Genome Survey Sequencing of Nomocharis forrestii, Assembly of Its Complete Chloroplast Genome and Analysis of Simple Sequence Repeat (SSR) Markers

Da Zhang
School of Agriculture, Yunnan University
LianLian Li
School of Agriculture, Yunnan University
XueWei Wu (wuxuewei@ynu.edu.cn)
School of Agriculture, Yunnan University
TianXi Wang
School of Agriculture, Yunnan University
Na Ping
School of Agriculture, Yunnan University
HaiYing Liu
School of Agriculture, Yunnan University
YongPing Li
School of Agriculture, Yunnan University
YiPing Zhang
Flower Research Institute, Yunnan Academy of Agricultural Sciences
LiHua Wang
Flower Research Institute, Yunnan Academy of Agricultural Sciences

Research Article

Keywords: Nomocharis forrestii, SSR marker, chloroplast genome, genome survey

Posted Date: December 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-124261/v1

Genome Survey Sequencing of *Nomocharis forrestii*, Assembly of Its Complete Chloroplast Genome and Analysis of Simple Sequence Repeat (SSR) Markers

Da Zhang [a*], LianLian Li [a*], XueWei Wu [a], TianXi Wang [a], Na Ping [a], HaiYing Liu [a], YongPing Li [a], YiPing Zhang [b], LiHua Wang [b**]**

[a] *School of Agriculture, Yunnan University, Chenggong District, Kunming, Yunnan 650091, China*

[b] *Flower Research Institute, Yunnan Agriculture Academy of Science, Panlong District, 650025, Yunnan, China;*

[*] Both authors contributed equally in preparing the manuscript.

** Corresponding author: wuxuewei@ynu.edu.cn (X.W. Wu); 687514549@qq.com (L.H. Wang)

Abstract

Background: *Nomocharis* is a genus that is closely related to *Lilium* in the Liliaceae family. It's useful to study the influence of the uplift of the Qinghai-Tibet Plateau on plants and their biological diversity. *Nomocharis* is a genus of such plants, and research on this genus will be especially informative, considering the genetic diversity of flowers. However, the genetic information of *Nomocharis* has not been fully elucidated.

Results: To obtain a complete *Nomocharis* reference genome, the paper first performed a general survey. Next-generation sequencing (NGS) was utilized to perform *de novo* sequencing of the entire *Nomocharis forrestii* genome. In this study, the sequencing process yielded approximately 137.4 Gb of high-quality data, the total sequencing depth was approximately 63X, and the Q30 ratio was 91.95%; the estimated genome size was approximately 2.17 Gb; the repetitive sequence content was approximately 84.7%, the heterozygosity rate was 3.99%, and the estimated GC content of the genome was 43%. Furthermore, an annotated circular chloroplast gene map was generated, and a preliminary evolutionary analysis was performed. In addition, a total of 78,045

high-quality SSR markers were developed.

Conclusion: *Nomocharis forrestii* has a 2.17 Gb heterozygous genome, its SSR markers are predominantly dinucleotides, and its chloroplast genome shows that *Nomocharis forrestii* and *Lilium bakerianum* have the highest homology followed by *Lilium distichum*. To the best of our knowledge, this report describes the first *de novo* whole-genome sequencing and assembly process to be performed for *Nomocharis*. The results of this study may provide new resources for the future genetic analysis and molecular breeding of *Nomocharis*.

Keywords: *Nomocharis forrestii*; SSR marker; chloroplast genome; genome survey

Background

Nomocharis is a genus closely related to *Lilium* in the Liliaceae family[1]. There are 7 species in this genus and 6 species in China. Among these species, *Nomocharis pardanthina* and *Nomocharis meleagrina* are endemic to China[2]. *Nomocharis* is distributed on the southeastern margin of the Qinghai-Tibet Plateau, being concentrated in northwestern Yunnan and adjacent areas, and only *Nomocharis synaptica* appears in northeastern India. Since Franchet established *Nomocharis*[3] with *Nomocharis pardanthina* as the model species in 1889, whether the genus was established as an independent genus and the scope of the genus have been controversial. For an extended period, scholars have had different opinions, resulting in frequent changes in the ownership of certain species within this genus and *Lilium*[1]. *Nomocharis aperta* and *Nomocharis saluenensis* in *Nomocharis* and *Lilium lophophorum*, *Lilium souliei*, and *Lilium henrici* in the genus *Lilium* have been moved back and forth between the two genera. Previously, scholars believed that *Nomocharis* was a young taxonomy newly derived from *Lilium* during the uplift of the

47 Qinghai-Tibet Plateau[1,4]. Regardless of morphological characteristics[3,5,6], geographic
48 distribution[7] or molecular properties[8,9,10], both *Nomocharis* and *Lilium* are inextricably linked.
49 Therefore, research investigating the *Nomocharis* genome and chloroplast levels and analyzing the
50 relationship between the two genera is of great significance not only for elucidating the phylogeny
51 and evolution of *Nomocharis* and *Lilium* but also for studying the effects of the uplift of the
52 Qinghai-Tibet Plateau on plants and on overall biological diversity.

53 A number of molecular biology research methods, such as the use of molecular markers,
54 require large quantities of information regarding genomes and specific functional genes. To date,
55 there are no research reports describing the whole genome of *Nomocharis*, and the lack of a
56 reference genome has placed considerable restrictions on the research investigating *Nomocharis*
57 molecular biology and genomics. Therefore, an investigation of the *Nomocharis* genome is
58 essential [11-16].

59 Recently, next-generation sequencing (NGS) has developed rapidly. This technique provides
60 scientists with faster and less expensive sequencing. Among many NGS sequencing platforms,
61 Illumina is the most commonly used for molecular marker development [11-18]. This platform is
62 preferred not only because it can use RCA products [19] but also because it can use a bioinformatic
63 platform for *de novo* assembly without reference to the genome sequence [20]. SSR markers are the
64 most widely utilized molecular marker system. SSR markers for many species have been
65 developed through NGS. The increase in the density of molecular markers can further promote
66 molecular breeding and genome-wide association. Therefore, to study and provide resources for
67 the *Nomocharis* genome for future research, research on the *N. forrestii* genome was performed
68 using NGS technology. In addition, the whole genome sequence of *N. forrestii* will be employed

for the development of SSR markers after assembly.

The chloroplast is a very important plant organelle with its own genome, and it produces energy through photosynthesis. Because chloroplasts have a highly conserved structure, the chloroplast genome has not only been employed as an useful research model, especially in phylogeny[21], but it has also been utilized as a DNA barcode[22] and for species protection and genome evolution[23]. To the best of our knowledge, there is no prior report on the *Nomocharis* chloroplast genome sequence, and the complete *N. forrestii* chloroplast genome is presented in this article.

Result

Genome Sequencing and Sequence Assembly

To sequence *N. forrestii*, we extracted DNA from fresh leaves. Through Illumina sequencing, the original sequencing data were approximately 137.4 Gb, the total sequencing depth was approximately 63X, and the Q30 ratio was 92.07%. After sequencing quality control, total clean data were obtained, and the sample Q30 base percentage was not less than 90% with an approximately 63X depth of sequencing. From the 300-bp library obtained by sequencing, the first 10,000 reads were extracted and compared with the NT library. No abnormal comparisons, such as microorganisms and humans, were observed in the comparison results, and there was no contamination in the samples. The chloroplast data of the *Lilium bakerianum* chloroplast complete genome (NC_035592.1) with a genome size of 151,655 bp were utilized to evaluate the plastid content. Comparing the obtained high-quality data with the plastid sequence, the comparison indicated that the content of plastids was low, which did not affect the sequencing and assembly of

the subsequently analyzed genome. The statistics of the *N. forrestii* sequencing data are shown in Table 1.

Clean data of high-quality reads were assembled using SOAPdenovo software based on a De Bruijn graph. The total length of the obtained genome sequence was 689 Mb, and the specific assembly results are shown in Table 2.

Table 1. Statistics of *Nomocharis forrestii* sequencing data

Raw/Clean	Read Number	Base Number(bp)	GC Content(%)	Q30(%)
Raw	458,075,259	138,014,716,200	43.48	91.95
Clean	458,066,599	137,419,979,700	43.49	91.95

Table 2. Information on the assembled genome sequences of *Nomocharis forrestii*

	Scaffold		Contig	
	Size(bp)	Number	Size(bp)	Number
N50	233	1,623,797	231	1,757,298
N60	233	1,623,797	231	1,757,298
N70	167	2,179,339	164	2,326,551
N80	151	2,936,082	151	3,090,524
N90	134	3,726,880	134	3,885,675
Longest	112	4,675,373	112	4,838,894
Total	245,202	1	25,481	1

Genome Size Estimation and Genome Survey

Using 137.4 Gb data for 17-mer analysis, the total number of K-mers was determined to be 1.3 × 10^{11}, and the expected K-mer depth was observed to be 75. According to the formula (genome size = total number of K-mers/expected depth of K-mer), the genome size was calculated to be approximately 1.73 Gb, and the genome size was estimated to be 2.17 Gb by GenomeScope software(Fig. 1). According to our experience, for complex genomes, the results of K-mer calculations may be smaller because the homologous K-mer is overlooked; therefore, the results predicted by GenomeScope were considered to be more accurate. The genome size of *N. forrestii* was estimated to be 2.17 Gb.

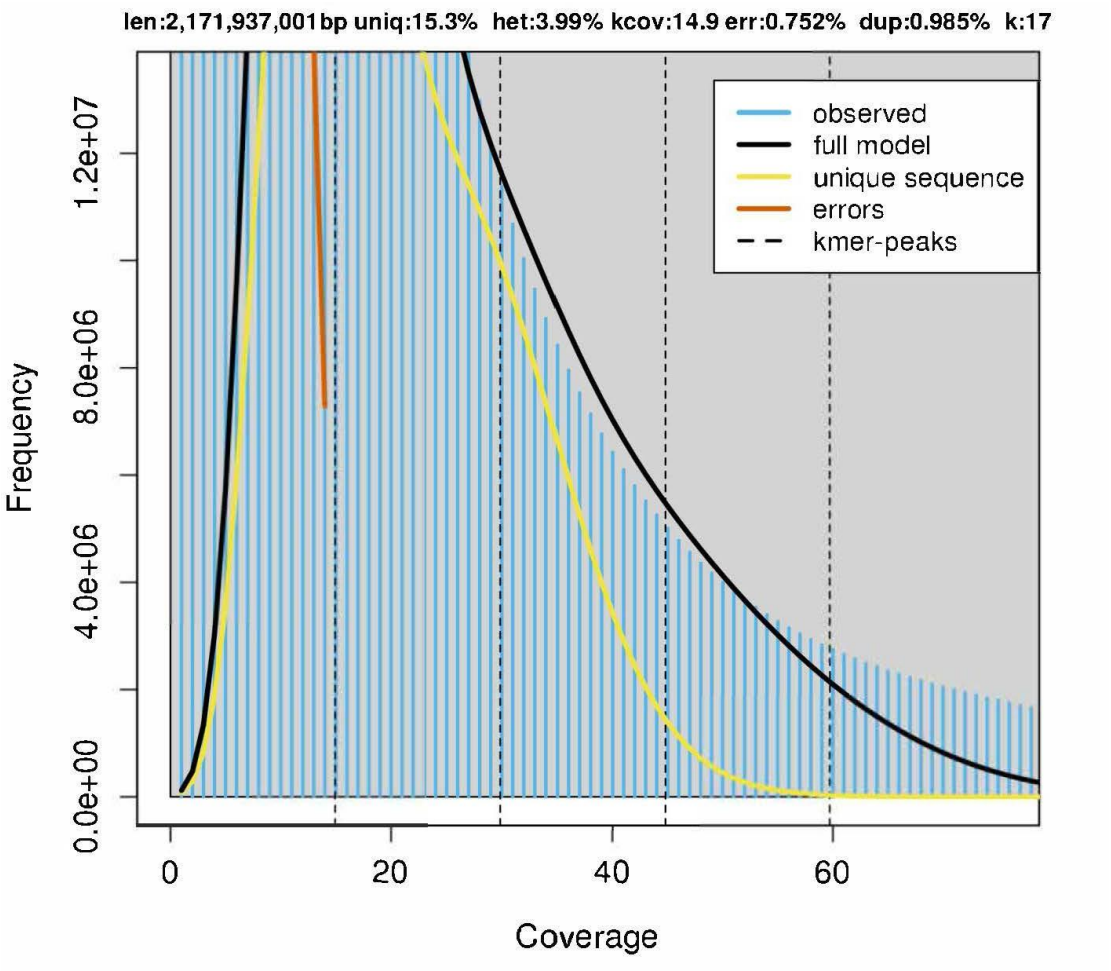

Fig. 1 K-mer distribution calculated by GenomeScope. The blue bar represents the observed K-mer

distribution; the black line represents no K-mer; the red line represents the error model distribution; and the yellow line represents the maximum K-mer coverage specified in the model.

Assembly of Chloroplast Genome

The fully annotated annotation results indicate that the sample chloroplast genome is a circular double strand. Similar to most higher plant chloroplast genomes, there are two inverted repeats (IRs), namely, IRAs and IRBs; between the inverted repeats, there is a large single-copy region (LSC) and a small single-copy region (SSC) (Fig. 2).

Fig. 2 Chloroplast genome of *Nomocharis forrestii.*

Statistics show that there are 116 genes in the chloroplast genome of *N. forrestii* samples, including 82 protein-coding genes (PCGs), 30 transfer RNA (tRNA) and 4 ribosomal RNA (rRNA) genes, of which 7 genes have more than 2 copies. The total GC content of the chloroplast genome was determined to be approximately 37.0%. All chloroplast genes and classifications are shown in Table 3.

The chloroplast genome plays an important role in the reconstruction of plant phylogeny and evolutionary history. In our research, we utilized whole-genome sequences from 25 kinds of chloroplasts (15 of which are Liliaceae) and constructed a phylogenetic tree using MEGAX software[24,25]. Using the neighbor-joining method[26], 1000 bootstrap test repeats draw a proportional evolutionary tree; taking the number of base substitutions at each site as the unit, the maximum likelihood method[27] is used to calculate the evolutionary distance (deleting all ambiguous positions of a single sequence pair). The branch length of the evolutionary tree is used to show the evolutionary distance of the phylogenetic tree, and the percentage of the replication tree of the bootstrap test is marked next to the branch. The use of the complete chloroplast genome sequence to reconstruct molecular phylogenetic relationships strongly supports the phylogenetic relationships of Liliaceae plants. In this study, it was observed that *Nomocharis forrestii* and *Lilium bakerianum* have the highest homology followed by *Lilium distichum* (Fig. 3).

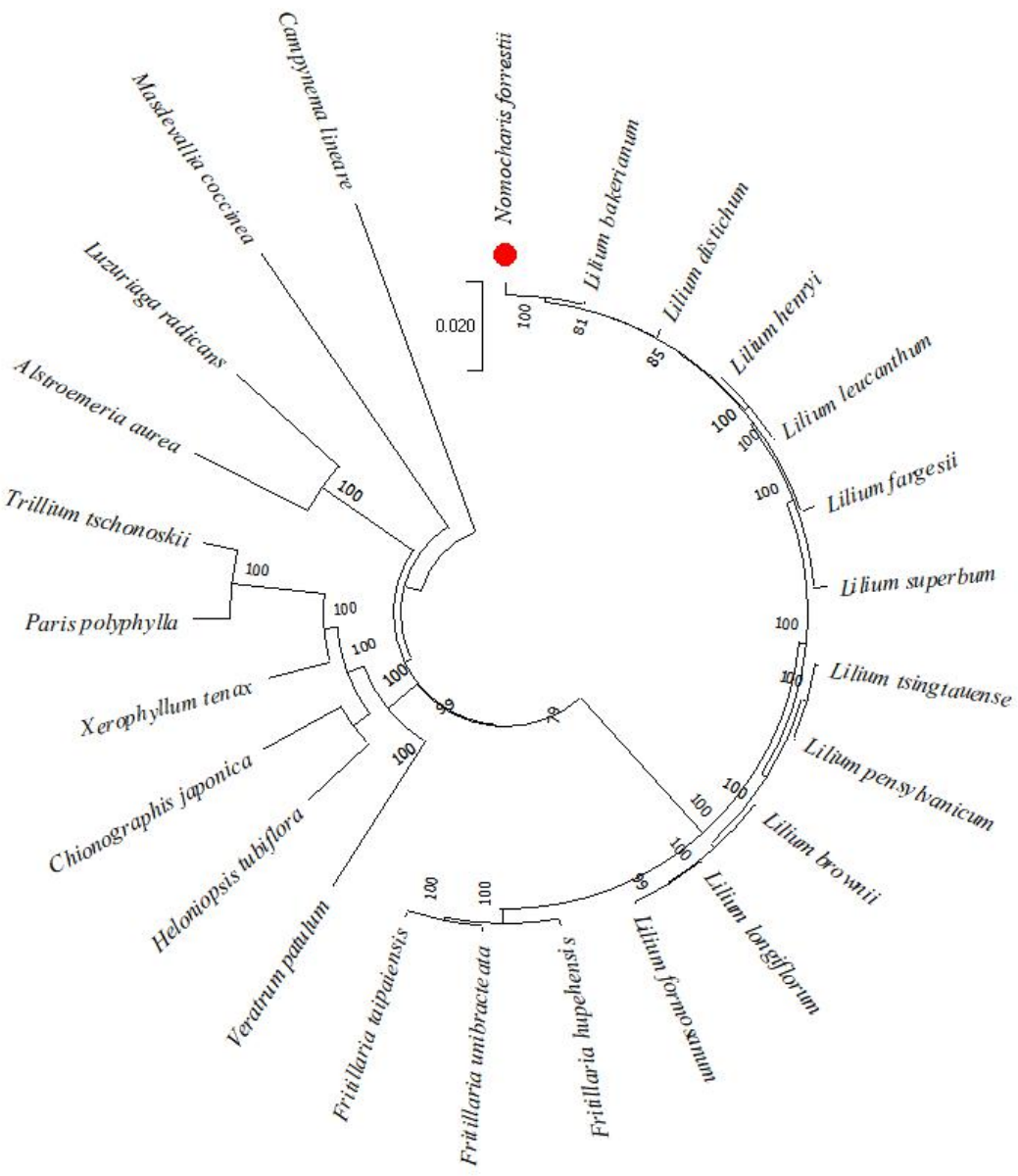

Fig. 3 Molecular phylogeny of 15 Liliaceae plants based on 25 whole chloroplast sequences

189 **Table 3.** Classification of *Nomocharis forrestii* chloroplast genes

Functions	**Family name**	**Name of Gene(s)**
	Subunits of ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Subunits of NADH dehydrogenase	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Subunits of cytochrome	*petA, petB, petD, petG, petL, petN*
Genes for photosynthesis	Subunits of photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Subunits of photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
	Subunit of rubisco	*rbcL*
	Subunit of Acetyl-CoA-carboxylase	*accD*
	c-type cytochrome synthesis gene	*ccsA*
Other genes		

	Envelop membrane protein	*cemA*
	Protease	*clpP*
	Maturase	*matK*
	Large subunit of ribosome	*rpl14, rpl16, rpl2, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36*
	DNA dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	Small subunit of ribosome	*rps11, rps12, rps14, rps15, rps16, rps18, rps19, rps2, rps3, rps4, rps7, rps8*
	rRNA Genes	*rrn16S, rrn23S, rrn4. 5S, rrn5S*
Self-replication	tRNA Genes	*trnI-GAU, trnA-UGC, trnR-ACG, trnN-GUU, trnL-UAG, trnN-GUU, trnV-GAC, trnL-CAA, trnY-GUA,trnT-UGU,trnN-GUU,trnT-GGU,trnL-CA A,trnA-UGC,trnR-UCU,trnW-CCA,trnG-GCC,trnC-GCA,trnE-UUC,trnP-UGG,trnQ-UUG,trnS-GGA,tr nI-CAU,trnS-GCU,trnL-UAA,trnS-UGA,trnV-UAC,t rnK-UUU,trnF-GAA,trnL-UAG,trnD-GUC,trnM-CA U,trnI-GAU,trnV-GAC,trnH-GUG,trnG-UCC,trnR-A CG*

Genes of unknown function	Conserved open reading frames	*ycf1, ycf2, ycf3, ycf4*

Genomic SSR Marker Development

We employed the MISA software to search the assembled scaffold for SSR marks. A total of 78,045 SSRs were identified from 2,847,542 scaffolds (Table 4). Regarding the base length of SSR repetitive sequences (not including single nucleotides), dinucleotides accounted for 89.7%, trinucleotides accounted for 8.97%, tetranucleotides accounted for 1.39%, pentanucleotides accounted for 0.17%, and hexanucleotides accounted for 0.30% (Fig. 4a).

Table 4. SSR types detected in the *Nomocharis forrestii* sequences

Item	Number	Percentage(%)
Total number of sequences examined	5,772,746	-
Total size of examined sequences (bp)	1,155,548,885	-
Total number of identified SSRs	63,472	100.00
Number of SSR containing sequences	53,523	84.33
Number of sequences containing more than 1 SSR	5,957	9.39
Number of SSRs present in compound formation	9,375	14.77
Mononucleotide	29,190	45.99
Dinucleotide	25,775	40.61

Trinucleotide	7,486	11.79
Tetranucleotide	513	0.81
Pentanucleotide	141	0.22
Hexanucleotide	367	0.58

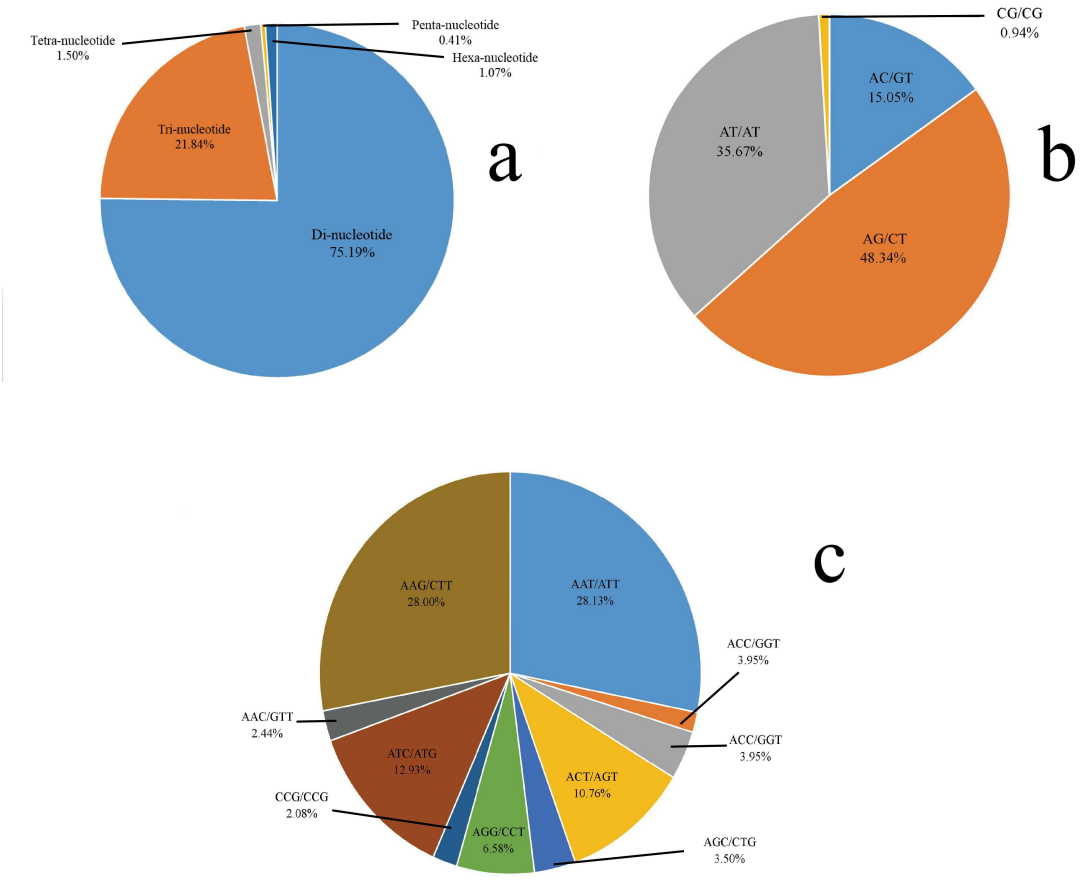

Fig. 4 Features of SSR markers. (a) The frequency of different SSR markers; (b) The frequency of different dinucleotide SSR markers; (c) The frequency of different trinucleotide SSR markers.

Among dinucleotide SSR markers, AT/AT repeat motifs accounted for 35.67%, AG/CT motifs accounted for 48.34%, AC/GT motifs accounted for 15.05%, and CG/CG motifs accounted for only 0.94% (Fig. 4b). Among the predominant trinucleotide SSR markers, the AAT/ATT repeat motif, AAG/CTT repeat motif and ATC/ATG repeat motif accounted for 28.13%, 28.00% and 12.93%, respectively (Fig. 4c).

SSR markers classified by the number of repeated motifs are summarized (Fig. 5). The SSR dinucleotide and trinucleotide markers were determined to be considerably more common than other SSR markers. In general, the number of SSR markers were observed to decrease as the length of the repeated motif increased.

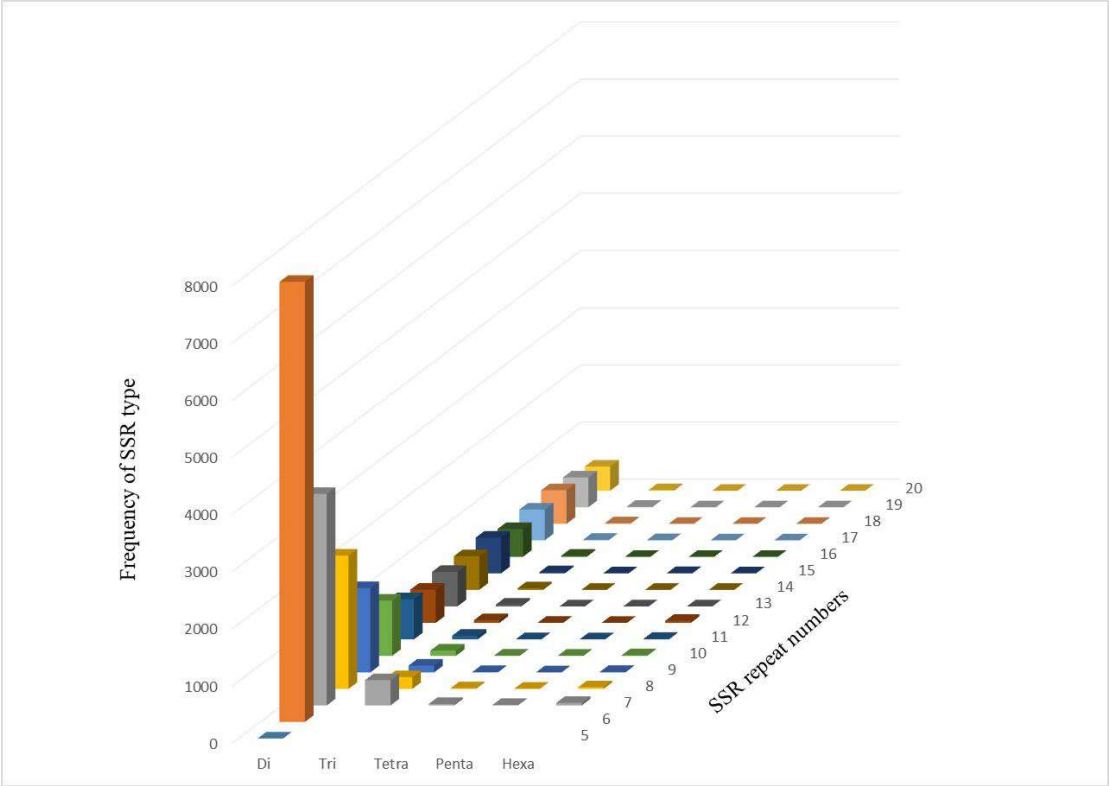

Fig. 5 Distribution and frequency of SSR motif repeats

Discussion

The genome of garlic, a member of the Liliaceae family, has been reported previously. The size of the sequenced garlic genome is 16.24 Gb, accounting for 96.1% of the total garlic genome [29]. Among the representative monocots, the genome size of indica rice is 430 Mb, and the functional coverage is 92%[30]; the genome size of japonica rice is 420 Mb, and the assembly coverage is 93%[31]. The genome size of maize is 2.3 Gb[32]. According to our genome survey data, using all clean data for Genome Scoper analysis, the estimated size of the *N. forrestii* genome was 2.17 Gb. Compared with the garlic genome, the whole genome of *N. forrestii* is small, but it is relatively large in monocots. With the development of NGS technology, whole-genome sequencing research has begun to be widely employed in horticultural plants, which may play an important role in understanding the key genes of *N. forrestii*.

GC content directly affects sequence bias [33]. GC content outside the 25-65% interval may cause sequence bias in Illumina sequencing. This problem is a notable one that affects the assembly of the genome [34]. The GC content of *N. forrestii* is 43.0%, which is higher than that of potato (34.8-36.0%)[35], *Luffa cylindrica* (37.9%)[36], and humans (41%) but lower than that of *Gracilariopsis lemaneiformis* (48%)[37].

From the 1,155,548,885-bp genome survey sequence, 34,552 SSRs without single nucleotide repeats were identified. Therefore, it is estimated that the distribution of SSRs in the genome of *N. forrestii* is approximately 29.90 SSR/Mb, which is considerably lower than the 135.50 SSR/Mb measured in *Arabidopsis*[38] and the 117.57 SSR/Mb detected in *Luffa cylindrica*. Among the dinucleotide repeat motifs, AG/CT accounted for 48.34%, which is the most abundant type, followed by AT/AT, accounting for 35.67%; in the trinucleotide repeat sequence, AAT/ATT and

AAG/CTT account for approximately the same proportion, being 28.13% and 28.00%, respectively; among other polynucleotide repeats, AAAT/ATTT, AAAAT/ATTTT and AAAAAG/CTTTTTT account for the highest proportions, and they are all A/T-rich motifs present in *N. forrestii*. This phenomenon is in keeping with the findings obtained by studies of other species, such as *L. cylindrica*[36], rice[39], and *Arabidopsis*[40].

Chloroplasts play important roles in the study of evolution and metabolism. The assembly and analysis of the whole chloroplast genome may also provide evidence to determine the evolutionary level and phylogeny of *N. forrestii*. The results of this study also indicate that *Nomocharis* evolved from the genus *Lilium.*

Conclusions

N. forrestii has a 2.17 Gb heterozygous genome, its SSR markers are predominantly dinucleotides, and its chloroplast genome shows that *N. forrestii* and *Lilium bakerianum* have the highest homology followed by *Lilium distichum*.

Materials and Methods

Materials

N. forrestii collected from Shangri-La, Yunnan, China. Centrifuges, reagents, and servers were provided by the School of Agricultural, Yunnan University.

Methods

Total genomic DNA extraction

Total genomic DNA was extracted from fresh leaves using the CTAB method [16].

Illumina Sequencing Data Analysis and Assembly

The Illumina HiSeq platform (Illumina Inc., San Diego, CA, USA) was used for genome sequencing. Sequencing was performed by Shaanxi Baiai Gene Information Technology Co., Ltd. Clean data were obtained through strict quality evaluation and data filtering of raw Illumina sequencing data. SOAPdenovo (https://github.com/aquaskyline/SOAPdenovo2) software [41] based on a De Bruijn Graph (version 1.05, BGI, Beijing, China) was employed to assemble clean data of high-quality reads. After assembly, the GC content information in the assembled genome was quantified.

Genome Size Estimation and Genome Survey

Clean data from high-quality reads were used for K-mer analysis. Based on the frequency distribution of K-mers (k = 17), we used GenomeScope (https://github.com/schatzlab/genomescope) to estimate the characteristics of the genome (genome size, duplicate content, and heterozygosity rate) [42]. Each read used 17 bp as the window and 1 bp as the step size to slide, and the total number of K-mers and the corresponding frequency were counted and calculated. Next, based on the K-mer depth distribution curve, the peak value (Peak_depth) was identified. Finally, according to the formula Genome Size = K-mer_num/Peak_depth, the genome size was calculated[16].

Assembly and analysis of chloroplast genome

The chloroplast genome was directly assembled with the help of NOVOPlasty (https://github.com/ndierckx/NOVOPlasty) [43] software; the reference sequence is NC_035592.1

of *L. bakerianum*. The chloroplast group genes of the samples were annotated with CPGAVAS (http://47.96.249.172:16014/analyzer/home) [24] software. The annotation results were plotted using OGDRAW (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html)[44]. MEGAX (https://www.megasoftware.net) was used to analyze the whole genome sequence of *N. forrestii* and 24 other chloroplasts using the neighbor joining method to analyze the evolutionary tree.

Identification and verification of SSRs

Using MISA software (https://webblast.ipk-gatersleben.de/misa), 2,847,542 scaffolds were utilized for genome SSR marker detection. We set the following search parameters for identification: di-, tri-, tetra-, penta- and hexanucleotide motifs have at least 6, 5, 4, 4, and 4 repeats, respectively, as described by previous authors [16,29,37,45,46].

Abbreviations:

SSR: Simple Sequence Repeats; NGS: next-generation sequencing; LSC: large single-copy region; SSC: small single-copy region; PCGs: protein-coding genes; tRNA: transfer RNA; rRNA: ribosomal RNA; MEGA: molecular evolutionary genetics analysis;

Declarations

Ethics approval and consent to participate

The plant materials used in this study were collected in Shangri-La, Yunnan and cultivated in the School of Agricultural, Yunnan University. They comply with national and international standards and local laws, and comply with the convention on trade in endangered species of wild

fauna and flora.

Consent for publication

All authors have read and approved the manuscript.

Availability of data and materials

All data generated or analysed during this study are included in this published article.

Competing interests

The authors declare that they have no competing interests.

Funding

Our work is supported by the Project of Ministry-Province Co-construct Grant Number C176280108 (Kunming, Yunnan) awarded to X.W. Wu. This research was carried out, reviewed references, received English editing service, and worked with two corresponding author to complete the manuscript to submit. Material treatment, data were recorded, tabulated and analyzed by two first authors who also assisted in preparing the draft of the manuscript.

Authors' contributions

Conceived and designed the experiments: X.W. Wu, L.H. Wang;

Performed the experiments: D. Zhang, L.L. Li;

Analyzed the data: D. Zhang, L.L. Li;

Contributed reagents/materials/analysis tools: D. Zhang, L.L. Li, X.W. Wu, T.X. Wang, N. Ping, H.Y. Liu, Y.P. Li, Y.P. Zhang and L.H. Wang; Wrote the paper: D. Zhang, L.L. Li, and X.W. Wu.

Acknowledgements

We thank for B.Q. Huang to help us review the manuscript. We are particularly grateful to R.K. Wang, Y.W. Zhang, and Q. Li for teach us using some analysis software.

References

1. Liang SY. Studies on the genus *Nomocharis* (Liliaceae). Bulletin of the of the Botanical Research.1984; 4:163-78.
2. Wan J, Zhou SD, Gao YD, He XJ. Karyotypes of Twenty-five Populations of Thirtee Species in *Nomocharis* and *Lilium*. Plant Diversity and Resources.2011; 33(5):477-94.
3. Sealy JR. *Nomocharis* and *Lilium*. Kew Bulletin.1950; 5(2):273-97.https://doi.org/10.2307/4117245.
4. Wu ZY, Li H, Yang CR. Cytogeography and phylogeny of Lilieae. Acta Botanica Yunnanica.1994; s6:101-12.
5. Sealy JR. A revision of the genus *Nomocharis* Franchet. Botanical journal of the Linnean Society.1983; 87(4):285-323.https://doi.org/10.1111/j.1095-8339.1983.tb00996.x.
6. Yu H, Huang RF, Dang CL. A study on the polymorphism of morphological character of *Nomocharis mairei*. Acta Botanica Yunnanica.1996; S8:48-58.
7. Liang SY. Chorology of Liliaceae (s. str.) and its bearing on the Chinese flora. Acta Phytotaxonomica Sinica.1995; 33(14):41.
8. Nishikawa T, Okazaki K, Uchino T, Arakawa K, Nagamine T. A molecular phylogeny of *Lilium* in the internal transcribed spacer region of nuclear ribosomal DNA. Journal of

Molecular Evolution.1999; 49(2):238-49.https://doi.org/10.1007/PL00006546

9. Hayashi K, Kawano S. Molecular systematics of *Lilium* and allied genera (Liliaceae): phylogenetic relationships among *Lilium* and related genera based on the *rbcL* and *matK* gene sequence data. Plant Species Biology.2000; 15(1):73-93.http://dx.doi.org/10.1046/j.1442-1984.2000.00025.x.

10. Cui GF, Wu LF, Wu XW, Zhang YP, Wang JH. Relationships among species in *Nomocharis* and *Lilium* based on ITS sequences [J]. Acta Botanica Boreali-Occidentalia Sinica.2008; 10.

11. Jiao Y, Jia HM, Li XW, Chai ML, Jia HJ, Chen Z, et al. Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (*Myrica rubra*). BMC genomics.2012; 13(1):201.

12. Zhou W, Hu YY, Sui ZH, Fu F, Wang JG, Chang LP, et al. Genome survey sequencing and genetic background characterization of *Gracilariopsis lemaneiformis* (Rhodophyta) based on next-generation sequencing. PLoS One.2013; 8(7):e69909.

13. Wei X, Wang L, Zhang Y, Qi X, Wang X, Ding X, et al. Development of simple sequence repeat (SSR) markers of sesame *(Sesamum indicum*) from a genome survey. Molecules.2014; 19(4):5150-62.

14. He Y, Xiao HT, Deng C, Xiong L, Nie H, Peng C. Survey of the genome of *Pogostemon cablin* provides insights into its evolutionary history and sesquiterpenoid biosynthesis. Scientific Reports.2016; 6:26405.http://doi.org/10.1038/srep26405.

15. Xu DP, Li Y, Meng X, Zhou T, Zhou Y, Zheng J, et al. Natural antioxidants in foods and medicinal plants: Extraction, assessment and resources. International journal of molecular sciences.2017; 18(1):96.http://doi.org/10.3390/ijms18010096.

16. Wang CR, Yan HD, Li J, Zhou SF, Liu T, Zhang XQ, et al. Genome survey sequencing of purple elephant grass (*Pennisetum purpureum* Schum 'Zise') and identification of its SSR markers. Molecular Breeding.2018; 38(7):94.http://doi.org/10.1007/s11032-018-0849-3.

17. Lu M, An HM, Li LL. Genome survey sequencing for the characterization of the genetic background of *Rosa roxburghii* tratt and leaf ascorbate metabolism Genes. PLoS One.2016; 11(2):e0147530.http://doi.org/10.1371/journal.pone.0147530.

18. Motalebipour EZ, Kafkas S, Khodaeiaminjan M, Çoban N, Gözel H. Genome survey of pistachio (*Pistacia vera* L.) by next generation sequencing: development of novel SSR markers and genetic diversity in Pistacia species. BMC genomics.2016; 17(1):998.http://doi.org/10.1186/s12864-016-3359-x.

19. Atherton RA, McComish BJ, Shepherd LD, Berry LA, Albert NW, Lockhart PJ. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant methods.2010; 6(1):1-6.https://doi.org/doi:10.1186/1746-4811-6-22.

20. Lin CS, Chen JJ, Huang YT, Chan MT, Daniell H, Chang WJ, et al. The location and translocation of *ndh* genes of chloroplast origin in the Orchidaceae family. Scientific reports.2015; 5:9040.http://doi.org/10.1038/srep09040.

21. Kocyan A, Zhang LB, Schaefer H, Renner SS. A multi-locus chloroplast phylogeny for the Cucurbitaceae and its implications for character evolution and classification. Molecular phylogenetics and evolution.2007; 44(2):553-77.https://doi.org/10.1016/j.ympev.2006.12.022.

22. Li XW, Yang Y, Henry RJ, Rossetto M, Wang YT, Chen SL. Plant DNA barcoding: from gene to genome. Biological Reviews.2015; 90(1):157-66.http://doi.org/10.1111/brv.12104.

23. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome biology.2016; 17(1):134.http://doi.org/10.1186/s13059-016-1004-2.

24. Zuo LH, Shang AQ, Zhang S, Yu XY, Ren YC, Yang MS, et al. The first complete chloroplast genome sequences of *Ulmus* species by de novo sequencing: Genome comparative and taxonomic position analysis. PLoS One.2017; 12(2):e0171264.http://doi.org/10.1371/journal.pone.0171264.

25. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution.2018; 35(6):1547-9.http://doi.org/10.1093/molbev/msy096.

26. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution.1987; 4(4):406-25.https://doi.org/10.1093/oxfordjournals.molbev.a040454.

27. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. evolution.1985; 39(4):783-91.

28. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences.2004; 101(30):11030-5.https://doi.org/10.1073/pnas.0404206101.

29. Sun XD, Zhu SY, Li NY, Cheng Y, Zhao J, Qiao XG, et al. A chromosome-level genome assembly of garlic (*Allium sativum*) provides insights into genome evolution and Allicin biosynthesis. Molecular Plant.2020; 13(9):1328-39.http://doi.org/10.1016/j.molp.2020.07.019.

30. Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). science.2002; 296(5565):79-92.https://doi.org/10.1126/science.1068037.

31. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science.2002; 296(5565):92-100.https://doi.org/10.1126/science.1068275.

32. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. science.2009; 326(5956):1112-5.http://doi.org/10.1126/science.1178534.

33. Cheung MS, Down TA, Latorre I, Ahringer J. Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic acids research.2011; 39(15):e103-e.https://doi.org/10.1093/nar/gkr425.

34. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology.2011; 12(2):1-14.http://doi.org/doi:10.1186/gb-2011-12-2-r18.

35. Hirakawa H, Okada Y, Tabuchi H, Shirasawa K, Watanabe A, Tsuruoka H, et al. Survey of genome sequences in a wild sweet potato, *Ipomoea trifida* (HBK) G. Don. DNA Research.2015; 22(2):171-9.http://doi.org/10.1093/dnares/dsv002.

36. Consortium PGS. Genome sequence and analysis of the tuber crop potato. Nature.2011; 475(7355):189.http://doi.org/10.1038/nature10158.

37. An JY, Yin MQ, Zhang Q, Gong DY, Jia XW, Guan YJ, et al. Genome survey sequencing of *Luffa Cylindrica* L. and microsatellite high resolution melting (SSR-HRM) analysis for genetic relationship of *Luffa* genotypes. International Journal of Molecular Sciences.2017; 18(9):1942.https://doi.org/10.3390/ijms18091942.

38. Zhao HS, Yang L, Peng ZH, Sun HY, Yue XH, Lou YF, et al. Developing genome-wide microsatellite markers of bamboo and their applications on molecular marker assisted taxonomy for accessions in the genus *Phyllostachys*. Scientific reports.2015; 5(1):1-10.http://doi.org/10.1038/srep08018.

39. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome research.2001; 11(8):1441-52.https://dx.doi.org/10.1101%2Fgr.184001.

40. Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular biology and evolution.2001; 18(7):1161-7.https://doi.org/10.1093/oxfordjournals.molbev.a003903.

41. Wang RK, Fan JS, Chang P, Zhu L, Zhao MR, Li LL. Genome survey sequencing of *Acer truncatum* bunge to identify genomic information, simple sequence repeat (SSR) markers and complete chloroplast genome. Forests.2019; 10(2):87.http://doi.org/10.3390/f10020087.

42. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics.2017; 33(14):2202-4.http://doi.org/10.1093/bioinformatics/btx153.

43. Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic acids research.2013; 41(13):e129-e.

44. Lohse M, Drechsel O, Kahlau S, Bock R. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic acids research.2013; 41(W1):W575-W81.

45. Wang RK, Liu P, Fan JS, Li LL. Comparative transcriptome analysis two genotypes of *Acer truncatum* Bunge seeds reveals candidate genes that influences seed VLCFAs accumulation.

Scientific reports.2018; 8(1):1-8.http://doi.org/10.1038/s41598-018-33999-3.

46. Li LL, Zhang HH, Liu ZS, Cui XY, Zhang T, Li YF, et al. Comparative transcriptome sequencing and de novo analysis of *Vaccinium corymbosum* during fruit and color development. BMC plant biology.2016; 16(1):223.http://doi.org/10.1186/s12870-016-0866-5.

Figures

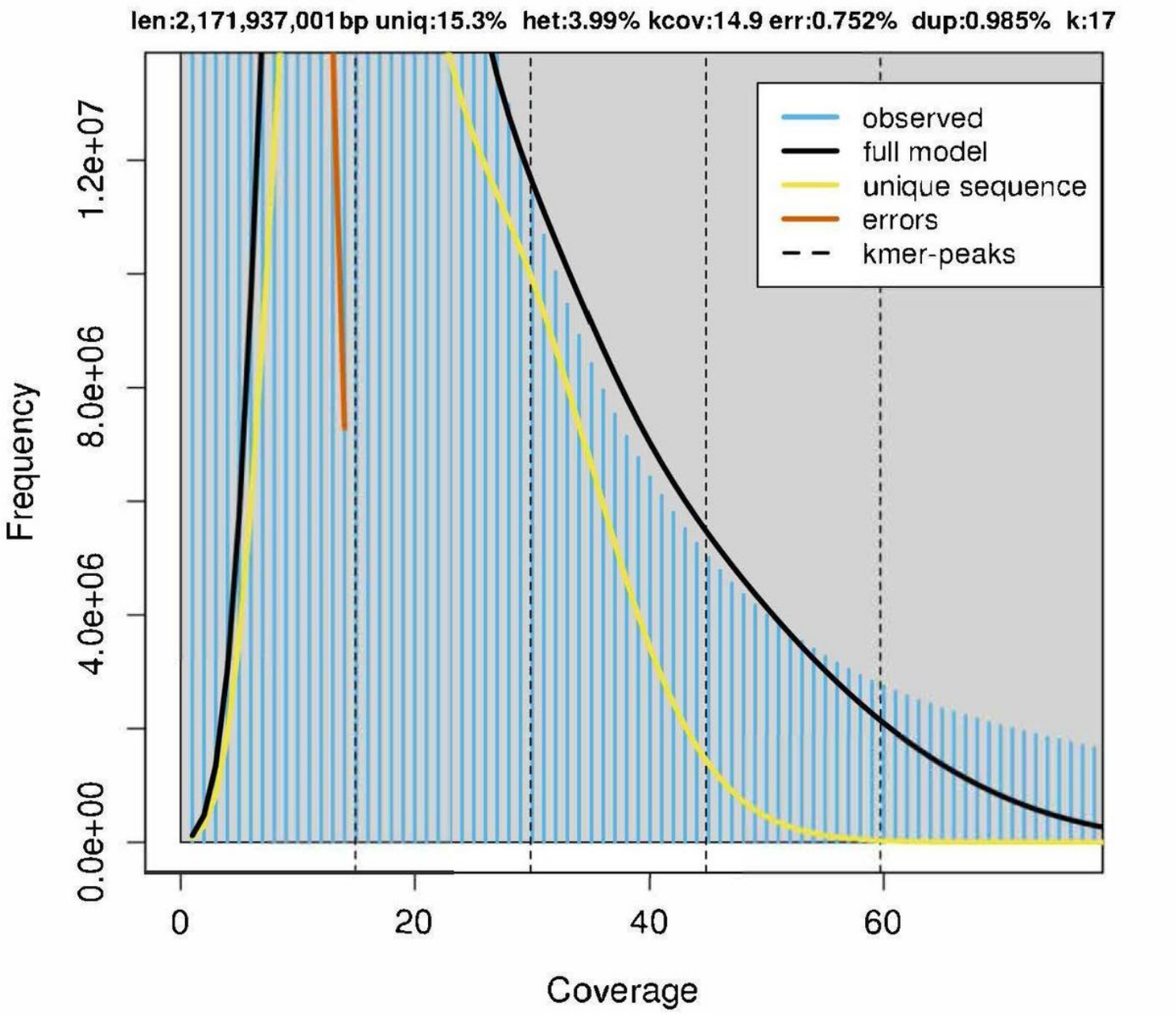

Figure 1

K-mer distribution calculated by GenomeScope. The blue bar represents the observed K-mer distribution;the black linerepresents no K-mer;the redline represents the errormodel distribution;and theyellowlinerepresentsthemaximum K-mercoveragespecifiedinthemodel

Figure 2

Chloroplast genome of Nomocharis forrestii

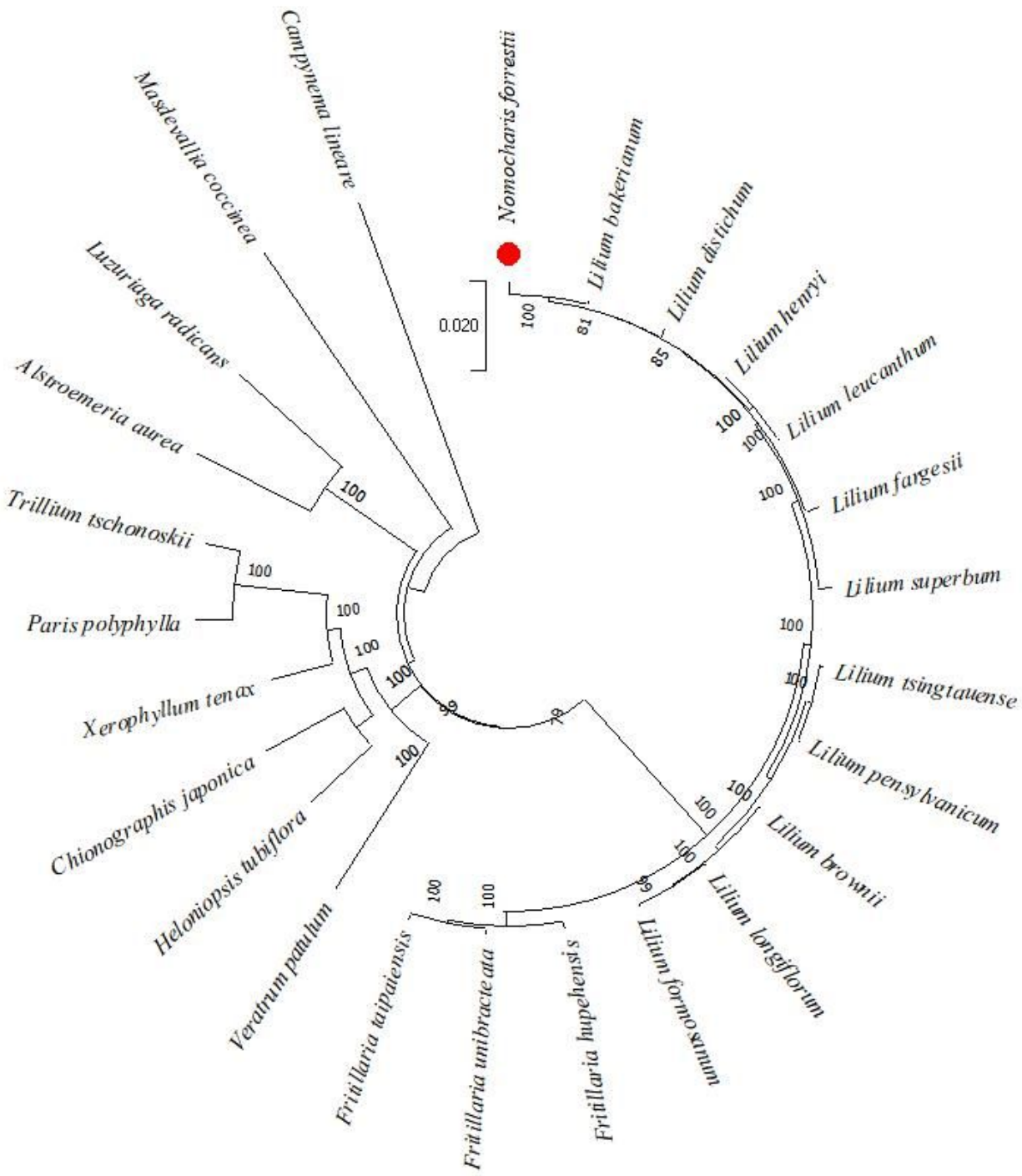

Figure 3

Molecular phylogeny of 15 Liliaceae plants based on whole chloroplast sequence

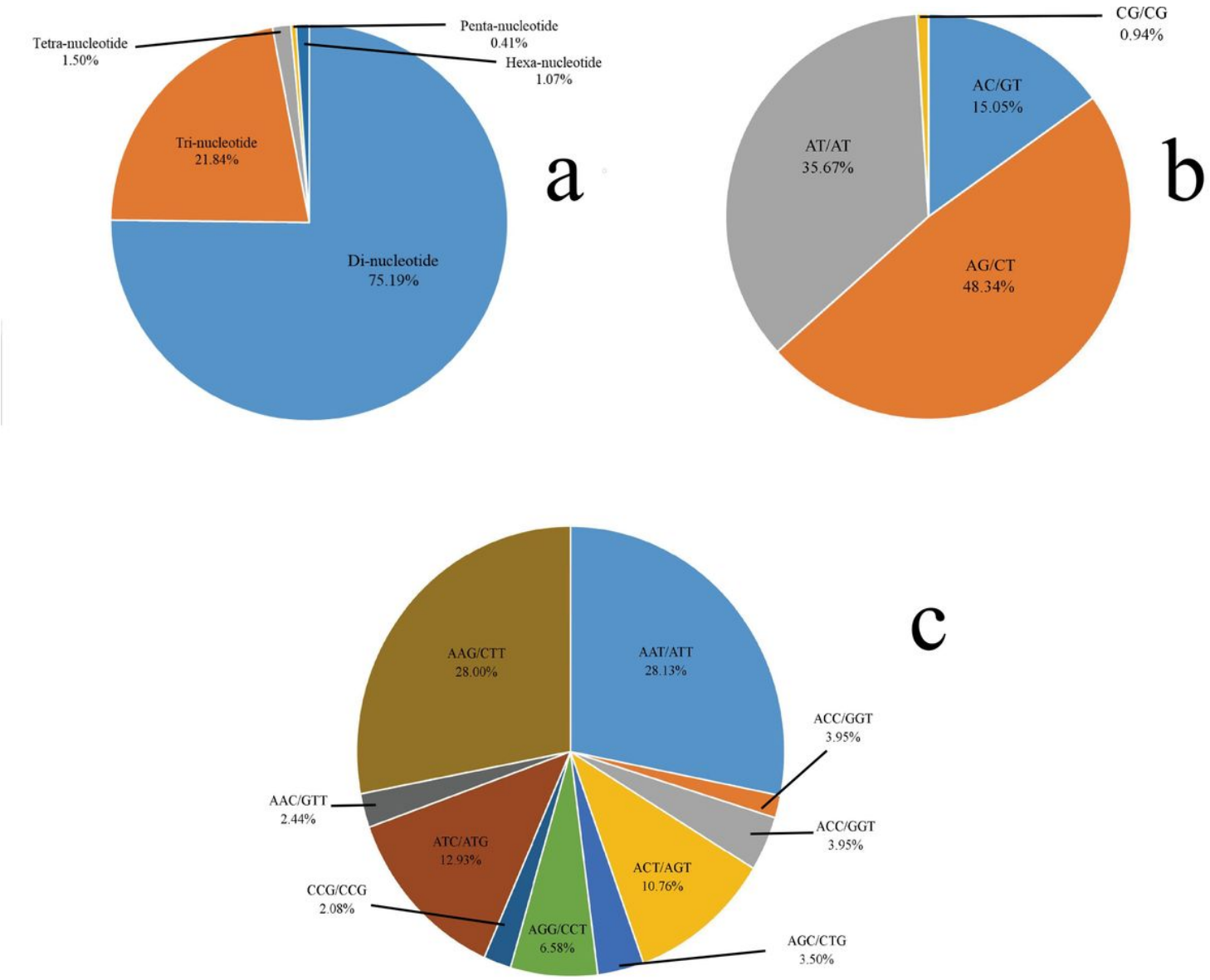

Figure 4

FeaturesofSSRmarkers.(a)ThefrequencyofdifferentSSRmarkers;(b)The frequencyofdifferentdinucleotideSSRmarkers;(c)ThefrequencyofdifferenttrinucleotideSSR markers

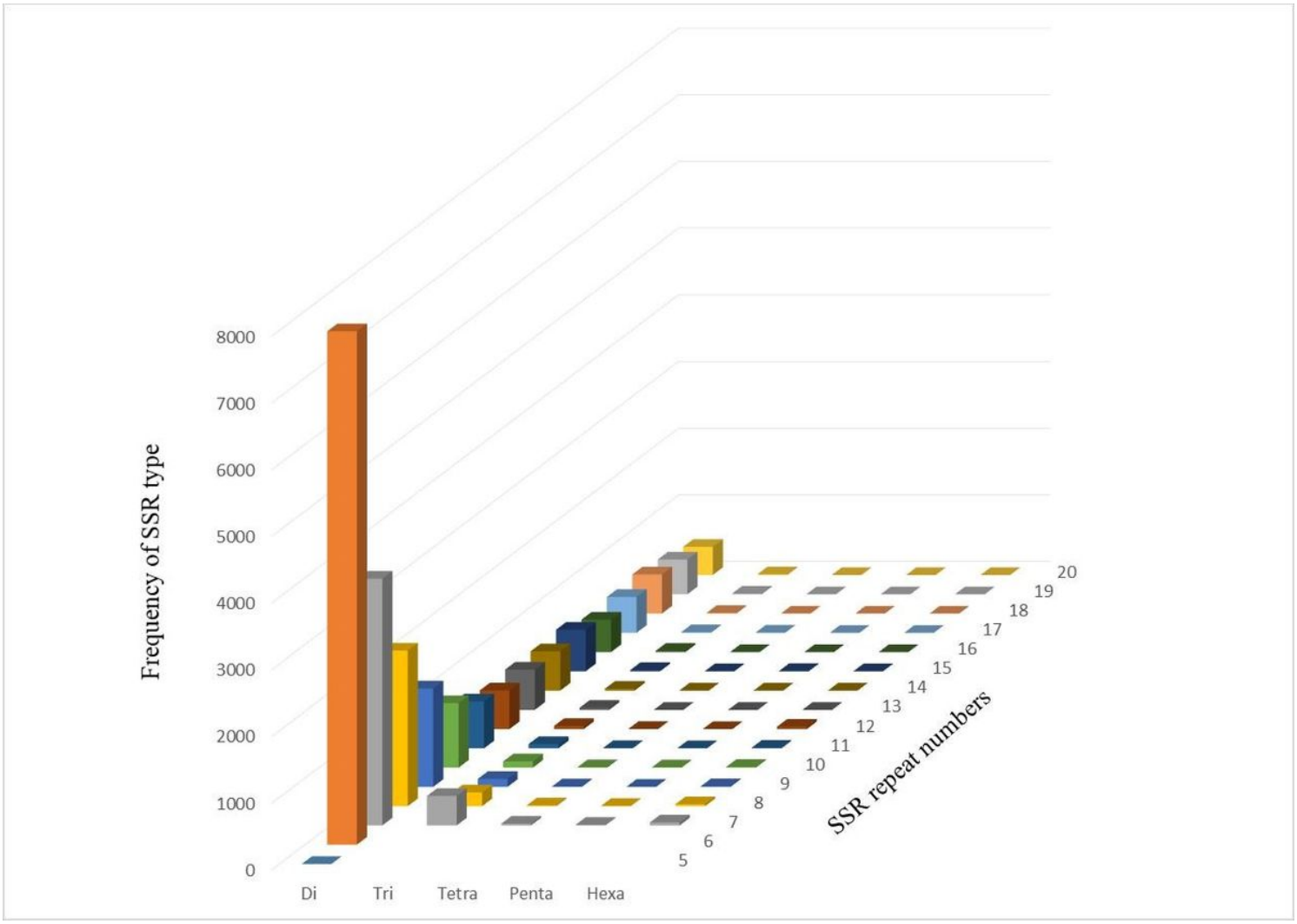

Figure 5

Distribution and frequency of SSR motif repeats