

# Analysis of Whole Genome Characteristics of *Helianthus annuus* Longkuiza 6 Chloroplast

Jun Ma <sup>a</sup>

Institute of Industrial Crops, Heilongjiang Academy of Agricultural Sciences, Xuefu Street, Harbin, Heilongjiang, China

**Keywords:** *Helianthus annuus*, Chloroplast Genome, Characteristics.

**Abstract:** The whole chloroplast genome was assembled and annotated on the leaf high-throughput sequencing data of the oil sunflower hybrid longkuiza 6. The results showed that the whole chloroplast genome was 151124bp in length, including a typical tetrad structure. Two single-copy regions were separated by a pair of inverted repeat regions. The lengths of LSC and SSC were 83548bp and 18308bp, respectively. The chloroplast genome encodes a total of 127 genes, including 84 protein-coding genes, 8 rRNA genes and 35 tRNA genes. In the chloroplast genome of *Helianthus annuus* longkuiza 6, there are a total of 18 genes containing introns, except for the *ycf3* and *clpP* genes, the rest contain 1 intron. In the protein coding gene of *H. annuus* longkuiza 6, leucine is the amino acid with the highest codon encoding rate, and the amino acid with the lowest encoding rate is cysteine.

## 1 INTRODUCTION

The chloroplast genome contains a large number of functional genes, which have important research value in species identification and system evolution (Dong 2021, Gao 2019). The chloroplast genome exists in a covalently closed double-stranded form and consists of four regions: a pair of inverted repeat regions, a large single-copy sequence region, and a small single-copy sequence region (Wen 2021, Redwan 2015). Chloroplast genome sequence has the characteristics of highly conservative, stable structure, slow molecular evolution rate and small molecular weight, which makes it play an important role in cytoplasmic inheritance, plant phylogeny, development of DNA barcode, genetic diversity and genetic relationship (Ng 2017, Coombe 2016). Therefore, the chloroplast genome has become one of the most effective tools for studying plant phylogeny.

*Helianthus annuus* belongs to the Asteraceae. *Helianthus* is an annual herbaceous plant and is an important oil crop (Liu 2020, Hussain 2017). Wild sunflower gradually evolved into cultivated sunflower after long-term natural selection and artificial domestication. This study analyzed the whole genome of the chloroplast of *H. annuus*

longkuiza 6, hoping to provide a reference for the classification, evolutionary analysis and utilization of resources.

## 2 MATERIALS AND METHODS

### 2.1 Experimental Materials

The *H. annuus* longkuiza 6 was selected for the experiment. The experimental materials were provided by the Economic Crop Research Institute of Heilongjiang Academy of Agricultural Sciences. During the experiment, fresh sunflower leaves were collected for follow-up experiments.

### 2.2 DNA Extraction and High-throughput Sequencing

Collect *H. annuus* longkuiza 6 samples, clean the surface of the leaves, extract the total DNA of *H. annuus* longkuiza 6 leaves with the CTAB method, and detect the concentration and purity by 1% agarose gel electrophoresis and UV spectrophotometer. The qualified samples were

<sup>a</sup> <https://orcid.org/0000-0002-3287-9940>

commissioned by BGI for high-throughput sequencing of *H. annuus* longkuiza 6 genome.

### 2.3 Assembly and Annotation of Chloroplast Genome

Filter the original data to obtain high-quality data. Use BLAT software to position and compare it with the chloroplast reference genome of closely related species, obtain the relative position between contig sequences, and correct assembly errors to obtain the full-length cpDNA framework. After that, use GapCloser software to fill high-quality short sequences into the gaps on the frame diagram sequence, and use first-generation sequencing to supplement and confirm the remaining gaps and suspicious regions. The *H. annuus* longkuiza 6 cpDNA was annotated by CPGAVAS2, a special annotation software for chloroplasts, and the annotated results were graphed and analyzed by OGDRAW software.

gene is a split gene, the 5'end is located in the LSC region, and the 3'end is located in the two IR regions.

Table 1: Basic characteristics of *H. annuus* longkuiza 6 chloroplasts

Chloroplasts feature	Numerical value
Length (bp)	151124
GC content (%)	38
AT content (%)	62
LSC length (bp)	83548
SSC length (bp)	18308
IR length (bp)	24634
Gene number	127
Gene number in IR regions	34
Protein-coding gene number	84
Protein-coding gene (%)	66.14
rRNA gene number	8
rRNA (%)	6.30
tRNA gene number	35
tRNA (%)	27.56

## 3 RESULTS AND DISCUSSION

### 3.1 *H. annuus* Longkuiza 6 Chloroplast Genome Structure and Gene Annotation

The assembled and spliced *H. annuus* longkuiza 6 chloroplast genome has a full length of 151124 bp and a GC content of 38%. It has a highly conservative and typical tetrad structure. The large single copy fragment (LSC) has a length of 83548bp, the small single copy fragment (SSC) has a length of 18308bp, and a pair of inverted repeat fragments (IRa and IRb) has a length of 24634bp. A total of 127 genes were annotated in the chloroplast genome sequence of *H. annuus* longkuiza 6, including 84 protein-coding genes, 35 tRNA genes and 8 rRNA genes (Table 1). The 84 protein-coding genes can be divided into three broad categories. The first category is self-replication-related genes, the second category is photosynthesis-related genes, and the third category is other protein-coding genes and genes with unknown functions.

The IR region contains 34 genes, which are 12 protein-coding genes (2 *rpl2*, 2 *rpl23*, 2 *ycf2*, 2 *ndhB*, 2 *rps7*, 2 *rps12*), 14 tRNA genes (2 *trnI-CAU*, 2 *trnI-GAU*, 2 *trnL-CAA*, 2 *trnV-GAC*, 2 *trnA*, 2 *trnR-ACG*, 2 *trnN-GUU*), and 8 ribosomal RNA genes (2 *rrn16s*, 2 *rrn23s*, 2 *rrn4.5s*, 2 *rrn5s*) (Figure 1). The *rps12*

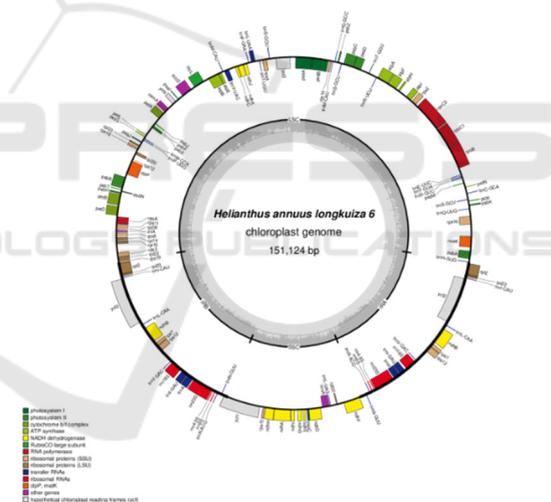


Figure 1: Gene cycle graph of the *H. annuus* longkuiza 6 chloroplast genome

### 3.2 Analysis of Introns and Exons of *H. annuus* Longkuiza 6 Chloroplast Genome

*H. annuus* longkuiza 6 chloroplast genome contains 18 genes containing introns, including 12 CDS and 6 tRNA. The *ycf3* and *clpP* genes contain 2 introns, and the other genes contain only one intron. In addition, the intron of *ndhA* is the longest at 1084 bp, and the intron of *trnL-UAA* is the shortest at 437 bp (Table 2).

Table 2: Genes with introns and exons of *H. annuus* longkuiza 6 chloroplast genome.

Gene	Start	End	Exon I	Intron I	Exon II	Intron II	Exon III
<i>rps16</i>	5216	6351	40	869	227		
<i>rpoC1</i>	16304	19105	432	732	1638		
<i>atpF</i>	27079	28346	145	713	410		
<i>ycf3</i>	42284	44234	124	699	230	745	153
<i>trnL-UAA</i>	47045	47566	38	437	47		
<i>trnV-UAC</i>	50816	51463	38	573	37		
<i>clpP</i>	69365	71327	71	751	294	621	226
<i>petB</i>	74278	75699	6	774	642		
<i>petD</i>	75894	77087	8	711	475		
<i>rpl2</i>	83705	85194	391	665	434		
<i>ndhB</i>	93287	95489	777	670	756		
<i>trnI-GAU</i>	101028	101878	42	774	35		
<i>trnA</i>	101943	102836	38	822	34		
<i>ndhA</i>	114693	116868	553	1084	539		
<i>trnA</i>	131839	132732	38	822	34		
<i>trnI-GAU</i>	132797	133647	42	774	35		
<i>ndhB</i>	139186	141388	777	670	756		
<i>rpl2</i>	149481	150970	391	665	434		

### 3.3 Codon Preference of *H. annuus* Longkuiza 6 Chloroplast Genome

In the *H. annuus* longkuiza 6 chloroplast genome, the most codon-encoded amino acid is leucine Leu, with 5111 (accounting for 10.58%); the least codon-encoded amino acid is cysteine Cys, with 537

(accounting for 1.11%). Among these codons, the most used codon is ATT, which encodes isoleucine Ile and occurs 1950 times, and the least used is TGC, which encodes Cys and 151 times. Except for tryptophan Trp and methionine Met, which have only one codon, the rest of the amino acids have 2-6 codons.

 Table 3: Codon Usage in the chloroplast genome of *H.annuus* longkuiza 6.

Amino acid	Codon	Number	Amino acid	Codon	Number	Amino acid	Codon	Number	
Val (V)	GTA	897	Met (M)	TTC	986	Gln (Q)	CAA	1291	
	GTC	345		ATG	1159		CAG	446	
	GTT	347							
Tyr (Y)	GTT	912	Lys (K)	AAG	723	Cys (C)	TGT	386	
	TAT	1482		AAA	1901		TGC	151	
	TAC	338		CTT	1115		GAT	1558	
Trp (W)	TGG	882	Leu (L)	CTG	322	Asp (D)	GAC	407	
				CTA	691				
				CTC	329				
				TTA	1530				
				TTG	1124				
				ATC	827		Asn (N)	AAC	552
				ATA	1269			AAT	1829
		ATT	1950						
Thr (T)	ACT	980	Ile (I)	CAT	849	Arg (R)	AGG	335	
	ACA	716		CAC	292		AGA	889	
	ACG	239					CGA	628	
	ACC	428					CGC	188	
	AGC	248					CGG	233	
	AGT	754					CGT	641	
Ser (S)	TCG	325	His (H)	Gly (G)	GGT	1037	Ala (A)	GCA	766
	TCA	753						GCC	396
	TCC	566						GGA	1236
	TCT	1068						GGC	363
	CCT	731						GAG	629
Pro (P)	CCG	328	Glu (E)	GAA	1839	Terminator	TAG	87	
	CCA	580					TGA	77	
	CCC	370					TAA	145	

## 4 CONCLUSIONS

In this paper, the structural characteristics of the whole chloroplast genome of *H. annuus* longkuiza 6 are studied. The results show that it is similar to the chloroplast genome structure of most plants. It has a highly conserved four-region structure, including a pair of 24634 bp inverted repeat regions (IR), a large single copy region (LSC) of 83548 bp and a small single copy region (SSC) of 18308 bp. *H. annuus* longkuiza 6 has annotated a total of 127 genes, including 84 protein-coding genes, 8 rRNA genes and 35 tRNA genes. There are 18 genes with introns, of which the *rps12* gene has a trans-spliced intron. The study of the whole genome structure and sequence information of the chloroplast of *H. annuus* longkuiza 6 laid the foundation for the study of its genetic background and the exploration of the relationship between system evolution. Using the chloroplast structural genome characteristics to explore can provide a basis for the relationship between sunflower phylogeny and evolution, and provide reference value for the future development, utilization and molecular evolution of primers for *Helianthus* plants.

## ACKNOWLEDGEMENT

The Heilongjiang Provincial Government Postdoctoral Funding (Grant Number LBH-Z17204).

## REFERENCES

- Coombe L., Warren R., Jackman S., et al. (2016). Assembly of the Complete Sitka Spruce Chloroplast Genome Using 10X Genomics' GemCode Sequencing Data. *J. Sci. PLoS one*, 11: 77-94.
- Dong F., Lin Z., Lin J., et al. (2021). Chloroplast Genome of Rambutan and Comparative Analyses in Sapindaceae. *J. Sci. Plants (Basel, Switzerland)*, 10: 283-294.
- Gao B, Yuan L, Tang T, et al. (2019). The complete chloroplast genome sequence of *Alpinia oxyphylla* Miq. and comparison analysis within the Zingiberaceae family. *J. Sci. PLoS One*, 14: 99-112.
- Hussain M., Rauf S., Riaz M., et al. (2017). Determination of drought tolerance related traits in *Helianthus argophyllus*, *Helianthus annuus*, and their hybrids. *J. Sci. Breeding science*, 67: 257-267.
- Liu Z., Gu W., Seiler G., et al. (2020). A Unique Cytoplasmic-Nuclear Interaction in Sunflower (*Helianthus annuus* L.) Causing Reduced-Vigor Plants and the Genetics of Vigor Restoration. *J. Sci. Frontiers in plant science*, 11: 110-123.
- Ng PK, Lin SM, Lim PE, et al. (2017). Complete chloroplast genome of *Gracilaria firma* (Gracilariaceae, Rhodophyta), with discussion on the use of chloroplast phylogenomics in the subclass Rhodymeniophycidae. *J. Sci. BMC Genomics*, 18: 40-47.
- Redwan R., Saidin A., Kumar S. (2015). Complete chloroplast genome sequence of MD-2 pineapple and its comparative analysis among nine other plants from the subclass Commelinidae. *J. Sci. BMC plant biology*, 15: 196-214.
- Wen F., Wu X., Li T., et al. (2021). The complete chloroplast genome of *Stauntonia chinensis* and compared analysis revealed adaptive evolution of subfamily Lardizabaloideae species in China. *J. Sci. BMC genomics*, 22: 161-174.