JOURNAL OF BACTERIOLOGY, June 1982, p. 1467-1471
0021-9193/82/061467-05$02.00/0

Vol. 150, No. 3

Primary Structure of the Chromosomal Origins (*oriC*) of *Enterobacter aerogenes* and *Klebsiella pneumoniae*: Comparisons and Evolutionary Relationships

JOSEPH M. CLEARY,† DOUGLAS W. SMITH, NANCY E. HARDING,† AND JUDITH W. ZYSKIND*

Department of Biology, University of California, San Diego, La Jolla, California 92093

Received 10 August 1981/Accepted 11 January 1982

The nucleotide sequences of the *Enterobacter aerogenes* and *Klebsiella pneumoniae* DNA replication origins (*oriC*) were determined and compared with those of *Escherichia coli* and *Salmonella typhimurium*. Four interrelated, 9-base-pair repeats were identified from the conserved regions within the minimal origin. Evolutionary rates calculated from the minimal origin sequences yielded a quantitative phylogenic tree which agreed with the taxonomic classification of these genera.

The chromosomal replication origin (*oriC*) is central to the events that encompass procaryotic DNA replication and cell division (4, 15). In the recent past, the *oriC* regions of *Escherichia coli* and *Salmonella typhimurium* have been isolated as autonomous replicons (10, 17, 18), and their nucleotide sequences have been determined (9, 16, 19). The minimal origin of *E. coli* has been delimited to 245 nucleotides by the ability of *oriC* to direct replication of chimeric plasmids (13).

To delineate those domains of *cis*-acting elements composing the chromosomal *oriC* region which are essential for interactions with molecular species required for the initiation process in *E. coli*, we have extended our study of the *oriC* locus to more distantly related members of the enteric family, namely, *Enterobacter aerogenes* and *Klebsiella pneumoniae*. Members of these two genera of gram-negative bacteria are grouped in the subfamily or tribe *Klebsielleae*, whereas *E. coli* and *S. typhimurium* are classified in a separate tribe, *Escherichieae* (1). We present here the DNA sequences of the *oriC* loci for *E. aerogenes* and *K. pneumoniae* with a comparison to those of *E. coli* and *S. typhimurium*.

Chromosomal DNAs from *E. aerogenes* and *K. pneumoniae* were digested with *Sal*I and ligated to pMK2004 (6) also digested with *Sal*I. Plasmids pNH3 and pNH62 containing *oriC* loci from *E. aerogenes* and *K. pneumoniae*, respectively, were isolated from colonies obtained after transformation of *E. coli* strain C2368 F^- *polA1 thy his rha* r_k^- m_k^- with the ligation mixtures. Restriction and genetic maps and properties of these plasmids will be published elsewhere. Plasmid pNH305 was constructed by inserting the fragments generated by *Pst*I digestion of pNH3 into the *Pst*I site of pMK2004. After transformation of C2368 selecting for kanamycin and tetracycline resistance, pNH305 was isolated. Restriction analysis of pNH305 showed that it consisted of pMK2004 plus a 2.35-kilobase-pair *Pst*I fragment containing the bacterial origin of *E. aerogenes*. Plasmid pJZ70 was constructed in a similar manner from pNH62 and consists of the bacterial origin of *K. pneumoniae* in a 4.75-kilobase-pair *Pst*I fragment inserted into the *Pst*I site of pMK2004.

The nucleotide sequencing strategy is shown in Fig. 1; restriction fragments from plasmids pNH305 for *E. aerogenes* and pJZ70 for *K. pneumoniae* were used. DNA fragments were end labeled with [γ-^{32}P]rATP and polynucleotide kinase (P-L Biochemicals) as described by Maxam and Gilbert (8). The [γ-^{32}P]rATP was synthesized as previously described (5). DNA sequences were determined by the Maxam and Gilbert chemical method (8), with strand-separating gels run as described (19). The entire sequence for both DNA strands within the minimal *oriC* region was determined for the two bacterial origins (Fig. 1). DNA isolation, enzymes, reaction conditions, and gel electrophoresis analyses of DNA restriction fragments were as described (19). DNA fragments for sequence analysis were isolated by either secondary cleavage or strand separation of the DNA after labeling with polynucleotide kinase. The sequences of all internal restriction sites were determined by overlapping fragments to eliminate the possibility of missing small fragments.

The nucleotide sequences of the *oriC* regions

† Present address: Kelco-Division of Merck and Co., Inc., San Diego, CA 92123.

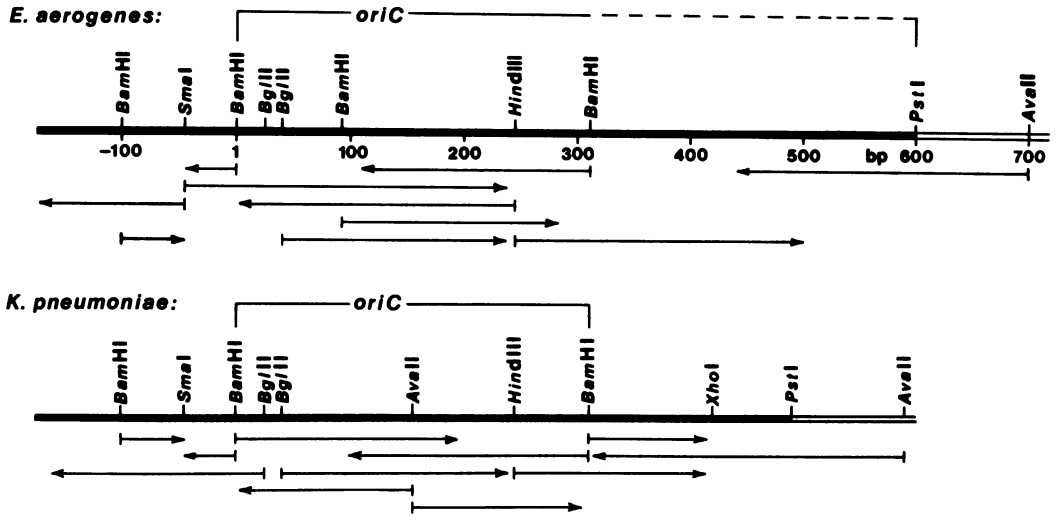

FIG. 1. Strategies for Maxam-Gilbert sequence determination of the *oriC* regions from *E. aerogenes* and *K. pneumoniae*. Tail of arrow, ^{32}P-labeled end of each fragment; arrowhead, extent to which sequence was determined from that fragment. Solid regions designate bacterial DNA, and open regions designate cloning vehicle DNA (pMK2004; see ref. 6).

from *E. aerogenes* and *K. pneumoniae* (Fig. 2) are almost entirely congruent. Only two single-base insertions in *E. aerogenes* at positions −7 and +279 prevent exact alignment of the 556 nucleotides, and both of these lie outside the minimal origin sequence (+23 to +267, determined for *E. coli* [13]). In addition, there are extensive regions of total homology separated mainly by clusters of nucleotide differences. The largest identical stretch is 49 nucleotides (positions −4 to +45). Within the presented 556 base pairs, there are 71 nucleotide differences, with 44 transitions and 27 transversions.

When the chromosomal origin regions of *E. aerogenes* and *K. pneumoniae* are compared with those of *E. coli* (9, 16) and *S. typhimurium* (19), the nucleotide differences in the *oriC* region are found mainly in clusters, as previously observed (19). These clustered differences in the *Klebsielleae* (*E. aerogenes* and *K. pneumoniae*) sequences predominantly overlap those found in the *Escherichieae* (*E. coli* and *S. typhimurium*) sequences, for example, between positions +71 and +78 (Fig. 2). The clusters of changes average 5 to 8 nucleotides and are interspersed with stretches of complete homology of up to 30 nucleotides. This pattern of differences is similar to the distribution of homology observed for transcriptional regulatory sequences in the *trp* operon promoter-operator region (14). With the exception of three single-base insertions and a single-base deletion, the four origin sequences are congruent. The *E. aerogenes* and *K. pneumoniae* sequences are greater than 87% homologous, those of *E. coli* and *S. typhimurium* are about 84% homologous, and those between the two tribes are 80 to 81% homologous, both within the minimal origin and in the total sequences presented.

One of the most striking observations concerning the conserved regions within the origins is that there are four 9-base-pair repeats, two in opposite orientation to the other two. Repeats between positions 78 and 88 and 258 and 268 are exact inverted repeats of each other. The two other related repeats are found between positions 184 and 194 and 219 and 229 and have a single-base difference with the first two repeats mentioned. The conservation and symmetrical location of these repeats suggest that their role in initiation is important.

As with *E. coli* and *S. typhimurium*, stop codons in all three coding frames within the *oriC* region of *E. aerogenes* and *K. pneumoniae* limit the size of possible protein products to small oligopeptides. Extensive potential intrastrand secondary structure possibilities exist, many of which are conserved among all four origins. The proposed stem-loop structure between positions 221 and 244 and the stem region of the suggested cloverleaf structure (19) are both completely conserved, although the arms of this cloverleaf structure are not. The extensive homology and regular spacing of GATC sites in the region between positions 20 and 70 make possible elaborate intrastrand structures. Of the 18 GATC sequences found in the *oriC* regions of both *E. coli* and *S. typhimurium*, 16 are con-

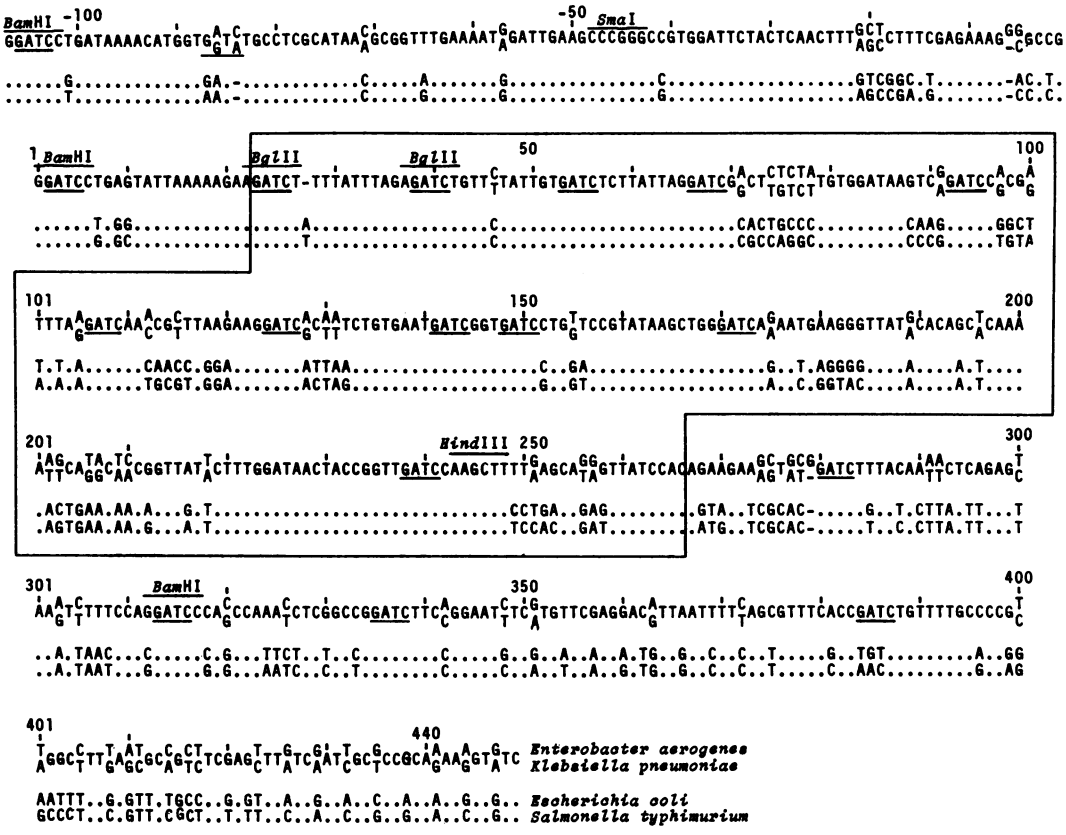

FIG. 2. Nucleotide sequences of the *E. aerogenes* and *K. pneumoniae oriC* regions. The first line of each row includes both the *E. aerogenes* and *K. pneumoniae* sequences. Identical nucleotides appear in the middle of this line; at positions where there are nucleotide differences between the two species, the nucleotide for *E. aerogenes* is given above and the nucleotide for *K. pneumoniae* is given below. For comparison, the sequences from *E. coli* (second line from bottom [9, 16]) and *S. typhimurium* (bottom line [19]) are shown; here dots are used when the nucleotide is identical in all four bacteria. A dash indicates a deletion in one species relative to others put in to optimize sequence alignment. Representative restriction sites common to *E. aerogenes* and *K. pneumoniae* are indicated. GATC sites present in either *E. aerogenes* or *K. pneumoniae* are underlined. The box encloses the minimal *E. coli* origin (13), and the numbering begins with the conserved *Bam*HI site used previously (9, 16) but refers to *E. aerogenes* nucleotides. The upper left end is the 5′ end.

served in *E. aerogenes* and *K. pneumoniae*, and all 11 GATC sites within the minimal origin are conserved.

Availability of nucleotide sequences for a given function such as the minimal *oriC* region also permits calculations of evolutionary rates between these species. Among the several ways to do this, the recent method of Kimura (7) accounts separately for transition and transversion type substitutions. Assuming equal probability for the substitution of any base at a given site, twice as many transversion-type substitutions are possible as transition-type substitutions. His method can be applied to both coding and noncoding regions (7). Two parameters, P, the fraction of nucleotides showing transition-type substitutions, and Q, the fraction of nucleotides showing transversion-type substitutions, are used to calculate K, the evolutionary distance (total number of substitutions) per site. K provides a relative evolutionary distance between the species, assuming a constant mutation rate in time. Results for K and its standard error are shown in Table 1 for the minimal origin region (Fig. 2). Comparison of the K values for each pair of *oriC* sequences unambiguously shows that the two members of each tribe are more closely related to each other than to either of the members of the other tribe, in agreement with the Bergey classification of these four genera (1). The K values can be used to obtain a quantitative phylogenic tree (Fig. 3). Thus, *E.*

TABLE 1. Evolutionary rate calculations via pairwise comparisons[a]

Sequence compared	n	n_P	n_Q	K(P,Q)	σ(P,Q)
E. coli, *S. typhimurium*	245	18	18	0.164	0.028
E. aerogenes, *K. pneumoniae*	244	16	14	0.135	0.025
S. typhimurium, *E. aerogenes*	245	16	29	0.210	0.033
S. typhimurium, *K. pneumoniae*	245	24	27	0.245	0.036
E. coli, *E. aerogenes*	245	22	23	0.212	0.033
E. coli, *K. pneumoniae*	245	27	20	0.224	0.035

[a] n, Number of nucleotide base pairs in the region compared; n_P, number of transition-type differences; n_Q, number of transversion-type differences; K(P, Q), evolutionary distance per site; σ(P, Q), standard error in K.

coli and *S. typhimurium* have diverged from a common *Escherichieae* tribe ancestor, called E; *E. aerogenes* and *K. pneumoniae* have diverged from a common *Klebsielleae* tribe ancestor, called K; and the ancestors E and K have diverged from a common *Enterobacteriaceae* ancestor, called A. Other methods for calculating evolutionary rates, e.g., that of Hori and Osawa (3), yielded similar results. The numbers indicated in Fig. 3 are linearly related to the K values, and hence measure the number of base substitutions occurring since divergence from the previous ancestor. Thus, the *E. aerogenes* minimal origin has suffered fewer changes than the *K. pneumoniae* origin since divergence of each from ancestor K, owing to either a difference in mutation rate or different times of divergence (or both).

The two regions outside the minimal origin sequence possess distinctly different features. The minus region (positions 22 to −107) contains clustered nucleotide changes resembling those found within the minimal origin. The degree of homology among all four organisms of this region (83%) is somewhat greater than that of the minimal origin sequence (79%). In contrast, the plus region (positions 267 to 449) contains a considerable number of nucleotide changes (only 54% conservation among all four species), with very little clustering of these changes. This region in *E. coli* was originally postulated on the basis of DNA sequence analysis to encode the carboxy terminal portion of a structural gene (2). Recently, Hansen et al. have demonstrated that a 15.5-kilodalton protein is synthesized in vivo from plasmids carrying only this region of the *E. coli* chromosome (1a). Comparison of the *Escherichieae* and *Klebsielleae* sequences in this region (Fig. 3) suggests that *S. typhimurium*, *E. aerogenes*, and *K. pneumoniae* code for similar proteins. Between the termination codon (TTA in the *Escherichieae* sequences) at position +291 and position +449, 65% of the nucleotide differences (41 of 63) among all four bacterial sequences are in the third position of the predicted *E. coli* reading frame, and no other stop codons are found in this frame. A stretch of 10 consecutive codons between positions +377 and +348 contain a change at every third position. Where other nucleotide sequences coding for proteins have been compared, for example, in the *trpG* region (12), most of the nucleotide substitutions are in the third position of the codons, reflecting synonymous codon usage. The nucleotide sequence for 52 amino acids at the carboxy termi-

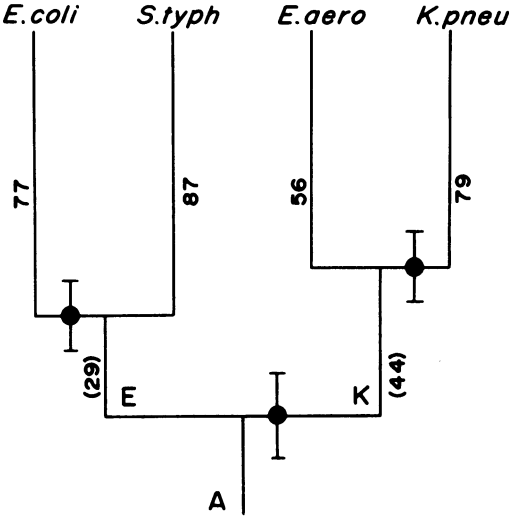

FIG. 3. Phylogenetic tree determined from the minimal origin regions. The K values × 1,000 (see Table 1 for K values) were used to determine relative times of divergence of each bacterial species from their respective ancestors (E, K, and A; see text). Since the K value for each pair of species represents the relative evolutionary distance between these two species, the point of divergence of each pair from its ancestor on the vertical axis is drawn as 1/2 × K × 1,000. The numbers indicated were calculated from the set of six equations resulting from each pair of species. Each number is the average of the set of values obtained from all possible solutions of the six simultaneous equations with five independent variables. These numbers are different from the expected values of 1/2 × K × 1,000; possible interpretations of these differences are discussed in the text. The numbers 29 and 44 are in parentheses since only their sum can be determined from these equations; this sum has been arbitrarily divided to give a common point in time (vertical axis) for divergence of E and K from A.

nal end of the 15.5-kilodalton *E. coli* protein is included in Fig. 2. The reading frames in the *Klebsielleae* sequences do not end at the *Escherichieae* stop codon but at the termination signal, TCA (+294), located one codon upstream. Such an alteration of the termination codon position is also observed in the *trpA* gene sequence between *K. aerogenes* and *E. coli* (11). Although these results suggest that the *E. coli* 15.5-kilodalton protein is conserved in the other species, its function is unknown and does not appear to be essential for replication at *oriC* (1a).

The sequence comparison of these four naturally occurring bacterial replication origins which are capable of functioning in *E. coli* yields a pattern of homology seen as a clustering of nucleotide differences interspersed with conserved regions up to 30 base pairs long. Information presented here will provide a foundation for the determination of function of these sequences, for example, as protein binding sites or RNA synthesis templates. This comparative study has also permitted a detailed analysis of base changes (mutational events) occurring within a large DNA regulatory sequence which probably does not code for a gene product.

This work was supported by Public Health Service grant GM21978 and postdoctoral traineeship GM07199 from the National Institutes of Health to J.M.C.

LITERATURE CITED

1. **Cowan, S. T.** 1974. Family I. *Enterobacteriaceae*, p. 290–340. *In* R. E. Buchanan and N. E. Gibbons (ed.), Bergey's manual of determinative bacteriology, 8th ed. The Williams and Wilkins Co., Baltimore.

1a. **Hansen, F. G., S. Koeford, F. von Meyenburg, and T. Atlung.** 1981. Transcription and translation events in the *oriC* region of the *E. coli* chromosome. Symp. Mol. Cell Biol. **22**:37–55.

2. **Hirota, Y., M. Yamada, A. Nishimura, A. Oka, K. Sugimoto, K. Asada, and M. Takanami.** 1981. The DNA replication origin (*ori*) of *Escherichia coli*: structure and function of the *ori*-containing DNA fragment. Prog. Nucleic Acid Res. Mol. Biol. **26**:33–47.

3. **Hori, H., and S. Osawa.** 1979. Evolutionary change in 5s RNA secondary structure and a phylogenic tree of 54 5s RNA species. Proc. Natl. Acad. Sci. U.S.A. **76**:381–385.

4. **Jacob, F., S. Brenner, and F. Cuzin.** 1964. On the regulation of DNA replication in bacteria. Cold Spring Harbor Symp. Quant. Biol. **26**:329–348.

5. **Johnson, R. A., and T. F. Walseth.** 1979. The enzymatic preparation of [α-^{32}P]ATP, [α-^{32}P]GTP, [^{32}P]cAMP and [^{32}P]cGMP, and their uses in the assay of adenylate and guanylate cyclases and cyclic nucleotide phosphodiesterases. Adv. Cyclic Nucleotide Res. **10**:135–167.

6. **Kahn, M., R. Kolter, C. Thomas, D. Figurski, R. Meyer, E. Remaut, and D. Helinski.** 1979. Plasmid cloning vehicles derived from plasmids ColE1, F, R6K and RK2. Methods Enzymol. **68**:268–280.

7. **Kimura, M.** 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

8. **Maxam, A., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.

9. **Meijer, M., E. Beck, F. G. Hansen, H. E. N. Bergmans, W. Messer, K. von Meyenburg, and H. Schaller.** 1979. Nucleotide sequence of the origin of replication of the *Escherichia coli* K-12 chromosome. Proc. Natl. Acad. Sci. U.S.A. **76**:580–584.

10. **Messer, W., H. E. N. Bergmans, M. Meijer, J. E. Womack, F. G. Hansen, and K. von Meyenburg.** 1978. Minichromosomes: plasmids which carry the *E. coli* replication origin. Mol. Gen. Genet. **162**:269–274.

11. **Nichols, B. P., M. Blumenberg, and C. Yanofsky.** 1981. Comparison of the nucleotide sequence of *trpA* and immediately beyond the *trp* operon of *Klebsiella aerogenes*, *Salmonella typhimurium* and *Escherichia coli*. Nucleic Acids Res. **9**:1743–1756.

12. **Nichols, B. P., G. R. Miozzari, M. van Cleemput, G. N. Bennett, and C. Yanofsky.** 1980. Nucleotide sequences of the *trpG* regions of *Escherichia coli*, *Shigella dysenteriae*, *Salmonella typhimurium* and *Serratia marcescens*. J. Mol. Biol. **142**:503–517.

13. **Oka, A., K. Sugimoto, M. Takanami, and Y. Hirota.** 1980. Replication origin of the *Escherichia coli* K12 chromosome: the size and structure of the minimum DNA segment carrying the information for autonomous replication. Mol. Gen. Genet. **178**:9–20.

14. **Oppenheim, D. S., G. N. Bennett, and C. Yanofsky.** 1980. *Escherichia coli* RNA polymerase and *trp* repressor interaction with the promoter-operator region of the tryptophan operon of *Salmonella typhimurium*. J. Mol. Biol. **144**:133–142.

15. **Pritchard, R. H., P. T. Barth, and J. Collins.** 1969. Control of DNA synthesis in bacteria. Symp. Soc. Gen. Microbiol. **19**:263–297.

16. **Sugimoto, K., A. Oka, H. Sugisaki, M. Takanami, A. Nishimura, S. Yasuda, and Y. Hirota.** 1979. Nucleotide sequence of *Escherichia coli* K-12 replication origin. Proc. Natl. Acad. Sci. U.S.A. **76**:575–579.

17. **Yasuda, S., and Y. Hirota.** 1977. Cloning and mapping of the replication origin of *E. coli*. Proc. Natl. Acad. Sci. U.S.A. **74**:5458–5462.

18. **Zyskind, J. W., L. T. Deen, and D. W. Smith.** 1979. Isolation and mapping of plasmids containing the *Salmonella typhimurium* origin of DNA replication. Proc. Natl. Acad. Sci. U.S.A. **76**:3097–3101.

19. **Zyskind, J. W., and D. W. Smith.** 1980. Nucleotide sequence of the *Salmonella typhimurium* origin of DNA replication. Proc. Natl. Acad. Sci. U.S.A. **77**:2460–2464.