American Journal of Biochemistry and Biotechnology

Original Research Paper

Comparative Analysis of the Complete Chloroplast Genome of the Alloplasmic Sunflower (*Helianthus* L.) Lines with Various CMS Types

[1]Kirill Azarin, [1]Maxim Makarenko, [1]Alexander Usatov, [1]Vladimir Khachumov, [1]Alexey Kovalevich, [2]Oleg Gorbachenko and [3]Vera Gavrilova

[1]Southern Federal University, Rostov-on-Don, Russia
[2]Zlhdanov Don Experiment Station, All Russia Research Institute of Oil Crops,pos. Oporny, Rostov region, Russia
[3]N. I. Vavilov Institute of Plant Genetic Resources, Saint Petersburg, Russia

Article history
Received: 22-12-2017
Revised: 22-02-2018
Accepted: 1-03-2018

Corresponding Author:
Kirill Azarin
Southern Federal University,
Rostov-on-Don, Russia
Email: sunlitart@yandex.ru

Abstract: The complete chloroplast genomes of sunflower fertile line HA89 and isonuclear CMS lines with four different cytoplasmic backgrounds (PET1, PET2, ANN2 and MAX) were sequenced. A total of 451 polymorphic sites, with including 58 SSRs, 317 SNPs and 76 microindels were revealed between the fertile and CMS cytotypes. Among the alloplasmic male-sterile lines, cpDNA of CMS-MAX had the largest number of polymorphisms. The lowest number of polymorphic sites was revealed in CMS-PET1. Like as CMS-PET1, CMS-PET2 was obtained as a result of interspecific crossing between *H. petiolaris* and cultivated sunflower *H. annuus*. Nevertheless, the number of INDELs and SNPs in CMS-PET2 chloroplast genome was more than 4-fold and 6.5-fold higher than that in cpDNA of CMS-PET1. The average frequency of SNPs and INDELs in the non-genic regions and genic regions were 0.0062 and 0.0046, respectively. Increased mutation rates were found in the *psbM-rpoB*, *rps16 intron*, *atpA-psbD*, *rps4-ndhJ* and *ndhc-atpE* non-coding regions, as well as in the *rpoC2*, *atpA*, *rbcL*, *ndhF* and *ycf1* genes. In addition to short insertions and deletions ranging from 1 to 5 bp, the relatively long INDELs (14-24 bp) unique for each CMS line were found. These insertions and deletions may be of use for PCR differentiation of the CMS lines due to differences in the amplicons length.

Keywords: Complete Chloroplast Genome, Alloplasmic Lines, CMS, Sunflower

Introduction

The success of a high productivity hybrids breeding with tolerant to a number of environmental factors largely depends on the genetic potential of the parental lines. In the overwhelming majority of the analysis of genetic diversity in plants, the main emphasis is on the combination of nuclear alleles (Pervaiz *et al*., 2015). At the same time, the potential of cytoplasmic variability and new cytoplasmic-nuclear combinations are practically not taken into account. The nuclear genome plays a significant role in the ontogenesis of plants, however, the effects of cytoplasmic genes has now been proven both on the expression of quantitative traits so on the adaptive potential of plants to extreme environmental factors (Mashkina *et al*., 2010). Indeed, plastid DNA, which account for only a few percent of the total cellular DNA, is involved in the realization of vital plant functions such as photosynthesis (Jansen and Ruhlman, 2012). Nuclear and cytoplasmic genomic interactions are confirmed by studies of the simultaneously variability both organelle and nuclear DNA (Russell *et al*., 2003). Mitochondrial and chloroplast DNA has a much lower level of variability compared to the nuclear genome. Along with this, the reduction of genetic diversity in the process of domestication and further selection is demonstrated. So, based on the restriction site polymorphism in cpDNA of 34 wild and cultivated lines of sunflower, Rieseberg and Seiler (1990) showed the monomorphism in cultivated lines for cpDNA phenotypes. Also, based on polymorphism of microsatellite loci of chloroplast genome in six Helianthus species and 46 lines of cultivated sunflower, it was demonstrated the absent cpDNA polymorphisms

within cultivated forms of the sunflower *H. annuus* (Markin *et al*., 2015). Currently, almost all commercial sunflower hybrids are obtained on the basis of CMS of only one type of PET1, which was discovered by P. Leclercq (1966) in the interspecific hybrid between *H. petiolaris* Nutt and *H. annuus* L. However, intensive use of only one CMS source makes cultivated hybrids extremely vulnerable to new strains of the pathogens (Levings, 1990; Liberatore *et al*., 2016). In connection with the decrease in cytoplasmic genetic diversity, accompanying the processes of domestication and artificial selection, it is necessary to introduce new plasmotype into cultivated plants. The study of the structural and functional organization of the organelle genomes is also relevant from the point of view of modern taxonomy and phylogenetics. Until now, only a few chloroplast markers are used in studies of plant diversity and phylogeny (Soejima and Wen, 2006; Schroeder *et al*., 2011). With the advent of Next-Generation Sequencing (NGS) technologies, a real opportunity to use complete chloroplast genomes for the taxonomic and evolutionary studies was appeared (McPherson *et al*., 2013). The complete sequences of cpDNA of sunflower HA89-alloplasmic lines with the types of cytoplasmic male sterility PET1, PET2, ANN2 and MAX will allow us to identify an additional source of genetic variation and make a significant contribution to the solution of the fundamental problems of biology, especially in the field of genetics and evolutionary genomics of plastids.

The aim of this study was to investigate the polymorphism of the complete chloroplast genomes of sunflower fertile line HA89 and isonuclear CMS lines with four different cytoplasmic backgrounds (PET1, PET2, ANN2 and MAX).

Materials and Methods

Plant Materials

The study was carried out on sunflower fertile line HA89 and isonuclear CMS lines with four different cytoplasmic backgrounds. The plant materials were obtained from the genetic collection of the N. I. Vavilov Institute of Plant Genetic Resources (Russia). CMS sources belonged to the following species of the genus Helianthus: H. petiolaris (PET1, PET2), H. annuus (ANN2) and H. Maximilliani (MAX).

Mitochondrial DNA Extraction, Genome Library Construction and Sequencing

First of all, from leaves of 14 day sunflower seedlings was extracted organelle fraction with reduced amount of nuclear DNA as has been described earlier (Makarenko *et al*., 2016). For each line we used the same quantity of leaf tissue from 5 plants. The DNA isolation from such fraction was performed with PhytoSorb kit (Syntol, Russia), according to the manufacture's instruction. The NGS library preparation was made using 1 ng of DNA and Nextera XT DNA Library Prep Kit (Illumina, USA) (Head *et al*., 2014). All the preparation steps were done pursuant to manual. For the qualitative control of libraries Bioanalyzer 2100 (Agilent, USA) was used. The libraries quantitation was performed with Qubit fluorimeter (Invitrogen, USA) and qPCR (van Dijk *et al*., 2014). For sequencing libraries were diluted up to concentration of 8 pM. Libraries were sequenced on different sequencing platforms. Fertile line, PET1, MAX1 NGS libraries were sequenced with NextSeq 500 sequencer using High Output v2 kit (Illumina, USA). PET2 and ANN2 libraries were sequenced with HiSeq2000 platform using TruSeq SBS Kit v3-HS (Illumina, USA).

Analysis of Sequence Data

Quality of reads was determined by FastQC (http://www.bioinformatics.babraham.ac.uk). Trimming of adapter-derived and low quality (Q-score below 25) reads was performed with Trimmomatic software (Bolger *et al*., 2014). Using Bowtie2 tool v 2.3.3 (Langmead and Salzberg, 2012) sequencing reads were aligned to reference sequences of chloroplast (NC_007977.1) genomes from NCBI databank. Variant calling was made by samtools/bcftools software (Li, 2011) and manually revised using IGV tool (Thorvaldsdóttir *et al*., 2013).

Results and Discussion

About 13.2, 14.7, 9.3, 6.9 and 10.2 GB of raw DNA sequences isolated from the fertile line HA 89, CMS-PET1, CMS-PET2, CMS-ANN2, CMS-MAX were obtained, respectively. The overall alignment rates for chloroplast genomes were 25-30% of total read number, depending on sample. The average read coverage was more than 1000 for chloroplast genomes. Based on aligning sequencing reads to reference sequences (NC_007977.1) we obtained the complete sequence of the chloroplast genome of the studied sunflower lines. The chloroplast genomes of the HA89, CMS-PET1, CMS-PET2, CMS-ANN2, CMS-MAX consisted of 151,094 bp, 151,110 bp, 151127, 151147 and 151,138 bp, respectively. The difference in size is due to the increased length of individual non-coding regions of the chloroplast genome. In most terrestrial plants, the chloroplast genome consists of a circular DNA molecule, which includes a Large Single Copy region (LSC) and a small Single Copy Region (SSC) separated by two copies of Inverted Repeats (IR). The content, order and organization of chloroplast genes are usually highly conserved, which makes chloroplast genomes invaluable for genetic and phylogenetic studies. Chloroplast genomes of the researched sunflower lines are constructed on the conservative type and contain an LSC, an SSC and

a pair of IRs region. The lengths of the LSC regions ranged from 83,527 of fertile line to 83,605 bp of CMS-ANN2. While the lengths of the SSC varied in five cytotypes from 19,113 of CMS-MAX to 19,147 bp of fertile line. The IR regions are characterized by the smallest length change (Table 1). The content and order of the genes in the investigated cytotypes was identical to the previously sequenced sunflower cpDNA (NC_007977.1). Regions of high G+C content are more sensitive to mutation (Smith *et al*., 2002). The total G+C content had insignificant variations from 37.60 to 37.62% (Table 1), which is comparable with the chloroplast genomes of other Angiosperms (Jansen and Ruhlman, 2012).

Comparative chloroplast genome analyses of alloplasmic CMS lines of sunflower revealed 451 polymorphic sites. Among them, 58 sites were microsatellite loci, represented exclusively by mononucleotide repeats, are located in non-coding cpDNA regions (Table 2). Chloroplast microsatellites due to uniparental inheritance have been widely used in the analysis of genetic diversity, differentiation and population structure (Provan, 2000; Flannery *et al*., 2006). Also, 317 SNPs and 76 microindels were identified. The ratio of SSRs, SNPs and INDEL variations were 12.9, 70.2 and 16.9%, respectively. SNPs are represented by six types of nucleotide substitutions: A/G (26.9%), A/C (14.4%), A/T (10.5%), T/C (29.7%), T/G (12.5%), G/C (6.0), whereas SSRs were as follows: (A) 6-30 (46.6%), (T) 6-31 (46.6%), (C) 6- 11 (5.1%), (G) 7-9 (1.7%). Our results are well consistent with the data from other studies that chloroplast microsatellite loci are mainly represented by short polyT and polyA repeats, which in turn makes a significant contribution to the prevalence of AT nucleobases (62.38-62.40%). As is typical for other flowering plants (Ni *et al*., 2016; Shen *et al*., 2017), the most number of polymorphic sites is located in large single copy region; the lowest number of polymorphisms is located in inverted repeat region of chloroplast genomes. Of the 317 SNPs detected, 120 were located in the coding regions. Moreover, 59 substitutions were non-synonymous (Table 3). The highest number of non-synonymous SNPs was identified in CMS-MAX (36 substitutions), whereas the lowest value of substitutions was in CMS-PET1 (3 substitutions). 21 and 23 non-synonymous substitutions were identified in CMS-PET2 and CMS-ANN2 cpDNA, respectively. Chloroplast genome of the HA89 fertile line was largely similar to the reference sequence (NC_007977.1). Only 2 INDELs in coding regions and 1 SNP, 4 INDELs in intergenic regions were identified in the fertile line (Table 4). Among the alloplasmic male-sterile lines, cpDNA of CMS-MAX had the largest number of polymorphisms in comparison with the reference sequence. The lowest number of polymorphisms was revealed in CMS-PET1 (23 SNPs and 10 indel mutations throughout the genic and non-genic regions). Like as CMS-PET1, CMS-PET2 was obtained as a result of interspecific crossing between *H. petiolaris* and cultivated sunflower *H. annuus*. It is interesting to note, that the number of INDELs and SNPs in CMS-PET2 chloroplast genome were more than 4-fold and 6.5-fold higher than that in cpDNA of CMS-PET1. Also, the high frequency of SNPs and indel mutations was detected in the CMS-ANN2 cytotype (Table 4). Previously, analysis of restriction site polymorphisms in the chloroplast DNA of the accessions of wild and cultivated *H. annuus*, including the lines CMS 89 and HA89, demonstrated that all accessions of domesticated sunflower had an *H. annuus* cpDNA (Rieseberg and Seiler, 1990). The authors explained the absence of *H. petiolaris* cDNA in cultivated sunflower in that the original H. petiolaris source for CMS 89 was a hybrid or introgressive population of *H. annuus* and *H. petiolaris*. This was subsequently confirmed by the analysis of seven individuals from the source population of *H. petiolaris*, where all seven accessions were morphologically *H. petiolaris* but had the cpDNA of *H. annuus* (Rieseberg and Seiler, 1990). Also, the high frequency of SNPs and INDEL mutations was detected in the CMS-ANN2 cytotype (Table 4). The ANN2 cytotype was derived from a crossing between wild and cultivated sunflower *H. annuus* (Serieys, 1984; Skoric *et al*., 2012). Previously, the comparative analysis of the complete chloroplast genomes of cultivated and wild sunflower *H. annuus* revealed only 43 variant sites, including 22 SNPs and 21 polymorphic SSR loci (Makarenko *et al*., 2016).

The average frequency of SNPs and INDELs in the intergenic regions was 0.0062. Herewith, a twofold increase in frequency was identified in the intergenic region between the *psbI* and *petN* genes of the CMS-MAX line. In addition, the frequency was higher in such non-genic regions as the *psbM-rpoB*, *atpA-psbD*, *rps4-ndhJ*, *ndhc-atpE* and rps16 intron (Fig. 1).

Table 1: The features of chloroplast genomes of sunflower fertile line HA89 and 4 alloplasmic CMS lines

Type of cytoplasm	Size (bp)	LSC (bp)	SSC (bp)	IR (bp)	GC Content (%)
Fertile line	151,094	83,527	19,147	24,210×2	37.62
CMS- PET1	151,110	83,545	19,145	24,210×2	37.61
CMS-PET2	151,127	83,565	19,138	24,212×2	37.60
CMS-ANN2	151,147	83,605	19,118	24,212×2	37.61
CMS-MAX	151,138	83,601	19,113	24,212×2	37.61

Table 2: Polymorphic microsatellite loci in the chloroplast genomes of the sunflower alloplasmic CMS lines

Position in reference genome	Repeat MotifRef	RF	PET1	PET2	ANN2	MAX1	Localization	Region
206	(A)11			(A)9	(A)9	(A)9	*rpl2-psbA*	LSC
370	(A)7					(A)9	*rpl2-psbA*	LSC
1667	(A)6			(A)7	(A)8	(A)9	*psbA-matK*	LSC
1991	(T)9					(T)13	*psbA-matK*	LSC
2032	(T)12			(T)9	(T)14	(T)10	*psbA-matK*	LSC
3953	(T)7					(T)8	*matk-rps16*	LSC
4437	(A)7				(A)6		*matk-rps16*	LSC
4788	(A)8					(A)9	*matk-rps16*	LSC
5450	(C)11	(C)9	(C)8	(C)10	(C)9		*rps16 intron*	LSC
5692	(T)13		(T)11	(T)7	(T)7	(T)7	*rps16 intron*	LSC
5718	(C)7				(C)11	(C)8	*rps16 intron*	LSC
7720	(A)8					(A)9	*rps16-psbK*	LSC
7944	(T)8			(T)11	(T)9		*psbK-psbI*	LSC
8294	(T)8				(T)10	(T)9	*psbK-psbI*	LSC
8552	(A)8			(A)7	(A)7	(A)7	*psbI-petN*	LSC
8880	(T)6					(T)7	*psbI-petN*	LSC
9882	(A)8		(A)9	(A)9	(A)9	(A)10	*psbI-petN*	LSC
12759	(A)9					(A)10	*psbM-rpoB*	LSC
12984	(T)15		(T)11	(T)9	(T)10	(T)14	*psbM-rpoB*	LSC
16964	(A)8			(A)9	(A)9		*rpoC1 intron*	LSC
17424	(G)9		(G)8	(G)7	(G)7		*rpoC1 intron*	LSC
25296	(T)8			(T)9		(T)10	*atpI-atpH*	LSC
25466	(A)10		(A)13	(A)21-23	(A)19	(A)20	*atpI-atpH*	LSC
30166	(A)10		(A)12				*atpA-psbD*	LSC
35019	(A)9			(A)16-17	(A)16	(A)8	*psbC-psbZ*	LSC
35398	(A)18		(A)30	(A)24	(A)30	(A)18-22	*psbC-psbZ*	LSC
35885	(A)9		(A)8	(A)7	(A)7	(A)7	*psbZ-rps14*	LSC
44324	(A)8			(A)10	(A)14	(A)10	*ycf3-rps4*	LSC
46721	(A)15			(A)13	(A)22	(A)11	*rps4-ndhJ*	LSC
50049	(T)8				(T)9		*ndhc-atpE*	LSC
50163	(T)10		(T)12		(T)15		*ndhc-atpE*	LSC
50764	(T)11		(T)8	(T)7	(T)7	(T)7	*ndhc-atpE*	LSC
51778	(T)10		(T)9	(T)21	(T)22	(T)12	*ndhc-atpE*	LSC
54313	(A)16	(A)15	(A)15	(A)17	(A)21	(A)16-17	*atpB-rbcL*	LSC
58257	(A)7					(A)8	*accD-psaI*	LSC
60017	(C)6		(C)7	(C)7			*ycf4-cemA*	LSC
60433	(T)6			(T)7	(T)8	(T)9	*ycf4-cemA*	LSC
62683	(T)22		(T)20	(T)15	(T)16		*petA-psbJ*	LSC
64145	(A)9		(A)8	(A)14			*psbE-petL*	LSC
64939	(A)23				(A)11		*psbE-petL*	LSC
65714	(T)16				(T)21	(T)13	*petG-psaJ*	LSC
69743	(A)6					(A)7	*clpP intron*	LSC
69936	(T)10			(T)13	(T)11	(T)9	*clpP intron*	LSC
70545	(A)15		(A)20	(A)7	(A)7	(A)10	*clpP intron*	LSC
70584	(T)9		(T)10	(T)10			*clpP intron*	LSC
73405	(T)7			(T)8	(T)8	(T)9	*psbB-psbT*	LSC
77234	(A)10			(A)11		(A)9	*petD-rpoA*	LSC
79932	(T)22		(T)15	(T)19	(T)17		*rps8-rpl14*	LSC
80456	(T)24					(T)27	*rpl14-rpl16*	LSC
82083	(T)8					(T)9	*rpl16-rps3*	LSC
83644	(T)8			(T)9	(T)9		*rps19-rpl2*	IRA
101481	(T)9					(T)10	*rrn16-rrn23*	IRA
119344	(T)7			(T)8	(T)8		*psaC-ndhD*	SSC
120979	(A)16		(A)13	(A)20	(A)19	(A)10	*ndhD-ccsA*	SSC
122970	(T)13		(T)12	(T)18	(T)19	(T)13	*ccsA-rpl32*	SSC
128212	(T)31			(T)21			*ndhF-rrn5*	IRB
133146	(A)9					(A)10	*rrn23-rrn16*	IRB
150984	(A)8			(A)9	(A)9		*rpl2-rps19*	IRB

Table 3: Nonsynonymous substitution in the chloroplast genomes of the sunflower alloplasmic CMS lines

Position	Ref	PET1	PET2	ANN2	MAX1	Gene	Region
2984	A				G	*matk Phe195Ser*	LSC
14496	A		G			*rpoB Gln481Arg*	LSC
17643	A				C	*rpoC1 Met210Leu*	LSC
17814	C				T	*rpoC1 Arg267Cys*	LSC
18307	G			A		*rpoC1 Arg431Gln*	LSC
20660	T	C				*rpoC2 Leu490Pro*	LSC
20750	C			T		*rpoC2 Thr520Ile*	LSC
20945	T			G		*rpoC2 Leu585Arg*	LSC
22083	C				A	*rpoC2 Asp964Glu*	LSC
24141	A	C				*rps2 Gln178His*	LSC
28592	A				G	*atpA Ile61Val*	LSC
41098	G				T	*psaA Leu144Ile*	LSC
48363	A		G	G		*ndhJ Val107Ala*	LSC
52386	C		T	T		*atpB Gly468Ser*	LSC
55291	T				A	*rbcL Cys247Asn along with 55292*	LSC
55292	G				A	*rbcL Cys247Asn along with 55291*	LSC
55393	T				G	*rbcL Ser281Ala*	LSC
55396	C				A	*rbcL Gln282Lys*	LSC
55534	T				G	*rbcL Ser328Ala*	LSC
55675	C				A	*rbcL Leu375Ile along with 55677*	LSC
55677	G				A	*rbcL Leu375Ile along with 55675*	LSC
55837	C				A	*rbcL Gln429Lys*	LSC
55996	G				T	*rbcL Asp482Tyr*	LSC
57148	A				C	*accD Glu191Ala*	LSC
65328	C				T	*petL Pro2Leu*	LSC
78756	T			C		*rps11 Lys8Arg*	LSC
79664	G		T	T	T	*rps8 Pro72Thr*	LSC
82899	C				T	*rpl22 AlThr*	LSC
89152	A				G	*ycf2 Ile1110Val*	IRA
108400	C		G	G		*ycf1 Asn271Lys*	SSC
108722	A				C	*ycf1 Ile379Leu*	SSC
109194	C			T		*ycf1 Pro536Leu*	SSC
109640	G		A	A		*ycf Asp685Asn*	SSC
109850	C			G		*ycf1 Arg755Gly*	SSC
110520	C		G	G		*ycf1 Thr978Arg*	SSC
110568	C		A	A	A	*ycf1 Thr994Lys*	SSC
110913	T				C	*ycf1 Val1109Ala*	SSC
110918	C		A	A	A	*ycf1 His1111Asn*	SSC
111048	C				A	*ycf1 Ser1154Tyr*	SSC
111059	C			G		*ycf1 Pro1158Ala*	SSC
111605	T		G	G	G	*ycf1 Tyr1340Asp*	SSC
111659	C		T	T	T	*ycf1 Pro1358Ser*	SSC
111983	A		C	C	C	*ycf1 Ile1466Leu*	SSC
114382	T		C	C		*ndhH Val298Ala*	SSC
114585	C		T			*ndhH Pro366Ser*	SSC
114863	C		T			*ndhA Pro64Leu*	SSC
120314	T		C	C	C	*ndhD Leu296Pro*	SSC
121676	A				C	*ccsA Asn167Lys*	SSC
121813	G				A	*ccsA His122Tyr*	SSC
122087	C		A	A		*ccsA Leu30Phe*	SSC
125593	T	C	C	C	C	*ndhF Leu477Ser*	SSC
125888	A		T			*ndhF Gln573His*	SSC
125944	G				A	*ndhF Ser592Asn*	SSC
125949	A		T			*ndhF Asn594Tyr*	SSC
126087	G		T	T	T	*ndhF Val640Phe*	SSC
126194	C				A	*ndhF Phe675Leu*	SSC
145483	T				C	*ycf2 Ile1110Val*	IRB

Table 4: Number of SNPs and INDELs identified between the complete chloroplast genome sequences of the reference sequences (NC_007977.1) and HA89-alloplasmic CMS lines

	Type of sequence							
	Genic		Intronic		Intergenic		Total	
Type of cytoplasm	SNP	Indel	SNP	Indel	SNP	Indel	SNP	Indel
Fertile line	0	2	0	0	1	4	1	6
CMS-PET1	9	2	4	0	10	8	23	10
CMS-PET2	54	3	13	5	87	33	154	41
CMS-ANN2	52	3	10	4	80	32	142	39
CMS-MAX	71	3	11	2	114	38	196	43

Table 5: The relatively long INDELs (14-24 bp) in the chloroplast genomes of the sunflower alloplasmic CMS lines

Position in ref	Type	PET2	ANN2	MAX1	Localization
23377-78	INS	CTTTCTTTT ATCTTTTA			*rpoC2-rps2*
32066-67	INS		TACATAAAAATT AGAATAGAAAG		*atpA-psbD*
49986-50004	DEL			GTTTCCTTTT GCGGGGCGT	*ndhc-atpE*
61429-30	INS			TCCAGTAAATAA CAGAATCGTGGA	*cemA-petA*
73364-65	INS	AGGGTACCAGAGAA			*psbB-psbT*
78307-08	INS		TCTCTTGTTTCTCTTGC		*rpoA-rps11*
126449-50	INS	TTTAATTGAGTCTTTAATTCT			*ndhF-rrn5*

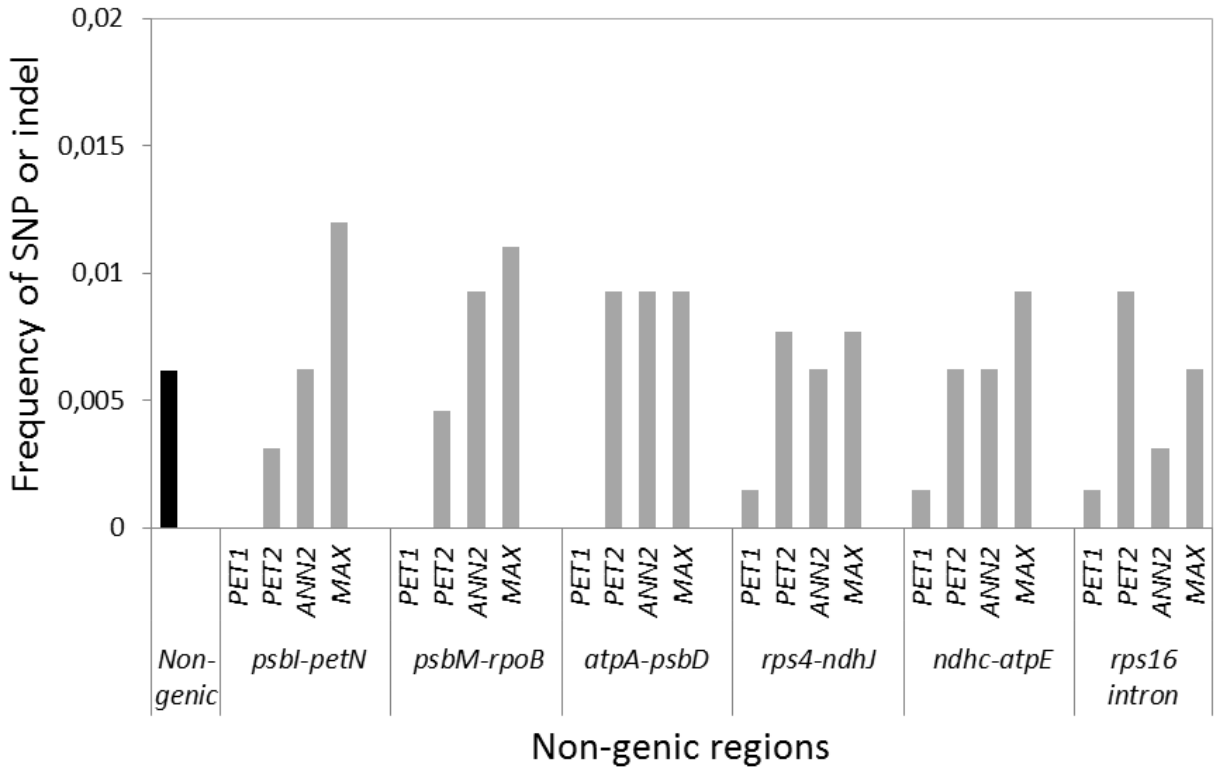

Fig. 1: Frequency of SNPs and INDELs in the highly variable non-genic regions of chloroplast genome sequences in the alloplasmic CMS lines relative to the fertile line. The grey bars indicate the frequencies of highly variable non-genic regions. The black bars indicate the average frequency of SNPs or INDELs in the non-genic regions, in total

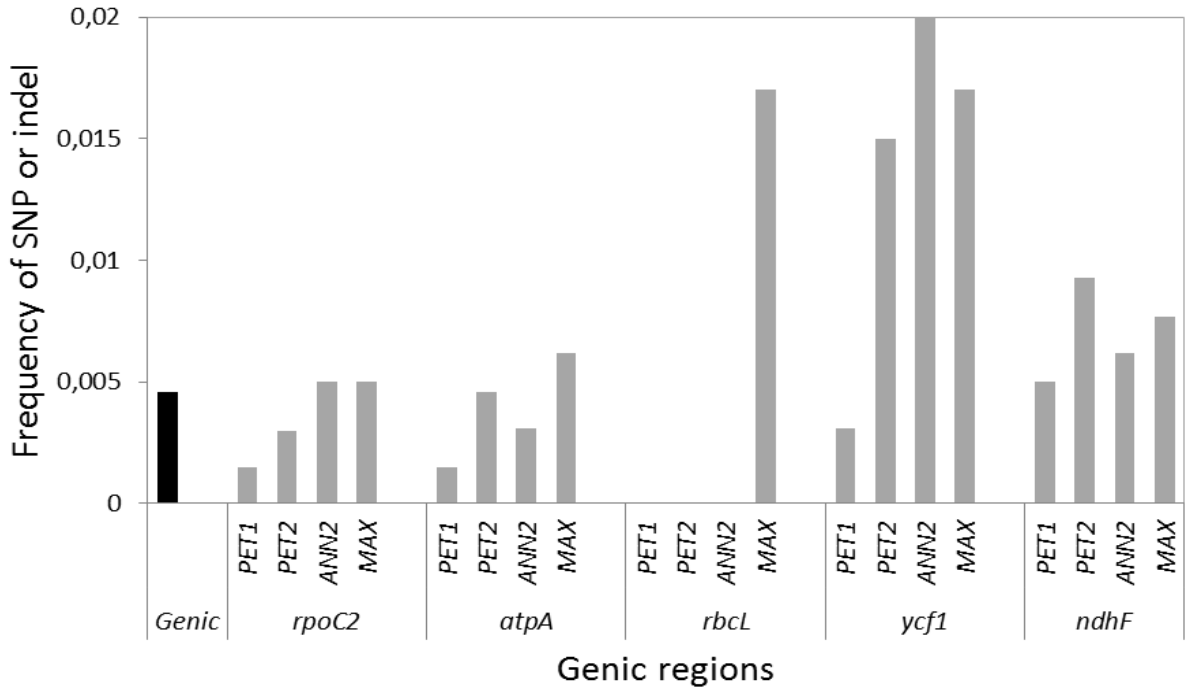

Fig. 2: Frequency of SNPs and INDELs in the highly variable genic regions of chloroplast genome sequences in the alloplasmic CMS lines relative to the fertile line. The grey bars indicate the frequencies of highly variable protein-coding genes. The black bars indicate the average frequency of SNPs or INDELs in the genic regions, in total

Meantime, the average frequency of SNPs and INDELs in the genic regions was 0.0046. Some chloroplast genes such as *rpoC2*, *atpA*, *rbcL*, *ndhF* and *ycf1* were distinguished by a higher level of polymorphism. Interestingly, the highest increase in frequency was characterized for the *ycf1* gene, with a 4.3-fold increases in frequency observed in the CMS-ANN2 line (Fig. 2). In a study devoted to a comparative analysis of the complete chloroplast genome sequences of the male-fertile line and two CMS lines of onions, it was reported that the average frequency of SNPs and INDELs in the non-genic regions was 0.0057, while the mean frequency in the genic regions was 0.0016 (Kim *et al*., 2015). Among intergenic regions, the most polymorphic was the regions between the *ndhF-rpl32*, *petN-psbM* and *trnS-trnG* genes. Meanwhile, as in our study, an increase in frequency of SNPs and INDELs in the genic regions was observed in the *ndhF* (3.5-fold) and *ycf1* (4.2-fold times) genes (Kim *et al*., 2015). It was shown that the *rpo* genes are highly variable and reliable phylogenetic markers, effective in the reconstruction of interrelations of species belonging to the same genus (Krawczyk and Sawicki 2013). A high level of polymorphism of *ndh*, *rpoC2,* rbcL and *ycf1* genes is also demonstrated in other studies (Wei *et al*., 2017; Joseph *et al*., 2013; Benkeblia, 2014).

In addition to short insertions and deletions (ranging from 1 to 5 bp) of bases in the chloroplast genome of CMS lines, the INDELs within the range of 14-24 bp in length were also found in the investigated cpDNA (Table 5). These insertions and deletions may be appropriate to use for PCR differentiation of the CMS lines due to differences in the amplicons length. Indeed, CMS-PET1 is widely used for the commercial production of F1 hybrid sunflower seeds. By now, this type of CMS has been fairly well studied. In particular, its molecular genetic bases are known and STS markers have been developed for the mitochondrial orf522 gene, which make it possible to distinguish between fertile line and sterile line which contained CMS-PET1 (Schnabel *et al*., 2008). The molecular basis of the CMS types like PET2, ANN2 and MAX1 is insufficiently studied, which is an obstacle to their introduction into commercial breeding. In our study, we detected INDELs specific for the chloroplast DNA of the CMS-PET2, CMS-ANN2 and CMS-MAX1 cytotypes (Table 5). The design a pair of primers for the conserved flanking regions of the INDELs will allow to development of PCR markers for identification of various types of sunflower CMS. These markers are a prerequisite for the development of highly productive heterotic sunflower hybrids on the basis of new CMS sources.

Conclusion

The comparative analysis of complete chloroplast genomes of fertile line HA89 and 4 alloplasmic CMS lines (PET1, PET2, ANN2, MAX1) revealed a total of 451 polymorphic sites, with including 58 SSRs, 317 SNPs and 76 microindels. Chloroplast microsatellite loci are mainly represented by short polyT and polyA repeats. Of the 317 SNPs detected, 120 were located in the coding regions, 59 of these substitutions are non-synonymous. Among the alloplasmic male-sterile lines, cpDNA of CMS-MAX was characterized by the largeast number of polymorphisms in comparison with the sequence of fertile line. The lowest number of polymorphic sites was revealed in CMS-PET1. Like as CMS-PET1, CMS-PET2 was obtained as a result of interspecific crossing between *H. petiolaris* and cultivated sunflower *H. annuus*. Nevertheless, the number of INDELs and SNPs in CMS-PET2 chloroplast genome was more than 4-fold and 6.5-fold higher than that in cpDNA of CMS-PET1. The average frequency of SNPs and INDELs in the non-genic regions and genic regions were 0.0062 and 0.0046, respectively. Increased mutation rates were found in the *psbM-rpoB, rps16 intron*, *atpA-psbD, rps4-ndhJ* and *ndhc-atpE* non-coding regions, as well as in the *rpoC2*, *atpA*, *rbcL*, *ndhF* and *ycf1* genes. In addition to short insertions and deletions ranging from 1 to 5 bp, the relatively long INDELs (14-24 bp) unique for each CMS line were found. These insertions and deletions may be of use for PCR differentiation of the CMS lines due to differences in the amplicons length.

Acknowledgement

This research was supported by a project of the Ministry Education and Science of Russia, no. 6.929.2017/4.6.

Funding Information

The funders of this manuscript were the Ministry Education and Science of Russia, no. 6.929.2017/4.6.

Author's Contributions

Kirill Azarin: Wrote the first draft of the manuscript, participated in the data analysis. Author read and approved the final manuscript.

Maxim Makarenko: Participated in the laboratory study, data analysis. Author read and approved the final manuscript.

Alexander Usatov: Designed experiments. Author read and approved the final manuscript.

Vladimir Khachumov: Participated in the laboratory study, data analysis. Author read and approved the final manuscript.

Alexey Kovalevich: Participated in the laboratory study, data analysis. Author read and approved the final manuscript.

Oleg Gorbachenko: Designed experiments. Author read and approved the final manuscript.

Vera Gavrilova: Collected and analyzed data. Author read and approved the final manuscript.

Ethics

The authors state that this article conforms to the ethical standards specified by the American Journal of Biochemistry and Biotechnology.

References

Benkeblia, N., 2014. Omics Technologies and Crop Improvement. 1st Edn., CRC Press, Taylor and Francis Group, ISBN-10: 1466586699, pp: 392.

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

Flannery, M.L., F.J. Mitchell, S. Coyne, T.A. Kavanagh and J.I. Burke *et al*., 2006. Plastid genome characterisation in Brassica and Brassicaceae using a new set of nine SSRs. Theor. Applied Genet., 113: 1221-1231. DOI: 10.1007/s00122-006-0377-0

Head, S.R., H.K. Komori, S.A. LaMere, T. Whisenant and F. Van Nieuwerburgh *et al*., 2014. Library construction for next-generation sequencing: Overviews and challenges. Biotechniques, 56: 61-64. DOI: 10.2144/000114133

Jansen, R.K. and T.A. Ruhlman, 2012. Plastid Genomes of Seed Plants. In: Genomics of Chloroplasts and Mitochondria, Bock, R. and V. Knoop, (Eds.), Springer Science and Business Media, Dordrecht, ISBN-10: 9400729200.

Joseph, B., J.A. Corwin, B. Li, S. Atwell and D.J. Kliebenstein, 2013. Cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation in the metabolome. Life, 2: e00776- e00776. DOI: 10.7554/eLife.00776

Kim, S., J.Y. Park and T. Yang, 2015. Comparative analysis of the complete chloroplast genome sequences of a normal male-fertile cytoplasm and two different cytoplasms conferring cytoplasmic male sterility in onion (*Allium cepa* L.). J Hortic Sci. Biotechnol., 90: 459-473. DOI: 10.1080/14620316.2015.11513210

Krawczyk, K. and J. Sawicki, 2013. The uneven rate of the molecular evolution of gene sequences of DNA-dependent RNA polymerase i of the genus *Lamium* L. Int. J. Mol. Sci., 14: 11376-11391. DOI: 10.3390/ijms140611376

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

Leclercq, P., 1969. Une sterilite male chez le tournesol. Ann. Amelior Plant., 19: 99-106.

Levings, C.S., 1990. The texas cytoplasm of maize: Cytoplasmic male sterility and disease susceptibility. Science, 250: 942-947. PMID: 17746917

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

Liberatore, K.L., S. Dukowic-Schulze, M.E, Miller, C. Chen and S.F. Kianian, 2016. The role of mitochondria in plant development and stress tolerance. Free Radic. Biol. Med., 100: 238-56. DOI: 10.1016/j.freeradbiomed.2016.03.033

Makarenko, M.S., A.V. Usatov, N.V. Markin, K.V. Azarin and O.F. Gorbachenko *et al*., 2016. Comparative genomics of domesticated and wild sunflower: Complete chloroplast and mitochondrial genomes. OnLine J. Biol. Sci., 16: 71-75. DOI: 10.3844/ojbsci.2016.71.75

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

Mashkina, E.V., A.V. Usatov and M.V. Skorina, 2010. Comparative analysis of thermotolerance of sunflower chlorophyll mutants. Russ. J. Genet., 46: 178-184. DOI: 10.1134/S1022795410020079

McPherson, H., M. van der Merwe, S.K. Delaney, M.A. Edwards and R.J. Henry *et al*., 2013. Capturing chloroplast variation for molecular ecology studies: A simple next generation sequencing approach applied to a rainforest tree. BMC Ecol. DOI: 10.1186/1472-6785-13-8

Ni, L.H., Z.L. Zhao, H.X. Xu, S.L. Chen and G. Dorje, 2016. The complete chloroplast genome of Gentiana straminea (Gentianaceae), an endemic species to the Sino-Himalayan subregion. Gene, 577: 281-288. DOI: 10.1016/j.gene.2015.12.005

Pervaiz, T., X. Sun, Y. Zhang, R. Tao and J. Zhang *et al*., 2015. Association between chloroplast and mitochondrial DNA sequences in Chinese *Prunus* genotypes (*Prunus persica, Prunus domestica and Prunus avium*). BMC Plant Biol., 15: 4-4. DOI: 10.1186/s12870-014-0402-4

Provan, J., 2000. Novel chloroplast microsatellites reveal cytoplasmic variation in Arabidopsis thaliana. Mol. Ecol., 9: 2183-2185. DOI: 10.1046/j.1365-294X.2000.105316.x

Rieseberg, L.H. and G.J. Seiler, 1990. Molecular evidence and the origin and development of the domesticated sunflower (Helianthus annus, Asteraceae). Econ. Bot., 44: 79-91. DOI: 10.1007/BF02860477

Russell, J.R., A. Booth, J.D. Fuller, M. Baum and S. Ceccarelli *et al*., 2003. Patterns of polymorphism detected in the chloroplast and nuclear genomes of barley landraces sampled from Syria and Jordan. Theor Applied Genet., 107: 413-21. DOI: 10.1007/s00122-003-1261-9

Schroeder, H., A. Höltken and M. Fladung, 2011. Chloroplast SNP-marker as powerful tool for differentiation of Populus species in reliable poplar breeding and barcoding approaches. BMC Proc., 5: P56- P56. DOI: 10.1186/1753-6561-5-S7-P56

Serieys, H., 1984. Wild *Helianthus* Species, a Potential Source of Androsterilities. 2nd Edn., Eucarpia Meeting on the Sunflower, Leningrad.

Shen, X., M. Wu, B. Liao, Z. Liu and R. Bai *et al*., 2017. Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant artemisia annua. Molecules. DOI: 10.3390/molecules22081330

Schnabel, U., U. Engelmann and R. Horn, 2008. Development of markers for the use of the PEF1 cytoplasm in sunflower hybrid breeding. Plant Breed, 127: 587-591. DOI: 10.1111/j.1439-0523.2008.01516.x

Skoric, D., G. Seiler, Z. Liu, C.C. Jan and J. Miller *et al*., 2012. Sunflower genetics and breeding. Int. monografy. Novi Sad.

Smith , N. G. C. , M. T. Webster , and H. Ellegren, 2002. Deterministic mutation rate variation in the human genome. Genome Research, 12: 1350 – 1356. PMCID: PMC186654

Soejima, A. and J. Wen, 2006. Phylogenetic analysis of the grape family (Vitaceae) based on three chloroplast markers. Am. J. Bot., 93: 278-87. DOI: 10.3732/ajb.93.2.278

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

van Dijk, E.L., Y. Jaszczyszyn and C. Thermes, 2014. Library preparation methods for next-generation sequencing: Tone down the bias. Exp. Cell Res., 322: 12-20. DOI: 10.1016/j.yexcr.2014.01.008

Wei, R., Y.H. Yan, A. Harris, J.S. Kang and H. Shen *et al*., 2017. Plastid phylogenomics resolve deep relationships among eupolypod II ferns with rapid radiation and rate heterogeneity. Genome Biolo. Evolut., 9: 1646-1657. DOI: 10.1093/gbe/evx107