Analyzing genetic diversity of chloroplast genomes in Liliales

Do Hoang Dang Khoa

Hi-Tech Institute, Nguyen Tat Thanh University
dhdkhoa@ntt.edu.vn

Abstract

Liliales is a monocotyledonous order and contains both photosynthetic and mycoheterotrophic species that distribute locally or worldwide. In this study, the genetic diversity of chloroplast genomes in Liliales was explored regarding their nucleotide diversity and repeated composition. The analysis of nucleotide diversity revealed various hotspots in large and small single-copy regions whereas the IR regions had low sequence divergence. Although each family has specific hotspots, the *rps15-ycf1* region was commonly found as a highly variable area in the cpDNA of observed taxa. In the cpDNA of Liliales, mononucleotide simple sequence repeat (SSR) is the most common type. The majority of SSRs are located in non-coding regions. Similarly, more long repeats were found in non-coding areas than in coding sequences. Additionally, the complement repeat exceeds forward type in the cpDNA of Liliales. The highest number of long repeats was found in *Corsia dispar* whereas that of SSRs was detected in *Smilax china*. The results of nucleotide diversity and repeat analyses provided fundamental information for further studies on population genetics, molecular marker development and evolutionary history of Liliales.

Received 26.01.2022
Accepted 15.03.2022
Published 06.04.2022

Keywords
Liliales,
chloroplast genome,
nucleotide diversity,
Pi values, repeat

1 Introduction

Liliales, a monocotyledonous order of angiosperms, includes nine families of over 1500 species [1]. The families of Liliales distribute worldwide or locally. For example, Smilacaceae species are widespread from Australia, Europe, to Africa and Asia whereas the monotypic family, Petermanniaceae, can only be found in Australia. Additionally, there are two types of plants in Liliales, including mycoheterotrophic type in Corsiaceae and photosynthetic type in the remained families [2]. Because of their opposite patterns of lifestyle and distribution, Liliales is a good model to explore the evolutionary history of land plants. Previously, biogeography and divergent time estimation of Liliales were conducted [3]. The results showed a divergent time of 124 million years ago (mya) from other monocots and the families were splitted approximately 113 mya. In addition to divergent time, the origin of Liliales was found in Australia where the ancestors of Liliales would then be widespread and evolved [3]. Beside order level, the divergent time estimation and biogeography of each family of Liliales were approached. Liliaceae originated from temperate Asia in the late Cretaceous (85 mya) to occupy the northern hemisphere [4]. Meanwhile, the members of Melanthiaceae used the Bering Land Bridge to migrate from North America to East Asia around 92.1 mya [5]. Colchicaceae arose in Australia 67 mya and migrated to Africa and North America [6]. Smilacaceae is an interesting family of Liliales that has many fossil records for elucidating the evolutionary history of Liliales [7-8]. These findings suggested an interesting evolutionary history of Liliales, especially at genomic level.

https://doi.org/10.55401/733csy82

Chloroplast genome (cpDNA) is one of three existing genomes (including mitochondrial, nucleus and chloroplast genomes) in most land plants. Typically, cpDNA has a quadripartite structure which includes one large single copy (LSC) and a small single copy (SSC) separated by two inverted repeat (IR) regions [9]. Also, cpDNA contains 80 protein-coding genes, 30 tRNAs and four rRNAs and some of the protein-coding genes are related to photosynthesis. These genomic data are crucial for elucidating the phylogeny of land plants; therefore, the 1000 plant genomes project was conducted, followed by another 10 000 plant genomes project [10-11]. As a result, a billion years of evolutionary history of plants was explored [12]. Additionally, cpDNA is a useful source for mining molecular markers for population genetics and plant identification [13–17]. In Liliales, the cpDNA sequences from all families were reported [18-21]. These data provided essential information for elucidating the evolution of Liliales [4–8,22]. Although the complete cpDNA of Liliales have been reported, there has been no compilation of data for nucleotide diversity and repeat composition among Liliales families. Therefore, in this study, the available cpDNA data of Liliales were combined to locate the highly variable regions. Additionally, the simple sequence repeat (SSR) and long repeat were screened across cpDNA of Liliales. These new results will add insights into the evolutionary history of Liliales.

2 Materials and methods

2.1 Sampling chloroplast genome data

Complete chloroplast genome (cpDNA) sequences of Liliales were searched on NCBI (National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/)) using the keywords “Liliales, chloroplast, complete genome”. The search results revealed all complete chloroplast genomes of Liliales, especially duplicated data for a species. Therefore, only one complete sequence (without unknown nucleotide in the genome) of a species was randomly selected as a presentative because the similarity of the duplicated data is usually over 99.5 % (data not shown). Then, the selected complete genomes were downloaded under the GenBank (full) format which includes various information of chloroplast genomes such as gene content, gene location, length, GC content, etc. All the data were imported to the Geneious Prime program for further analysis.

2.2 Nucleotide diversity analysis

To calculate the nucleotide diversity (Pi values, resulted from estimating the average number of nucleotide differences per site among DNA sequences) among chloroplast genomes of Liliales. The higher nucleotide diversity is, the higher genetic variation is detected in the genome. DnaSP 6 program was employed. First, the Pi values were estimated at familial level, making the complete cpDNA within each family of Liliales aligned using the MAUVE program embedded in Geneious Prime. The aligned sequences were then imported to DnaSP6 for Pi value calculation and sliding window analysis with the window size of 2 000 and the step size of 100. Among the families of Liliales, there is a monotypic family labeled as Petermanniaceae. Additionally, only one complete cpDNA was reported in Campynemataceae. The Corsiaceae includes heterotrophic species (i.e., *Corsia dispar* and *Arachnitis uniflora*) which exhibit extreme structural changes. Therefore, the nucleotide diversity analysis was not conducted for Petemanniaceae, Campynemataceae and Corsiaceae in this study.

2.3 Examination of repeat structure and microsatellites

For screening repeat number and location in cpDNA of Liliales, REPuter program was used with the minimum length of 20 bp for forward, reverse, and complement repeats. Meanwhile, Phobos program embedded in Geneious Prime was used for identifying microsatellites number and location with the minimum lengths of 10 bp for mono-, 12 bp for di-, 15 bp for tri-, 16 bp for tetra-, 20 bp for penta-, and 24 bp hexanucleotide repeats. A representative species of each genus in Liliales was selected randomly from available data for examing the repeat content in this study.

3 Results and discussion

3.1 Features of chloroplast genome in Liliales

On the NCBI database, 177 out of over 1500 species of Liliales have records of complete chloroplast genome (Table 1). Liliaceae family have the highest number of complete chloroplast genomes (105 species), followed by Melanthiaceae (49 species), Colchicaceae (9 species), Smilacaceae (4 species), Alstroemeriaceae (3 species), Philesiaceae (3 species), Corsiaceae (2 species) and one species each for Petermanniaceae and Campynemataceae. The lengths of chloroplast genomes range from 24 846 bp (Arachnitis uniflora, Corsiaceae) to 163 860 bp (Paris liiana, Melanthiaceae). The GC content of Liliales is 37 % on average. Although Corsia dispar has a reduced size of cpDNA as found in Arachnitis uniflora, the GC content of the former is 30.8 %, which is lower than the latter (37.1 %) and other observed species (Table 1). Most of cpDNAs of Liliales encode 80 proteins, 30 tRNAs and four rRNAs (Table 1, Table 2). However, there are only 79 protein-coding genes in Amana species and Chionographis japonica, of which infA and rps16 were lost, respectively. In contrast to other species, two members of Corsiaceae exhibited an extreme loss of protein-coding gene and tRNA (Table 2). Specifically, Corsia dispar has 30 protein-coding genes and 24 tRNAs whereas Arachnitis uniflora includes 16 protein-coding genes and 5 tRNAs. Notably, these two species still have four rRNAs that are commonly found in other Liliales taxa (Table 2).

There are two groups of species in Liliales according to cpDNA structure. The first group contains photosynthetic species that has typical structure of cpDNA including one large single copy (LSC), a small single copy (SSC) and two inverted repeat (IR) regions and contains approximately 80 protein-coding genes, 30 tRNA-coding genes and four rRNA-coding genes. The second group includes mycoheterotrophic species that exhibited an extreme loss of genes and significant changes of genome structure. Although Arachnitis uniflora has fewer genes than Corsia dispar, the former cpDNA has a typical quadripartite structure that was not found in the latter. This phenomenon suggested different stages of change in chloroplast genomes of mycoheterotrophic species. Previously, the plastid genomes of Ericaceae revealed the loss of genes related to photosynthesis whereas the other genes were remained [23]. Similarly, different numbers of gene loss were found in orchids that provided a scenario of 5 steps for the loss of plastid genes [24-25].

Table 1 Comparison of the features of plastid genomes from ten families of Liliales

Family	Species	Accession number	Length (bp)	GC content (%)	Gene content
					(Protein coding/ tRNA/rRNA)
Liliaceae (105 species)	Amana anhuiensis	KY101423	150 842	36.7	79/30/4
	Amana baohuaensis	MT898423	150 757	36.7	79/30/4
	Amana edulis	KY401425	151 136	36.7	79/30/4
	Amana erytgronioides	KY401421	150 858	36.7	79/30/4
	Amana kuocangshanica	KY401423	151 058	36.7	79/30/4
	Amana wanzhensis	KY401422	150 913	36.7	79/30/4
	Calochortus uniflorus	MK673754	155 794	37.4	80/30/4
	Calochortus venustus	MT261150	155 688	37.4	80/30/4
	Cardiocrinum cathayanum	KX575836	152 415	37.1	80/30/4
	Cardiocrinum cordatum	KX575837	152 410	37.1	80/30/4
	Cardiocrinum giganteum	KX528334	152 653	37.1	80/30/4
	Clintonia udensis	MT261153	156 214	37	80/30/4
	Erythronium japonicum	MT261155	151 416	36.6	80/30/4
	Erythronium sibiricum	KX644899	151 034	36.7	80/30/4
	Fritillaria anhuiensis	MH593363	152 119	37	80/30/4
	Fritillaria cirrhosa	KF769143	151 991	36.9	80/30/4
	Fritillaria crassicaulis	MK258147	151 852	37	80/30/4
	Fritillaria dajinensis	MH244913	151 991	36.9	80/30/4

	Fritillaria davidii	MK158145	152 044	37	80/30/4
	Fritillaria delavayi	MN480806	151 938	37	80/30/4
	Fritillaria eduardii	MF947708	152 224	37	80/30/4
	Fritillaria hupehensis	NC024736	152 145	37	80/30/4
	Fritillaria karelinii	KX354691	152 118	36.9	80/30/4
	Fritillaria maximowiczii	MK258138	152 434	37.1	80/30/4
	Fritillaria meleagroides	MF947710	151 846	37	80/30/4
	Fritillaria pallidiflora	MG211822	152 078	37	80/30/4
	Fritillaria persica	MF947709	151 803	37	80/30/4
	Fritillaria prewalskii	MH244908	151 983	36.9	80/30/4
	Fritillaria sichuanica	MH244907	151 958	37	80/30/4
	Fritillaria sinica	MH244912	152 064	36.9	80/30/4
	Fritillaria taipaiensis	KC543997	151 693	37	80/30/4
	Fritillaria tortifolia	MG211819	152 005	37	80/30/4
	Fritillaria thungergii	MH244914	152 160	37	80/30/4
	Fritillaria unibracteata	MH244909	151 058	37	80/30/4
	Fritillaria unibracteata var wabuensis	KF769142	151 009	37	80/30/4
	Fritillaria ussuriensis	MT261156	152 156	37	80/30/4
	Fritillaria verticillata	MG211823	151 959	37	80/30/4
	Fritillaria walujewii	MG211820	151 920	36.9	80/30/4
	Fritillaria yuminensis	MG200070	151 813	37	80/30/4
	Fritillaria yuzhongensis	MK258139	151 652	37	80/30/4
	Gagea triflora	MT261157	150 345	37	80/30/4
	Lilium bulbiferum	MW465412	152 690	37	80/30/4
	Lilium amabile	MT261159	152 614	37	80/30/4
	Lilium amoenum	MT880912	152 280	37	80/30/4
	Lilium bakerianum	KY748301	151 655	37.1	80/30/4
	Lilium brownii	KY748296	152 677	37	80/30/4
	Lilium callosum	MT261160	152 630	37	80/30/4
	Lilium candidum	MK753244	152 101	37	80/30/4
	Lilium cernuum	MT261161	152 553	37	80/30/4
	Lilium davidii var. uniclolor	MK954110	152 659	37	80/30/4
	Lilium distichum	NC029937	152 598	37.1	80/30/4
	Lilium duchartei	KY748300	152 287	37	80/30/4
	Lilium fargesii	KX592156	153 235	36.0	80/30/4
	Lilium formosanum	MT261162	152 610	37	80/30/4
	Lilium gongshanense	MK493297	151 974	37	80/30/4
	Lilium hansonii	MT261163	152 168	37	80/30/4
	Lilium henricii	MH136807	152 784	37	80/30/4
	Lilium henryi	KY748302	153 119	37	80/30/4
	Lilium japonicum	MT261164	152 613	37.1	80/30/4
	Lilium lancifolium	MH177880	152 479	37	80/30/4
	Lilium lankongense	MK757466	152 611	37	80/30/4
	Lilium leichtlinii var. maximowiczii	MK753242	152 604	37	80/30/4
	Lilium leucanthum	KY748299	152 935	37	80/30/4

	Lilium longiflorum	KC968977	152 793	37.02	80/30/4
	Lilium lophophorum	MK493298	152 382	37	80/30/4
	Lilium martagon var. pilosiusculum	MF964219	152 816	37	80/30/4
	Lilium matagense	MN745201	152 402	37	80/30/4
	Lilium meleagrinum	MK493299	152 197	37	80/30/4
	Lilium nanum	MK493300	152 417	37	80/30/4
	Lilium nepalense	MK493301	152 316	37	80/30/4
	Lilium pardalinum	MH029495	151 969	37	80/30/4
	Lilium pardanthinum	MG704135	152 718	37	80/30/4
	Lilium pensylvanicum	MK493295	152 058	37.1	80/30/4
	Lilium primulinum var. ochraceum	KY7482988	152 036	37	80/30/4
	Lilium pumilum	MK954109	152 591	37	80/30/4
	Lilium philadelphicum	KY940847	152 175	37.1	80/30/4
	Lilium regale	MK493302	153 082	37	80/30/4
	Lilium rosthornii	MW136390	152 956	37	80/30/4
	Lilium sargentiae	MK493303	153 129	37	80/30/4
	Lilium souliei	MW007720	152 326	37	80/30/4
	Lilium speciosum var. gloriosoides	MN509267	152 912	37.02	80/30/4
	Lilium sulphureum	MK493304	153 107	37	80/30/4
	Lilium superbum	NC026787	152 069	37	80/30/4
	Lilium taliense	KY009938	153 055	36.9	80/30/4
	Lilium tsingtauense	KU230438	151 983	37	80/30/4
	Lilium washintonianum	MH590100	151 967	37.1	80/30/4
	Lilium xanthellum	MN745202	151 967	37.1	80/30/4
	Lloydia tibetica	MK673752	150 379	36.9	80/30/4
	Medeola virginiana	MK673752	153 914	37	80/30/4
	Nomocharis aperta	MK493293	152 042	37	80/30/4
	Nomocharis pardanthina	NC_038193	152 718	37	80/30/4
	Notholirion bulbuliferum	MN509268	153 019	37.1	80/30/4
	Notholirion campanulatum	MK673746	153 169	37	80/30/4
	Notholirion macrophyllum	MH011354	152 143	37.1	80/30/4
	Prosartes lanuginosa	MK673749	158 265	37	80/30/4
	Scoliopus bigelovii	MK673747	154 698	37.2	80/30/4
	Streptopus ovalis	MT261171	157 359	37.1	80/30/4
	Tulipa altaica	MK673755	146 887	37.1	80/30/4
	Tulipa buhseana	MT316022	152 062	36.6	80/30/4
	Tulipa iliensis	MW077740	152 073	36.6	80/30/4
	Tulipa patens	MT327740	152 050	36.7	80/30/4
	Tulipa sylvestris	MT261172	151 940	36.7	80/30/4
	Tulipa thianschanica	MT327741	152 122	36.6	80/30/4
	Tricyrtis formosana	MK673751	156 018	37.3	80/30/4
	Tricyrtis macropoda	MT261173	155 453	37.4	80/30/4
Smilacaceae	Smilax china	HM536959	157 878	37.25	80/30/4
	Smilax glyciphylla	MT261169	158 922	36.9	80/30/4
	Smilax microphylla	MW423607	158 246	37.1	80/30/4

	Smilax nipponica	MT261170	158 178	37.1	80/30/4
Philesiaceae	Ripogonum scandens	MT261167	160 287	37.6	80/30/4
	Philesia magellanica	MT261166	158 786	37.6	80/30/4
	Lapageria rosea	MT261158	160 054	37.5	80/30/4
Melanthiaceae (49 species)	Chionographis japonica	KF951065	154 646	37.7	79/30/4
	Heloniopsis tubiflora	KM078036	157 940	37.5	80/30/4
	Paris axialis	MN125591	156 821	37.4	80/30/4
	Paris bashanensis	MN125580	157 320	37.7	80/30/4
	Paris birmanica	MN125580	157 857	37.3	80/30/4
	Paris caobangensis	MN125593	158 256	37.2	80/30/4
	Paris caojianensis	MZ147601	163 853	37	80/30/4
	Paris cronquistii	KX784041	157 710	37.3	80/30/4
	Paris daliensis	MN125574	158 118	37.3	80/30/4
	Paris delavayi	MN125581	158 575	37.2	80/30/4
	Paris dulongensis	MN125566	157 342	37.4	80/30/4
	Paris dunniana	KX784042	157 984	37.2	80/30/4
	Paris fargesii	KX784043	157 518	37.3	80/30/4
	Paris forrestii	KX784044	158 345	37.3	80/30/4
	Paris incompleta	MN125572	157 610	37.7	80/30/4
	Paris japonica	MH796668	155 957	37.6	80/30/4
	Paris liiana	MT857225	163 860	37	80/30/4
	Paris luquanensis	KX784045	158 451	37.3	80/30/4
	Paris marei	KX784046	157 891	37.3	80/30/4
	Paris marmorata	KX784047	157 566	37.3	80/30/4
	Paris polyphylla var chinensis	KX784048	158 307	37.2	80/30/4
	Paris polyphylla var yunnanensis	KX784049	157 547	37.3	80/30/4
	Paris qiliangiana	MN125576	158 354	37.2	80/30/4
	Paris quadrifolia	KX784051	157 097	37.7	80/30/4
	Paris rugosa	MN125570	157 239	37.4	80/30/4
	Paris stigmatosa	MN125570	157 239	36.8	80/30/4
	Paris tengchongensis	MN125584	157 150	37.4	80/30/4
	Paris tetraphylla	MN125596	156 567	37.5	80/30/4
	Paris thibetica	MN125596	157 389	37.4	80/30/4
	Paris undulata	MN125586	158 286	37.2	80/30/4
	Paris vaniotii	MN125567	156 846	37.4	80/30/4
	Paris verticillata	KJ433485	157 379	37.6	80/30/4
	Paris vietnamensis	KX784050	158 224	37.2	80/30/4
	Paris xichouensis	MN125585	158 225	37.3	80/30/4
	Paris yanchii	MN125582	157 918	37.3	80/30/4
	Trillium camschatcense	MN125568	156 139	37.5	80/30/4
	Trillium cuneatum	NC027185	156 610	37.5	80/30/4
	Trillium decumbens	NC027282	158 552	37.7	80/30/4
	Trillium govanianum	MH796670	157 379	37.7	80/30/4
	Trillium maculatum	KR780075	157 359	37.5	80/30/4
	Trillium tschonoskii	KR780076	156 852	37.5	80/30/4

	Veratrum japonicum	MG940972	151 791	37.7	80/30/4
	Veratrum mengtzeanum	MW147219	153 705	37.8	80/30/4
	Veratrum oxysepalum	MW147219	153 705	37.7	80/30/4
	Veratrum patulum	KF437397	153 699	37.7	80/30/4
	Veratrum taliense	MN125578	151 909	37.8	80/30/4
	Xerophyllum tenax	KM078035	156 746	37.8	80/30/4
	Ypsilandra thibetica	MH796671	157 613	37.5	80/30/4
	Ypsilandra yunnanensis	MH796672	158 806	37.4	80/30/4
Alstroemeriaceae	Alstroemeria aurea	KC968976	155 510	37.26	80/30/4
	Bomarea edulis	KM233641	154 925	38.2	80/30/4
	Luzuriaga radicans	KM233640	157 885	38.1	80/30/4
Colchicaceae (9 species)	Androcymbium greuterocymbium	MT261148	154 804	37.6	80/30/4
	Colchicum autumnale	KP125337	156 462	37.6	80/30/4
	Disporum cantoniense	MW759302	156 688	37.6	80/30/4
	Disporum sessile	MN332241	159 102	37.3	80/30/4
	Gloriosa superba	KP125338	157 924	37.6	80/30/4
	Iphgenia indica	MT012417	158 319	37.4	80/30/4
	Tripladenia cunninghamii	MT261174	155 652	37.6	80/30/4
	Uvularia grandiflora	MT261175	157 025	37.6	80/30/4
	Wurmbea burtii	MT261176	155 297	37.7	80/30/4
Petermanniaceae	Petermannia cirrhosa	MT261165	156 852	38	80/30/4
Campynemataceae	Campynema lineare	MT261151	156 305	36.9	80/30/4
Corsiaceae	Corsia dispar	MT261154	63 172	30.8	30/24/4
	Arachnitis uniflora	MT261149	24 846	37.1	16/05/4

Table 2 Gene contents in the chloroplast genomes of Liliales taxa

Group of gene		Name of gene(common)
RNA genes	Ribosomal RNAs	rrn4.5(x2) xz, rrn5(x2) xz, rrn16(x2) xz, rrn23(x2) xz
	Transfer RNAs	trnA-UGCa(x2) x, trnC-GCAxz, trnD-GUCx, trnE-UUCxz, trnF-GAAx, trnfM-CAUxz, trnG-GCC, trnG-UCCa, trnH-GUGx, trnI-CAU(x2) x, trnI-GAUa(x2), trnK-UUUax, trnL-CAA(x2) x, trnL-UAAax, trnL-UAGx, trnM-CAUx, trnN-GUU(x2) x, trnP-UGGx, trnQ-UUGxz, trnR-ACG(x2) x, trnR-UCU, trnS-GCUx, trnS-GGAx, trnS-UGAx, trnT-GGUx, trnT-UGUx, trnV-GAC(x2) x, trnV-UACa, trnW-CCAxz, trnY-GUAx
Protein genes	Photosystem I	psaAx, psaB, psaC, psaI, psaJ
	Photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ
	Cytochrome	petA, petB, petD, petG, petL, petN
	ATP synthase	atpA, atpBx, atpEx, atpFa, atpH, atpIx
	Rubisco	rbcL
	NADH dehydrogenase	ndhAa,ndhBa(x2)x,ndhC,ndhD,ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK
	ATP-dependent protease subunit P	clpPaxz

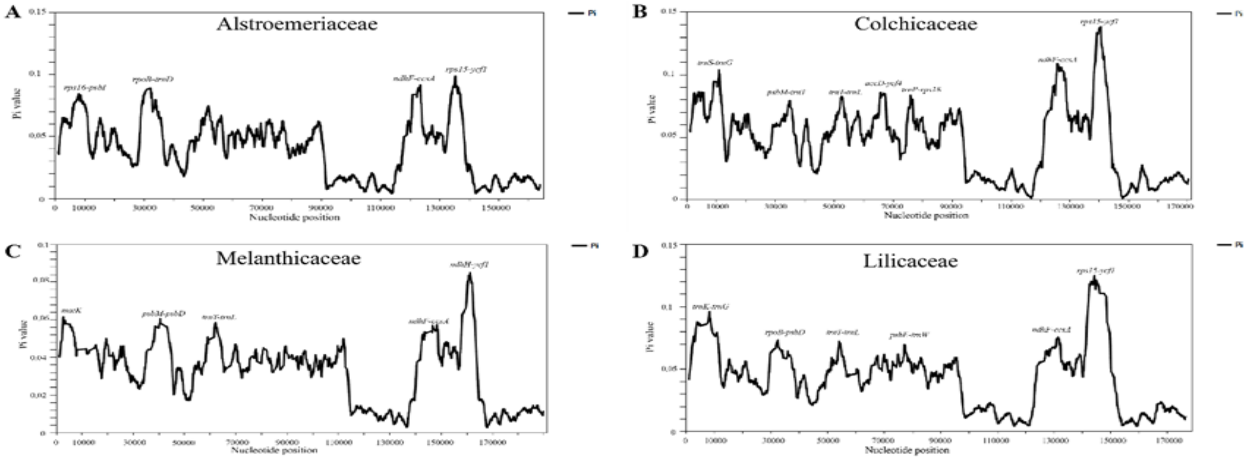
	Chloroplast envelope membrane protein	cemA
Ribosomal proteins	large units	rpl2a(x2) xz, rpl14xz, rpl16xz, rpl20xz, rpl22x, rpl23 (x2), rpl32x, rpl33x, rpl36x
	small units	rps2xz, rps3xz, rps4xz, rps7(x2) xz, rps8xz, rps11xz, rps12a(x2) xz, rps14xz, rps15x, rps16x, rps18x, rps19(x2) xz
Transcription	RNA polymerase	rpoA, rpoB, rpoC1a,rpoC2
/translation	Initiation factor	infA
	Miscellaneous proteins	accDxz, ccsA, matKx
	Hypothetical proteins & Conserved reading frame	ycf1x, ycf2 (x2) x, ycf3, ycf4x, ycf15
a: gene has intron; x2: gene has two copies; x : remained in Corsia dispar; z : remained in Arachnitis uniflora		

In the first stage, the ndh genes were lost, followed by the disappearance of photosynthetic genes. In the third stage, the genes for RNA polymerase were not found. In the fourth and fifth stages, genes for ATP synthase and other functions were lost, respectively. In Corsiaceae of Liliales, the cpDNA of Arachnitis uniflora and Corsia dispar are in the third and the fourth stages (Table 2). However, there are only two out of 26 species of Corsiaceae that have available cpDNA on NCBI data. Therefore, further studies that cover all species of Corsiaceae should be conducted to provide a better understanding of the evolutionary history of mycoheterotrophic species in Liliales.

Among photosynthetic species of Liliales, there are records of infA and rps16 loss in cpDNA sequences of Amana and Chionographis species (Table 1). Previously, the loss of infA in cpDNA was detected in many angiosperms and the intact infA was found in nucleus [26]. Similarly, the loss of rps16 was also found in other angiosperms that was compensated by a copy of rps16 in the nucleus genome [27-28]. In Liliales, the loss of gene was recorded but there is no study on the effect of that loss in cpDNA. Therefore, further studies on the impact of gene loss in photosynthetic as well as mycoheterotrophic species of Liliales should be conducted.

3.2 Nucleotide diversity patterns in chloroplast genomes of Liliales

The nucleotide diversity analysis revealed different Pi values among families of Liliales (Figure 1). The high Pi values (> 0.1) were recorded in Alstroemeriaceae, Colchicaceae and Liliaceae (Figure 1A, 1B, 1D) whereas Philesiaceae and Smilacaceae have smaller Pi values (< 0.04) (Figure 1E, 1F). In Melanthiaceae, the high Pi values range from 0.04 to 0.1 (Figure 1C). In Alstroemeriaceae, the high Pi values were found in rps16-psbI, rpoB-trnD, ndhF-ccsA and rps15-ycf1 (Figure 1A). In Colchicaceae, trnS-trnG, psbM-trnT, trnT-trnL, accD-ycf4, trnP-rps18, ndhF-ccsA and rps15-ycf1 exhibit high Pi values (Figure 1B). In Melanthiaceae, five regions including matK, psbM-psbD, trnT-trnL, ndhF-ccsA and ndhH-ycf1 have high Pi values.



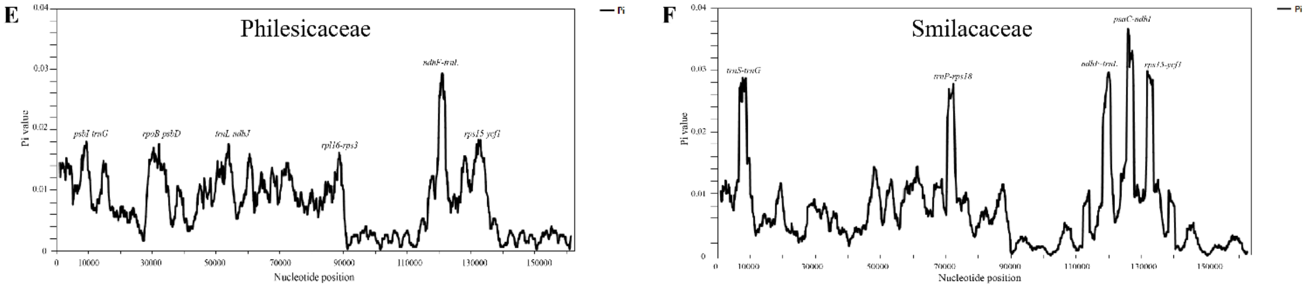


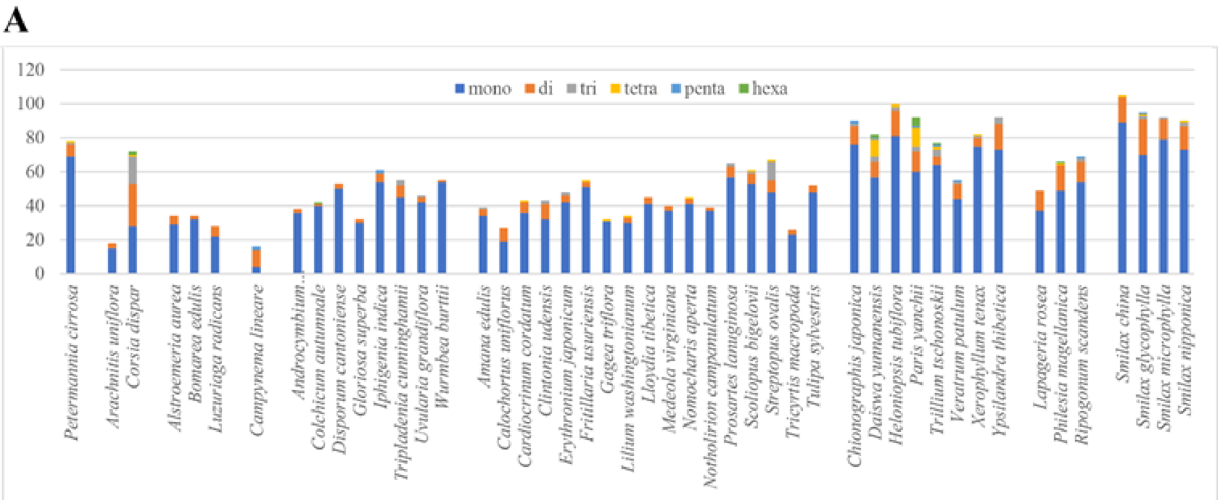
Figure 1 Sliding window analysis of the whole chloroplast genomes of Liliales species. (window length: 2 000 bp, step size: 100 bp). X-axis: position of nucleotide, Y-axis: Pi values of each window.
A. Alstroemeriaceae; B. Colchicaceae; C. Liliaceae; D. Melanthiaceae; E. Philesiaceae; F: Smilacaceae

In Liliaceae, trnK-trnG, rpoB-psbD, trnT-trnL, psbE-trnW, ndhF-ccsA and rps15-ycf1 regions showed high Pi values. In Philesiaceae, high Pi values were found in psbI-trnG, rpoB-psbD, trnL-ndhJ, rpl16-rps3, ndhF-trnL and rps15-ycf1 (Figure 1E). In Smilacaceae, five regions have high Pi values including trnS-trnG, trnP-rps18, ndhF-trnL, psaC-ndhI and rps15-ycf1 (Figure 1F). Most of high Pi values were found in non-coding regions but some coding regions such as matK and ycf1 also had high nucleotide diversity.

Similar to Liliales, nucleotide diversity has been explored in cpDNA of various angiosperms. For example, in Paris species (Melanthiaceae), divergent hotspots were found in both coding regions (rpoC1 and ycf2) and non-coding regions (trnS-trnG, rpl32-trnL, etc.) [21]. In other land plants such as species of Symplocos, Avena and Senecioneae, various hotspots with high Pi values were located in different regions of cpDNA [29-31]. The information of hotspots is a useful source for developing molecular markers in angiosperms [32]. In Liliales, highly variable regions were identified for each family, except Corsiaceae, Petermanniaceae and Campynemataceae due to the lack of data and mycoheterotrophic lifestyle. Among the hotspots, ycf1 is the common region in observed families. However, this region should be verified with the lack of data from Campynemataceae and Petermanniaceae in further studies.

3.3 Comparison of Repeat composition in chloroplast genomes in Liliales

The analysis of SSRs in cpDNA of Liliales resulted in different numbers of repeats among families (Figure 2A). The highest number of SSRs was found in Smilax china (105 records) whereas Campynema lineare and Arachnitis uniflora have 16 and 18 SSRs in cpDNA, respectively. Among six types of SSR, the mononucleotide repeat (A/T) is the most abundant (2191 records) followed by dinucleotide (AT/TA/GC/CG) type (339 records).



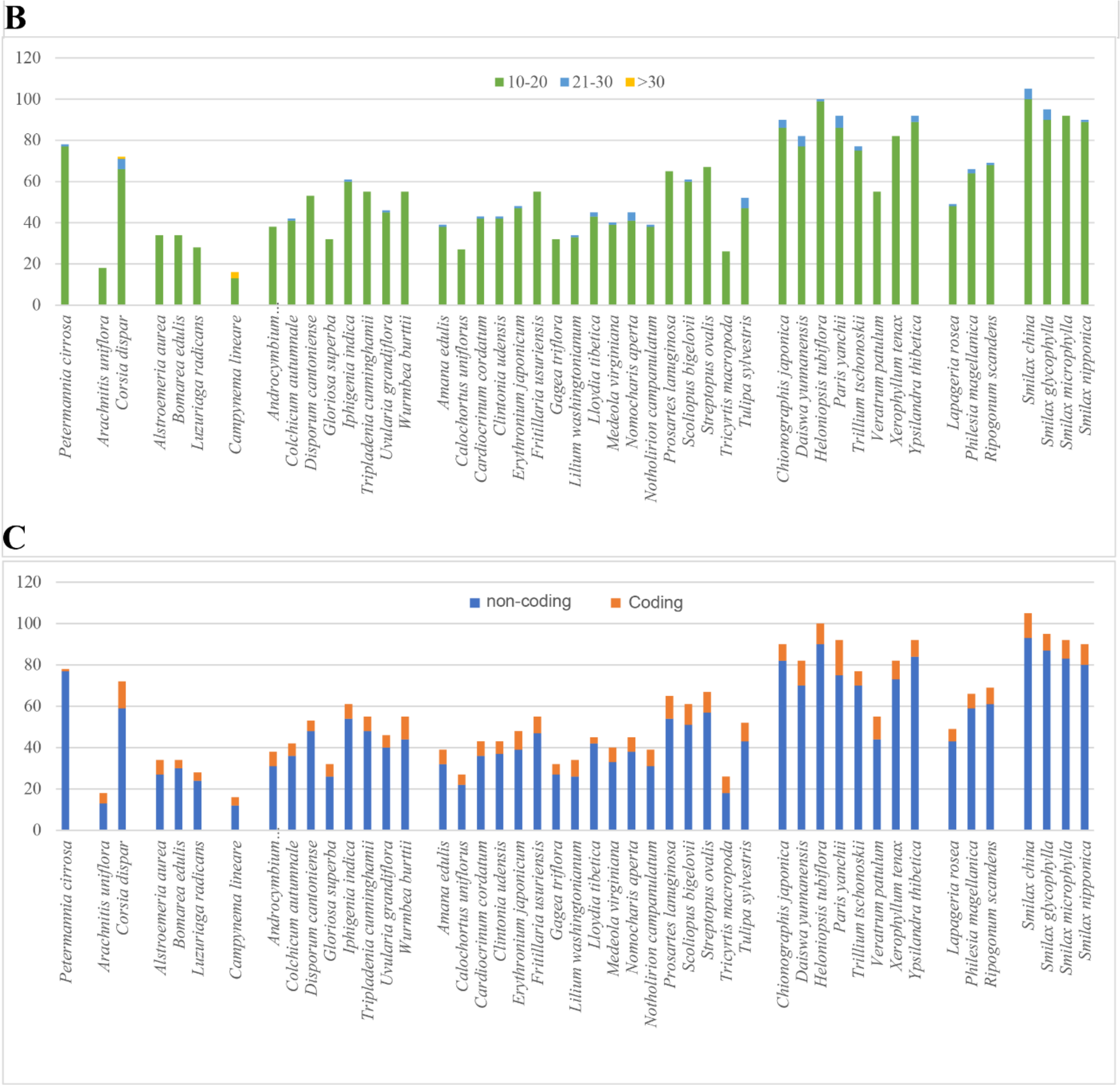
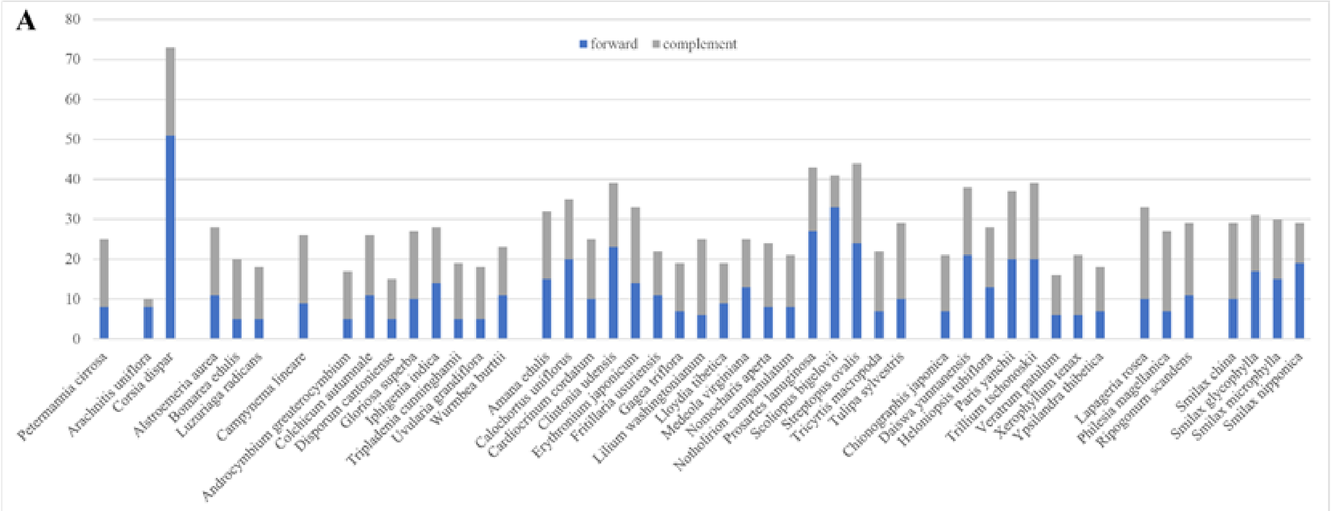


Figure 2 Quantity of SSR in chloroplast genomes of Liliales. A. Types of SSR; B. Length of SSR; C. Location of SSR.



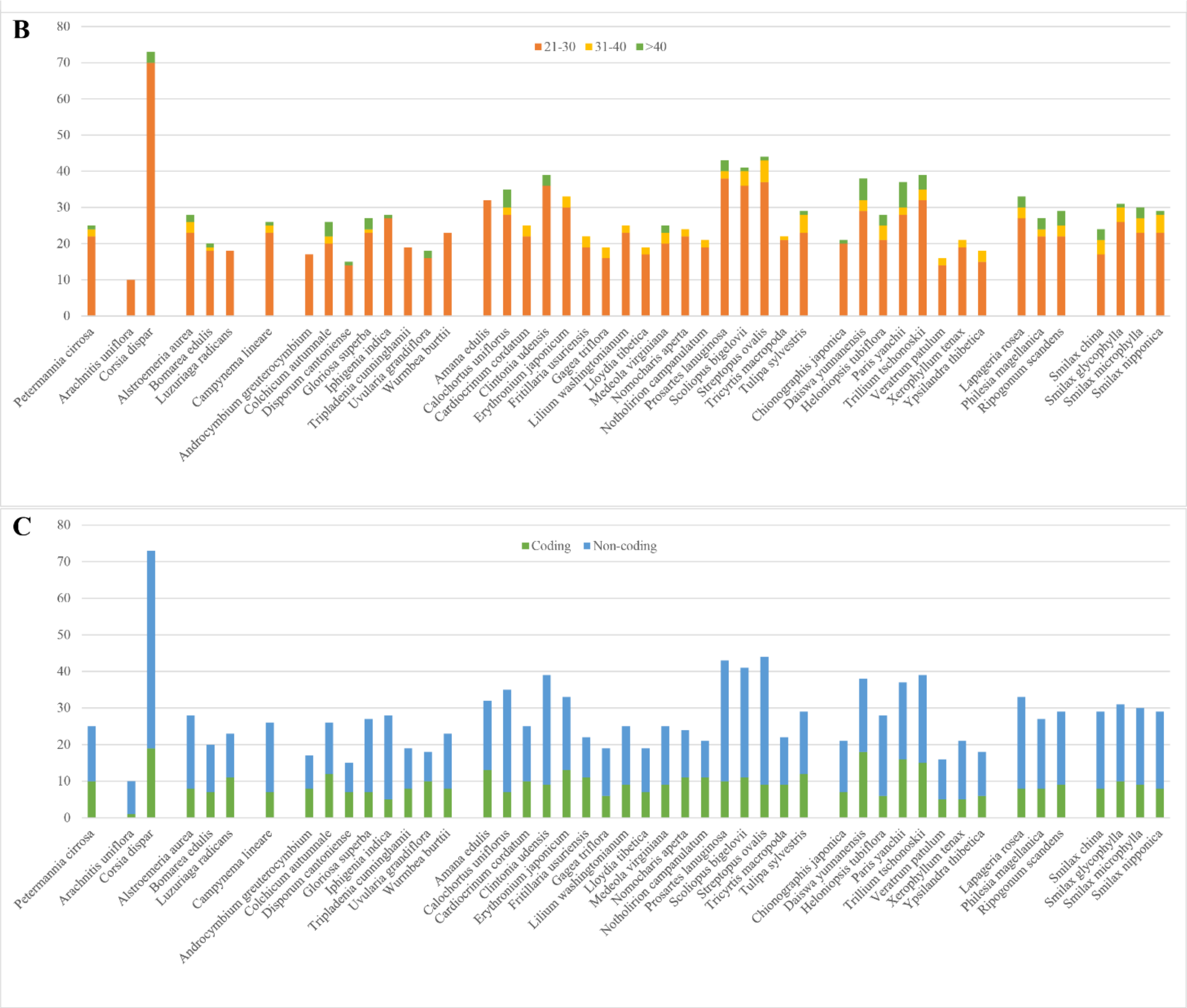


Figure 3 Quantity of long repeat in chloroplast genomes of Liliales.
A. Types of repeat; B. Length of repeat; C. Location of repeat.

The types of tri-, tetra-, penta- and hexanucleotide are not common in Liliales, except Melanthiaceae members of which cpDNAs have a total of 59 records of these four types (AAT/ACAT/ATATC/AAAAT/AAAGAG).

Although Melanthiaceae members possessed a larger number of repeats compared to other Liliales’s families, the repeats in the chloroplast genome do not affect the morphological characteristics of Melanthiaceae, encoded by nuclear genes. The lengths of SSR varied across Liliales taxa (Figure 2B). Most of SSRs (2 591 units) have the lengths of up to 20 bp whereas only 64 SSRs have the lengths from 21 to 30 bp. Although Campynema lineare has the smallest number of SSRs in comparison to other taxa, it contains three SSRs of which the length is over 30 bp. The location of SSRs is mainly in non-coding regions; however, 13.6 % of SSR was found in coding regions (Figure 2C). In Liliales cpDNAs, the coding regions containing SSRs are rpoC1, rpoC2, rpoB, ycf1, cemA, psbD, psbC, psbF, accD, ndhF, ndhG, ndhI, rps2, rps7, rps14, rps19, rps3 and atpB.

Among surveyed species of Liliales, there are 597 records of forward repeats and 700 complement repeats in cpDNA (Figure 3A). The highest number of repeats was detected in Corsia dispar (73 repeats) whereas Arachnitis uniflora only has 10 repeats including eight forward and two complement units. In most cpDNAs, the complement repeat exceeded; however, more forward repeats were recorded in cpDNAs of Corsia dispar, Arachnitis uniflora, Calochortus uniflorus, Clintonia udensis, Medeola virginiana, Prosartes lanuginosa, Scoliopus bigelovii, Streptopus ovalis, Daiswa yunnanensis, Paris yanchii, Trillium tschonoskii, Smilax glycophylla and Smilax nipponica (Figure 3A). The lengths of repeats are mostly shorter than 30 bp (Figure 3B). Only 13 % of repeat has the length over 30 bp. Similar to SSRs, the repeats are located mainly in non-coding regions (Figure 3C). In coding areas, the forward and complement repeats were found in psaA, psaB, rpoC2, ycf1, ycf2, ndhF, ndhI, trnS, trnnfM, and trnG.

In chloroplast genomes, SSRs and repeats are useful information for tracking the evolution of the plants. The SSRs can be used to develop molecular markers for population genetics and identification of plants [17,33]. Additionally, SSR markers can be used for testing the breeding of plants [34,35]. Beside SSRs, the repeats are important factors affecting the structure of cpDNA during the evolutionary history [36,37]. Repeat is also the cause of new repeated generations in cpDNA [38]. In Liliales, Corsia dispar, a mycoheterotrophic species that is in the third stage of cpDNA structural change, has the highest number of repeats and does not have a typical quadripartite structure, suggesting the high impact of repeats on the plastid genome structure of Corsia species. The mycoheterotrophic lifestyle does not require photosynthesis; therefore, genes related to performing and controlling photosynthetic progress are not necessary in the plastid genome of Corsia species. Consequently, various genes in plastid were lost during evolutionary history. In the plastid genome, repeats initiated the deletion of genes as well as non-coding regions. At the present, only one complete plastid genome of Corsia has been reported. Therefore, more samples of Corsia should be sampled to investigate the effectiveness of repeats in the structural change of Corsiaseae of which Arachnitis uniflora has smallest number of repeats and remains the typical quadripartite structure.

4 Conclusions

Complete chloroplast genomes of Liliales were surveyed and the analysis of nucleotide diversity revealed various hotspots among the families of Liliales in both coding and non-coding regions. Additionally, various types of repeats were identified in representative species of Liliales that are crucial sources for further studies on population genetics and development of molecular markers. Last but not least, more samples of Corsiaceae, Campynemataceae and Colchicaceae should be collected to cover the gaps within those families for fulfilling the complete evolutionary history of the chloroplast genome in Liliales.

Acknowledgement

This study was supported by the Foundation for Science and Technology Development of Nguyen Tat Thanh University under the grant 2021.01.211/HĐ-KHCH

References

1.P. F. S. The Angiosperm Phylogeny Group, M. W. Chase, M. J. M. Christenhusz, M. F. Fay, J. W. Byng, W. S. Judd, D. E. Soltis, D. J. Mabberley, A. N. Sennikov, P. S. Soltis, “An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV,” *Bot. J. Linn. Soc.*, vol. 181, no. 1, pp. 1–20, May 2016.

2.J. David L. and G. Bruce, “Corsia dispar D.L.Jones & B.Gray (Corsiaceae), a new species from Australia, and a new combination in Corsia for a New Guinea taxon,” *Austrobaileya*, vol. 7, no. 4, pp. 717–722, 2008.

3.T. J. Givnish *et al.*, “Phylogenomics and historical biogeography of the monocot order Liliales: out of Australia and through Antarctica,” *Cladistics*, vol. 32, no. 6, pp. 581–605, Dec. 2016.

4.J. S. Kim and J.-H. Kim, “Updated molecular phylogenetic analysis, dating and biogeographical history of the lily family (Liliaceae: Liliales),” *Bot. J. Linn. Soc.*, vol. 187, no. 4, pp. 579–593, Jul. 2018.

5.C. Kim, S.-C. Kim, and J.-H. Kim, “Historical Biogeography of Melanthiaceae: A Case of Out-of-North America Through the Bering Land Bridge,” *Front. Plant Sci.*, vol. 10, Apr. 2019.

6.J. Chacón and S. S. Renner, “Assessing model sensitivity in ancestral area reconstruction using L <scp>agrange</scp> : a case study using the Colchicaceae family,” *J. Biogeogr.*, vol. 41, no. 7, pp. 1414–1427, Jul. 2014.

7.Z. Qi *et al.*, “Phylogenetics, character evolution, and distribution patterns of the greenbriers, Smilacaceae (Liliales), a near-cosmopolitan family of monocots,” *Bot. J. Linn. Soc.*, vol. 173, no. 4, pp. 535–548, Dec. 2013.

8.C. Chen *et al.*, “Understanding the formation of Mediterranean–African–Asian disjunctions: evidence for Miocene climate-driven vicariance and recent long-distance dispersal in the Tertiary relict S milax aspera (Smilacaceae),” *New Phytol.*, vol. 204, no. 1, pp. 243–255, Oct. 2014.

9.H. Daniell, C.-S. Lin, M. Yu, and W.-J. Chang, “Chloroplast genomes: diversity, evolution, and applications in genetic engineering,” *Genome Biol.*, vol. 17, no. 1, p. 134, Dec. 2016.

10. E. J. Carpenter *et al.*, “Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP),” *Gigascience*, vol. 8, no. 10, Oct. 2019.

11. S. Cheng *et al.*, “10KP: A phylodiverse genome sequencing plan,” *Gigascience*, vol. 7, no. 3, Mar. 2018.

12. M. A. Gitzendanner, P. S. Soltis, G. K.-S. Wong, B. R. Ruhfel, and D. E. Soltis, “Plastid phylogenomic analysis of green plants: A billion years of evolutionary history,” *Am. J. Bot.*, vol. 105, no. 3, pp. 291–301, Mar. 2018.

13. H. Robert J, *Molecular Markers in Plants*. Oxford, UK: Blackwell Publishing Ltd., 2012.

14. K. Semagn, å Bjørnstad, and M. N. Ndjiondjop, “An overview of molecular marker methods for plants,” *African J. Biotechnol.*, vol. 5, no. 25, pp. 2540–2568, 2006.

15. T. N. Vu *et al.*, “Molecular markers for analysis of plant genetic diversity,” *Vietnam J. Biotechnol.*, vol. 18, no. 4, pp. 589–608, May 2021.

16. J. Hyun, H. D. K. Do, J. Jung, and J.-H. Kim, “Development of molecular markers for invasive alien plants in Korea: a case study of a toxic weed, Cenchrus longispinus L., based on next generation sequencing data,” *PeerJ*, vol. 7, p. e7965, Nov. 2019.

17. M. L. C. Vieira, L. Santini, A. L. Diniz, and C. de F. Munhoz, “Microsatellite markers: what they mean and why they are so useful,” *Genet. Mol. Biol.*, vol. 39, no. 3, pp. 312–328, Aug. 2016.

18. H. D. K. Do, C. Kim, M. W. Chase, and J. Kim, “Implications of plastome evolution in the true lilies (monocot order Liliales),” *Mol. Phylogenet. Evol.*, vol. 148, p. 106818, Jul. 2020.

19. H. D. K. Do, J. S. Kim, and J.-H. Kim, “Comparative genomics of four Liliales families inferred from the complete chloroplast genome sequence of Veratrum patulum O. Loes. (Melanthiaceae),” *Gene*, vol. 530, no. 2, 2013.

20. S.-C. Kim, J. S. Kim, and J.-H. Kim, “Insight into infrageneric circumscription through complete chloroplast genome sequences of two Trillium species,” *AoB Plants*, vol. 8, p. plw015, 2016.

21. Y. Song *et al.*, “Chloroplast Genomic Resource of Paris for Species Discrimination,” *Sci. Rep.*, vol. 7, no. 1, p. 3427, Dec. 2017.

22. H. D. K. Do and J.-H. Kim, “The implication of plastid transcriptome analysis in petaloid monocotyledons: A case study of Lilium lancifolium (Liliaceae, Liliales),” *Sci. Rep.*, vol. 9, no. 1, 2019.

23. T. Braukmann and S. Stefanović, “Plastid genome evolution in mycoheterotrophic Ericaceae,” *Plant Mol. Biol.*, vol. 79, no. 1–2, pp. 5–20, May 2012.

24. S. W. Graham, V. K. Y. Lam, and V. S. F. T. Merckx, “Plastomes on the edge: the evolutionary breakdown of mycoheterotroph plastid genomes,” *New Phytol.*, vol. 214, no. 1, pp. 48–55, Apr. 2017.

25. C. F. Barrett and J. I. Davis, “The plastid genome of the mycoheterotrophic Corallorhiza striata (Orchidaceae) is in the relatively early stages of degradation,” *Am. J. Bot.*, vol. 99, no. 9, pp. 1513–1523, Sep. 2012.

26. R. S. Millen *et al.*, “Many Parallel Losses of infA from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus,” *Plant Cell*, vol. 13, no. 3, pp. 645–658, Mar. 2001.

27. A. A. Alqahtani and R. K. Jansen, “The evolutionary fate of rpl32 and rps16 losses in the Euphorbia schimperi (Euphorbiaceae) plastome,” *Sci. Rep.*, vol. 11, no. 1, p. 7466, Dec. 2021.

28. M. Ueda *et al.*, “Substitution of the Gene for Chloroplast RPS16 Was Assisted by Generation of a Dual Targeting Signal,” *Mol. Biol. Evol.*, vol. 25, no. 8, pp. 1566–1575, Apr. 2008.

29. S.-C. Kim, J.-W. Lee, and B.-K. Choi, “Seven Complete Chloroplast Genomes from Symplocos: Genome Organization and Comparative Analysis,” *Forests*, vol. 12, no. 5, p. 608, May 2021.

30. Q. Liu, X. Li, M. Li, W. Xu, T. Schwarzacher, and J. S. Heslop-Harrison, “Comparative chloroplast genome

analyses of Avena: insights into evolutionary dynamics and phylogeny," *BMC Plant Biol.*, vol. 20, no. 1, p. 406, Dec. 2020.

31. A. W. Gichira, S. Avoga, Z. Li, G. Hu, Q. Wang, and J. Chen, "Comparative genomics of 11 complete chloroplast genomes of Senecioneae (Asteraceae) species: DNA barcodes and phylogenetics," *Bot. Stud.*, vol. 60, no. 1, p. 17, Dec. 2019.

32. W. Dong, J. Liu, J. Yu, L. Wang, and S. Zhou, "Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding," *PLoS One*, vol. 7, no. 4, p. e35071, Apr. 2012.

33. C. Li, Y. Zheng, and P. Huang, "Molecular markers from the chloroplast genome of rose provide a complementary tool for variety discrimination and profiling," *Sci. Rep.*, vol. 10, no. 1, p. 12188, Dec. 2020.

34. A.-H. Yang, J.-J. Zhang, X.-H. Yao, and H.-W. Huang, "Chloroplast microsatellite markers in Liriodendron tulipifera (Magnoliaceae) and cross-species amplification in L. chinense," *Am. J. Bot.*, vol. 98, no. 5, pp. e123–e126, May 2011.

35. M. Yang *et al.*, "Genetic linkage maps for Asian and American lotus constructed using novel SSR markers derived from the genome of sequenced cultivar," *BMC Genomics*, vol. 13, no. 1, p. 653, Dec. 2012.

36. F. Yue, L. Cui, C. W. DePamphilis, B. M. Moret, and J. Tang, "Gene rearrangement analysis and ancestral order inference from chloroplast genomes with inverted repeat," *BMC Genomics*, vol. 9, no. S1, p. S25, Mar. 2008.

37. J. D. Palmer, B. Osorio, J. Aldrich, and W. F. Thompson, "Chloroplast DNA evolution among legumes: Loss of a large inverted repeat occurred prior to other sequence rearrangements," *Curr. Genet.*, vol. 11, no. 4, pp. 275–286, Jan. 1987.

38. H. D. K. Do and J.-H. Kim, "A dynamic tandem repeat in monocotyledons inferred from a comparative analysis of chloroplast genomes in melanthiaceae," *Front. Plant Sci.*, vol. 8, 2017.

Phân tích đa dạng di truyền bộ gen lục lạp ở bộ Loa kèn

Đỗ Hoàng Đăng Khoa

Viện Kĩ thuật Công nghệ cao, Đại học Nguyễn Tất Thành
dhdkhoa@ntt.edu.vn

Tóm tắt Bộ Loa kèn là một bộ thực vật một lá mầm và bao gồm cả loài thực vật tự dưỡng và dị dưỡng cộng sinh với nấm; phân bố rộng khắp hoặc cục bộ tại một số vùng nhất định. Trong nghiên cứu này, da dạng di truyền của bộ Loa kèn được khảo sát thông qua phân tích đa dạng nucleotide và thành phần các loại trình tự lặp trong bộ gen lục lạp. Kết quả phân tích đa dạng nucleotide cho thấy rất nhiều trình tự có biến động cao trong vùng trình tự đơn lớn (LSC) và vùng trình tự đơn nhỏ (SSC) trong khi vùng trình tự lặp đảo thì có mức biến động thấp. Mặc dù từng họ trong bộ Loa kèn có các trình tự biến động đặc trưng nhưng vùng trình tự *rps15-ycf1* có biến động cao được tìm thấy hầu hết trong các bộ gen lục lạp. Trong bộ gen lục lạp của bộ Loa kèn, các trình tự lặp đơn giản (SSR) loại nucleotide đơn là loại phổ biến và hầu hết các trình tự SSR nằm ở vùng không mã hóa. Tương tự như vậy, các trình tự lặp dài cũng chủ yếu được tìm thấy ở vùng không mã hóa. Ngoài ra, trình tự lặp đảo là trình tự lặp phổ biến hơn so với trình tự lặp liên tục trong bộ gen lục lạp của bộ Loa kèn. Số lượng trình tự lặp dài cao nhất được tìm thấy trong bộ gen lục lạp của loài *Corsia dispar* trong khi trình tự lặp đơn giản được xác định nhiều nhất trong loài *Smilax china*. Các kết quả nghiên cứu đa dạng nucleotide và trình tự lặp sẽ cung cấp các thông tin nền tảng cho các nghiên cứu tiếp theo trong lĩnh vực di truyền quần thể, chỉ thị phân tử và lịch sử tiến hóa của bộ Loa kèn.

Từ khóa bộ Loa kèn, bộ gen lục lạp, đa dạng nucleotide, giá trị Pi, trình tự lặp.