Complete Mitochondrial Genome Sequence of Sunflower (*Helianthus annuus* L.)

Christopher J. Grassa,ᵃ* Daniel P. Ebert,ᵃ* Nolan C. Kane,ᵃ* Loren H. Rieseberg[a,b]

Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canadaᵃ; Department of Biology, Indiana University, Bloomington, Indiana, USAᵇ

* Present address: Christopher J. Grassa, Department of Organismic and Evolutionary Biology and Harvard University Herbaria, Harvard University, Cambridge, Massachusetts, USA; Daniel P. Ebert, The Beef Industry Centre, University of New England, Armidale, New South Wales, Australia; Nolan C. Kane, EBIO Department, University of Colorado, Boulder, Colorado, USA.

C.J.G. and D.P.E. contributed equally to this article.

This is the first complete mitochondrial genome sequence for sunflower and the first complete mitochondrial genome for any member of *Asteraceae*, the largest plant family, which includes over 23,000 named species. The master circle is 300,945-bp long and includes 27 protein-coding sequences, 18 tRNAs, and the 26S, 5S, and 18S rRNAs.

Address correspondence to Christopher J. Grassa, cj_grassa@fas.harvard.edu.

We present the sunflower's (*Helianthus annuus* L.) complete mitogenome based on the male-fertile oil-seed line HA412. The annual sunflowers, including the wild *H. annuus*, are endemic to North America and are adapted to a wide variety of habitats (1). Together, they are an important model system for studying evolution and ecology, particularly reticulate evolution and the genetic and ecological processes leading to speciation (2). Sunflower is also a globally important hybrid oilseed crop (3, 4) with production valued at $20 billion annually. Mitochondrial-based cytoplasmic male sterility is employed for hybrid production.

Leaf tissue from 10-day-old seedlings was enriched for mitochondria by centrifugation, to a purity of over 99% mitochondrial DNA. DNA was sequenced on 1/48th of an Illumina lane, producing 2,727,097,000 bp of sequence data. Reads were trimmed for quality and plastid contamination (5) with Trimmomatic (6) and BWA (7), and then assembled with SOAPdenovo (8). The *de novo* assembly was digested *in silico* and aligned (9) to a previously published restriction map (10). The genome was finished by hand using Illumina and Roche 454 reads and annotated using the Mitofy software (11).

The genome's master replication circle is 300,945 bp in length with a G+C content of 45%. It includes a large repeat 12,933 bp in length and two single-copy regions, measuring 51,681 bp and 223,398 bp. Alignments of short reads to the reference support the hypothesis that this structural configuration is rare. Rather, the genome's predominant configuration is two equimolar circular chromosomes, each containing one copy of the large repeat and either the large or small single copy sequence (10). The genome contains a total of seven sequences at least 200 bp in length, repeated with at least 98% identity, and several other smaller repetitive sequences. A 265-bp repeat is present in three copies.

The genome includes 18 tRNA loci. Six are similar to those commonly found in plant plastids, but are not perfectly identical to those of the sunflower's plastid. There are two *tRNA-fM* loci and one plastid-like *tRNA-M* locus. The *tRNA-I* locus contains a CAU anticodon, suggesting that it is modified posttranscription. The 26S, 5S, and 18S rRNAs are present. The genome includes at least 27 protein-coding sequences. Two genes, *rps3* and *mttB*, begin with an ATT start codon. Five sequences are homologous to the sunflower's plastid genome.

Just 25,611 bp, approximately 8.5%, of the genome could be functionally annotated. An additional 8,149 bp appear to be pseudogenes in various states of decay. The sunflower's mitogenome is repetitive and sparsely populated with genes. This is typical for a plant, but stands in stark contrast with the streamlined mitogenomes of animals. This reference is expected to facilitate the guided assembly of the mitogenomes of hundreds of sequenced sunflower accessions, as well as other *Asteraceae*, and will be an important resource for plant breeders and evolutionary biologists.

Accession number(s). This organelle genome project has been deposited in GenBank under the accession number KF815390. The version described in this paper is the first version, KF815390.1.

REFERENCES
1. **Renaut S, Grassa CJ, Moyers BT, Kane NC, Rieseberg LH.** 2012. The population genomics of sunflowers and genomic determinants of protein

evolution revealed by RNAseq. Biology **1**:575–596. http://dx.doi.org/10.3390/biology1030575.

2. **Andrew RL, Kane NC, Baute GJ, Grassa CJ, Rieseberg LH**. 2013. Recent nonhybrid origin of sunflower ecotypes in a novel habitat. Mol Ecol **22:**799–813. http://dx.doi.org/10.1111/mec.12038.

3. **Baute GJ, Kane NC, Grassa CJ, Lai Z, Rieseberg LH**. 2015. Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. New Phytol **206:**830–838. http://dx.doi.org/10.1111/nph.13255.

4. **Hulke BS, Grassa CJ, Bowers JE, Burke JM, Qi L, Talukder ZI, Rieseberg LH**. 2015. A unified single nucleotide polymorphism map of sunflower (*Helianthus annuus* L.) derived from current genomic resources. Crop Sci **55:**1696–1702. http://dx.doi.org/10.2135/cropsci2014.11.0752.

5. **Timme RE, Kuehl JV, Boore JL, Jansen RK**. 2007. A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. Am J Bot **94:**302–312. http://dx.doi.org/10.3732/ajb.94.3.302.

6. **Bolger AM, Lohse M, Usadel B**. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:**-2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.

7. **Li H, Durbin R**. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics **25:**1754–1760. http://dx.doi.org/10.1093/bioinformatics/btp324.

8. **Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J**. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience **1:**18. http://dx.doi.org/10.1186/2047-217X-1-18.

9. **Smith TF, Waterman MS**. 1981. Identification of common molecular subsequences. J Mol Biol **147:**195–197. http://dx.doi.org/10.1016/0022-2836(81)90087-5.

10. **Siculella L, Palmer JD**. 1988. Physical and gene organization of mitochondrial DNA in fertile and male sterile sunflower. CMS-associated alterations in structure and transcription of the *atp*A gene. Nucleic Acids Res **16:**3787–3799. http://dx.doi.org/10.1093/nar/16.9.3787.

11. **Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD**. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol **27:**1436–1448. http://dx.doi.org/10.1093/molbev/msq029.