Plant Diversity of Central Asia

PDCA ISSN 2181-2403
PDCA ISSN 2181-239X

Complete chloroplast genomes of ten species from subgenus *Allium* (*Allium*, Amaryllidaceae)

Abdukhalimkhon Munavvarov[1,2,3], Ziyoviddin Yusupov[1,3*], Ibrokhimjon Ergashov[1,2,3], Komiljon Sh. Tojibaev[3], Tao Deng[1*], Hang Sun[1]

[1]*CAS Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China*

[2]*University of Chinese Academy of Sciences, Beijing, China*

[3]*International Joint Lab for Molecular Phylogeny and Biogeography, Institute of Botany, Academy of Sciences of Uzbekistan, Tashkent, 100125, Uzbekistan*

✉ yusupov@mail.kib.ac.cn; dengtao@mail.kib.ac.cn

ABSTRACT

Subgenus *Allium* is the largest group within the genus *Allium* (Amaryllidaceae) which detailed phylogenetic classification is still missing despite intensive phylogenetic research. Reconstruction of the subgenus phylogeny is impossible without proper coverage of the subgenus members representing Central Asia, one of the centers of genus (and subgenus) diversity. We sequenced whole chloroplast genomes of 10 species from subgenus *Allium*, mainly from Central Asia, which, together with 7 species available at NCBI provided us with a representative subgenus sampling for making inferences about the whole plastome-based phylogeny of subgenus *Allium*. The phylogenetic relationships among the representatives of subgenus *Allium* in our phylogenetic tree not only contradicted the current sectional classification within a subgenus, but even doubted the validity of the current subgenera recognized in genus *Allium*. The detected numerous mismatches between the species positions in the phylogenetic tree and the current species taxonomic position may suggest that the importance of the morphological traits used previously for defining sections in subgenus *Allium* needs reconsideration. Analysis of single nucleotide polymorphism in the studied genomes identified the genes *ycf1*, *ndhF*, *rpoC2*, *matK* and *ndhD* as the most variable regions (number of SNPs ≥ 30).

Key words: phylogeny; plastome; nucleotide polymorphism; pseudogene

Introduction

The genus *Allium* L., a complicated polymorphous genus, is one of the largest monocotyledonous genera containing more than 900 species (Herden et al. 2016). They are characterized by bulbs enclosed in membranous (sometimes fibrous) tunics, free or almost free tepals (Friesen et al. 2006), and the number of species in the genus is still growing due to the new species discoveries (Friesen et al. 2021b; Bartolucci et al. 2022; Xie et al. 2022). The latest and widely accepted classification of *Allium* by Friesen et al. (2006) was based on ITS sequences. However, although the ITS-based phylogenetic tree of Friesen et al. (2006) in general agrees well with the morphological

http://doi.org/10.54981/PDCA/vol1_iss2/a3
Received: 24 May 2022; Accepted: 24 August 2022 Academic editor: Hyeok Jae Choi
Plant Diversity of Central Asia 2 (2022) 67–81

data, because of the small size of ITS (less than 700 bp), the resolution of the ITS-based phylogenetic analysis is insufficient for resolving positions of many clades. Similarly, the chloroplast DNA (cpDNA) fragments (e.g. *matK*, *rps16*, and *trnL–trnF*) widely used to infer the phylogeny of *Allium* (Li et al. 2010; Herden et al. 2016; Li et al. 2016; Hauenschild et al. 2017), suffer the same problem. Fortunately, recent technological developments made possible routine sequencing of the whole chloroplast genomes, greatly increasing resolution power in phylogenetic studies. Because of the conserved structure, small effective population size, lack of recombination and usually uniparental inheritance (Daniell et al. 2016), the complete cp sequences have been widely used for phylogenetic reconstructions (Jansen et al. 2007; Moore et al. 2010; Malé et al. 2014). Since 2017 when Lee et al. (2017) reported the first complete chloroplast genome of *Allium*, the number of published cp genomes of *Allium* species grew rapidly (Huo et al. 2019; Xie et al. 2019; Yusupov et al. 2021). The plastome-based phylogenetic studies have been performed at the sectional (Xie et al. 2019; Friesen et al. 2021a), subgenus (Yang et al. 2020), and the whole genus level (Xie et al. 2020a). In the latter study, from analyses of complete chloroplast genomes representing 39 species of *Allium*, 125 species from 19 other monocot families, and three fossil calibration points, the authors inferred that *Allium* originated in the late Eocene (ca. 42 Mya), and diverged into three major evolutionary lineages from the early Eocene to the middle Miocene. Despite these advances in understanding *Allium* evolution, relationships among species of *Allium* are still not well resolved, especially at subgeneric level, and more plastimes of *Allium* species are required for reconstruction of the genus phylogeny. This especially applies to subgenus *Allium*, the largest subgenus within *Allium* comprising more than 375 species and 35 subspecies (Khassanov 2018). Although this subgenus was a subject of many phylogenetic studies, a detailed phylogenetic classification of this taxonomically complex subgenus is still missing (Khassanov 2018). Reconstruction of the subgenus phylogeny is impossible without proper coverage of the subgenus members representing Central Asia, one of the centers of genus (and subgenus) diversity (Xie et al. 2020b).

Currently, only a few plastomes of species from subgenus *Allium* representing Central Asian region are available at NCBI. In this study, we examined the complete chloroplast genomes of ten species representing 6 sections from subgenus *Allium*. Eight species, *A. filidens* Regel (sect. *Allium*), *A. michaelis* F.O. Khass. & Tojibaev and *A. michaelis* F.O. Khass. & Tojibaev (sect. *Brivedentia*), *A.caesium* Schrenk (sect. Caerulea), *A. confragosum* Vved. and *A. sabulosum* Steven ex Bunge (sect. *Eremoprasum*), *A. haneltii* F.O.Khass. & R.M.Fritsch (sect. *Haneltia*), *A.pallasi* Murray (two samples) (sect. *Pallasia*), were collected from Central Asia, and the remaining two species, *A. tanguticum* Regel (sect. *Pallasia*) and *A. glomeratum* Prokh. (sect. *Caerulea*) were collected from China. The main purposes of this study were to (1) explore and compare the organization of the complete cp genomes of these species, (2) to analyze the subgenus *Allium* phylogeny using newely aquired and existing sequences; and (3) identify genetic markers within the plostome useful for studying *Allium* phylogeny.

Materials and Methods

Taxon sampling and DNA extraction

Fresh leaves of all studied species were collected in natural populations, and the information about collection places is presented in Table 1. The flower morphological characters, filaments and bulbs of six species representing six diferent sections of *Allium* are illustrated in Fig. 1. Voucher specimens were deposited in KUN and TASH. The total genomic DNA was extracted from leaf tissues with the CTAB method (Doyle 1991).

Chloroplast plastomes sequencing, assembling, and annotation

DNA was sheared to construct a 350 bp (insert size) sequence library using the Genomic DNA Sample Prep Kit (Illumina) according to the manufacturer's protocol and then sequenced using 150 paired-end reads on the Illumina HiSeq 4000 at Beijing Novogene Bioinformatics Technology Co., Ltd, Beijing, China.

For raw data processing, we used the Next Generation Sequencing (NGS) QC Tool Kit with default settings (Patel & Jain 2012), and the resulting clean reads were assembled in NovoPlasty v.3.8.3 (Dierckxsens et al. 2017) with k-mer sizes ranging from 21 to 39 and the seed sequence of *rbcL* from *Allium fistulosum* (Yusupov et al. 2019). Gene annotation was performed in Geneious v.10.0.2 using *A. ferganicum* (NC_057975.1) as the reference,

Table 1. Vouchers, collection places and location coordinates for the sequenced species

Voucher	Species Name	Region	Latitude	Longitude
ZD18475	*Allium caesium*	Namangan district, Uzbekistan	41.259216	71.896181
ZD0505	*Allium confragosum*	Khujand region, Tajikistan	40.284675	69.428585
ZD0513	*Allium filidens*	Khujand region, Tajikistan	40.334085	69.619864
ZD0323	*Allium haneltii*	Namangan district, Uzbekistan	40.958604	70.871283
ZD0300	*Allium michaelis*	Namangan district, Uzbekistan	40.907722	70.909296
ZD0522	*Allium nikolai*	Surkhandarya district, Uzbekistan	38.2553	67.4986
ZD1053	*Allium pallasii*	Almaty district, Kazakhstan	43.855587	75.470149
ZD1102	*Allium pallasii*	Almaty district, Kazakhstan	43.472233	78.637764
ZD0422	*Allium sabulosum*	Kyzylkum region, Uzbekistan	42.0192	61.0244
KBG17366	*Allium tanguticum*	Hanyan, Qinghai province, China		
KBG17350	*Allium glomeratum*	Wulumuqi, Xinjiang Autonomous Region, China		

and start and stop codons and intron/exon boundaries for protein-coding genes were checked manually (Kearse et al. 2012). The circular map of these genomes was drawn using OrganellarGenomeDRAW (OGDRAW) (Greiner et al. 2019).

Simple sequence repeats (SSRs)

The identification of simple sequence repeats (SSRs) was performed using MIcroSAtellite (MISA) web tool (Beier et al. 2017). The search conditions for SSRs were set to isolate perfect mono-, di-, tri-, tetra-, penta- and hexanucleotide motifs with minimum 10, 5, 4, 3, 3 and 3 repeats, respectively. The REPuter program (Kurtz et al. 2001) was used to identify repeats: forward, reverse, palindrome, and complement sequences. The following settings for repeat identification were used: (1) hamming distance equal to 3; (2) minimal repeat size set to 30 bp; and (3) maximum computed repeats set to 90 bp. Tandem repeats were identified using the web-based Tandem Repeats Finder (https://tandem.bu.edu/trf/trf.html). The alignment parameters match, mismatch, and indels were set to 2, 7 and 7, respectively, and the minimum alignment score and maximum period size were set to 80 and 500, respectively. Other parameters were set to default values.

Comparative analysis

MAFFT (Katoh & Standley 2013) was used for multiple alignment of the plastomes. Then the intergenic spacer regions, intronic regions, and protein-coding regions were extracted in Geneious v.10.0.2 (Kearse et al. 2012), and nucleotide variability (*Pi*) was calculated for each region using DnaSP v.5.1 software (Librado & Rozas 2009). Single nucleotide polymorphisms (SNPs) were assessed for 10 newly sequenced Allium cp genomes by Geneious v.10.0.2 (Kearse et al. 2012). SNP variability was assessed at several levels: 1) Total cp genome; 2) LSC; 3) SSC; 4) IR; 5) LSC-SSC; 6) LSC-IR; and 7) SSC-IR. The IR/SSC and IR/LSC boundary analyses among *Allium* species were performed in Geneious v.10.0.2 (Kearse et al. 2012). Pairwise chloroplast genomic alignment among ten species was compared by mVISTA in Shuffle-LAGAN mode (Frazer et al. 2004), and *Allium cepa* (MK335926) was used as a reference.

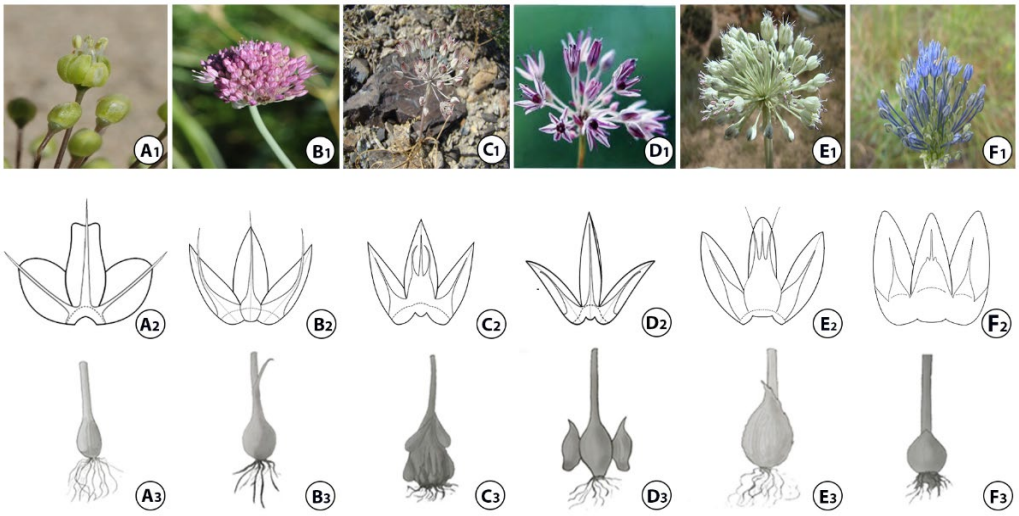

Fig. 1. The flower morphological characters, filaments and bulbs of the subgenus *Allium* species. **(A)** *A. sabulosum* (sec. *Eremoprasum*, photo by V. Epiktetov); **(B)** *A. pallasii* (sec. *Pallasia*, photo by V. Epiktetov); **(C)** *A. michaelis* (sec. *Brivedentia*, photo by G. Lazkov); **(D)** *A. haneltii* (sec. *Haneltia*, photo by F, Khassanov); **(E)** *A. filidens* (sec. *Allium* photo by G. Lazkov); **(F)** *A. caesium* (sec. *Coerulea*, photo by L. Valdshmit).

Phylogenetic analysis

The dataset for phylogenetic analysis included newly sequenced cp genomes of 10 *Allium* species plus 53 available *Allium* and 14 non-*Allium* sequences from NCBI. The sequences were aligned using MAFFT (Katoh & Standley 2013). The phylogenetic tree was constructed based on 77 protein coding genes (CDS). After extraction, the protein coding gene sequences were aligned using MUSCLE v.3.6 (Edgar 2004) and then concatenated as a super locus of coding DNA sequences using SequenceMatrix v.1.8 (Vaidya et al. 2011). Maximum parsimony (MP) was performed using PAUP* version 4.10 (Swofford 2003). All characters were equally weighted, gaps were treated as missing and character states were treated as unordered. A heuristic search was performed with TBR branch swapping and the Multrees option, and random stepwise addition with 1000 replications. All analyses used the best-fitting models of nucleotide substitutions selected in jModelTest v.2.1.4 (Darriba 2012) under the Akaike information criterion (AIC). Maximum likelihood (ML) analyses were conducted using RAxML v.8.0 (Stamatakis 2014) based on the best-fit GTR + G model and 1000 bootstrap replicates. Bayesian analyses were performed with MrBayes v.3.2 (Ronquist

2011). iTOL v.4.2.3. (https://itol.embl.de/) was used to visualize phylogeny.

Results

Plastome of Allium species

The characteristics of 11 newly sequenced chloroplast genomes representing 10 species are presented in Table 1. Total plastome size ranged from 151,465 bp (*Allium pallasii*) to 153,406 bp (*Allium glomeratum*). The chloroplast genome structure of subgenus *Allium* species is represented in Fig. 2 and Suppl. Table 2. The structure of the chloroplast genomes of *Allium* is conserved for overall size and the order and size of each gene and intergenic region (Suppl. Table 1 and Fig. 3). The cp genome consisted of a pair of IR regions (25,698–26,540 bp) separated by the LSC (80,412– 82,720 bp) and SSC (13,504–18,140 bp) regions. The GC content ranged 36.6 – 37.8%. A total of 133-135 genes were found in the studied *Allium* species, of which 85-89 are protein-coding genes, 38 are tRNA genes, and 8 are rRNA genes (Suppl. Table 1). Ten genes, namely *rps16*, *rps2*, *accD*, *rpl20*, *infA*, *rpl22*, *rpl2*, *rpl23*, *rpl32*, and *ndh* were present as pseudogenes or lost their function completely (Fig. 4).

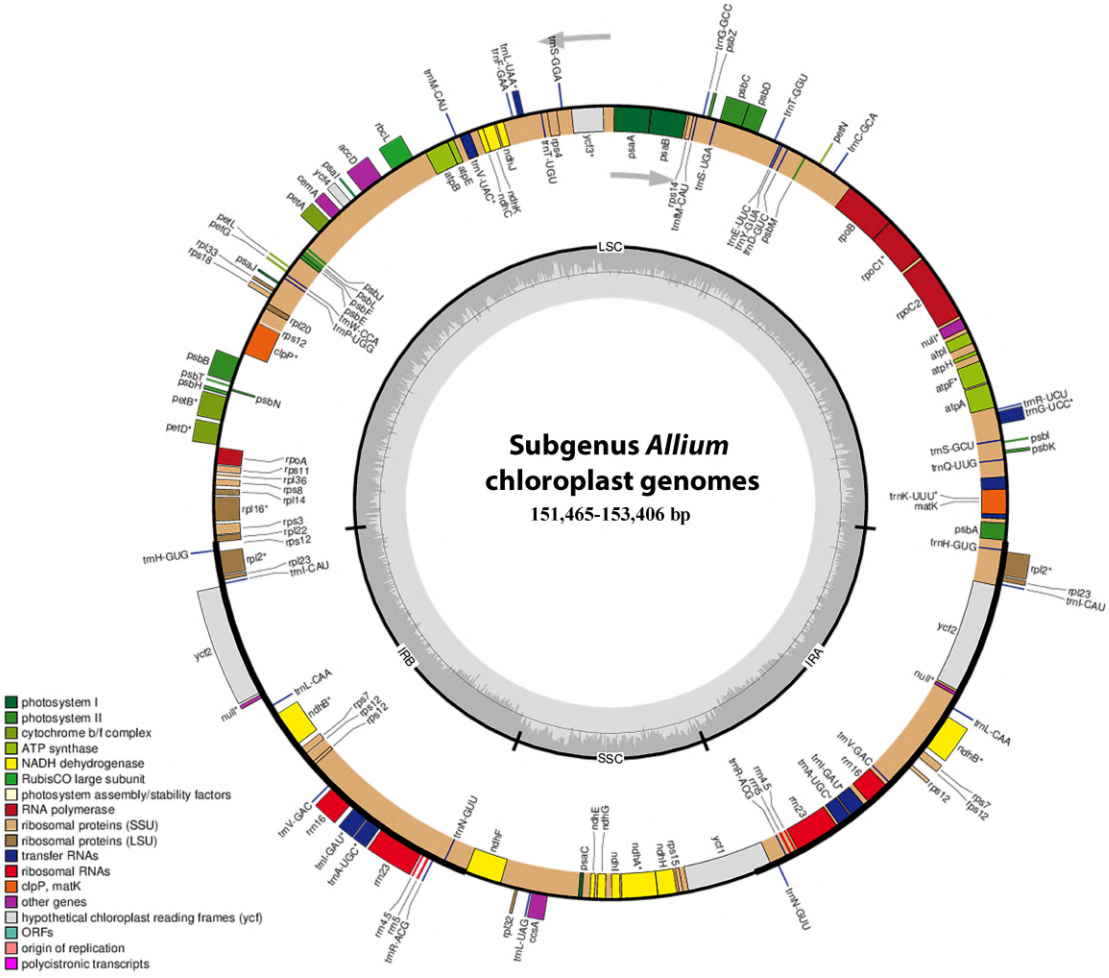

Fig. 2. Chloroplast genome structure of subgenus *Allium* species. Genes shown outside the circles are transcribed clockwise, while those drawn inside are transcribed counter clockwise. Genes are color-coded according to their functional group. The positions of the large single-copy (LSC), small single-copy (SSC), and inverted repeat (IR: IRA and IRB) regions are shown in the inner circles.

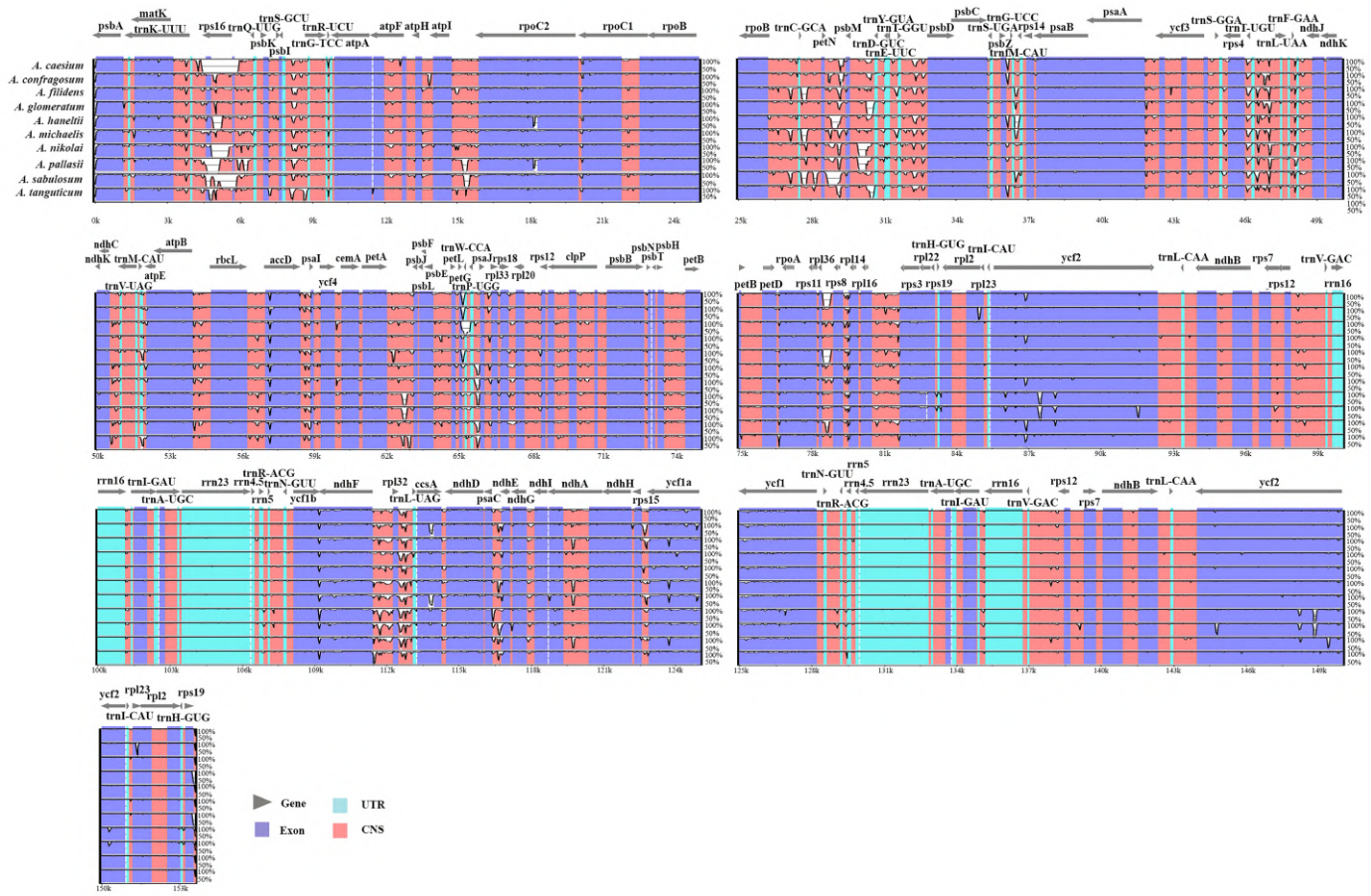

Fig. 3. VISTA-based identity plot for 10 *Allium* species using *A. cepa* as a reference.

Analysis of SSRs and single nucleotide polymorphisms

The total number of SSRs in the chloroplast genomes of *Allium* ranged from 67 (*A. michaelis*) to 77 (*A. pallasii*). Six types of SSRs were found. The largest group of SSRs were mononucleotide repeats ranging from 49 (*A. pallasii*) to 67 (*A. glomeratum* and *A. schoenoprasoides*), making up more than half of the detected SSRs. Di-nucleotides (especially AT) were the second most predominant, varying from 14 to 19 (*A. confragosum* and *A. michaelis*, respectively). Tetranucleotide repeats were the next most common SSRs (112 repeats overall), followed by trinucleotide (25) and pentanucleotide (17) categories (Fig. 5). Hexanucleotide repeats were found only in *A. filidens*, *A. tanguticum* and *A. pallasii*, one hexanucleotide repeat in each species.

The majority of SSRs were in the LSC region, ranging from 52 (A. pallasii) to 77 (*A. michaelis*). In total, 209 and 46 SSRs werefound in the SSC and IR regions, respectively (Fig. 5). Three types of repeats, comprising dispersed, tandem, and palindromic repeats were analyzed in the newly sequenced *Allium* species. Of the 593 repeats found, dispersed repeats were the most frequent (240), followed by palindromic repeats (195), and tandem repeats (158) (Fig. 5). The species with the largest number of repeats was *A. sabulosum* which had 42 dispersed, 26 palindromic, and 17 tandem repeats. The lowest number of repeats was found in *A. pallasii* (7 tandem repeats and 10 dispersed repeats). Comparisons of plastomes revealed 2,319, 1,541, 604 and 186 SNPs in the total genome sequence, LSC, SSC and IR regions, respectively.

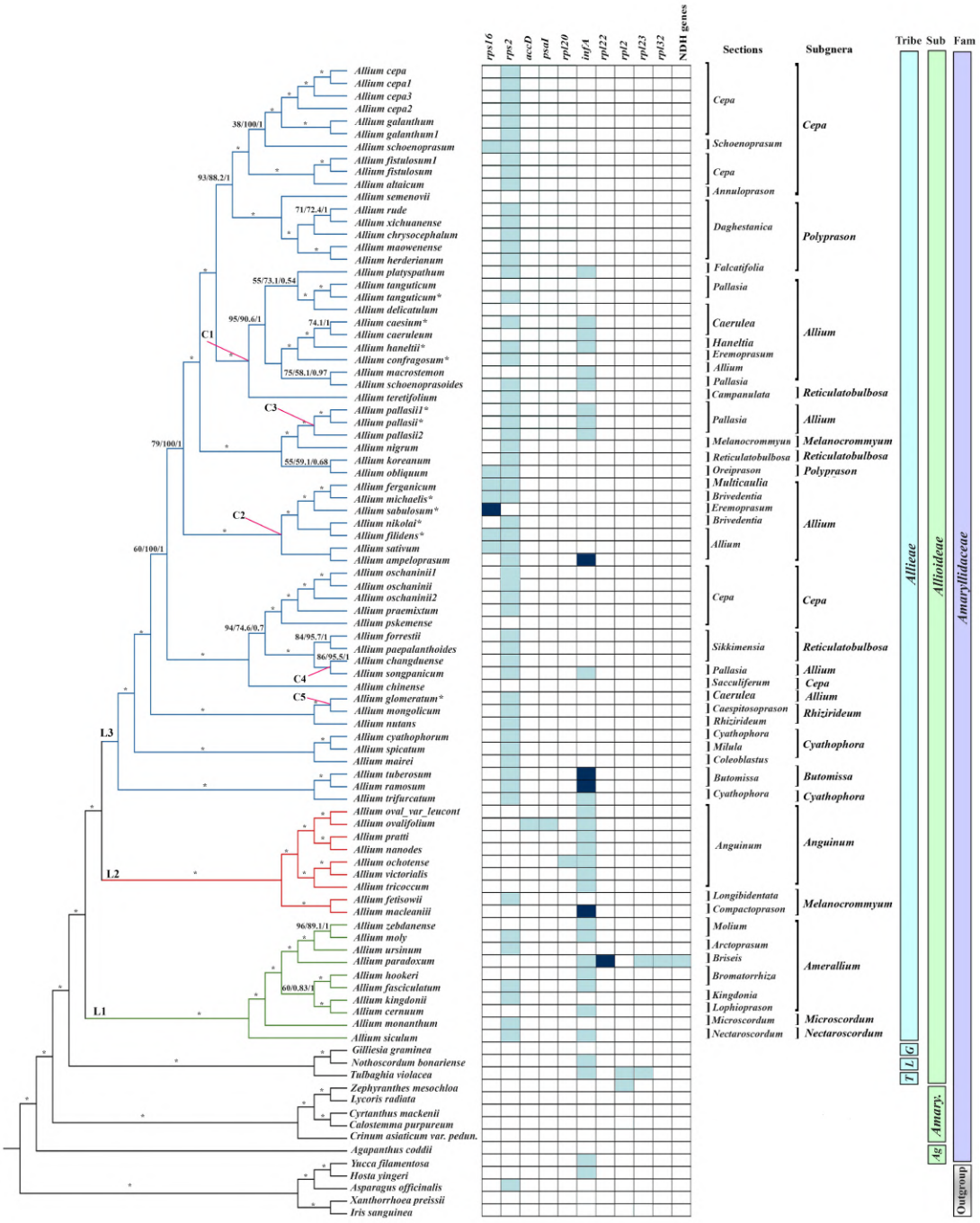

Fig. 4. Phylogenetic tree showing relationships of 92 chloroplast genomes from Allioideae and related taxa inferred using 77 protein-coding regions, and expression of gene losses/pseudogenisation. Newly sequenced species are indicated by an asterisc. Numbers above the branches follow the order: BS (bootstrap support)/MP (bootstrap values)/ PP (posterior probability); values of 100% are denoted by an asterisc. The evolutionary lines 1-3 are in green, red and blue, respectively. In the table, light blue indicates pseudogenisation and dark blue - loss of the gene. C1 – C5 indicate Clades 1-5 respectively. Abbreviations: Sub: Subfamily; Fam: Family; Amary: Amaryllidoideae; T: Tulbaghieae; L: Leucocoryneae; G: Gilliesieae; Ag: Agapanthoideae.

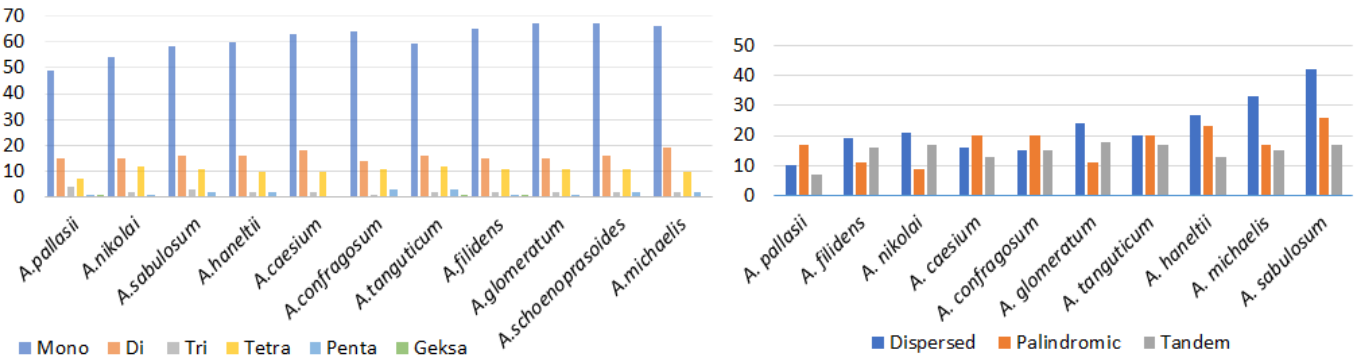

Fig. 5. SSR locus analysis of 11 *Allium* chloroplast genomes: SSRs by number of nucleotides (left) and by types of SSRs (dispersed, palindromic and tandem repeats) (right).

IR contraction/expansion and nucleotide diversity

We compared the IR/SSC and IR/LSC boundaries of the 10 *Allium* species (Fig. 6). Although the genome structure and the number and sequence of genes were highly conserved among *Allium* species, the contraction and expansion of IR boundaries was observed in terms of structure and size. The gene *rpl22* was located at the LSC-IRA boundary in the majority of the analyzed sequences; in only three sequences (*A. glomeratum*, *A. nikolai*, and *A. tanguticum*) rpl22 was located in the LSC region. The LSC-IRB junction was located adjacent to genes *rps19* and *psbA*. In *A. glomeratum*, *A. nikolai* and *A. tanguticum* the gene *rps19* expanded into the LSC region by 3 bp, 5 bp, and 3 bp, respectively. The gene *ycf1* was present in the cp genome by duplicates having different lengths. The shorter one ranged from 1,071 to 1,290 bp, and the longer one ranged from 5,271 to 5,370 bp. The SSC-IRB junctions were embedded in the *ycf1* gene. The length of *ycf1* in the SSC region varied from 4,018 to 4,300 bp, and the overlap between *ycf1* and the IRB region varied from 1,071 to 1,268 bp. The *ndhF* gene flanked the junction between SSC and IRA in *A. caesium* and *A. pallasii*, but the *ndhF* gene was located at a region with a length of 5-67 bp away from SSC/IRA boundary in other species.

The analysis of nucleotude diversity revealed that intergenic spacer regions on average had higherpolymorphism (0.0326) than protein-coding sequences (0.0298) or intronic regions (0.0249). Fig. 7 shows polymorphism in 46 regions with nucleotide diversity > 0.02 bp, excluding loci smaller than 200 bp. Among them, seven regions, *trnC-GCA-petN*, *rps15-ycf1*, *ndhA*, *ccsA-ndhD*, *psaC-ndhE*, *ndhF-rpl32*, and *rpl32-trnL-UAG*, showed the highest polymorphism (nucleotide diversity > 0.04).

Phylogenetic analysis

In the CDS-based phylogenetic tree showing phylogenetic relationships among 93 species from Allioideae and related taxa, the 18 species from subgenus *Allium* formed several clades (Fig. 4). The largest clade includes *A. schoenoprasoides* and *A. tanguticum* from sect. *Pallasia*, *A. delicatulum*, *A. caesium* and *A. caeruleum* from sect. *Coerulea*, *A. haneltii* from sect. *Haneltia*, *A. confragosum* from section *Eremoprasum*, and *A. macrostemon* from sect. *Allium* (Fig. 4, C1). The second by size clade includes *A. filidens*, *A. sativum* and *A. ampeloprasum* from section *Allium*, *A. michaelis* and *A. nikolai* from section *Brivedentia*, *A. ferganicum* from sect. *Multicaulia* and *A. sabulosum* from sect. *Eremoprasum* (Fig 4. C2). The three species, *A. pallasii* and *A. songpanicum* from sect. *Pallasia* and *A. glomeratum* from sect. *Coerulea* were clustered together with the species from other subgenera (Fig. 4, C3, C4 and C5).

Fig. 6. Comparison of the junctions between the LSC, SSC and IR regions among 10 *Allium* chloroplast genomes

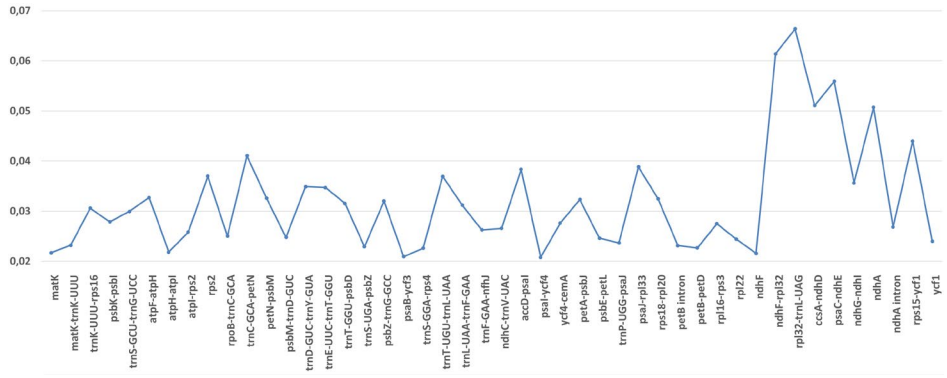

Fig. 7. Plastome regions' polymorphism. Regions with length <200 bp, and nucleotide diversity <0.02 are not shown.

Discussion

The ten newly sequenced species belonged to six sections of subgenus *Allium*, namely, sect. *Allium* (*A. filidens*), sect. *Brivedentia* (*A. michaelis* and *A. nikolai*), sect. *Coerulea* (*A. caesium* and *A. glomeratum*), sect. *Eremoprasum* (*A. confragosum* and *A. sabulosum*), sect. *Haneltia* (*A. haneltii*) and sect. *Pallasia* (*A. pallasii* and *A. tanguticum*). Of these, the last two species from sect. *Pallasia* were the same with those analyzed by Xie et al. (2020b) using, like the present study, the whole chloroplast genome. Positions of these two species in the phylogenetic tree in our study and Xie et al. (2020b) were similar in that the latter two species representing the same section were not clustered together but scattered along the tree. Similar scattering of closely related species representing the same section was found in our phylogenetic tree for *A. caesium* and *A. glomeratum* from sect. *Coerulea* and *A. confragosum* and *A. sabulosum* from sect. *Eremoprasum*. Moreover, similar to Xie et al. (2020b), the other members of their respective clades were the species from not only different sections but even subgenera.

Our study revealed numerous mismatches between the species positions in the phylogentic tree and the current species taxonomic position. Some of the observed mismatches between the species positions in the phylogenetic tree and morphology-based classification can be probably explained by a closer look at the species morphological differences. For example, *A. confragosum* was placed in the section *Eremoprasum* based on morphology (Khassanov 2018); but in our phylogenetic tree *A. confragosum* was not a member of the same clade with the type species of sect. *Eremoprasum A. sabulosum*. Instead, it was clustered together with *A. caesium*, *A. haneltii* and *A. caeruleum* from different sections. Khassanov (2018) placed *A. confragosum* and *A. sabulosum* into the same section because they share such traits as tripartite inner filaments and outer tunics splitting at base with crests. However, *A. confragosum* differs from *A. sabulosum* by having persistent spathe. In terms of the persistent spathe, *A. confragosum* resembles *A. haneltii*, *A. caeruleum* and *A. caesium*. The latter fact may suggest that the importance of the morphological traits used previously for defining sections in subgenus *Allium* needs reconsideration.

Our phylogenetic tree is the first that has a representative number of species (18) for making the first conclusions about the whole plastome-based phylogeny of *subgenus Allium*. The phylogenetic relationships among the representatives of subgenus *Allium* in our phylogenetic tree not only contradicted the current sectional classification within a subgenus, but even doubted validity of the current subgenera recognized in genus *Allium*. Of the five clades that contained representatives of subgenus *Allium*, in only one clade all the members were from the subgenus *Allium*. In the four clades, other clade members represented such subgenera as *Polyprason*, *Reticulatobulbosa* and *Cepa*.

The other existing molecular-based phylogeny of *Allium* is one utilizing ITS sequences (Friesen et al. 2006). Although in the study of Friesen et al. (2006) sections of subgenus *Allium* were represented by a single species each, which could potentially affect the robustness of the results, the ITS-based tree agrees much better with the morphology-based classification than those utilizing selected cp markers or the whole chloroplast genome. In the ITS-based tree of Friesen et al. (2006), the subgenus *Allium* forms a distinct clade except for *A. turkestanicum* from sect. *Mediasia*. In the study of Abugalieva et al. (2017) that utilized ITS, *A. sabulosum* and *A. caesium*, in contrast to our results, were clustered together. This discrepancy between the two marker classes needs close attention. We earlier hypothesized (Yusupov et al. 2021) that hybridization could play a role in formation of many *Allium* species, and a difference in the two marker types inheritance (maternal vs. biparental) can be, at least partly, responsible for the differences in species phylogenetic positions in the two types of trees.

Beside inferring subgenus *Allium* phylogeny, our goal was to analyze the chloroplast genome structure differences within a subgenus. Pseudogenization or total loss of genes are frequently found in plastomes (Haberle et al. 2008; Guisinger et al. 2011; Schwarz et al. 2015). We have detected the loss of *rpl22*, pseudogenization of *rps2, accD, psaI*, *rpl20, rpl23*, *rpl32* and NDH group of genes, and both loss and pseudogenization for such genes as *rps16* and *infA.* Earlier, loss/pseudogenization of genes in *Allium* was shown for *rps16, rps2*, *infA* and *rpl32* (Xie et al. 2019), and *rps2*, *accD*, *accD* and *ndh* (Scobeyeva et al. 2021). In addition, we have detected some new pseudogenization or deletion in the functionality of *psaI* (*A. ovalifolium*), *rpl20* (*A. ochotense*) and *rpl22* genes (*A. paradoxum*).

Similar to previous studies in *Allium* (Huo et al. 2019; Xie et al. 2020b), comparisions of the IR/SSC and IR/LSC boundaries showed similar characteristics between the studied species, with only slight differences in the position of *rpl22*, *rps19*, *ycf1* and *ndhK* genes. We also identified in all studied species the trnH-rps19 cluster to be in the IR regions of subgenus *Allium* species corresponding to Wang et al. (2008) type III expansion and contraction of IR/LSC junctions. These findings may be useful for understanding the genome structure of subgenus *Allium* species.

Acknowledgments

This study was supported by the «Tree of life: monocots of Uzbekistan». The authors are grateful to the curators of TASH and KUN herbaria.

References

Abugalieva S, Volkova L, Genievskaya Y, Ivaschenko A, Kotukhov Y, Sakauova G, Turuspekov Y (2017) Taxonomic assessment of *Allium* species from Kazakhstan based on ITS and matK markers. *BMC Plant Biology* 17: 51-60.

Bartolucci F, Iocchi M, De Castro O, Conti F (2022) *Allium ducissae* (*A. subgen. Polyprason*, Amaryllidaceae) a New Species from the Central Apennines (Italy). *Plants* 11: 426.

Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33: 2583-2585.

Daniell H, Lin C-S, Yu M, Chang W-J (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology* 17: 1-29.

Darriba DT, Guillermo L Doallo, Ramón Posada, David (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9: 772-772.

Dierckxsens N, Mardulyn P, Smits G (2017) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* 45: e18-e18.

Abugalieva S, Volkova L, Genievskaya Y, Ivaschenko A, Kotukhov Y, Sakauova G, Turuspekov Y (2017) Taxonomic assessment of *Allium* species from Kazakhstan based on ITS and matK markers. *BMC Plant Biology* 17: 51-60.

Bartolucci F, Iocchi M, De Castro O, Conti F (2022) *Allium ducissae* (*A. subgen. Polyprason*, Amaryllidaceae) a New Species from the Central Apennines (Italy). *Plants* 11: 426.

Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33: 2583-2585.

Daniell H, Lin C-S, Yu M, Chang W-J (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology* 17: 1-29.

Darriba DT, Guillermo L Doallo, Ramón Posada, David (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9: 772-772.

Dierckxsens N, Mardulyn P, Smits G (2017) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* 45: e18-e18.

Doyle J (1991) DNA protocols for plants: a CTAB total DNA isolation. In *Molecular techniques in taxonomy* (eds Hewitt GM, Johnston A), pp. 283–293. Springer, New York.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792-1797.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Research* 32: W273-W279.

Friesen N, Fritsch RM, Blattner FR, Botany E (2006) Phylogeny and new intrageneric classification of *Allium* (Alliaceae) based on nuclear ribosomal DNA ITS sequences. *Aliso* 22: 372-395.

Friesen N, Smirnov SV, Leweke M, Seregin AP, Fritsch RM (2021a) Taxonomy and phylogenetics of *Allium* section *Decipientia* (Amaryllidaceae): morphological characters do not reflect the evolutionary history revealed by molecular markers. *Botanical Journal of the Linnean Society* 197: 190-228.

Friesen N, Vesselova P, Osmonaly B, Sitpayeva G, Luferov A, Shmakov A (2021b) *Allium toksanbaicum* (Amaryllidaceae), a new species from Southeast Kazakhstan. *Phytotaxa* 494: 251–267.

Greiner S, Lehwark P, Bock R (2019) OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Research* 47: W59-W64.

Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Molecular Biology and Evolution* 28: 583-600.

Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *Journal of Molecular Evolution* 66: 350-361.

Hauenschild F, Favre A, Schnitzler J, Michalak I, Freiberg M, Muellner-Riehl AN (2017) Spatio-temporal evolution of *Allium* L. in the Qinghai–Tibet-Plateau region: Immigration and in situ radiation. *Plant Diversity* 39: 167-179.

Herden T, Hanelt P, Friesen N (2016) Phylogeny of *Allium* L. subgenus *Anguinum* (G. Don. ex WDJ Koch) N. Friesen (Amaryllidaceae). *Molecular Phylogenetics and Evolution* 95: 79-93.

Huo Y, Gao L, Liu B, Yang Y, Kong S, Sun Y, Yang Y, Wu X (2019) Complete chloroplast genome sequences of four *Allium* species: comparative and phylogenetic analyses. *Scientific Reports* 9: 1-14.

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

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772-780.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647-1649.

Khassanov FO (2018) Taxonomical and ethnobotanical aspects of *Allium* species from Middle Asia with particular reference to subgenus *Allium*. The *Allium* Genomes. Springer, pp. 11-21.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* 29: 4633-4642.

Lee J, Chon J, Lim J, Kim E-K, Nah G (2017) Characterization of complete chloroplast genome of *Allium victorialis* and its application for barcode markers. *Plant Breeding Biotechnology* 5: 221-227.

Li Q-Q, Zhou S-D, He X-J, Yu Y, Zhang Y-C, Wei X-Q (2010) Phylogeny and biogeography of *Allium* (Amaryllidaceae: *Allieae*) based on nuclear ribosomal internal transcribed spacer and chloroplast rps16 sequences, focusing on the inclusion of species endemic to China. *Annals of Botany* 106: 709-733.

Li Q-Q, Zhou S-D, Huang D-Q, He X-J, Wei X-Q (2016) Molecular phylogeny, divergence time estimates and historical biogeography within one of the world's largest monocot genera. *AoB Plants* 8: plw041

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451-1452.

Malé PJG, Bardon L, Besnard G, Coissac E, Delsuc F, Engel J, Lhuillier E, Scotti-Saintagne C, Tinaut A, Chave J (2014) Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular Ecology Resources* 14: 966-975.

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

Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS One* 7: e30619.

Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants. In *Plant diversity and evolution: Genotypic and phenotypic variation in higher plants* (ed. Henry RJ), pp. 45-68. CAB International.

Ronquist F, Huelsenbeck J, Teslenko M, Nylander JAA (2011) MrBayes version 3.2 manual: tutorials and model summaries. *Available online: mrbayes. sourceforge. net/mb3. 2_manual. pdf (accessed on 19 October 2018)*.

Schwarz EN, Ruhlman TA, Sabir JS, Hajrah NH, Alharbi NS, Al-Malki AL, Bailey CD, Jansen RK (2015) Plastid genome sequences of legumes reveal parallel inversions and multiple losses of rps16 in papilionoids. *Journal of Systematics and Evolution* 53: 458-468.

Scobeyeva VA, Artyushin IV, Krinitsina AA, Nikitin PA, Antipin MI, Kuptsov SV, Belenikin MS, Omelchenko DO, Logacheva MD, Konorov EA (2021) Gene loss, pseudogenization in plastomes of genus *Allium* (Amaryllidaceae), and putative selection for adaptation to environmental conditions. *Frontiers in Genetics* 12: 674783.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312-1313.

Swofford DL, Sullivan J (2003) Phylogeny inference based on parsimony and other methods using PAUP*. In *The phylogenetic handbook: A practical approach to DNA and protein phylogeny* (eds Salemi M and Vandamme A-M), pp. 160–206, Cambridge University Press.

Vaidya G, Lohman DJ, Meier R (2011) SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics* 27: 171-180.

Xie D-F, Cheng R-Y, Price M, Chen J-P, Lei J-Q, Zhang Y-Y, He X-J (2022) *Allium heterophyllum* (Amaryllidaceae), a new species from Henan, China. *PhytoKeys* 190: 53.

Xie D-F, Tan J-B, Yu Y, Gui L-J, Su D-M, Zhou S-D, He X-J (2020a) Insights into phylogeny, age and evolution of *Allium* (Amaryllidaceae) based on the whole plastome sequences. *Annals of Botany* 125: 1039-1055.

Xie D-F, Yu H-X, Price M, Xie C, Deng Y-Q, Chen J-P, Yu Y, Zhou S-D, He X-J (2019) Phylogeny of Chinese *Allium* species in section *Daghestanica* and adaptive evolution of *Allium* (Amaryllidaceae, *Allioideae*) species revealed by the chloroplast complete genome. *Frontiers in Plant Science* 10: 460.

Xie F-M, Xie D-F, Xie C, Yu Y, Zhou S-D, He X-J (2020b) Adaptation Evolution and Phylogenetic analyses of species in Chinese *Allium* section *Pallasia* and related species based on complete chloroplast genome sequences. *BioMed Research International* 2020: 8542797.

Yang X, Xie D-F, Chen J-P, Zhou S-D, Yu Y, He X-J (2020) Comparative analysis of the complete chloroplast genomes in *Allium* Subgenus *Cyathophora* (Amaryllidaceae): phylogenetic relationship and adaptive evolution. *BioMed Research International* 2020:1732586.

Yusupov Z, Deng T, Liu C, Lin N, Tojibaev K, Sun H (2019) The complete chloroplast genome of *Allium fistulosum*. *Mitochondrial DNA Part B* 4: 489-490.

Yusupov Z, Deng T, Volis S, Khassanov F, Makhmudjanov D, Tojibaev K, Sun H (2021) Phylogenomics of *Allium* section *Cepa* (Amaryllidaceae) provides new insights on domestication of onion. *Plant Diversity* 43: 102-110.

Supplementary tables

Suppl. Table 1. Comparison of plastomes of 11 species of subgenus *Allium*

	Genome length (bp)				GC content (%)				Accession no.
Species	Total	LSC	SSC	IR	Total	LSC	SSC	IR	
Allium caesium	151 533	80 507	18 018	26 504	36.9	34.7	29.4	42.6	OM912567
Allium confragosum	153 119	82 270	17 969	26 440	36.7	34.5	29.4	42.7	OM912568
Allium filidens	153 150	82 195	17 917	26 519	36.6	34.4	29.1	42.6	OM912569
Allium glomeratum	153 406	82 530	18 042	26 417	36.8	34.7	29.4	42.7	OM912570
Allium haneltii	152 344	81 271	18 031	26 521	36.7	34.5	29.3	42.6	OM912571
Allium michaelis	153 134	81 958	18 064	26 556	36.7	34.5	29.1	42.6	OM912572
Allium nikolai	152 485	81 714	17 907	26 432	36.7	34.5	29.2	42.6	OM912573
Allium pallasii1	151 610	80 762	17 756	26 546	36.7	34.4	29.2	42.5	OM912575
Allium pallasii	151 465	80 699	17 590	26 588	36.7	34.4	29.2	42.5	OM912574
Allium sabulosum	151 627	80 412	18 033	26 591	36.7	34.5	28.9	42.6	OM912576
Allium tanguticum	152 980	82 261	17 889	26 415	36.8	34.5	29.5	42.7	OM912577

Abbreviations: IR, inverted repeat; LSC, large single copy; SSC, small single copy

Suppl. Table 2. List of genes identified in plastomes of 11 species of subgenus *Allium*

Genes' category	Gene group	Gene name
Self-replication	Ribosomal RNA genes	*rrn4.5* x 2, *rrn5* x 2, *rrn16* x 2, *rrn23* x 2
	Transfer RNA genes	*trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-GCC, trnG-UCCa, trnH-GUG* x 2, *trnK-UUUa, trnL-UAAa, trnL-UAG, trnM-CAU,* *trnP-UGG, trnQ-UUG, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-UGU, trnT-GGU, trnV-UACa, trnY-GUA, trnW-CCA,* *trnfM-CAU, trnA-UGCa* x 2, *trnI-CAU* x 2, *trnI-GAUa* x 2, *trnL-CAA* x 2, *trnN-GUU* x 2, *trnR-ACG* x 2, *trnV-GAC* x 2
	Ribosomal protein (small subunit)	*rps2, rps3, rps4, rps7* x 2, *rps8, rps11, rps12b* x 2, *rps14, rps15, rps16a, rps18, rps19*
	Ribosomal protein (large subunit)	*rpl2a* x 2, *rpl14, rpl16a, rpl20, rpl22, rpl23* x 2, *rpl32, rpl33, rpl36*
	RNA polymerase	*rpoA, rpoB, rpoC1a, rpoC2*
	Translational initiation factor	*infA*
Genes for photosynthesis	Subunits of photosystem I	*psaA, psaB, psaC, psaI, psaJ, ycf3b, ycf4*
	Subunits of photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
	Subunits of cytochrome	*petA, petBa, petDa, petG, petL, petN*
	Subunits of ATP synthase	*atpA, atpB, atpE, atpFa, atpH, atpI*
	Large subunit of Rubisco	*rbcL*
	Subunits of NADH dehydrogenase	*ndhAa, ndhBa* x 2, *ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
Other genes	Maturase	*matK*
	Envelope membrane protein	*cemA*
	Subunit of acetyl-CoA	*accD*
	Synthesis gene	*ccsA*
	ATP-dependent protease	*clpPb*
	Component of TIC complex	*ycf1* x 2
Genes of unknown function	Conserved open reading frames	*ycf2* x 2, *ycf15* x 2

Abbreviations: x 2 Two gene copies in the IR regions; *a* With one intron; *b* With two introns.