OnLine Journal of Biological Sciences

Original Research Paper

Comparative Genomics of Domesticated and Wild Sunflower: Complete Chloroplast and Mitochondrial Genomes

[1]M.S. Makarenko, [1]A.V. Usatov, [1]N.V. Markin, [1]K.V. Azarin, [2]O.F. Gorbachenko and [1]N.A. Usatov

[1]*Department of Genetics, Southern Federal University, Rostov-on-Don, Russia*
[2]*Zhdanov Don Experiment Station, All Russia Research Institute of Oil Crops, pos. Oporny, Rostov region, Russia*

Article history
Received: 17-11-2015
Revised: 18-01-2016
Accepted: 11-02-2016

Corresponding Author:
M.S. Makarenko,
Department of Genetics,
Southern Federal University,
Rostov-on-Don, Russia
Email: mcmakarenko@yandex.ru

Abstract: The entire chloroplast and mitochondrial genomes of domesticated and wild type sunflower were sequenced**.** The comparative analysis of chloroplast genomes revealed 43 variant sites, including 21 polymorphic SSR loci and 22 SNPs. About 14 variant sites were found by collation of mitochondrial DNA (mtDNA), among them 4 SSRs, 8 SNPs and 2 deletions. About 9 SNPs were located in coding region of chloroplast DNA (cpDNA) and single SNP was mapped in mitochondrial gene. Only three SNPs caused amino acid changes: Two SNPs in cpDNA and one mtDNA SNP. Despite the fact that sunflower mitochondrial genome sequence is twice as long as chloroplast genome sequence, mtDNA has one third as much variant sites than cpDNA.

Keywords: Whole Genome Sequencing, cpDNA, mtDNA, Sunflower

Introduction

Whole Genome Sequencing (WGS) is now a common technique in contemporary plant research (Balakrishnan *et al*., 2015). WGS provides an opportunity to investigate nucleotide diversity much faster and more accurate than hybridization-based methods (molecular beacons, microarrays etc.), enzyme-based methods (RFLP, many PCR methods, HRM) or Sanger sequencing technique. Due to WGS recently there has been rapidly increasing number of genomes in GeneBank, predominantly small genomes as bacterial or organelle genomes (Phan and Nguyen, 2013). Mostly, WGS data are used in phylogenetic analysis. For plant phylogenetic analysis, typically chloroplast DNA (cpDNA) is used. Whole chloroplast genomes of buckwheat (Logacheva *et al*., 2008), citrus (Carbonell-Caballero *et al*., 2015), mulberry (Kon and Yang, 2015) etc. have revealed relationships between valuable cultivated plants and their wild ancestries. For understanding of plant phylogeny, mitochondrial DNA (mtDNA) sequences also are significant (Knoop *et al*., 2011), although, for such purpose, complete mitochondrial genomes are used less often than chloroplast. Relatively few papers provide phylogenetic analysis on data of both extranuclear genomes. Moreover, phylogenetic trees based on chloroplast genomes and mitochondrial genomes of the same objects may have quite different topologies (Bock *et al*., 2014). In addition, WGS data make possible creation of DNA markers (SSR, CAPS etc). As well as complete genome sequences can be used for developing specific transformation vectors (Chen *et al*., 2011) or other genetic engineering applications.

In the present work we studied the polymorphism of the chloroplast and mitochondrial genomes of *Helianthus annuus*. We sequenced the cpDNA and mtDNA of wild and domestic sunflower and identified the polymorphic loci that can be used as DNA targets for extranuclear genomes genotyping. These data will also be useful for future phylogenetic analysis.

Materials and Methods

Plant Material

The study was carried out on two forms of *Helianthus annuus*: Cultivated line (№ 3629) and wild type (№ 398941). Sample seeds were received from seed bank of the N.I. Vavilov Research Institute of Plant Industry. The origin of cultivated line is Rostov region (Russia), the ancestor of inbreeded cultivated line 3629 was high oil variety of Zhdanov Don Experiment Station collection. Since 1965 domesticated line has been cultivated at the station of

Southern Federal University, in strict isolation. Wild type sunflower has its origin from California region (USA), while it has cultivated in isolation conditions of Vavilov Research Institute of Plant station in Krasnodar region (Russia) since 1977. Both types of sunflower are highly inbred lines. Domesticated sunflower has stem without lateral shoots and single large inflorescence. However, wild sunflower has fruticose habit and large number of small inflorescences (40-80).

Mitochondrial and Chloroplast DNA Isolation

Before performing DNA extraction, chloroplast and mitochondrial fractions were isolated from 10 day sunflower seedlings according to the method of Triboush *et al*. (1998) with our modifications. Briefly, 1 g of leaves was homogenized by mortar and pestle in STE buffer (0,4M sucrose, 50 mMTris pH 7.8, 20 mM EDTA-Na2, 0.2% bovine serum albumin, 0.2% b-mercaptoethanol) and then centrifuged. The organelle fractions were isolated by centrifuging the homogenate at 2,000 g for 15 min, discarding the pellet and centrifuging the supernatant at 14,000 g for 15 min. DNA was extracted from the precipitate by PhytoSorb kit (Syntol, Russia), according to the manufacture's instruction.

NGS Library Preparation

For library preparation 40 ng of DNA were sheared using Covaris S220 system. NEBNext Ultra DNA Library Prep Kit (New England Biolabs, UK) was used for further manipulation. All library preparation steps were done pursuant to manual. According to Agilent 2100 Bioanalyzer data, NGS libraries length was, mainly, 450-550 bp. Libraries were quantified using Qubit (Invitrogen, USA) fluorimeter and qPCR, then diluted up to final concentration of 8 pM.

Sequencing and NGS Data Analysis

Diluted libraries were clustered on a paired-end flow cell using cBot instrument and sequenced in 100 cycles using HiSeq2000 sequencer with TruSeqSBSKitv3-HS (Illumina, USA). A total number of 2,806,411 100-bp paired reads were generated for domesticated sunflower and 2,058.566 reads for wild type. Quality of reads was determined by Fast QC. Trimming of adapter-derived and low quality (Q-score below 30) reads was performed with Trimmomatic software (Bolger *et al*., 2014). Using Bowtie2 tool (Langmead and Salzberg, 2012) sequencing reads were aligned to reference sequences (NCBI accessions NC_007977.1 and NC_023337.1). Variant calling was made by samtools/bcftools software (Li, 2011) and manually revised using IGV tool (Thorvaldsdóttir *et al*., 2013).

Results

Obtained NGS data allowed us to get complete sequences of domesticated and wild sunflower extranuclear genomes. The overall alignment rate for both genomes was more than 50% of total read number. The average read coverage was more than 800 for chloroplast genomes and more than 100 for mitochondrial genomes. These data were sufficient for a qualitative variant calling. Comparative analysis of chloroplast genomes of domestic and wild sunflower revealed 43 variant sites (Table 1). Among them 21 (48.8%) variations were in simple sequence repeats length and 22 (51.2%) were SNPs. Most presented polymorphic sites (20 SSRs, 16 SNPs) were located in large single copy region of chloroplast, the other 7 (1 SSR, 6 SNPs) sites were mapped in small single copy region. It is interesting to note, that inverted repeat region of chloroplast genomes had identical sequences.

The most common polymorphic mononucleotide repeat was poly T- 47.6%, then followed poly A- 38.1%, poly C- 9.5% and poly G- 4.8%. The percentage of transitions was 59.1%, among them 40.9% (9 substitutions) A/G and 18.2% (4 substitutions) C/T. 40.9% transversions were as follows: A/C- 22.7% (5 substitutions), A/T- 13.6% (3 substitutions), G/T- 4.6% (1 substitution). Remarkable that C/G tranversion was not presented among 22 SNPs.

About 9 SNPs were located in Intergenic Region (IGR), 4 SNPs were presented in noncoding gene regions. 7 SNPs of coding regions were synonymous and only 2 were nonsynonymous (Table 1). Comparison of cultivated lines 3629 and HA383 (presented in GenBank) in our previous study has revealed 12 polymorphic sites- 8 SSRs and 4 SNPs (Markin *et al*., 2015).

Comparative analysis of mitochondrial genomes of domestic and wild sunflower revealed 14 variant sites (Table 2). Among them 4 (28.6%) variations were in simple sequence repeat length, 8 (57.1%) were SNPs and 2 sites (14.3%) had deletions.

The variable mononucleotide repeats were presented by two poly T, one poly C and one poly A. Among single nucleotide substitutions there were 2 transition mutations: A/G and C/T. Transversion mutations were as follows: 3 G/T, 2 A/C and 1 C/G. Seven out of eight SNPs were located in IGR and the last one was mapped in coding DNA sequence. This SNP results in amino acid change at 232-d position (Ser232Tyr) of protein encoded by *nad6*.

Comparison of complete mitochondrion sequences of cultivated lines 3629 and HA412 (presented in GenBank) allowed detecting 6 polymorphic sites: 5 SSRs and 1 SNP.

Table 1. Polymorphic sites of chloroplast genomes of domestic and wild type sunflower

Position in reference genome	Variation type	Domesticated line	Wild type	Localization	Substitution type
206	SSR	$(A)_{11}$	$(A)_{12}$	IGR (trnH-GUG-psbA)	
1991	SSR	$(T)_{9}$	$(T)_{10}$	trnK-UUU (intron)	
2032	SSR	$(T)_{12}$	$(T)_{13}$	trnK-UUU (intron)	
5450	SSR	$(C)_{9}$	$(C)_{8}$	rps16 (intron)	
5653	SNP	A	C	rps16 (intron)	
5692	SSR	$(T)_{13}$	$(T)_{11}$	rps16 (intron)	
9883	SSR	$(A)_{8}$	$(A)_{9}$	IGR (trnC-GCA-petN)	
12984	SSR	$(T)_{15}$	$(T)_{11}$	IGR (trnE-UUC-rpoB)	
16887	SNP	C	A	rpoC1 (intron)	
17424	SSR	$(G)_{9}$	$(G)_{8}$	rpoC1 (intron)	
20660	SNP	T	C	rpoC2	nonsynonymous (Leu490Pro)
24141	SNP	A	C	rps2	nonsynonymous (Gln178His)
25466	SSR	$(A)_{10}$	$(A)_{13}$	IGR (atpI-atpH)	
28373	SSR	$(T)_{15}$	$(T)_{16}$	IGR (atpF-atpA)	
29701	SNP	G	A	atpA	synonymous
30166	SSR	$(A)_{10}$	$(A)_{12}$	IGR (trnR-UCU-trnG-UCC)	
35885	SSR	$(A)_{9}$	$(A)_{8}$	IGR (psbZ-trnG-GCC)	
39980	SNP	G	A	psaA	synonymous
41995	SSR	$(T)_{9}$	$(T)_{10}$	IGR (psaA-ycf3)	
43668	SNP	T	A	ycf3 (intron)	
46980	SNP	A	G	IGR (trnT-UGU-trnL-UAA)	
50163	SSR	$(T)_{10}$	$(T)_{12}$	IGR (ndhC-trnV-UAC)	
50764	SSR	$(T)_{11}$	$(T)_{8}$	IGR (ndhC-trnV-UAC)	
51778	SSR	$(T)_{10}$	$(T)_{9}$	IGR (trnM-CAU-atpE)	
54286	SNP	A	T	IGR (atpB-rbcL)	
54313	SNP	G	A	IGR (atpB-rbcL)	
58654	SNP	A	G	IGR (accD-psaI)	
60017	SSR	$(C)_{6}$	$(C)_{7}$	IGR (ycf4-cemA)	
64145	SSR	$(A)_{9}$	$(A)_{8}$	IGR (psbE-petL)	
64668	SNP	C	T	IGR (psbE-petL)	
70584	SSR	$(T)_{9}$	$(T)_{10}$	clpP (intron)	
73600	SSR	$(A)_{10}$	$(A)_{8}$	psbT	synonymous
74157	SNP	T	A	IGR (psbH-petB)	
76370	SNP	A	C	petD (intron)	
80718	SNP	G	A	rpl16	synonymous
81057	SNP	G	A	IGR (rpl16-rps3)	
81555	SNP	A	G	IGR (rpl16-rps3)	
108541	SNP	C	A	ycf1	synonymous
108697	SNP	G	A	ycf1	synonymous
117833	SNP	C	T	ndhG	synonymous
118438	SNP	G	T	IGR (ndhG-ndhE)	
122970	SSR	$(A)_{13}$	$(A)_{12}$	IGR (trnL-UAG-rpl32)	
123004	SNP	T	C	IGR (trnL-UAG-rpl32)	

Table 2. Polymorphic sites of mitochondrial genomes of sunflower domestic line and wild type.

Position in reference genome	Variation type	Domesticated line	Wild type	Localization	Substitution type
36360	SNP	T	G	IGR (nad4L-atp8)	
46039	Deletion	A	-	IGR (rpl5-trnD)	
49272	SSR	(C)11	(C)9	IGR (rpl5-trnD)	
75332	SNP	A	C	IGR (rpl10-trnM)	
116777	SNP	G	T	IGR (atp9-rps4)	
169028	SNP	G	T	nad6	nonsynonymous (Ser232Tyr)
170184	SSR	(T)13	(T)12	IGR (trnP-trnF)	
178406	SSR	(T)9	(T)8	IGR (trnS-cob)	
190813	SNP	G	A	IGR (cob-ccmFc)	
202672	SNP	T	C	IGR (orf873-atp1)	
209335-6	Deletion	AA	--	IGR (atp1-ccmFn	
230112	SNP	A	C	IGR (ccmFn-rpl16)	
248266	SSR	(A)11	(A)10	IGR (rpl16-matR)	
269062	SNP	G	C	IGR I(trnW-atp6)	

Discussion

According to data obtained from extranuclear genomes analysis of domestic and wild sunflower, a few assumptions could be established. The first one is that mtDNA has less total number of polymorphic sites, than cpDNA. In chloroplast genome 0.146 SNP accounted for 1 kb of sequence, in mitochondrial genome this characteristics is 0.027 (5.4 fold lower). However, this may be due to conservatism of mtDNA, because, plant mitochondrial genes evolve slowly than chloroplast genes (Page and Holmes, 1998).

Sunflower mitochondrial genome contains about 22.8 kb of CDS and frequency of SNP in CDS was 0.04/1kb. In Chloroplast CDS (total CDS length is 78.5 kb) 0.11 SNP accounted for 1 kb of sequence, so the difference of SNP frequency in coding region is 2.75 fold. Another assumption could be made, that sunflower chloroplast CDS evolve approximately 2.5-3 faster than mitochondrial CDS. For real establishment of this supposition we have not enough data, but according to published data the rate of substitutions in mitochondrial and chloroplast angiosperms genes is 1:3 (Drouin *et al*., 2008; Duminil, 2014). It is interesting to note, that the frequency of nonsynonymous SNP is, conversely, 1.6 fold higher in mtDNA (0.04/1kb), than in cpDNA (0.025/1kb). Although this feature may be present due to lack of data.

The sequence data of genomes have been obtained using only one inbred domestic line and one wild line and could not demonstrate the variety of all wild genotypes. However, we would expect the same SNP, especially in mitohondrial DNA, in other wild types of sunflower with different origins. So the Russian cultivated line 3629 and American cultivated line HA383 (NCBI accession NC_007977.1) has only 4 polymorphic SNP sites in cpDNA and mtDNA comparison of 3629 and HA412 (NCBI accession NC_023337.1) lines revealed only one SNP. The revealed polymorphic sites could be useful for molecular markers development. In future studies we plan to investigate these polymorphic sites in wild types of sunflower with diverse ancestry.

Conclusion

The comparative analysis of domesticated and wild sunflower chloroplast genomes revealed 43 variant sites, including 21 polymorphic SSR loci and 22 SNPs. About 14 variant sites were found by collation of mitochondrial DNA (mtDNA), among them 4 SSRs, 8 SNPs and 2 deletions. About 9 SNPs were located in multiple coding regions of chloroplast DNA and the frequency of SNP in CDS was 0.11/1kb. A single SNP in mitochondrial gene was detected, so the frequency of SNP in CDS was 0.04/1kb. Only three SNPs caused amino acid changes-two cpDNA SNP and one mtDNA SNP. Despite the fact that the complete mitochondrial genome sequence is twice as long as chloroplast genome sequence, mtDNA has one third as much variant sites than cpDNA.

Acknowledgement

This research was supported by Ministry of Education and Science of Russian Federation, project no. 40.91.2014/K.

Author's Contributions

All the six authors equally participated in the laboratory study, data analysis and the entire process of the article preparation**.**

Ethics

This article is original and contains unpublished material. The authors declare that there is no conflict of interest regarding publication of this paper. The authors declare that no ethical issues are going to arise after the work has been published.

References

Balakrishnan, K.N., A.A. Abdullah, Y. Abba, J.A. Bala and F.J. Abdullah *et al*., 2015. Closing the Gaps in rat cytomegalovirus ALL-03 (Malaysian Strain) genomic scaffold. Am. J. Anim. Vet. Sci. 10: 133-140. DOI: 10.3844/ajavsp.2015.133.140

Bock, D.G., N.C. Kane, D.P. Ebert and L.H. Rieseberg, 2014. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: Neither from Jerusalem nor an artichoke. New Phytol., 201: 1021-1030. DOI: 10.1111/nph.12560

Bolger, A.M., M. Lohse and B. Usadel, 2014. Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics, 30: 2114-20. DOI: 10.1093/bioinformatics/btu170

Carbonell-Caballero, J., R. Alonso, V. Ibanez, J. Terol and M. Talon *et al*., 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus *Citrus.* Molecular Biol. Evolut., 32: 2217-2218. DOI: 10.1093/molbev/msv101

Chen, G.Q., R. Thilmony and J. Lin, 2011. Transformation of lesquerella fendleri with the new binary vector pGPro4-35S. OnLine J. Biol. Sci., 11: 90-95. DOI: 10.3844/ojbsci.2011.90.95

Drouin, G., H. Daoud and J. Xia, 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Molecular Phylogenetics Evolut., 49: 827-831. DOI: 10.1016/j.ympev.2008.09.009

Duminil, J., 2014. Mitochondrial genome and plant taxonomy. Meth. Molecular Biol., 1115: 121-140. DOI: 10.1007/978-1-62703-767-9_6

Knoop, V., U. Volkmar, J. Hecht and F. Grewe, 2011. Mitochondrial Genome Evolution in the Plant Lineage. In: Plant Mitochondria, Kempken, F. (Ed.), Springer, ISBN: 978-0-387-89780-6, pp: 3-29.

Kon, W. and J. Yang, 2015. The complete chloroplast genome sequence of *Morusmongolica* and a comparative analysis within the Fabidae clade. Current Genet. DOI: 10.1007/s00294-015-0507-9

Langmead, B. and S. Salzberg, 2012. Fast gapped-read alignment with Bowtie 2. Nat. Meth., 9: 357-359. DOI: 10.1038/nmeth.1923

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27: 2987-2993. DOI: 10.1093/bioinformatics/btr509

Logacheva, M.D., T.H. Samigullin, A. Dhingra and A.A. Penin, 2008. Comparative chloroplast genomics and phylogenetics of Fagopyrumesculentum ssp. ancestrale -a wild ancestor of cultivated buckwheat. BMC Plant Biol., 8:59. DOI: 10.1186/1471-2229-8-59

Markin, N.V., A.V. Usatov, M.D. Logacheva, K.V. Azarin and O.F. Gorbachenko *et al*., 2015. Study of chloroplast DNA polymorphism in the sunflower (*Helianthus* L.). Russian J. Genet., 51: 745-751. DOI: 10.1134/S1022795415060101

Page, R.D.M. and E.C. Holmes, 1998. Molecular Evolution: A Phylogenetic Approach. 1st Edn., Wiley-Blackwell, ISBN-10: 0865428891, pp: 352.

Phan, T.H. and D.L. Nguyen, 2013. The bacterial chromosomal sequence and related issues. Am. J. Virol., 2: 1-7. DOI: 10.3844/ajvsp.2013.1.7

Thorvaldsdóttir, H., J.T. Robinson and J.P. Mesirov, 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Briefings Bioinform., 14: 178-192. DOI: 10.1093/bib/bbs017

Triboush, S.O., N.G. Danilenko and O.G. Davydenko, 1998. A method for isolation of chloroplast DNA and mitochondrial DNA from sunflower. Plant Molecular Biol. Reporter, 16: 183-189. DOI: 10.1023/A:1007487806583