Distribution and Abundance of Insertion Sequences Among Natural Isolates of *Escherichia coli*

Stanley A. Sawyer,*,† Daniel E. Dykhuizen,* Robert F. DuBose,* Louis Green,* T. Mutangadura-Mhlanga,‡ David F. Wolczyk§ and Daniel L. Hartl*

**Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110,* †*Department of Mathematics, Washington University, St. Louis, Missouri 63130,* ‡*Department of Biological Sciences, University of Zimbabwe, Mount Pleasant, Harare, Zimbabwe and* §*Department of Biology, Reed College, Portland, Oregon 97202*

Manuscript received July 7, 1986
Revised copy accepted October 11, 1986

ABSTRACT

A reference collection of 71 natural isolates of *Escherichia coli* (the ECOR collection) has been studied with respect to the distribution and abundance of transposable insertion sequences using DNA hybridization. The data include 1173 occurrences of six unrelated insertion sequences (IS*1*, IS2, IS*3*, IS*4*, IS*5* and IS*30*). The number of insertion elements per strain, and the sizes of DNA restriction fragments containing them, is highly variable and can be used to discriminate even among closely related strains. The occurrence and abundance of pairs of unrelated insertion sequences are apparently statistically independent, but significant correlations result from stratifications in the reference collection. However, there is a highly significant positive association among the insertion sequences considered in the aggregate. Nine branching process models, which differ in assumptions regarding the regulation of transposition and the effect of copy number on fitness, have been evaluated with regard to their fit of the observed distributions. No single model fits all copy number distributions. The best models incorporate no regulation of transposition and a moderate to strong decrease in fitness with increasing copy number for IS*1* and IS*5*, strong regulation of transposition and a negligible to weak decrease in fitness with increasing copy number for IS*3*, and less than strong regulation of transposition for IS2, IS*4* and IS*30*.

INSERTION sequences are a special class of transposable elements found in prokaryotes that have been studied extensively in *Escherichia coli* K12 (CALOS and MILLER 1980). They usually range in size from 1 to 2 kilobasepairs (kb) and contain perfect or nearly perfect inverted terminal repeat sequences. The terminal sequences flank a unique central sequence with at least one long open reading frame coding, presumably, for the transposase protein.

One evolutionary implication of insertion sequences derives from their mutagenic activity in causing insertion mutations. The ability of insertion sequences to inactivate genes is well known and used, for example, in the technique of transposon mutagenesis (BOTSTEIN and SHORTLE 1985). The insertion of elements into suitable sites in the genome can also result in the activation of cryptic genes (REYNOLDS, FELTON and WRIGHT 1981) or the overexpression of genes nearby (SAEDLER *et al.* 1980; CIAMPI, SCHMID and ROTH 1982; MILLER *et al.* 1984).

Insertion sequences also play an important role in the evolution of transposons and plasmids. A pair of insertion sequences flanking a central sequence can transpose as a unit, and such composite transposons containing antibiotic-resistance genes, for example, Tn*5*, Tn*9* and Tn*10*, are well documented (CALOS and MILLER 1980). Insertion sequences and transposons can also transpose into plasmids and remold their structure, and in general change the genetic capabilities of plasmids. By providing sites for replicon fusion they can also enable the transfer of nonconjugative plasmids during conjugation.

A third evolutionary role of insertion sequences is that of selfish or parasitic DNA sequences (DOOLITTLE and SAPIENZA 1980; ORGEL and CRICK 1980), which are maintained in the population, even in the face of adverse natural selection, as a result of their ability to transpose and become horizontally transmitted among strains by means of hitchhiking in plasmids. In theory, the number of insertion sequences and their locations within the genome can change rapidly. Therefore, the finding of close similarity or identity among strains in the number and position of insertion elements provides strong evidence of recent common ancestry. Insertion sequences in the genome are rapidly evolving molecular markers that may be of important practical value in epidemiology in tracing the place of origin or genetic ancestry of bacterial isolates (GREEN *et al.* 1984).

The distribution of numbers of insertion sequences in the genome is also of some interest in understanding the population dynamics of the elements. We have studied the distribution of six insertion sequences (IS*1*, IS*2*, IS*3*, IS*4*, IS*5* and IS*30*) among a reference collection of 71 strains of *E. coli* (the ECOR collection), representing isolates from humans and animals of diverse geographical locations. The distributions were tested against nine branching process models differing in strength of regulation of transposition and effect of copy number in reducing fitness. Maximum likelihood estimates of the parameters in the best fitting models are provided.

MATERIALS AND METHODS

Strains: Numbers of insertion sequences were determined among strains in the ECOR reference collection of *E. coli.* Details on the sources of the strains and the electrophoretic types are given in OCHMAN and SELANDER (1984) and SELANDER, CAUGANT and WHITTAM (1986).

DNA preparation: DNA from the ECOR strains was extracted by the method described in HARTL *et al.* (1983), which enriches for chromosomal DNA. Plasmid DNA was prepared by the alkaline lysis method of KADO and LIU (1981), except that after 20–30-min incubation at 65° in the alkaline lysis solution, ½ volume of 3 M NaOAc (pH = 5.3) was added and the solution incubated in ice for 60 min prior to phenol and chloroform extraction. Chromosomal and plasmid DNA preparations were digested with *Eco*RI, and the restriction fragments were fractionated by electrophoresis in adjacent lanes of 0.7% agarose gels and transferred to nitrocellulose (Schleicher and Schuell BA85) or nylon membranes (Amersham Hybond) according to the manufacturer's instructions and hybridized with appropriate nick translated probes (SOUTHERN 1975). Electrophoresis of chromosomal and plasmid restriction fragments permitted identification of a small number of bands in a few chromosomal preparations that actually resulted from contaminating plasmid DNA. DNA hybridizations were also carried out with undigested plasmid DNA in order to determine whether insertion sequences were contained in large (greater than 25 kb) or in small (less than 25 kb) plasmids.

Insertion sequence probes: IS*1* was probed with a 0.636-kb internal fragment isolated from pBRG36 (BIEL, ADELT and BERG 1984), kindly provided by D. E. BERG. Plasmid pBRG36 is a pBR322::Tn*9* derivative, which was digested with *Pvu*II, *Tth*111I, and *Eco*RI to separate the *Pvu*II/*Tth*111I fragment of IS*1* from a comigrating fragment from Tn*9*. The IS*2* probe was a 0.717-kb *Hpa*I/*Hin*dIII internal fragment isolated from pBRK10 (KLAER *et al.* 1981), kindly provided by P. STARLINGER. The IS*3* probe was a 0.916-kb *Hin*dIII internal fragment isolated from plasmid pSH2 (ACHTMAN *et al.* 1978), which was subcloned into the *Hin*dIII site of pBR322. The IS*30* probe was a 1.1-kb internal *Bgl*II/*Hin*cII fragment isolated from plasmid pAW522 (DALRYMPLE, CASPERS and ARBER 1984), kindly provided by W. ARBER and B. DALRYMPLE.

Two probes were used for the insertion element γδ. The first was a 0.8-kb *Sal*I/*Hin*dIII internal fragment including most of the resolvase, and the second was a 1.7-kb *Hin*dIII/*Sal*I fragment including much of the transposase. Both fragments were isolated from plasmid pUF6 (STOKES and HALL 1984), a pBR322::γδ plasmid kindly provided by B. G. HALL. These probes hybridize very weakly to transposon Tn*3* (data not shown), which is related to γδ.

The probes used for IS*4* and IS*5* have been described previously (GREEN *et al.* 1984; DYKHUIZEN *et al.* 1985). All probes were isolated from agarose by means of the glass powder method (HARTL *et al.* 1983) and nick-translated with α-^{32}P-dATP to a specific activity greater than 10^8 dpm/μg according to the method of RIGBY *et al.* (1977).

Hybridization: Hybridization were carried out at 65° for 16–20 hr in 0.7 M sodium ion buffer (moderate stringency). Occasional weakly hybridizing bands were seen for all insertion sequences. These were always greatly in the minority (for example, 2/68 bands with IS*30*), and they were not included in the totals. We have not determined whether the weak hybridization signals result from fragments of the insertion sequences being probed or different insertion sequences with less homology to the probe.

Since most strains contain fewer than 10–15 chromosomal copies of any insertion sequence, estimates of the number of copies obtained by counting the number of hybridization bands on gels are quite accurate and reliable. Beyond about 15 elements occasional ambiguity creeps in, and the highest numbers (relevant only for IS*1*) are probably accurate to within ±2 elements. However, virtually all of our statistical procedures are on aggregated data (for example, aggregated by combining all strains containing five or more elements), and ascertainment errors of this size would have had no effect.

RESULTS

Genetic variation and correlation in insertion sequences

Genomic location of γδ-related sequences: Only four ECOR strains contained DNA sequences that hybridized with either of the γδ probes. In all four strains the γδ-related sequence was contained in a large plasmid. The location of these sequences in plasmids is expected based on the close evolutionary relationship between γδ and Tn*3*, as indicated by their ability to undergo homologous recombination (REED 1981), and the strong preference of Tn*3* to transpose into plasmid replicons (HEFFRON 1983). Since both the 0.8 γδ resolvase probe and the 1.7-kb γδ transposase probe hybridized with the element found in ECOR strain 35, the element in this strain is evidently complete. However, only the transposase probe hybridized with the element in ECOR strains 10, 31 and 8. The hybridizing sequences in these three strains thus seem to be atypical γδ elements missing at least the resolvase sequence. Moreover, the hybridization signal in ECOR strain 8 was notably weaker than in the others, suggesting the presence of a related but nonidentical element.

IS elements in strain identification: Insertion elements are dynamic in the genome, and the rate of transposition is typically at least an order of magnitude greater than the rate of deletion (EGNER and BERG 1981; FOSTER *et al.* 1981). For example, the rate of transposition of IS*50* is about 10^{-3} per infected cell per generation (HARTL *et al.* 1983), whereas the rate of deletion is about 10^{-5} per element per generation (EGNER and BERG 1981). Both of these rates are much greater than the rates of nucleotide substitution that alter restriction sites or the electrophoretic mobility of proteins. However, since insertion sequences are inherited, strains that share a recent common ancestor should also tend to share some or all of the sites at which insertion sequences are present in the genome. These shared sites would result in restriction fragments of identical size hybridizing with the insertion sequence probe in Southern blots (insertion sequence bands). Novel bands not shared between closely related strains may result from combinations of transposition, deletion, or mutations in restriction sites. Considering the greater rate of transposition than of deletion or point mutation, novel bands found in strains sharing a very recent common ancestor will usually result from transposition.

Rates of transposition are sufficiently great that the similarity in position of insertion sequences in the genome may serve as a useful, rapidly evolving genetic marker of recent common ancestry among strains. This point was examined by comparing the sizes of restriction fragments bearing insertion sequences among strains in the ECOR collection which are closely related as evidenced by their identical or nearly identical electrophoretic mobilities of enzyme proteins.

Selander, Caugant and Whittam (1986) have identified 17 ECOR strains that can be assigned to seven groups in which all members of each group demonstrate an identical electrophoretic mobility of each of 35 enzymes. The most similar of these strains in terms of insertion sequences were ECOR strains 21 and 22, which were isolated from a single steer; these strains have 98% of their insertion sequence bands in common. In the other groups many differences in insertion sequence bands were found. On average, strains that are electrophoretically identical at 35 enzyme loci share just 43% of their insertion sequence bands.

Selander, Caugant and Whittam (1986) also identified 17 strains comprising eight groups in which the strains within groups differ in electrophoretic mobility in only one enzyme. The most similar strains in these groups were ECOR strains 26 and 27; however, these strains contained only two insertion sequences (both IS*3*), which occurred in restriction fragments of identical size. In other groups the differences in insertion sequence bands were much greater than found among electrophoretically identical strains. With strains that differ in one enzyme, an average of only 28% of the insertion sequence bands are in common. The distribution of similarity based on insertion sequence bands is shown in Figure 1.

With few exceptions, ECOR strains that differ in the electrophoretic mobility of two or more enzymes share no common insertion sequence bands. For example, among 13 ECOR strains in four groups that differ within groups in the electrophoretically mobility of two enzymes (Selander, Caugant and Whittam 1986), three of the groups contained no identical insertion sequence bands. However, strains in the fourth group (ECOR 38, 39, 40 and 41) shared an average of 25% of their bands, suggesting relatively recent common ancestry. Strain 39, which differs at two electrophoretic loci from the others, shares as many insertion sequence bands as they do, suggesting that the electrophoretic differences in strain 39 might have been introduced by recombination. These strains also provide evidence of the deletion of insertion elements. For example, ECOR strains 38, 39 and 41 each contained a single IS*3* fragment identical in size, whereas strain 40 contained no IS*3* elements; yet shared bands containing other insertion sequences in these strains imply that IS*3* must have been present in an ancestor of strain 40.

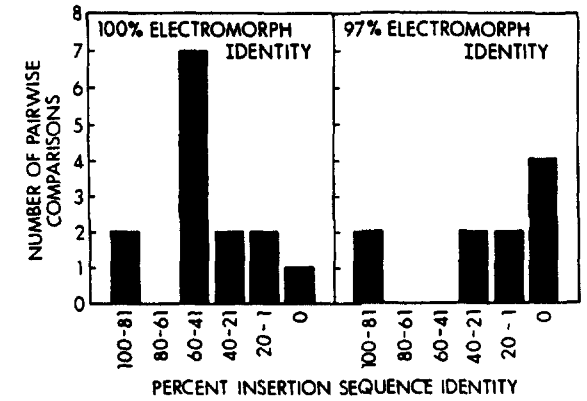

Figure 1.—Proportion of insertion sequence bands in common among pairs of ECOR strains that are identical in the electrophoretic mobility of 35 enzymes (*left panel*) or among 34 of 35 enzymes (*right panel*).

Five groups containing 21 strains differ within groups in the electrophoretic mobility of three enzymes (Selander, Caugant and Whittam 1986). No comparisons demonstrate any insertion sequence bands in common except for ECOR strains 49 and 50, which share two identical bands each for IS*1*, IS*5* and IS*30*, in addition to other bands that hybridize with these elements. Considering just these three elements, the proportion of insertion sequence bands in common in ECOR strains 49 and 50 is 76%. Although strain 49 lacked copies of IS*3* and IS*4*, strain 50 contained one copy of IS*3* and 11 copies of IS*4*. The latter finding suggests that the number of elements can increase rapidly in a relatively short time after infection.

Chromosomal *vs.* plasmid copies: Data were collected and analyzed for the number of chromosomal copies of each IS element present in each strain, as well as for the combined number of chromosomal and plasmid copies of the IS elements. (Plasmid copies were uncorrected for possible differences in plasmid abundance, *e.g.*, high copy number plasmids *vs.* low copy number plasmids). Among a total of 1173 observed insertion sequence bands, 117 (10%) were

TABLE 1

Distribution of chromosomal copy numbers

IS	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	Total
IS*1*	11	14	8	6	7	1	6	2	0	0	1	0	0	0	0	1	1	2	1	1	3	0	1	0	1	2	1	1	0	71
IS2	28	8	12	5	5	1	3	0	2	1	0	2	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	71
IS*3*	23	10	19	10	6	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	71
IS*4*	43	5	5	3	5	2	0	0	2	0	0	3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	71
IS*5*	46	12	3	2	2	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	71
IS*30*	36	16	13	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	71

contained in plasmids. The proportion of insertion sequence bands in plasmids ranged from a low of 3% with IS*4* to a high of 16% with IS*5* (excluding $\gamma\delta$). However, since the results and conclusions were completely unaffected by the inclusion of the plasmid data, we will present and discuss only the analysis based on the number of IS elements present in the chromosome. Complete data on the numbers of elements present in the chromosome and plasmids among the 71 strains are given in the APPENDIX. The distribution of the number of chromosomal copies is given in Table 1. For the purposes of curve fitting, strains with five or more copies were grouped together. The grouped distributions are displayed in Table 2.

Stratification in the sample: The ECOR strains were chosen from as wide a variety as possible of naturally occurring *E. coli* sources. In that respect, the ECOR collection is better regarded as a representative sample of *E. coli* rather than as a random sample. For example, the three subgroups of strains designated type I, type II and type III, which have been identified by means of factor analysis of measures of genetic difference (WHITTAM, OCHMAN and SELANDER 1983), are not represented in the ECOR collection in the same proportions as they would occur in random samples. The ECOR strains include isolates from both human and nonhuman sources. The strains isolated from humans are further stratified in that some derive from Sweden and others from North America, and some are from pathogenic and others are from nonpathogenic isolates. Tests were carried out to determine whether the occurrence of IS elements in the strains was correlated with the obvious stratifications in the sample, namely (a) type I *vs.* type II *vs.* type III; (b) human origin *vs.* animal origin, and, within human isolates; (c) Swedish origin *vs.* North American origin; and (d) pathogenic *vs.* nonpathogenic isolates.

We constructed contingency tables for each of these possible stratifications for the presence or absence of each IS element (see Table 3). The (two-sided) Fisher exact test was used to test the 2×2 tables and the Pearson χ^2 test for the 2×3 comparisons. Two highly significant ($P < 0.01$) associations were found, namely IS*3* with type I strains and IS*4* with isolates from humans. There were four other significant ($P < 0.05$) associations: IS*1* and IS*30* with type I strains, IS*4* with

TABLE 2

Observed distributions

	IS*1*	IS2	IS*3*	IS*4*	IS*5*	IS*30*
Chromosomal copies						
0	11	28	23	43	46	36
1	14	8	10	5	12	16
2	8	12	19	5	3	13
3	6	5	10	3	2	2
4	7	5	6	5	2	2
≥5	25	13	3	10	6	2
Max. No. of copies	27	17	6	14	21	5
Mean (all strains)	6.37	2.72	1.68	2.00	1.18	0.93
Mean (infected strains)	7.53	4.49	2.48	5.07	3.36	1.89
No. strains with ≥1 plasmids	22	13	9	5	14	6

TABLE 3

Heterogeneity tests[a]

	IS					
	IS*1*	IS2	IS*3*	IS*4*	IS*5*	IS*30*
Type:	0.028 (I)	—	0.004 (I)	0.020 (III)	0.052 (I)	0.015 (I)
An/Hum:	0.055 (H)	—	0.041 (An)	0.008 (H)	—	—
NA/Swe:	—	0.094 (Sw)	0.052 (NA)	0.101 (Sw)	—	—
Path:	—	—	—	—	—	—

[a] An = animal origin, Hum = human origin; NA = North American isolate, Swe = Swedish isolate; Path = pathogenic *vs.* nonpathogenic isolate. Statistical tests were Pearson *P*-values for types and Fisher two-sided exact test for others. — = not significant at $P = 0.11$ Favored strata indicated in parentheses.

type III strains, and IS*3* with isolates from animals. Five nearly significant ($0.05 \leq P < 0.11$) associations were also found: IS*5* with type I strains and IS*1* with isolates from humans, and, within the human strains, IS2 and IS*4* with Swedish isolates, and IS*3* with isolates from North America. All other comparisons had higher *P*-values. Within-strain copy numbers of the IS elements were also tested for stratification by nonparametric procedures, and yielded similar results. Note that while these associations are exact for the ECOR data set, they may overstate the degree of stratification among wild *E. coli* strains. This is due to the fact that with 24 independent tests of association one might expect 1.2 significant associations ($P = 0.05$) purely on the basis of chance. However, the probability that five out of six associations would be significant for types purely by chance is $<10^{-5}$.

Pairwise correlations in presence or absence of insertion elements: With six IS elements there are 15 possible pairwise associations. Using the entire data set, five of these associations were significant (using the Fisher exact test). However, most of the correlations were in the direction expected from stratification in the sample. For example, IS*1* and IS*30* are correlated within strains ($P < 0.001$), but both of these elements occur preferentially in type I strains. When the 15 possible associations were analyzed separately for type I, type II and type III strains, only one association was significant within type I and none were significant with type III. Although four pairs were significantly correlated within type II strains, type II is a more heterogeneous group than the others (WHITTAM, OCHMAN and SELANDER 1983), and the correlations are most easily explained as resulting from unrecognized substratification in the type II strain.

The same pattern is obtained for pairwise Spearman rank correlations of copy numbers rather than 2 × 2 contingency tables. Eight of the possible 15 pairwise Spearman correlations are significant, and six of the eight are highly significant. (All significant correlations are positive.) However, the number drops to one within type I and to two within type III. In contrast, human isolates by themselves and animal isolates by themselves show about the same correlational structure as the entire data set, so that human *vs.* animal stratification, or the decreased sample sizes of the individual types, do not appear to be factors. Spearman correlations were also calculated for doubly infected strains for the 15 pairs of IS elements. Only three of these were significant, and two of the three became nonsignificant when the analysis was carried out within each type. Curiously, IS*4* and IS*5* are significantly positively correlated in the entire data set, even though IS*4* occurs preferentially in type III and IS*5* occurs preferentially in type I. In general, it appears that the IS elements occur independently among isolates of *E. coli* except for effects secondary to type, at least insofar as this analysis permits one to determine.

Multiple correlations in presence or absence of insertion elements: When multiple correlations are considered, there is a highly significant tendency for strains infected with one IS element to also be infected with other, different IS sequences. Multiple associations were tested by the nonparametric procedures described below. In general, multiple associations of the IS elements were highly significant within all but one stratum of each of the stratification schemes considered above, and generally were significant within all strata. Thus, stratification cannot account for the multiple correlation. Such a phenomenon might result from any combination of the following: multiple simultaneous infection of strains by large conjugative plasmids containing two or more related IS sequences, differential susceptibility to infection by different types of plasmids among the strains, differential susceptibility to infection or transposition of the IS sequences among the strains, or historical accident in which the descendants of a strain that was multiply infected early in its lineage are overrepresented in the sample. Whatever its source or sources, the multiple correlation, while significant, is sufficiently weak so that most pairwise correlations remain nonsignificant.

Multiple associations were first tested by analysis of a 71 × 6 (strain × insertion sequence) matrix of entries either 1 or 0 according to whether or not a particular strain contains a particular insertion sequence. The null hypothesis is that there is no significant variation in the average numbers of 1's among the strains (*i.e.*, no significant association in the presence of unrelated insertion sequences). An appropriate nonparametric statistical test of this hypothesis is the Friedman index test for this array, which in the context of all 0's and 1's is called the Cochran test (LEHMANN 1975). Rejection of the null hypothesis would imply that certain strains contain a greater number of unrelated insertion sequences, and other strains a smaller number, than would be expected by chance.

In standard terminology, the data for the insertion sequences are "blocks" and the data for the strains are "treatments." In our situation the number of blocks is small (six) while the number of treatments is large (71), which is the opposite of what is assumed in the large-sample approximation for the Friedman index. Thus, we used a randomization procedure in which the Cochran statistic for the reference data was calculated and compared with the scores obtained in simulations in which the entries within each of five of the six blocks were randomly permuted. Since only two among 100,000 simulations yielded a Cochran statistic as great as that observed, the null hypothesis can be rejected. In a second test of association the entries in the test matrix were replaced with the within-block ranks or midranks of the actual copy numbers, and the test statistic replaced with the Friedman index. In this case, no simulations among 100,000 yielded a Friedman index as great as the one observed, reinforcing the previous conclusion. Significant or highly significant Cochran and Friedman indexes were also obtained in tests when restricted to strains of type II, strains of type III, strains isolated from human hosts only, and strains from animal hosts. In addition, the Friedman index within type I was significant. Thus, the multiple association observed is unlikely to result from stratification in the sample.

A quantitative measure of this multiple correlation can be obtained as follows. Consider a randomly chosen pair of strains and a randomly chosen pair of IS elements for which the copy numbers of the two strains are distinct. The probability that the strain with the higher copy number for one IS element also has the higher copy number for the other IS element is 61%. This proportion is highly significant as measured by the randomization procedure described above. Within types, the ratio was 57% for type I ($P \simeq 0.02$), 65% for type II ($P < 10^{-3}$), and 58% for type III ($P \simeq 0.02$), all of which are significant.

When the comparison is restricted to doubly infected pairs of strains for two different IS elements with distinct copy numbers, the ratio is 55%, which is not significant ($P \simeq 0.065$). In this sense, there is no significant tendency for large copy numbers of unrelated insertion sequences to occur together in the same strain, apart from the tendency of the IS elements to occur together in the first place.

Models and estimation

Models: We use a continuous-time multitype branching process to model copy numbers of insertion sequences within hosts, where the type of a host is the copy number of a particular insertion sequence (SAWYER and HARTL 1986, Section 4). Uninfected hosts (*i.e.*, type 0) are assumed to transform to type 1 (one copy of the element) at the rate u per generation. Infected individuals of type (or copy number) n change to type $n + 1$ at a rate given by a function $T(n)$ of n. Deletion of IS elements within hosts was ignored since (1) available evidence indicates that deletion occurs at a substantially smaller rate than transposition (EGNER and BERG 1981; FOSTER *et al.* 1981) and (2) adequate fits to the data are obtained without deletion. CANNINGS and GREGORY (1986) have analogous theoretical models with IS deletion included.

Individuals infected with the element are assumed to have either a slower growth rate or a higher death rate (or both) with respect to uninfected individuals. Specifically, we assume that the cell division rate minus the cell death rate for hosts with copy number n is $R - D(n)$, where $D(0) = 0$ and $D(n) > 0$ for $n \neq 0$. As long as (1) $R > u$, (2) uninfected hosts do not become extinct, (3) $T(n) + D(n) - u$ is bounded from below by a positive number, and (4) the infinite series below which defines the normalization constant L converges, then the asymptotic numbers of individuals in the population with copy number n will be independent of R and proportional to

$$\mu(n) = \frac{1}{T(n) + D(n) - u} \prod_{k=1}^{n-1} \frac{T(k)}{T(k) + D(k) - u} \qquad (n \geq 1),$$

with $\mu(0) = 1/u$. In this case the equilibrium frequency of copy number n in the population will be given by

$$\lambda(n) = \mu(n)/L, \qquad L = \sum_{k=0}^{\infty} \mu(k) \qquad (n \geq 0).$$

This is proven in SAWYER and HARTL (1986, Appendix) under the additional assumption that $\{T(n) + \mathrm{D}(n)\}$ is strictly increasing. However, if Theorem 2.2 of SAWYER (1976) is used to estimate the asymptotic covariances of the theoretical copy numbers instead of the arguments in SAWYER and HARTL (1986), it follows that this additional assumption can be removed.

The choices we consider for $T(n)$ are $T(n) = T$ (constant), $T(n) = Tn^{1/2}$ (root), $T(n) = T/n$ (harmonic), $T(n) = T/n^{1/2}$ (inverse root), $T(n) = Tn$

(linear), and $T(n) = Tn^2$ (quadratic). Given one of these choices for $T(n)$, and an analogous choice for $D(n)$ (*i.e.*, $D(n) = D$, $D(n) = Dn^{1/2}$, and so forth), the asymptotic frequency distribution $\{\lambda(n)\}$ defined above depends on two independent parameters, which we can take as u/D and T/D.

SAWYER and HARTL (1986) also consider a class of models in which individuals that die are replaced by uninfected individuals, thus holding the total population size constant. These models will not be pursued here. In the branching process model, individuals that die are not replaced, which is more-or-less equivalent to replacing them by individuals chosen at random from the rest of the population.

Model fitting: We examined nine models with various different choices for the rates of transposition and death. It is convenient to refer to these models in terms of the functional dependence of transposition and death rates on n, namely constant (C), root (R), harmonic (H), inverse root (D), linear (L), and quadratic (Q). Using this notation, with the rate of transposition designated first and the death rate second, the nine models are CC, HC, CR, RC, RR, DR, RL, LL and LQ.

For each IS element, the strains were grouped into six classes according to whether the number of IS elements was 0, 1, 2, 3, 4 or ≥5. Maximum likelihood estimates for the parameters u/D and T/D were calculated for each of the nine models for each IS element for both grouped and ungrouped data. Estimates based on the grouped data do not use the tails of copy number distributions, but permit a standard test of goodness of fit (the ungrouped data has too many empty cells). Table 4 compares the theoretical and observed distributions for the grouped data for each IS element for the models CC, RL, RR, CR, LL and LQ, except that RL and RR are replaced by HC and DR for IS*3*. The right-most column of Table 4 contains the nominal Fisher chi square goodness-of-fit P-values for the fitted distributions for the grouped data, which in this case has three degrees of freedom. The models which significantly do not fit are flagged with an asterisk. The quality of fit for the ungrouped data seemed heuristically about the same as that for the grouped data, although the goodness-of-fit test could not be used.

The models were also compared directly by means of the log likelihood scores for the ungrouped data (see Table 5). In part A of Table 5 the models for each IS element are ranked according to their log likelihoods. The lower part of each table gives the difference between these log likelihoods for each IS element and that of the rank one model, in the order of the corresponding ranks. Asterisks indicate the models that could be rejected for that insertion sequence using the goodness of fit test for the grouped data (see Table 4). The two right-most columns give the mean of the within-IS ranks of the models, and the within-model ranks or midranks of those scores, for all IS elements with IS*3* excluded or included, respectively. The corresponding columns in the lower part give the differences between the sum of the log likelihoods and that of the best model, in the order of the sum of the log likelihoods. On the basis of the sum of the log likelihoods for all six IS elements, CR was the best model for the grouped data and LQ was the best model for the ungrouped data.

Although there is no direct way to test goodness of fit for the ungrouped data, models with relatively high log likelihoods can be rejected by the following test. Model fitting with the ungrouped data can be considered within the context of a more general model with $T(n) = Tn^a$, $D(n) = Dn^b$, and four parameters u/D, T/D, a, and b. Suppose that a particular model (*i.e.*, particular values of a and b) is assumed. If the log likelihood for a particular IS element is maximized by varying all four parameters (including a and b), and the result compared with the log likelihood maximized over u/D and T/D with a and b fixed at their assumed values, then twice the difference between the two log likelihood scores should be approximately a chi square random variable with two degrees of freedom (RAO 1973). Thus twice the difference in fitted log likelihood scores between a given model and the highest ranking model for that IS element can conservatively be compared with a chi square distribution with two degrees of freedom. In particular, a difference in log likelihoods of 3.00 is significant ($P = 0.05$), and 4.61 is highly significant ($P = 0.01$). Note that every model in Table 5 which is rejected by goodness of fit for the grouped data is also rejected by this test, so that the ungrouped data provides greater discrimination between the models. In addition, the essential pairwise independence of the IS elements makes it unlikely that estimates based on the ungrouped data are overly influenced by a few outliers with high copy numbers.

The results of the model fitting group the six IS elements into three categories, which may represent the effects of differing biological mechanisms. We now discuss the results for these three categories separately.

IS*1* and IS*5*: Using the grouped data, only HC and DR can be rejected for IS*1* and IS*5*. However, when the tail of the copy number distribution for IS*1* is taken into account, all models are rejected except LL, LQ, and the biologically unreasonable model RC (Table 5). Only HC, DR and CR can be rejected for the ungrouped data for IS*5*, but models of the type CC, RL or RR tend not to fit well ($P < 0.09$ for CC and RL). On the basis of the ungrouped data, the models LL and LQ are the two best for IS*1* and are in the top four (with good fits) for IS*5*.

TABLE 4

Fitted distributions for *grouped* data

	Copies						
	(0)	(1)	(2)	(3)	(4)	(≥5)	(PVAL)
For IS*1*:							
Observed:	11	14	8	6	7	25	
CC:	11.00	11.93	9.56	7.66	6.14	24.72	0.778
RL:	11.62	10.68	9.26	8.10	6.92	24.41	0.616
RR:	11.28	11.94	9.29	7.56	6.18	24.75	0.807
CR:	12.09	8.73	9.44	9.13	8.10	23.51	0.186
LL:	10.83	14.37	8.89	6.48	5.01	25.41	0.818
LQ:	11.00	13.42	8.92	6.91	5.62	25.13	0.901
For IS*2*:							
Observed:	28	8	12	5	5	13	
CC:	28.00	10.84	8.11	6.06	4.53	13.45	0.413
RL:	27.75	11.02	8.02	6.19	4.78	13.24	0.385
RR:	27.67	11.89	7.83	5.67	4.23	13.70	0.289
CR:	28.03	9.26	8.44	7.08	5.58	12.61	0.501
LL:	27.39	13.88	7.34	4.85	3.49	14.05	0.102
LQ:	27.41	13.25	7.58	5.32	4.00	13.45	0.175
For IS*3*:							
Observed:	23	10	19	10	6	3	
CC:*	23.00	18.95	11.47	6.94	4.20	6.44	0.004
HC:	23.00	13.15	14.99	10.55	5.60	3.71	0.568
DR:	22.52	13.77	14.76	10.54	5.69	3.72	0.485
CR:	22.06	16.91	13.18	8.70	5.08	5.06	0.085
LL:*	21.55	22.26	10.60	6.04	3.69	6.86	0.000
LQ:*	21.60	21.06	11.52	7.01	4.26	5.55	0.003
For IS*4*:							
Observed:	43	5	5	3	5	10	
CC:	43.00	6.00	4.71	3.70	2.91	10.67	0.602
RL:	42.82	6.59	4.60	3.56	2.83	10.61	0.530
RR:	42.83	6.97	4.51	3.31	2.54	10.84	0.380
CR:	42.95	5.34	4.81	4.13	3.41	10.36	0.778
LL:	42.69	8.37	4.25	2.80	2.04	10.85	0.118
LQ:	42.65	8.23	4.40	3.01	2.27	10.44	0.200
For IS*5*:							
Observed:	46	12	3	2	2	6	
CC:	46.00	8.48	5.60	3.70	2.45	4.76	0.278
RL:	46.27	8.31	5.27	3.61	2.49	5.05	0.307
RR:	46.16	9.02	5.14	3.30	2.21	5.17	0.468
CR:	46.28	7.43	5.80	4.19	2.84	4.46	0.108
LL:	46.12	9.94	4.64	2.80	1.87	5.63	0.737
LQ:	46.21	9.17	4.73	3.08	2.16	5.66	0.590
For IS*30*:							
Observed:	36	16	13	2	2	2	
CC:	36.00	18.05	8.74	4.23	2.05	1.93	0.322
RL:	35.93	17.76	8.99	4.49	2.14	1.68	0.332
RR:	35.87	18.63	8.32	4.06	2.03	2.09	0.257
CR:	36.04	16.78	9.80	4.87	2.15	1.35	0.377
LL:	35.76	19.65	7.59	3.64	1.91	2.45	0.147
LQ:	35.86	18.50	8.40	4.26	2.14	1.83	0.253

* Goodness of fit test is significant at level 0.05.

If either the LL or the LQ model is accepted for IS*1* or IS*5*, then there are some interesting biological implications. When the rate of transposition is given by $T(n) = nT$ (L), then the probability of transposition *per element* in the genome is $T(n)/n = T$, a constant, irrespective of the number of elements. This means that transposition of these elements is unregulated. The models LL and LQ imply a moderate to strong detrimental effect of insertion sequences on fitness, with $D(n) = nD$ (LL model) or $D(n) = n^2D$ (LQ model).

IS*3*: The IS*3* element is anomalous in that the only three models (HC, DR and CR) that fit the grouped data do quite poorly (especially HC and DR) when applied to the other insertion elements. The distribution of copy numbers of IS*3* has the odd feature of being strongly bimodal with modes at $n = 0$ and at $n = 2$, which most models have difficulty emulating.

TABLE 5

A. Ranks of log likelihoods of *ungrouped* data

Model	IS*1*	IS2	IS*3*	IS*4*	IS*5*	IS*30*	Mean (excluding IS*3*)	Mean (all)
CC	5.0	1.0	5.0*	2.0	5.0	5.0	3.60 (1)	3.83 (1.5)
HC	9.0*	9.0	1.0	9.0	9.0*	4.0	8.00 (9)	6.83 (9)
CR	7.0	7.0	3.0	3.0	7.0	1.0	5.00 (6)	4.67 (5)
RC	3.0	6.0	9.0*	7.0	1.0	9.0	5.20 (7)	5.83 (7)
RR	4.0	2.0	7.0*	4.0	3.0	7.0	4.00 (3.5)	4.50 (4)
DR	8.0*	8.0	2.0	8.0	8.0*	3.0	7.00 (8)	6.17 (8)
RL	6.0	4.0	4.0*	1.0	6.0	2.0	3.80 (2)	3.83 (1.5)
LL	1.0	5.0	8.0*	6.0	2.0	8.0	4.40 (5)	5.00 (6)
LQ	2.0	3.0	6.0*	5.0	4.0	6.0	4.00 (3.5)	4.33 (3)

B. Differences in log likelihood in order of ranks, for each IS and for sums

1	0.00	0.00	0.00	0.00	0.00	0.00	0.00 (LQ)	0.00 (LQ)
2	0.39	0.09	0.23	0.14	0.04	0.26	1.93 (LL)	3.07 (RR)
3	0.91	0.34	2.50	0.27	1.15	0.27	2.78 (RR)	3.19 (CC)
4	3.60	0.47	4.48*	0.60	1.24	0.43	3.46 (CC)	5.31 (LL)
5	3.86	1.35	6.19*	0.91	2.42	0.46	4.39 (RC)	5.72 (RL)
6	7.52	2.22	6.46*	2.45	2.87	0.53	7.70 (RL)	10.22 (RC)
7	17.72	2.37	6.75*	2.66	5.95	0.75	22.89 (CR)	18.93 (CR)
8	40.49*	9.95	9.85*	3.25	14.44*	1.50	64.98 (DR)	58.75 (DR)
9	44.69*	12.60	12.29*	5.77	18.53*	2.02	78.61 (HC)	72.15 (HC)

* Goodness of fit test with grouped data results in rejection of model at level of significance 0.05.

However, the HC and DR models fit very well. Biologically, $T(n) = T/n$ (H) or $T(n) = T/n^{1/2}$ (D) imply a quite strong regulation of transposition, with the rate of transposition per element being proportional to $1/n^2$ or $1/n^{3/2}$, respectively. At the same time, the assumption that $D(n) = D$ (HC model) or $D(n) = Dn^{1/2}$ (DR model) implies at most a mild effect of the insertion element on fitness.

IS2, IS*4* and IS*30*: These three insertion elements are similar in that they are very amenable to model fitting. No models can be rejected for the grouped data, and only HC and DR are inconsistent for IS2 and IS*4* with the ungrouped data. The best models for the ungrouped data are CC and RR (IS2), RL and CC (IS*4*), and CR and RL (IS*30*). There seems to be little basis for choosing among the models CC, RL and RR. These models share in the assumption that transposition is regulated, with the rate of transposition per element proportional to $1/n$ in the CC and CR models, which may be regarded as moderate regulation, and $1/n^{1/2}$ in the RL and RR models, which may be regarded as weak regulation. The models also incorporate a dependence of fitness on the number of elements, either nil (CC), weak (RR and CR) or moderate (RL). However, the model LQ also provides a good fit for the ungrouped data for these three distributions.

A unitary model: For grouped data, the models CC, RL and RR fit all of the IS elements except IS*3* quite well ($P > 0.25$ in all cases). Although CR does not do quite as well for IS*1* and IS*5* ($P < 0.25$) on grouped data, it is the only model that provides an acceptable fit for the grouped data for all six insertion sequences. The CR model also provides the best fit for IS2, IS*4* and IS*30* on the basis of P-values for grouped data. Its behavior for ungrouped data is respectable for all IS elements except for IS*1* and IS*5*, on which it fares quite badly. The CR model combines a moderate regulation of transposition, with the rate of transposition per element proportional to $1/n$, and a weak effect of number of elements on fitness, with $D(n) = Dn^{1/2}$.

There is no comparable unitary model for the ungrouped data. The same conservative chi square test (with two degrees of freedom) also applies to the sum of log likelihoods across models. The only models that fit the ungrouped data for all IS elements except IS*3* under this criterion are LQ, LL and RR, all of which are decisively rejected by IS*3*.

Parameter estimation: Table 6 contains some typical maximum likelihood estimates and 95% confidence intervals for u/D and T/D. Estimates are included for the ungrouped data for the six models CC, RL, RR, CR, LL and LQ (RL and RR are replaced by HC and DR for IS*3*). The 95% confidence intervals are valid if the underlying model fits. Estimates are also included for u/AvD, where AvD is per-host death rate averaged over the host population. These estimates were included in order to have a measure of the balance between the infection rate u and mean IS-induced death rate (or rate of growth inhibition) which would be comparable over different models.

TABLE 6

The 95% Confidence Intervals for *ungrouped data*

	First row: u/D		Second row: u/AvD		Third row: T/D	
	IS*1*	IS2	IS*3*[a]	IS*4*	IS*5*	IS*30*
CC:	0.85 ± 0.08	0.61 ± 0.11	0.68 ± 0.11	0.39 ± 0.11	0.35 ± 0.11	0.49 ± 0.12
	0.85 ± 0.08	0.61 ± 0.11	0.68 ± 0.11	0.39 ± 0.11	0.35 ± 0.11	0.49 ± 0.12
	1.01 ± 0.61	1.38 ± 0.61	0.48 ± 0.24	2.47 ± 1.12	1.53 ± 0.76	0.45 ± 0.24
RL:	6.66 ± 1.62	2.75 ± 0.85	0.68 ± 0.11	1.98 ± 0.85	1.26 ± 0.58	0.93 ± 0.29
	0.88 ± 0.22	0.61 ± 0.19	0.68 ± 0.11	0.39 ± 0.17	0.37 ± 0.17	0.49 ± 0.16
	22.73 ± 8.38	10.43 ± 4.75	0.87 ± 0.44	15.10 ± 8.75	8.84 ± 5.69	1.75 ± 1.00
RR:	2.05 ± 0.34	1.17 ± 0.27	1.03 ± 0.18	0.81 ± 0.27	0.59 ± 0.21	0.65 ± 0.17
	0.86 ± 0.14	0.60 ± 0.14	0.67 ± 0.12	0.39 ± 0.13	0.36 ± 0.13	0.49 ± 0.13
	4.59 ± 1.55	3.18 ± 1.30	1.74 ± 0.72	4.76 ± 2.41	2.87 ± 1.60	0.86 ± 0.46
CR:	2.12 ± 0.34	1.20 ± 0.27	1.03 ± 0.18	0.83 ± 0.27	0.62 ± 0.22	0.65 ± 0.17
	0.89 ± 0.14	0.62 ± 0.14	0.67 ± 0.12	0.40 ± 0.13	0.38 ± 0.13	0.49 ± 0.12
	11.48 ± 3.75	5.60 ± 2.22	1.35 ± 0.56	8.44 ± 4.04	4.98 ± 2.69	1.01 ± 0.53
LL:	6.36 ± 1.98	2.73 ± 1.04	1.70 ± 0.49	2.01 ± 1.08	1.18 ± 0.63	0.93 ± 0.32
	0.84 ± 0.26	0.61 ± 0.23	0.69 ± 0.20	0.40 ± 0.21	0.35 ± 0.19	0.50 ± 0.17
	11.29 ± 4.85	6.98 ± 3.74	2.29 ± 1.11	10.19 ± 7.23	5.65 ± 4.22	1.57 ± 0.99
LQ:	102.68 ± 59.7	22.16 ± 13.6	5.65 ± 2.63	17.14 ± 14.2	9.54 ± 8.32	2.45 ± 1.38
	0.84 ± 0.49	0.61 ± 0.37	0.74 ± 0.34	0.41 ± 0.34	0.33 ± 0.29	0.51 ± 0.29
	231.76 ± 156	71.59 ± 50.5	10.33 ± 6.0	102.66 ± 95.5	62.98 ± 62.7	6.04 ± 4.50

[a] For IS*3*, RL estimates are replaced by HC and RR estimates by DR.

The estimates for T/D are highly variable across IS elements and are also highly model dependent. In contrast, the estimates for u/AvD vary over a comparatively narrow range for all models and all IS elements. One might infer from this that the six IS elements have somewhat similar infection rates and costs to their hosts, and that the difference in their distributions may be principally due to the difference in their transposition mechanisms (Sawyer and Hartl 1986).

DISCUSSION

Some pairs of strains in the ECOR collection differ in the electrophoretic mobility of zero or one enzyme among 35 examined, but demonstrate many differences in the sizes of restriction fragments that hybridize with insertion sequence probes. Strains that differ in no enzymes share an average of 43% of their insertion sequence bands, as compared with 28% among strains differing in one enzyme. Our previous suggestion (Green *et al.* 1984) that the number and sizes of DNA restriction fragments containing insertion sequences could be used to distinguish among closely related strains has thus been verified.

In virtually every case of closely related strains, including strains that are indistinguishable in the electrophoretic mobility of all 35 enzymes, the strains were found to differ in the number of elements, or in the sizes of element-containing DNA fragments, with respect to at least one insertion sequence. These results are consistent with the hypothesis that insertion sequence numbers and positions in the genome change much more rapidly in evolution than do the electrophoretic mobilities of proteins. Consequently, insertion sequence comparisons may be useful in determining the detailed phylogenetic history of a group of strains known to be related because of identity or close similarity in their electrophoretic type. Such an approach holds promise as a method of highly precise strain identification in bacterial epidemiology and could also be used for strain identification in monitoring the survival or spread of genetically engineered bacteria that are released into the environment. A second application of insertion sequence comparisons is in the detection of recombinant strains. Strains that are similar in insertion sequences but differ in several electrophoretic alleles may have originated by means of recombination between two clonal types, as suggested in our data for ECOR strain 39.

The ECOR reference collection is stratified in several ways, mainly with respect to type I *vs.* type II *vs.* type III, but also human *vs.* animal origin, pathogenic *vs.* nonpathogenic isolates, and North American *vs.* Swedish sources. Presence of certain of the insertion sequences is significantly correlated with this stratification. For example, IS*1*, IS*3* and IS*30* are significantly associated with type I strains, IS*4* with type III strains, IS*4* with strains from humans, and IS*3* with strains from animals. There is no significant stratification with respect to North American *vs.* Swedish or pathogenic *vs.* nonpathogenic sources, except that IS*3* favors North American and IS2 favors Swedish sources at a level of significance of 10%. Overall, there are significant correlations between many of the insertion elements (all significant pairwise correlations are positive), but most vanish when the correlations are calculated within types. Hence, the correlations appear to be secondary to the association of certain insertion elements with certain types of strains. Any heterogeneous collection of *E. coli* strains would be expected to show significant pairwise correlations between insertion elements as a result of recognized or unrecognized stratification in the sample. However, there is a highly significant positive multiple correlation among the insertion sequences taken as a group which persists even within the obvious strata in the ECOR collection.

Nine different related branching process models were used to match the observed copy number distributions of the IS elements. With the exception of IS*1* and IS*3*, most insertion sequence distributions were consistent with most of these simple models. The elements IS*1* and IS*5* have distributions that are most compatible with the LL or LQ models. These models imply a lack of regulation of transposition and a moderate to strong detrimental effect of copy number on fitness. The copy number distribution of IS*1* is consistent only with the models LL, LQ, and the biologically unreasonable RC. Absence of regulation in the transposition of IS*1* might account for the very large number of copies of this and a related element found in isolates of *Shigella* (OHTSUBO *et al.* 1981).

The distribution of IS*3* was more compatible with models with much stronger regulation and less of a detrimental effect of copy number on fitness. Two models (HC and DR) fit the distribution of IS*3* quite well; most of the other models were rejected by goodness of fit tests. The HC model, which fits the best, implies a strong regulation of transposition and fitnesses that are unaffected by increasing copy number. The DR model combines a slightly weaker regulation of transposition with a weak effect of copy number on fitness.

Four models (CC, RL, RR and LQ) fit the distributions of IS*2*, IS*4* and IS*30* quite well. The models CC and RL were the two best models in terms of the sum of the ranks of the log likelihoods for ungrouped data with IS*3* either included or excluded. (LQ was the best model in terms of the sums of the log likelihoods themselves.) The models CC, RL and RR combine weak to moderate regulation of transposition with nil to moderately harmful effects of copy number on fitness.

The CC model is the simplest model in that both $T(n)$ and $D(n)$ are constants, and it is the best model in terms of the sum of the ranks of the log likelihoods excluding IS*3*. In the CC model, the probability of transposition per element is proportional to $1/n$ (moderate regulation), and fitness is independent of the number of copies of the element. However, the CC model is rejected in the case of IS*3* ($P < 0.005$). Indeed, only one model (CR) is acceptable with respect to all six insertion sequences. This model has a moderate regulation of transposition and a weak effect of copy number on fitness (reduction in fitness proportional to $n^{1/2}$).

The results with model fitting serve to emphasize the fact that many two-parameter distributions are very pliable and can be bent to fit a variety of empirical distributions. Thus, the fitting of models to the distributions of numbers of transposable elements cannot be expected to be a powerful tool for revealing details as to the mechanisms of transposition or effects on fitness. However, considering the types of models that fit the data, it seems reasonable to conclude that, with the possible exception of IS*1*, and to a lesser extent IS*5*, transposition is mildly regulated and that harmful effects on fitness increase mildly, if at all, with increasing copy number. This conclusion suggests that detailed experiments to study the regulation and fitness effects of individual insertion sequences would be of considerable interest.

This work was supported by National Institutes of Health grant GM30201 (D.E.D. and D.L.H.), National Science Foundation grant DMS-8504315 (S.A.S.) and National Institutes of Health Training Grant GM08036 (R.F.D.).

LITERATURE CITED

ACHTMAN, M., S. A. SKURRAY, R. THOMPSON, R. HELMUTH, S. HALL, L. BEUTIN and A. J. CLARK, 1978 Assignment of *tra* cistrons to *Eco*RI fragments of F sex factor DNA. J. Bacteriol. **133:** 1383–1392.

BIEL, S. W., G. ADELT and D. E. BERG, 1984 Transcriptional control of IS*1* transcription in *Escherichia coli.* J. Mol. Biol. **174:** 251–264.

BOTSTEIN, D. and D. SHORTLE, 1985 Strategies and applications of in vitro mutagenesis. Science **229:** 1193–1201.

CALOS, M. P. and J. H. MILLER, 1980 Transposable elements. Cell **20:** 579–595.

CANNINGS, C. and K. GREGORY, 1986 Birth-death-immigration-catastrophe processes, as models for transposon copy number. Theor. Pop. Biol. In press.

CIAMPI, M. S., M. B. SCHMID and J. R. ROTH, 1982 Transposon Tn*10* provides a promoter for transcription of adjacent sequences. Proc. Natl. Acad. Sci. USA **79:** 5016–5020.

DALRYMPLE, B., P. CASPERS and W. ARBER, 1984 Nucleotide sequence of the prokaryotic mobile genetic element IS*30*. EMBO J. **3:** 2145–2149.

DOOLITTLE, F. W. and C. SAPIENZA, 1980 Selfish DNA, the phenotype paradigm and genome evolution. Nature **284:** 601–603.

DYKHUIZEN, D. E., S. A. SAWYER, L. GREEN, R. D. MILLER and D. L. HARTL, 1985 Joint distribution of insertion elements IS*4* and IS*5* in natural isolates of *Escherichia coli.* Genetics **111:** 219–231.

EGNER, C. and D. E. BERG, 1981 Excision of transposon Tn*5*. Proc. Natl. Acad. Sci. USA **78:** 459–463.

FOSTER, T. J., V. LUNDBLAD, S. HANLEY-WAY, S. M. HALLING and N. KLECKNER, 1981 Three Tn*10*-associated excision events:

relationship to transposition and role of direct and inverted repeats. Cell **23:** 215–227.

Green, L., R. D. Miller, D. E. Dykhuizen and D. L. Hartl, 1984 Distribution of DNA insertion element IS*5* in natural isolates of *Escherichia coli.* Proc. Natl. Acad. Sci. USA **81:** 4500–4504.

Hartl, D. L., D. E. Dykhuizen, R. D. Miller, L. Green and J. de Framond, 1983 Transposable element IS*50* improves growth rate of *E. coli* cells without transposition. Cell **35:** 503–510.

Heffron, F., 1983 Tn*3* and its relatives, pp. 223–260. In: *Mobile Genetic Elements,* Edited by J. Shapiro. Academic Press, New York.

Kado, C. I. and S. T. Liu, 1981 Rapid procedure for detection and isolation of large and small plasmids. J. Bacteriol. **145:** 1365–1373.

Klaer, R., S. Kuhn, E. Tillmann, H.-J. Fritz and P. Starlinger, 1981 The sequence of IS*4*. Mol. Gen. Genet. **181:** 169–175.

Lehmann, E. L., 1975 *Nonparametrics: Statistical Methods Based on Ranks,* p. 262. Holden-Day, Oakland.

Miller, R. D., D. E. Dykhuizen, L. Green and D. L. Hartl, 1984 Specific deletion occurring in the directed evolution of 6-phosphogluconate dehydrogenase in *Escherichia coli.* Genetics **108:** 765–772.

Ochman, H. and R. K. Selander, 1984 Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157:** 690–693.

Ohtsubo, H., K. Nyman, W. Doroszkiewicz and E. Ohtsubo, 1981 Multiple copies of iso-insertion sequences of IS*1* in *Shigella dysenteriae* chromosome. Nature **292:** 641–643.

Orgel, L. E. and F. H. C. Crick, 1980 Selfish DNA: the ultimate parasite. Nature **284:** 604–607.

Rao, C. R., 1973 *Linear Statistical Inference and Its Applications,* Ed. 2, p. 417. John Wiley & Sons, New York.

Reed, R. R., 1981 Resolution of cointegrates between transposons gamma-delta and Tn*3* defines the recombination site. Proc. Natl. Acad. Sci. USA **78:** 3428–3432.

Reynolds, A. E., J. Felton and A. Wright, 1981 Insertion of DNA activates the cryptic *bgl* operon in *E. coli* K12. Nature **293:** 625–629.

Rigby, P. W. J., M. Dieckmann, C. Rhodes and P. Berg, 1977 Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase. I. J. Mol. Biol. **113:** 237–251.

Saedler, H., G. Cornelis, J. Cullum, B. Schumacher and H. Sommer, 1980 IS*1*-mediated DNA rearrangements. Cold Spring Harbor Symp. Quant. Biol. **45:** 93–98.

Sawyer, S., 1976 Branching diffusion processes in population genetics. Adv. Appl. Prob. **8:** 659–689.

Sawyer, S. and D. L. Hartl, 1986 Distribution of transposable elements in prokaryotes. Theor. Pop. Biol. **30:** 1–16.

Selander, R. K., D. A. Caugant and T. S. Whittam, 1986 Genetic structure and variation in natural populations of *Escherichia coli.* In: *Escherichia coli and Salmonella typhimurium, Cellular and Molecular Biology,* Edited by J. L. Ingraham. American Society for Microbiology, Washington, D.C. In press.

Southern, E. M., 1975 Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98:** 503–517.

Stokes, H. W. and B. G. Hall, 1984 Topological repression of gene activity by a transposable element. Proc. Natl. Acad. Sci. USA **81:** 6115–6119.

Whittam, T. S., H. Ochman and R. K. Selander, 1983 Multilocus genetic structure in natural populations of *Escherichia coli.* Proc. Natl. Acad. Sci. USA **80:** 1751–1755.

Communicating editor: B. S. Weir

Appendix

Insertion Sequences in ECOR Strains of E. coli

	Chromosomal copies						Plasmid copies					
ECOR	IS*1*	IS*2*	IS*3*	IS*4*	IS*5*	IS*30*	IS*1*	IS*2*	IS*3*	IS*4*	IS*5*	IS*30*
1	4	0	1	0	1	1	3	0	0	0	1	0
2	2	1	2	0	0	1	1	2	0	0	0	0
3	4	0	2	1	5	2	2	1	0	0	0	0
4	18	2	2	0	0	0	0	0	0	0	0	0
5	1	0	2	0	2	0	0	0	0	0	0	0
6	24	8	2	0	0	0	0	0	0	0	0	0
7	2	0	2	0	0	0	3	1	0	0	0	0
8	4	0	2	0	9	4	0	0	3	0	1	0
9	17	0	0	0	0	2	1	1	0	0	0	1
11	6	11	1	4	2	2	0	0	0	0	1	0
12	1	17	2	4	5	3	0	0	0	0	0	0
13	16	9	6	1	4	0	2	0	0	1	1	0
14	19	0	4	4	1	5	0	0	0	0	0	0
15	15	0	4	0	0	2	1	0	0	0	0	0
16	17	12	4	0	0	0	1	0	0	0	0	0
17	20	1	5	0	0	1	0	0	0	0	0	0
18	27	0	1	0	1	0	1	0	0	0	1	0
19	25	3	2	0	6	1	0	0	0	0	0	0
20	20	3	2	0	1	2	5	0	0	0	1	2
21	20	2	2	0	1	2	5	0	0	0	1	2
22	6	0	2	4	2	0	2	0	0	0	0	0
23	10	6	6	1	21	4	11	1	1	0	1	1
24	2	1	0	0	0	1	0	0	0	0	0	0
25	3	3	2	0	0	2	0	2	1	0	0	0
26	0	0	2	0	0	0	0	0	0	0	0	0
27	0	0	2	0	0	0	0	0	2	0	0	0

Appendix *Continued*

	Chromosomal copies						Plasmid copies					
ECOR	IS*1*	IS*2*	IS*3*	IS*4*	IS*5*	IS*30*	IS*1*	IS*2*	IS*3*	IS*4*	IS*5*	IS*30*
28	1	0	3	0	0	0	0	0	1	0	0	0
29	2	0	1	0	0	0	0	0	1	0	0	0
30	1	0	3	0	1	0	3	1	0	0	0	0
31	7	2	0	0	3	2	2	0	2	1	0	0
32	1	2	3	0	0	1	0	0	0	0	0	0
33	1	2	3	0	0	1	0	0	0	0	0	0
34	1	2	2	0	0	0	2	1	0	1	0	0
35	26	6	0	13	1	2	1	0	0	0	1	0
36	25	6	0	14	3	2	1	0	0	0	1	0
37	7	0	0	0	0	0	4	0	1	0	0	0
38	6	4	0	8	0	1	0	0	0	0	1	0
39	4	4	0	12	1	1	0	0	0	0	0	0
40	4	4	0	11	0	0	0	0	0	0	3	0
41	5	3	0	11	0	1	0	0	0	0	0	0
42	2	0	2	0	0	0	0	0	1	0	0	0
43	3	1	1	0	0	5	0	1	0	1	0	1
44	6	11	0	0	0	0	0	1	0	0	0	0
45	0	0	2	0	0	0	0	0	0	0	0	0
46	0	2	1	0	1	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	1	0	0	0	0	0
49	4	13	0	0	2	3	0	0	0	0	1	0
50	6	14	1	11	4	2	0	0	0	0	1	0
51	1	2	3	1	0	1	0	0	0	0	0	0
52	1	3	3	2	0	1	0	0	0	0	0	0
53	1	0	3	4	0	0	0	0	0	0	0	0
54	1	0	3	2	0	0	0	0	0	0	0	0
55	2	8	0	2	0	2	0	0	0	0	0	0
56	1	1	0	3	0	1	0	0	0	0	0	0
57	2	4	1	0	1	1	0	0	0	0	0	0
58	0	0	4	0	0	0	0	0	0	0	0	0
59	3	0	0	0	0	0	0	0	0	0	0	1
60	1	2	3	0	0	0	0	0	0	0	0	0
61	3	4	0	5	1	0	0	0	0	0	0	0
62	3	5	0	5	1	0	4	2	0	0	0	0
63	0	1	3	2	0	0	0	0	0	0	0	0
64	4	2	0	3	0	0	0	0	0	0	0	0
65	0	0	1	2	0	0	0	0	0	0	0	0
66	0	1	0	0	0	0	0	0	0	0	0	0
67	0	0	0	0	0	0	0	0	0	0	0	0
68	1	2	4	0	0	0	0	0	0	0	0	0
69	6	0	4	1	0	1	0	0	0	0	0	1
70	2	2	1	8	0	0	0	0	0	0	0	0
71	22	1	2	3	6	1	2	1	0	1	0	0
72	3	0	0	0	0	0	0	2	0	0	0	0