The complete chloroplast genome of Jerusalem artichoke (*Helianthus tuberosus* L.) and *ycf2* gene comparative analysis

Qiwen Zhong Equal first author, , **Shipeng Yang** Corresp., Equal first author, , **xuemei Sun** , **Lihui Wang** , **Yi Li** Corresp.

Corresponding Authors: Shipeng Yang, Yi Li
Email address: qhyysp@163.com, jerusalemyys@aliyun.com

Jerusalem artichoke (*Helianthus tuberosus* L.) is widely cultivated in Northwest China, which has become an emerging economic crop with rapid development. Because of its elevated inulin content and high resistance, it is widely used in functional food, inulin processing, feed, and ecological management. In this study, Illumina sequencing technology was utilized to assemble and annotate the complete chloroplast genome sequences of Jerusalem artichoke. The total length was 151,431 bp, including four conserved regions: A pair of reverse repeat regions (IRa 24,568 bp and IRb 24,603 bp), a large single-copy region (LSC, 83,981 bp), and a small single-copy region (SSC, 18,279 bp). The genome had a total of 115 genes, with 19 present in the reverse direction in the IR region. 36 simple sequence repeats (SSRs) were identified in the coding and non-coding regions, most of which were biased towards A/T bases. 32 SSRs were distributed in the non-coding regions. Comparative analysis of the chloroplast genome sequence of Jerusalem artichoke and other species of the composite family revealed the chloroplast genome sequences of plants of the composite family to be highly conserved. Differences were observed in 24 gene loci in the coding region, with the degree of differentiation of the *ycf2* gene being the most obvious. Phylogenetic analysis showed *Helianthus petiolaris subsp. fallax* had the closest relationship with Jerusalem artichoke, both members of the *Helianthus* genus. Insights from our assessment of the complete chloroplast genome sequences of Jerusalem artichoke will aid in the in-depth study of the evolutionary relationship of the composite family, and provide significant sequencing information for the genetic improvement of Jerusalem artichoke.

The complete chloroplast genome of Jerusalem artichoke (*Helianthus tuberosus* L.) and *ycf2* gene comparative analysis

Qiwen Zhong[1,2,3¶], Shipeng Yang[2,3¶], Xuemei Sun[1,2,3], Lihui Wang[2,3], Yi Li[1*]

[1]Qinghai Key Laboratory of Qinghai-Tibet Plateau Biological Resources, Northwest Institute of Plateau Biology , Chinese Academy of Sciences, Xining 810000, Qinghai, China
[2]Vegetable Genetics and Physiology Laboratory, Agriculture and Forestry Sciences of Qinghai University, Xining 810016, Qinghai, China
[3]The Open Project of State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining 810016, Qinghai, China

Corresponding Author:
Yi Li[1]

Email address: jerusalemyys@aliyun.com

Abstract

Jerusalem artichoke (*Helianthus tuberosus* L.) is widely cultivated in Northwest China~~,~~ which has become an emerging economic crop with rapid development. Because of its elevated inulin content and high resistance, it is widely used in functional food, inulin processing, feed, and ecological management. In this study, Illumina sequencing technology was utilized to assemble and annotate the complete chloroplast genome sequences of Jerusalem artichoke. The total length was 151,431 bp, including four conserved regions: A pair of reverse repeat regions (IRa 24,568 bp and IRb 24,603 bp), a large single-copy region (LSC, 83,981 bp), and a small single-copy region (SSC, 18,279 bp). The genome had a total of 115 genes, with 19 present in the reverse direction in the IR region. 36 simple sequence repeats (SSRs) were identified in the coding and non-coding regions, most of which were biased towards A/T bases. 32 SSRs were distributed in the non-coding regions. Comparative analysis of the chloroplast genome sequence of Jerusalem artichoke and other species of the composite family revealed the chloroplast genome sequences of plants of the composite family to be highly conserved. Differences were observed in 24 gene loci in the coding region, with the degree of differentiation of the *ycf2* gene being the most obvious. Phylogenetic analysis showed *Helianthus petiolaris subsp. fallax* had the closest relationship with Jerusalem artichoke, both members of the *Helianthus* genus. Insights from our assessment of the complete chloroplast genome sequences of Jerusalem artichoke will aid in the in-depth study of the evolutionary relationship of the composite family, and provide significant sequencing information for the genetic improvement of Jerusalem artichoke.

Keywords: Chloroplast genome, *Helianthus tuberosus* L., *Asteraceae*, Illumina sequencing, Phylogeny

Introduction

Jerusalem artichoke (*Helianthus tuberosus* L.) is a species of the composite family [1] native to North America, mainly distributed in the temperate zone of 40-55 °C north latitude and the temperate region with the approximate similar latitude in the southern hemisphere. Jerusalem artichoke was brought to China via Europe in the 17th century. It has been grown on a small scale as a pickled vegetable in various regions of China. Jerusalem artichoke is highly resistant and can be grown in saline, alkaline, dry and low temperature conditions. Therefore, it is widely cultivated in various regions of China, especially in the Qinghai plateau in recent years. To date, most research on Jerusalem artichoke has focused on ecological management, feed research and development, and the processing of inulin products. Studies centered on the improvement of saline land in the Songnen Plain have recognized Jerusalem artichoke as an excellent improved crop, which has already been initially grown in saline-alkali grassland[2]. The overground part of Jerusalem artichoke is tall, making it an easily accessible source of animal feed. Furthermore, its leaves are especially particularly nutritious compared with other feed ingredients, being rich in lysine and methionine, and having a dry matter content of protein as high as 20%, of which 5% to 6% corresponds to lysine, an essential amino acid [3]. Jerusalem artichoke also utilizes fructan as a source of carbon, instead of starch, as most crops. Fructan can be processed or modified, which providing raw materials for the production of bioethanol, paper, and healthcare products[4-6].

The composite family is the largest group of dicotyledonous chrysanthemums, encompassing 25,000-30,000 species distributed throughout the world. 52 species and a large number of subspecies have been recognized in the *Helianthus* genus, including Jerusalem artichoke. The morphology of these plants is complex and diverse, leading to difficulties in identification and evolutionary analysis. Jerusalem artichoke is a hexaploid species (2n = 6x = 102), which reproduces mainly through vegetative propagation by tubers[7]. The evolutionary assessment of this plant is controversial, with its ancestral species being uncertain. Hybridization experiments between Jerusalem artichoke and *Helianthus annuus L.* have confirmed homologous genes between these species. It is generally believed that the chromosome number of triploid hybrid (AAB) in Jerusalem artichoke is doubled. Moreover, cytogenetic studies have demonstrated two of the three genomes of Jerusalem artichoke are homologous[8-10]. The diploid (2n = 2x = 34) B genome is provided by the immediate ancestor of *Helianthus annuus L.*, while the autotetraploid (2n = 4x = 68) A genome is provided by the crop in the composite family[11-13]. *Helianthus hirsutus*is is regarded as the most likely tetraploid ancestor [13]; while *Helianthus grosseserratus*, and *Helianthus giganteus* are viewed as the most likely diploid ancestors. Sequencing of related species using partial mitochondrial genomes as well as 35 s and 5 s ribosomal DNA has shown the origin of Jerusalem artichoke is very rich and probably linked to the hybridization of tetraploid Hairy *Helianthus annuus* L. and diploid Sawtooth *Helianthus annuus L*. [13, 14]. With the development of high-throughput sequencing technology, chloroplast phylogenetic genome evaluation has become a hot topic in the evolutionary research of plants in recent years. Plenty of phylogenetic information is contained in the chloroplast genome, providing a broad data platform for the study of phyletic evolution, and thereby verifying and extending the results of previous studies. The chloroplast genome sequencing of 8 *Helianthus* species has been completed. However, this aspect remains unexplored concerning Jerusalem artichoke.

Thus, in this study, we report the complete chloroplast genome sequencing, assembly and comparative analysis of Jerusalem artichoke. This data will help elucidate the evolutionary history of Jerusalem artichoke and its phylogenetic position in the composite family. In addition, it will lay a foundation for further studies of population genetics and other molecular aspects of Jerusalem artichoke based on chloroplast DNA sequencing.

Materials & Methods

Samples and genome sequencing

Fresh tender leaves of Jerusalem artichoke were obtained from the experimental base of the Qinghai Academy of Agricultural and Forestry Sciences (N36°43′51, E101°45′24). Chloroplast DNA was extracted through an improved high-throughput chloroplast genome extraction method [15]. Illumina HiSeq PE150 paired-end sequencing technology was used to establish the library for sequencing. The library was of the DNA small fragment type with 350 bp, with a read length was 150 bp.

Chloroplast genome assembly and annotation

FastQC was used for the quality filtering of clean data. SOAPdenovo software was used for pre-assembly [16]; while SPAdes v3.6.2 (http://bioinf.spbau.ru/spades) was used for sequence assembly [17]. The sequence of the chloroplast genome of *Helianthus annuus L.* was used as a reference to determine the location of the chloroplast genome. Gapcloser [18] and GapFiller [19] software for repairing gaps; and PrInSeS-G was then used for sequence correction. DOGMA software (http://dogma.ccbb.utexas.edu/) [20] was used for annotation. The gene region and protein coding sequence were manually adjusted according to the initiation codon and termination codon sequences. tRNA was entered into tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/) for annotation [21]. rRNA was submitted to the RNAmmer 1.2 Server (http://www.cbs.dtu.dk/services/RNAmmer/) for prediction. The resulting sequence information and annotation results were submitted to Genebank, with the sequence number of MG696658. The Organellar Genome DRAW software (http://ogdraw.mpimp-golm.mpg.de/index.shtml) [22] was used to render a complete circular chloroplast genome map.

Repeats and SSRs analysis

The chloroplast genome was entered into REPuter [23]to identify forward and reverse repeat sequences. Simple sequence repeats (SSRs) searching was identified by MIcroSAtellite (MISA) software based on perl script (http://pgrc.ipk-gatersleben.de/misa/). The number of repeats from mononucleotide to hexanucleotide was set to 10, 5, 4, 3, 3 and 3.

Comparative analysis of different *Asteraceae* plastomes

The LAGAN model in the mVISTA software [24] was used to perform a comparative analysis of the chloroplast genome of Jerusalem artichoke with *Carthamus tinctorius* (KX822074.1), *Ageratina adenophora* (JF826503.1), *Guizotia abyssinica* (EU549769.1). *Lactuca sativa* (NC_007578.1), *Helianthus argophyllus* (KU314500.1), *Helianthus debilis* (KU312928.1), and *Helianthus petiolaris subsp. fallax* (KU295560.1). After screening for the quality of the original chloroplast genome data of Jerusalem artichoke, the final constructed sequence (the gene sequence extracted from the annotation) and the established chloroplast genome of 15 plant species were compared by Blast++. HomBlocks [25] was used to construct a Circos map (http://circos.ca/) to find the reception, relative position and link color of genes. This was then standardized according to the length of all alignment regions. Coloring was performed in accordance with the long, medium, relative short, and short sequence lengths (pink, orange, green, and blue, respectively). COBALT (https://www.ncbi.nlm.nih.gov/tools/cobalt/cobalt.cgi?CMD=Web) was utilized to compare the differential protein sequence *ycf2*.

Phylogenetic analysis

The following 15 species of the composite family were used for the phylogenetic analysis of Jerusalem artichoke: *Ageratina adenophora* (JF826503.1), *Carthamus tinctorius* (KX822074.1), *Guizotia abyssinica* (NC_010601.1), *Jacobaea vulgaris* (NC_015543.1), *Lactuca sativa* (NC_007578.1), *Helianthus annuus* (NC_007977.1), *Helianthus petiolaris subsp. fallax* (KU295560.1), *Helianthus argophyllus* (KU314500.1), *Helianthus debilis* (KU312928.1), *Helianthus annuus cultivar line HA383* (DQ383815.1), *Helianthus petiolaris* (KU310904.1), *Helianthus praecox* (KU308401.1), *Helianthus annuus subsp. Texanus* (KU306406.1), *Mikania micrantha* (NC_031833.1), and *Taraxacum Mongolicum* (NC_031396.1). MAFFT 7.388 [26] was used to compare 16 chloroplast genome sequences. A phylogenetic tree was constructed with the method of maximum-likelihood and Bayesian, respectively. The GTRGAMMAI model was used in the ML Tree, and RAxML v8.1.24 [27] was used to construct the tree. Parameters were set to search for 30 repeats, and the tree with the maximum likelihood value was used. In addition, Bootstrap was set to run 1000 times to detect the credibility of each branch. To build the Bayesian tree, the nucleotide substitution model GTR+I+G in Bayesian analysis was selected according to BIC in the jModelTest 2.1.7 software [28]. MrBayes 3.2 [29] was used for calculations, employing the Markov chain Monte Carlo methodology. Four Markov chains were initialized at the same time. The random tree was marked as the initial tree, and one was saved every 500 trees for a total of 5,000,000 trees. The first 20% of the Burin-in trees was discarded. The remaining trees were used to calculate the posterior possibility of the consistent tree and each branch.

Results and Discussion

Genome organization and gene features

The chloroplast genome of Jerusalem artichoke had a total length of 151,431 bp. The genome was composed of four parts: A pair of reverse repeat regions, IRa (24,568 bp) and IRb (24,603 bp), separated by a large single-copy region LSC (83981bp) and a small single-copy region SSC (18,279 bp) (Fig. 1). Genes in the coding regions accounted for 55.45% of the genome, including protein-coding genes, tRNA genes and rRNA genes. The chloroplast genome of Jerusalem artichoke had a total guanine-cytosine content (G-C content) of 37.6%; with GC in the IR region corresponding to 43.2%, and GC in the LSC and SSC regions being 35.6% and 31.3%, respectively. This may be due to the fact that the IR region contained four high-GC rRNA genes [30]. High G-C content made conservatism in the IR regions higher than that in the large single-copy (LSC) and small single-copy (SSC) regions [31].

The chloroplast genome of Jerusalem artichoke contained 115 genes, including 84 protein-coding genes CDS, 27 tRNA genes and 4 rRNA genes distributed in the IR region. Furthermore, this region encompassed 19 inverse genes, including 8 CDS genes (*ycf2*, *ndhB*, *rps7*, *rps12*, *ycf15*, *ycf1*, *rpl2*, and *rpl23*), 7 tRNA genes, and 4 rRNA genes. The 115 genes contained 60 Protein synthesis and DNA replication genes, 44 Photosynthesis genes, 6 Miscellaneous group genes and 5 pseudogenes of unknown function genes (Table 1). The sequence and composition of chloroplast genes of Jerusalem artichoke were similar to those of other crops of the composite family [32].

Introns play an important role in selective gene splicing. In the chloroplast genome of Jerusalem artichoke, 16 intron-containing genes were annotated, 11 of which were protein-encoding and 5 were tRNA genes. Of the 16 intron genes, the intron sequence in *trnK-UUU* was the longest (2,528 bp), while the intron in the trnL-UAA gene was the smallest (436 bp). There were two introns in the *clpP*, *ycf3* and *rps12* genes, whereas the other genes contained only one intron (Table 2).

Repeats and SSRs analysis

Because the chloroplast genome was simple, relatively conservative and maternal, chloroplast SSR were highly efficient molecular markers. Moreover, chloroplast simple sequence repeats (cpSSRs) have been widely used previously in crossbreeding, biogeography, and population genetics studies [33]. Distribution of cpSSR in Jerusalem artichoke was analyzed, revealing 36 different SSR loci in its chloroplast genome. Among them, 32 SSR were composed of A or T, 2 were composed of C, and only 1 was composed of G; indicating the chloroplast genomic SSR of Jerusalem artichoke are biased towards A/T bases. This is consistent with the chloroplast genomes of most angiosperms[31, 34]. In regards to repeat length, most SSR had 10-20 bp, while fewer had less than 10 bp, indicating the SSR segment of the Jerusalem artichoke chloroplast genome is short. However, the long repeated sequence might promote the rearrangement of the chloroplast genome, causing an increase in population genetic diversity [35]. This may be related to the vegetative propagation of Jerusalem artichoke, which greatly reduces the probability of genetic variation. Assessment of SSR distribution found 32 SSR in the non-coding region of the chloroplast genome. The non-coding region mainly includes intergenic spacer (IGS) and introns, accounting for 68% and 20% of the distribution, respectively. In the coding region, there are SSR only in the *rpoC2*, *cemA*, and *ycf1* genes. These repetitive structures provide valuable information resources for the future development of molecular markers in the study of the phylogenetic evolution and population genetics of Jerusalem artichoke.

Comparative analysis of different composite chloroplast

Comparative analysis with the plastomes of other species of the composite family revealed only small differences in plastome size and composition in comparison to that of Jerusalem artichoke (Table 3). There were very few inconsistencies in the types and number of chloroplast genes in several species of the composite family, and the performance was very conserved. The chloroplast genome of Jerusalem art ichoke ranked 5th in the aligned genomes of the 8 chloroplast genomes of the composite family. Length variation in the sequence may be caused by the difference in length between the LSC and IR regions. The chloroplast genome size of 8 crops of the composite family was approximately 150 kb, with a GC content of approximately 37.5%. The number of coding protein genes ranged between 79-89. All of these genomes had 4 rRNA-coding genes and 20-30 tRNA-coding genes. The plastome of Jerusalem artichoke was 327 bp longer than that of *Helianthus petiolaris subsp. fallax* (a crop in the same genus), mainly in the LSC region. In addition, it had 5 more protein-coding genes than that of *Helianthus petiolaris subsp. fallax*, with no difference in the number of rRNA- and tRNA-coding genes. The length variations of the chloroplast genomes of 8 species of the composite family correlated with the lengths of the IR regions, indicating the length of IR region had a significant effect on the length of genome [36].

The genomic sequences of 8 composite species were analyzed by the mVISTA software, detecting the variations of the sequences (Fig. 3). Results showed there was less variation between

Jerusalem artichoke, *Helianthus petiolaris subsp. fallax* and *Helianthus debilis* and *Helianthus argophyllus*. Compared with *Ageratina adenophora*, partial structure was lacking in Jerusalem artichoke. Comparative analysis of coding regions in the chloroplast genome of plants in the composite family showed Jerusalem artichoke and *Helianthus petiolaris subsp. fallax* had the least differences. As a whole, the chloroplast genome of crops in the composite family tends to be conserved. mVISTA analysis showed the coding region was more conserved than the non-coding region, which is consistent with reports on crops in the composite family *such as Cynara cardunculus* [32] and *Ageratina adenophora* [37]. The *ycf2* gene showed the greatest degree of differentiation. In addition, there was a gene deletion in the crops of genus *Helianthus.*

Based on the results of mVISTA, a systematic comparative analysis was performed in a coding region with small variation amplitude [38]. As shown in Figure 4, there were differences among 8 species of the composite family in the following 24 gene loci: *trnN-GUU*, *trnR-ACG*, *trnA-UGU*, *ycf68*, *trnL-GAU*, *trnV-GAC, ycf15*, *rps7*, *ndhB*, *trnL-CAA*, *ycf2*, *trnL-CAU*, *rpl23*, *rpl2*, *rps19*, *rps12*, *rpl20*, *rps18, rpl33*, *trnP-UGG*, *petL*, *trnG-UCC*, *trnS-GCU*, and *trnC-GCA*. The discovery of these differential genes provides valuable phylogenetic information for the further evaluation of the composite family. At present, many different gene regions are considered potential tools for phylogenetic analysis. These DNA domains will play an important role in the application of molecular phylogeny in this species [37].

The *ycf2* gene is the largest known plastid gene in angiosperms [39]. Although the *ycf2* gene can be used to predict phylogenetic relationships[40], its function remains unclear. In many studies, the *ycf2* gene has become an alternative choice for the assessment of plant sequence variation and phylogenetic evolution. Our results showed the *ycf2* gene segment had large deletion and inconsistency. The *ycf2* gene of Jerusalem artichoke and seven other composite species was compared. Four species of genus *Helianthus* had 152 amino acid sequence deletions of *ycf2* gene in the segment 308-460. In addition, only *Helianthus petiolaris* had 12 amino acid sequence deletions in the segment 1524-1536 among four *Helianthus* species. There were 12 amino acid sequence deletions in the segment 1641-1653 of *Ageratina adenophora* and *Lactuca sativa,* as well as in the segment 1641-1664 of *Guizotia abyssinica*. In addition, there were some amino acid site differences. Ultimately, the was greatest similarity was observed between the *ycf2* genes of Jerusalem artichoke and *Helianthus petiolaris subsp. fallax*, except for the presence of 5 additional amino acids in the initial site of *ycf2* in the Jerusalem artichoke plastome. This suggests the *ycf2* gene is very conserved in the evolution of the species within the composite family. The *ycf2* gene appears to gradually degenerate compared in gramineous crops, with only 734 bp remaining in rice and wheat [41]. The results of phylogenetic tree analysis using partial angiosperm *ycf2* genes were consistent with those obtained from the whole plastid genome data phylogenetic tree analysis. This provides even more precise details for evolutionary evaluation.[38]

Phylogenetic analysis

The composite family is one of the largest families in the plant kingdom, and the chloroplast genome plays an important role in plant classification and phylogenetic analysis. To date, abundant research has evaluated the phylogeny of crops in the composite family. Notably, study of the evolution of the *Aster spathulifolius* chloroplast genome has revealed it bears its closest relationship with *Jacobaea vulgaris* [42-44]. To assess the phylogenetic relationships of Jerusalem artichoke, the chloroplast genomes of 15 species of the composite family were compared globally. *Jacobaea vulgaris* was taken as an outgroup, and then RAxML and Bayesian evolutionary trees were constructed respectively. The resulting phylogenetic trees constructed by the two methods shared the same topological structure (Figure 6). All species in the composite family formed three highly supported evolutionary clades: Members of the genus *Helianthus* are included in the first branch, including some *Helianthus annuus L.* species, subspecies and Jerusalem artichoke, as well as *Eupatorieae* and *Millerieae*. On the evolutionary branches of the genus *Helianthus*, Jerusalem artichoke and *Helianthus petiolaris subps. fanax* are in the closest relationship. The common node bootstrap is fully resolved. *Lactuca sativa* and *Taraxacum offcinale* of Crepidinae are contained in the second branch, while *Jacobea vulgaris* is clustered in Senecioninae alone, which is consistent with previous reports on the uncertainty of the evolution of the Senecioninae tribe [38]. In the group of the composite in which the number of involved species more than or equal to 2, it can be seen that genetic relationship of Jerusalem artichoke is more closely to other species of composite family, genus *Helianthus*. At the same time, Jerusalem artichoke is also the earliest isolated species of the genus *Helianthus*. This provides a theoretical basis for the further study of the relationship between phylogenetic branches of Jerusalem artichoke in the composite family.

Conclusions

In this study, the complete chloroplast genome sequence of Jerusalem artichoke was successfully assembled, annotated and analyzed. The chloroplast genome of plants in the composite family is relatively conservative. Variations of the chloroplast genome are scarce between Jerusalem artichoke and plants in the same genus. Compared with composite plants belonging to other genera, we found deletions in the chloroplast genome of Jerusalem artichoke. The identification of repetitive sequences in the chloroplast genome of Jerusalem artichoke, especially SSR, will be helpful for the development of molecular markers, the study of population genetics and the phylogenetic analysis of Jerusalem artichoke. Phylogenetic analysis of plants in the composite family shows Jerusalem artichoke and Helianthus petiolaris subsp. fallax share the closest relationship, both belonging to the composite family, genus Helianthus. Completion of the sequencing of the chloroplast genome will provide key genetic information for further research on Jerusalem artichoke and deepen our understanding on the evolutionary history of the chloroplast genome and phylogenetic position of Jerusalem artichoke. In addition, it may be useful for various molecular biology applications of Jerusalem artichoke in the future.

Acknowledgements

Funding

This project was supported by the National Natural Science Foundation of China (31460523, 31660588, 31660569, 31760600), The Open Project of Qinghai Key Laboratory of Qinghai-Tibet Plateau Biological Resources (2017-ZJ-Y10), Fundamental Research Program of

Qinghai(2016-ZJ-751, 2017-ZJ-Y18). There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

National Natural Science Foundation of China: 31460523, 31660588, 31660569, 31760600

The Open Project of Qinghai Key Laboratory of Qinghai-Tibet Plateau Biological Resources: Grant No. 2017-ZJ-Y10

Fundamental Research Program of Qinghai: 2016-ZJ-751, 2017-ZJ-Y18

Competing interests

The authors declare that they have no competing interests.

Author Contributions

_ Qiwen Zhong performed the experiments, wrote the paper.

_ ShipengYang analyzed the data, prepared figures and/or tables.

_ Xuemei Sun performed the experiments.

_ Lihui Wang contributed reagents/materials/analysis tools.

_ Yi Li conceived and designed the experiments, reviewed drafts of the paper.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences: NCBI: MG696658

The sequences have been provided as Supplemental Dataset Files.

References

1. Garcia PM, Hayashi AH, Silva EA, Figueiredo-Ribeiro Rde C, Carvalho MA. Structural and metabolic changes in rhizophores of the Cerrado species Chrysolaena obovata (Less.) Dematt. as influenced by drought and re-watering. Frontiers in plant science. 2015;6:721. doi: 10.3389/fpls.2015.00721. PubMed PMID: 26442035; PubMed Central PMCID: PMC4585265.

2. Yan X, Li Y, Wang Y. Jerusalem artichoke,an optimal plant for the improvement of alkalic grassland in Songnen Plain,China. Journal of Natural Science of Heilongjiang University. 2008.

3. Rawate PD, Hill RM. Extraction of a high-protein isolate from Jerusalem artichoke (Helianthus tuberosus) tops and evaluation of its nutrition potential. Journal of Agricultural and Food Chemistry. 1985;33(1):29-31. doi: 10.1021/jf00061a008.

4. Wyse DL, Young FL, Jones RJ. Influence of Jerusalem Artichoke (Helianthus tuberosus) Density and Duration of Interference on Soybean (Glycine max) Growth and Yield. Weed Science. 2017;34(2):243-7. Epub 06/12. doi: 10.1017/S0043174500066753.

5. Wang Y-Z, Zou S-M, He M-L, Wang C-H. Bioethanol production from the dry powder of Jerusalem artichoke tubers by recombinant Saccharomyces cerevisiae in simultaneous saccharification and fermentation. Journal of Industrial Microbiology & Biotechnology. 2015;42(4):543-51. doi: 10.1007/s10295-014-1572-7.

6. Saengkanuk A, Nuchadomrong S, Jogloy S, Patanothai A, Srijaranai S. A simplified spectrophotometric method for the determination of inulin in Jerusalem artichoke (Helianthus tuberosus L.) tubers. European Food Research and Technology. 2011;233(4):609. doi: 10.1007/s00217-011-1552-3.

7. Baldini M, Danuso F, Turi M, Vannozzi GP. Evaluation of new clones of Jerusalem artichoke (Helianthus tuberosus L.) for inulin and sugar yield from stalks and tubers. Industrial Crops and Products. 2004;19(1):25-40. doi: https://doi.org/10.1016/S0926-6690(03)00078-5.

8. Kostoff D. Autosyndesis and structural hybridity in F1-hybrid Helianthus tuberosus L. x Helianthus annuus L. and their sequences. Genetica. 1939;21(5):285-300. doi: 10.1007/bf01508121.

9. Atlagić J, Dozet B, ŠKorić D. Meiosis and Pollen Viability in Helianthus tuberosus L. and its Hybrids with Cultivated Sunflower. Plant Breeding. 1993;111(4):318-24. doi: doi:10.1111/j.1439-0523.1993.tb00648.x.

10. Kostoff D. A Haploid Plant of Nicotiana sylvestris. Nature. 1934;133:949. doi: 10.1038/133949b0.

11. Heiser CB, Smith DM. SPECIES CROSSES IN HELIANTHUS: II. POLYPLOID SPECIES. Rhodora. 1964;66(768):344-58.

12. Heiser CB, Smith DM, Clevenger SB, Martin WC. THE NORTH AMERICAN SUNFLOWERS (HELIANTHUS). Memoirs of the Torrey Botanical Club. 1969;22(3):1-218.

13. Bock DG, Kane NC, Ebert DP, Rieseberg LH. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. New Phytologist. 2014;201(3):1021-30. doi: doi:10.1111/nph.12560.

14. Timme RE, Simpson BB, Linder CR. High-resolution phylogeny for Helianthus (Asteraceae) using the 18S-26S ribosomal DNA external transcribed spacer. American Journal of Botany. 2007;94(11):1837-52. doi: doi:10.3732/ajb.94.11.1837.

15. Shi C, Hu N, Huang H, Gao J, Zhao Y-J, Gao L-Z. An Improved Chloroplast DNA Extraction Procedure for Whole Plastid Genome Sequencing. PloS one. 2012;7(2):e31468. doi: 10.1371/journal.pone.0031468.

16. Lee WI, Lee G. From natural language to shell script: A case-based reasoning system for automatic UNIX programming. Expert Systems with Applications. 1995;9(1):71-9. doi: https://doi.org/10.1016/0957-4174(94)00050-6.

17. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology. 2012;19(5):455-77. doi: 10.1089/cmb.2012.0021. PubMed PMID: 22506599.

18. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1(1):18. doi: 10.1186/2047-217x-1-18.

19. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biology. 2012;13(6):R56. doi: 10.1186/gb-2012-13-6-r56.

20. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004;20(17):3252-5. doi: 10.1093/bioinformatics/bth352.

21. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Research. 2016;44(W1):W54-W7. doi: 10.1093/nar/gkw413.

22. 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. doi: 10.1093/nar/gkt289.

23. Kurtz S, ., Choudhuri JV, Ohlebusch E, ., Schleiermacher C, ., Stoye J, ., Giegerich R, . REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research. 2001;29(22):4633-42.

24. Frazer KA, Lior P, Alexander P, Rubin EM, Inna D. VISTA: computational tools for comparative genomics. Nucleic Acids Research. 2004;32(Web Server issue):W273.

25. Bi G, Mao Y, Xing Q, Cao M. HomBlocks: A multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. Genomics. 2018;110(1):18-22. doi: https://doi.org/10.1016/j.ygeno.2017.08.001.

26. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics. 2017:bbx108-bbx. doi: 10.1093/bib/bbx108.

27. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3. doi: 10.1093/bioinformatics/btu033.

28. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods. 2012;9:772. doi: 10.1038/nmeth.2109 https://www.nature.com/articles/nmeth.2109#supplementary-information.

29. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Systematic Biology. 2012;61(3):539-42. doi: 10.1093/sysbio/sys029.

30. Asaf S, Khan AL, Khan AR, Waqas M, Kang S-M, Khan MA, et al. Complete Chloroplast Genome of Nicotiana otophora and its Comparison with Related Species. Frontiers in plant science. 2016;7(843). doi: 10.3389/fpls.2016.00843.

31. Yang Y, Yuanye D, Qing L, Jinjian L, Xiwen L, Yitao W. Complete Chloroplast Genome Sequence of Poisonous and Medicinal Plant Datura stramonium: Organizations and Implications for Genetic Engineering. PloS one. 2014;9(11):e110656. doi: 10.1371/journal.pone.0110656.

32. Curci PL, De Paola D, Danzi D, Vendramin GG, Sonnante G. Complete Chloroplast Genome of the Multifunctional Crop Globe Artichoke and Comparison with Other Asteraceae. PloS one. 2015;10(3):e0120589. doi: 10.1371/journal.pone.0120589.

33. Bayly MJ, Rigault P, Spokevicius A, Ladiges PY, Ades PK, Anderson C, et al. Chloroplast genome analysis of Australian eucalypts – Eucalyptus, Corymbia, Angophora, Allosyncarpia and Stockwellia (Myrtaceae). Molecular Phylogenetics and Evolution. 2013;69(3):704-16. doi: https://doi.org/10.1016/j.ympev.2013.07.006.

34. Raveendar S, Na Y-W, Lee J-R, Shim D, Ma K-H, Lee S-Y, et al. The Complete Chloroplast Genome of Capsicum annuum var. glabriusculum Using Illumina Sequencing. Molecules. 2015;20(7):13080. PubMed PMID: doi:10.3390/molecules200713080.

35. Qian J, Song J, Gao H, Zhu Y, Xu J, Pang X, et al. The Complete Chloroplast Genome Sequence of the Medicinal Plant Salvia miltiorrhiza. PloS one. 2013;8(2):e57607. doi: 10.1371/journal.pone.0057607.

36. Guo H, Liu J, Luo L, Wei X, Zhang J, Qi Y, et al. Complete chloroplast genome sequences of Schisandra chinensis: genome structure, comparative analysis, and phylogenetic relationship of basal angiosperms. Science China Life Sciences. 2017;60(11):1286-90. doi: 10.1007/s11427-017-9098-5.

37. Nie X, Lv S, Zhang Y, Du X, Wang L, Biradar SS, et al. Complete Chloroplast Genome Sequence of a Major Invasive Species, Crofton Weed (Ageratina adenophora). PloS one. 2012;7(5):e36869. doi: 10.1371/journal.pone.0036869.

38. Doorduin L, Gravendeel B, Lammers Y, Ariyurek Y, Chin-A-Woeng T, Vrieling K. The Complete Chloroplast Genome of 17 Individuals of Pest Species Jacobaea vulgaris: SNPs, Microsatellites and Barcoding Markers for Population and Phylogenetic Studies. DNA Research. 2011;18(2):93-105. doi: 10.1093/dnares/dsr002.

39. Drescher A, Ruf S, Calsa T, Carrer H, Bock R. The two largest chloroplast genome ‐ encoded open reading frames of higher plants are essential genes. The Plant Journal. 2000;22(2):97-104. doi: doi:10.1046/j.1365-313x.2000.00722.x.

40. Drescher A, Ruf S, Calsa T, Carrer H, Bock R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. The Plant Journal. 2000;22(2):97-104. doi: doi:10.1046/j.1365-313x.2000.00722.x.

41. Matsuoka Y, Yamazaki Y, Ogihara Y, Tsunewaki K. Whole Chloroplast Genome Comparison of Rice, Maize, and Wheat: Implications for Chloroplast Gene Diversification and Phylogeny of Cereals2003. 2084-91 p.

42. Choi KS, Park S. The complete chloroplast genome sequence of Aster spathulifolius (Asteraceae); genomic features and relationship with Asteraceae. Gene. 2015;572(2):214-21. doi: https://doi.org/10.1016/j.gene.2015.07.020.

43. HUANG J-L, SUN G-L, ZHANG D-M. Molecular evolution and phylogeny of the angiosperm ycf2 gene. Journal of Systematics and Evolution. 2010;48(4):240-8. doi: doi:10.1111/j.1759-6831.2010.00080.x.

44. SOLTIS DE, SOLTIS PS, CHASE MW, MORT ME, ALBACH DC, ZANIS M, et al. Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. Botanical Journal of the Linnean Society. 2000;133(4):381-461. doi: doi:10.1111/j.1095-8339.2000.tb01588.x.

Figure 1

Gene map of the *Helianthus tuberosus* L. chloroplast genome.

Genes drawn outside of the circle are transcribed counter-clockwise, while genes shown on the inside of the circle are transcribed clockwise. Genes belonging to different functional groups are color-coded. The darker gray in the inner circle indicates GC content, while the lighter gray corresponds to AT content.

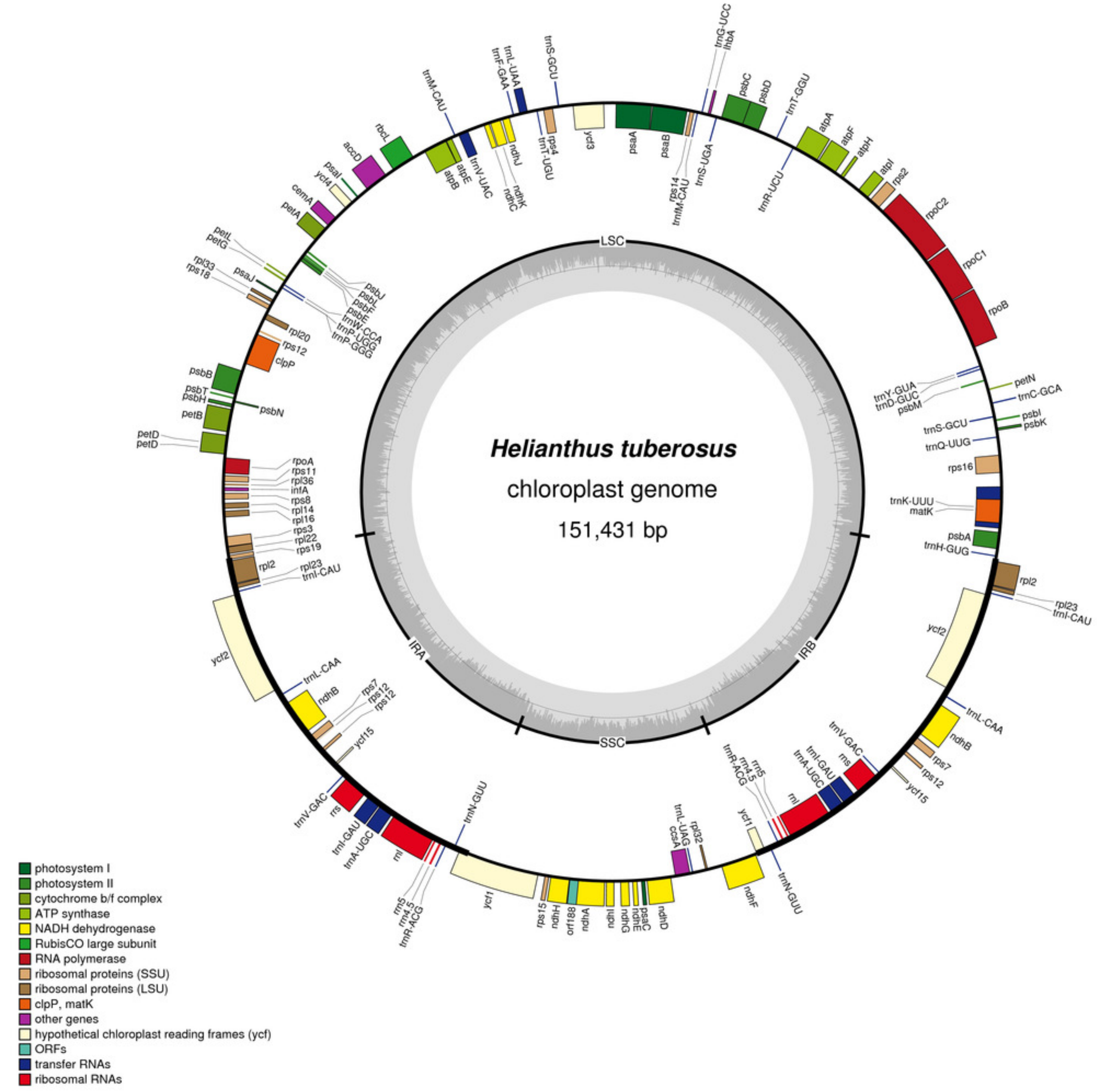

Helianthus tuberosus
chloroplast genome
151,431 bp
LSC
IRA
IRB
SSC
trnG-UCC
ttnA
psbC
psbD
trnT-GGU
atpA
atpF
atpH
atpI
rps2
rpoC2
rpoC1
rpoB
trnY-GUA
trnD-GUC
psbM
petN
trnC-GCA
psbI
psbK
trnS-GCU
trnQ-UUG
rps16
trnK-UUU
matK
psbA
trnH-GUG
rpl2
rpl23
trnI-CAU
ycf2
trnL-CAA
ndhB
rps7
rps12
ycf15
trnV-GAC
rrs
trnI-GAU
trnA-UGC
rrl
rrn5
rrn4.5
trnR-ACG
ycf1
trnN-GUU
ndhF
rpl32
trnL-UAG
ccsA
ndhD
psaC
ndhE
ndhG
ndhI
ndhA
orf188
ndhH
rps15
trnS-GCU
trnL-UAA
trnF-GAA
trnM-CAU
rbcL
accD
psaI
ycf4
cemA
petA
petL
petG
rpl33
rps18
psaJ
psbJ
psbL
psbF
psbE
trnW-CCA
trnP-UGG
trnP-GGG
rpl20
rps12
clpP
psbB
psbT
psbH
psbN
petB
petD
rpoA
rps11
rpl36
infA
rps8
rpl14
rpl16
rps3
rpl22
rps19
atpE
atpB
trnV-UAC
ndhJ
ndhK
ndhC
rps4
trnT-UGU
ycf3
psaA
psaB
rps14
trnfM-CAU
trnS-UGA
trnR-UCU
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)
ORFs
transfer RNAs
ribosomal RNAs

Figure 2

Distribution frequency in *Helianthus tuberosus* L. cp genome

a The frequency of repeats, length of repeats; Number of repeats. b The percentage distribution of gene area

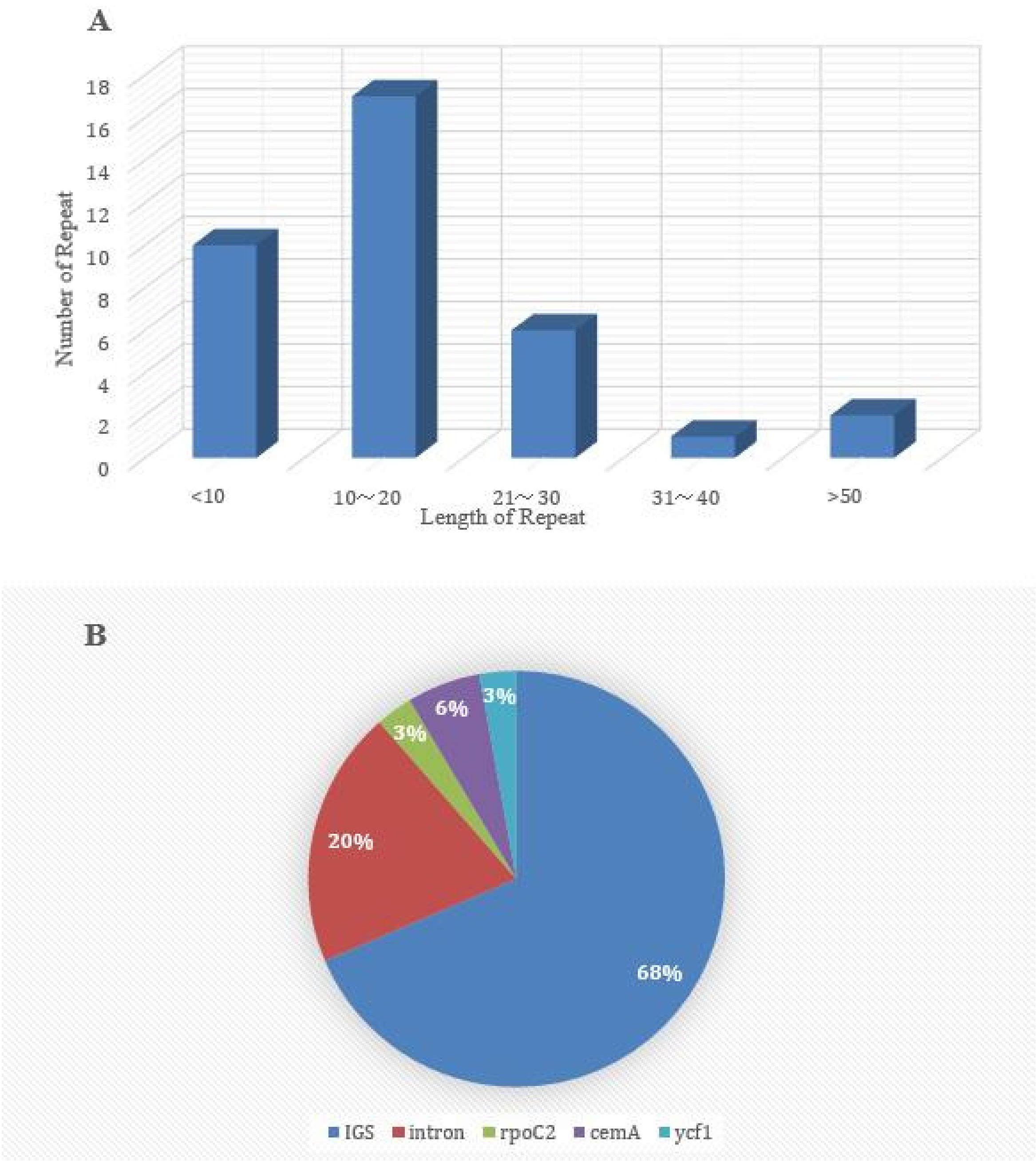


Figure 3

Percent identity plot for the comparison of 8 composite chloroplast genomes.

The whole chloroplast genome was divided into four parts, and the gene names are displayed in sequence on the top line of each part (arrows indicate the transcriptional direction). The sequence similarity of the alignment region of Jerusalem artichoke and seven other species is shown as the filling color in each black stripe. The x-axis indicates the position of the chloroplast genome at a certain site, and the y-axis indicates the average sequence identity percentage (50-100%) with Jerusalem artichoke on the position of a species at a certain position (50-100%). The coding sequences (exons), rRNA, tRNA and the conserved non-coding sequences (CNS) in the genomic region are represented with different colors.

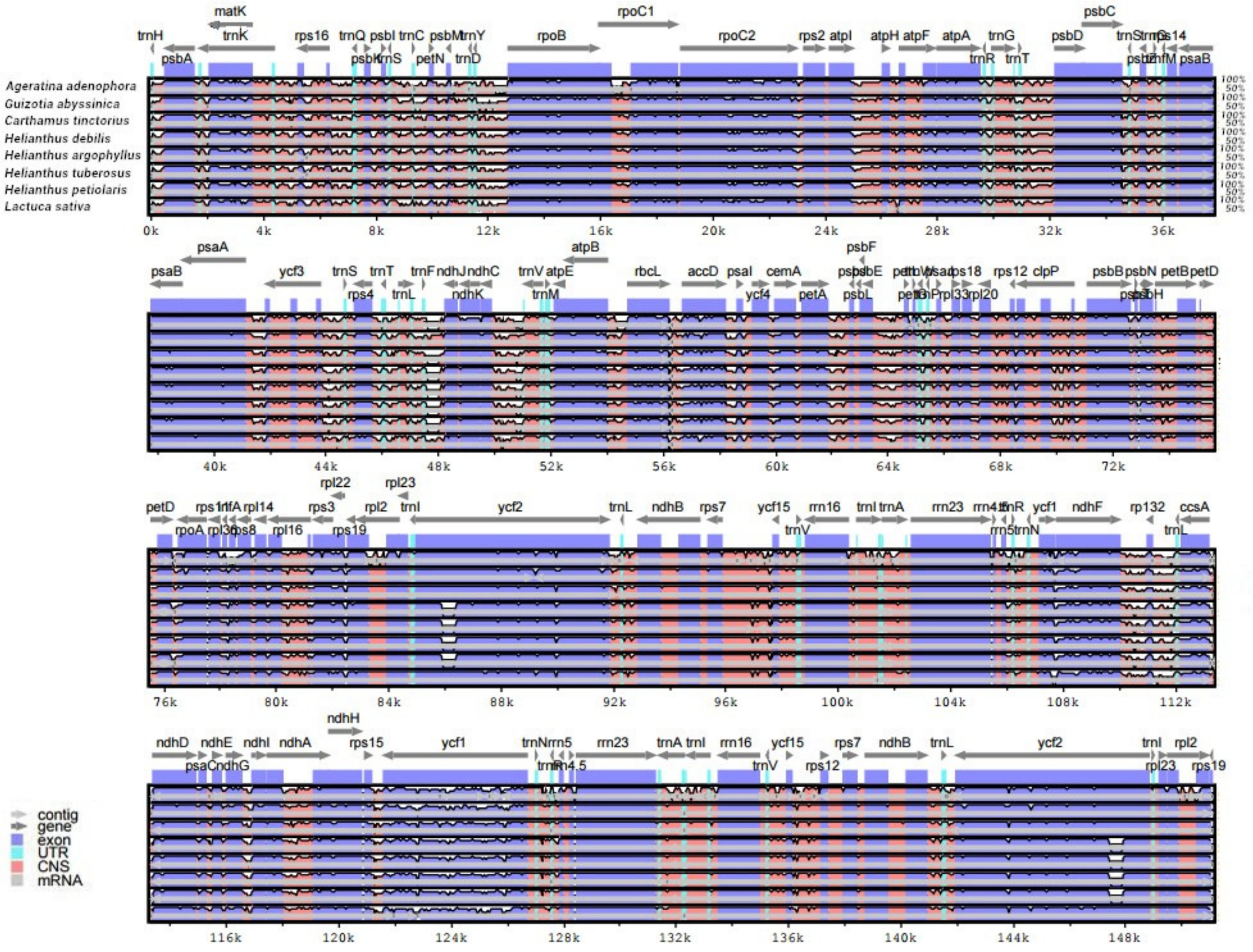


Figure 4

Comparison of the similarity of chloroplast genomes between Jerusalem artichoke and seven other species of crops in the composite family.

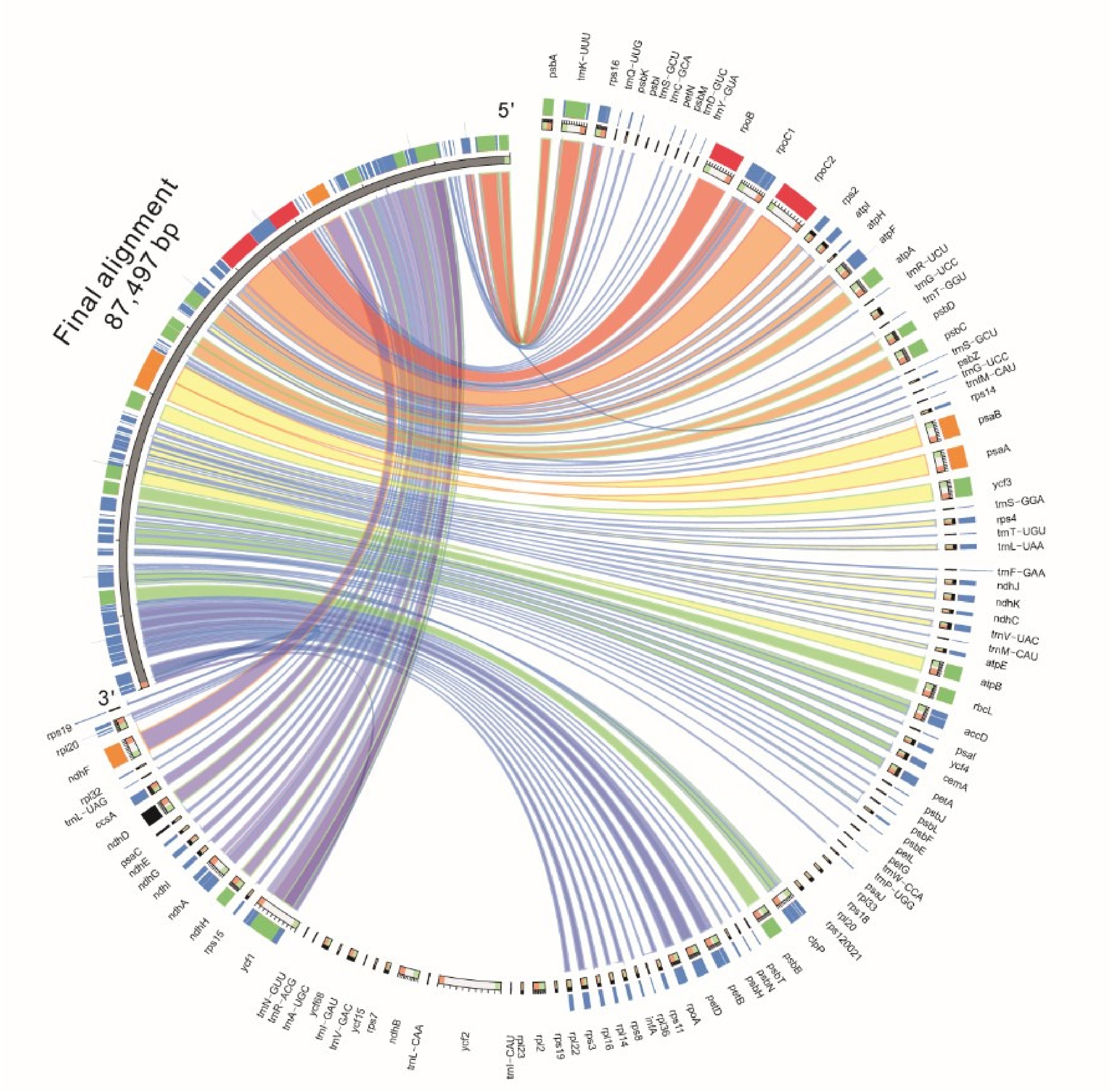
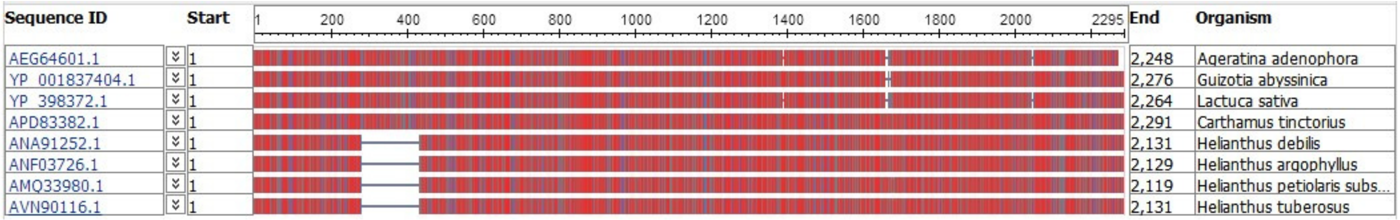


Figure 5

Comparison of the *ycf2* gene sequence in chloroplast genomes between Jerusalem artichoke and seven other species of crops in the composite family.

The white vacancy corresponds to the missing amino acid sequence.



Sequence ID	Start	End	Organism
AEG64601.1	1	2,248	Ageratina adenophora
YP_001837404.1	1	2,276	Guizotia abyssinica
YP_398372.1	1	2,264	Lactuca sativa
APD83382.1	1	2,291	Carthamus tinctorius
ANA91252.1	1	2,131	Helianthus debilis
ANF03726.1	1	2,129	Helianthus argophyllus
AMQ33980.1	1	2,119	Helianthus petiolaris subs...
AVN90116.1	1	2,131	Helianthus tuberosus

Figure 6

Molecular phylogenetic tree of 16 composite species based on a neighbor joining analysis.

Numbers above and below nodes are bootstrap support values 50%.

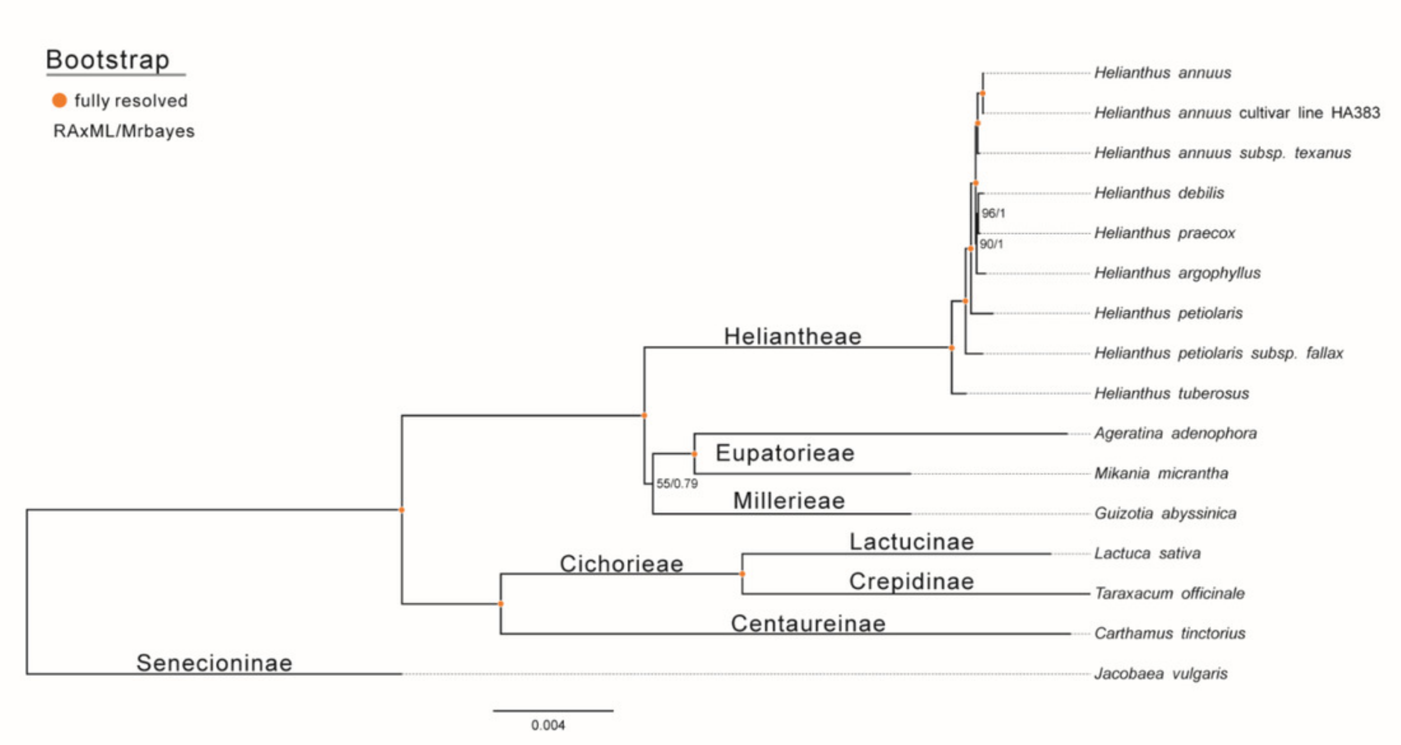


Table 1(on next page)

List of genes in the chloroplast genome of *Helianthus tuberosus* L.

Table 1 List of genes in the chloroplast genome of *Helianthus tuberosus* L.

	Groups of genes	***Names of genes***
Protein synthesis and DNA replication	Ribosomal RNAs	*16S r RNA(2×), 23S r RNA(2×), 4.5S r RNA(2×), 5S r RNA(2×)*
	Transfer RNAs	*trnQ-TTG, trnL-TAG, trnD-GTC, trnS-GGA ,trnE-TTC, trnS-GCT, trnY-GTA, trnV-GAC, trnP-TGG, trnH-GTG, trnF-GAA, trnN-GTT, trnT-TGT, trnW-CCA, trnS-TGA, trnV-GAC, trnL-CAA(2×), trnM-CAT(2×), trnC-GCA, trnI-CAT, trnT-GGT, trnI-CAT, trnR-ACG, trnN-GTT, trnR-TCT, trnR-ACG, trnG-GCC*
	Ribosomal protein small subunit	*rps7, rps14,rps12, rps2, rps4, rps12, rps7, rps11, rps16, rps12, rps19(2×), rps3, rps15, rps8, rps19*
	Ribosomal protein large subunit	*rpl14, rpl23, rpl36, rpl2, rpl20, rpl2, rpl32, rpl16, rpl33, rpl23, rpl22*
	Subunit s of RNA polymerase	*rpoB, rpoC(2×), rpoA,*
Photosynthesis	Photosystem I	*psaC, psaA, psaB, psaI, psaJ*
	Photosystem II	*psbZ, psbK, psbB, psbI, psbF, psbN, psbL, psbJ, psbC, psbE, psbM, psbH, psbA, psbD, psbT,*
	Cytochrome b/f complex	*petA, petD, petL, petB, petG, petN*
	ATP synthase	*atpE, atpH, atpA, atpI, atpF, atpB*
	NADH-dehydrogenase	*ndhJ, ndhA, ndhK(2×), ndhG, ndhI, ndhB(2×), ndhH, ndhE, ndhD, ndhC, ndhF,*
	Large subunit Rubisco	*rbcL*
Miscellaneous group	Translation initiation factor IF-1	*infA*
	Acetyl-CoA carboxylase	*accD*
	Cytochrome c biogenesis	*ccsA(2×)*

	Maturase	*matK*
	ATP-dependent protease	*clpP*
	Inner membrane protein	*cemA*
Pseudogenes of unknown function	Conserved hypothetical chloroplast open reading frame	*ycf15(4×), ycf4, ycf3, ycf1(2×), ycf2(2×)*

Table 2(on next page)

Characteristics of genes including introns and exons in the chloroplast genome of *Helianthus tuberosus* L.

Table 2 Characteristics of genes including introns and exons in the chloroplast genome of *Helianthus tuberosus* L.

Gene	Region	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
*trn*K-UUU	LSC	51	2528	36		
*rps*16	LSC	29	864	226		
*rpo*C1	LSC	431	733	1727		
*atp*F	LSC	144	714	391		
*ycf*3	LSC	152	746	229	700	123
*trn*L-UAA	LSC	36	436	49		
*trn*V-UAC	LSC	36	574	37		
*clp*P	LSC	68	792	290	624	227
*pet*B	LSC	5	775	641		
*pet*D	LSC	8	712	473		
*rpl*2	LSC	392	663	434		
*ndh*B	IR	755	671	776		
*trn*I-GAU	IR	41	776	34		
*trn*A-UGC	IR	37	822	34		
ndhA	SSC	552	1095	538		
rps12	LSC-IR	113		230		29

Table 3(on next page)

Comparison of cp genomes among 8 composite species

Table 3 Comparison of cp genomes among 8 composite species

Species	Size(bp)				G+C(%)	Total number of genes			GeneBank accessions
	Total	LSC	IR	SSC		Protein-coding genes	rRNAs	tRNAs	
Carthamus tinctorius	153675	83606	25407	19156	37.4	89	4	30	KX822074
Ageratina adenophora	150689	84815	23755	18358	37.5	80	4	28	JF826503
Guizotia abyssinica	150689	82855	24777	18277	37.3	79	4	29	HQ234669
Lactuca sativa	152772	84105	25034	18599	37.5	78	4	20	DQ383816
Helianthus tuberosus	151431	83981	24568	18279	37.6	84	4	27	MG696658
Helianthus argophyllus	151862	83845	24588	18149	37.6	80	4	27	KU314500
Helianthus debilis	151678	83799	24502	18121	37.6	82	4	27	KU312928
Helianthus petiolaris subsp. fallax	151104	83530	24633	18308	37.6	79	4	27	KU295560

2