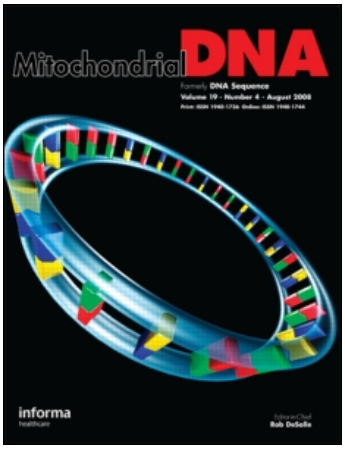


Mitochondrial DNA

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713640135

Simple sequence repeats in different genome sequences of Shigella and comparison with high GC and AT-rich genomes

Ashraf Hosseini [a]; Suvidya H. Ranade [b]; Indira Ghosh [c]; Pramod Khandekar [d]
[a] Institute of Bioinformatics and Biotechnology, University of Pune, Pune, India [b] Department of Zoology, University of Pune, Pune, India [c] Bioinformatics Center, University of Pune, Pune, India [d] Off Campus Unit, Pravara Rural University, Nashik, India

To cite this Article Hosseini, Ashraf, Ranade, Suvidya H., Ghosh, Indira and Khandekar, Pramod(2008) 'Simple sequence repeats in different genome sequences of Shigella and comparison with high GC and AT-rich genomes', Mitochondrial DNA, 19: 3, 167 — 176
To link to this Article: DOI: 10.1080/10425170701461730
URL: http://dx.doi.org/10.1080/10425170701461730

DNA Sequence, June 2008; 19(3): 167–183

informa
healthcare

FULL LENGTH RESEARCH PAPER

Simple sequence repeats in different genome sequences of Shigella and comparison with high GC and AT-rich genomes

ASHRAF HOSSEINI[1†], SUVIDYA H. RANADE[2], INDIRA GHOSH[3], & PRAMOD KHANDEKAR[4]

[1]*Institute of Bioinformatics and Biotechnology, University of Pune, Pune 411007, India,* [2]*Department of Zoology, University of Pune, Pune 411007, India,* [3]*Bioinformatics Center, University of Pune, Pune 411007, India, and* [4]*Off Campus Unit, Pravara Rural University, Sinnar, Nashik, India*

(Received 11 September 2006; revised 4 May 2007; accepted 17 May 2007)

Abstract

Simple sequence repeats (SSRs) are omnipresent in prokaryotes and eukaryotes, and are found anywhere in the genome in both protein encoding and noncoding regions. In present study the whole genome sequences of seven chromosomes (*Shigella flexneri 2a str301* and *2457T*, *Shigella sonnei*, *Escherichia coli k12*, *Mycobacterium tuberculosis*, *Mycobacterium leprae* and *Staphylococcus saprophyticus*) have downloaded from the GenBank database for identifying abundance, distribution and composition of SSRs and also to determine difference between the tandem repeats in real genome and randomness genome (using sequence shuffling tool) of the organisms included in this study.

The data obtained in the present study show that: (i) tandem repeats are widely distributed throughout the genomes; (ii) SSRs are differentially distributed among coding and noncoding regions in investigated *Shigella* genomes; (iii) total frequency of SSRs in noncoding regions are higher than coding regions; (iv) in all investigated chromosomes ratio of Trinucleotide SSRs in real genomes are much higher than randomness genomes and Di nucleotide SSRs are lower; (v) Ratio of total and mononucleotide SSRs in real genome is higher than randomness genomes in *E. coli K12*, *S. flexneri str 301* and *S. saprophyticus*, while it is lower in *S. flexneri str 2457T*, *S.sonnei* and *M. tuberculosis* and it is approximately same in *M. leprae*; (vi) frequency of codon repetitions are vary considerably depending on the type of encoded amino acids.

Keywords: *SSR, microsatellite, Shigella, comparison*

Introduction

The genus *Shigella* an etiological agent of bacillary dysentery, identified in 1890's, a very important member of the family *Enterobacteriaceae* is classified into four etiologically important species viz., *Shigella flexneri*, *Shigella dysenteriae*, *Shigella sonnei* and *Shigella boydii* (Hale 1991).

Simple sequence repeats (SSRs), or microsatellites, are the genetic loci where one or a few bases are tandemly repeated for varying numbers of times (Levinson and Gutman 1987). Repetitive DNA consists of simple homopolymeric tracts of a single nucleotide type (poly (A), poly (G), poly (T), or poly(C)) or of large or small numbers of several multimeric classes of repeats. These multimeric repeats are built from identical units (homogeneous repeats), mixed units (heterogeneous repeats), or degenerate repeat sequence motifs (Jeffreys et al. 1985). SSRs have been extensively studied in eukaryote genomes and are well-established targets for pedigree analysis (Jeffreys et al. 1986). But little is currently known about microsatellites in simple organisms (Field and Wills 1996). Bacterial SSR-type DNA can be divided into four main categories. First, dispersed repeat motifs that generally do not

Correspondence: P. Khandekar, Off Campus Unit, MBA Bioscience Management, School of Bio Science Management, Pravara Rural University, Sinnar, Nashik, India. Tel: 919890008949. E-mail: pramodvidya@sify.com
[†]Tel: 91 9822759492. E-mail: ashosaini@yahoo.co.in

ISSN 1042-5179 print/ISSN 1029-2365 online
DOI: 10.1080/10425170701461730

occur in tandem have been identified. Although these repeats occur throughout genomes of a multitude of microorganisms, they are sometimes organized in tandem as well. The homopolymeric tracts form a second class. Multimers of one of the four nucleotides are peculiar sequence elements that are frequently encountered in the genome of *Saccharomyces cerevisiae*, for instance. These homogeneous stretches can amount to as much as 42 nucleotides. Third, short-motif SSRs are identified. With repeat units differing from 2 to 6 bases, it is this class of repeats that is most liable to unit number variation at a given locus. Particularly, when these short-motif repeats are located within genes and are not 3 or 6 nucleotides long, they can drastically affect the coding potential of a given transcript. Fourth, repeats harboring more than eight nucleotides per unit form a separate category. (Belkum et al. 1998).

Bell (1996) suggested that the abundance and length distribution of SSRs across the genome could result from unbiased single-step random walk processes. Some investigators considered SSRs to be selectively neutral sequences randomly or almost randomly distributed over the euchromatic genome (Schlötterer and Wiehe 1999; Schlötterer 2000).

Initial studies of humans reported a higher mutation rate of tetranucleotide repeats (Weber and Wong 1993), whereas a later study that compared microsatellite variability in different human populations found strong evidence for an inverse correlation of microsatellite repeat unit length and mutation rate (Chakraborty et al. 1997). Prokaryotic and eukaryotic repeat families are clustered to nonhomologous 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 (Kashi et al. 1997; King and Soller 1999; Wren et al. 2000).

In a variety of organisms, it has been demonstrated that microsatellite mutation rates are positively correlated with repeat number (Wierdl et al. 1997; Schlötterer et al. 1998). In prokaryotes, strong positive selective pressures are associated with highly mutable microsatellite tracts that control pathogenicity (Moxon et al. 1994). The increasing availability of prokaryotic genome sequences has shown that SSRs are also widespread in prokaryotes and that there is extensive variation in their length, number and distribution (Cox and Mirkin 1997; Field and Wills l998; Gur-Arie et al. 2000; Coenye and Vandamme 2003; Yang et al. 2003).

In the present study we have analyzed distribution and composition of SSRs in the entire genomes of three strain of *Shigella*, and compared with *E. coli k12*, *GC rich* (*M. tuberculosis* and *M. leprae*) and also AT rich genomes (*S. saprophytycus*).

Materials and methods

DNA sequences

The whole genome sequence of *S. flexneri 2a str301* (NC_004337), *S. flexneri 2a str2457T* (NC_004741), *S. sonnei Ss046* (NC_007384), *E. coli K12* (NC_000913), *M. tuberculosis* CDC1551(NC_002755.2), *M. leprae* TN (NC_002677) and *S. saprophyticus* subsp. *saprophyticus ATCC 15305*(NC_007350) were downloaded from the GenBank database.

Analysis of SSRs

In this study, we have used two software for identifying SSRs. One software was developed by Gur-Arie et al. downloadable from ftp://ftp.technion.ac.il/supported/biotech/ssr.exe to screen the entire genome of the organisms included in this study for SSRs with minimal number of three repeats for chromosomes, minimal motif length of one and minimal length of whole SSR array two. The second software was MICAS (microsatellite Analysis Server) an Interactive web-based server to find non-redundant microsatellites in coding and noncoding region of genome sequence, downloadable from http://210.212.212.7/MIC/gr-ve.html or http://www.cdfd.org.in/micas.

Estimation of the excess of tandem repeats with different factors

To determine difference between the observed and the expected number of tandem repeats in entire genome of the organisms included in this study, distribution of SSRs between coding and non-coding regions of the genome and compare SSRs distributions with random expectations in coding and non coding regions, SPSS 11.0.1, SAS 9.1 and Sequence Shuffling Tool (http://bcf.arl.arizona.edu/resources/online_tools/shuffle.php) have been used. Statistical significance was tested by (χ^2 test and two-tailed *t*-tests.

Results

Distribution of SSRs

By a computer-based screen three chromosomes of *Shigella*, we found large number of SSRs with motif length 1–9 bp scattered through out genome (Table I).

The number of mononucleotide SSRs decreased rapidly with increasing size of the repeat unit, and there is an almost perfect and highly significant linear relationship between the logarithm of the number of mononucleotide repeats and the repeat size $P < 0.0001$ for all genomes).

Mononucleotide SSRs constituted the majority of SSRs in all three *Shigelle* genomes, with the majority of mononucleotide SSRs being ≤ 6 bp.

Table I. Frequency of SSRs ≥ 3 bp in three chromosomes of *Shigella*.

Chromosome Length		*S. sonnei* 4,825,260 bp	Sh.f 2457T 4,599,554 bp	Sh.f 301 4,607,203 bp
Total SSRs	*N*	822,647	807,557	789,224
	%	17.0	17.6	17.1
SSRs in coding region	*N*	639,622	620,468	584,896
	%	16.4	17.5	15.9
SSRs in non coding region	*N*	183,025	187,089	204,328
	%	19.7	17.7	22.2
GC%		50.8	50.9	50.9

As mononucleotide repeat number grow higher, they occur more in noncoding regions than coding regions, but it is no markedly difference in the repeat of di and tri nucleotide SSRs (Tables SI–SIII).

In two strain of *S. flexneri* (301 and 2457T), coding regions contain less dinucleotide and tetranucleotide SSRs than trinucleotide SSRs. In coding regions of *S. sonnei* trinucleotide and tetranucleotide SSRs are more represented than dinucleotide repeats (Tables SI–SIII).

SSR tracts were more distributed over intergenic regions than coding regions as found in other prokaryotes (Gur-Arie 2000; Coenye and Vandamme, P.google html) in following cases: in mononucleotide SSRs ranging in length from 3–10 bp of *S. sonnei*, 3 bp and 5–10 bp of *sh.f 301* and 4–10 bp of *Sh.f2457T.* Dinucleotide SSRs ranging in length ≥ bp of *Sh.f 301* and ≥ bp of *Sh.f2457Tand S. sonnei.* Tetranucleotide SSRs of *Sh.f 301.* Pentanucleotide SSRs *of Sh.f 301* and *Sh.f2457T.* Hexa, Hepta and Octanucleotide SSRs of *S. sonnei.*

For other SSRs there is a deviation from this trend as they are more likely to be in coding regions such as: mononucleotide SSRs ranging in length 4 bp of *Sh.f 301* and 3 bp of *Sh.f 2457T.*

Dinucleotide SSRs ranging in length 3–4 bp of *Sh.f2457T* and *S. sonnei.* Trinucleotide SSRs of all 3 *Shigella* genomes. Tetranucleotide SSRs of *Sh.f 2457T* and *S. sonnei*. Pentanucleotide SSRs of *S. sonnei*. Hexanucleotide SSRs in *Sh.f 301* and *Sh.f 2457T* (Table SI–SIII).

Frequency of SSRs

Total number of SSRs in whole genome, coding regions and noncoding regions, with minimal repeat number 3 for *Shigella* chromosomes has shown in Table I.

In all investigated *Shigella* chromosomes, total frequency of SSRs in noncoding regions are higher than coding regions. Also the frequency of total SSRs in whole genome and coding regions of *Sh.f 2457T* are more than *Sh.f 301* and *S. sonnei*, however in noncoding region occurrence is higher in *S. sonnei* (Table I).

There is significant differences between frequency of total SSRs and also mononucleotide SSRs in coding regions and noncoding regions of three chromosomes of *Shigella* by χ^2 test ($P = 0.0001$).

The frequency of total SSRs (24%), mononucleotide SSRs (21.68%) and Dinucleotide SSRs (1.2%) are higher in genome of *S. saprophyticus* (AT-rich) than other genomes.

Frequency of total SSRs (15.5%) and mononucleotide SSRs (13.04%) is lower in *M. tuberculosis* (GC-rich) and frequency of dinucleotide SSRs is lower in *Sh.fla str301*(0.96%). Though frequency of triplet SSRs in *M. tuberculosis* is much more than other genomes (Table II). The distribution of mononucleotide SSRs over different length categories are significantly different between investigated genomes by χ^2 test ($P = 0.0001$).

The upper limit SSRs

The upper limits for mononucleotide SSRs and any given SSRs are shown in Table III.

Frequency of SSRs in real genomes/randomness genomes in different chromosomes

Total number of SSRs observed in seven computer generated random genomes by using shuffle tool (with the same overall nucleotide frequency as the original genome) were higher than expected by chance alone in *E. coli k12*, *Sh.f 301* and *S. saprophyticus* but it is lower in *Sh.f 2457T*, *S. sonnei*, and specially *M. tuberculosis*, and it is approximately same in *M. leprae* (Table IV).

There was significant difference between frequency of total number of SSRs observed in real genomes and observed in randomness genomes of investigated chromosomes by χ^2 test ($P < 0.0001$).

Ratio of mononucleotide composition in real genome/randomness genome show that there is overrepresentation of A/T mononucleotide SSRs in *Shigella* species and *E. coli K12* (with 50–51% GC content), however in *M. tubeculosis* (with 65% GC content) A/T mononucleotide SSRs in real genome

Table II. Comparison of number of SSRs in three *Shigella* genomes, *Ecolab k12*, *M. tuberculosis*, *M. leprae* and *S. saprophyticus*.

Genomes	*Sh.F 2a. St 2457T*	*Sh.F 2a. St 301*	*Ecolab K12*	*S. sonnei*	*M. tuberculosis*	*M. leprae*	*S. saprophyticus*
Length	4,599,554 bp	4,607,203 bp	4,639,675 bp	4,825,260 bp	4,403,837 bp	3,268,203 bp	2,516,575 bp
No of copy per nucleotide	*N* (%)	*N* (%)	*N* (%)	*N* (%)	*N* (%)	*N* (%)	*N* (%)
GC%	50.9	50.9	50	50.8	65	57	33
Mono	211,892(15.72)	218,725(16.1)	225,446(16.52)	222,843(15.73)	177,013(13.04)	144,775(14.16)	156,616(21.68)
3	151,802(9.9)	157,238(10.2)	163,345(10.6)	158,475(9.85)	141,748(9.7)	124,828(11.46)	105,847(12.6)
4	41,614(3.6)	42,441(3.7)	42,901(3.7)	45,633(3.8)	28,472(2.6)	13,633(1.7)	33,105(5.3)
5	12,916(1.4)	13,613(1.5)	13,837(1.5)	13,178(1.37)	5830(0.44)	5266(0.81)	12,562(2.5)
6	3950(0.52)	4071(0.5)	4123(0.53)	4481(0.56)	818(0.11)	851(0.16)	3591(0.86)
7	1357(0.2)	1142(0.17)	1000(0.15)	848(0.12)	138(0.022)	161(0.034)	1360(0.38)
8	221(0.038)	188(0.02)	217(0.037)	201(0.033)	6	30(0.007)	138(0.04)
9	32(0.006)	24(0.003)	22(0.004)	25(0.0047)	1	3	13
>10	8	8	1	2	0	3	0
Di	7297(0.98)	7289(0.96)	7575(1.06)	7610(0.97)	8526(1.19)	5263(0.98)	4955(1.2)
3	6863(0.90)	6824(0.88)	7081(0.92)	7112(0.88)	7880(1.07)	4978(0.91)	4641(1.1)
4	412(0.072)	437(0.075)	465(0.08)	469(0.078)	606(0.11)	261(0.063)	296(0.09)
5	21(0.0046)	23(0.005)	28(0.006)	28(0.006)	40(0.009)	24(0.007)	16(0.006)
>6	1	1	1	1	0	11	2
Tri	1770(0.356)	1812(0.356)	2401(0.468)	1845(0.365)	3998(0.738)	1542(0.426)	1297(0.469)
3	1715(0.34)	1756(0.34)	2335(0.45)	1789(0.33)	3794(0.78)	1502(0.41)	1263(0.45)
>4	55(0.016)	56(0.016)	66(0.018)	56(0.015)	204(0.058)	40(0.016)	34(0.019)
Tetra	49	49	43	35	64	32	35
Penta	2	2	0	0	7	7	2
Hexa	10	9	3	9	18	4	3
3	7	6	0	9	16	3	3
>4	3	3	0	0	2	1	
Hepta	0	0	0	2	0	1	1
Octa	0	0	2	1	0	0	1
Nona	3	3	0	0	55	0	0

Table III. Upper limits for length of any given SSRs and mononucleotide SSRs in seven genomes.

	Upper limits for any given SSRs					Upper limits for mononucleotide SSRs		
Chromosomes	Motif	Motif length	No. of repeats	Total length	Position	Nucleotide	No of repeat	Position
E. coli K12	ATGAAATG	8	6	48	1,197,686	G	10	379,246
	GCACTATG				2,763,443			
Sh.f 301	TAATGATTT	9	12	108	3,106,844	A	43	3,080,936
Sh.f 2457T	TAATGATTT	9	6	54	3,095,543	A	32	3,071,772
S. sonnei	AGAAAGC	7	14	98	4,203,630	T	29	255,910
M. tuberculosis	ACGGCGGCA	9	7	63	3,921,959	G	9	976,892
M. leprae	GCACCT	6	7	42	1,816,862	G	22	229,636
Staphylococcus saprophyticus	TAAGGAT	7	4	28	564,760	A/T	9	12 position

Table IV. The ratio of SSRs in real genome/randomness genome by shuffling tool in different chromosomes.

			Ratio of SSRs in real genome/randomized genome for								
								Mononucleotide composition			
Chromosomes	Length (bp)	GC%	Total SSRs	Mononucleotide SSRs	Dinucleotide SSRs	Trinucleotide SSRs	Tetranucleotide SSRs	A	C	G	T
E. coli K12	4,639,675	50	1.05	1.06	0.76	3.13	0.84	1.46	0.66	0.74	1.46
Sh.f 301	4,607,203	50.9	1.01	1.02	0.72	2.81	1.1	1.4	0.67	0.68	1.41
Sh.f 2457T	4,599,554	50.9	0.99	0.99	0.72	2.68	0.86	1.34	0.67	0.68	1.34
Sh.sonnei	4,825,260	50.8	0.98	0.99	0.72	2.65	0.79	1.41	0.53	0.7	1.41
M. tuberculosis	4,403,837	65	0.72	0.73	0.70	3.8	0.59	1.01	0.7	0.66	1.02
M. leprae	3,268,203	57	1.00	1.01	0.68	2.86	0.89	1.17	0.77	1.05	1.27
S. saprophyticus	2,516,575	33	1.08	1.09	0.67	2.39	0.54	1.13	0.75	0.81	1.14

is approximately same as randomness genome (Table IV).

There is significant difference between frequency of mononucleotide repeats of *Shigella* genomes and *E. coli K12* with GC-rich and AT-rich genomes by χ^2 test ($P < 0.0001$).

In all investigated chromosomes ratio of real genome to randomness genome for Trinucleotide SSRs are much higher than 1 and Di nucleotide SSRs are lower and also Tetranucleotide SSRs are lower than 1. Average of ratio of real genome to randomness genomes in 7 chromosomes for Di, Tri and Tetranucleotide SSRs are 0.71, 2.9 and 0.8, respectively (Table IV).

There is significant difference between ratio of real genome to randomness genomes in seven chromosomes for Di, Tri and Tetranucleotide SSRs by χ^2 test ($P < 0.0001$).

Density of SSRs:

The distribution of SSRs in chromosome of *S. flexneri str 301* and *2457T* is nearly same except in some regions like 1 020 000–1 060 000, 3 560 000–3 640 000 and 4 180 000–4 240 000 that are more variable than the other regions. The occurrence of SSRs is higher in the hypothetical proteins than RNA associated genes. These regions show more variability on the number of ORFs vary from 29 to 5 and some shifting of hypothetical protein and RNA associated genes occur in these regions of these strains.

Average of SSR density in *Sh.f 2a str2457T* is more than *Sh.f 2a strBOl*, it is 1000.497 and 1000.07 per 20 kb and standard deviation is 50.17 and 47.7, respectively (Figure Sl).

Composition of mononucleotide SSRS:

The A/T composition of mononucleotide repeats in *Shigella* genomes is significantly higher than the overall composition (and, consequently, an under representation of G and C mononucleotide SSRs), however this difference can exclusively be attributed to noncoding regions, difference is significant with X2 test ($P < 0.0001$). Frequency of C mononucleotide SSRs in coding and noncoding regions of *S. sonnei* is more than *S. flexneri 301* and *2457T* and frequency of T mononucleotide SSRs in coding regions of *S. flexneri 2457T* is more than *S. flexneri 301* and *S. sonnei* (Figure l).

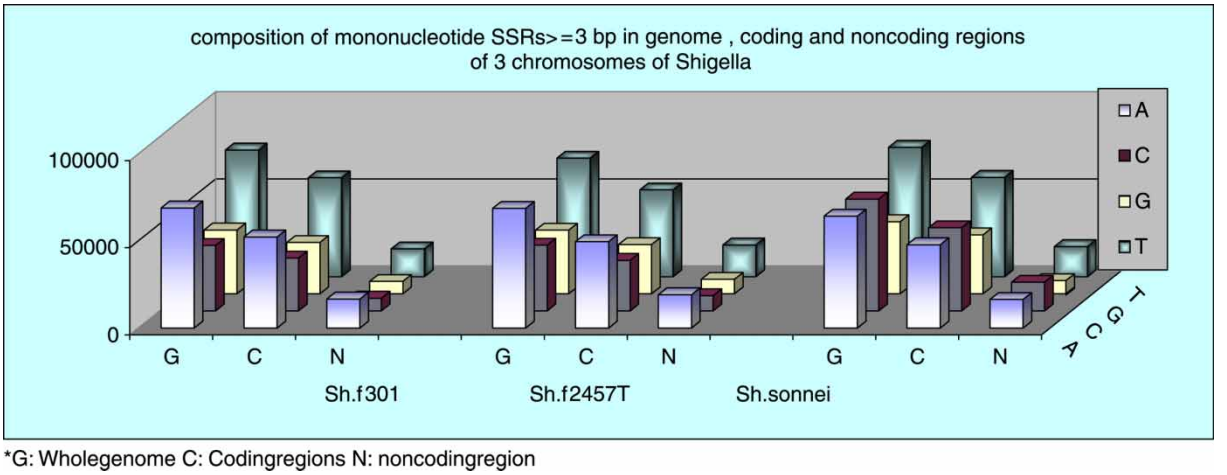
In the genome of *Sh.f 301*, *Sh.f2457T*, *S. sonnei*, *E. coli K12* and *S. saprophysticus* as repeat number became higher, frequency of A and T has increased. Difference is significant with X2 test ($P < 0.001$). But no such trend is observable for *M. tuberculosis* and *M. leprae* (Figure 2).

Frequency of dinucleotide SSRs

In all three genomes of *Shigella*, frequency of GC/CG in coding region is higher and frequency of AT/TA is lower. But frequency of GC/CG in *S. sonnei* is higher than two other strains of *S. flexneri*. Difference is significant with C2 test ($P < 0.01$). Frequency of AC/CA in noncoding region is higher in three *Shigella* genomes but in *S. sonnei* difference is more. Frequency of CT/TC in noncoding region is higher than coding regions in all chromosomes (Figure 3).

Frequency of trinucleotide SSRs:

The trinucleotide SSRs can be grouped into 10 motif subclasses, each representing six overlapping and complementary unit patterns. The total trinucleotide SSRs and groups number 9 and 10 including alanine and arginine are predominant in coding regions of *S. flexneri str 301* and *2457T* and *S. sonnei*. The trinucleotide SSRs Groups number 9 and 10 in coding and noncoding regions of chromosomes, are over



*G: Wholegenome C: Codingregions N: noncodingregion

Figure 1. Composition of mononucleotide SSRs in genome, coding regions and noncoding regions of three *Shigella* genomes.

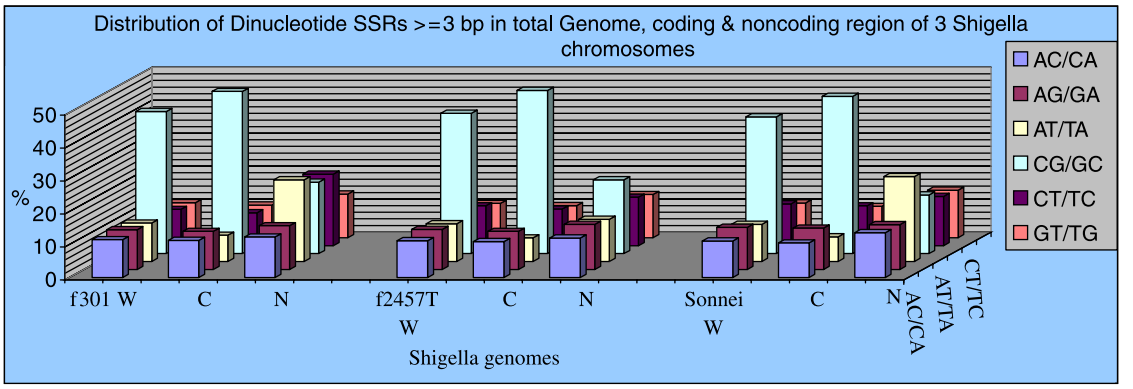


Figure 2. Frequency of dinucleotide SSRs.

represented, and group's number 5 and 6 are under represented in coding and noncoding regions of all *Shigella* chromosomes (Figure 4). There is significant difference between distribution of trinucleotide SSRs in coding region and noncoding region of both strains of *S. flexneri* by χ^2 test ($P < 0.001$).

Codon repetitions in complete genome sequences:

In *Sh.f 2a.str 2457T*, *Sh.f 2a.str 301*, *S. sonnei* and *E. coli K12* repetitions of Alanine (271, 287, 298, 318 time, respectively) are predominant, followed by Arginine (236, 220, 255, 246 time respectively), Glutamine (174, 173, 161, 163 time, respectively), leucine and valine. In *M. tuberculosis* Arginine repetitions (1310 time) are predominant, followed by Alanine (958 time), Valine (287 time) and Serine (235 time). In *M. leprae* Arginine repetitions (276 time) are predominant, followed by Alanine (268 time), Valine (177 time), Threonine (117 time) and Serine (104 time). In *S. saprophyticus*, Isoleucine repetitions (267 time) are predominant, followed by Tyrosine

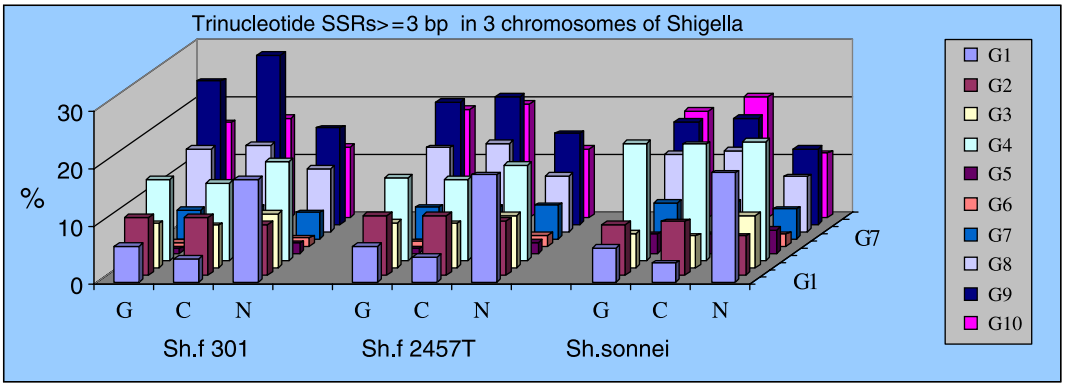


Figure 3. Frequency of dinucleotide SSRs.

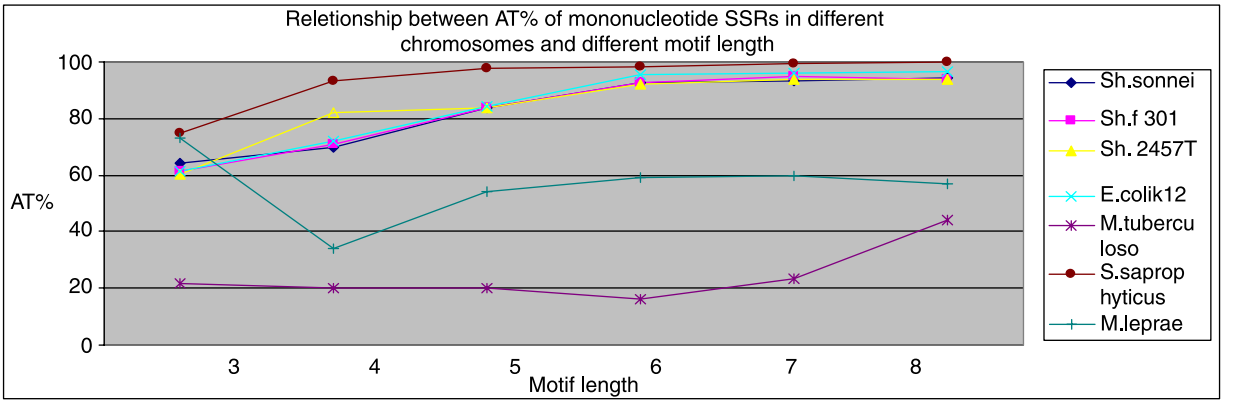


Figure 4. AT% of mononucleotide SSRs.

(133 time), Serine (96 time) and Leucine (66 time) (Table SVI).

Frequency of tetranucleotide SSRs

In *Sh.f2a StSOl* most frequency of tetra nucleotide SSRs are GCTG (5 time), TGGC (5 time), and CTGG (4 time). But in *Sh.f 2a str 2457T* most frequency of tetra nucleotide SSRs are CAGC(6 time), CCAG (5 time), TGGC (5 time), and CCGA (4 time) and most of them are in coding region. The tetranucleotide SSRs are predominant in coding regions of *S. sonnei* and *Sh.f 2a str 2457T* and noncoding regions of *Sh.f 2a str 301*. They have been found in both coding hydrophobic and hydrophilic aminoacids. They are presented in different genes with different function. Most of them are found in insertion sequences, putative proteins, hypothetical proteins and some genes with transposon-related functions and related to enzyme–tRNA synthetase (Table SIV).

Frequency of longer unit SSRs

Frequency of pentanucleotide, hexanucleotide, heptanucleotide, octanucleotide and nonanucleotide repeats are shown in Table II. Frequency of pentanucleotide, hexanucleotide and nonanucleotide repeats are more represented in *M. tuberculosis* than other investigated genomes. In *Sh.fstr301* and *2457T* the hexanucleotide SSRs are predominant in coding regions but in *S. sonnei* it is predominant in noncoding regions. From 14 larger unit SSRs in *Sh.f str301* only 2 SSRs and from 16 larger unit SSRs in *Sh.fstr2457T* only three SSRs are presented in intergenic regions. Most larger unit SSRs are presented in some ORFs that can have important product including:NADH dehydrogenase, DNA mismatch repair protein and ATP-dependent dsDNA exonuclease. Amino acid translated of longer unit tandem repeats are shown in (Table SV).

Discussion

Our data show that trinucleotide SSRs in coding regions of three investigated *Shigella* genomes are over represented, whereas dinucleotide and tetranucleotide repeats are underrepresented. It has been reported that triplet repeats show approximately twofold greater frequency in exonic regions than in intronic and intergenic regions in all human chromosomes except the Y chromosome (Subramanian et al. 2003). Such dominance of triplets over other repeats in coding regions may be explained on the basis of the suppression of non-trimeric SSRs in coding regions, possibly caused by frame shift mutations (Metzgar et al. 2000).

Frequency of codon repetitions in complete genome sequences of *Sh.f 2a.str 2457T* and *301*, *S. sonnei* and *E. coli K12* are approximately same. Codon repetitions are comparatively more numerous in *M. tuberculosis* (arginine 1310 time) than in other investigated genomes, (even *M. leprae* with arginine repetition 276 time) since the comparatively frequency of microsatellites is very low. High frequency of arginine and alanin in *M. tuberculosis* and *M. leprae* and low frequency of them in *S. saprophyticus* can be related to GC 11 content of the different codons, but over representation of these codon repetitions and some differences between frequency of codon repetition in *Shigella* and *E. coli* (with 50–51%GC) cannot be related to GC%.

Frequencies of codon repetitions are low in *S. saprophyticus* since the microsatellites is more frequent. While in all investigated genomes except *S. saprophyticus* arginine and alanine are predominant in *S. saprophyticus* isoleucine and tyrosine are predominant and arginine and alanine are very low abundant. Within a trinucleotide repeat class, frequencies of different codon repeats vary considerably depending on the type of encoded amino acid, while in *S. saprophyticus* the frequency of hydrophobic is higher than small/hydrophilic amino acids in other investigated genomes frequency of small/hydrophilic basic amino acids repetitions are more than hydrophobic amino acids, this might play an important role in the structure and function of the encoded proteins in these genomes.

It should be noted that frequencies of trinucleotide repeats in the chromosome sequences also include those occurring in the coding regions and could be partially limited by selection at the protein level (Mukund et al. 2001).

As mononucleotide repeat number became higher, representation of SSRs in noncoding regions has increased in 3 investigated *Shigella* genomes, which can be due to the fact that longer mononucleotide SSRs has more opportunity to undergo slipped-strand mispairing and there will be more mutability in their length than in shorter mononucleotide SSRs. This could help to explain why these are over represented in non-coding regions of the genome as selection has ample opportunity to operate against these larger repeats that would cause frame shift and nonsense mutations in coding regions (Coenye and Vandamme 2005).

DNA strand slippage can occur during transient dissociation and reannealing in the repeat region, and this could be a deceptive event for DNA processing machinery leading to expansions or deletions in the repeat tracks. It has been suggested that if the nucleotides on the single strand are self-complementary, they can base pair to form loops or hairpins and stabilize strand slippage (Gacy et al. 1995; Moore et al. 1999).

The upper limits for length of any given SSRs was higher in *Sh.f 301* (108 bp) and for mononucleotide

SSRs was higher in *S. sonnei* (29 bp). The upper limits for length of any given SSRs and mononucleotide SSRs in *S. saprophyticus* was lower (28 and 9, respectively). It has been proposed that these limit to repeat lengths are evidence for the fact that the increase of repeat length by mutations is counteracted by selection (through a mechanism acting on the length of the SSR sequence itself and/or through a mechanism acting on gene expression as affected by the SSR) (Gur-Arie et al. 2000). If this is true, our data suggest that, these mechanisms are less active in *Shigella* genomes than *S. saprophyticus*. The over representation of poly (A) and poly (T) mononucleotide repeats in all *Shigella* sp can be explained by the fact that strand separation for these poly (A) and poly (T) tracts is considerably easier than for poly (G) or poly(C) tracts, increasing the possibility of slipped strand mispairing. In this study in theree investigated *Shigella* genomes, *E. coli K12*, *M. tuberculosis* and *M. leprae* CG/GC dinucleotide SSRs are more frequent compared with other diucleotide repeats followed by GT/TG dinucleotide repeats, and AT/TA dinucleotide repeats are extremely rare. In *S. Saprophyticus* AT/TA are predominant followed by AC/CA and GT/TG dinucleotide repeats (AT reach genome).

It is evident that in human and Drosophila chromosomes, AC dinucleotide repeats are more frequent, followed by AT and AG repeats. In contrast, Arabidopsis chromosomes contain more AT repeats, followed by AG repeats. However, in the yeast genome, AT repeats seems to be predominant compared with other dinucleotide repeats. Interestingly, GC dinucleotide repeats are extremely rare in all of the eukaryotic genomes studied. Lower frequencies of CpG dinucleotides in vertebrate genomes have been attributed to methylation of cytosine, which, in turn, increases its chances of mutation to thymine by deamination (Schorderet and Gartler 1992).

However, it has been observed that similar to our study: TA is underrepresented in almost all prokaryotic genomes; which could be due to the fact that (i) TA forms the thermodynamically least stable DNA (allowing unwinding of the helix), (ii) RNases preferentially degrade UA dinucleotides in mRNA, and/or (iii) TA is part of many regulatory sequences. This may explain why TA/AT in dinucleotide SSRs is lower than GC/CG.

Our data indicate that when the GC content of the genome is high, the average and standard deviation of mononucleotide SSRs is lowest. This is confirmed by the observation of Coenye and Vandamme (2005). Who have shown that the GC content of mononucleotide SSRs is highest when the repeat density is lowest and repeat density is significantly higher in organisms with an intracellular or strictly parasitic lifestyle.

These observations suggest 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 C + G content, and it might be responsible for the marked differences observed in G + C content of these mononucleotide SSRs, as it would be too costly to have many poly (G) and/or poly(C) SSRs in genomes with a high density of mononucleotide SSRs (Coenye and Vandamme 2005).

While density of SSR in *E. coli* is more than *S. flexneri str 301*, *S. flexneri str 2457T* and *S. sonnei* it is very low in *M. tuberculosis* and there is similarity between distributions of SSR during the genome of these organisms in most of positions.

The observed similarities such as distribution of SSRs in the genome and representation of various types of SSRs indicate that investigated *Shigella* genomes and *E. coli K12* have shared a similar evolutionary history. Although there are some differences between investigated *Shigella* genomes and *E. coli K12* such as frequency of total SSRs, the upper limits for mononucleotide SSRs and any given SSRs, average of SSR density, frequency of mononucleotide SSRs and dinucleotide SSRs.

There are some differences between SSRs in protein coding regions of investigated *Shigella* genomes such as frequency of total, mononucleotide, dinucleotide and trinucleotide SSRs, composition of mononucleotide, dinucleotide and trinucleotide SSRs and distribution of tetranucleotide SSRs.

These variations be attributable to differences in gene expression and regulation of gene expression. Because 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. There are numerous lines of evidence indicating that changes in lengths of triplet or amino acid repeats could affect protein function, and frameshifts within coding regions caused by SSR expansion or contraction could (1) cause gain of function and loss of function or gene silencing; and (2) induce novel protein, bacterial pathogenesis, and virulence. This study also has shown some differences between distribution of SSR across coding and noncoding regions, and differential distributions of various repeats observed in different genome sequences suggest that apart from the nucleotide composition of repeats, the characteristic DNA replication/repair/recombination machinery might have an important role in the evolution of SSRs. For example the low frequency of dinucleotide and tetranucleotide repeats and the enhanced frequency of triplet repeats in the coding sequences of investigated organisms are signs of the effects of selection, indicating that those SSRs are selected against possible frameshift mutation. The action of selection on triplet repeats is best seen by considering their distribution between the different strands and reading frames within ORFs, and between coding and noncoding regions of the genome. Numerous

investigations have indicated that triplet repeats show strong reading frame and strand preferences (e.g. Richard and Dujon 1997; Alba et al. 1999) caused by selection. Strongly biased distributions of triplet repeats and amino acid repeats have also been found in different functional protein groups and cell locations.

Acknowledgements

The authors wish to express deep grateful to Prof. A.S. Kolaskar for contribution criticism and carefully reading of manuscript. One of the authors "Ashraf Hosseini" would sincerely thanks to Kulkarni–Kale Urmila and Fakher Rahim of Bioinformatics centre for training and making available the facility at Bioinformatics Centre, and Prof. Dilip Deobagker for his helpful discussion and comments. The work was supported by grant of Institute of Bioinformatics and Biotechnology, University of Pune.

References

Alba MM, Santiba'n ~ ez-Koref MF, Hancock JM. 1999. Amino acid reiterations in yeast are over represented in particular classes of proteins and show evidence of a slippage-like mutational process. J Mol Evol 49:789–797.

Belkum AV, Scherer S, Alphen OV, Verbrugh H. 1998. 2172/98 Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev 1092.

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci USA 94:1041–1046.

Coenye T, Vandamme P. 2003. Simple sequence repeats and compositional bias in the bipartite *Ralstonia solanacearum* GMI1000 genome. BMC Genomics 4:10.

Coenye T, Vandamme P. 2004. Abundance, distribution and composition of simple sequence repeats in the genomes of *ε-Proteobacteria*: Implications for genome diversity of *Helicobacter pylori*, Google automatically html

Coenye T, Vandamme P. 2005. Characterization of mononucleotide repeats in sequenced prokaryotic genomes. DNA Res 12:221–233.

Cox R, Mirkin S. 1997. Characteristic enrichment of DNA repeats in different genomes. Proc Natl Acad Sci USA 94:5237–5242.

Field D, Wills C. 1996. Proc R Soc Lond 263:209–215.

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 mutation pressures and a variety of selective forces. Proc Natl Acad Sci USA 95:1647–1652.

Gacy AM, Goellner G, Juranic N, Macura S, McMurray CT. 1995. Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. Cell 81:533–540.

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. Genet Res 10:62–71.

Hale T. 1991. Genetic basis of virulence in Shigella species. Microbiol Rev 55:206–224.

Jeffreys AJ, Wilson V, Thein SL. 1985. Hypervariable "minisatellite" regions in human DNA. Nature 314:67–73, 88,1D6.

Jeffreys AJ, Wilson V, Thein SL, Weatheral DJ, Ponder BAJ. 1986. DNA fingerprints and analysis of multiple markers in human pedigrees. Am J Hum Genet 39:11–24.

Kashi Y, King DG, Soller M. 1997. Simple sequence repeats as a source of quantitative genetic variation. Trends Genet 13:74–78.

King DG, Soller M. 1999. Variation and fidelity: The evolution of simple sequence repeats as functional elements in adjustable genes. In: Wasser SP, editor. Evolutionary theory and processes: Modern perspectives., p 65–82.

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221.

Metzgar D, Bytof J, Wills C. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10:72–80.

Moore H, Greenwell PW, Liu CP, Arnheim NT, Petes D. 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. Proc Natl Acad Sci USA 96:1504–1509, [Abstract/Free Full Textl].

Moxon E, Rainey P, Nowak M, Lenski R. 1994. Curr Biol 4:24–33.

Mukund V. Katti, Prabhakar K. Ranjekar, Vidya S. Gupta. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167.

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

Schlötterer C. 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma 109:365–371.

Schlötterer C, Wiehe T. 1999. Microsatellites, a neutral marker to infer selective sweeps. In: Goldstein DB, Schlötterer C, editors. Microsatellites: Evolution and applications., p 238–247.

Schloötterer 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.

Schorderet DF, Gartler SM. 1992. Analysis of CpG suppression in methylated and nonmethylated species. Proc Natl Acad Sci USA 89:957–961.

Subramanian S, Mishra RK, Singh L. 2003. Genome wide analysis of microsatellite repeats in humans: Their abundance and density in specific genomic regions. Genome Biol 4:R13.

Wren JD, Forgacs E, Fondon J, et al. 2000. Repeat polymorphisms within gene regions: Phenotypic and evolutionary implications. Am J Hum Genet 67:345–356.

Weber JL, Wong C. 1993. Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128.

Wierdl M, Dominska M, Petes TD. 1997. Instability in yeast: Dependence on the length of the microsatellite. Genetics 146:769–779.

Yang J, Wang J, Chen L, Yu J, Dong J, Yao Z, Shen Y, Jin Q, Chen R. 2003. Identification and characterisation of simple sequence repeats in the genomes of *Shigella* species. Gene 322:85–92.

Table SI. Distribution of simple sequence repeats among coding and noncoding regions of the *Sh.f 2a.str 301* chromosome.

		Coding regions		Noncoding regions	
Repeat unit	Total No.	No.	%	No.	%
Mono	212,907	158,878	74.6	54,029	25.4
3 bp	152,361	116,657	76.6	35,704	23.4
4 bp	41,556	29,818	71.8	11,738	18.2
5 bp	13,472	9105	67.6	4367	32.4
6 bp	4200	2674	63.7	1526	36.3
7 bp	1106	546	49.4	560	50.6
8 bp	180	64	35.6	116	64.4
9 bp	26	11	42.3	15	57.7
≥ 10 bp	6	3	50.0	3	50.0
Di	7054	5265	74.6	1789	25.4
3 bp	6569	4924	74.9	1674	25.1
4 bp	431	326	75.6	105	24.4
≥ 5 bp	25	15	60.0	10	40.0
Tri	2199	1854	84.3	354	15.7
3 bp	2145	1800	83.9	345	16.1
≥ 4 bp	54	45	83.3	9	16.7
Tetra	49	36	73.9	13	26.1
Penta	2	0	0	2	100.0
Hexa	9	9	100.0	0.0	0.0
3 bp	6	6	85.7	1	14.3
≥ 4 bp	3	3	100.0	0	0.0
Genome partition			*80.0*		*20%*

Table SII. Distribution of simple sequence repeats among coding and noncoding regions of the *Sh.f 2a. str 2457T* chromosome.

		Coding regions		Non-coding regions	
Repeat unit	Total no	No.	%	No.	%
Mono	218,707	169,035	77.3	49,672	22.7
3 bp	157,238	124,006	78.9	33,232	21.1
4 bp	42,441	31,830	75.0	10,611	25.0
5 bp	13,613	9734	71.5	3879	28.5
6 bp	4071	2759	67.8	1312	32.2
7 bp	1142	619	55.1	505	44.9
8 bp	188	75	39.9	113	60.1
9 bp	24	10	41.7	14	58.3
≥ 10 bp	8	2	25.0	6	75.0
Di	7004	5454	77.3	1550	22.7
3 bp	6553	5097	77.6	1456	22.4
4 bp	426	342	79.3	84	20.7
≥ 5 bp	25	15	60.0	10	40.0
Tri	2200	1846	83.9	354	16.1
3 bp	2144	1796	83.7	348	16.36
≥ 4 bp	56	50	92.6	6	7.4
Tetra	49	39	79.6	10	20.4
Penta	2	1	50.0	1	50.0
Hexa	10	8	88.8	1	11.2
3 bp	8	7	87.5	1	12.5
≥ 4 bp	2	2	100.0	0	0.0
Genome partition			*77.0*		*23%*

Table SIII. Distribution of simple sequence repeats among coding and noncoding regions of the *S. sonnei* chromosome.

		Coding regions		Non-coding regions	
Repeat unit	Total no	NO.	%	NO.	%
Mono	222,843	173,628	77.9	49,215	22.1
3 bp	158,475	125,172	79.0	33,303	21.0
4 bp	45,633	34,491	75.6	11,142	24.4
5 bp	13,178	10,247	77.8	2931	22.2
6 bp	4481	3104	69.3	1377	30.7
7 bp	848	527	62.1	321	37.9
8 bp	201	81	40.3	120	59.7
9 bp	25	6	24.0	19	76.0
≥10 bp	2	0	0.0	2	100.0
Di	7610	6027	79.2	1583	20.8
3 bp	7112	5647	79.4	1465	20.6
4 bp	469	360	79.4	109	20.6
≥5 bp	29	20	69.0	9	31.0
Tri	1845	1592	86.3	253	13.7
3 bp	1789	1548	86.4	241	13.6
≥4 bp	56	44	78.6	12	21.4
Tetra	35	30	85.7	5	14.3
Penta	0	0	0.0	0	0.0
Hexa	9	6	66.7	3	33.3
Hepta	3	0	0.0	3	100.0
Octa	1	0	0.0	1	100.0
Genome partition			*80.8*		*20%*

Table SIV. Distribution of tetranucleotide SSRs (3 time repeated) in coding regions of *S. flexneri str301 and 257T.*

ORF	Gene position in the genome Sh.f301	Length	Sh.f2457T	Length	Product	Repeat unit		Position of SSR in the genome sh.f 301	Position of SSR in the genome sh.f 2457T	Position of SSR in the gene of sh.f 301	Position of SSR in the gene of sh.f **2457T**
YhbN	3,335,549–3,336,106	557 bp	–		Hypothetical protein	AAAC	3 times	3,335,562	–	13	–
YrbK	–	575 bp	3,326,122–3,326,697		Hypothetical protein	AAAC	3 times	–	3326,659	–	537
FrwC	C 4,157,597–4,158,676	1079 bp	3,614,542–3,615,621	1079 bp	PTS system, fructose-like enzyme II component	AATT	3 times	4,158,667	3,614,540	1070	−2
SlyX	3,446,918..3,447,136	218 bp	–		Hypothetical protein	AGCC	3 times	3,447,079	–	161	
YfhH	2,683,559..2,684,479	920 bp	–		Hypothetical protein	AGTC	3 times	2,683,768	–	202	
YfhG	–		C 2,667,251..2,667,964	713 bp	Putative alpha helix protein	AGTC	3 times	–	2,677,336	–	1085
SF2615	2,690,826..2,691,236	412 bp	–		IS2 OrfA	CAGC	3 times	2,690,998		170	–
S1578	–		1,538,659..1,539,069	510 bp	IS2OrfA protein	CAGC	3 times	–	1,538,813	–	154
MurA	C 3,327,416..3,328,675	6551 bp	C 3,318,533..3,319,792	1259 bp	UDP–*N*-acetylglucosamine	CAGC	3 times	3,327,537	3,318,634	121	101
S2787	–	–	2,684,272..2,684,682	390 bp	IS2 OrfA protein	CAGC	3 times	–	2,684,425	–	153
S3760	–	–	3,655,595..3,656,005	410 bp	IS2 OrfA protein	CAGC	3 times	–	3,655,747	–	152
YhhF	–	–	C 4,198,492..4,199,088	596 bp	Hypothetical protein	CAGC	3 times	–	4,198,624	–	132
S1969	–	–	C 1,906,298..1,907,566	1268 bp	Putative tail fiber protein	CAGT	3 times	–	1,906,637	–	339
YhiL	–	–	4,174,918..4,176,108	1190 bp	Pseudo	CATA	3 times	–	4,175,676	–	758
YihP	–	–	3,693,359..3,694,754	1395 bp	Pseudo	CCAA	3 times	–	3,694,022	–	663
NarX	C 1,271,634..1,273,430	1796 bp	C 1,271,349..1,273,145	1796 bp	Nitrate/nitrite sensor protein	CCAG	3 times	1,273,331	1,273,031	1697	1682
GlyA	C 2,669,126..2,670,379	1253 bp	C 2,662,713..2,663,966	1253 bp	Serine hydroxymethyltransferase	CCAG	3 times	2,669,228	2,662,796	102	83
LepB	C 2,701,550..2,702,524	974 bp	C 2,694,994..2,695,968	974 bp	Leader peptidase	CCAG	3 times	2,702,344	2,695,769	794	775
GlyS	C 3,703,991..3,706,060	2069 bp	–		Glycyl-tRNA synthetase beta subunit	CCAG	3 times	3,704,942	–	951	–
YjeA	4,487,030..4,488,007	977 bp	4,479,223..4,480,200	977 bp	Lysyl-tRNA synthetase	CCAG	3 times	4,487,678	4,479,851	648	628
BglJ	4,571,170..4,571,697	522 bp	4,563,323..4,563,850	527 bp	2-Component transcriptional regulator	CGAC	3 times	4,571,432	4,563,564	162	241
Alas	C 2,792,881..2,795,511	2630 bp	2,787,515..2,790,145	2630 bp	Alanyl-tRNA synthetase	CGCA	3 times	2,795,302	2,789,917	2421	2402
S1350	1,397,033..1,397,398	365 bp	–		Hypothetical protein	CTGG	3 times	1,397,186	–	156	–
RnfD	1,685,860..1,686,918	1058 bp	1,725,498..1,726,556	1058 bp	Electron transport complex protein RnfD	CTGG	3 times	1,686,540	1,726,160	680	662
YhcI	C 3,355,571..3,356,479	908 bp	3,346,653..3,347,561	908 bp	*N*-acetylmannosamine kinase	CTGG	3 times	3,356,342	3,347,404	771	751
HpaI	C 4,551,428..4,552,231	803 bp	4,543,581..4,544,384	803 bp	2,4-Dihydroxyhept-2-ene-1,7-dioic acid aldolase	CTGG	3 times	4,552,033	4,544,165	605	584
S1360	1,403,317..1,404,585	1268 bp	–		Putative tail fiber protein	GACT	3 times	1,404,227	–	910	–
RbsB	3,945,759..3,946,649	890 bp	–		Similar to *E. coli* K12 D-ribose periplasmic binding protein /pseudo	GCCA	3 times	3,946,308	–	549	–
YieL	–	–	C 3,926,012..3,927,187	1175 bp	Putative xylanase	GCCA	3 times	–	3,926,272	–	260
YhjC	–	–	C 4,130,019..4,130,990	971 bp	Putative transcriptional regulator LYSR-type	GCCA	3 times	–	4,130,026	–	7
CcmF	C 2,314,962..2,316,905	1943 bp	C 2,294,419..2,296,362	1943 bp	Cytochrome *c*-type biogenesis	GCCC	3 times	2,315,057	2,294,495	95	76
S1580	C1,611,148..1,611,393	245 bp	–		IS2 ORF1	GCTG	3 times	1,611,381	–	233	–

Table SIV – *continued*

ORF	Gene position in the genome Sh.f301	Length	Sh.f2457T	Length	Product	Repeat unit		Position of SSR in the genome sh.f 301	Position of SSR in the genome sh.f 2457T	Position of SSR in the gene of sh.f 301	Position of SSR in the gene of sh.f **2457T**
HisC	2,098,736..2,099,806	1070 bp	2,081,649..2,082,719	1070 bp	Histidinol–phosphate aminotransferase	GCTG	3 times	2,099,327	2,082,221	591	572
S2694	C 2,770,551..2,770,961	410 bp	C 1,735,555..1,735,965	410 bp	Insertion sequence 2 OrfA protein	GCTG	3 times	2,770,784	1,735,770	233	215
YgiA	3,172,732..3,172,992	260 bp	–		Hypothetical protein	GCTG	3 times	3,172,981	–	249	–
S4097	C 4,256,114..4,256,524	410 bp	–		Insertion sequence 2 OrfA protein	GCTG	3 times	4,256,347	–	233	–
S1508	1,540,750..1,541,475	725 bp	–		Hypothetical protein	GGCG	3 times	1,541,016	–	266	–
YhhF	3,574,915..3,575,511	596 bp	–		Hypothetical protein	GGCT	3 times	3,575,353	–	438	–
FhuF	C 4,571,735..4,572,523	788 bp	–		Hypothetical protein	GTTG	3 times	4,572,279	–	544	–
S2880	–	–	C 2,765,187..2,765,597	410 bp	IS, phage, Tn; Transposon-related	GCTG	3 times	–	2,765,401	–	214
YgiB	–	–	3,161,510..316,221	704 bp	Hypothetical protein	GCTG	3 times	–	3,161,624	–	114
RbsB	–	–	C 3,826,645..3,827,535	890 bp	Ribose high-affinity ABC transporter permease component	GCTG	3 times	–	3,826,959	–	314
GlyS	–	–	4,069,158..4,071,227	2069 bp	Glycyl-tRNA synthetase beta subunit	GCTG	3 times	–	4,070,250	–	1092
YicF	–	–	3,983,771..3,985,453	1682 bp	DNA ligase	GGAA	3 times	–	3,983,899	–	128
S1625	–	–	1,579,126..1,579,851	725 bp	Hypothetical protein	GGCG	3 times	–	1,579,374	–	248
SlyX	–	–	4,325,568..4,325,786	218 bp	Hypothetical protein	GGCT	3 times	–	4,325,600	–	32
FhuF	–	–	C 4,563,888..4,564,676	788 bp	Hypothetical protein	GTTG	3 times	–	4,564,411	–	523
S3505	C 3,597,496..3,599,271	1775 bp	–	–	Hypothetical protein	TATG	3 times	3,598,302	–	796	–
WcaJ	C 2,125,692..2,127,086	1394 bp	2,108,558..2,109,952	1394 bp	Putative colanic acid biosynthsis UDP-glucose lipid carrier transferase	TCGC	3 times	2,127,059	2,109,906	1367	1348
LysR	2,939,382..2,940,317	935 bp	–	–	Positive regulator for lysine	TGCC	3 times	2,939,678	–	296	–
SodC	C 1,702,236..1,702,772	536 bp	1,741,826..1,742,362	536 bp	Superoxide dismutase precurs	TGGC	3 times	1,702,529	1,742,101		
YaeG	174,104..175,278	1174 bp	–	–	Similar to *E. coli* O157:H7 regulator protein of D-galactarate, /pseudo	TGGC	3 times	174,132	–	28	–
YaeG	–	–	173,609..174,765	1156 bp	S0157"/pseudo	TGGC	3 times	–	173,620	–	11
DsdA	2,482,265..2,483,593	1328 bp	2,460,404..2,461,732	1328 bp	D-serine dehydratase	TGGC	3 times	2,483,383	2,461,503	1118	1099
Ppx	2,613,299..2,614,840	1541 bp	2,591,432..2,592,973	1541 bp	Exopolyphosphatase	TGGC	3 times	2,614,788	2,592,902	1489	1470
YbaN	427,413..427,790	377 bp	427,214..427,591	377 bp	Hypothetical protein	TGGC	3 times	427,448	427,234	35	20
YdgB	1,655,613..1,656,335	722 bp	1,695,251..1,695,973	722 bp	Short chain dehydrogenase	TTAC	3 times	1,656,269	1,695,889	656	638
YhjC	3,642,502..3,643,473	941 bp	–	–	Putative transcriptional regulator LYSR-type	TGGC	3 times	3,643,441	1,736,20	939	–
YieL	3,848,072..3,849,247	1175 bp	–	–	Putative xylanase	TGGC	3 times	3,848,962	–	890	–
YicF	C 3,789,804..3,791,486	1682 bp	–	–	DNA ligase	TTCC	3 times	3,791,333	–	1529	–
S3949	C 4,078,619..4,079,872	1253 bp	–	–	Putative permease	TTGG	3 times	4,079,172	–	553	–
WcaJ	–	–	C 2,108,558..2,109,952	1394 bp	Putative colanic acid biosynthsis UDP-glucose	TCGC	3 times	–	210,9906	–	1348
LysR	–	–	2,934,063..2,934,998	935 bp	Positive regulator for lysine	TGCC	3 times	–	293,4340	–	277

Table SV. Distribution of longer unit SSRs in coding regions of *S. flexneri str301 and 2457T.*

ORF *S. flexneri str301*	Gene position in the genome	Length	Product	Repeat unit	No of repeat	Position of SSR in Genome	Gene	A.A translated of tandem repeats
HcaD	2,656,900..2,658,122	1222 bp	Similar to *E. coli K12* ferredoxin reductase	CGCAG	6 times	2,657,393	493	Not translated
NfrB	C 484,845..487,082	2237 bp	Bacteriophage N4 receptor	ACGCG	3 times	485,684	839	PRRVA
YidR	3,883,444..3,884,718	1274 bp	Hypothetical protein	ACCAAT	4 times	388,3522	78	TNTNTNTNTNTNT
SbcC	C 342,118..345,261	3143 bp	ATP-dependent dsDNA exonuclease	AGCGCC	3 times	342,374	256	LALALAL
LldP	3,749,997..3,751,652	1655 bp	L-lactate permease	CACTGG	3 times	3,751,288	1291	LALALA
HycC	2,815,733..2,817,550	1817 bp	NADH dehydrogenase subunit N	CCCAGC	8 times	2,816,652	919	GLGLGLGLGLGLGL
RplW	C 3,435,766..3,436,068	302 bp	50S ribosomal protein L23	CGACTT	3 times	3,435,900	134	EVEVEV
HyfF	2,594,974..2,596,560	1586 bp	NADH dehydrogenase subunit N	GCGCTG	3 times	2,596,147	1173	ALALALA
MutL	4,502,231..4,504,078	1847 bp	DNA mismatch repair protein	GCTGGC	3 times	4,502,434	203	LALALA
HemX	C 3,996,216..3,997,433	1217 bp	Uroporphyrinogen III methylase	GGTGCA	8 times	3,996,259	43	PAPAPAPAPAPAPAPA
PpiA	C 3,460,935..3,461,516	591 bp	Peptidyl–prolyl *cis–trans* isomerase A (rotamaseA"	AGAAAGAGC	3 times	3,461,454	519	ALSALSALS
PpdC	C 2,922,766..2,923,116	350 bp	Prepilin peptidase dependent protein C"	CACCATCAG	4 times	2,923,021	255	MVLMVLMVLMVL
S. flexneri str2457T								
NfrB	C 484,648..486,885	2237 bp	Point mutation/pseudo	ACGCG	3 times	485,472	824	Not translated
HemX	3,775,878..3,777,077	1199 bp	Uroporphyrinogen III methylase	ACCTGC	5 times	3,776,988	1110	PAPAPAPAPA
SbcC	C 341,252..344,395	3140 bp	ATP-dependent dsDNA exonuclease	AGCGCC	3 times	341,509	257	LALALAL
HycC	C 2,810,368..2,812,161	1793 bp	NADH dehydrogenase subunit N	CCCAGC	4 times	2,811,268	900	LGLGLGLG
RplW	4,336,637..4,336,939	302 bp	50S ribosomal protein L23	GAAGTC	3 times	433,6773	136	EVEVEV
LldP	C 4,023,565..4,025,220	1655 bp	L-lactate permease	GCCAGT	3 times	4,023,897	332	LALALAL
HyfF	2,573,107..2,574,693	1586 bp	NADH dehydrogenase subunit N	GCGCTG	3 times	257,4261	1154	AL ALALA
MutL	4,494,390..4,496,237	1847 bp	DNA mismatch repair protein	GCTGGC	3 times	4,494,573	183	LALALA
YidR	C 3,889,875..3,891,149	1274 bp	Hypothetical protein	GTATTG	4 times	3,891,032	1157	TNTNTNTNTNTNT
S0917	C 903,683..904,321	638 bp	Putative bacteriophage protein	TCTTCC	3 times	903,885	202	EEEEE
PpdC	C 2,917,457..2,917,807	350 bp	Prepilin peptidase dependent protein C	CACCATCAG	4 times	2,917,693	236	VLMVLMVLMVLMV
YnfL	C 1,682,957..1,683,868	911 bp	Putative LYSR-type transcriptional regulator	GCCAGTTCA	3 times	1,683,267	310	LAELAELAELA
PpiA	4,311,188..4,311,769	581 bp	Peptidyl–prolyl *cis–trans* isomerase (rotamaseA)	AGCTCTTTCT	3 times	4,311,210	22	ALSALSALS

Table SVI. Codon repetition in the different investigated genomes.

	Amino acid	Hydrophobicity	Codon	Sh.f2457T	Sh.sonnei	Sh.301	K12	*M. leprae*	mt	staph
1	Threonine	Hydrophilic	ACA/ACC/ACG/ACT	85	93	90	91	117	216	55
2	Isoleucine	Hydrophobic	ATA/ATC/ATT/TTA/TTG	110	139	128	156	37	18	**267**
3	Methionine	Hydrophobic	ATG	47	50	39	60	14	16	52
4	Glutamine	Hydrophilic	CAA/CAG	174	161	173	163	72	130	55
5	Arginine	Hydrophilic basic	CGA/CGC/CGG/CGT	236	255	220	246	**276**	**1310**	4
6	Leucine	Hydrophobic	CTA/CTC/CTG/CTT	112	119	129	134	55	128	66
7	Glutamate	Hydrophilic	GAA/GAG	61	62	65	68	39	71	41
8	Aspartate	Hydrophilic	GAC/GAT	57	71	74	68	58	93	27
9	Alanine	Hydrophobic	GCA/GCC/GCG/GCT	**271**	**298**	**287**	**318**	268	958	25
10	Valine	Hydrophobic	GTA/GTC/GTG/GTT	116	107	121	131	177	287	50
11	Tyrosine	Hydrophobic	TAT/TAC	26	14	25	30	6	1	133
12	Serine	Hydrophilic	TCA/TCC/TCG/TCT	132	150	111	127	104	235	96
13	Cysteine	Hydrophilic	TGC/TGT	72	64	77	76	76	128	43
14	Tryptophan	Hydrophobic	TGG	73	66	77	85	67	91	18
	Total			1572	1657			1385	3882	1021

Table SVII. Codon repetitions of predominant amino acids in different investigated chromosomes.

Amino acid	Codon	Sh.f 2457T	Sh.sonnei	*M. leprae*	mt	K12	staph	Sh.301
Arginine	CGA	9	16	80	224	10	1	13
	CGC	127	138	75	522	135		106
	CGG	71	73	44	454	80	1	69
	CGT	29	28	49	110	21	1	22
Alanine	GCA	64	69	57	92	70	19	65
	GCC	55	57	63	388	61	1	53
	GCG	109	116	98	383	130	2	119
	GCT	43	56	50	95	57	2	50
Isolucine	ATA	21	32	10	0	30	80	23
	ATC	58	68	11	13	70	91	65
	ATT	31	35	4	0	42	98	36
	TTA	0	0	0	0	0	0	0
	TTG	0	0	0	0	0	0	0
Tyrosine	TAC	3	2	4	1	3	5	2
	TAT	23	16	4	0	27	131	23

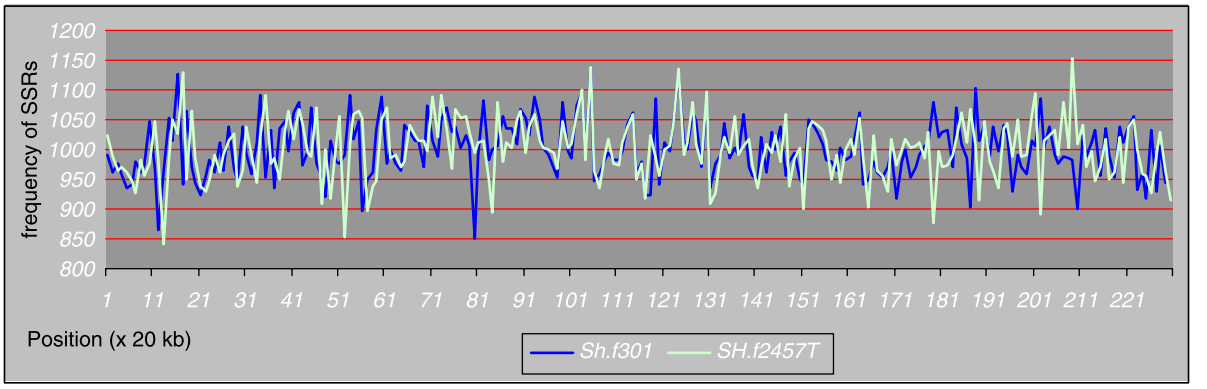


Figure S1. Distribution of SSRs in the genomes of *S. flexneri 2a strain 301* and *2457T.*