RESEARCH ARTICLE **(Open Access)**

Mining and Survey of Simple Sequence Repeats in Expressed Sequence Tags of Tomato Species

SAIDA SHARIFOVA[1*], SABINA MEHDIYEVA[2], ILHAM SHAHMURADOV[3]

[1]Department of Biotechnology, Genetic Resources Institute / Azadlig 155, AZE1106, Baku, Azerbaijan
[2]Department of Cytogenetics, Genetic Resources Institute / Azadlig 155, AZE1106, Baku, Azerbaijan
[3]Department of Bioinformatics, Institute of Botany / AZ1073, Baku, Azerbaijan

Abstract

Expressed Sequence Tags (ESTs) of three tomato species were computationally mined for simple sequence repeats (SSRs). A total of 4,490, 291 and 1,270 simple sequence repeats identified in analyzed non-redundant ESTs of *Solanum lycoperisicum*, *Solanum pennellii* and *Solanum habrochaites*, respectively. In *S. lycopersucum*, 416 sequences contained more than one SSR and 264 motifs were present in compound formation. 24 and 137 EST sequences contained more than one SSR, and 16 and 93 motifs were found in compound formation in *S. pennellii* and S. *habrochaites*, respectively. The frequency of repeats within all retrieved *S. lycopersicum* EST sequences were 7.6%, whereas this number was corresponded to 6.5% in *S. pennellii* and 9% in *S. habrochaites*. An average density was one SSR per 9 kb in *S. lycopersicum*, per 7.9 kb in *S. pennellii* and per 9.4 kb in *S. habrochaites*. AT/AT, AG/CT and AAG/CTT motifs, considering sequence complementary, detected more frequently among all types of identified repeats.

Keywords: SSRs, tomato, wild, repeats, unigene, ESTs.

1. Introduction

The Lycopersicon section of Solanum contains domesticated tomato (*Solanum lycopersicum* L.) and its 12 closest wild relatives, that all native to western South America [21]. A cultivated tomato (*S. lycopersicum* L.) is an important vegetable crop with a worldwide production of around 162 million tons in 2012 (FAOSTAT, 2012). However, cultivated tomato is genetically poor than those of wild species, due to population bottlenecks [1, 23]. Hence, wild tomato species harboring many valuable genes are frequently used in cultivated tomato breeding programs [22].

Solanum pennellii Corell. and *Solanum habrochaites* S. Knapp & D.M. Spooner have both self-incompatible and self-compatible populations and they are considered valuable species for improvement of cultivated tomato germplasms. *S. pennellii* is an important donor for its extreme stress tolerance [3, 8]. *S. habrochaites* possess many important traits for disease resistance, cold tolerance and etc. [8].

All of the 13 species of Lycopersicon section have been proposed for sequencing by the SOL-100 project (http://solgenomics.net/organism/sol100/view). *S. lycopersicum* cv. Heinz 1706 and *S. pimpinellifolium* LA1589 already have been sequenced and assembled by the International Tomato Genome Sequencing Consortium [27]. Furthermore, whole genome re-sequencing and transcriptome sequencing of cultivated and several wild tomato species carried out by different researchers [3]. Such projects generate large amounts of sequence information that are stored in different databases and extensively used for mining useful genes and molecular markers [12].

This work represents an attempt on computationally mining and examining of abundance and types of SSRs in ESTs of cultivated and two wild species of tomato, retrieved from the GenBank at the National Center for Biotechnology Information (NCBI).

2. Materials and Methods

EST sequences of *S. lycopersicum*, *S. pennellii* and *S. habrochaites* species of tomato deposited at the NCBI (http://www.ncbi.nlm.nih.gov/) were retrieved and assembled with the CAP3 assembler for identification of non-redundancy using the default

*Corresponding author: Saida Sharifova; E-mail: saidasharifzade@yahoo.com

(Accepted for publication June 20. 2016)

value [11]. The non-redundant unigene sequences were screened for the presence of EST-SSR motifs using the MIcro SAtellite identification tool (MISA) (http://pgrc.ipk-gatersleben.de/misa) [26]. SSR detection criteria was fixed at 7, 5, 4, 3, and 3 repeat units for di-, tri-, tetra-, penta-, and hexanucleotide motifs respectively. Mononucleotide repeats were not included in the SSR search criteria. For compound repeats the maximum default interruption length was set at 100 bp.

3. Results and Discussions

300,422 EST sequences of *S. lycopersicum*, 10,946 of *S. pennellii* and 26,019 of *S. habrochaites* were used to computational analysis of frequency and types of EST-SSRs. The total size of examined sequences was 157,531,085 bp, 4,910,677 bp and 18,854,172 bp, respectively. Since, random sequencing within cDNA libraries usually results in a high proportion of redundant ESTs, we performed SSR search after elimination of redundancy. All of the downloaded EST sequences were assembled with the CAP3 sequence assembly program resulting in production of 59,466 unigens including 21,354 contigs and 38,112 singletons in *S. lycopersicum*, 4,486 unigens consisted of 1,163 contigs and 3,323 singletons in *S. pennellii*, and 14,104 unigens comprising 3,194 contigs and 10,910 singletons in *S. habrochaites*, respectively (Table 1). The total size of unigens was 40,559,149 bp, 2,309,828 bp and 11,926,425 bp after sequence assembly.

Table 1. Results of microsatellite search in three tomato species

Parameters	Plant species		
	Solanum lycopersicum	*Solanum pennellii*	*Solanum habrochaites*
Total number of sequences examined:	300422	10946	26019
Total number of non-redundant sequences:	59466	4486	14104
Total number of identified SSRs:	4490	291	1270
Total number of SSR containing sequences:	3989	264	1037
number of sequences containing more than 1 SSR:	416	24	137
number of SSRs present in compound formation:	264	16	93
dinucleotide motifs:	843	37	222
trinucleotide motifs:	2071	143	399
tetranucleotide motifs:	216	11	55
pentanucleotide motifs:	611	24	440
hexanucleotide motifs:	749	76	154

A total of 4,490 EST-derived simple sequence repeats within 3,989 sequences were identified after mining of 59,466 non-redundant unigens of cultivated tomato (*S. lycoperisicum* L.). 416 sequences contained more than one SSR and 264 motifs were found in the compound formation (Table 1). A total of 291 SSRs within 264 unigens were identified in 4,486 non-redundant sequences of *S. pennellii*, while 1,270 SSR repeats were found in 1,037 different unigens among of 14,104 sequences of *S. habrochaites*. 24 and 137 sequences were carrying more than one SSR and 16 and 93 motifs were found in the compound formation in *S. pennellii* and S. *habrochaites*, respectively (Table 1).

Development of molecular markers from the transcribed part of genome becomes simple and cheap approach recent years, which allow to track coding regions of genome [28]. Especially, *in silico* mining of sequence data accumulated in public databases make this approach much more handy and powerful [9]. In this research, we tried to mine and examine abundance and types of SSRs in ESTs of cultivated and two wild species of tomato and compare their presence in transcribed parts of genome.

The frequency of simple sequence repeats in ESTs reflects the density of SSRs in the transcribed regions of the genome. The frequency of mentioned repeats within all retrieved *S. lycopersicum* EST sequences in our research was 7.6%, while this number was corresponded to 6.5% in *S. pennellii* and 9% in *S. habrochaites*. Only a small fractions of screened non-redundant ESTs (5.2% in *S. lycopersicum*, 9.5% in *S. pennellii* and 12% in *S. habrochaites*) from the respective species contained SSR repeats, giving an average density of one SSR per 9 kb in *S. lycopersicum*, per 7.9 kb in *S. pennellii* and per 9.4 kb in *S. habrochaites* species. In other reports, an EST-SSRs have been observed to be correspond to one every 11.1 kb in tomato, 3.8 kb in pepper, 14.7 kb in lettuce, 13.8 kb in *Arabidopsis thaliana*, 3.4 kb in rice, 8.1 kb in maize, 7.4 kb in soybean, 20.0 kb in cotton and 14.0 kb in poplar [4, 13, 30, 31].

The observed variations in the frequency and density among different studies were considered mainly due to the criteria used to identify SSRs, size of data set, database mining tools and EST sequence redundancy removal criteria [28]. Several authors show that, differently sized genomes may also contribute to affecting repetitiveness of microsatellites [2, 10, 32].

We classified all repeat types, taking sequence complementarity into account, and examined their occurrence (Figure 1). Trinucleotide repeats in our survey were the most abundant class of microsatellites in *S. lycopersicum* and *S. pennellii* species which constituted 46.1% and 49.1% of all detected SSRs respectively. The most abundant trinucleotide motifs were followed by dinucleotide in cultivated tomato (18.8%) while the second most frequent repeat types in *S. pennellii* were hexamers (26.1%). The most common type of repeat in *S. habrochaites* was the pentamers, which constituted 34.6% of all SSRs detected, followed by trimers (31.4%), dimers (17.5%), hexamers (12.1%) and tetramers (4.3%) (Figure 1).

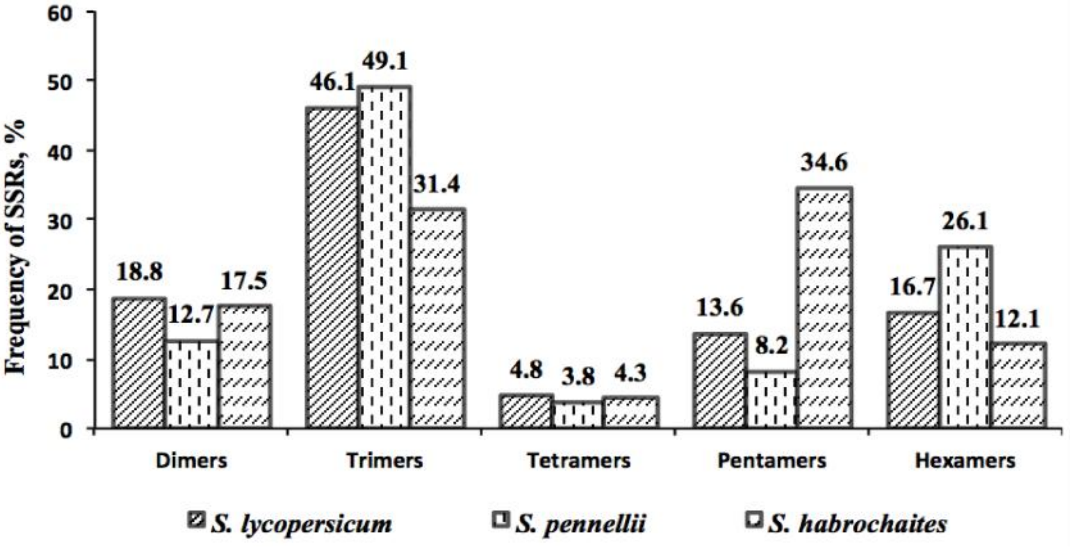

Figure 1. Frequency of SSRs in ESTs of three tomato species

Varshney and et al. reported that, trinucleotide repeats in plants were the most common, followed by either dimers or tetramers [29]. Other researches have shown that triplet motifs predominantly associated with the coding regions [18]. However, the reference genome of tomato (Heinz 1706) were mined for microsatellite repeats by Suresh and et al. and much more abundance of dinucleotide repeats (60.18%) than that of trimers (19.56%) and other repeats found [25]. In the present study, we found that the dominant type of EST-SSR repeat units was trimers in *S. lycopersicum* and *S. pennellii*, whereas the most common type was pentamers in *S. habrochaites*. The second common motifs were dimers, hexamers and trimers, respectively.

It is considered that SSRs within genes are subjected to stronger selective pressure than other genomic regions. However, expansion or deletions of trinucleotide and hexanucleotide repeats in coding region do not perturb reading frames and therefore less frequency of other repeat types in coding regions should be a result of negative selection against frameshift mutations [15, 18].

Identified repeats in *S. lycopersicum* were contained 11 different types of dinucleotides, 60 types of trinucleotides, 73 types of tetranucleotides, 211 types of pentanucleotides, and 411 types of hexanucleotide motifs. Among all types of repeats identified in *S. lycopersicum* ESTs, AT/AT, AG/CT, AAG/CTT and AAT/ATT units, considering sequence complementary, were detected more frequently than other motifs. AT/AT, AAG/CTT, AAT/ATT, AAAT/ATTT, AAAAT/ATTTT and AAAAAT/ATTTTT motifs comprised the 49%, 31.2%, 14.1%, 24%, 18% and 7% of all identified dimers, trimers, tetramers, pentamer and hexamers, respectively.

We observed 8 types of dinucleotides, 44 different types of trinucleotides, 10 different types of tetranucleotides, 21 different types of pentanucleotides and 66 different types of hexanucleotide motifs in examined *S. pennellii* ESTs.

The most abundant repeat classes among all types of identified motifs in *S. pennellii*, when considering sequence complementary, were AAG/CTT (16%), AT/AT (9%) and AGC/CTG (7%). Frequency of AAG/CTT and AGC/CTG motifs among the trinucleotides were 34% and 15%, while this number was equal to 70% and 22% for the most frequent dimers AT/AT and AG/CT. The AAAT/ATTT (45%) was the most frequent among tetramers, whereas AAAAT/ATTTT (29%) and AAAAAT/ATTTTT (5%) were the most dominant among penta and hexamers, respectively.

S. habrochaites, EST-SSRs were contained 10 types of dinucleotides, 48 different types of trinucleotides, 28 different types of tetranucleotides, 98 different types of pentanucleotides and 113 different types of hexanucleotide motifs. Dominant types of motifs among all were AG/CT (9.8%), AAG/CTT (7.6%), and ACC/GGT (7.5%). The AG/CT (56%) and AT/AT (38%) motif comprised the majority of the dinucleotide repeats. In term of trinucleotide motifs, the most common types were AAG/CTT and ACC/GGT, each representing 24% of all trimers. AATT/AATT (24%), AAAT/ATTT (22%), and AGGG/CCCT (22%) had the highest occurence among identified types of tetramers, whereas AATAT/ATATT (45%), AAAAAG/CTTTTT (8.4%) and AAAAAT/ATTTTT (8%) were the most commons among penta and hexamers.

Among trinucleotide repeats, AAG/CTT made up the highest proportion in all three species examined in our study, which is in agreements with other researches, since several authors had been identified that AAG/CTT were the most frequent trinucleotide motif in majority of plants [14]. According to Lopez and et al., trinucleotide repeats abundance in exons might be useful in evolutionary and conservation studies [17].

The dinucleotides AG/CT and AT/AT also were the most abundant microsatellites in EST sequences, which is consistent with previous surveys. For example, the higher frequency of AG/CT (27,7%) and AAG/TTC (17.37%) have been observed in cultivated peanut and its wild species [16]. High frequency of ACC/GGT (24.7%), AAC/GTT (5.9%), AAG/CTT (17.4%), AAT/ATT (13.3%) trinucleotide repeats were found in lettuce (*Lactuca sativa* L.) [24]. Besides, these motifs were the dominant motifs (AG/CT, 33.8%, AAG/CTT, 13.9%) identified in Chinese cabbage ESTs [7]. The AG/CT and AT/AT motifs were the most common dinucleotide repeats in citrus with 54.4% and 22.3%, respectively [20]. Several researchers observed higher frequency for GA/CT repeat than that of AT repeats in exons and ESTs of *Arabidopsis thaliana* and cereals [13, 19] and seem to be characteristic of the plant genomes [5, 6]. We observed such a result in *S. habrochaites* where AG/CT repeats made up 56% of all dinucleotide motifs identified. It was also observed that AT-rich repeat patterns were the most abundant among penta- and hexanucleotides in all three species.

The expressed sequence tags (ESTs) databases are important resources to develop functional SSR marker. The availability of a number of EST-SSRs and other functional markers contribute to the functional diversity analysis, comparative mapping, marker-assisted selection and etc. Obtained results on survey of simple sequence repeats demonstrate their abundance in expressed parts of genome and potential of ESTs for development of microsatellite markers by mining of available tomato databases. Nevertheless, since the conserved nature of the EST-SSRs might limit their polymorphism, experimental validation both the polymorphic nature and transferability of identified SSRs should be confirmed in future laboratory researches.

4. References

1. Bai Y, Lindhout P: **Domestication and breeding of tomatoes: what have we gained and what can we gain in the future?** *Ann Bot-London* 2007, **100**(5): 1085-1094.
2. Behura SK, Severson DW: **Motif mismatches in microsatellites: insights from genome-wide investigation among 20 insect species.** *DNA Res.* 2014, dsu036.

3. Bolger A, Scossa F, Bolger M.E, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sorensen I, Lichtenstein G, Fich EA: **The genome of the stress-tolerant wild tomato species *Solanum pennellii*.** *Nature genetics* 2014, **46**(9): 1034-1038.
4. Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R: **Computational and experimental characterization of physically clustered simple sequence repeats in plants.** *Genetics* 2000, **156**(2): 847-854.
5. Chen CM, Hsiao MC, Pai TW, Cheng R, Tzou WW, Chang MDT: **Identify SSR regulators for functional gene sets through cross-species comparison. Lecture notes in operations research 9, optimization and systems biology.** 2008, 204-211.
6. Condit R, Hubbell SP: **Abundance and DNA sequence of two-base repeat regions in tropical tree genomes.** *Genome* 1991, **34**(1): 66-71.
7. Ding Q, Li J, Wang F, Zhang Y, Li H, Zhang J, Gao J: **Characterization and development of EST-SSRs by deep transcriptome sequencing in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*).** *Int. J. of Genomics* 2015, 11p.
8. Ercolano MR, Sebastiano A, Monti L, Frusciante L, Barone A: **Molecular characterization of *Solanum habrochaites* accessions.** *J. Genet. Breed.* 2005, **59**(1): 15.
9. Gupta PK, Rustgi S: **Molecular markers from the transcribed/expressed region of the genome in higher plants.** *Func Integr Genomics* 2004, **4**(3): 139-162.
10. Han B, Wang C, Tang Z, Ren Y, Li Y, Zhang D, Dong Y, Zhao X: **Genome-Wide Analysis of Microsatellite Markers Based on Sequenced Database in Chinese Spring Wheat (*Triticum aestivum* L.).** *PloS One* 2015, **10**(11), p.e0141540.
11. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res.* 1999, ***9:*** 868-877.
12. Jones N, Ougham H, Thomas H: **Markers and mapping: we are all geneticists now.** *New Phytol.* 1997, **137**(1): 165-177.
13. Kantety RV, La Rota M, Matthews DE, Sorrells ME: **Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat.** *Plant Mol Biol,* 2002, **48**(5-6): 501-510.
14. Kumpatla SP, Mukhopadhyay S: **Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species.** *Genome* 2005, **48**(6), 985-998.
15. Li YC, Korol AB, Fahima T, Nevo E: **Microsatellites within genes: structure, function, and evolution**. *Mol Biol Evol,* 2004, **21**(6): 991-1007.
16. Liang X, Chen X, Hong Y, Liu H, Zhou G, Li S, Guo B: **Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and *Arachis* wild species**. *BMC Plant Biol,* 2009, **9**(1), 35.
17. Lopez L, Barreiro R, Fischer M, Koch MA: **Mining microsatellite markers from public expressed sequence tags databases for the study of threatened plan**ts. *BMC genomics* 2015, **16**(1), 781.
18. Metzgar D, Bytof J, Wills C: **Selection against frameshift mutations limits microsatellite expansion in coding DNA.** *Genome Res.* 2000, **10**(1), 72-80.
19. Morgante M, Hanafe M, Powell W: **Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes.** *Nat Genet.* 2002, **30**(2): 194-200.
20. Palmieri DA, Novelli VM, Bastianel M, Cristofani-Yaly M, Astúa-Monge G, Carlos EF, Oliveira ACD: **Machado Ma, Frequency and distribution of microsatellites from ESTs of citrus.** *Genet Mol Biol.* 2007, **30**(3): 1009-1018.
21. Peralta IE, Spooner DM: **Morphological characterization and relationships of wild tomatoes (Solanum L. Sect. Lycopersicon).** In *Monographs In Systematic Botany*: Missouri Botanical Garden Press; 2005: **(104**): 227-257.
22. Ranjan A, Ichihashi Y, Sinha NR: **The tomato genome: implications for plant breeding, genomics and evolution.** *Genome Biol.* 2012, **13:** 167.
23. Rick CM: **Natural variability in wild species of Lycopersicon and its bearing on tomato breeding**. *Genet Agrar* (Italy) 1976.
24. SIMKO I: **Development of EST-SSR markers for the study of population structure in lettuce (*Lactuca sativa* L.).** *J Hered.* 2009, **100**(2), 256-262.
25. Suresh BV, Roy R, Sahu K, Misra G, Chattopadhyay D: **Tomato genomic resources database: an integrated repository of useful**

tomato genomic information for basic and applied research. *PloS one* 2014, **9**(1): e86387.

26. Thiel T, Michalek W, Varshney R, Graner A: **Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.).** *Theor Appl Genet.* 2003, **106**(3): 411-422.
27. Tomato Genome Consortium: **The tomato genome sequence provides insights into fleshy fruit evolution.** *Nature* 2012, **485**(7400): 635-641.
28. Varshney RK, Graner A, Sorrells ME: **Genic microsatellite markers in plants: features and applications.** *Trends Biotechnol.* 2005, **23**(1), 48-55.
29. Varshney RK, Thiel T, Stein N, Langridge P, Graner A: ***In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species.** *Cell Mol Biol Lett.* 2002, **7**(2A), 537-546.
30. Yi G, Lee JM, Lee S, Choi D, Kim BD: **Exploitation of pepper EST–SSRs and an SSR-based linkage map.** *Theor Appl Genet.* 2006, **114**(1), 113-130.
31. Yu JK, La Rota M, Kantety RV, Sorrells ME: **EST derived SSR markers for comparative mapping in wheat and rice.** *Mol Genet Genomics* 2004, **271**(6), 742-751.
32. Zhao X, Tian Y, Yang R, Feng H, Ouyang Q, Tian Y, Tan Z, Li M, Niu Y, Jiang J, Shen G: **Coevolution between simple sequence repeats (SSRs) and virus genome size.** *BMC Genomics* 2012, **13**(1), 435.