndh*F genes existed at the bottomed at first and ended up in a loss after gradually falling apart [12,13]. This region were known to be helpful to analysis of inter- genus evolution. The *ycf*1 gene is also be found the most divergent of all the genes in *A.adenophora* and *P.argentatum* [18]. So *ycf*1 gene may be the best suited gene for the phylogenetic analysis even though it was no effect to some species of Asteraceae. The *mat*K gene was used to analyze eight *Asteraceae* species, and it had no use to difference Parthenium with Lactuca subfamilies [20]. Even it can provide the sufficient information to differentiate three *Parthenium* species, the matK-barcode did not differentiate *P. argrntatum* or *P. argentatum* or *P. agentatum* lines from each other [12]. Using the combined barcodes, such as *mat*K and *psb*A-*trn*H, the additional differentiation at the some species level and below [12]. The genes *ndh*F and *trn*L-F were also chosen for the phylogenetic analysis of the 90 species in the Asteraceae family [25]. Other DNA barcodes were found in the Asteraceae phylogenetic research such as *trn*SUGA-*trn*fMCAU and *trn*SGCU-*trn*CGCA, *rps*32-*trn*L and *psb*A-*trn*H and other more genes were shown in **Table 2** [2,3,7,13,28,29]. In **Figure 3**, the combination of *ndh*C, *ndh*A and *ndh*G were used to analysis 27 *Asteraceae* species, seven species in *Helianthus*, two in *Chrysanthemum* and four in Cynara subfamily can be clustered in one group and be differentiated at species level. However, it also separated two *Eupatorium* species in to two groups. In Curci's research, whole cp sequence provided a higher phylogenetic resolution than using a subset of variable characters in Cynara [11]. With the more and more cp genomes registered in Genebank, The efficacy of the whole cp genome may be a super-barcode alongside with the reduction of sequencing costs of the Asteraceae family.

Table 2. DNA barcodes were used for phylogenetic tree in Asteraceae species.

列1	列2
Paper	DNA barcodes for phylogenetic tree
Kumar et al. [12]	matK, psbA-trnH, conbined matK and psbA-trnH
Garcia et al. [28]	trnSUGA-trnfMCAU, trnSGCU-trnCGCA
Doorduin et al. [2]	ndhC-trnV, ndhC-atpE, rps18-rp120,clpP, psbM-trnD, petN-psbM, rps8-rps14, ycf1, ycf3-trnS, ndhA, petD,petB, ndhI, rps8-rps3, rps15, rpoC1, psbB, rpoC2, nshG, rpoB, cemA, psaC, combined regions
Nie et al. [7]	atpA, atpB, matK, petA, petB, petD, petG, petN, psaA, psaB, psbA, psbB, psbC, psbD, psbE,psbF, psbH, psbI, psbJ, psbK, psbN, psbT, rpoB, rpoC1, rpoC2,rps8, rps11, rps14, ycf3, ndhA, ndhD, dhH, ndhF, rpoA
Riggins et al. [29]	rps32-trnL,psbA-trnH
Liu et al. [25]	ndhF, trnL-F
Zhang et al. [6]	ccsA-trnL, trnG-trnfM, rpl33-rps18, lhbA-trnG,rpoC2-rps2,cemA-petA,ndhG-ndhE,psbK-psb1,rpl16-rps3,clpP,matK,ycf 3,rps15,psbH,psbI,rbcL,ycf4,ndhK,atpF,rpl20,ndhI,rps8,rpoA,infA,cemA,rps14,ndhG,ndhH,combined regions
Choi et al. [13]	accD, atpB, atpE, cemA, clpP, infA,matK, ndhC, ndhJ, ndhK, petA, petB, petD, petG, petL, psaA, psaB, psaI, psaJ,psbA, psbB, psbC, psbD, psbF, psbH, psbI, psbK, psbL, psbN, psbT, psbZ, rbcL,rpl14, rpl16, rpl20, pl22, rpl23, rpl33, rpl36, rpoA, rps3, rps4, rps8, rps11,rps14, rps16, rps18, rps19 and ycf2
Curci et al. [3]	matK,ndhD,ndhF,ndhI,rncL,rpoB and the first exon of rpoC1

Perspectives

With the uncovered information of twenty-seven Asteraceae whole cp genomes in Genebank, we can get the following conclusion: From the size of cp genome, these are the larger cp genomes of Asteraceae compared with other plants. The Asteraceae cp genomes form a double stranded, circular molecule, which is highly conserved in size, structure and gene contents same with other plants. Pseudogenes can be found in most *Asteraceae* species and the genes are inconvenient. For the DNA barcodes, there were still no some special gene or combined genes can be the suitable DNA barcodes to discriminate all Asteraceae plants at the species level and below. But, with the more and more cp genomes registered in Gene bank, the efficacy of the whole cp genome may be a super-barcode alongside with the reduction of sequencing costs of the Asteraceae family.

ACKNOWLEGEMENTS

This work was supported national Science Foundation of China (No. 31360173).

REFERENCES

1. Maia GL, et al. Flavonoids from Praxelis clematidea R.M. King and Robinson modulate bacterial drug resistance. Molecules. 2011;16:4828-4835.
2. Doorduin L, et al. The complete chloroplast genome of 17 individuals of pest species Jacobaea vulgaris: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA Res. 2011;18:93-105.
3. Curci PL, et al. Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other Asteraceae. PLoS One. 2015;10:e0120589.
4. Curci PL, et al. Development of chloroplast genomic resources for Cynara. Mol Ecol Resour. 2016;16:562-573.
5. Timme RE, et al. A comparative analysis of the Lactuca and Helianthus (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. Am J Bot. 2007;94:302-312.
6. Zhang Y, et al. Complete chloroplast genome sequences of Praxelis (Eupatorium catarium Veldkamp), an important invasive species. Gene.. 2014;549:58-69.
7. Nie X, et al. Complete chloroplast genome sequence of a major invasive species, crofton weed (Ageratina adenophora). PLoS One. 2012;7:e36869.
8. Dyall SD, et al. Ancient invasions: from endosymbionts to organelles. Science. 2004;304:253-257.
9. Shinozaki K, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 1986;5:2043-2049.
10. Li X, et al. Plant DNA barcoding: from gene to genome. Biol Rev Camb Philos Soc. 2015;90:157-166.
11. Curci PL and Sonnante G. The complete chloroplast genome of Cynara humilis. Mitochondrial DNA. 2015 .
12. Kumar S, et al. Comparative analysis of the complete sequence of the plastid genome of Parthenium argentatum and identification of DNA barcodes to differentiate Parthenium species and lines. BMC Plant Biol. 2009;9:131.
13. Choi KS and Park S. The complete chloroplast genome sequence of Aster spathulifolius (Asteraceae); genomic features and relationship with Asteraceae. Gene. 2015;572:214-221.
14. Dempewolf H, et al. Establishing genomic tools and resources for Guizotia abyssinica (L.f.) Cass.-the development of a

library of expressed sequence tags, microsatellite loci, and the sequencing of its chloroplast genome. Mol Ecol Resour. 2010;10:1048-1058.

15. Markina NV, et al. Study of Chloroplast DNA Polymorphism in the Sunflower (Helianthus L.). Genetika. 2015;51:873-880.
16. Raubeson LA and Jansen RK. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science. 1992;255:1697-1699.
17. Tsumura Y, et al. Chloroplast DNA inversion polymorphism in populations of Abies and Tsuga. Mol Biol Evol. 2000;17:1302-1312.
18. Kim KJ, et al. Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Mol Biol Evol. 2005;22:1783-1792.
19. Martin GE, et al. The first complete chloroplast genome of the Genistoid legume Lupinus luteus: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. Ann Bot. 2014;113:1197-1210.
20. Yang M, et al. The complete chloroplast genome sequence of date palm (Phoenix dactylifera L.). PLoS One. 2010;5:e12762.
21. Wiegert KE, et al. Evolution of the chloroplast genome in photosynthetic euglenoids: a comparison of Eutreptia viridis and Euglena gracilis (Euglenophyta). Protist. 2012;163:832-843.
22. Kuang DY, et al. Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics. Genome 54: 663-673.
23. Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet. 2011;19:253-272.
24. Poliseno L, et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465:1033-1038.
25. Liu Y, et al. Complete chloroplast genome sequences of Mongolia medicine Artemisia frigida and phylogenetic relationships with other plants. PLoS One. 2013;8:e57533.
26. Drescher A, et al. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 2000;22:97-104.
27. Panero JLand Funk VA. The value of sampling anomalous taxa in phylogenetic studies: major clades of the Asteraceae revealed. Mol Phylogenet Evol. 2008;47:757-782.
28. Garcia S, et al. A molecular phylogenetic approach to western North America endemic Artemisia and allies (Asteraceae): untangling the sagebrushes. Am J Bot. 2011;98:638-653.
29. Riggins CW and Seigler DS. The genus Artemisia (Asteraceae: Anthemideae) at a continental crossroads: molecular insights into migrations, disjunctions, and reticulations among Old and New World species from a Beringian perspective. Mol Phylogenet Evol. 2012;64:471-490.