Cell, Vol. 90, 205–206, July 25, 1997, Copyright ©1997 by Cell Press

Letter to the Editor

The Recombination Hot Spot Chi Is Embedded within Islands of Preferred DNA Pairing Sequences in the E. coli Genome

Genetic recombination is the exchange of homologous DNA. Although it occurs anywhere along chromosomes, certain loci display elevated frequencies of exchange (hot spots). In E. coli, χ (5′-GCTGGTGG-3′) is a hot spot (Lam et al., 1974; Smith et al., 1981). Stimulation of recombination requires RecBCD and is to the 5′-side of χ (Ennis et al., 1987; Myers et al., 1995). These characteristics arise from the regulatory nature of χ: it attenuates the 3′→5′ nuclease, but not the helicase, activity and activates a weaker 5′→3′ nuclease activity of RecBCD (Dixon and Kowalczykowski, 1991; Anderson and Kowalczykowski, 1997a). These modifications create ssDNA with χ at its 3′-terminus.

In addition, χ is a preferred DNA pairing sequence for RecA. GT-rich sequences, of which χ is a subset, bind preferentially to RecA and show enhanced homologous pairing (Dixon and Kowalczykowski, 1995; Tracy and Kowalczykowski, 1996). Finally, χ-containing ssDNA is a target for loading of RecA by the translocating RecBCD (Anderson and Kowalczykowski, 1997b). Thus, χ is central to recombination, coordinating the activities of RecA and RecBCD. The importance of χ to E. coli is illustrated by the large number present in its genome; 1009 copies are present (between 142 and 262 are expected depending on whether the distribution of mono-, di-, or trinucleotides is considered), making it one of the most overrepresented motifs (Cardon et al., 1993).

The pivotal role of this 8 nucleotide sequence in recombination is even more surprising given its simplicity; however, it is unclear whether genomic context around χ potentiates its effectiveness. Therefore, we aligned the sequences flanking all 1009 χ sites and searched for statistically significant biases. Table 1 shows that the 50 bases flanking χ deviate from the overall base composition of the E. coli genome: A residues decrease to 22.3%, and G residues increase to 27.9%. The deviations from the mean for A and G (−2.32% and +2.53%, respectively) are significant since they fall outside of the 99% confidence intervals (Table 1). Remarkably, when the analysis was extended to 250 positions on either side of χ, the bias towards G and away from A was still apparent (Table 1). When 500 positions were examined using a sliding window of 30 positions, the average composition for both A and G was returning to its genomic mean (not shown). Linear extrapolation of this average suggests that the nucleotide bias will extend to about 400 bp on either side of χ (not shown), encompassing about 17% of the genome.

In addition to the compositional bias, a striking pattern of nucleotide occurrence is revealed (Figure 1). As expected, A is underrepresented within this region, with only one position enriched for A; the most significant underrepresentation of A occurs as a contiguous block of 4 to the immediate left of χ. For C, no positions deviate significantly, with the exception of those close to χ, and substantial underrepresentation at 4 of the 5 positions to the 3′-side of χ. The G residues show an interesting repeated pattern: almost every third G is significantly underrepresented and is followed by an overrepresented pair of G's. The T residues also have an unusual pattern, with nearly every third position significantly overrepresented and the next two underrepresented. Thus, though T is not enriched on average, its periodic overrepresentation every third position is unusual. When extended to 230 positions surrounding χ, the patterns of over- and underrepresentation remain the same (not shown). Thus, (1) A is avoided to each side of χ; (2) C is distributed randomly, except very close to χ; (3) G is overrepresented, displaying a characteristic GG repeat every 3 nucleotides; and (4) T is overrepresented at every third position, alternating with the GG dinucleotide repeats. Thus, the distribution of G and T creates a repeating pattern of the trinucleotides: TGG, GGT, or GTG, which extends for at least 230 bp flanking χ (not shown). Moreover, positions adjacent to χ display the most significant biases, suggesting that χ could be larger than the canonical 8-base sequence: 5′-C(/T)GCTGGTGGC(/T)GG-3′.

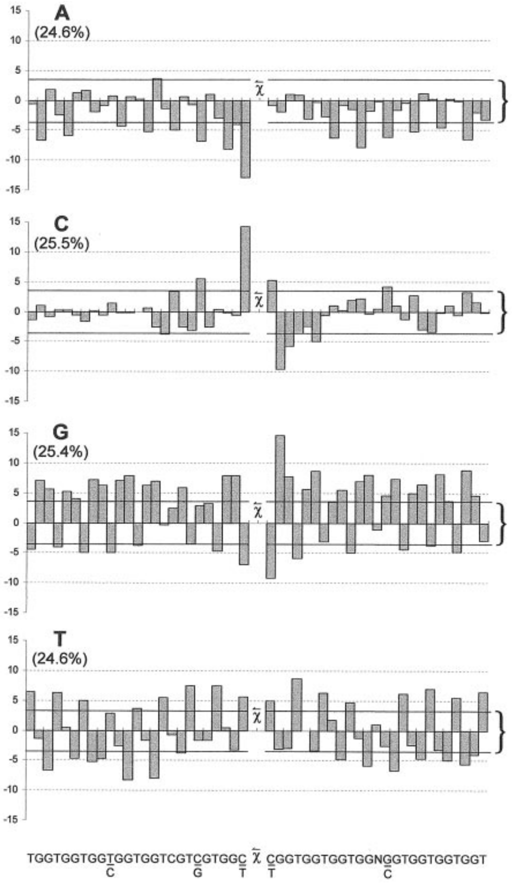

Figure 1. Deviations from the Genomic Mean of A, C, G, and T at 50 Positions Surrounding the 1009 Aligned χ Sites

Arrows indicate the active orientation of χ. The 99% confidence interval is shown. The most overrepresented bases at each position are indicated at the bottom. Deviations are in percent.

Table 1. Average Composition of the Indicated Regions Flanking the 1009 χ Sites, Compared to the E. coli Genome

	Average Composition (%)			
		Size of the Region Surrounding χ		Size of the Region Surrounding the CTGCAG Control Sequence
Base	E. coli	50 bp	500 bp	50 bp
A	24.6	22.3 (± 0.5)	22.6 (± 0.2)	24.2
C	25.4	25.3 (± 0.5)	25.3 (± 0.2)	25.8
G	25.4	27.9 (± 0.5)	27.8 (± 0.2)	25.6
T	24.6	24.5 (± 0.5)	24.3 (± 0.2)	24.3

The composition of the E. coli genome (GenBank accession number U00096), and the compositions of the bases surrounding χ and the control sequence, were determined using the Wisconsin Package Version 9.0 (Genetics Computer Group [GCG], Madison, Wisconsin). The 99% confidence intervals are shown in parenthesis.

Two-thirds of genomic χ sites have the trinucleotide CTG (which is, by far, the most frequent codon in E. coli) placed in frame (not shown). This raises the possibility that codon usage bias may contribute to the pattern. Therefore, we examined the nucleotide composition surrounding the hexameric control sequence 5′-CTGCAG-3′. This sequence was selected because it occurs nearly the same number of times (958) as χ and it contains the CTG codon followed by another very frequent codon, CAG, making it likely to be in frame as well. Table 1 shows that the average composition around this sequence is comparable to the E. coli average. Other controls addressing the potential bias of codon usage confirmed that the paucity of A and the abundance of G residues (particularly the GG dinucleotide repeats) are unique to χ (not shown).

Our finding that χ is surrounded by almost a kilobase of TGG-enriched DNA raises several interesting issues. In vitro selection showed that ssDNA with a high GT and low A content is preferred by RecA: the most preferred trinucleotides are TGG, GTG, GTT, GGT, and TGT; preferential binding also correlated with increased homologous pairing function (Tracy and Kowalczykowski, 1996). It is therefore apparent that χ sites are embedded throughout the genome in DNA sequences that are preferred for recombination by RecA. We propose that these regions constitute "recombination islands" that contribute to increased binding by RecA and, thus, to increased initiation of homologous recombination at χ.

The large size of these recombination islands (∼400 bp on each side of χ) makes it unlikely that they evolved after appearance of χ in the genome. We favor the idea that χ evolved in regions with biased nucleotide composition and intrinsic hot spot activity. Cells may have used the ability of these ancestral GT-rich regions to undergo genetic exchange at higher than "average" frequencies as a means of restoring genomic integrity after DNA breakage. This hypothesis is supported by observations that GT-rich loci are recombinogenic in other organisms. This behavior is especially true for microsatellite DNA (Jeffreys et al., 1985; Wahls et al., 1990), as well as for interstitial telomeric sequences in eukaryotic genomes (Biessmann and Mason, 1994). In addition, the constant regions of immunoglobulin heavy chains undergo rearrangement via homologous recombination at loci (switch sequences) that are GT-rich (Singer and Berg, 1991).

Finally, the RecA-preferred pairing sequences themselves possess an increased intrinsic ability to invade homologous DNA (R. B. T. and S. C. K., unpublished data), suggesting that RecA adopted use of these sequences. Furthermore, yeast Rad51, a RecA homolog, has a similar preference for GT-rich sequences (R. B. T. and S. C. K., submitted). This suggests that GT-rich regions may be used universally to stimulate homologous recombination. Clearly, E. coli adopted a strategy of embedding a regulatory sequence, χ, central to the recombination process, within loci of high intrinsic homologous pairing proficiency.

Robert B. Tracy,*†§ **Frédéric Chédin,***§
and Stephen C. Kowalczykowski*‡
*Sections of Microbiology
and of Molecular and Cellular Biology
†Graduate Group in Microbiology
University of California
Davis, California 95616

References

Anderson, D.G., and Kowalczykowski, S.C. (1997a). Genes Dev. *11*, 571–581.

Anderson, D.G., and Kowalczykowski, S.C. (1997b). Cell *90*, 77–86.

Biessmann, H., and Mason, J.M. (1994). Chromosoma *103*, 154–161.

Cardon, L.R., Burge, C., Schachtel, G.A., Blaisdell, B.E., and Karlin, S. (1993). Nucleic Acids Res. *21*, 3875–3884.

Dixon, D.A., and Kowalczykowski, S.C. (1991). Cell *66*, 361–371.

Dixon, D.A., and Kowalczykowski, S.C. (1995). J. Biol. Chem. *270*, 16360–16370.

Ennis, D.G., Amundsen, S.K., and Smith, G.R. (1987). Genetics *115*, 11–24.

Jeffreys, A.J., Wilson, V., and Thein, S.L. (1985). Nature *314*, 67–73.

Lam, S.T., Stahl, M.M., McMilin, K.D., and Stahl, F.W. (1974). Genetics *77*, 425–433.

Myers, R.S., Stahl, M.M., and Stahl, F.W. (1995). Genetics *141*, 805–812.

Singer, M., and Berg, P. (1991). Genes and Genomes (Mill Valley, CA: University Science Books).

Smith, G.R., Kunes, S.M., Schultz, D.W., Taylor, A., and Triman, K.L. (1981). Cell *24*, 429–436.

Tracy, R.B., and Kowalczykowski, S.C. (1996). Genes Dev. *10*, 1890–1903.

Wahls, W.P., Wallace, L.J., and Moore, P.D. (1990). Cell *60*, 95–103.

Acknowledgments

This work was supported by NIH grants AI-18987 and GM-41347.

‡To whom correspondence should be addressed.
§Both authors contributed equally to this work.