Mining of simple sequence repeats in chloroplast genome sequence of *Trifolium subterraneum*

Deepika Srivastava and Asheesh Shanker

Department of Bioscience and Biotechnology, Banasthali University, Banasthali-304022, Rajasthan, India
ashomics@gmail.com

Abstract: Simple sequence repeats (SSRs), also known as microsatellites, are found in DNA sequences and consist of short repeating motifs of 1-6 nucleotides. These repeats are ubiquitous and play an important role in the development of molecular markers. Therefore, the present analysis was conducted to identify SSRs in chloroplast genome of *Trifolium subterraneum*. A total of 77 SSRs (including 3 compound SSRs) were identified with an average length of 12.79 bp in 144.76 kb sequence mined. Depending upon the repeat unit, SSRs varied in length from 12 to 27 bp. The identified SSRs showed a density of 1 SSR/1.88 kb. Mononucleotides (38, 49.35%) were found to be the most abundant repeat, followed by dinucleotide (15, 19.48%), trinucleotide (13, 16.88%) and tetranucleotide (11, 14.29%). The penta and hexanuleotide repeats were not detected in chloroplast genome of *Trifolium subterraneum*.

[Srivastava D and Shanker A. ***In silico* mining of simple sequence repeats in chloroplast genome of *Trifolium subterraneum*.** *Researcher* 2015;7(4):63-66]. (ISSN: 1553-9865). http://www.sciencepub.net/researcher. 10

Key words: Chloroplast; Data Mining; Microsatellites; Simple sequence repeats

1. Introduction

Trifolium subterraneum is a species of clover which is grown commercially for animal fodder. Microsatellites or simple sequence repeats (SSRs) are short repeat motifs (1-6 bp) present in DNA sequences. These repeats are ubiquitous and found in both coding/non-coding regions of genome (Shanker et al., 2007). SSRs have been considered as molecular markers of choice in many plant genomes (Cardle et al., 2000).

Chloroplasts are cytoplasmic organelles present in green plants and contain their own autonomously replicating genome which encodes a number of components for the process of photosynthesis. The adequate number of available complete chloroplast genome sequences makes it feasible to use them for various purposes. SSR mining is one of them. In the recent past SSR specific databases have been developed including MitoSatPlant (Kumar et al., 2014) and ChloroSSRdb (Kapil et al., 2014).

Despite all these efforts a detailed analysis of SSRs in chloroplast genome of *Trifolium subterraneum* is not available. Therefore, in the present study chloroplast genome sequence of *Trifolium subterraneum* was mined for the identification of chloroplast simple sequence repeats (cpSSRs).

2. Materials and Methods

2.1. Chloroplast genome sequence retrieval

The complete organellar genome sequences of angiosperms are available at National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov). The chloroplast genome sequence of *Trifolium subterraneum* (Accession number: NC_011828) was downloaded from NCBI in FASTA and GenBank format.

2.2. Simple sequence repeats mining

MISA, (http://pgrc.ipk-gatersleben.de/misa), was used for the detection of SSRs. The minimum repeat size was considered as >= 12-mono, >= 6-di, >= 4-tri, >= 3-tetra, penta and hexa nucleotide, respectively. The maximum difference taken between two SSRs was kept 0.

2.3. Analysis of mined cpSSRs

Data generated after SSR mining was analyzed for frequency & distribution of SSRs in coding and non-coding regions of cpDNA. The information about coding, non-coding and coding-non-coding regions was taken from GenBank files. SSRs were classified as coding and non-coding on the basis of their presence in coding and non-coding regions (Kapil et al., 2014).

3. Results and Discussion

The present analysis deals with the identification of chloroplast simple sequence repeats (cpSSRs) in *Trifolium subterraneum*.

A total of 77 SSRs (including 3 compound SSRs) were identified with an average length of 12.79 bp in 144.76 kb sequence mined. Mononucleotides (38, 49.35%) were found to be the most abundant repeat, followed by dinucleotide (15, 19.48%), trinucleotide (13, 16.88%) and tetranucleotide (11, 14.29%). Pentanucleotide and hexanucleotide repeats were not detected in chloroplast genome of *Trifolium subterraneum*. The distribution of mined cpSSRs is presented in figure 1. Among mononucleotide repeats presence of only A/T motifs showed consistency with SSR analysis of other organelle genomes

(Rajendrakumar et al., 2008; Melotto-Passarin et al., 2011).

The chloroplast genome of *Trifolium subterraneum* contains 1 SSR/1.88 kb sequence mined. The density of cpSSRs in this study found to be higher than the density of EST-SSRs in barley, maize, wheat, rye, sorghum and rice (1 SSR/6.0 kb; Varshney et al., 2002), cotton and poplar (1 SSR/20 kb and 1 SSR/14 kb respectively; Cardle et al., 2000), Unigenes sequences of *Citrus* (1 SSR/12.9 kb; Shanker et al., 2007a) and cpSSRs of *Nothoceros aenigmaticus* (1SSR/3.65kb; Shanker, 2015). Moreover, the density of SSRs in *Trifolium* was higher when compared to the cpSSRs of rice (1SSR/6.5 kb; Rajendrakumar et al., 2007) however, lower than the cpSSRs density in family Solanaceae (1 SSR/1.26kb; Tambarussi et al., 2009). The variation in SSR density might be due to different parameters including minimum length of SSRs taken, the amount of data analyzed and genomic composition of the sequence mined.

The identified SSRs motif, their length, start-end position and the region in which they lie is presented in table 1.

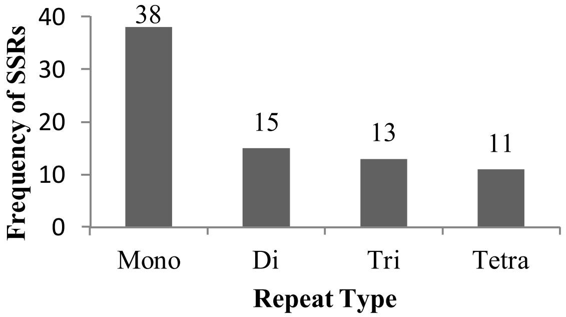

Figure 1. Frequency distribution of various repeat types.

Table 1. Identified SSRs motif, their length, start-end position in chloroplast genome of *Trifolium subterraneum*.

S. No.	Motif	Length	Start	End	Region
1	(ATT)4	12	1821	1832	Non-coding
2	(AATA)3	12	2333	2344	Non-coding
3	(T)12	12	4118	4129	Non-coding
4	(TA)6	12	4623	4634	Non-coding
5	(T)13	13	5775	5787	Non-coding
6	(A)12	12	6105	6116	Non-coding
7	(AT)6	12	12863	12874	Non-coding
8	(T)12	12	14424	14435	Non-coding
9	(ATT)4	12	15020	15031	Non-coding
10	(TA)6	12	15387	15398	Non-coding
11	(A)13	13	15877	15889	Non-coding
12	(A)13	13	20835	20847	Non-coding
13	(A)12	12	20939	20950	Non-coding
14	(AT)8	16	23147	23162	Non-coding
15	(T)12	12	23218	23229	Non-coding
16	(A)12	12	23231	23242	Non-coding
17	(TAT)4	12	24968	24979	Non-coding
18	(ATA)5	15	25645	25659	Non-coding
19	(TTTA)3	12	28344	28355	Non-coding
20	(T)12	12	28653	28664	Non-coding
21	(A)13	13	28953	28965	Non-coding
22	(T)15	15	29007	29021	Non-coding
23	(T)12	12	32366	32377	Coding
24	(T)13	13	34267	34279	Coding
25	(A)13	13	37758	37770	Non-coding
26	(ATT)4	12	37852	37863	Non-coding
27	(TA)9	18	39090	39107	Non-coding
28	(T)12	12	39689	39700	Non-coding
29	(T)12	12	39801	39812	Non-coding
30	(ATTT)3	12	41954	41965	Non-coding
31	(TA)10	20	42442	42461	Non-coding
32	(TA)6	12	43069	43080	Non-coding

33	(GATA)3	12	43240	43251	Non-coding
34	(ATA)4(AT)6*	21	45251	45271	Compound/non-coding
35	(AT)6	12	45292	45303	Non-coding
36	(AT)6	12	45536	45547	Non-coding
37	(TA)7	14	45746	45759	Non-coding
38	(T)16	16	46532	46547	Non-coding
39	(T)12	12	49908	49919	Non-coding
40	(AT)6	12	50358	50369	Non-coding
41	(ACTA)3	12	50533	50544	Non-coding
42	(T)12	12	63584	63595	Non-coding
43	(TAT)4(ATA)5*	27	63680	63706	Compound/non-coding
44	(A)13	13	66428	66440	Non-coding
45	(TTTA)3	12	84353	84364	Non-coding
46	(TA)8	16	84378	84393	Non-coding
47	(T)12	12	86236	86247	Non-coding
48	(TCT)4	12	89015	89026	Non-coding
49	(TAT)4	12	89992	90003	Non-coding
50	(TAT)4	12	90468	90479	Non-coding
51	(TAT)4	12	90691	90702	Non-coding
52	(CTAC)3	12	93276	93287	Coding
53	(AT)6	12	104236	104247	Non-coding
54	(A)13	13	104265	104277	Non-coding
55	(A)13	13	104290	104302	Non-coding
56	(A)12	12	109632	109643	Coding-non-coding
57	(ATAG)3	12	109656	109667	Non-coding
58	(TATT)3	12	113059	113070	Coding
59	(AT)6	12	114488	114499	Coding-non-coding
60	(TATT)3	12	114773	114784	Non-coding
61	(T)12	12	117409	117420	Non-coding
62	(A)12	12	117478	117489	Non-coding
63	(A)13	13	117786	117798	Non-coding
64	(T)12	12	120011	120022	Non-coding
65	(CTTT)3	12	120039	120050	Non-coding
66	(A)13	13	121315	121327	Non-coding
67	(A)13	13	121849	121861	Non-coding
68	(T)12	12	124227	124238	Non-coding
69	(T)13	13	125845	125857	Non-coding
70	(T)12	12	130832	130843	Non-coding
71	(T)18	18	132793	132810	Non-coding
72	(A)15	15	132812	132826	Non-coding
73	(TAT)5(T)13*	27	142164	142190	Compound/non-coding
74	(T)14	14	142393	142406	Non-coding

It is evident from this table that the majority of SSRs were found in non-coding region of the chloroplast genome. This non random distribution of cpSSRs towards non-coding regions showed consistency with earlier studies Solanaceae (Daniell et al., 2006), Asteraceae (Timme et al., 2007), Fabaceae (Saski et al., 2008) and *Saccharum* (Melotto-Passarin et al., 2011).

4. Conclusion

The identified SSRs will be useful for the development of SSR markers, which help in genetic diversity studies and reveals variation in genomes. Moreover, the study provides scientific base for phylogenetics and evolutionary genetics studies on different *Trifolium* species in future.

Acknowledgment

We are thankful to Prof. Aditya Shastri, Vice Chancellor, Banasthali Vidyapith and Prof. Vinay Sharma, HOD, Deptt. of Bioscience and Biotechnology, Banasthali Vidyapith for providing necessary facilities. Financial support from DBT, Center for Bioinformatics to DS and UGC-MRP (F.No. 42-138/2013) to AS is also acknowledged.

Correspondence to
Asheesh Shanker,
Deptt. of Bioscience and Biotechnology,
Banasthali University, Banasthali-304022,
Rajasthan (INDIA).
Email: ashomics@gmail.com

References

1. Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D and Waugh R. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 2000; 156:847-854.
2. Daniell H, Lee SB, Grevich J, Saski C, Quesada-Vargas T, Guda C, Tomkins J and Jansen RK. Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. Theoretical and Applied Genetics 2006; 112: 1503-1518.
3. Kapil A, Rai PK and Shanker A. ChloroSSRdb: a repository of perfect and imperfect chloroplastic simple sequence repeats (cpSSRs) of green plants. Database: The Journal of Biological Databases and Curation 2014; doi:10.1093/database/bau107.
4. Kumar M, Kapil A and Shanker A. MitoSatPlant: Mitochondrial microsatellites database of viridiplantae. Mitochondrion 2014; 19: 334-337.
5. Melotto-Passarin DM, Tambarussi EV, Dressano K, de Martin VF and Carrer H. Characterization of chloroplast DNA microsatellites from *Saccharum* spp. and related species. Genetics and Molecular Research 2011; 10: 2024-2033.
6. Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K and Sundaram RM. Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. Bioinformatics 2007; 23: 1-4.
7. Saski C, Lee SB, Daniell H, Wood TC, Tomkins J, Kim HG, Jansen RK. Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. Plant Molecular Biology 2005; 59: 309-322.
8. Shanker A, Singh A and Sharma V. *In silico* mining in expressed sequences of *Neurospora crassa* for identification and abundance of microsatellites. Microbiological Research 2007; 162: 250- 256.
9. Shanker A, Bhargava A, Bajpai R, Singh S, Srivastava S and Sharma V. Bioinformatically mined simple sequence repeats in UniGene of *Citrus sinensis*. Scientia Horticulturae 2007a; 113:353-361.
10. Shanker A. *In silico* mining of simple sequence repeats in chloroplast genome of *Nothoceros aenigmaticus*. Researcher 2015; 7:24-27.
11. Tambarussi EV, Melotto-Passarin DM, Gonzalez SG, Brigati JB, de Jesus FA, Barbosa AL, Dressano K and Carrer H. *In silico* analysis of simple sequence repeats from chloroplast genomes of Solanaceae species. Crop Breeding and Applied Biotechnology 2009; 9:344-352.
12. Timme R, Kuehl EJ, Boore JL and Jansen RK. A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. American Journal of Botany 2007; 94: 302-312.
13. Varshney RK, Thiel T, Stein N, Langridge P and Graner A. *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cellular & Molecular Biology Letters 2002; 7: 537-546.

4/10/2015