BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(6)

Mining and survey of simple sequence repeats in wheat rust *Puccinia* sp

Rajender Singh*, Bharati Pandey, Mohd Danishuddin, Sonia Sheoran, Pradeep Sharma, Ravish Chatrath

Directorate of Wheat Research, Karnal – 132001, India; Rajender Singh - Email – rajenderkhokhar@yahoo.com; *Corresponding author

Received November 02, 2011; Accepted November 08, 2011; Published November 20, 2011

Abstract:
The abundance and inherent potential for extensive allelic variations in simple sequence repeats (SSRs) or microsatellites resulted in valuable source for genetic markers in eukaryotes. In this study, we analyzed and compared the abundance and organisation of SSR in the genome of two important fungal pathogens of wheat, brown or leaf rust (*Puccinia triticina*) and black or stem rust (*Puccinia graminis f. sp. tritici*). *P. triticina* genome with two fold genome size as compared to *P. graminis tritici* has lower relative abundance and SSR density. The distribution pattern of different SSR motifs provides the evidence of greater accumulation of dinucleotide followed by trinucleotide repeats. More than two-hundred different types of repeat motifs were observed in the genomes. The longest SSR motifs varied in both genomes and some of the repeat motifs are found in higher frequency. The information about survey of relative abundance, relative density, length and frequency of different repeat motifs in *Puccinia* sp. will be useful for developing SSR markers that could find several applications in analysis of fungal genome such as genetic diversity, population genetics, race identification and acquisition of new virulence.

Background:
Simple-sequence repeats (SSRs) or microsatellites, a class of repetitive DNA sequences, consist of 1-6 bp motifs repeated in tandem arrays with identical, composite or degenerate motifs **[1]**. SSRs are ubiquitous, diverse and their distribution is non-random in the genomes of eukaryotes. SSR experience an extremely high rate of reversible and length altering mutations, resulting in extensive allelic variations in repeat number which can be functionally significant **[2]**. Because of their high abundance within the genome and high efficiency in detecting variation, SSR markers have become a widespread analytical tool in various organisms for identification of individuals, species and varieties; evolution, ecology and population studies; as markers for traits and genes; and in the generation of genetic maps. SSRs are also considered to be potential mutators and provide in-exhaustive source of genetic variation for rapid evolutionary adaptation **[2]**. The polymorphism in SSRs is generally believed to be the result of DNA polymerase slippage and unequal recombination **[3, 4]**. SSR markers have been developed for many species of plants, animals and fungi from genomic DNA through the construction of SSR enriched libraries. SSRs have been isolated and characterized in several fungi using library enrichment method; however, fungal microsatellites indeed appeared difficult to isolate using enriched libraries **[5]**. This approach is also labour intensive and time consuming. However in recent years, with the establishment of several sequencing projects in crop plants, animals and microorganisms resulted in a wealth of DNA sequence information. This sequence data for expressed sequence tags (ESTs), genes and cDNA clones can be downloaded from various databases in public domain and by using suitable computer programs these can be scanned for identification of SSRs, referred as EST-SSRs or genic microsatellites. 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.

The rusts of wheat (*Triticum aestivum L.)* are a group of obligate biotrotrophic basidiomycete fungi, causing severe disease in most of the areas of world where wheat is grown **[6]**. Wheat is host to three different rust fungi, black or stem rust (*Puccinia graminis f. sp. tritici*), brown or leaf rust (*P. triticina*), and yellow or stripe rust (*P. striiformis f. sp. tritici*) causing serious losses in wheat production. Rust populations world-wide are highly diverse for virulence phenotype or races. Up to 70 different leaf rust races are identified on annual basis in US **[7]**. In France and Australia, 30-50 and 10-15 races are detected annually **[8, 9]**. The effect of evolutionary processes in fungal pathogen populations may occur more rapidly and display larger effects in agricultural systems than in wild ecosystems because of human involvement by plant breeding and crop management **[10].** The evolutionary forces in agricultural systems are generally stronger due to common farming practices by growing genetically uniform food crops or cultivars across large areas, leading to a rapid change in genetic composition of pathogen population. An understanding of evolutionary processes in plant pathogens has received increased interest from scientific community. Virulence tests are commonly used to detect the pathogen variations and a number of races have been identified. However, virulence tests are subjected to availability of host selection pressure. The DNA based molecular markers provides a powerful tool for virulence evaluation in wheat rust and has been used in diversity analysis **[11, 12]**, virulence evaluation **[10, 13, 14]** and genetic structure of rust races **[15]**.

Large scale genome sequencing projects on a growing number of organisms are providing the opportunity to evaluate the abundance and relative distribution of SSRs in different genera based on available genome sequences. There have been few reports on meaningful comparison of SSR motifs in fungal species **[16, 17]**. Recent availability of genome sequence information of *P. triticina* and *P. graminis f. sp. tritici* have provided the opportunity to study the genome- wide distributional pattern of SSRs in wheat rust pathogens. In this study, we report a comparative assessment of distribution of SSRs between these two species.

Methodology:

Source of sequence data

The nuclear and mitochondrial genome sequences of *P. triticina* and *P. graminis f. sp. tritici* available in *Puccinia* group database of Broad Institute of MIT and Harvard, Cambridge (http://www.broadinstitute.org/) were used for the present study.

Mining of SSRs and comparative analysis

DNA sequences were searched to identify SSRs using Simple Sequence Repeat Identification Tool (SSRIT) which is available at GRAMENE web site http://www.gramene.org/db/searches/ssrtool **[18].** The program was run online and the parameters were set for detection of perfect di-, tri-, tetra-, penta-, and hexanucleotide motifs with a minimum of 6 repeats. The data were processed and counted with Microsoft Excel 2007. The total number of repeats has been normalized in terms of relative abundance and relative density for accurate comparison of repeat types between genomes of different sizes. Relative abundance reveals the frequency of occurrence of particular repeat type in the genome, while relative density reveals the length of sequence in base pair contribute by each repeat type to total sequence analyzed **[17]**. 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 and discussion:

Abundance and density of SSR

We have characterized and analyzed perfect simple sequence repeats in the nuclear and mitochondrial genome sequences of two *Puccinia* species; *P. triticina* and *P. graminis f. sp. tritici*. Genomic sequence data of 162.95 Mb size of *P. triticina* assembled into 38,776 contigs and further assembled into 24,423 scaffolds or supercontigs from Broad Institute (http://www.broadinstitute.org/) was used to search for di-, tri-, tetra-, penta- and hexanucleotide motifs with a repeat of ≥6 times. The SSRIT detected 4,814 SSRs in 1,357 supercontigs from genomic database of *P. triticina*. Dinucleotides repeats were the most abundant (3124) repeats in the genome accounting 65% of SSRs followed by trinucleotides SSRs (25%) **(Table 1, see supplementary material)**. Tetra-, penta- and hexanucleotide repeats were the least frequent repeats accounting 10% of SSRs. The abundance of di- and tri-nucleotide repeats has also been reported expressed sequence tag (EST) derived SSRs in *P. triticina* **[19]**.

Similarly, genome sequence of *P. graminis f. sp. tritici* assembled into 4,557 contigs and further assembled into 392 scaffolds of 88.64 Mb size (http://www.broadinstitute.org/) was analyzed for SSR motifs. We mined 5829 SSRs in 221 supercontigs in the genome sequence. The mitochondrial genome of *P. triticina* and *P. graminis f. sp. tritici* possessed 17 and 11 SSRs only. Dinucleotides repeats were the most abundant repeats in the genome accounting 61% of repeats followed by trinucleotides SSRs (33%). Tetra-, penta- and hexanucleotide repeats were the least frequent repeats accounting only 6% of total SSRs **(Table 1 see supplementary material)**. In a previous study 60,579 EST sequences of *P. graminis f. sp. tritici* were screened for tandemly repeated di- and tri-nucleotide units using bioinformatics approach and identified 708 unisequences containing putative SSRs with six or more repeat units **[11]**.

The total lengths of SSRs represented in *P. triticina* and *P. graminis f. sp. tritici* genomes are 109,937 bp and 117,162 bp representing 0.07% and 0.13% of total DNA sequence analyzed, respectively. *P. triticina* genome has less number of SSRs than *P. graminis tritici* even it has two fold genome size compared to *P. graminis f. sp. tritici*. In a comparision of similar sized genomes, N. Crassa has five fold SSR abundance over that of *F. graminearum* **[17]**. These results indicate that the total SSR contents in fungi are not influenced by the genome size.The relative abundance and density of SSR in *P. graminis f. sp. tritici* genome were 29.39 SSR/Mb and 1,322 bp/Mb, respectively as compared to relative abundance of 65.76 SSR/Mb and SSR density of 674 bp/Mb, respectively in the genome of *P. triticina* **(Table 2 see supplementary material)**. The relative abundance

and density vary among different fungal species and is neither inversely nor directly proportional to the genome size in fungi **[17]**.

Most common and longest SSR

Twelve types of dinucleotide repeat motifs were found in the genome. The CT/TC dinucleotide repeat motif was the most predominant followed by AG/GA motif in *P. triticina*, whereas AG/GA repeats motif was most frequent followed by CT/TC motif in *P. graminis f. sp. tritici* **(Table 3 see supplementary material).** The predominance of AG/GA/CT/TC dinucleotide motifs were also observed in *P. triticina* **[19]**. High frequency of AG repeats has also been reported earlier in *P. graminis f. sp. tritici* **[11]**, *M. grisea*, *N. grassa* **[17]** and *F. graminearum* [20].The CG/GC repeat motif was very rare in both rust genomes. Low abundance of CG/GC repeats has also been reported in other fungi **[17, 20]** and neither of these repeats was found in S. pombe and E. cuniculi **[17]**. However, *P. chrysosporium* and *A. gossypii* were unusual as CG repeat motif was in greater abundance than any other non-mononucleotide repeat motif **[16]**. It was proposed that CG repeats may be infrequent because they must possess a deleterious structural effect on DNA conformation **[21]**. The longest dinucleotide repeat motifs were found to be CA (464 bp) and TG (124 bp) in *P. triticina* and *P. graminis f. sp. tritici*, respectively **(Table 4 see supplementary material)**. The dinucleotide repeats are longer in larger genome of *P. triticina* compared to *P. graminis f. sp. tritici* and also reported in N. crassa and M. crisea **[17]**.

Trinucleotide repeats were found in significant frequency in rust genomes. Among trinucleotide repeats, 60 different types of repeat motifs were identified and the CAG repeat motif was predominant followed by CAC and CAA in *P. triticina* genome **(Table 3 see supplementary material)**. In *P. graminis f. sp. tritici* genome, 58 types of trinucleotide motifs were present and CAA repeat motif was most predominant followed by GTT and TGT. For the trinucleotide SSRs, differences in frequency of repeat motif types were also observed in other fungi. The CAA repeat appeared 152 times in N. crassa genome but only 17 times in F. graminearum. Similarly the group GTT/TGT/TTG occurs 361 times in *N. crassa* **[17]**. The AAC was the most frequent repeat motif in *P. graminis f. sp. tritici*, *S. cerevisiae* and *C. albicans* **[11, 16]**. The frequency distribution by repeat types shows major differences in various genomic regions and among texa **[22]**. Tri-nucleotide repeats have been found to be common feature in EST-derived SSRs. High frequency of these repeats in coding regions could be due to mutation and selection pressure for specific amino acids **[23]**. The abundance of trinucleotide repeats EST-SSR is likely due to suppression of other kind of repeats in the coding region, which reduces the frame-shift mutations in the coding regions **[24]**. The longest trinucleotide repeat motifs were found to be TAC (132 bp) and AAC (144 bp) in *P. triticina* and *P. graminis f. sp. tritici*, respectively **(Table 4 see supplementary material)**. In *N. crassa*, the TTA repeat motif was 279 bp and the longest in observed in fungi was (AAT)129 in *C. neoformans var. gattii* **[16, 17]**. Tetra- to hexanucleotide repeats were less frequent in the rust genomes, however, the repeat types were more diverse compared to di- and trinucleotide repeats. *P. triticina* and *P. graminis f. sp. tritici* showed 56 and 65 types of tetranucleotide repeat motifs, respectively. The most common tetranucleotide repeat was found to be GTTG motif which occurred 34 times in *P. triticina* **(Table 3, see supplementary material)**. It was observed that exons contain almost no tetranucleotide repeats compared to introns and intergenic regions **[22]**. The longest tetranucleotide repeat motifs were found to be AAAC (396 bp) and AGAA (312 bp) in *P. triticina* and *P. graminis tritici*, respectively **(Table 4 see supplementary material)**. The longest tetranucleotide SSR observed was AAAT (404bp) in *C. neoformans var. gattii* **[16]**.

P. triticina contained 73 types of pentanucleotide repeats as compared to 55 in *P. graminis f. sp. tritici*. The most common repeats were found to be GTGTT and AACAA in *P. triticina* and *P. graminis f. sp. tritici*, respectively, occurring 22 times in the genome **(Table 3 see supplementary material)**. The longest repeat motifs were found to be ATTGT (315 bp) and TTGTT (350 bp) in *P. triticina* and *P. graminis f. sp. tritici*, respectively **(Table 4 see supplementary material)**. Similarly, the longest pentanucleotide motif observed was AATGT (404bp) in *C. neoformans var. gattii* **[16]**.

P. triticina and *P. graminis f. sp. tritici* genome possessed 54 and 29 types of hexanucleotide repeats. The longest hexanucleotide repeat motifs in *P. triticina* and *P. graminis f. sp. tritici* were found to be GGGTTA (330 bp) and TTTTTC (270 bp), respectively. Four of five longest repeats were represented by GGGTTA repeat motif in *P. triticina* **(Table 4 see supplementary material)**. The longest hexanucleotide repeat observed was GCCTGA (462) in *M. grisea* **[16]**. The present study revealed the relative abundance and density of SSR motifs in the genome of two important rust pathogens of wheat. SSRs were more abundant in *P. graminis f. sp. tritici* (one SSR/15.2 kb) than *P. triticina* (one SSR/34.1 kb). It was also observed that genome size did not correlate with SSR abundance as *P. triticina* genome is two-fold in size but possess less than half of SSR density and abundance. Similar trends were also observed in other genomes and was suggested that the evolution of SSR may differ in genomes of different sizes **[16]**. Taxon- specific variations were also detected in frequency of SSR motifs **[22]**. When the number of tandem repeats of SSR motifs was examined, it was observed that shorter number of repeats (6-9 times) predominated with 84% and 88% of all SSRs in *P. triticina* and *P. graminis f. sp. tritici*. Similar trends were also observed in other fungal species **[16]**.

Some of the most abundant SSRs have been implicated with regulatory roles in gene expression. The polyglutamine (CAG)n and polyproline (CCN)n repeats were reported in protein coding regions of over 67 different transcription factors **[25]**. In regulatory regions, changes in SSR length will necessary change the length of DNA in that region, thereby altering the local spatial relationship of transcription factor interactions **[2]**. Non-triplet SSRs cause frameshifts in coding region resulting in inactivation of gene expression.

Conclusion:

SSRs provide a ready and virtually inexhaustible supply of new quantitative variations for rapid evolutionary adaptation **[25]** and gene-associated tandem repeats function as facilitators of evolution and enabling rapid evolution of new forms **[26]**. Thus, SSRs are a major source genetic variation has broad implications for understanding the molecular process of evolutionary adaptation, including the evolutionary control of the mutation process itself. The clonally reproducing rust

pathogen rapidly evolves through stepwise mutational acquisition of new virulence **[27]** and thus, resulting arms race between plant breeders and the pathogen is characterized by frequent and rapid resistance breakdowns. The fitness advantage of being able to grow on a resistant cultivar is so strong that a new virulent mutant pathotype can replace an existing dominant pathotype within few years **[28]**. Therefore knowledge of evolutionary change in the pathogen can help plant breeders to develop more efficient strategies of rust resistance management in wheat. This comparative information on the nature of SSR motifs might be best to target for developing molecular markers in rust fungi for strain typing, population genetics, phylogenetics, genetic mapping and evolutionary studies.

Acknowledgement:
The financial support for Agri-Bioinformatics Promotion Program provided by Bioinformatics Initiative Division, Department of Information Technology, Ministry of Communications & Information Technology, Government of India, New Delhi is gratefully acknowledged.

References:

[1] Tautz D & Renz M. *Nucleic Acids Res*. 1984 **12**:.4127 [PMID: 6328411]
[2] Kashi Y & King DG. *Trends Genet* 2006 **22**:253 [PMID: 16567018]
[3] Schlötterer C & Tautz D. *Nucleic Acids Res* 1992 **20**:211 [PMID: 1741246]
[4] Jakupciak JP & Well RD. *J Biol Chem* 1999 **274**:23468 [PMID: 10438526]
[5] Dutech C *et al*. *Fungal Genet Biol* 2007 **44**: 933 [PMID: 17659989]
[6] Kolmer JA. *Curr Opin Plant Biol* 2005 **8**:441 [PMID: 15922652]
[7] Kolmer JA *et al*. *Plant Disease* 2007 **91**:979.
[8] Goyeau H *et al*. *Phytopathology* 2006 **96**:264 [PMID: 18944441]
[9] Park RR. *Australasian Plant Pathology* 1996 **25**:12.
[10] Hovmøller MS & Justesen AF. *Mol Ecol*. 2007 **16**:4637 [PMID: 17887968]
[11] Zhong S *et al*. *Phytopathology* 2009 **99**:282 [PMID: 19203281]
[12] Ordoñez ME & Kolmer JA. *Phytopathology* 2007 **97**:574 [PMID: 18943576]
[13] Ordoñez ME & Kolmer JA. *Phytopathology* 2009 **99**:750 [PMID: 19453235]
[14] Admassu B *et al*. *Journal of Phyopathology* 2010 **158**:806.
[15] Visser B *et al*. *Mol Plant Pathol*. 2009 **10**:213 [PMID: 19236570]
[16] Lim S *et al*. *Fungal Genet. Biol*. 2004 **41**:1025 [PMID: 15465391]
[17] Karaoglu H *et al*. *Mol. Biol. Evol*. 2005 **22**:639 [PMID: 15563717]
[18] Temnykh S *et al*. *Genome Res*. 2001 **11**:1441 [PMID: 11483586]
[19] Wang X *et al*. *Canadian Journal of Plant Pathology* 2010 32:98.
[20] Singh R *et al*. *Bioinformation* 2011 **5**:402 [PIMD: 21423884]
[21] Stallings RL (1992). *Genomics*. 1992 **13**:890.
[22] Tóth G *et al*. *Genome Res*. 2000 **10**:967 [PMID: 10899146]
[23] Morgante M *et al*. *Nat Genet*. 2002 **30**:194 [PMID: 11799393]
[24] Metzgar D *et al*. *Genome Res*. 2000. **10**: 72 [PMID: 10645952]
[25] Kashi Y *et al*. *Trends Genet*. 1997 **13**: 74 [PMID: 9055609]
[26] Fondon JW 3rd & Garner HR. *Proc Natl Acad Sci USA*. 2004 **101**:18058 [PMID: 15596718]
[27] Wellings CR & McIntosh RA. *Plant Pathology* 1990 **39**:316.
[28] Enjalbert J *et al*. *Molecular Ecology* 2005 **14**:2065.

Edited by P Kangueane
Citation: Singh *et al.* Bioinformation 7(6): 291-295 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: Simple sequence repeats in nuclear and mitochondria genome of *P. triticina* and *P. graminis tritici*

Organism	Genome size	SSR repeats					
		Di	Tri	Tetra	Penta	Hexa	Total
P. triticina	162.95 Mb	3124	1175	193	176	117	4785
P. graminis tritici	88.64 Mb	3586	1899	133	164	47	5829
P. triticina (M)*	75.83 kb	6	4	3	3	1	17
P. graminis tritici (M)*	79.75 kb	6	1	0	4	0	11

*M- Mitochondria

Table 2: Relative abundance and density* of SSRs in genome of *P. triticina* and *P. graminis tritici*

Organism	Di	Tri	Tetra	Penta	Hexa	Total
P. triticina	19.17 (310)	7.21 (163)	1.18 (62)	1.08 (73)	0.72 (66)	29.36 (674)
P. graminis tritici	40.46 (618)	21.42 (456)	1.50 (62)	1.85 (152)	0.53 (34)	65.76 (1322)

*Relative density of SSR (in paranthesis)

Table 3: Most frequent SSR motifs* in genome of *P. triticina* and *P. graminis tritici*

Organism	Di	Tri	Tetra	Penta	Hexa
P. triticina	TC (517)	CAG (59)	GTTG (34)	GTGTT (22)	AGGGTT (9)
	AG (480)	CAC (52)	CAAC (30)	TAGCG (21)	GGGTTA (9)
	GA (327)	CAA (50)	TACA (14)	CGCTA (15)	AACCCT (7)
	AT (322)	TGG (37)	AAAC (9)	CTACG (7)	CCCTAA (6)
	CT (297)	TGT (37)	ATGT (8)	GTTGT (7)	CCTAAC (6)
P. graminis tritici	AG (604)	CAA (145)	TTTC (9)	AACAA (22)	TGTTGG (9)
	TC (568)	GTT (107)	AAAG (6)	TTTTG (14)	
	AT (409)	TGT (99)	GAAA (6)	AAAAG (8)	
	CT (397)	TGA (84)		ACAAC (8)	
	GA (391)	AAC (82)		TGTGT (8)	

*Number in parenthesis represents the occurrence of repeat motifs.

Table 4: Longest SSR motifs* in genome of *P. triticina* and *P. graminis tritici*

Organism	Di	Tri	Tetra	Penta	Hexa
P. triticina	CA (232)	TAC (44)	AAAC (99)	ATTGT (63)	GGGTTA (55)
	CA (174)	TAC (41)	CGTA (86)	ACACA (51)	TTAGGG (45)
	AC (172)	ACA (37)	GAAA (81)	GTGTT (50)	GGGTTA (43)
	AC (168)	TAG (37)	TACA (77)	AAAAT (46)	GGGTTA (42)
	GT (135)	CTC (36)	ACGT (75)	TGTGT (41)	GGGTTA (40)
P. graminis tritici	TG (62)	AAC (48)	AGAA (78)	TTGTT (70)	TTTTTC (45)
	TC (48)	TGT (45)	CAAT (72)	AATAA (67)	CCAGCA (34)
	GT (48)	ATC (36)	TGAT (66)	AGAAA (59)	AACCCT (34)
	GT (40)	AAG (35)	AGAA (64)	TTTTG (58)	TTTCTT (33)
	CT (39)	AAT (30)	GAAA (49)	CTTTT (54)	AAAACA (28)

*Number in parenthesis represents the number of repeats in longest SSR