Analysis of Genome Characteristics of *Helianthus annuus* J-01 Chloroplast

Jun Ma¹,a*

¹Institute of Industrial Crops, Heilongjiang Academy of Agricultural Sciences, Harbin, Heilongjiang, China
*a-e-mail: wangyimajun@126.com

*Corresponding Author: a-zhangyuehua@jmsu.edu.cn

Abstract. *Helianthus annuus* J-01 was sequenced using a high-throughput sequencing platform, and the structural characteristics of the whole chloroplast genome sequence were analyzed. The results showed that the *H. annuus* J-01 chloroplast genome, like most higher plants, has a typical circular double-stranded tetrad structure. The chloroplast genome size is 151142bp, GC content is 38.99%, and a total of 127 genes have been annotated, including 84 protein-coding genes, 35 tRNA genes and 8 rRNA genes. A total of 18 genes in the *H. annuus* J-01 chloroplast genome contain introns, of which *ycf3* and *clpP* genes contain 2 introns. In protein-coding genes, leucine is the amino acid with the highest codon encoding rate, and the amino acid with the lowest encoding rate is cysteine. The A/T content of the third codon is 69.32%, and the third codon of the *H. annuus* J-01 chloroplast genome coding gene prefers to use A/T.

1. Introduction

Chloroplast is an essential organelle in plant cytoplasm. It can undergo photosynthesis and provide energy for plant growth and development. It is the basis of plant growth and development and the main factor of productivity [1-2]. In plants, the chloroplast genome contains a large amount of genetic information and is highly conserved; the self-replication and evolution of its genome remain relatively independent of species. Therefore, the chloroplast genome is often used to explore the occurrence, development and evolution of plant genomics and bioinformatics [3-4]. Thus, the complete sequence of the chloroplast genome has become a valuable tool suitable for studying molecular phylogeny and molecular ecology.

*Helianthus annuus* belongs to the Asteraceae *Helianthus* is an annual herbaceous plant and is one of the four major oil crops in the world[5]. This study takes *H. annuus* J-01 chloroplast as the research object, and conducts genomics research on it, determines the characteristics of the *H. annuus* J-01 chloroplast genome, and provides a reference for the subsequent systematic evolution and genetic diversity analysis of *Helianthus*.

2. Materials and Methods

2.1. Experimental Materials

*H. annuus* used in the experiment was planted in the Science and Technology Park of Jiamusi University, number J-01, and fresh *H. annuus* J-01 leaves were used as test materials for *H. annuus* J-01 chloroplast genome extraction.
2.2. Total DNA extraction and sequencing
Taking fresh *H. annuus* J-01 leaves, washing them with sterile water, add to liquid nitrogen and extracting the total DNA by using a plant DNA extraction kit and using agarose gel electrophoresis to detect the quality of the DNA. After reaching the sequencing requirements, *H. annuus* J-01 DNA will be sent to BGI for high-throughput sequencing, and the rest will be frozen for use.

2.3. Splicing and annotation of chloroplast genome sequence
Using stqc software to filter the original data to remove linkers on reads, bases with a quality of less than 30 at both ends, sequences containing ambiguous base N, and sequences with a length of less than 60 bp to obtain the desired high-quality data. Using SOAP denovo software to assemble the obtained Clean Data into the contig sequence, using BLAT to locate it on the chloroplast reference genome of the closed-source species, to get the relative position between the contig sequences, and then splice and correct the contig to get the full-length frame map. Fill in the gap in the sequence to get a circular chloroplast genome to complete the map sequence. Using CpGAVAS for gene annotation and get annotation results. Finally, using Organellar Genome DRAW generates a complete annotated physical map of the chloroplast genome of circular *H. annuus* J-01.

3. Results and discussion

3.1. Basic characteristics of *H. annuus* J-01 chloroplast genome
*H. annuus* J-01 chloroplast genome is a typical circular double-stranded tetrad structure, with a total length of 151142 bp, GC content is 38.99%. It includes a small single-copy fragment (SSC, 18312 bp), a large single-copy fragment (LSC, 83534 bp) and a pair of inverted repeats (IRa and IRb, 24648 bp) (Figure 1). *H. annuus* J-01 chloroplast genome has a total of 127 genes annotated, including 84 protein-coding genes, 35 tRNA genes, and 8 rRNA genes (Table 1).

The 84 protein-coding genes can be divided into three major categories. The first category includes 29 genes related to self-replication, including three subunits encoding RNA polymerase (large subunit, small subunit and DNA-dependent RNA polymerase). The second category contains 45 genes related to photosynthesis (ATP synthase subunit, NADH-dehydrogenase subunit, cytochrome b/f complex subunit, photosystem I subunit, photosystem II subunits, ribulose diphosphate oxygenase/carboxylase subunits). The third category includes 6 other genes encoding proteins and 4 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 *rrn16*, 2 *rrn23*, 2 *rrn4.5*, 2 *rrn5*). The *rps12* gene is a split gene, and its 5’end is located in the LSC region, and the 3’end is located in the two IR regions. This phenomenon is consistent with other plants[6].
Figure 1. Gene cycle graph of the *H. annuus* J-01 chloroplast genome

### Table 1. Basic characteristics of *H. annuus* J-01 chloroplasts

| Chloroplasts feature | Numerical value |
|----------------------|-----------------|
| Length (bp)          | 151,142         |
| GC content (%)       | 38.99           |
| AT content (%)       | 61.01           |
| LSC length (bp)      | 83,534          |
| SSC length (bp)      | 18,312          |
| IR length (bp)       | 24,648          |
| 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           |

3.2. *H. annuus* J-01 chloroplast genome intron information

A total of 18 genes in the *H. annuus* J-01 chloroplast genome contain introns, and the *ycf3* and *clpP* genes contain 2 introns. In addition, the intron of *ndhA* is the longest at 1088 bp, and the intron of *trnL-UAA* is the shortest at 437 bp (Table 2).
Table 2. Genes with introns of H. annuus J-01 chloroplast genome and lengths of exons and introns

| Gene     | Start  | End    | Exon I | Intron I | Exon II | Intron II | Exon III |
|----------|--------|--------|--------|----------|---------|-----------|----------|
| rps16    | 5216   | 6349   | 40     | 867      | 227     |           |          |
| rpoC1    | 16302  | 19103  | 432    | 732      | 1638    |           |          |
| atpF     | 27077  | 28344  | 145    | 713      | 410     |           |          |
| ycf3     | 42283  | 44233  | 124    | 699      | 230     | 745       | 153      |
| trnL-UAA | 47044  | 47565  | 38     | 437      | 47      |           |          |
| trnV-UAC | 50815  | 51462  | 38     | 573      | 37      |           |          |
| clpP      | 69363  | 71325  | 71     | 751      | 294     | 621       | 226      |
| petB     | 74276  | 75697  | 6      | 774      | 642     |           |          |
| petD     | 75892  | 77085  | 8      | 711      | 475     |           |          |
| rpl2     | 83691  | 85180  | 391    | 665      | 434     |           |          |
| ndhB     | 93273  | 95475  | 777    | 670      | 756     |           |          |
| trnL-GAU | 101014 | 101864 | 42     | 774      | 35      |           |          |
| trnA     | 101929 | 102821 | 38     | 821      | 34      |           |          |
| ndhA     | 114693 | 116872 | 553    | 1088     | 539     |           |          |
| trnA     | 131858 | 132750 | 38     | 821      | 34      |           |          |
| trnL-GAU | 132815 | 133665 | 42     | 774      | 35      |           |          |
| ndhB     | 139204 | 141406 | 777    | 670      | 756     |           |          |
| rpl2     | 149499 | 150988 | 391    | 665      | 434     |           |          |

3.3. Codon Usage in the chloroplast genome of H. annuus J-01

Statistics of the codons of all protein-coding genes in the H. annuus J-01 chloroplast genome found that Leu is the amino acid with the highest coding rate, and 10.61% (5099) of the codons are involved in coding; Cys is the amino acid with the lowest coding rate. 1.13% (541) of the codons are involved in coding, and the A/T content of the third codon is 69.32%. This codon preference is universally present in the chloroplast genomes of other higher plants (Table 3)[7].

Table 3. Codon Usage in the chloroplast genome of H. annuus J-01

| Amino acid | Codon | Number | Amino acid | Codon | Number | Amino acid | Codon | Number |
|------------|-------|--------|------------|-------|--------|------------|-------|--------|
| Val (V)    | GTA   | 881    | GTC        | 341   |        | GTG        | 349   |        | GTT    | 904   |
| Tyr (Y)    | TAT   | 1484   | TAC        | 336   |        | TGT        | 1484  |        |        |       |
| Trp (W)    | TGG   | 862    |            |       |        |            |       |        |        |       |
| Thr (T)    | ACT   | 978    | Ile (I)    | ATC   | 825    | Asn (N)    | AAC   | 552    |        |       |
| Ser (S)    | AGC   | 246    | His (H)    | CAT   | 837    | Arg (R)    | AGG   | 341    |        |       |
| Pro (P)    | CCT   | 729    | Gly (G)    | GGT   | 1025   | Ala (A)    | GCA   | 758    |        |       |
|            | CCG   | 328    |            | GGG   | 556    |            | GCC   | 392    |        |       |
|            | CCA   | 580    |            | GGA   | 1220   |            | GCG   | 279    |        |       |
4. Conclusions
The whole plant cp genome includes approximately 63-209 genes, mostly concentrated between 110-130 genes. The composition and arrangement are highly conserved, including photosynthesis genes, cp transcriptional expression-related genes and some other protein-coding genes[8]. In this study, the whole cp genome of *H. annuus* J-01 was assembled and analyzed, and the results showed that its whole length was 151142 bp, with a typical tetrad structure and gene sequencing. That is, it contains 4 different regions: a pair of 24648 bp inverted repeat region (IR), a large single-copy region (LSC) of 83534 bp and a small single-copy region (SSC) of 18312 bp. With a typical four-part structure, two single-copy regions are separated by a pair of inverted repeats (IRs). A total of 127 genes were annotated, including 87 protein-coding genes, 35 tRNAs, and 8 rRNAs. The use of codon 3 in the chloroplast genome protein-coding gene of *H. annuus* J-01 has apparent A/T preference, similar to other higher plants. In this study, the structural characteristics of the *H. annuus* J-01 chloroplast genome were analyzed, which provided reference value for the development, utilization and molecular evolution of primers for *Helianthus* plants in the future.

Acknowledgement
The Heilongjiang Provincial Government Postdoctoral Funding (Grant Number LBH-Z17204); “Supported by China Agriculture Research System of MOF and MARA”.

References
[1] Daniell H, Jin S, Zhu XG, et al. (2021) Green giant-a tiny chloroplast genome with mighty power to produce high-value proteins: history and phylogeny. J. Sci. Plant Biotechnol J, 19: 430-447.
[2] Zoschke R, Bock R. (2018) Chloroplast Translation: Structural and Functional Organization, Operational Control, and Regulation. J. Sci. Plant Cell, 30: 745-770.
[3] Zhang W, Zhao Y, Yang G, et al. (2019) Determination of the evolutionary pressure on Camellia oleifera on Hainan Island using the complete chloroplast genome sequence. J.Sci. PeerJ, 43: 724-733.
[4] Xiao-Ming Z, Junrui W, Li F, et al. (2017) Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. J. Sci. Sci Rep, 7:1555-1567.
[5] Vangelisti A, Mascagni F, Giordani T, et al. (2019) Arbuscular mycorrhizal fungi induce the expression of specific retrotransposons in roots of sunflower (*Helianthus annuus* L.). J. Sci. PloS one, 14: 78-86.
[6] 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.
[7] Asaf S, Khan AL, Khan MA, et al. (2018) Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species. J. Sci. PLoS One, 13: 77-89.
[8] 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.