ISSN: 0976-3104
REGULAR ISSUE
Ashraf et al.

RESEARCH ARTICLE OPEN ACCESS

SIMPLE SEQUENCE REPEATS IN SPECIFIC GENE GROUPS OF SHIGELLA GENOME

Hosseini Ashraf[1-2], Indira Ghosh[3], Pramod Khandekar[4], Mohammad Hiresh Ayoubian[5]
[1]*Tehran University of Medical Science, School of Allied Medical Sciences, Tehran, IRAN*
[2]*Institute of Bioinformatics and Biotechnology, University of Pune, Pune, INDIA*
[3]*School of Information Technology, JNU, New Delhi, INDIA*
[4]*Dept. Of Biotechnology, Sinhagad Engineering College, Sinhagad Institute of Technology, Vadguon, Pune, India.*
[5]*Dept. Of Biotechnology, University of Pune, Pune, INDIA*

ABSTRACT

In this paper we attempt to analyze the phenomenon of simple sequence repeats (SSRs) variations in Clusters of Orthologous Groups of proteins (COGs) and horizontal transfer genes (HGT) of Shigella flexneri. We have performed a detailed comparative study of the distribution of SSRs in different gene clusters. According to our finding SSR elements in Shigella pathogenicity islands (PAIs) are significantly overrepresented than in other gene clusters of Shigella particularly in sfII islands which have implications in Shigella virulence and also in virulence genes of 2 Shigella plasmids. The trinucleotide groups of SSRs, the codon repetitions and the amino acid repeats have biased distribution in different gene clusters. Data have been found in this study are subject to; (I) Strong selection of SSRs in PAIs which have important roles in Shigella virulence and (II) important roles of SSRs in determining protein function and genetic development.

Received on: 7th-July-2012
Revised on: 28th-July-2012
Accepted on: 4th– Sept-2012
Published on: 4th –Feb-2013

KEY WORDS

SSR; Shigella; gene clusters; PAI; COGs

Corresponding author: Email: ashosaini@yahoo.co.in; ash-hosseini@tums.ac.ir; Tel: +91-989189995351; Fax: 0982188622533

[I] INTRODUCTION

Simple sequence repeats (SSRs) may provide an evolutionary advantage; they may function as evolutionary tuning knobs by allowing fast adaptation to new environments [1, 2]. Numerous lines of evidence have demonstrated that genomic distribution of simple sequence repeats (SSRs) is nonrandom, presumably because of their effects on chromatin organization, regulation of gene activity, recombination, DNA replication, cell cycle, mismatch repair (MMR) system, etc [3]. Recently, however, many reports have demonstrated that a large number of SSRs are located in transcribed regions of genomes, including protein-coding genes and expressed sequence tags (ESTs) [4], although in general, repeat numbers and total lengths of SSRs in these regions are small [5,6]. Debates over whether SSRs play any functional role in organism development, adaptation, survival, and evolution are never-ending. The currently available information on the location of specific SSRs in known genes and ESTs permits the unraveling of the biological significance of SSR distribution, expansion, and contraction in the functioning of the genes themselves [7].

Prokaryotic and eukaryotic repeat families are clustered to non homologous proteins. This may indicate that repeated sequences emerged after these two kingdoms had split. The eukaryotes incorporating more repeats may have an evolutionary advantage of faster adaptation to new environments [8]. In a variety of organisms, it has been demonstrated that microsatellite mutation rates are positively correlated with repeat number [9]. In prokaryotes, strong positive selective pressures are associated with highly mutable microsatellite tracts that control pathogenicity [10].

The presence of SSRs in prokaryotes is rare, but most that do occur are related to pathogenic organisms; their variation in repeat numbers can also cause phenotypic changes [11]. Haemophilus influenzae (Hi), an obligate upper respiratory tract commensal/pathogen, uses phase variation (PV) to adapt to host environment changes. Switching occurs by slippage of SSR repeats within genes coding for virulence molecules. When SSR repeats lie within protein coding regions, UTRs, and introns, any changes by replication slippage and other mutational mechanisms may lead to changes in protein function [12].

Shigella is an important human pathogen, responsible for the majority of cases of endemic bacillary dysentery prevalent in developing nations [13]. Shigellosis is common among children less than five years of age in developing countries and in persons who travel from industrialized to less developed countries [14].

THE IIOAB JOURNAL
ISSN: 0976-3104

In this paper, we attempt to analyze the phenomenon of SSR variation in clusters of orthologous groups of proteins (COGs) and horizontal transfer genes (HGT) of *Shigella flexneri* as it is common among children in developing countries. We have performed a detailed comparative study of the distribution of SSRs in different gene clusters.

[II] MATERIALS AND METHODS

2.1. DNA sequences

Genome sequences were obtained from ftp://ncbi.nlm.nih.gov/genbank/genomes/.
Gene groups Distributed by COGs (Clusters of Orthologous Groups of proteins) were obtained from http://www.ncbi.nlm.nih.gov/sutils/coxik.cgi?gi=257.
Pathogenicity islands of *Shigella flexneri* 2a str 301 were determined using http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=NC_004337.
Pseudogenes, plasmid virulence genes, and chromosome virulence genes were determined by using SHIBASE, an integrated database for comparative genomics of *Shigella*, at http://www.mgc.ac.cn/ShiBASE/VFs.htm

2.2. Analysis methods

We used the software developed by Gur-Arie et al [15] to screen the COGs, *Shigella* islands, pseudogenes, and virulence genes of *Shigella* for SSRs with a motif length between 1 and 10 bp and a minimal number of three repeats and an entire SSR array length of at least two. This software can be downloaded from ftp://ftp.technion.ac.il/supported/biotech/ssr.exe.

Filter of DNA is done by using http://puma.icb.usp.br/sms/filter_dna.html, for remove non-DNA characters from text including digits and blank spaces from a sequence.
Shuffle nucleic acid sequence has done to produce a randomized sequence with the same overall composition as the original sequence by useing sequence shuffling tool downloaded from http://bcf.arl.arizona.edu/resources/online_tools/shuffle.php.

2.3. Calculation of the expected number of mononucleotide SSRs

To compare the observed number of SSRs with the expected number, we calculated the expected number of homo-oligomer tracts of t bases in a sequence of length N using the formula given by De Wachter T1,t = ∑ P ti (1-Pi)×[(N-t-1)(1-Pi)+2] (summed over i = 1 – 4), with pi being the frequency of each base in the sequence (R).

[III] RESULTS

The Using above mentioned computer programs, we analyzed most important gene groups/mechanisms of *Shigella* which have a important role in *Shigella* function and pathogenicity. The frequency of total SSRs (%5.49) and mononucleotide SSRs (%5.28), with a minimal repeat of 3 in *Shigella* islands, is higher than the other clusters. The frequency of dinucleotide SSRs in *Shigella* islands and transcription genes (%0.17) as well as trinucleotide SSRs in translation genes (%0.06) are higher, followed by intracellular trafficking and secretion genes (%0.053). The frequency of total SSRs (%4.23) and mononucleotide SSRs (%4.18) is lower in transport genes **[Table–1]**. The ratio of observed mononucleotide SSRs to expected mononucleotide SSRs for A and T is greater than 1 and for C and G smaller than 1, except for IS elements and A8 in cell wall/membrane genes and T8 in translation genes. The ratios of observed mononucleotide SSRs to expected mononucleotide SSRs for A3–A5 and T3–T5 for different gene clusters are nearly identical but increases with increasing motif length differences. The ratio of A and T mononucleotide SSRs with increasing motif length was shown to increase, except for A8 and T8, which decrease. The ratios of A6–A8 of *Shigella* islands are much higher than those of other gene groups. The ratio of observed mononucleotide SSRs to expected mononucleotide SSRs in IS elements is totally different from other gene groups **[Figure–1]**. Genes with a high GC content have a lower mononucleotide and dinucleotide SSR density than genes with a low GC content. The numbers of mononucleotide and dinucleotide SSRs are negatively correlated with GC content (Figure 2**)**. The A/T compositions of mononucleotide repeats in all investigated genes are much higher than the C/G compositions of those repeats. Differences were significant by X^2 test ($p < 0.0001$) **[Figure–3]**. In dinucleotide SSRs the frequency of GC/CG in all investigated genes except *Shigella* islands, are higher while frequency of AT/TA in most of genes are lower. Most difference between frequency of CG/GC and AT/TA has observed in the amino acid transport genes (%62.9 and %4.3 respectively) and least difference has observed in the *Shigella* islands (%23.8 and %15.0 respectively). The frequency of AC/GT dinucleotide repeats is higher in intracellular trafficking and secretion genes and *Shigella* islands than other genes. The frequency of AG/CT dinucleotide repeats is higher in posttranslational modification, protein turnover, chaperones genes and signal transduction mechanisms genes. The frequency of TC/GA dinucleotide repeats is higher in signal transduction mechanisms genes and repair, recombination, replication genes. The frequency of TG/CA dinucleotide repeats is higher in repair, recombination, replication genes and *Shigella* islands **[Figure–4]**.

The codon repetition of CAG for (Gln) and GCG for (Ala) are most abundant in all gene clusters except translation genes. In translation genes ACG for (Thr) and CGC for (Arg) are the most abundant. Alanine and Arginine repetition are strongly overrepresented in most of the gene clusters. Alanine and serine repetition in the signal transduction mechanism genes as well as alanine and glutamine in the replication recombination and repair genes are strongly overrepresented. Arginine and glutamine in the inter trafficking and secretion genes are overrepresented.

In contrast the frequency of amino acid repeats in *Shigella* islands is nearly identical and there is no overrepresentation of alanine or arginine such as other gene groups. Differences between frequency of amino acid repeats in *Shigella* islands

THE IIOAB JOURNAL
ISSN: 0976-3104

and other gene groups were significant by $\chi 2$ test ($P<0.001$) **[Figure–5]**.

The frequency of mononucleotide SSRs in Sf II island are higher than other SIs gene groups and it is lower in Sci islands **[Table–2]**. The percentage of mono, di, tri and total SSRs ≥3bp in mxi-spa genes are more than vir genes in both investigated plasmids **[Table–3]**.

Table: 1. Frequency of SSRs in different gene clusters of *Shigella.f 2a str 301*

Genes	Length bp		Total SSRs		Mono nucleotide Repeats		Di nucleotide Repeats		Tri nucleotide Repeats	
	N	%	N	%	N	%	N	%	N	%
IS	303097	6.6	13447	4.43	13080	4.31	322	0.11	45	0.015
SI	430105	9.3	23620	**5.49**	22731	**5.28**	741	**0.17**	141	0.032
PS	363034	7.9	17330	4.77	16720	4.6	476	0.13	131	0.036
RRR	423084	9.3	18387	4.34	17587	4.16	446	0.11	93	0.022
TrS	135834	3.0	5975	4.4	5699	4.2	197	0.15	75	**0.06**
TrC	192078	4.2	9021	4.7	8621	4.5	322	**0.17**	74	0.04
CWMM	244116	5.4	11843	4.85	11349	4.65	384	0.16	97	0.04
EP	292344	6.3	13151	4.5	12555	4.3	446	0.15	134	0.046
TR	1093803	23.7	46376	**4.23**	44479	**4.18**	1645	0.15	468	0.043
InCTS	95832	2.1	4468	4.66	4405	4.6	107	0.11	51	**0.053**
PTMPTC	113073	2.4	5167	4.57	4812	4.3	136	0.12	58	0.05
STM	137145	3.0	6478	4.73	6232	4.53	191	0.14	53	0.04
tRNA	7475	0.16	326	4.36	317	4.24	8	0.11	1	0.013
rRNA	32821	0.71	1723	5.25	1675	5.1	43	0.13	5	0.015

IS: Insertion Sequences; **PS**: Pseudogenes; **SI**: *Shigella* islands; **RRR**: Repair, recombination, replication; **TrS**: Translation; **TrC**: Transcription; **CWMM**: Cell wall/ Membrane Mechanism; **EP**:Energy production; **TR**: Transport Genes; **InCTS**: Intracellular trafficking and secretion; **PTMPTC**: Posttranslational modification, protein turnover, chaperones; **STM**: Signal transduction mechanisms

Table: 2. SSRs in 4 groups of *Shigella* Islands with sizes >1 kb in chromosome of *Shigella.f 2a str 301*

Mono nucleotide Repeats	Sci islands 21440bp		ipaH islands 98767bp		SHI-1 &2 80483bp		Sf II island 28913bp	
	N	%	N	%	N	%	N	%
Mono	834	**3.89**	4918	4.98	3982	4.95	1482	**5.13**
Di	30	0.14	138	0.14	135	0.17	57	0.2
Tri	8	0.037	35	0.035	30	0.037	8	0.03
Tetra	0	0.0	2	0.002	0	0.0	0	0.0
Total	872	**4.1**	5093	5.16	4147	5.15	1547	**5.35**

Table: 3. Frequency of SSRs in Vir genes and Mxi Spa genes of plasmids pCP301 and pSD1_197

Plasmids	pCP301						pSD1_197					
Virulence genes	Mxi Spa genes 25551bp		Vir genes 32551bp		Total 58102bp		Mxi Spa genes 25448bp		Vir genes 27293bp		Total 52741bp	
SSRs	N	%	N	%	**N**	**%**	N	%	N	%	**N**	**%**
Mono≥3bp	1571	**6.1**	1781	5.5	3352	5.77	1566	**6.2**	1463	5.36	3029	5.74
Di≥3bp	70	0.27	66	0.2	136	0.23	68	0.3	54	0.2	122	0.23
Tri-Hexa≥3bp	12	0.05	6	0.02	18	0.03	8	0.03	8	0.03	16	0.03
Total≥3bp	1653	**6.47**	1853	5.69	3506	6.03	1642	**6.45**	1525	5.59	3167	6.0

Vir genes including: icsA, ipaA, ipaB, ipaC, ipaD, ipaH, ipah1.4, ipaH4, ipaH7.8, ipaH9.8, ipaJ, ipgA, ipgB1, ipgB2, ipgC, ipgD, ipgE, repA, repB, virA, virB, virK.
Mxi-spa gene including: MxiA, MxiC, MxiD, MxiE, MxiG, MxiH, MxiI, Mxij, MxiK, MxiL, MxiM, MxiN, ospB, ospC1, ospC2, ospC4, ospD1, ospD2, ospD3, spa13, spa15, spa24, spa29, spa32, spa33, spa40, spa47, spa orf10

THE IIOAB JOURNAL
ISSN: 0976-3104

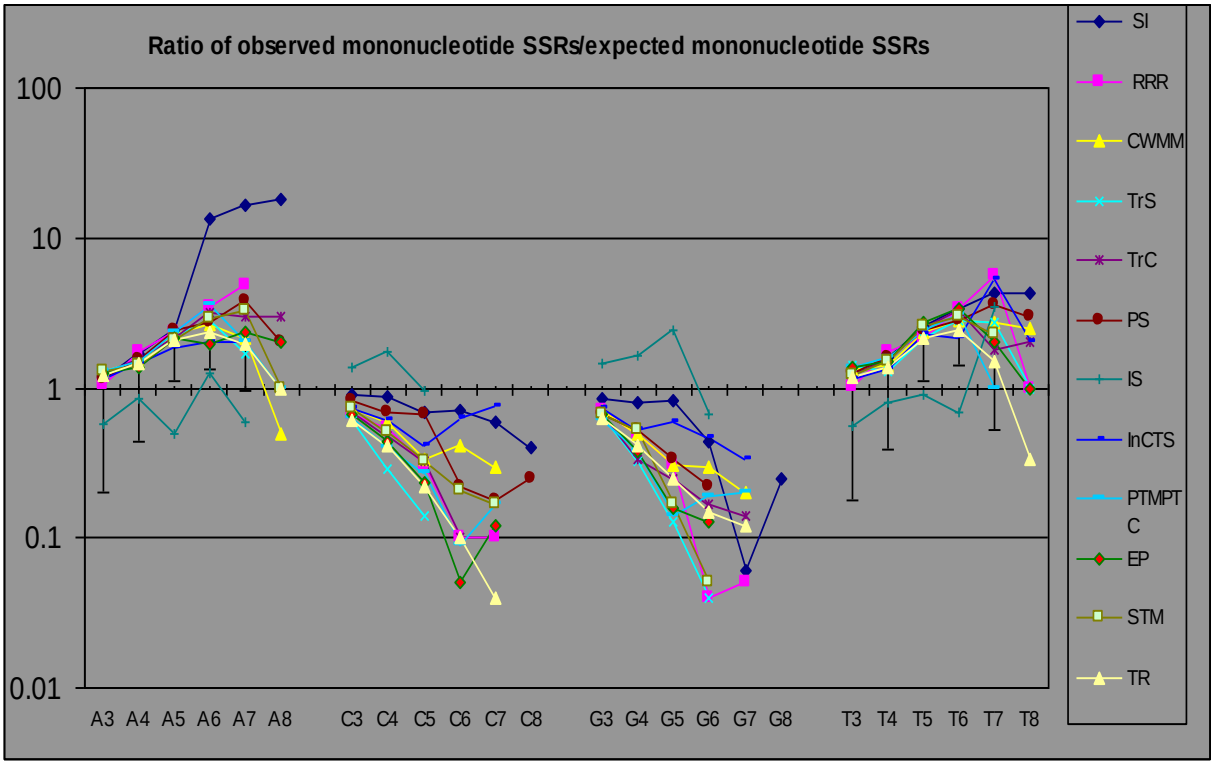

Fig: 1. Ratio of observed Mononucleotide SSRs /expected Mononucleotide SSRs

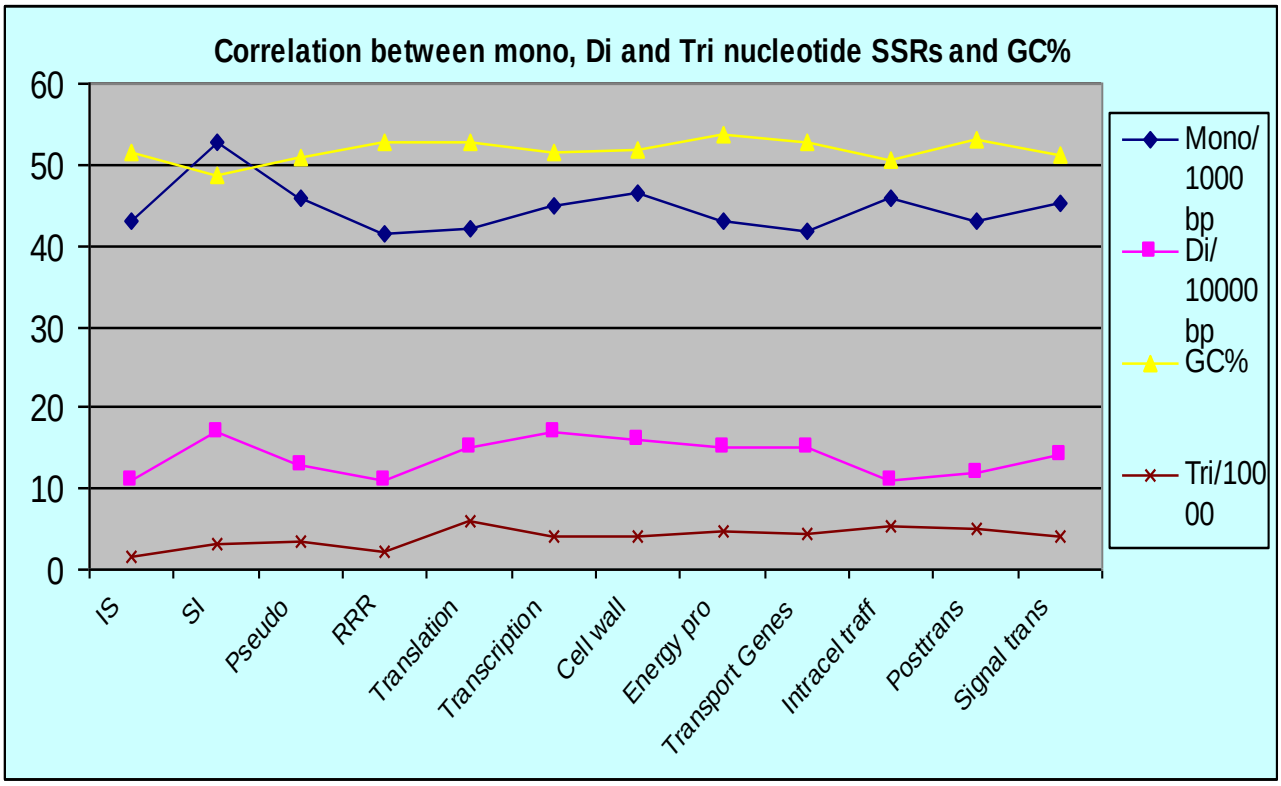

Fig: 2. Correlations between mono, di and trinucleotide SSRs with GC%in different gene clusters

THE IIOAB JOURNAL
ISSN: 0976-3104

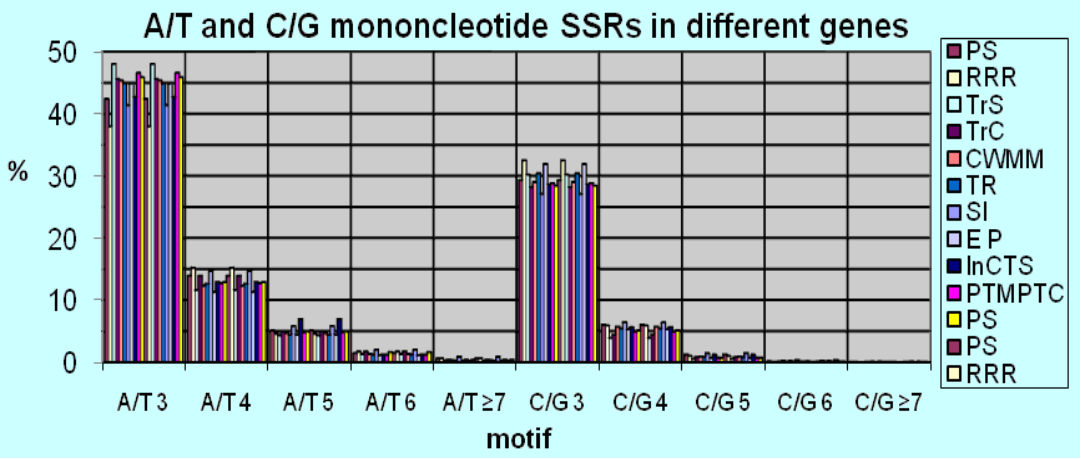

Fig: 3. A/T and G/C mononucleotide SSRs in different gene clusters of *Shigella.f 2a str301*

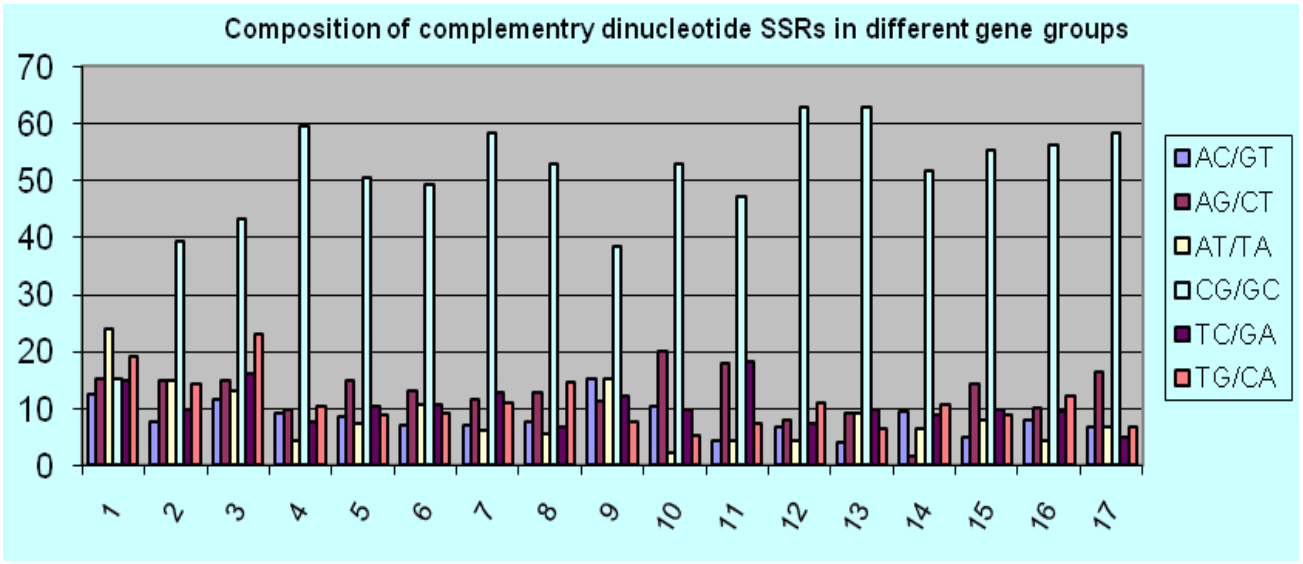

Fig: 4. Composition of complementary dinucleotide SSRs in different gene clusters of *Shigella.f 2a str 301*
1:SI 2:PS 3:RRR 4:TrS 5:TrC 6:CWMM 7:TR 8:EP 9:InCTS 10:PTMPTC 11:STM 12:Amino Acid Transport 13;Coenzime Transport 14:Carbohydrate Transport 15:Inorganic Ion T 16;Lipid Transport 17:Nucleotide Transport

[IV] DISCUSSION

4.1. Association of SSRs with protein function

Although in prokaryotes SSRs are not as abundant as in eukaryotes, most of the SSRs in bacteria are located in virulence genes and/or regulatory regions, and they affect pathogenesis and bacterial adaptive behavior, indicating the signature of natural selection [12, 16]. This hypothesis has been supported by our finding. For instance in our finding SSR in *Shigella* Pathogenicity islands are overrepresented than in other gene groups of *Shigella*, particularly ipaH islands, sf II island and SHI-1&2 which have implications in *Shigella* virulence. Thus, phage-mediated horizontal DNA transfer appears to be one of the major routes by which *Shigella flexneri* gains virulence determinants.

However regarding to overrepresentation of SSRs in ipaH, there is evidence that *S. flexneri* expresses more IpaH within host cells, and the proteins penetrate the host cell nuclei [17]. This, and the fact that all IpaH proteins have a leucine-rich repeat region found in a diverse group of proteins from bacteria and eukaryotes [18], implies that IpaH might be involved in manipulating host gene expression.

Biased distribution of Codon repetition and amino acid repeats

THE IIOAB JOURNAL
ISSN: 0976-3104

have been found in different gene groups, suggesting that repeats of these kinds are subject to strong selection. Functional associations of amino acid repeats for such a scenario to be valid, amino acid repeats of this kind must be associated in some way with protein function.

Our data has shown the overrepresentation of SSRs in ipaH genes and mxi- spa genes of Plasmids pCP301 and pSD1_197. While the Ipa proteins are essential for the invasion of epithelial cells, and their secretion is mediated by the proteins encoded at the mxi and spa loci [19, 20], the SSRs overrepresentation indicates opportunity of adaptability under different system environment.

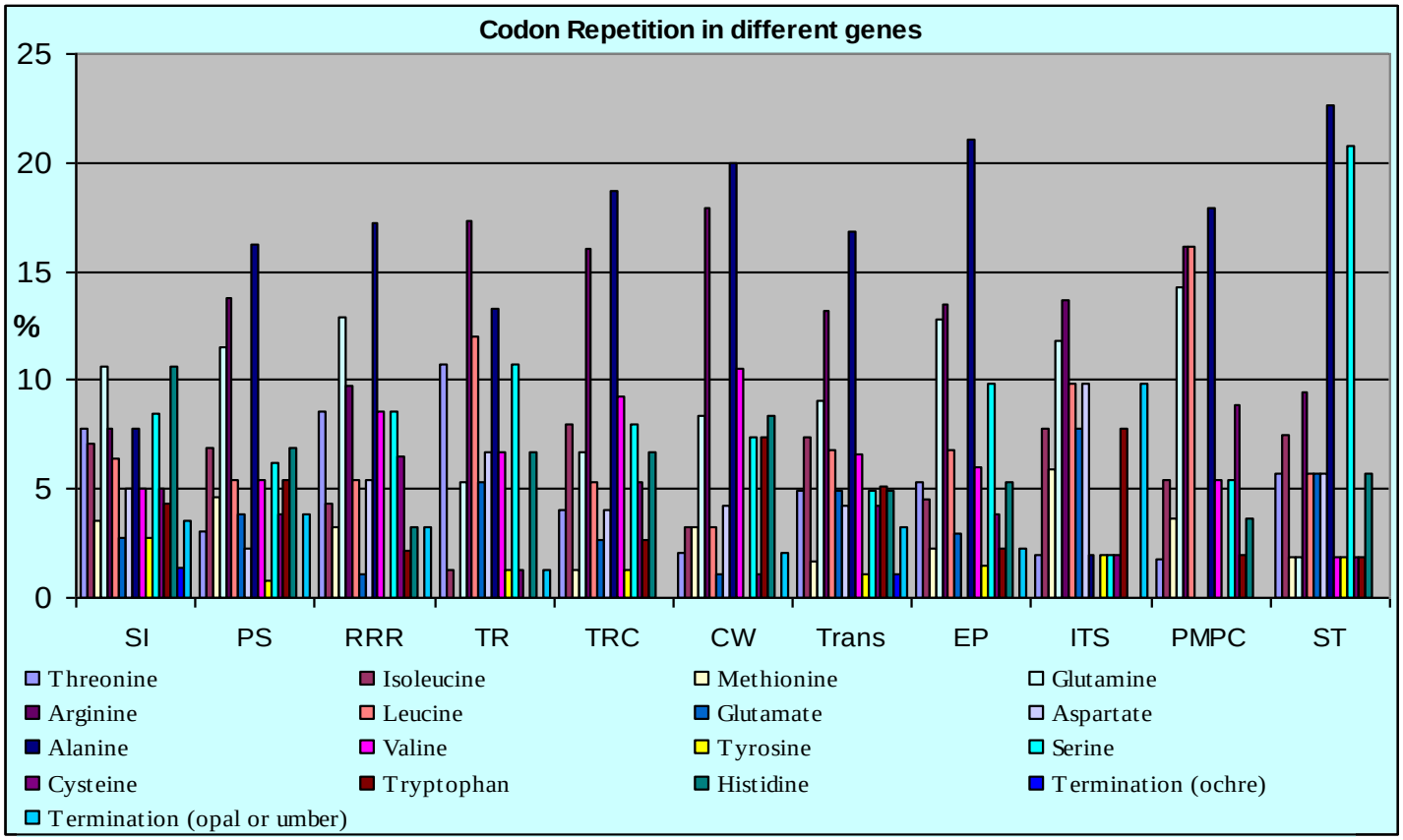

Fig: 5. Frequency of Codon repetition in different gene groups of *Shigella.f 2a str 301*

4.2. SSRs in Shigella Pathogenicity islands >1 kb

The Overrepresentation of total SSRs (%5.49) and mononucleotide SSRs (%5.28) in *Shigella* islands compare with other gene groups (%4.47 and %4.28 respectively), more occurrence of SSRs with increasing size of motif length in *Shigella* islands compare to other gene groups, differences on amino acid repeats pattern and codon repeats frequency between *Shigella* islands and other gene groups, high frequency of AT/TA dinucleotide and low frequency of GC/CG dinucleotide SSRs than the other gene clusters, shows differences between genetic structure of *Shigella* islands (horizontal genes transfer) and other COG genes. These differences may be associated with their evolution or may be generated after integration of PAI-specific DNA regions into the host genome via recombination [21]. It has been speculated that changes in length of repeats in such systems could alter their behavior and therefore contribute to their evolutionary diversification [22, 23], perhaps involving molecular co evolution between proteins [23].

Strongly biased distributions of the all SSR elements in *Shigella* islands have been found in this study emphasize the importance of SSRs in these PAIs which have important roles in *Shigella* virulence.

4.3. SSRs in specific gene groups located in large Shigella Islands>1 kb

Our investigation of SSRs in Sci islands, ipaH islands, SHI-1 and 2 islands and sf II islands, and their implications on the role of virulence, clearly shows differences between frequencies of total and mononucleotide SSRs and the composition of mononucleotide and dinucleotide SSRs between them, particularly between Sci islands and sf II island. This may be

THE IIOAB JOURNAL
ISSN: 0976-3104

associated with their evolution because these islands share homology with genes of different phages. For example, chromosomal ipaH islands are originally linked with phage P27. The Sci island possesses a typical structure of PAI–inserted at an asp-tRNA, and ends with an IS629 on the other side. It also carries paralogs of the Salmonella sci CDEFF operon of unknown function and of phages P22 and HK620. The Sf II island has been demonstrated to be a lysogenic phage required for the expression of the type II antigen. It appears to be one of the major routes by which *S. flexneri* gains virulence determinants. Significant differences occur between SSR elements in the Sf II island with other SIs gene groups by X2 test (P=0.01), including overrepresentation of mononucleotide SSRs, a higher frequency of A/T mononucleotide SSRs and higher frequency of AT/TA dinucleotide SSRs. This may be associated with gene function. By analyzing the Sf II island we found that this island involves 37 proteins, including 12 hypothetical proteins (%5.1 SSR), 12 IS elements (%5.6 SSR), four phage integrases (%6.0 SSR), two putative glucosyl tranferases (%6.9 SSR), and seven other genes (%4.5 SSR). Our investigation has shown that SSR in putative glucosyl tranferases and phage integrases is higher than the other genes, implying the involvement of these genes in pathogenic activities.

4.4. Correlation of GC content and SSRs

Our data clearly indicate that the GC content of mononucleotide SSRs is highest when the repeat density is lowest. This suggests that there could be other reasons for the tremendous overrepresentation of poly (A) and poly (T) mononucleotide SSRs. It has been suggested that the higher energy cost of G and C over A and T/U could be the reason for the high variation seen in genomic G+C content. Indeed, the synthesis of GTP requires an additional NAD compared with AMP, while the synthesis of CTP from UTP requires an additional ATP molecule. In addition, due to its central role in metabolism, ATP is abundantly present in the cell [24, 25].

[V] CONCLUSION

In conclusion, the present study has shown biased distribution of trinucleotide groups of SSRs, Codon repetition and amino acid repeats in different gene clusters in *Shigella*. Significant differences between SSR patterns in *Shigella* Pathogenicity islands with other gene groups of *Shigella* are also manifested. The overrepresentation of SSRs in ipaH genes and Mxi, Spa gene particularly plasmid ipaH genes are correlated with pathogenicity of *Shigella*.

Strongly biased distributions of the all SSR elements in *Shigella* islands have been found in this study, emphasize the importance of SSRs in these PAIs which have important roles in *Shigella* virulence.

Study has suggested that SSRs in different positions of a gene can play important roles in determining protein function, genetic development, and regulation of gene expression.

CONFLICT OF INTERESTS

Authors declare no conflict of interests

FINANCIAL DISCLOSURE

The work was carried out without any financial support

ACKNOWLEDGEMENT

Help and support of Director, research fellows, faculty members and staff of Bioinformatics Centre, University of Pune is gratefully acknowledged.

REFERENCES

[1] Kashi Y, King D, Soller M. [1997] Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet* 13:74-78.
[2] Trifonov EN. [2003] Tuning function of tandemly repeating sequences: a molecular device for fast adaptation. Pp. 1–24 in S. P. Wasser, ed., Evolutionary theory and processes: nodern horizons, papers in honor of Eviatar Nevo. Kluwer Academic Publishers. Amsterdam, The Netherlands.
[3] Li, YC, Korol AB. Fahima T, Beiles A, and Nevo E. [2002] Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol* 11:2453-2465.
[4] Morgante M, Hanafey M, Powell W. [2002] Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* 30:194-200.
[5] Kantety RV, Rota M L, Matthews DE, Sorrells ME. [2002] Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol Biol* 48:501–510.
[6] Thiel T, Michalek W, Varshney RK, Graner A. [2003] Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley. *Theor Appl Genet* 106:411-422.
[7] Whittam TS. [1996] Genetic variation and evolutionary processes in natural populations in Escherichia coli and Salmonella typhimurium: Cellular and molecular biology, 2nd edition (eds). 2708–2720
[8] Wren JD, Forgacs E, Fondon WJ, Pertsemlidis A, Cheng SY, Gallardo T, Williams RS, Shohet RV, Minna JD and. Garner HR. [2000] Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am J Hum Genet* 67:345–356.
[9] Schlo¨tterer C, Ritter R, Harr B, Brem G [1998] Micromutation rate of a long microsatellite allele in Drosophila melanogaster provides evidence for allele-specific mutation rates. *Mol Biol Evol* 15: 1269–1274.
[10] Moxon E, Rainey P, Nowak M, Lenski R [1994] Adaptive evolution of highly mutable loci in pathogenic bacteria *Curr Biol* 4: 24–33.
[11] van Belkum A, Scherer S, van Alphen L, and Verbrugh H. [1998] Short-sequence DNA repeats in prokaryotic genomes. Microbiol. *Mol Biol Rev* 62:275–293.
[12] Hood DW, Deadman ME, Jennings MP, Bisercic M, Fleischmann RD, Venter JC, Moxon ER. [1996] DNA repeats

THE IIOAB JOURNAL
ISSN: 0976-3104

identify novel virulence genes in Haemophilus influenzae. *Proc Natl Acad Sci USA* 93:11121-11125.

[13] Kotloff, KL, Winickoff JP, Ivanoff B, Clemens JD, Swerdlow DL, Sansonetti PJ, Adak G K and Levine MM [1999] Global burden of *Shigella* infections: implications for vaccine development and implementation of control strategies. *Bull WHO* 77: 651-666.

[14] Shlim DR, Hoge CW, Rajah R, Scott RM, Pandy P, Echeverria P, 1999. Persistent high risk of diarrhea among foreigners in Nepal during the first 2 years of residence. *Clin Infect Dis* 29: 613–616.

[15] Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y [2000] Simple sequence repeats in Escherichia coli: abundance, distribution, composition and polymorphism. *Gen Res* 10:62-71.

[16] Field D, Wills C, [1998] Abundant microsatellite polymorphisms in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutationpressures and a variety of selective forces. *Proc Natl Acad Sci USA* 95:1647-1652.

[17] Toyotome T, Suzuki T, Kuwae A, Nonaka T, Fukuda H, Imajoh-Ohmi S, Toyofuku T, Hori M, Sasakawa C. [2001] *Shigella* protein IpaH9.8 is secreted from bacteria within mammalian cells and transported to the nucleus. *J Biol Chem* 276: 32071-32079.

[18] Buchanan SG, Gay NJ. [1996] Structural and functional diversity in the leucine-rich repeat family of proteins. *Prog Biophys Mol Biol* 65: 1-44.

[19] Blocker A, Gounon P, Larquet E, Niebuhr K, Cabiaux V, Parsot C, Sansonetti PJ. [1999] The tripartite type III secreton of *Shigella flexneri* inserts IpaB and IpaC into host membranes. *J Cell Biol* 147:683-693.

[20] Ménard R, Sansonetti PJ, Parsot C, Vasselon T. [1994] Extracellular association and cytoplasmic partitioning of the IpaB and IpaC invasins of *S. flexneri. Cell* 79: 515–525.

[21] Hancock JM. [1999] Microsatellites and other simple sequences: genomic context and mutational mechanisms. Pp. 1–9 in D. B. Goldstein and C. Schlo¨ tterer, eds., Microsatellites: evolution and applications. Oxford University Press, Oxford, U.K.

[22] Richard GF, Dujon B. [1997] Trinucleotide repeats in yeast. *Res Microbiol* 148:731-744.

[23] Hancock JM. [1993] Evolution of sequence repetition and gene duplications in the TATA-binding protein TBP (TFIID). Nuc. Acids Res. 21: 2823-2830.

[24] Bentley SD, Parkhill J. [2004] Comparative genomic structure of prokaryotes, *Annu. Rev. Genet.* 38, 771–791.

[25] Rocha EP, Danchin A. [2002]. Base composition bias might result from competition for metabolic resources, *Trends Genet* 18, 291–294.