Available online at www.sciencedirect.com
ScienceDirect
Gene 410 (2008) 18–25
GENE
www.elsevier.com/locate/gene

# Distribution and evolution of short tandem repeats in closely related bacterial genomes

Edit Kassai-Jáger [a], Csaba Ortutay [b], Gábor Tóth [c], Tibor Vellai [a], Zoltán Gáspári [d,*]

[a] *Department of Genetics, Eötvös Loránd University, Pázmány Péter sétány 1/C, H-1117 Budapest, Hungary*
[b] *Institute of Medical Technology, FI-33014 University of Tampere, Tampere, Finland*
[c] *Bioinformatics Group, Agricultural Biotechnology Center, Szent-Györgyi Albert u. 4, H-2100 Gödöllő, Hungary*
[d] *Institute of Chemistry, Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest, Hungary*

Received 24 July 2007; received in revised form 8 November 2007; accepted 16 November 2007
Available online 28 November 2007
Received by J. Jurka

**Abstract**

Simultaneous identification and comparison of perfect and imperfect microsatellites within a genome is a valuable tool both to overcome the lack of a consensus definition of SSRs and to assess repeat history. Detailed analysis of the overall distribution of perfect and imperfect microsatellites in closely related bacterial taxa is expected to give new insight into the evolution of prokaryotic genomes. We have performed a genome-wide analysis of microsatellite distribution in four *Escherichia coli* and seven Chlamydial strains. Chlamydial strains generally have a higher density of SSRs and show greater intra-group differences of SSR distribution patterns than *E. coli* genomes. In most investigated genomes the distribution of the total lengths of matching perfect and imperfect trinucleotide repeats are highly similar, with the notable exception of *C. muridarum*. Closely related strains show more similar repeat distribution patterns than strains separated by a longer divergence time. The discrepancy between the preferred classes of perfect and imperfect repeats in *C. muridarum* implies accelerated evolution of SSRs in this particular strain. Our results suggest that microsatellites, although considerably less abundant than in eukaryotic genomes, may nevertheless play an important role in the evolution of prokaryotic genomes and several gene families.


*Keywords:* Genome evolution; Microsatellite; SSR; *Chlamydia*; *Escherichia*; Prokaryotes

## 1. Introduction

Microsatellites – simple sequence repeats (SSRs) – are of great practical and theoretical importance in eukaryotes (Ellegren 2004; Kashi and King, 2006). In prokaryotes, their abundance is relatively low (van Belkum et al., 1998; Eckert and Yan, 2000; Metzgar et al., 2001; Schlotterer et al., 2006; Mrazek et al., 2007), they nevertheless contribute to genome polymorphism in bacteria (Lindstedt, 2005). *Escherichia coli* O157:H7 VNTR repeats have been recently monitored (Noller et al., 2003). Since there is no consensus definition of microsatellites (Ellegren, 2004), it is not straightforward to compare SSRs identified in different studies. We have recently introduced a new approach, the separate identification and subsequent comparison of perfect and imperfect SSRs (Gáspári et al., 2007) to overcome this difficulty. Our approach is also expected to yield information about the history of the repeats if we assume that the majority of imperfect repeats containing a perfect core is a remnant of a longer perfect stretch. In this paper we apply our approach to related bacterial taxa to assess the intra- and inter-group similarities of genomic repeat distributions. Parallel investigation of related genomes using multiple SSR detection methods, combined with standardized SSR classification (Jurka and Pethiyagoda, 1995; Tóth et al.,

*Abbreviations:* SSR, simple sequence repeat; bp, base pair(s); Mbp, megabase(s) or 1 million bp.

* Corresponding author. Institute of Chemistry, Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest, Hungary. Tel.: +36 1 20090555x1408; fax: +36 1 3722620.
*E-mail address:* szpari@chem.elte.hu (Z. Gáspári).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.11.006

2000) is expected to yield a biologically relevant picture of the significance of SSRs in the bacterial strains under study.

The two bacterial groups selected for the present survey include *E. coli* and Chlamydial strains. Chlamydiales comprise a monophyletic group that is phylogenetically well separated from other bacterial taxa (Stephens et al., 1998; Kalman et al., 1999; Read et al., 2000, 2003; Shirai et al., 2000; Chen et al., 2007). Their genome evolution has recently been investigated by bioinformatic methods (Ortutay et al., 2003; McNally and Fares, 2007). These features make these genomes ideal for a comparative analysis.

We chose *E. coli* genomes as our other target group because these bacteria are among the most widely studied prokaryotes, with well-described genetics. The increasing number of *E. coli* strains with known genomes offers a unique opportunity to analyze SSR evolution in these very closely related bacteria (Blattner et al., 1997; Hayashi et al., 2001; Perna et al., 2001; Welch et al., 2002). We analyzed 4 *Escherichia* genomes used also in other studies (Azad and Lawrence, 2007; McNally and Fares, 2007).

## 2. Materials and methods

### 2.1. Genomes used for this study

Complete genome sequences for various strains of *E. coli*, *Chlamydia muridarum*, *C. trachomatis*, *Chlamydophila pneumoniae* and *C. caviae* were downloaded from NCBI GenBank (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The genome sequences used for this study are summarized in Table 1.

### 2.2. SSR extraction and classification

SSR extraction and classification was performed as described previously (Gáspári et al., 2007), using in-house programs and Tandem Repeats Finder (TRF; Benson, 1999). Repeats with 1–6 bp units and with a minimum length of 12 bp were considered. Repeat unit classes were standardized as described earlier (Jurka and Pethiyagoda, 1995; Tóth et al., 2000; Gáspári et al., 2007), e.g. the class '**acg**' represents all of its permutant and/or reverse complement sequences (**acg**=**cga**=**gac**=**cgt**=**gtc**=**tcg**). To identify imperfect repeats corresponding to perfect ones (i.e. to select perfect and imperfect repeats at identical loci within a selected genome), all imperfect repeats found around the location of each perfect repeat were selected. If there were multiple imperfect repeats matching the perfect one, priority was given to repeats with a repeat class identical to that of the perfect repeat. If no such imperfect repeat was found, the lengths of the repeated units were considered in a way that one of them must be a multiple of the other (e.g., a perfect repeat with unit length 6 can match an imperfect one with unit length 3). It is important to stress that in this context, 'matching' perfect and imperfect repeats are located in the same genome, and no systematic attempt was made to find homologous repeats in related genomes.

### 2.3. Data processing and evaluation

All data were stored in MySQL tables for subsequent analysis. Matching perfect and imperfect repeats were identified as being located at the same chromosomal position within a genome. Therefore, these repeats were identified by two independent methods. SSRs were assigned to coding or non-coding regions according to the 'CDS' records in the NCBI annotation. Orthologous genes in related strains were identified using the KEGG Database (http://www.genome.jp/kegg, Kanehisa et al., 2006). Gene sequences were aligned using ClustalW (Thompson et al., 1994). All other computations were performed using in-house PERL programs. Amino acid repeats encoded by trinucleotide SSRs were identified by mapping the repeat position onto the coding nucleotide and the corresponding translated protein sequences in the annotation of the GenBank files. Identical amino acids coded by at least 50% of the bases in the repeat sequence were included in the statistics. This was important to characterize imperfect repeats and also to account for the fact that repeat units may not coincide with codons (even a perfect repeat may code for more than one amino acid types).

### 2.4. Comparison of repeat distributions

To assess the differences between perfect and imperfect repeat distribution patterns and to compare the differences of

Table 1
Genomes used for this study

| Strain | Accession (GenBank) | GN | RefSeq identifier | Total length (bp) | Length of coding regions (bp) |
|---|---|---|---|---|---|
| *Chlamydia muridarum* Nigg | AE002160.2 | cmu | NC_002620 | 1,072,950 | 961,248 |
| *Chlamydia trachomatis* ser. D | AE001273.1 | ctr | NC_000117 | 1,042,519 | 936,164 |
| *Chlamydophila caviae* GPIC | AE015925.1 | cca | NC_003361 | 1,173,390 | 1,046,055 |
| *Chlamydophila pneumoniae* AR39 | AE002161.1 | cpa | NC_002179 | 1,229,853 | 1,090,813 |
| *Chlamydophila pneumoniae* CWL029 | AE001363.1 | cpn | NC_000922 | 1,230,230 | 1,085,960 |
| *Chlamydophila pneumoniae* J138 | BA000008.3 | cpj | NC_002491 | 1,226,565 | 1,097,297 |
| *Chlamydophila pneumoniae* TW-183 | AE009440.1 | cpt | NC_005043 | 1,225,935 | 1,102,622 |
| *Escherichia coli* K12 | U00096.2 | eco | NC_000913 | 4,639,675 | 4,048,916 |
| *Escherichia coli* O157:H7 | BA000007.2 | ecs | NC_002695 | 5,498,450 | 4,819,150 |
| *Escherichia coli* O157:H7 EDL933 | AE005174.2 | ece | NC_002655 | 5,528,445 | 4,820,481 |
| *Escherichia coli* CFT073 | AE014075.1 | ecc | NC_004431 | 5,231,428 | 4,600,495 |

GN: Genome identifier according to the KEGG Database.