Distribution of Repetitive Sequences on the Leading and Lagging Strands of the *Escherichia coli* Genome: Comparative Study of Long Direct Repeat (LDR) Sequences

Mitsuoki Kawano,1,† Shigehiko Kanaya,1,† Taku Oshima,2 Yasushi Masuda,2 Takeshi Ara,3 and Hirotada Mori1,3,*

Research and Education Center for Genetic Information, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101, Japan,1 CREST, JST (Japan Science and Technology),2 and Institute for Advanced Biosciences, Keio University, 14-1 Baba, Tsuruoka, Yamagata 997-0035, Japan3

(Received 8 November 2001; revised 23 January 2002)

Abstract

In the present study, we developed a method for detecting sequences whose similarity to a target sequence is statistically significant and we examined the distribution of these sequences in the *E. coli* K-12 genome. Target sequences examined are as follows: (i) short repeat: Crossover hot-spot instigator (Chi) sequence, replication termination (*Ter*) sequence, and DnaA binding sequence (DnaA box); (ii) potential stem-loop structure repeats: palindromic unit (PU), boxC sequences, and intergenic repeat unit (IRU); (iii) potential RNA coding repeats: rRNAs, PAIR, TRIP, and QUAD; and (iv) potential protein coding repeats: insertion elements (ISs) and Long Direct Repeats (LDRs). We also examined the distribution of these sequences on leading and lagging strands. We obtained another four statistically significant LDR sequences with more than 187 bp matched to LDR-A near the LDR loci, suggesting that these regions might be used as high recombination hot spots for LDR. Adaptation of individual LDRs to *E. coli* genome is also discussed on the basis of codon usage.

Key words: *Escherichia coli* K-12 (W3110); repetitive sequence; leading and lagging strand; codon usage; LDR

1. Introduction

The distribution of both large and small repetitive sequences in *E. coli* is an important source of genetic plasticity. In the case of structural genes, such as those for tRNAs and rRNAs, the advantages given by sequence multiplicity might include higher expression levels.1,2 Repetitive sequences such as those of rRNA genes may have been established early with respect to species divergence, and, in the case of tRNA genes, the arrangement would tend to be conserved among individuals throughout the species. The sequences may have become amplified by recent rearrangement (e.g., tandem duplications),3 and such arrangements can be highly specific for either the leading or the lagging strand.

Repetitive sequences derived from accessory elements such as insertion sequences (ISs) can vary considerably with regard to number and chromosomal position in individuals of the species.4 Illegitimate, site-specific, and homologous recombination all might have roles in the establishment of sequence repetition in different circumstances, and at least in the case of very small repetitive sequences convergent evolution should also be considered.5 Repetitive sequences can allow for recombination between dispersed homologous sequences, which can lead to chromosomal rearrangements such as duplication, deletion, inversion, and transposition. Clearly, genetic interaction between repetitive sequences is a major contributor to genome plasticity. The topic of repetitive sequences in *E. coli* has been reviewed extensively.5–11

The organization of genes around the replication origin of *E. coli* indicates that the number of genes transcribed away from *oriC* is much larger than the number of those transcribed toward it.12 In a recent analysis using genetic algorithms, it was proposed that if a mutator and a normal DNA polymerase existed in a bacterial cell during replication, that is, if there was unequal fidelity in the synthesis of the leading and lagging strands, the bacterial population could further expand its genetic diversity and preserve advantageous genotypes.13
The purpose of the present study was to characterize the distribution of repetitive sequences on the leading and lagging strands of the E. coli genome. Using standard statistics for global similarities, we examined the entire E. coli K-12 (W3110) genome for sequences similar to a target sequence. Using this method, we detected sequences whose similarity to repetitive sequences in the E. coli genome was statistically significant, and we examined the distribution of these sequences on the leading and lagging strands. The repetitive sequences examined are short genomic repeats such as Crossover hot-spot protein-coding repeats such as insertion elements (ISs) and Long Direct Repeats (LDRs).

2. Materials and Methods

2.1. Southern blot analysis

Genomic DNA from Escherichia coli W3110, Salmonella typhimurium LT2, Kluyvera ascorbata, Citrobacter freundii, and Leclercia adecarboxylata was kindly supplied by Dr. H. Kasai. Genomic DNA was digested with EcoRV, and the DNA was loaded into each lane of a 0.8% (w/v) agarose gel. After electrophoresis, DNA was blotted on a positively charged nylon membrane (Roche Diagnostics), and hybridized with the palindromic unit (PU; also called Repetitive Extra-genic Palindromic [REP] sequence),

intergenic repeat unit (IRU; also called ERIC sequence),

and the boxC sequence;

potential RNA coding repeats such as rRNAs, PAIR repeats, TRIP repeats, and QUAD repeats;

and potential protein-coding repeats such as insertion elements (ISs) and

2.2. DNA sequences

DNA sequences were obtained from GenoBase (version 3.0; http://ecoli.aiast-nara.ac.jp/) for E. coli K-12 (W3110), from GenBank web site (http://www.ncbi.nlm.nih.gov:80/) for E. coli O157, from the Washington University Consortium sequencing project web site (http://genome.wustl.edu/gsc/Projects/bacteria.shtml) for S. typhimurium LT2 and S. paratyphi A, from the Sanger Centre sequencing project web site (http://www.sanger.ac.uk/Projects/Syphi/) for S. typhi CT18, and from the University of Illinois sequencing project web site (http://www.salmonella.org/) for S. enteritidis, S. dublin, and S. pullorum.

2.3. Method for detecting statistically significantly similar sequences of a repetitive sequence

Score-based methods for comparing DNA sequences were developed and used for searching sequence databases. In the E. coli genome, the occurrences of four nucleotides (A, T, G, and C) are nearly identical, so we assumed that the probability of having identical nucleotides at any sites between two sequences is 1/4, independent of the four nucleotides. In the present study, we used binomial statistics to detect sequences in the E. coli genomes whose similarity to a target sequence was statistically significant. By scanning the E. coli genome sequence with an N$_i$-base window, we counted the number of k-base matched sequences to the ith target sequence. Here, N$_i$ represents the length in nucleotides (nt) of the ith target sequence. The frequency of k-base matched sequences is statistically estimated by Eq. (1).

$$ f(k) = \frac{\binom{N_i}{k}}{u^k v^{N_i-k}} $$

Here, u represents the probability of identical nucleotides at a position in the N$_i$-base window between sequences retrieved and targeted, and v ($= 1 - u$) represents the probability of different nucleotides at a position between them.

When N$_i$ is larger than 30, the binomial distribution f(k) denoted by Eq. (1) is approximately represented by a normal distribution with an average of N$_i$u and a variance of N$_i$uv. The statistical occurrence of k-base matched sequences to the target sequence can be estimated by the average (av[N$_i$]) and standard deviation (SD[N$_i$]) = $\sqrt{N_i w v}$. When border$_w[N_i]$ is denoted by av[N$_i$] + w SD[N$_i$], where w is an arbitrary coefficient, the statistically expected occurrence of sequences with nucleotides matched to a target sequence larger than border$_1[N_i]$ is 0.159 (the expected number in the E. coli genome = 4,639,221 × 0.159 for an individual strand), the expected occurrence of sequences with those larger than border$_2[N_i]$ is 0.023 (the expected number = 4,639,221 × 0.023), and so on. When the expected number of sequences larger than border$_w[N_i]$ in the E. coli genome is smaller than 1, if sequences with identical nucleotide number larger than border$_w[N_i]$ are detected, they are significantly similar to the target sequence. In the present study, we selected 5.2 for w, that is, the product of genome size and expectation occurrence (4,639,221 × 1.993 × 10$^{-7}$ = 0.92) is smaller than 1, and referred to sequences with more than border$_{w-1}[N_i]$ bases matched to the target sequence as statistically significantly similar sequences.


3. Results and Discussion

3.1. Distribution of short repeats on leading and lagging strands

To characterize the distribution of repeat elements and to detect sequences highly similar to those on the leading and lagging strands, we examined the occurrences of k-base matched sequences to a target (or consensus) sequence on those strands. The consensus sequences used in the present study are listed in Table 1. First, we examined repetitive sequences related to replication, such as the Chi sequence, DnaA box, and Ter element.

The oriC region shows remarkable symmetry. Genes tend to be transcribed divergently from oriC, and Chi octa-nucleotide (5′-GCTGGTGG-3′) recombination hot spots are oriented symmetrically with respect to oriC.21 Figure 1(a) shows a clear difference in the occurrences of the k-base matched sequences to the Chi sequence between the leading and lagging strands. Sequences that are highly homologous to Chi and Ter tend to be located on the leading strand and those highly homologous to DnaA box tend to be located on the lagging strand. As shown in Table 1, the number of G nucleotides is much larger than the number of C nucleotides in the Chi and Ter consensus sequences; in contrast, the number of G nucleotides is much smaller than the number of C nucleotides in the DnaA box. Asymmetrical nucleotide usage has been reported between the leading and lagging strands, and a high percentage of G nucleotides was observed on the leading and lagging strand in several bacteria.26,27 This is reflected in the distribution of the small repeats on the leading and lagging strands. Figure 1(d) and (e) show distribution of G and T octamers on leading and lagging strands. Biased usage of sequences highly similar to G octamers is much stronger than that of sequences highly similar to T octamers. The genome GC skew contributed strongly to the distribution bias of these elements between the leading and lagging strands, and is reflected in the distribution around the peak of \( N_i/4 \) (referred to as GC skew background). In Fig. 1, the bias of DNA sequences in the E. coli genome towards being in the same orientation as the direction of replication is associated with the G content of their sequence,28 and is mainly reflected in the distribution around the peak of \( N_i/4 \).

3.2. Distribution of sequences similar to PU, boxC, and IRU

Sequences similar to the repetitive elements PU, boxC, and IRU are shown in Fig. 2. Sequences whose similarity is statistically significant are separated from the GC skew background centered at \( N_i/4 \) are distributed (left panels in Fig. 2). PU is a component of BIME (Bacterial interspersed mosaic element), which is a class of highly repetitive sequences found initially in the genome of E. coli.29 BIME sequences appear to participate in seemingly disparate functions such as transcription,18,30–32 translation,33 and chromosome organization and stability, though it is still unknown whether BIMEs are essential for bacterial viability. BIMEs are specific binding sites of nucleotide-associated proteins, such as DNA gyrase, DNA Pol I, and IHF.34,35 Sequences whose similarity to PU consensus sequences is statistically significant are distributed separate from the GC skew background (Fig. 2a), and the number of such sequences with larger than 19 nt (= border\(_{5.2}\) [28]) is 71 on leading strand and 67 on lagging strand. This indicates that the orientation of PU is independent of the orientation of DNA replication.

The boxC sequences are thought to be mobile elements because they are located differently in various strains of
Figure 1. Distribution of k-matched sequences on the leading and lagging strands of the *E. coli* genome. (a) Chi, (b) DnaA box, (c) Ter, (d) G octamer, and (e) T octamer. “Diff.” represents the number of k-matched sequences on the leading strand minus that on the lagging strand. The x-axis represents the number of matched nucleotides to a target sequence, and the y-axis indicates the frequencies of individual matched sequences.

the same species. Kunisawa and Nakamura proposed the idea that they are also involved with regulating the stabilization of upstream mRNA. Fourteen sequences on the leading strand and 20 on the lagging strand were similar to the boxC consensus sequence with more than 28 nt (= border3.2[50]) (Fig. 2b). This indicates that the distribution of boxC sequences is biased weakly between the leading and lagging strands. A similar situation was also obtained for IRU (Fig. 2c). It has been proposed that IRU affects genome stability by promoting rearrangements or by providing hot spots for recombination. Interestingly, the distribution of similar sequences of boxC and IRU was biased between leading and lagging strands separated from the GC skew distribution. These biased distributions might contribute to the rearrangement difference between the leading and lagging strands.
Figure 2. Distribution of k-matched sequences on the leading and lagging strands of the *E. coli* genome. (a) PU, (b) boxC, and (c) IRU (see legend of Fig. 1).

Figure 3. Distribution of k-matched sequences on the leading and lagging strands of the *E. coli* genome. (a) IS1, (b) IS2, and (c) QUAD1c (see legend of Fig. 1). Arrows indicate statistically significantly matched sequences on the lagging strand.
3.3. Distribution of sequences similar to coding sequences of ISs and rRNAs

Several IS elements are present in multiple copies in E. coli. Approximately 60 insertion elements have been detected in E. coli K-12 (W3110) (see GenoBase 3.0). Eighteen insertion elements (8 for IS1, 5 for IS3, 3 for IS30, 1 for IS150 and 1 for IS911) possess ORFs with identical transcription orientation within ISs. Of these, 11 (61%) are located on the lagging strand ($p < 0.05$, $\chi^2$-test); that is, the transcription orientation and replication orientation of IS elements tend to be opposite. Figure 3 shows the distribution of IS1- and IS3-like sequences, which have much larger copy numbers than the other ISs mentioned above. There are more similar sequences located on the lagging strand than on the leading strand (see arrows in Figs. 3a, b). Transcription from QUAD, a potential RNA-coding repeat, was observed by DNA microarray analysis and Northern blot analysis. Repetitive sequences of QUADs may form stable stem-loop structures, and these sequences are distributed equally on leading and lagging strands (Fig. 3c).

In E. coli K-12, five rRNA operons are located clockwise from the origin of replication, and two are located counterclockwise. Each is oriented so that it is transcribed in the same direction as it is replicated. The advantage conferred by the number and the orientation of rRNA operons could simply be transcription level. Under conditions of rapid growth, rRNA expression comprises over half of the cell’s transcription activity. Figure 4 shows distributions of sequences similar to three rRNAs (rrfA, rrlA, rrsA corresponding to 5S, 23S, and 16S rRNAs). Statistically significantly similar sequences of rRNAs separated from the distribution of the GC skew were detected on the leading strands, but not on the lagging strands. All of these correspond to whole or partial sequences of rRNAs (9 for rrfA, 15 for rrlA, and 7 for rrsA). In general, highly expressed genes tend to be located on the leading strand, and the genes for all rRNAs are located on leading strand.

The number of 16S rRNAs is the same as that of 23S rRNAs in E. coli K-12, but the number of 5S rRNAs is different. We detected two similar sequences to 5S rRNA with more than 55 nt ($= \text{border}_{5.2}(120)$) using the nucleotide sequence of rrfA. All of these se-
sequences are also located on the leading strand (Fig. 4a) when their orientation is defined by the orientation of the rRNAs. The outside circle in Fig. 5 shows genome positions for the rrf genes and the two additional similar sequences. These two sequences correspond to repeats designated as TRIP1a and TRIP1c by Rudd (1999); TRIP1a may form a stable stem-loop structure.22 Thus, two sequences (TRIP1a, and TRIP1c) may have important roles associated with 5S rRNAs. The 5S rRNAs, including the two sequences, are distributed symmetrically on the leading strand at the center of oriC.

3.4. Characteristics of Long Direct Repeats in Escherichia coli K-12

Four copies of LDRs have been found in the E. coli genome.27,38 Three of them have repetitive units of 530 bp (LDR-A, LDR-B, and LDR-C) and exist at the 27 min region of the chromosome as tandem repeats without any spacer sequence. Another copy, with 450 bp units (LDR-D), exists at the 80 min region (see the inside circle of Fig. 5; GenoBase 3.0). We obtained four statistically significantly similar sequences with more than 187 bases matched to LDR-A (corresponding to border52[535]), located in both strands (see arrows in Fig. 6). Three of them are located near LDR-D, and one is located near LDR-A, B, and C. These two regions might be used as high recombination spots for LDR.

Using the gene finding program, GeneMark,39 and a genetic analysis, Kawano et al. detected two transcripts with opposite orientation in the LDR-D region and predicted a small ORF with 35 amino acids in each of the repeating units (unpublished results). The small antisense RNA was located in the strand complementary to the ldrD mRNA leader region. A similar genetic structure was found in the hok/sok system in plasmid R1, which functions for the stable maintenance of low copy plasmids by post-segregational killing (PSK).40 The small ORFs for several bacteria are shown in Fig. 7. The alignment of 18 LDRs for E. coli K-12, E. coli O157:H7, S. typhimurium LT2, S. typhi CT18, S. paratyphi A, S. dublin, S. enteritidis and S. pullorum indicates that these LDRs are very highly conserved.

We examined codon usage heterogeneity for ldr genes by Z1-parameter analysis based on the parameters of E. coli K-12 genes, which could be applicable for codon usage heterogeneity in Salmonella genes because codon usage heterogeneity in E. coli is fundamentally identical to that in S. typhimurium.41,42 A large, positive Z1 value indicates high adaptation of codon usage of the E. coli translation system, and a negative Z1 represents low adaptation.2 The ldra, B, and C genes of E. coli K-12 and O157 have positive Z1 values (ID = 1–3 in Fig. 7), whereas the ldrD genes of E. coli K-12 and O157, and the ldr genes for Salmonella have negative Z1 values (ID = 4–9 in Fig. 7). These data support the idea that the former set is adapted to the host genome, but the latter is not. The ldra genes for all Salmonella analyzed have the smallest Z1 and these genes are located at the 9 min region in Salmonella genomes but not at the identical site in E. coli genomes. This suggests that
Distribution of Repetitive Sequences in E. coli

Figure 6. Distribution of k-matched sequences on the leading and lagging strands of the E. coli genome. (a) LDR-A, and (b) LDR-D (see legend of Fig. 1). Arrows show the border of a statistically significantly matched sequence determined by statistics (see Materials and Methods).

![Figure 6](image)

Figure 7. Base conservation of 18 ldr genes for several species based on their relative positions and the genetic structure of LDR-D. ID 1, ldrB for E. coli K12, ldrB homologs E. coli O157, and ldrC homologs for E. coli O157; ID 2, ldrA homologs for E. coli O157; ID 3, ldrA and ldrC for E. coli K12; ID 4, ldrD homologs for S. typhi CT18; ID 5, ldrD homologs for S. dublin and S. typhimurium LT2; ID 6, ldrD homologs for S. paratyphus A; ID 7, ldrD for E. coli K12; ID 8, ldrD homologs for E. coli O157; ID 9, ldrA homologs for S. typhi CT18; S. dublin, S. typhimurium LT2, S. enteridis, and S. paratyphus A, and S. pullorum. The underlined codon corresponds to the optimal translation codon for E. coli K-12. Base conservation is represented as the ratio of the largest occurrence among the four types of nucleotides to the total number at individual positions.

![Figure 7](image)

these repetitive units underwent different evolutionary pressures during the long history of bacterial evolution. The LDR families were also detected in the genomes of Salmonella typhimurium LT2, Kluyvera ascorbata, Citrobacter freundii and Leclercia adecarboxylata by Southern blot analysis using an LDR-D probe (Fig. 8). It should be stressed that LDR-homologous sequences are highly conserved and distributed in various enteric bacteria, and therefore might have an important physiological role.

Acknowledgements: We wish to acknowledge gratefully Dr. Hiroaki Kasai for genomic DNAs. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology of Japan and a grant from CREST and ACT, JST (Japan Science and Technology).
Figure 8. Southern blot analysis of EcoRV-restricted genomic DNA from *Escherichia coli* W3110 (Ec), *Salmonella typhimurium* LT2 (St), *Klebsiella ascorbata* (Ka), *Citrobacter freundii* (Cf), and *Leclercia adecarboxylata* (La). The DIG-labeled LDR-D RNA probe was used at 37°C.

References


35. Boccard, F. and Prentki, P. 1993, Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3' end of transcription units, *EMBO J.*, 12, 5019–5027.


