FEBS Letters 585 (2011) 1072–1076

FEBS Letters

journal homepage: www.FEBSLetters.org

Compound microsatellites in complete *Escherichia coli* genomes

Ming Chen [a,b], Guangming Zeng [a,b,*], Zhongyang Tan [c,*], Min Jiang [a,b], Jiachao Zhang [a,b], Chang Zhang [a,b], Lunhui Lu [a,b], Yuzhen Lin [a,b], Jun Peng [c]

[a] College of Environmental Science and Engineering, Hunan University, Changsha 410082, China
[b] Key Laboratory of Environmental Biology and Pollution Control, Hunan University, Ministry of Education, Changsha 410082, China
[c] College of Biology, State Key Laboratary for Chemo/Biosensing and Chemometrics, Hunan University, Changsha 410082, China

ARTICLE INFO

Article history:
Received 4 January 2011
Revised 28 February 2011
Accepted 2 March 2011
Available online 4 March 2011

Edited by Takashi Gojobori

Keywords:
Comparative genomics
Compound microsatellite
Microsatellite
Simple sequence repeat
Escherichia coli

ABSTRACT

Compound microsatellites consisting of two or more repeats in close proximity have been found in eukaryotic genomes. So far such compound microsatellites have not been investigated in any prokaryotic genomes. We have therefore examined compound microsatellites in 22 complete genomes of *Escherichia coli*, which is one of the ideal model organisms to analyze the nature and evolution of prokaryotic compound microsatellites. Our results indicated that about 1.75–2.85% of all microsatellites could be accounted as compound microsatellites with very low complexity, and most compound microsatellites were composed of very different motifs. Compound microsatellites were significantly overrepresented in all surveyed genomes. These results were dramatically different from those in eukaryotes. We discussed the possible reasons for the observed divergence.

1. Introduction

Microsatellites are repetitive DNAs composed of tandemly repeated short motifs of 1–6 base pairs (bp) [1,2]. As the whole genome sequence data of many organisms are now available and various analytical tools of microsatellites are developed, microsatellites have attracted a lot of attention with respect to their origin, distribution, roles and evolution [3]. Strand-slippage theories are frequently used to explain microsatellite distribution but alone insufficient, and other possible factors such as the interplay of the repeat type, the genomic position of the microsatellite and the genetic-biochemical background of the cell are thought to contribute to the observed divergence of microsatellite distribution in different organisms [4]. There are evidences that microsatellites play important roles in gene regulation, transcription and protein function [5,6]. Because of high polymorphisms, microsatellites have been used as markers for studies of population genetic and linkage association [7,8]. Microsatellites are ubiquitous in eukaryotic [4] and prokaryotic [9] genomes but relatively rare in Human Immunodeficiency Virus Type 1 (HIV-1) genomes [1]. It has been suggested that the difference of repeat frequencies between coding and non-coding regions arises from increased selection in coding regions [10,11]. Interestingly, microsatellites are more abundant in coding regions than in non-coding regions in eukaryotes [4,12], but on the contrary microsatellites are more predominant in coding regions in some prokaryotes [13,14]. Microsatellites are variable in length and hence may affect local DNA structure or the encoded proteins [9]. Main features of microsatellites include non-random distribution, high polymorphisms and diversity [3].

Over the past years, the distribution, evolution and roles of microsatellites in eukaryotic and prokaryotic genomes have been well documented, but compound microsatellites are still rarely reported. Compound microsatellites are composed of two or more microsatellites being found directly adjacent to each other. The study associated with compound microsatellites could provide a very good insight into the imperfection and evolution of microsatellites [15]. Some compound repeats such as $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ have been observed in human genome, and these repeats exhibit high polymorphisms [16]. It has been estimated that ~10% of microsatellites can be categorized as compound microsatellites in human genome [16]. A detailed analysis of a compound dinucleotide repeat $(CG)_m-(CA)_n$ in a marker *D18S58* also showed both repeats varied in length [17]. A survey of compound

Abbreviations: *E. coli*, *Escherichia coli*; HIV-1, Human Immunodeficiency Virus Type 1

* Corresponding authors. Address: College of Environmental Science and Engineering, Hunan University, Changsha 410082, China (G. Zeng).
E-mail addresses: zgming@hnu.cn (G. Zeng), zhongyang@hnu.cn (Z. Tan).

0014-5793/$36.00 © 2011 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2011.03.005

microsatellites in eight eukaryotic genomes showed that about 4–25% of all microsatellites had a composite motif [15]. Similar studies are needed for prokaryotic genomes to investigate whether the distribution of compound microsatellites in prokaryotic genomes is in agreement with that of eukaryotic genomes. Microsatellites have been extensively studied in *Escherichia coli (E. coli)* genomes [13], suggesting compound microsatellites may be present in *E. coli* genomes. Thus, the prokaryotic model organism *E. coli* is an excellent system and ideal model organism to study the distribution rules, possible roles and evolution of prokaryotic compound microsatellites. In this study, we sampled from available complete *E. coli* genome sequences, and then we screened these sequences for the presence, location, frequency and density of compound microsatellites in both coding and non-coding regions. Our results indicate about 1.75–2.85% of all microsatellites can be classified as compound microsatellites, and the compound microsatellite distribution of *E. coli* is dramatically different from that of eukaryote.

2. Materials and methods

2.1. Genome sequences

E. coli is a Gram-negative rod-shaped bacterium consisting of various strains and serotypes which are harmless or pathogenic [18]. Each strain has its unique characteristics at the molecular level [19]. In the present study, we analyzed compound microsatellites in 22 complete *E. coli* genomes, ranging from 4 639 675 bp (NC_000913) to 5 572 075 bp (NC_011353). All these genomes have been listed in 'Genome-level Extraction Mode' of IMEx [20], and were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/). The features of these *E. coli* genomes including size, GC content and coding density are present in Table 1.

2.2. Compound microsatellites

To aid our systematic analysis of compound microsatellites, we used a program called IMEx [20]. Some of previous studies focused on microsatellites with lengths of 12 bp or more [4], and we also used this value in our analysis. We detected microsatellites and compound microsatellites using the 'Genome-level Extraction Mode' of IMEx. The parameters were set as follows: Type of Repeat: imperfect; Repeat Size: all; Minimum Repeat Number: 12, 6, 4, 3, 3, 3; Max. distance allowed between any two SSRs (dMAX): 10. Other parameters were default. No compound microsatellites were standardized. This is because we want to observe real composition and evolutionary dynamics of compound microsatellites.

2.3. Analysis of compound microsatellite representation

Compound microsatellite representation was measured as the ratio of the observed number of compound microsatellites to the expected number of compound microsatellites (Obs/Exp). The expected number of compound microsatellites was calculated using the formula given by Kofler et al. [15].

$$C_{\exp}(m_1, m_2) = 2 * \mathrm{dMAX} * N_{m_1} * N_{m_2}/(G - N_m * L_m)$$

where $C_{\exp}(m_1, m_2)$ is the expected number of compound microsatellites having the motif m_1–m_2 (m_1 is non-identical to m_2) in a genome of length G, dMAX is the maximum distance between adjacent microsatellites constituting the compound microsatellite (bp), L_m is the average length of a microsatellite (bp), N_{m_1}, N_{m_2} and N_m are the numbers of m_1, m_2 and all microsatellites, respectively. We tested the significance of compound microsatellites based on a both sided Poisson Distribution, by calculating cumulative probability $P(X \geqslant C_{\exp}(m_1, m_2))$ [15].

3. Results and discussion

When detecting compound microsatellites in a genome, the dMAX is the most influential parameter [15]. Therefore, it is necessary to determine the impact of dMAX on the identification of compound microsatellites. We assessed this impact in three different *E. coli* strains (*E. coli* 536, *E. coli* 55 989 and *E. coli APEC O1*) using the percentage of individual microsatellites being part of a compound microsatellite (cSSR-%). Our results indicated the cSSR-% dramatically increased with the dMAX, but this increase was not completely linear (Supplementary Fig. 1). Thus, it is very difficult to select an optimal dMAX for the identification of compound

Table 1
List of analyzed *Escherichia coli* genomes and overrepresentation of compound microsatellites.

No.	Acc. no.	Size (bp)	GC content (%)	Coding density (%)	N[a]	Obs/exp[b]	P[c]
S1	NC_008253	4 938 920	50	87	31	660	0
S2	NC_011748	5 154 862	50	86	36	444	0
S3	NC_008563	5 082 025	50	86	32	696	0
S4	NC_010468	4 746 218	50	86	28	509	0
S5	NC_004431	5 231 428	50	87	34	576	0
S6	NC_009801	4 979 619	50	85	33	418	0
S7	NC_011745	5 209 548	50	85	34	607	0
S8	NC_009800	4 643 538	50	86	31	596	0
S9	NC_011741	4 700 560	50	87	32	360	0
S10	NC_011750	5 132 068	50	86	37	712	0
S11	NC_011601	4 965 553	50	85	34	642	0
S12	NC_002655	5 528 445	50	87	40	500	0
S13	NC_011353	5 572 075	50	83	40	506	0
S14	NC_002695	5 498 450	50	85	41	513	0
S15	NC_011742	5 032 268	50	87	30	667	0
S16	NC_011415	4 887 515	50	88	31	492	0
S17	NC_010498	5 068 389	50	87	40	588	0
S18	NC_011751	5 202 090	50	87	49	476	0
S19	NC_007946	5 065 741	50	88	28	667	0
S20	NC_010473	4 686 137	50	83	34	340	0
S21	NC_000913	4 639 675	50	85	34	362	0
S22	AC_000091	4 646 332	50	86	35	398	0

[a] Number of compound microsatellites.
[b] Observed number of compound microsatellites/expected number of compound microsatellites.
[c] Significance of obs/exp on the basis of a Poisson distribution (see Section 2).

microsatellites. In the present study, we examined the distribution of compound microsatellites with dMAX being set to 10 bp. This threshold value has been used for studying compound microsatellites in eight eukaryotes, and was thought to be able to provide the maximum sensitivity for identification of compound microsatellites by allowing for mismatches in microsatellite-search [15].

3.1. Occurrence of compound microsatellites

We analyzed compound microsatellites in 22 different *E. coli* genomes. As a result, 28–49 compound microsatellites were found in each of surveyed genomes (Table 1). *E. coli UTI89* (NC_007946) and *E. coli ATCC 8739* (NC_010468) had the lowest number of compound microsatellites, whereas *E. coli UMN026* (NC_011751) had the highest. Compound microsatellites were 340–712-fold overrepresented in complete *E. coli* genomes, which suggested that they were not likely to emerge by chance. To allow comparison among genome sequences of different sizes, we normalized the total numbers for all compound microsatellites as percentage or number of repeats per Mb of sequence (compound microsatellite density). A more-or-less similar density of microsatellites (644.6–675.7 microsatellites/Mb) and compound microsatellites (5.5–9.4 compound microsatellites/Mb) was observed in analyzed *E. coli* genomes, respectively (Supplementary Table 1; Fig. 1A), regardless of whether these genomes belonged to different strains or substrains. The complete genome of *E. coli UMN026* (NC_011751) had the highest compound microsatellite density (9.4 compound microsatellites/Mb) and *E. coli UTI89* (NC_007946) showed the lowest (5.5 compound microsatellites/Mb). In an attempt to analyze the distribution of compound microsatellites more clearly, we characterized the difference between the occurrences of compound microsatellites in coding and non-coding regions. Our analysis indicated that compound microsatellites were richer in coding regions than in non-coding regions (Fig. 1B). A higher number of compound microsatellites observed in coding sequences than in non-coding sequences may be attributed to very high coding densities of *E. coli* genomes (Table 1). The highest cSSR-% was found in NC_011751 (2.85%), followed by the NC_010498 (2.43%), and the lowest was in NC_010468 (1.75%) (Fig. 1C).

Despite overall similarity of compound microsatellite distribution (number, cSSR-% and density), the completely same distribution pattern was not observed between any two surveyed *E. coli* genomes. These differences might be caused by the parameters 'genome size' and 'microsatellite density'. It appears that the longer genomes contain more microsatellites than do the shorter genomes [21]. In contrast to this, microsatellite frequency is negatively correlated with genome size in plants [11,22]. However, more and more evidences indicate that the total microsatellite contents are not directly proportional to the genome sizes in many organisms [1,12]. Over the past years much has been learned about the relationship between genome size and distribution of microsatellites, but still very little is known about the association of genome size with distribution of compound microsatellites. To test whether the genome size had a significant influence on number and density of compound microsatellites, linear regression was carried out. Number of compound microsatellites ($R^2 = 0.294$, $P < 0.05$) weakly correlated with the genome size. Weaker correlation was observed between density of compound microsatellites and genome size ($R^2 = 0.023$, $P < 0.05$) in analyzed *E. coli* genomes. The parameter 'microsatellite density' had a weakly positive influence on number ($R^2 = 0.128$, $P > 0.05$) and density ($R^2 = 0.214$, $P < 0.01$) of compound microsatellites. However, microsatellite density had a significant influence on compound microsatellite density in eukaryotes [15]. Weaker association of microsatellite density with the distribution of compound microsatellites in *E. coli* genomes than in eukaryotic genomes may be due to the fact that these two types of organisms have different DNA polymerases and replication methods [15].

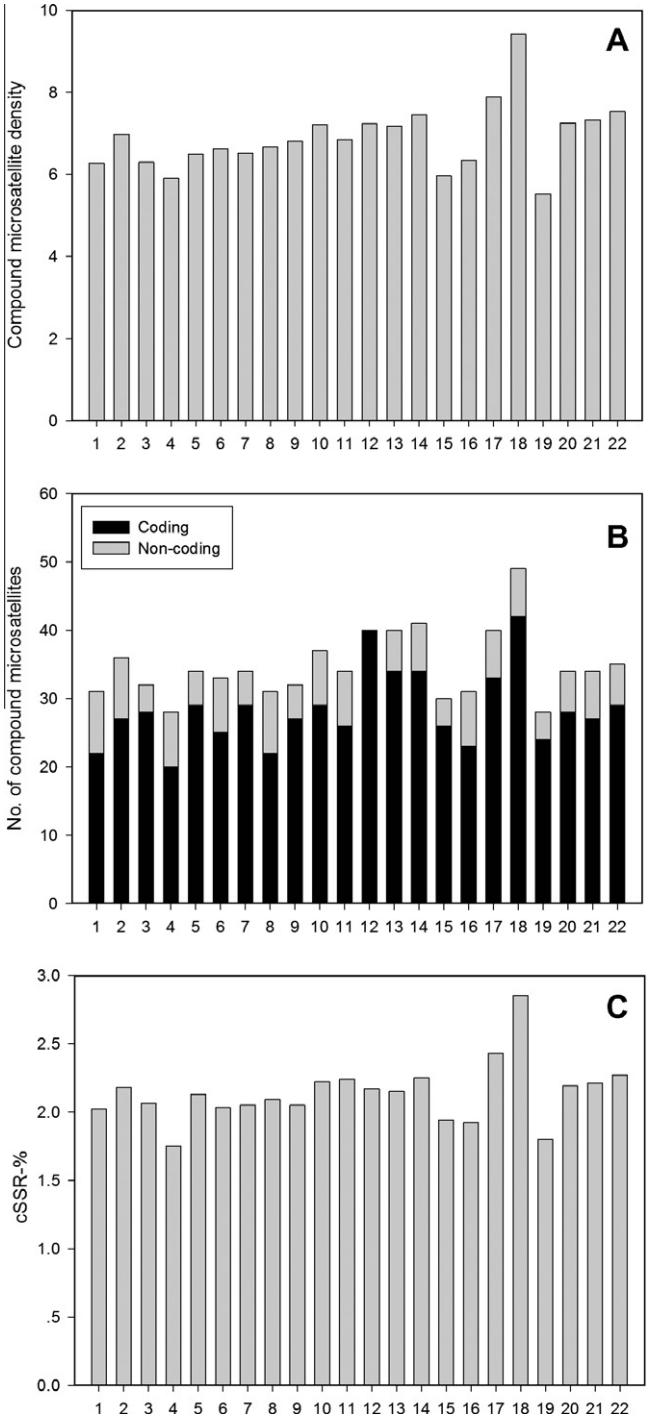

Fig. 1. Distribution of compound microsatellites in analyzed *Escherichia coli* genomes. (A) Compound microsatellite density (number of compound microsatellites/Mb). (B) Occurrence of compound microsatellites in coding regions and non-coding regions. (C) Percentage of individual microsatellites being part of a compound microsatellite (cSSR-%). *x*-Axis: S1–S22 complete *E. coli* genomes.

3.2. Compound microsatellite complexity

Compound microsatellites are composed of two or more adjacent individual microsatellites (cSSRs). Therefore, the

composition of compound microsatellites is indeed very complicated due to variable number of cSSRs. We term the $m_1-x_n-m_2$ '2-microsatellite' and the $m_1-x_n-m_2-x_t-m_3$ '3-microsatellite' compound microsatellite [15]. Analysis of compound microsatellite complexity indicated that all surveyed *E. coli* genomes were rich in '2-microsatellite' compound microsatellites, followed by '3-microsatellite' compound microsatellites. In general, the number of compound microsatellites became less and less with the increase of complexity in each surveyed complete genome (Table 2). No very large compound microsatellites were observed in *E. coli* genomes. The largest compound microsatellites were composed of three cSSRs, showing a lower complexity in *E. coli* than in eukaryotes [15]. To date, there are no available explanations for the lower complexity of compound microsatellites in *E. coli*. Genome size may be taken into account, due to the fact that larger genome generally contains relatively more microsatellites and thus may increase the frequency of adjacent microsatellites by chance. Furthermore, among *E. coli* genomes, most microsatellites occurred in the coding regions in which selection against framshift mutation limited expansion of microsatellites other than triplet repeats [11], and therefore there might be length and complexity constraints in relation to functional importance as compared to higher eukaryotes.

3.3. Motif and structure of compound microsatellites

Compound microsatellite is believed to mostly originate from imperfection in microsatellites in eukaryotes, and its individual microsatellite motifs exhibit very high similarity [15]. To assess whether the individual microsatellite motifs constituting the compound microsatellite in *E. coli* genomes were also very similar, we investigated the composition and structure of compound microsatellites (Supplementary Table 2). Among all 22 complete *E. coli* genomes, the most abundant compound microsatellite motifs were $(GGT)_4$-x_6-$(GCC)_4$, $(GCG)_4$-x_5-$(GGC)_4$, $(CCTGA)_3$-x_0-$(TGCA)_3$ and $(ATCC)_3$-x_9-$(AATG)_3$. CTG-CAG compound microsatellite composed of self complementary motifs has been proposed to be created by recombination [23]. However, our study showed no compound microsatellites contained self complementary motifs, suggesting these compound microsatellites were not likely to be derived from recombination in *E. coli* genomes. Most compound microsatellites were composed of very distinct motifs with two or more different bases. In sharp contrast to this, almost all compound microsatellites consisting of two cSSRs had very similar motifs in eukaryotes [15].

In conclusion, the knowledge about the biological significance of compound microsatellites is blank to date in prokaryotes. Comparative analysis of compound microsatellites in various *E. coli* genomes is possible to be helpful in understanding their origin and roles. Our study showed compound microsatellites were common features of diverse *E. coli* genomes. Our analyses also indicated that compound microsatellites were diverse in the form of motif and complexity. It has been demonstrated that microsatellite distribution of eukaryotes is significantly different from that of prokaryotes [4,9,14]. Likewise, we also observed dramatic differences among compound microsatellite distribution between *E. coli* and eukaryotes in three aspects [15]. First, in general, the number of compound microsatellites in *E. coli* genome was relatively less than that in eukaryotic genomes. This can be expected, since the genome sizes of eukaryotic genomes are generally larger than those of *E. coli*. Second, weaker association of microsatellite density with the distribution of compound microsatellites was observed in *E. coli* genomes than in eukaryotic genomes. Third, compound microsatellites showed a lower complexity in *E. coli* than in eukaryotes. Dramatic differences found in compound microsatellite distribution between *E. coli* and eukaryotes suggested that fundamental differences between these two types of organisms in the mechanisms of formation and fixation of compound microsatellites. We have speculated these differences might partly arise from their different genome features, mismatch repair systems and replication methods. Further studies related to compound microsatellites in whole prokaryotes are in progress to systematically and roundly reveal the nature and evolutionary dynamics of prokaryotic compound microsatellites.

Table 2
Compound microsatellite complexity in analyzed *Escherichia coli* genomes.

c.c.[a]			c.c.[a]			c.c.[a]		
No.	2	3	No.	2	3	No.	2	3
S1	28	3	S9	31	1	S17	38	2
S2	34	2	S10	35	2	S18	48	1
S3	28	4	S11	31	3	S19	25	3
S4	28	0	S12	39	1	S20	33	1
S5	32	2	S13	39	1	S21	33	1
S6	31	2	S14	40	1	S22	34	1
S7	32	2	S15	27	3			
S8	29	2	S16	30	1			

[a] Compound microsatellite complexity which indicates the number of cSSRs [15].

Acknowledgements

We thank editor and anonymous reviewers for their valuable comments. The study was financially supported by the National Natural Science Foundation of China (50608029, 50978088, 50808073, 51039001), Hunan Provincial Innovation Foundation for Postgraduate, the National Basic Research Program (973 Program) (2005CB724203), Program for Changjiang Scholars and Innovative Research Team in University (IRT0719), the Hunan Provincial Natural Science Foundation of China (10JJ7005), and the Hunan Key Scientific Research Project (2009FJ1010).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2011.03.005.

References

[1] Chen, M., Tan, Z., Jiang, J., Li, M., Chen, H., Shen, G. and Yu, R. (2009) Similar distribution of simple sequence repeats in diverse completed human immunodeficiency virus type 1 genomes. FEBS Lett. 583, 2959–2963.
[2] Chen, M., Tan, Z., Zeng, G. and Peng, J. (2010) Comprehensive analysis of simple sequence repeats in pre-miRNAs. Mol. Biol. Evol. 27, 2227–2232.
[3] Archak, S., Meduri, E., Kumar, P.S. and Nagaraju, J. (2007) InSatDb: a microsatellite database of fully sequenced insect genomes. Nucleic Acids Res. 35, D36–D39.
[4] Toth, G., Gaspari, Z. and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10, 967–981.
[5] Kashi, Y. and King, D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22, 253–259.
[6] Usdin, K. (2008) The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res. 18, 1011–1019.
[7] Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W. (2002) Genetic structure of human populations. Science 298, 2381–2385.
[8] Abdurakhmonov, I.Y. et al. (2005) Simple sequence repeat marker associated with a natural leaf defoliation trait in tetraploid cotton. J. Hered. 96, 644–653.
[9] Mrazek, J., Guo, X. and Shah, A. (2007) Simple sequence repeats in prokaryotic genomes. Proc. Natl. Acad. Sci. USA 104, 8472–8477.
[10] Metzgar, D., Bytof, J. and Wills, C. (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. 10, 72–80.
[11] Ellegren, H. (2004) Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. 5, 435–445.
[12] Karaoglu, H., Lee, C.M. and Meyer, W. (2005) Survey of simple sequence repeats in completed fungal genomes. Mol. Biol. Evol. 22, 639–649.

[13] Gur-Arie, R., Cohen, C.J., Eitan, Y., Shelef, L., Hallerman, E.M. and Kashi, Y. (2000) Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. Genome Res. 10, 62–71.
[14] Li, Y.C., Korol, A.B., Fahima, T. and Nevo, E. (2004) Microsatellites within genes: structure, function, and evolution. Mol. Biol. Evol. 21, 991–1007.
[15] Kofler, R., Schlotterer, C., Luschutzky, E. and Lelley, T. (2008) Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. BMC Genomics 9, 612.
[16] Weber, J.L. (1990) Informativeness of human (dC-dA)n·(dG-dT)n polymorphisms. Genomics 7, 524–530.
[17] Bull, L.N., Pabon-Pena, C.R. and Freimer, N.B. (1999) Compound microsatellite repeats: practical and theoretical features. Genome Res. 9, 830–838.
[18] Hudault, S., Guignot, J. and Servin, A.L. (2001) *Escherichia coli* strains colonising the gastrointestinal tract protect germfree mice against *Salmonella typhimurium* infection. Gut 49, 47–55.
[19] Grozdanov, L., Raasch, C., Schulze, J., Sonnenborn, U., Gottschalk, G., Hacker, J. and Dobrindt, U. (2004) Analysis of the genome structure of the nonpathogenic probiotic *Escherichia coli* strain Nissle 1917. J. Bacteriol. 186, 5432–5441.
[20] Mudunuri, S.B. and Nagarajaram, H.A. (2007) IMEx: imperfect microsatellite extractor. Bioinformatics 23, 1181–1187.
[21] Hancock, J.M. (2002) Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. Genetica 115, 93–103.
[22] Morgante, M., Hanafey, M. and Powell, W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat. Genet. 30, 194–200.
[23] Jakupciak, J.P. and Wells, R.D. (1999) Genetic instabilities in (CTG.CAG) repeats occur by recombination. J. Biol. Chem. 274, 23468–23479.