AcademicPres

Available online at www.notulaebiologicae.ro

Print ISSN 2067-3205; Electronic 2067-3264

Not Sci Biol, 2012, 4(4):127-131

Notulae Scientia Biologicae

Computational Mining and Genome Wide Distribution of Microsatellite in *Fusarium oxysporum* f. sp. *lycopersici*

Sudheer KUMAR*, Deepak MAURYA, Shalini RAI, Prem Lal KASHYAP, Alok Kumar SRIVASTAVA

*National Bureau of Agriculturally Important Microorganisms (NBAIM), Mau, Uttar Pradesh, 275101, India; sudheer.nbaim@gmail.com (*corresponding author)*

Abstract

Simple sequence repeat (SSR) is currently the most preferred molecular marker system owing to their highly desirable properties *viz.*, abundance, hyper-variability, and suitability for high-throughput analysis. Hence, in present study an attempt was made to mine and analyze microsatellite dynamics in whole genome of *Fusarium oxysporum* f. sp. *lycopersici*. The distribution pattern of different SSR motifs provides the evidence of greater accumulation of tetra-nucleotide (3837) repeats followed by tri-nucleotide (3367) repeats. Maximum frequency distribution in coding region was shown by mono-nucleotide SSR motifs (34.8%), where as minimum frequency is observed for penta-nucleotide SSR (0.87%). Highest relative abundance (1023 SSR/Mb) and density of SSRs (114.46 bp/Mb) were observed on chromosome 1, while least density of SSR motifs was recorded on chromosome 11 (7.40 bp/Mb) and 12 (7.41 bp/Mb), respectively. Maximum trinucleotide (34.24%) motifs code for glutamic acid (GAA) while GT/CT were the most frequent repeat of dinucleotide SSRs. Most common and highly repeated SSR motifs were identified as (A)64, (T)48, (GT)24, (GAA)31, (TTTC)24, (TTTCT)28 and (AACCAG)27. Overall, the generated information may serve as baseline information for developing SSR markers that could find applications in genomic analysis of *F. oxysporum* f. sp. *lycopersici* for better understanding of evolution, diversity analysis, population genetics, race identification and acquisition of new virulence.

Keywords: amino acid, chromosome, codon, microsatellite, SSR

Introduction

Fusarium oxysporum f. sp. *lycopersici*, the cause of tomato crown and root rot is an important soil-borne fungus and reduce crop productivity by 10-50% (Borrero *et al.,* 2004). The use of resistant varieties is the most economical and effective way to manage the disease. However, new races of pathogen have been emerged that overcome resistance in currently growing tomato cultivars (Mishra *et al.,* 2010). Therefore, knowledge of the genetic variation within and among populations is an important component to understand the population biology of *F. oxysporum* f. sp. *lycopersici* for developing strategies to enhance the durability of resistance. Virulence tests are commonly used to detect the pathogen variations (Elias *et al.,* 1991) and three distinct races (1, 2 and 3) of *F. oxysporum* f. sp. *lycopersici* have been identified (Cai *et al.*, 2003). However, these tests are subjected to availability of host selection pressure, tedious, inconclusive and preclude nonpathogenic strains. To circumvent these problems, DNA based molecular markers have been used in diversity analysis, virulence evaluation and genetic structure of pathogen races (Lievens *et al.*, 2009).

Simple sequence repeat (SSR) or microsatellite markers have become a preferred choice in recent years for several uses due to their multi-allelic nature, co-dominant inheritance, high abundance, hyper variance, extensive genome coverage, reproducibility, and discriminatory power (Mahfooz *et al.*, 2012). Except for some nuclear restriction fragment length polymorphism (RFLP) (Rosewich *et al.*, 1999) and RAPD (random amplified polymorphic DNA) markers (Balmas *et al.*, 2005), limited molecular markers were available for *F. oxysporum* f. sp. *lycopersici* genetic studies. Nevertheless, recent availability of genome sequence information of *F. oxysporum* f. sp. *lycopersici* has provided the opportunity to study the genome wide distributional pattern of SSRs motifs in tomato root rot pathogen. This study describes comprehensive report on mining and analysis of microsatellite dynamics in *F. oxysporum* f. sp. *lycopersici* using bioinformatics approaches.

Materials and methods

Retrieval of sequences

The nuclear and mitochondrial genome sequences of *F. oxysporum* f. sp. *lycopersici* (strain 53 4287, race 2, VCG 0030) available in *Fusarium* Comparative Database of Broad Institute of MIT and Harvard, Cambridge (http://www.broadinstitute.mit.edu/) were used for the present study.

Received 11 September 2012; accepted 02 November 2012

Microsatellite mining

The retrieved sequences were analyzed for repeat patterns using WebSat (SSR finder program) (Martins *et al.*, 2009). The generated data was further used for screening of SSR containing sequences by Simple Sequence Repeat Identification Tool (SSRIT). The program was run online and the parameters were set for detection of perfect di-, tri-, tetra-, penta- and hexa-nucleotide motifs with a minimum of six repeats. The data were processed and counted with Microsoft Excel 2007.

Statistical analysis

The analysis of SSRs was done based on their types (mono- to hexa-nucleotides), number of repeats, frequency of occurrences of each SSR motif and their distribution in the sequence. The relative abundance and density were calculated by following formulas:

Relative abundance = Number of SSRs / Length of sequence analyzed (Mb)*;*

Relative density = Length of SSR (bp) / Length of sequence analyzed (Mb).

Results

Abundance and density of microsatellite

Total genome sequence data (59.9 Mb) of *F. oxysporum f. sp. lycopersici* was assembled into 423 scaffolds and used to explore mono-, di-, tri-, tetra-, penta- and hexa-nucleotide motifs with a repeat of ≥6 times. A total 13864 SSRs were identified from whole genome data of *F. oxysporum* f. sp. *lycopersici* (Tab. 1). The relative abundance and density of SSRs were 231.45 SSR/Mb and 2643.73bp/Mb, respectively (Tab. 1).

The number of repeat units in di-, tri-, tetra-, penta- and hexanucleotides ranged from 10 to 46, but the majority of SSRs (78.25%) had 6 to 31 repeat units. Tetra- (3837) and tri-nucleotide (3367) repeats were the most abundant repeats in the genome accounting 27.67 % and 24.28 % SSRs followed by penta- (18.09%) and tri-nucleotides (17.10%) repeats (Tab. 2). Maximum frequency distribution in coding region was shown by tri-nucleotide SSR motifs i.e. 34.8% in contig, whereas minimum frequency are observed for penta-nucleotide SSR in contig (0.87%). Mono-, di-, tetra- and penta-nucleotide repeats were the least frequent accounting less than 10% of SSRs in the coding region of genome. Among the mononucleotides, polyT (64) was the longest SSR motif (Tab. 2). The number of repeat units ranged from 22 to 64 among mono-nucleotides, but majority of repeats had 22 repeat units. Some of the highly repeated sequences identified were (A)64, (T)48, (GT)24, (GAA)31, (TTTC)24, (TTTCT)28 and (AACCAG)27 (Tab. 2).

Tab. 1. Number and distribution of SSRs in whole genome of *F. oxysporum* f. sp. *lycopersici*

Total size covered by examined sequences (Mb)	59.9
Total number of SSR identified	13864
Perfect SSR	13608
Compound SSR	256
Imperfect SSR	139
Total relative abundance (SSR/Mb)	231.45
Total relative density (bp/Mb)	2643.73

Microsatellite distribution on chromosome

Chromosome 1 possessed highest number of SSRs (7008) and chromosome 13 had the least number of SSRs (119) (Tab. 3). Four hundred and twenty three SSRs were identified in the contigs not mapped to any chromosome. Tetra-nucleotides repeats were the most abundant (3837) repeats in all the chromosomes accounting 27.67% of SSRs. Maximum relative abundance (1023 SSR/Mb) and density of SSRs (114.46 bp/Mb) was occurred on chromosome 1 and followed by chromosome 2 (148 SSR/Mb, 18.1 bp/Mb), chromosome 7 (147 SSR/Mb, 16.55 bp/Mb) , chromosome 8 (146 SSR/Mb, 17.17 bp/Mb), chromosome 15 (145 SSR/Mb, 18.6 bp/Mb) and 14 (144 SSR/Mb, 21.6 bp/Mb), respectively (Tab. 3). Least relative abundance of repeat motifs were occurred on chromosome 12 (68 SSR/Mb) and 13 (68 SSR/Mb), while least density was observed on chromosome 11 (7.40 bp/Mb) and 12 (7.41 bp/Mb), respectively (Tab. 3).

Frequency of microsatellite classes

SSRs were categorized into three groups based on length of SSR tracts (Fig. 1). Class I, II and III SSRs contain perfect repeats ≥10, 10-20 and <20 nucleotides in length, respectively. Out of 13180 SSRs, 269 repeats (2.04%) were categorized as Class I SSRs. About 9.87% and 88.08% SSRs in *F. oxysporum* f. sp. *lycopersici* genome were classified in Class II and Class III, respectively.

Tab. 2. Distribution of SSR motifs in coding and non-coding regions for *F. oxysporum* f. sp. *lycopersici*

Motif length	Total	Coding region		Non-coding region		Longest SSR motifs *
		Number	%	Number	%	
Mono	2371	106	4.5	2265	95.5	A (48), T (64), G (22), C (22)
Di	522	17	3.25	505	96.75	AT (19), TG (19), GT (24), CT (20)
Tri	3367	1172	34.8	2195	65.2	CAG (10), TTC (17), GAA (31), ATT (25)
Tetra	3837	209	5.4	3608	94.6	CTTG (10), TTTC (24), GTAG (10), AAAG (19)
Penta	2509	22	0.87	2487	99.13	CTCTT (11), TTTCT (28), GAGAA (6), AGCAT (11)
Hexa	1258	248	19.7	1010	80.3	CTAACC (20), TGGCTC (12), GGGTTA (16), AACCAG (27)

Number in parenthesis represents the number of repeats in longest SSR

Tab. 3. Distribution of SSRs in different chromosomes identified from public available whole genome database of *F. oxysporum* f. sp. *lycopersic*

Chromosome	Mono	Di	Tri	Tetra	Penta	Hexa	Size (Mb)	SSR (Mb)	bp (Mb)	SSR
1	1102	296	1777	1802	1358	673	6.85	1023	114.46	7008
2	140	35	216	176	176	88	5.58	148	18.1	831
3	204	19	128	202	96	40	5.63	122	13.23	689
4	110	21	155	148	86	52	5.21	109	12.86	572
5	91	25	160	231	119	72	4.91	142	17.07	698
6	128	12	96	158	85	34	4.59	111	11.95	513
7	99	25	153	180	126	60	4.35	147	16.55	643
8	92	15	139	165	105	67	3.98	146	17.17	583
9	65	18	110	122	89	38	3.3	133	17.22	442
10	75	10	97	128	63	36	2.9	141	21.59	409
11	30	4	56	59	19	9	2.34	75	7.402	177
12	26	10	43	46	19	8	2.23	68	7.41	152
13	10	2	44	36	14	13	1.75	68	8.86	119
14	36	10	45	99	45	17	1.65	144	21.6	252
15	63	8	75	147	41	19	2.43	145	18.6	353
Scaffold	100	12	73	138	68	32	3.75	112	7.592	423
Total	2371	522	3367	3837	2509	1258	59.1	234	2643.73	13864

Codon repetition and amino acid distribution

Tri-nucleotide SSRs are triplet codon that code for a particular amino acid. It was observed that out of all triplet codons of contig sequences, GAA (encoding glutamic acid) repetitions are predominant (34.24%) and followed by ATT (encoding isoleucine) and TTC (encoding phenylalanine) (Tab. 4). Analysis of all coded amino acid in contigs sequences demonstrated that the serine (391) and leucine (386) had the highest occurrence followed by arginine (248) (Tab. 4). Tryptophane (57), methionine (58) and asparagine (59) were the least occurred amino acids in the ESTs of *F. oxysporum* f. sp. *lycopersici* (Tab. 4).

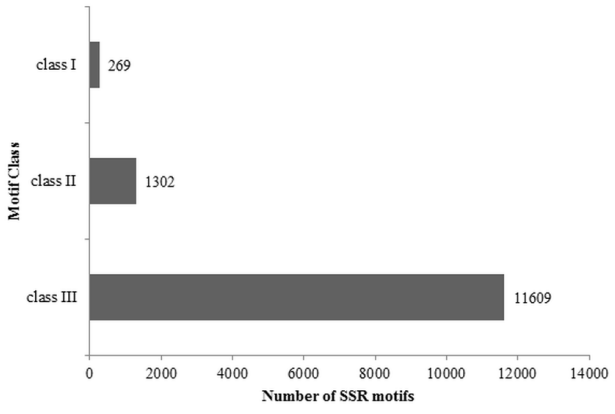

Fig. 1. Frequency of Class I, II and III SSRs in *F. oxysporum* f. sp. *lycopersici* genome

Discussion

Simple sequence repeat (SSR) is currently the most preferred molecular marker system owing to their highly desirable properties *viz.*, abundance, hyper-variability, and suitability for high-throughput analysis. Several studies have shown the importance of using microsatellites to understand epidemiological processes in plant pathogenic fungi (Breuillin *et al.*, 2006; Lievens *et al.*, 2009). Their co-dominance, high polymorphism, and ease of scoring allow inferences of population genetic parameters such as gene flow, effective population size, or reproductive system, to be made with high accuracy (Mahfooz *et al.*, 2012). Most importantly, microsatellite sequences obtained through *in silico* mining have more or less the same utility and potential comparative with those derived from a genomic library. However, the negligible cost of *in silico* mining and high abundance of microsatellites in different sequence resources make this approach extremely attractive for the generation of microsatellite markers. Therefore, in present study, computational approaches were employed to mine and analyze genome wide distribution patter of microsatellite in *Fusarium oxysporum* f. sp. *lycopersici.*

The present study clearly demonstrates that the distribution of microsatellites in the genome is non-random, presumably because of their effects on chromatin organization, regulation of gene activity, recombination, DNA replication, cell cycle, mismatch repair system etc. (Li *et al.*, 2002, 2004). Coding regions are mostly dominated by tri- and hexa- repeats, whereas di-, teta-, and hexa- nucleotide repeats are often found in non-coding regions. Similar, distribution pattern of SSR motifs and predominance of tri- and hexa- motifs in the coding region was reported by Mahfooz *et al.* (2012). These tri- and hexa- SSR motifs in the coding regions are translated into amino acids repeats, which possibly contribute to the biological function of the protein (Kim *et al.*, 2008). Di-nucleotide motifs are often found in the exonic region of *F. oxysporum* (Mahfooz *et al.*, 2012), however, (GT)n repeats were also com-

Tab. 4. Different types of amino acid and their distribution in *F. oxysporum* f. sp. *lycopersici* genome

Amino Acid	Total Number	Nucleotide sequence
Alanine	154	GCT/GCC/GCA/GCG
Arginine	248	AGA/AGA/AGG/CGT/CGC/CGG/CGA
Asparagine	59	AAC/AAT
Aspartic Acid	140	GAT/GAC
Cysteine	129	TGT/TGC
Glutamic Acid	214	GAG/GAA
Glutamine	145	CAA/CAG
Glycine	121	GGT/GGC/GGA
Histadine	113	CAC/CAT
Isoleucine	146	ATA/ATC/ATT
Leucine	386	CTT/CTG/CTA/CTC/TTA/TTG
Lysine	88	AAG
Methionine	58	ATG
Phenylalanine	97	TTC
Proline	117	CCT/CCG/CCA
Serine	391	TCA/TCC/TCT/TCG/AGT/AGC
Stop Codon	139	TAA/TAG/TGA
Threonine	100	ACT/ACA/ACG/ACC
Tryptophan	57	TGG
Tyrosine	90	TAT/TAC
Valine	160	GTC/GTA/GTG/GTT

mon in the *F. oxysporum* f. sp. *lycopersici*. Stallings *et al.* (1991) reported that (GT)n repeat is able to enhance the gene activity from a distance independent of its orientation. However, more effective transcription enhancement results from GT repeats being closer to promoter region.

The frequency distribution by repeat types shows major differences in various genomic regions (Tóth *et al.,* 2000). Tri-nucleotide repeats have been found to be common feature in EST-derived SSRs in present study. High frequency of these repeats in coding regions could be due to mutation and selection pressure for specific amino acids (Morgante *et al.,* 2002). The abundance of trinucleotide repeats EST-SSR is likely due to suppression of other kind of repeats in the coding region, which reduces the frameshift mutations in the coding regions (Metzgar *et al.,* 2000). GAA repeats are very abundant in *F. oxysporum* f. sp. *lycopersici* coding regions, and found very rare in the exons of *F. graminearum* coding region exons (Singh *et al.,* 2011 a), and CTT repeat motif, relatively abundant in *F. graminearum* exons, are uncommon in *F. oxysporum* f. sp. *lycopersici*. These differences could be due to differences in the slippage process, or they may reflect the low GC content of the genome (Richard and Dujon, 1997). The chromosomal location and distribution of SSR-motifs was also predicted in the present study. EST-SSRs appear to be dispersed unevenly across the *F. oxysproum* f. sp. *lycopersici* genome, and there is a higher density of EST-SSRs on chromosome 1. This observation were consistent with the observation of Singh *et al.* (2011 a), where they mentioned that the SSR repeat motif density in *F. graminearum* genome was higher in chromosome 1 relative to other chromosomes.

Role of microsatellites in regulation of gene expression and in the evolution of gene regulation are well documented (Li *et al*., 2002, 2004). Polyleucine and polarginine repeats were reported as abundant amino acids in coding regions of *F. oxysporum* f. sp. *lycopersici*. In regulatory regions, changes in SSR motif length will necessary change the length of DNA in that region, thereby altering the local spatial relationship of transcription factor interactions (Kashi and King 2006).

Microsatellites, generally, show a decrease in abundance with increasing repeat length (Grover *et al.,* 2007) and similar results were obtained in present study, where hexa-repeats were found least abundant in the genome. The longest hexanucleotide repeat motifs in *F. oxysporum* f. sp. *oxysporum* were found to be GGGTTA and similar repeat motif was reported in *P. triticina* and *P. graminis f. sp. tritici* (Singh *et al.,* 2011 b). The rationale behind the categorization of SSR motifs on the basis of length of SSR tracts (Class I, II and III) is that longer perfect repeats are highly polymorphic as noticed in case of *F. graminearum* (Singh *et al.,* 2011 a) and *Fusarium oxysporum* (Mahfooz *et al.,* 2012). Microsatellites in Class III tended to be less variable, representing sites where SSR expansion may occasionally occur but its probability is limited due to a smaller chance of slipped-strand impairing over the shorter SSR template (Temnykh *et al.,* 2001).

In conclusion, the present study has summarized information on cataloging SSRs along with their genomic and chromosomal positions, distribution and dynamics in the

genome of *F. oxysporum* f. sp. *lycopersici*. This information will be very useful for in-depth understanding of fungus evolutionary leading to the formation of repeats in the genome, diversity analysis, population genetics, race identification and acquisition of new virulence.

Acknowledgements

The authors gratefully acknowledge the financial assistance under project 'Outreach project on *Phytophthora, Fusarium* and *Ralstonia* disease in horticulture and field crops' from Indian Council of Agricultural Research (ICAR), India.

References

Balmas V, Scherm B, Di Primo P, Rau D, Marcello A, Migheli Q (2005). Molecular characterization of vegetative compatibility groups in *Fusarium oxysporum* f. sp. *radicis-lycopersici* and f. sp. *lycopersici* by random amplification of polymorphic DNA and microsatellite-primed PCR. European J Plant Pathol 111:1-8.

Borrero C, Trillas MI, Ordales J, Tello JC, Aviles M (2004). Predictive factors for the suppression of *Fusarium* wilt of tomato in plant growth media. Phytopathology 94:1094-1101.

Breuillin F, Dutech C, Robin C (2006). Genetic diversity of the chestnut blight fungus *Cryphonectria parasitica* in four French populations assessed by microsatellite markers. Mycol Res 110:288-296.

Cai G, Gale LR, Schneider RW, Kistler HC, Davis RM, Elias KS, Miyao EM (2003). Origin of race 3 of *Fusarium oxysporum* f. sp. *lycopersici* at a single site in California. Phytopathology 93:1014-1022.

Elias KS, Schneider RW, Lear MM (1991). Analysis of vegetative compatibility groups in nonpathogenic populations of *Fusarium oxysporum* isolated from symptomless tomato roots. Can J Bot 69:2089-2094.

Grover A, Aishwarya V, Sharma PC (2007). Biased distribution of microsatellite motifs in the rice genome. Mol Genet Genomics 277:469-480.

Kashi Y, King DG (2006). Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22(5):253-259.

Kim T-S, Booth JG, Gauch HG, Sun Q, Park J, Lee Y-H, Lee K (2008). Simple sequence repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference. BMC Genomics 9:31.

Li Y-C, KorolAB, Fahima T, Nevo E (2004). Microsatellites within genes: Structure, function, and evolution. Mol Biol Evol 21(6):991-1007.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: A review. Mol Ecol 11(12):2453.

Lievens B, van Baarlen P, Verreth C, van Kerckhove S, Rep M, Thomma BP (2009). Evolutionary relationships between *Fusarium oxysporum* f. sp. *lycopersici* and *F. oxysporum* f. sp. *radicis-lycopersici* isolates inferred from mating type, elongation factor-1alpha and exopolygalacturonase sequences. Mycol Res 113(10):1181-91.

Mahfooz S, Maurya DK, Srivastava AK, Kumar S, Arora DK (2012). A comparative *in silico* analysis on frequency and distribution of microsatellites in coding regions of three *formae speciales* of *Fusarium oxysporum* and development of EST-SSR markers for polymorphism studies. FEMS Microbiol Lett 328(1):54-60.

Martins WS, Lucas DCS, Neves KFS, Bertioli DJ (2009). WebSat - A web software for microsatellite marker development. Bioinformation 3(6):282-283.

Metzgar D, Bytof J, Wills C (2000). Selection against frame shift mutations limits microsatellite expansion in coding DNA. Genome Res 10:72-80.

Mishra KK, vr A, Pandey KK (2010). RAPD based genetic diversity among different isolates of *Fusarium oxysporum* f. sp. *lycopersici* and their comparative biocontrol. World J Microbiol Biotechnol 26(6):1079-1085.

Morgante M, Hanafey M, Powell W (2002). Microsatellites are preferentially associate with non-repetitive DNA in plant genomes. Nat Genet 30:194-200.

Richard G-F, Dujon B (1997). Trinucleotide repeats in yeast. Res Microbiol 148:731-744.

Rosewich UL, Pettway RE, Katan T, Kistler HC (1999). Population genetic analysis corroborates dispersal of *Fusarium oxysporum* f. sp. *radicis-lycopersici* from Florida to Europe. Phytopathology 89:623-630.

Singh R, Pandey B, Danishuddin M, Sheoran S, Sharma P, Chatrath R (2011b). Mining and survey of simple sequence repeats in wheat rust *Puccinia* sp. Bioinformation 7(6):291-295.

Singh R, Sheoran S, Sharma P, Chatrath R(2011a). Analysis of simple sequence repeats (SSRs) dynamics in fungus *Fusarium graminearum*. Bioinformation 5(10):402-404.

Stallings RL, Ford AF, Nelson D, Torney DC, Hildebrand CE, Moyzis RK (1991). Evolution and distribution of (GT) n repetitive sequences in mammalian genomes. Genomics 10:807-815.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch SR (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441-1452.

Tóth G, Gáspári Z, Jurka J (2000). Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10:967-981.