Repeated Sequences

SOPHIE BACHELLIER, ERIC GILSON, MAURICE HOFNUNG, AND CHARLES W. HILL

112

INTRODUCTION

Sequence repetition in *Escherichia coli* and *Salmonella typhimurium* (official designation, *Salmonella enterica* serovar Typhimurium) is encountered in many contexts. Sources of repeated sequences include structural genes, accessory genetic elements, sequence motifs, and genetic duplications. On a random basis, there is a high probability that any sequence of 12 or fewer bases will be repeated at least once in a chromosome this size. Therefore, for a sequence repetition to be considered significant, it must be either substantially larger than 12 bases or present in many copies.

Certain genetic issues relate especially to repeated sequences. One issue has to do with advantages conferred by sequence multiplicity. In the case of structural genes, benefits might include greater expression or differential control. Benefits may also be derived from functional diversity if repeated genes are similar but not identical. In the case of sequence motifs, the same structural domain can be incorporated into diverse gene products or the motif can invoke the same process at many chromosomal locations.

A second issue has to do with the source of the DNA comprising the repeated sequences and also the mechanism for generation of the repetition. As to source, we generally recognize two alternatives. The repeated sequences may be DNA common to the species, or they may derive from DNA introduced through horizontal transfer (e.g., accessory genetic elements [23]). The repeated sequences may have been established early with respect to species divergence, in which case the arrangement would tend to be conserved among individuals throughout the species (e.g., rRNA genes). Alternatively, the sequences may have become amplified by recent rearrangement (e.g., tandem duplications), and such arrangements can be highly specific for individual strains or clones. Repeated sequences derived from accessory elements (e.g., insertion sequences [ISs]) can vary considerably as to numbers and chromosomal positions in individuals of the species. As to possible mechanisms for establishment of sequence repetition, illegitimate, site-specific, and homologous recombination all might have roles in different circumstances, and at least in the case of very small repeated sequences, convergent evolution also should be considered.

A third issue concerns genetic interaction between repeated sequences. Repeated sequences provide homology for ectopic recombination, i.e., recombination between dispersed homologous sequences. This will lead to chromosomal rearrangement such as duplication, deletion, inversion, and transposition (Fig. 1). The nature of the rearrangement will depend on the positions and orientations of the repeated sequences and also on whether the repeated sequences are on the same or different chromatids. The schemes drawn in Fig. 1 imply reciprocal crossover events, but nonreciprocal events may happen as well. Specifically in the case of direct repeats, nonreciprocal crossover may leave one product rejoined but the other a nonviable, incomplete fragment. Recombination between inverted repeats, however, must be reciprocal if a circular chromosome is to be recovered. Each type of rearrangement has distinct consequences, with duplication and deletion obviously affecting dosage. Inversion changes gross positional relationships between genes. This may be especially important when the position of a gene relative to the replication origin and terminus is altered. The schemes depicted in Fig. 1 do not exhaust the configurations of possible interactions. For example, the small circle excised in Fig. 1A could reinsert, either into the homologous portion of a different chromosome (yielding a tandem duplication) or into a third copy of the repeated sequence (yielding a transposition). Figure 1D depicts the

consequences of a crossover between inverted repeats on sister chromatids. The product is a complex chromosomal dimer that cannot be resolved to monomers by a single additional crossover between the large homologs (i.e., region 4 5 6 with 4 5 6 or region 1 2 3 with 1 2 3) since the large homologs are themselves inverted. It could be resolved by the less frequent crossover between appropriate copies of the repeated sequence (i.e., p [proximal] with d [distal] or p/d with p/d). This dimer would have two replication origins, and a cell containing it would stably inherit two copies of every gene. The genetic properties of such a cell would be striking, and the fact that such a mutant has not been reported likely indicates that a cell harboring such a chromosome is not viable.
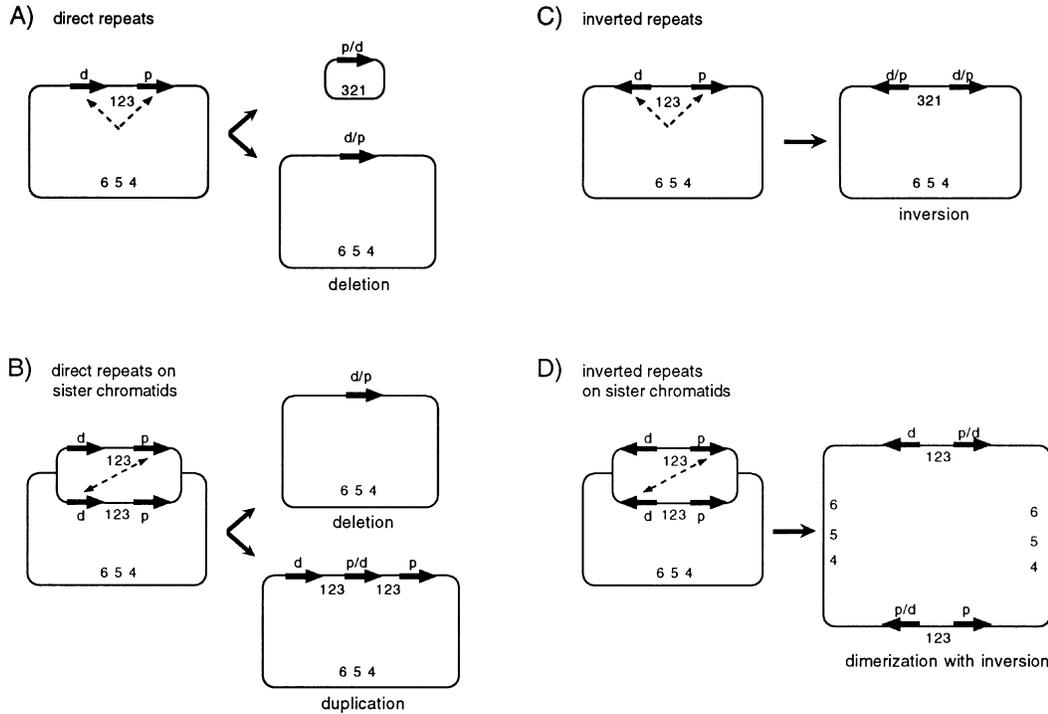

FIGURE 1 Crossover between sequence repeats.These schemes represent fundamental intersections between repeated sequences,and no mechanism is implied.The repeats are shown by the arrows:d (distal) and p (proximal).They may occur as direct repeats (A and B) or as inverted repeats (C and D).The crossover may be between repeats present on the same chromosome (A and C) or on sister chromatids (B and D).For schemes B and D ,the structures of the ultimate recombinant products assume completion of the replication cycle.The unique sequences separating the repeats are designated 1 2 3 and 4 5 6 ;1 2 3 may be deleted,duplicated,or inverted with respect to 4 5 6,depending on the configuration of the event.The repeated sequences,p and d,need not be identical,in which case the crossovers will produce hybrid sequences p/d and d/p,distinct from either of the parental repeats.

While rearrangements will often be deleterious, this is not true in all circumstances. Duplications will be advantageous under special conditions if they amplify a gene whose product is limiting (88, 186, 190). For example, growth on a limiting carbon source such as arabinose selects for cells with a large duplication that enhances transport of not only the compound used for selection, but several other compounds as well (186). An inherent property of a large tandem duplication is that it will be very unstable because of the extensive homology available for looping out. Therefore, when conditions change so that the large duplication becomes a burden, the duplication can be eliminated with high

frequency (5, 186). Duplication can serve as the first essential step toward acquisition of new capabilities when it is followed by divergence of the duplicated copies (77, 93, 120).

If the repeated sequences are not identical, recombination between them can either eliminate the differences or produce new combinations. Elimination of diversity can occur by gene conversion (in its strict sense) or by double crossovers between divergent repeated sequences on sister chromatids. Novel joints characteristic of rearrangements will generally be a hybrid gene (Fig. 1).

Clearly, genetic interaction between repeated sequences is a major contributor to genome plasticity. The topic of repeated sequences in *E. coli* and *S. typhimurium* has been reviewed extensively (3, 83, 109, 149, 159–161).

FACTORS AFFECTING ECTOPIC RECOMBINATION

Several interesting mechanistic questions relate to ectopic recombination, and much recent experimentation has focused on this deceptively complex problem. Important issues include which recombination pathways are involved, whether the crossover is reciprocal, and how the crossover frequency depends on physical parameters such as length, sequence, sequence divergence (for nonidentical repeated sequences), and distance between repeats.

recA Dependence

Repetitious structures were recognized early in the study of *E. coli* genetics. Early examples included lysogens with tandem prophages (138) and mutants with large chromosomal duplications (21, 22, 29, 88). A common property of these early mutants was their high degree of genetic instability. It was observed that the duplications were rapidly lost, presumably through crossover between the tandem repeats (Fig. 1A and B). Homologous recombination was directly implicated by the observation that a *recA* mutation stabilized duplications (4, 77, 174). In a specific example, *recA* reduced the 6% segregation frequency of a 164-kb tandem duplication at least 250-fold (77). It was quite surprising, therefore, when later studies using plasmid model systems found that deletion between repeats sometimes occurs efficiently in *recA* strains (33, 127, 130, 132). One way of resolving this apparent contradiction was to hypothesize a fundamental difference in recombination pathways operating in plasmids compared with those operating in the host chromosome. However, when deletion between 787-bp repeats was tested on both plasmid and chromosome, the process was *recA* independent in both settings (127). Interestingly, the deletion frequency on the chromosome was 2 orders of magnitude lower than on the plasmid even when the arrangement of the repeats was otherwise identical. The distance between the points of crossover strongly influences the degree of *recA* dependence (13). If the repeated sequences are small and in tandem so that the crossover positions are spaced at only 300 bp, deletion occurs in a *recA*-independent manner. As the distance is increased, *recA* dependence increases, and at a spacing of 4 kb, deletion is highly *recA* dependent. Apparently, deletion between direct repeats can occur by both *recA*-independent and *recA*-dependent pathways, but the effectiveness of the *recA*-independent pathway diminishes as the separation between the crossover sites increases. One model proposed for nonreciprocal, *recA*-independent deletion hypothesizes that the exchange is initiated between repeats on sister chromatids during passage of the replication fork (127). The authors speculate that the asymmetric association of the leading and lagging strands with respect to the polymerase dimer complex brings a proximal repeat located on one chromatid close to a distal repeat located on the sister chromatid. Strand switching ultimately results in one chromatid with a deletion, while the other retains the parental duplication. Such a mechanism would inherently operate most efficiently if the repeated sequences are small and, more importantly, if they are spaced with a short periodicity. Deletion between very large repeats or widely spaced repeats could require a different, *recA*-dependent mechanism. The *recA*-independent deletion between 18-bp direct repeats in *Bacillus subtilis* is also strongly dependent on the distance between the repeats (26). It increased 1,000-fold as the distance was increased from 33 to 2,313 bp, and the effect was observed in both plasmid and chromosomal systems.

Size

A priori there is a strong presumption that ectopic recombination frequency will increase with size of the repeated sequence, and this has been the general observation. Below a threshold of 20 to 40 bp, homologous recombination is very inefficient (101, 132, 180, 208). Above this threshold, its frequency increases with length. The segregation frequency of very large chromosomal duplications also increases with size. Segregation occurred at frequencies of 1.4 and 5.9% for 40- and 164-kb duplications, respectively, when measured under otherwise identical conditions (77, 81). At least in this case, the increase in frequency with size was remarkably proportional. Although recombination frequency falls sharply when the repeats are shorter than a threshold of about 20 bp (208), sequence repetitions as small as 8 to 17 bp are observed to recombine specifically (1, 41, 100, 194, 210). In some studies, recombination between very short sequences was observed to be *recA* independent (97, 129, 132).

The major pathway for reciprocal recombination is the RecBCD pathway (101). Consistent with the assumption that chromosomal inversion between inverted repeats must be fully reciprocal in order to produce a viable chromosome (Fig. 1C), inversion is reduced by a *recB* mutation (177). Duplications, on the other hand, which can be formed by nonreciprocal recombination, occur at comparable frequencies in *recB* and *recB*$^+$ backgrounds.

Sequence

Repeat size is actually a deceptive parameter to assess since increase in the size of the repeated sequence will generally require the introduction of completely new sequence, and it may also increase the spacing between the points of exchange. As described above, the spacing factor can be a very significant one (13). Sequence is also clearly important (184). The best understood example of sequence influencing crossover frequency involves the Chi sequence. This sequence motif strongly stimulates the RecBCD recombination pathway (see chapter 119). In the context of ectopic recombination, a Chi sequence located 2 kb away from one of two recombining homologies stimulated crossover 20-fold (45).

Sequence Divergence

Homologous recombination can occur between sequences which are similar but not identical. However, the frequency has been shown to be sensitive to small amounts of sequence divergence, and the effect can be quite strong (68, 180, 181). In one study, 10% sequence divergence reduced recombination 40-fold (180). The mismatch repair system eliminates recombination intermediates that contain mismatched heteroduplex (157, 181). Inactivation of mismatch repair enhances recombination between divergent sequences (131). Reduction of recombination by sequence divergence has important implications for the evolution of microbial genomes (154, 155). It provides a means of suppressing chromosomal rearrangement due to crossover between naturally repeated sequences, since repeats that are somewhat divergent will cause less rearrangement than equivalent perfect repeats (150). By the same token, sequence divergence, once established, would be less prone to elimination by gene conversion. This would be especially important if the divergence has functional significance. There are many examples of protein-encoding genes that show significant sequence similarity indicating a common origin, but the similarity is usually less than 80% (159). This degree of divergence would reduce crossover between these gene pairs to very low levels.

Mutagens

Environmental factors strongly affect ectopic recombination. Mutagens such as UV irradiation, nitrous acid, ethyl methanesulfonate, and niridazole greatly stimulate duplication frequency by recombination between large repeated sequences such the rRNA genes (76) and the *Rhs* elements (122, 189) as well as other duplications whose precise origins are not established (72, 86, 87). The frequency of large duplications induced by moderate doses of UV light is remarkable, with estimates ranging from 5 to 12% of the survivors (72, 76). The SOS response, even in the absence of mutagen damage to DNA, stimulates duplication formation (34).

Nonpermissive Inversions

Chromosomal inversion between inversely oriented homologies has been documented extensively (79, 81, 172). However, not all chromosomal intervals are permissive for inversion (107, 158, 175, 176, 218). In at least some cases, the nonpermissiveness is not due to lethality of the final product since the equivalent structure can be created by other means (175, 176). Furthermore, the same homologies, located at the same chromosomal positions but with direct orientation, can recombine. It appears that repeats placed at certain positions in the chromosome cannot participate in the reciprocal, intrachromosomal crossover necessary for inversion.

LARGE REPEATED SEQUENCES

In *E. coli* and *S. typhimurium,* instances of large repeats that have similarities greater than 95% are rather limited (16). Essential genes that occur as highly conserved repeated sequences are the rRNA genes, some tRNA genes, genes for the small stable RNA component of RNase P (104, 105), and the genes for a protein elongation factor (*tufA* and *tufB*) (2, 215). A number of other protein-encoding genes show clear homology, but their sequence divergence is substantial (159). Accessory elements are a significant source of repetitious DNA. Accessory elements that have been found in multiple copies include ISs and the composite *Rhs* elements.

rRNA Genes

The most significant large repeated sequence in *E. coli* is the array of rRNA genes. Each of the seven *rrn* operons encodes 16S, 23S, and 5S rRNAs, with the 16S and 23S sequences separated by a tRNA spacer (43). Five operons are located clockwise from the origin of replication, and two are counterclockwise (Fig. 2). Each is oriented so that it is transcribed in the same direction as it is replicated. All have tandem promoters, and some operons encode additional tRNA species at their distal end (see chapter 13). Under conditions of rapid growth, *rrn* expression accounts for over half of the cell's transcriptional activity. The significance of there being seven of these operons is a long-standing question. The fact that both *E. coli* and *S. typhimurium* have seven *rrn* operons (117) indicates that the arrangement is both ancient and selectively maintained. The advantage of seven could simply be one of gene dosage. Alternatively, the operons might encode rRNA molecules with (slightly) different functions, or they might be affected differently by changes in growth conditions. The existence of rRNA sequence divergence has been long recognized (47), but its extent has not been clear. The recent determination of the complete sequences of five of the seven (*rrnA, rrnB, rrnC, rrnE,* and *rrnH*) (17) may bring us closer to an understanding. Divergence is limited, and most is unique to an individual operon, but there is a prominent exception. The five available 23S sequences are divided into two classes (*rrnB-E* and *rrnA-C-H*) by a 13-bp segment containing eight differences (17). This segment comprises a stem-loop (bases 1722 to 1738), and six of the substitutions are actually three pairs of compensating stem changes (Fig. 3). The accumulation of so many compensating changes implies that the divergence of these two versions of this stem-loop is relatively ancient. If both versions have always been carried within the same chromosome, one or the other should have been eliminated through conversion unless there is a selective advantage to having both (183). An alternative possibility is that

the two versions originated in different lineages and came together relatively recently in an ancestor of the present day *E. coli* K-12 chromosome.

FIGURE 2 Map positions of large repetitions found in the *E.coli* chromosome. The spacer types and orientations of the *rrn* operons are shown.The tRNA spacer types are abbreviated as follows: i-a, $tRNA^{Ile}$ -$tRNA_{IB}^{Ala}$; g, $tRNA_2^{Glu}$; sl-g, *rsl*-$tRNA_2^{Glu}$. Three inversion mutations between *rrn* operons are indicated. The positions of the five K-12 *Rhs* elements and an H-rpt fragment are shown.

FIGURE 3 Alternative stem-loops of 23S RNA species. Nucleotides 1721 to 1740 of the 235 RNA of *rrnB* and *rrnE* are shown in panel A. The corrresponding nucleotides of *rrnA, rrnC*, and *rrnH* are shown in panel B. Positions that differ between the two are enclosed in boxes.

Differential Expression. The question of differential expression of individual *rrn* operons has been directly addressed (28). Significant differences exist with regard to promoter activity in minimal medium, response to amino acid starvation, and response to depletion of a transcription factor.

Compared with its expression at a standardized location, *rrnH* transcription was specifically reduced when the gene was placed at its normal chromosomal location. Although it is not clear how these observations relate to the fitness of the cell in natural settings, taken collectively they convincingly establish differential expression. An early attempt at assessing the necessity of all seven operons showed that deletion of *rrnE* had no observable detrimental effect even under conditions of maximum growth (42). More recently, individual *rrn* operons were sequentially disrupted within the chromosome and the mutants were tested for whether the expression of other *rrn* operons would increase to compensate for the loss (27). Disruption of one or two operons had minimal effect, but disruption of a third and fourth operon progressively reduced both growth and ribosome content. These experiments support a model for feedback control of rRNA synthesis (96) that compensates for changes in gene copy. The mechanism of compensation involved increases in transcription initiation of the remaining operons and, curiously, in the rate of transcription elongation (27). The problem of *rrn* operon redundancy is complicated by the diversity of the tRNA spacers. Consequently, a separate question concerns the importance of spacer multiplicity. To assess this problem, an *E. coli* mutant was constructed in which three of the spacers were replaced with   spacers, leaving only the *rrnG* operon with a   spacer (70). The growth rate of this mutant was approximately half that of the wild type in synthetic medium, and cultures were rapidly overgrown by variants that had converted one of their tRNA$^{\text{Ile}}$- spacers to a second   spacer. Taken all together, these observations suggest that gene dosage and differential expression are both factors in the natural selection that has established seven *rrn* operons in *E. coli* and *S. typhimurium.*

The *E. coli* spacer exists in two forms, which differ in that a 106-bp sequence (called *rsl*) of one is replaced by a 20-bp sequence in the other. The *rrnG* and *rrnB* operons of strain K-12 contain the *rsl* version, while *rrnC* and *rrnE* contain the alternate version. However, some K-12 derivatives have had the *rsl* form replaced by the alternate, presumably through recombination (71). No function has been assigned to the *rsl* sequence. It is clearly not essential, because simultaneous elimination of *rrnB* and *rrnG,* which eliminates both copies of *rsl,* does not cause loss of viability (27).

Genetic Exchange between rRNA Genes. Since the rRNA genes are such large sequence repeats, the extent of genetic exchange between them is an important issue. The rRNA genes are hot spots for chromosomal rearrangement in both *E. coli* (32, 76, 78) and *S. typhimurium* (5, 118). Rearrangements associated with *rrn* recombination include duplication (5, 76), deletion (32, 82), inversion (81), and transposition (82). Recombination between the *rrnB* and *rrnE* operons has been the focus of several studies. They have the same orientation, and with a spacing of only 39.5 kb, they are the most closely linked *rrn* operons. Recombination between them occurs spontaneously at a frequency of 0.1×10^{-3} to 0.2×10^{-3} (82). If all possible pairwise combinations of the seven *rrn* operons were to recombine at this frequency, 2 to 4% of the cells in an *E. coli* population would bear some sort of chromosomal rearrangement. However, a number of factors, such as orientation, distance, and possibly sequence divergence, probably affect the frequency of specific pairwise interactions. For example, recombination between *rrnD* and either *rrnB* or *rrnE* occurs at a frequency of only 10^{-5} (79). *rrnD* is oriented in opposition to *rrnB* and *rrnE* and is separated from them by about 18 min. Nevertheless, the frequency of 10^{-5} is still large for a spontaneous genotypic change, and it is many orders of magnitude larger than the frequency of base substitution. It should be noted that DNA-damaging agents such as UV irradiation can increase the frequency of *rrn* recombination by at least 2 orders of magnitude (76, 79).

Given the frequency of crossover between *rrn* operons, the apparent stability of the genetic map of the enteric bacteria is remarkable (160, 168). Despite the divergence of the *E. coli* and *S. typhimurium* genomes by roughly 1 million base substitutions, the gross linkage maps, particularly the positions of the *rrn* operons, are highly similar. At least three exchanges between *rrn* operons have, in fact, occurred either in the natural population or without immediate recognition in early laboratory stocks. *E. coli* K-12 and *S. typhimurium* LT2 differ by a reciprocal exchange of spacers between *rrnD* and *rrnB* (117), and there has been a conversion of the *rrnB* spacer carried by the subline of *E. coli* K-12 that contains Cavalli Hfr (see above). In addition, an inversion, IN(*rrnD-rrnE*)*1,* occurred in another *E. coli* K-12 subline that contains such commonly used strains as W3110 and W3102 (81). IN(*rrnD-rrnE*)*1* involves 18 min of the chromosome, and the replication origin is 6 min from the *rrnE* end (Fig. 2). The

IN(*rrnD-rrnE*)*1* mutation causes a 2.7% growth disadvantage, probably reflecting displacement of genes outside the inversion toward or away from the replication origin (79). A larger inversion, IN(*rrnG-rrnE*), which is much more asymmetric with respect to the origin, has severely detrimental effects (79). Another large and extremely asymmetric inversion, IN(29–78), also has severely detrimental effects (126). The IN(29–78) inversion was shown experimentally to affect relative gene dosages along the rapidly replicating chromosome (31). Gene expression can be affected by the distance of the gene from the origin in a manner that correlates with the predicted gene dosage gradient in a population of partially replicated chromosomes (173). Another possible effect of inversions is to alter the relative positions of the origin and the terminus of replication if they span the origin asymmetrically. This symmetry is important for cell growth (125). Consistent with these ideas is the observation that the IN(*rrnG-rrnH*) inversion is relatively harmless despite its huge size (70). While IN(*rrnG-rrnH*) includes half of the chromosome, it is virtually symmetrical with respect to the origin (Fig. 2). The implication is that although rearrangements mediated by *rrn* operon recombination, as well as those derived by other means, may be present at high levels in all populations, the combination of growth disadvantage and frequent reversion to wild type would prevent the establishment of mutant clones in nature (79). In contrast to the conservation of chromosomal organization seen for *E. coli* and *S. typhimurium,* recent results with *Salmonella typhi* show that its chromosomal organization differs considerably and that a series of inversions and transpositions between rrn operons can account for much of the rearrangement (S.-L. Liu and K. E. Sanderson, personal communication).

The efficiency of mating between *S. typhimurium* and *E. coli* is strongly reduced by the general level of sequence divergence (131, 157). A high proportion of the infrequent recombinants are merodiploid (9). If markers in the intervals between the *rrn* operons are selected in interspecies conjugation, merodiploid recombinants are particularly prevalent (119). The *rrn* operons evidently have two significant roles. First, the *rrn* operons of *E. coli* and *S. typhimurium* are much more conserved than most genes and are more efficient sites for crossover. Second, their repetition provides opportunity for unequal recombination leading to duplication. Consequently, the interspecies merodiploids can retain a full set of recipient genes and thus avoid possible incompatible combinations. ISs, which can be highly conserved between species (14, 15), might serve a similar role in facilitating interspecies transfer by providing long stretches of sequence identity.

tRNA Genes

Many tRNA genes are present in multiple copies in *E. coli,* and they are organized in a variety of ways (51, 103). The example of the tRNA genes in *rrn* operon spacers has already been mentioned. In addition, both *rrnC* and *rrnH* encode at their distal ends (43). Beyond these cases, identical tRNA genes sometimes occur as tandem repeats in the same transcription unit. Examples include (99), (92, 103), and (39), all of which occur as tandem triplications, , which occurs as a tandem quadruplication (103), and , which occurs as a tandem duplication. Interestingly, genes for and occur in additional single copies, unlinked to the triplicate versions (51). The *metT* operon presents a particularly complex arrangement. In this operon, duplicate copies of and occur in tandem, while duplicates of are separated by other tRNA genes. In those cases where the duplicate copies are in the same transcription unit, the advantage of the arrangement would seem to be simply the capacity for more product. When the genes are unlinked, the possibility of differential control should be considered. A correlation seems to exist between tRNA gene multiplicity and codon usage. For example, , which translates the preferred glycine codons GGU and GGC (178), is present in four copies, while the species responsible for the rarer codons, GGA and GGG, are present as single copies (85). Similarly , present in four copies, translates the preferred leucine codon CAG. The tRNA sequences of *S. typhimurium* and *E. coli* tend to be highly conserved and often identical. Recombination between tandem copies of tRNA genes has been observed (99, 164).

IS Elements

Several IS elements are present in multiple copies in both *E. coli* and *S. typhimurium* (see chapter 111). Independent natural isolates tend to have distinctive IS profiles, although closely related strains show statistically significant conservation of IS position (67, 114, 115, 170). The IS patterns of strain subclones can vary considerably (140, 199). Aside from potential to cause rearrangement through site-specific recombination, IS elements provide sites for homology-dependent chromosomal rearrangement. There are numerous reports of repeated ISs serving as endpoints for deletions, duplications, and large inversions (106, 126, 169, 192, 198, 199). Their role in F-factor integration is particularly important (30). At least for IS*5,* this recombination is homology dependent and *recA* dependent, and it does not depend on element-encoded functions (193). IS elements have been implicated in the creation of an interesting genetic redundancy found in *E. coli* K-12. Ornithine carbamoyltransferase of *E. coli* K-12 is unusual in that this enzyme is encoded by two unlinked loci, *argI* and *argF* (201). While clearly homologous, these genes share only 78% nucleotide identity. The *argF* locus is absent from *E. coli* B and W (116). In K-12, the *argF* locus is flanked by IS*1* elements (90, 216). This duplicate gene likely evolved in a related species and entered a progenitor of strain K-12 facilitated by the IS*1* elements.

Prophage Insertion Sites

Some classes of temperate phage insert into the host chromosome by site-specific mechanisms that recognize both phage and host attachment sites. These sites share a homologous core sequence (24). In the case of lambda, the identity segment is 15 bp. Generally speaking, the crossover occurs within the identity segment, and once integrated, the prophage is flanked by duplicate copies of the homology. A number of phage or phagelike elements integrate within host structural genes. The attachment site of these phages contains the 3′ end of the target gene so that it replaces the portion of the gene displaced by the integration. The consequence of this mechanism is that the target gene remains intact, and the prophage is flanked by duplications of the 3′ end of the gene. Both protein and tRNA genes are used for attachment. For example, e14 (80), phage 21 (24), and Atlas (24, 134) all integrate into protein genes. The gene used by both e14 and 21 is the *icd* (isocitrate dehydrogenase) locus. P22 inserts into a threonine tRNA gene (24), DLP12 inserts into an arginine tRNA gene (123), and P4 inserts into a leucine tRNA gene (151). In each case, the displaced 3′ end is replaced by an equivalent sequence from the phage.

Rhs Elements

A major source of repeated sequence in the *E. coli* K-12 genome is the *Rhs* element family (122). The five *Rhs* elements of strain K-12 have been mapped (Fig. 2) and sequenced (49, 166, 217). These unusual elements are complex composites of distinct components. Some components are conserved, while others are divergent or even unique. The largest *Rhs* element is 9.6 kb long, and collectively the elements comprise 0.8% of K-12 DNA. Recombination between conserved portions of *RhsA* and *RhsB* produces a characteristic duplication that includes 3% of the chromosome (hence the name "rearrangement hot spot") (25, 122). Generation of this specific duplication is *recA* dependent, and there is no indication that it requires a specific *Rhs* function (122).

***Rhs* Organization.** The most prominent *Rhs* component is a 3.7-kb *Rhs* core (Fig. 4A), and core homology is present in each of the five K-12 elements (Fig. 4B). This core maintains a single open reading frame (ORF) throughout its length, with the start codon coinciding with the first base of the homology. Remarkably, the respective core ORFs extend up to 177 codons beyond the homology. Thus, the *Rhs* elements are predicted to produce a set of roughly 160-kDa proteins with long, conserved N termini and shorter, dissimilar C termini. The core ORF is immediately followed by another ORF, termed the downstream ORF. Typically, the downstream ORFs contain from 100 to 200 codons, and each is predicted to have a signal sequence for export across the inner membrane (217). In two cases,

this capability was proven through protein fusions with alkaline phosphatase (84). Like the adjacent core extensions, most of the downstream ORFs are unique, showing no homology with other downstream ORFs. An additional *Rhs* component is an insertion sequence, the H-rpt. Classification of the H-rpt as an IS is based in part on its homology to IS elements in other organisms. In both *Vibrio cholerae* (191) and *Salmonella enterica* serovar Strasbourg (212), homologous sequences are found at the *rfb* locus, linked to determinants of O-antigen variation. An H-rpt homolog in *Aeromonas salmonicida* has been shown to have transposition activity (69). This element, ISAS*2,* is 57% similar to the H-rpt over 335 amino acids.

No individual *Rhs* element has precisely the structure depicted in Fig. 4A. The left half of the core of *RhsE* is deleted. The H-rpt can be absent, defective, or present in multiple copies. In some elements, the distal portion of the core is repeated one or more times, and each core repetition is accompanied by an additional core extension and downstream ORF (165). Not all natural *E. coli* strains have *Rhs* elements, and comparison of *E. coli* strains with and without *Rhs* elements has been used to define the boundaries (49). Different *Rhs* elements replace from 10 to 807 bp of the reference chromosome (49, 217), and *RhsD* replaces a 224-bp bacterial interspersed mosaic element (BIME) (166). A general observation of accessory elements is that their ends are related by some kind of sequence repetition, but no sequence similarities are observed when the termini of the *Rhs* elements are compared. This holds both when the left end of an element is compared with its right end and when the ends of different elements are compared.

The *Rhs* Core. The most striking feature of the *Rhs* core ORF is a peptide motif that is repeated 28 times (49). The motif can be written xxGxxRYxYDxxGRL(I or T)xxxx, and in one cluster, it repeats with an average periodicity of 21 amino acids. This large protein is predicted to be strongly hydrophilic, but it has a hydrophobic region near the N terminus that could serve as a membrane anchor. It has been proposed that the core proteins are ligand-binding proteins of the cell surface (217). This idea received strong support from the report of a wall-associated protein (WAPA) of *B. subtilis* (50). WAPA is an abundant, nonessential protein. It is derived from a giant precursor encoded by a 7,002-bp ORF, and its C-terminal domain contains 31 copies of a motif that closely resembles the *Rhs* motif. The *Rhs* motif also resembles motifs associated with bacterial cell surface proteins involved with carbohydrate binding (207, 211). If the core protein is a cell surface component, the mechanism of its export is problematic because it does not have a good signal sequence. A possible function of the downstream ORFs, which do have signal peptides, is to assist export of the core protein (84).

***Rhs* Origins**. The *Rhs* cores fall into three distinct subfamilies based upon sequence divergence. The *RhsA-B-C* and the *RhsD-E* subfamilies are about 22% divergent at the nucleotide level, while divergence within the subfamilies is limited to between 1 and 4% (166, 217). The prototype of a third subfamily, *RhsG,* has been detected in other strains of *E. coli,* and partial sequence analysis indicates that *RhsG* is about 22% divergent from each of the other two (84). This degree of mutual divergence is greater than that observed for homologous genes in *E. coli* and *S. typhimurium* (143). The cores of all three subfamilies have G+C contents of about 62%, while the core extensions and downstream ORFs are only 35% G+C. Both of these values are significantly different from the 51% value found for the average *E. coli* gene. Our picture of *Rhs* evolution is a complex one. Apparently, the cores evolved into three subfamilies in a high-GC-content species. Separately, the core extension/downstream ORF combinations evolved in a high-AT-content species, diverging to a much greater degree. Much more recently, the components joined and entered the *E. coli* species. It seems especially significant that despite the relatively ancient divergence of the *RhsA* and *RhsD* cores, both have the same 28 repetitions of the *Rhs* peptide motif (166). This finding suggests that the motif repetitions were in place at the time of divergence, and that strong selective pressures exist for maintaining this aspect of *Rhs* structure**.**
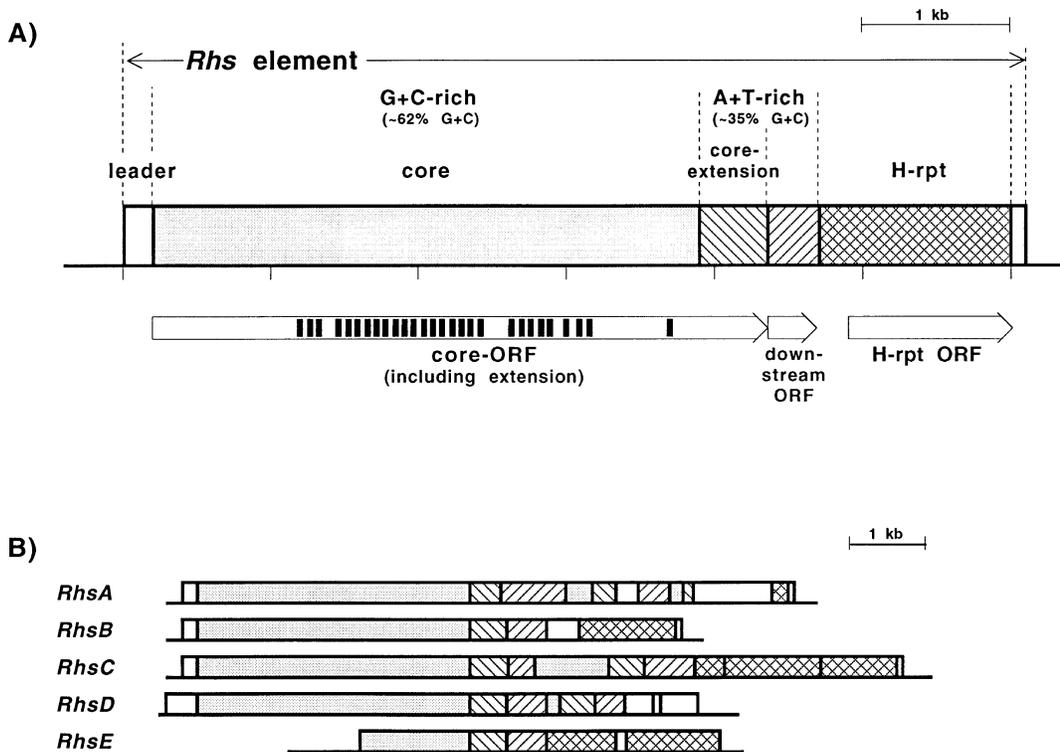
FIGURE 4 Structure of the *Rhs* elements. (A) The various *Rhs* components are shown in their usual positions in an idealized composite element. The boxes denoting different components are identified in the diagram,and the thin line denotes flanking genomic DNA. ORFs are indicated by the arrows,and the repetitions of the peptide motif are marked by bars within core ORF. (B) Schematic representation of the organization of the five *Rhs* elements of *E.coli* K-12. The symbols are in panel A. Note that no specific element has exactly the idealized structure depicted in panel A. Partial core repetitions found in *RhsA, RhsC, RhsD* generally contain the 3' end of the core and are followed by unique A+T-rich core extensions. The exceptional core fragment is a small piece from the 5' end found in *RhsD*.

SMALL EXTRAGENIC HIGHLY REPEATED SEQUENCES

In noncoding regions, the *E. coli* and *S. typhimurium* genomes contain a number of highly repetitive sequences (usually more than 30 to 50 occurrences per genome). Despite the fact that they are highly repeated, they do not constitute more than 2% of the total bacterial DNA. This explains why they were not detected by the classic C_{0t} analysis, which revealed eukaryotic highly recurrent sequences (19). They were essentially discovered in the last 12 years or so, usually by inspection or computer analysis of sequence data or by hybridization experiments. As the sequencing of the *E. coli* genome progresses to completion, computer analysis should lead to an exhaustive listing of such sequences and may reveal yet unsuspected relationships between them.

This section deals with the six classes of highly repetitive sequences which have been identified so far (Table 1). These sequences are rather short, in comparison with IS sequences, for example, do not usually encode proteins, and are in all cases except one (*iap* sequences) dispersed throughout the chromosome. They will be presented in the following order: BIMEs, with a review of their possible functions, intergenic repeat units (IRUs), box C sequences, RSA sequences, *iap* sequences, and Ter sequences (a class of *rho*-independent terminators).

BIMEs

Structure and Evolution. BIMEs are highly repetitive sequences found initially in the genomes of *E. coli* and *S. typhimurium* (64, 65). About 500 BIMEs are scattered over the whole bacterial chromosome, where they appear to be homogeneously distributed. BIMEs are found in extragenic locations at the 3′ ends of operons or between two genes of an operon, but rarely upstream of the first gene of an operon, and as far as we know, they are transcribed.

BIMEs are a mosaic combination of several short sequence motifs (64). One of these motifs, called PU (for palindromic unit) or REP (for repetitive extragenic palindromic) (Fig. 5), is present in all BIMEs and was the first to be described as a palindromic repetitive sequence (75). PUs occur usually in clusters in which they are associated with several sequences (called extra-PU sequences), belonging to a total of seven possible motifs.

BIMEs are composed of 10 short DNA motifs. The description of BIMEs in terms of their component motifs requires that each motif be clearly defined and named. This is done in the rest of this section, which primarily establishes the nomenclature.

Two types of PUs, Y and Z, have been distinguished according to the seventh position of the consensus, which is a G and a T, respectively. There are two PU motifs of the Y type, called Y and Y*, and two of the Z type, called Z^1 and Z^2 (8, 63). These motifs differ in size and in sequence (Fig. 5). Y* motifs are smaller than typical PUs (15 to 22 bp), are located between convergent operons, and are used as bidirectional *rho*-independent transcription terminators (see below).

When a BIME contains several PUs, successive occurrences of PUs are separated by short sequences (of up to 40 bp), which we called extra-PU sequences (64) (Table 2). Because of the strict alternation of successive PU orientations within clusters, extra-PU sequences are located either between the head ends of the two flanking PUs (and are called head internal sequences) or between the tail ends of the two flanking PUs (and are called tail internal sequences). Head internal sequences can be separated into two motifs, S (12- to 14-bp-long sequences) and L (32- to 34-bp-long sequences). There is a consensus for the L motif, which contains the consensus of an integration host factor (IHF)-binding site in its central part (18, 147). The S sequences are less conserved (Table 2). Two short motifs called s and l belong to the tail internal sequences. A third group of tail internal sequences called r is composed of a few sequences ranging from 18 to 31 bp which do not exhibit any sequence similarity. Two external motifs, flanking the tail end of the last PU, are present in a subset of BIMEs (reviewed in reference 65). They are juxtaposed either to a Z-type PU (called the A motif) or to a Y-type PU (called the B motif).

I   II   III   IV   V

```
        I       II                III    IV          V

Y    AA TGCCG GATGCG GCG-T G -AACGC CTTATCC GGCCTAC A
     CT                    A                        G

z¹   ATT GCCT GATGCG ---CTA--- CG- CTTATCA GGCCTAC A
                                                   G

z²   AAT GCCT GATGCGA CGCT G GCGCGT CTTATCA GGCCTAC A
     GTC                 T                          G
     ..................................................

PU   A T GCC G GATGCG G CG C NNNN G CG C CTTATC C GGCCTAC G
     T     T        A    T      A    T        A        A
```

Tail "T"   ▶   Head "H"

FIGURE 5 Three *E. coli* PU motifs. The consensus sequences of the three PU motifs are aligned. Y-type PUs are 38 bp long and posses a G at the seventh position. Z-type PUs possess a T at position 7 and are subdivided into z^i (33 bp long) and Z^2 (40 bp long). Below the three motifs is given the consensus previously found for PUs (57).Sequences similar among the three consensus sequences are boxed and similar among the three consensus sequences are boxed and numbered I to V. The lower part of the figure shows a schematic drawing of a PU;the triangle indicates its orientation, from the tail end to the head end.In previous publications, "tail" was named "left" and "head" was named "right".

TABLE 1   Main properties of extragenic highly repetitive sequences

Sequence	Size (bp)	Bacterial hosts	Tropism for location	Exchange
BIME	40–500	*Escherichia coli/Shigella* spp.; *Salmonella typhimurium; Levinea, Citrobacter, Klebsiella,* and *Enterobacter* spp.	IRU (*S. typhimurium*) Box C (*E. coli*)	IRU (*malE-F* in *K. oxytoca; glnA-L* in *S. typhimurium*) Box C (*araA-D* in *K. pneumoniae*)
Box C	56	*Escherichia coli/Shigella* spp.; *Klebsiella pneumoniae; Rhizobium* spp.	BIME (*E. coli*)	RSA (*envM* in *S. typhimurium*)
IRU	127	*Escherichia coli/Shigella* spp.; *Salmonella typhimurium; Citrobacter, Levinea, Erwinia, Klebsiella, Vibrio, Serratia, Yersinia,* and *Xenorhabdus* spp.	BIME (*S. typhimurium*) IRU (*X. luminescens*)	BIME (*glnA-L* in *S. typhimurium; malE-F* in *K. oxytoca*)
RSA	152	*Escherichia coli, Salmonella typhimurium, Erwinia carotovora*		Box C (*envM* in *E. coli*)
Iap	29	*Escherichia coli, Shigella dysenteriae, Salmonella typhimurium*		
Ter	30	*Escherichia coli*		

Two major BIME families. BIMEs with more than two PUs can be described as direct repetitions of a given association of PUs and extra-PU motifs. It is noteworthy that even for extra-PU motifs with poor sequence conservation (the S motif or the r sequences), the sequence similarity of two motifs within the same BIME is higher than for sequences originating from two different BIMEs (64).

BIME organization is variable, since these elements may contain different numbers of PUs (from 1 to 12) and since PUs can be associated with several motifs among a total of seven. However, *E. coli* BIMEs containing at least two PUs belong to two major BIME families, called BIME-1 and BIME-2 (Fig. 6) (8). Members of the BIME-1 family are composed of only two PUs, one Y and one Z[1], associated with an L motif. The external motifs A and B are frequently found in BIME-1. Members of

the BIME-2 family are BIMEs with 2 to 12 PUs, which are Y and Z^2, and are associated with the motifs S and s or l. External motifs are rarely present in BIME-2. Another difference between the two families is that BIME-1 members are mostly located after the last gene of an operon (18, 147), while BIME-2 members are located either between genes or after the last gene of an operon. However, the two families seem evenly distributed on the *E. coli* chromosome. The presence of two major BIME families on the *E. coli* chromosome could reflect a functional specialization of these sequences (see below).

TABLE 2 Extra-PU motifs in BIMEs

Extra-PU motifs[a]	Schematic representation	Motif name and size (bp)	Consensus[b]
Head internal sequences		**L**, 32–34	AAAGCATGCAAATTCAATATATAT
		S, 12–14	TGCAGRRATCR
			RWTYSGCAMCRACT
Tail internal sequences		**l**, 8–9	CGSWGCAC
		s, 1	C
		r, 18–30	
Tail external sequences		**A**, 20	TTTGCGTTTGTCATCAGTCT
		B, 20	AACAAAGCGCACTTTGTCAR
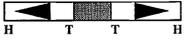
[a]Internal sequences are between two successive PUs, while external sequences form the border of the BIME, and flank only one PU. Head internal sequences (HIS) and tail internal sequences (TIS) are located respectively between the head ends and the tail ends of flanking PUs. Tail external sequences (TES) flank the tail end of the last PU of the BIME. In previous publications, tail was named "left" and head was named "right."
[b]R = A or G; Y = C or T; M = A or C; W = A or T.

BIMEs in other bacteria. BIME motifs were first identified in *E. coli* because of the high number of chromosomal sequences known in this species. However, BIME-like sequences were identified recently in other bacteria. For example, PUs are known in *S. typhimurium* and relatives and also in several *Klebsiella* species and relatives (6, 56, 62). In these bacteria, some structural features of the PU are identical to those of *E. coli,* but their sequences are species specific. The more divergent PU sequences are found in bacterial species that belong to phylogenetically more distant groups. For example, the major difference between *S. typhimurium* and *E. coli* PUs is the presence of an additional base pair in *S. typhimurium* sequences (underlined in Fig. 7), while *Klebsiella* PUs exhibit sequence singularities other than this additional base pair (not shown) (6). As in *E. coli,* there are several PU motifs in *S. typhimurium*; we were able to identify four different PU motifs (S. Bachellier, E. Gilson, and M. Hofnung, unpublished data), three of them being homologous to the *E. coli* Y, Z^1, and Z^2 motifs (Fig. 7). *S. typhimurium* PUs are associated with short sequences in a BIME-like structure. However, the sizes of extra-PU sequences are more variable than in *E. coli*; hence no consensus was determined, and no hybridization was obtained on the *S. typhimurium* DNA with the *E. coli* L motif as a probe (64). In conclusion, while BIMEs are present in *S. typhimurium,* their components, other than the PUs, are different in the two bacteria. This is not the case in *Klebsiella* species, in which the sizes of the extra-PU motifs are homogeneous and in some cases identical to the sizes of *E. coli* motifs (not shown) (6).

BIME intraspecific variations. To gain insight on the DNA rearrangements associated with BIME regions in *E. coli,* we undertook a systematic comparison, in 42 of the 72 strains forming the ECOR collection (144), of a subset of intergenic regions containing BIMEs in *E. coli* K-12. The observed BIME local variations between strains are described below.

BIMEs belonging to the BIME-2 family exhibit a polymorphism of repetition: the motifs present in the BIMEs are the same in all strains, but the number of repetitions of the BIME-2 basic motif combination is different than in *E. coli* K-12. As deduced from the phylogenetic relationships of ECOR strains (74), there can be either an increase or a decrease of this number of repetitions (Bachellier et al., unpublished data). Different mechanisms could explain such a result. It has been shown that DNA polymerase I (Pol I) can generate a polymorphism of repetition because of DNA strand slippage during a process called reiterative replication (108). It is worth noting that BIME DNA has affinity for Pol I (see
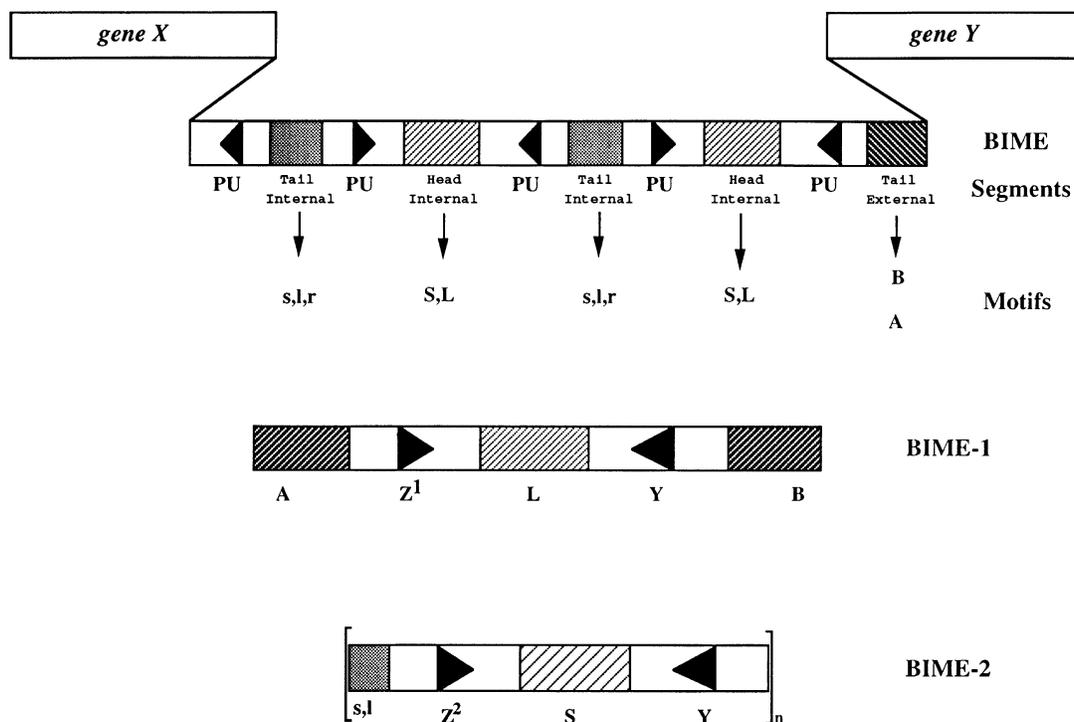
FIGURE 6 Two major BIME families. The upper part represents the schematic organization of BIMEs. Successive occurrences of PUs have alternate orientations and are separated by short extra-PU motifs. According to the ends of the PUs flanking these motifs, they are called head internal or tail internal sequences. In some BIMEs, the tail end of the last PU is flanked by a motif and hence called tail external. The lower part shows the structural organizations of the two major BIME families. BIME-1 is composed of two PUs (Z^1 and Y), separated by the L motif and often associated with the two external motifs. BIME-2 contains a variable number of motifs; the PUs are Z^2 and Y and are associated with the S motif and to either s or 1.

below) (61). This affinity could cause pausing of the polymerase at BIME DNA, favoring the slippage reaction. It is also possible that homologous recombination events occur within a single BIME, leading to the deletion of a part of the element.

Members of the BIME-1 family do not vary locally. The BIME structure is invariant, but in two cases, we observed no BIME in the intergenic regions of several ECOR strains, while the *E. coli* K-12 region had one (Bachellier et al., unpublished data). Such a result could be explained either by a deletion of the BIME in some ECOR strains or by an insertion of BIME-1 in K-12 and the other ECOR strains. As in the two intergenic regions the BIMEs are flanked by direct repetitions of non-BIME sequences, the deletions can easily be explained by homologous recombination events between the repeats. Conversely, it can be hypothesized that the direct repeats originate from the duplication of the BIME insertion site, as has been described for transposable elements. However, the sizes of the repeats in the two regions (30 and 34 bp) are larger than the average size of IS target site duplication, for example (3 to 13 bp) (54).

BIME spreading and sequence homogenization.

Spreading. The interspersed distribution of BIMEs on the bacterial chromosome suggests that these sequences have been dispersed. However, their transposition has never been reported, which could indicate that the mechanism(s) used for their spreading has been lost or occurs at a very low frequency. Many hypotheses for BIME spreading can be imagined, according to known mechanisms, such as (i) formation of a duplex DNA molecule (slDNA) in the presence of a stable stem-loop structure during replication (146) or (ii) reverse transcription of BIME-containing mRNAs by Pol I or specialized reverse transcriptases originating from retrons (113, 121) or group II introns (48), which are found in several E. coli strains. In addition, since BIMEs do not possess an ORF which could lead to the synthesis of proteins used for their own transposition, it can be hypothesized that there are on the chromosome a few copies of active or complete elements leading to the synthesis of proteins involved in their spreading. Such a phenomenon has already been described for several eukaryotic repeated sequences (for example, L1 [46]). It can also be hypothesized that BIME spreading relies on proteins of other transposable elements, like ISs, and/or on host proteins. It has indeed been shown that IS transposition requires some bacterial proteins, namely, DNA Pol I, DNA gyrase, and IHF (for a review, see reference 54; see also chapter 124), which are known to interact specifically with BIME DNA (see below).

Homogenization. The sequences of the BIME motifs exhibit a high level of species specificity (see previous section). This had already been reported for eukaryotic repeated sequences and attributed to "concerted" evolution (for reviews, see references 36 and 37). Two major models have been established to explain the mechanisms of sequence homogenization. The first is a succession of nucleotide sequence variations of the repeats, but new sequences appear in the genome, from intact copies, through a mechanism of duplicative transposition. The second model, called gene conversion, involves nonreciprocal information transfer between two repeats (reviewed in reference 36). In the case of BIMEs, the presence of different extra-PU motifs could avoid frequent exchanges between nonidentical BIMEs, leading to the fixation of the two major families (BIME-1 and BIME-2).

BIMEs as Multifunctional Genetic Elements. BIME sequences appear to participate in seemingly disparate functions: transcription, translation, chromosome organization, and stability. However, it is still unknown whether BIMEs are essential for bacterial viability. If such a role exists, it does not require all of the BIMEs present on the chromosome since the removal of one has no effect on cell growth (187).

Here, we summarize work on the multiple processes in which BIMEs have been shown to be involved. Finally, we present a model that attempts to explain this functional diversity in terms of different combinations of BIME motifs and of sequence context. This leads to the idea that noncoding repeated DNA can be a source of different functions through a number of sequence variations within or around BIMEs.

BIMEs and gene expression. BIMEs were first described as potential regulatory sequences because of their palindromic nature and ability to form stable stem-loop structures in transcribed RNA (75). In fact, among the BIMEs examined, many but not all participate in the stabilization of the 3′ end of mRNA and subsequently in the expression of the upstream gene; a small subset is implicated in a *rho*-independent transcription termination event, and only one seems to be involved in the translational control at the ribosome binding site.

mRNA turnover and retroregulation. In numerous operons, the gene located upstream of BIMEs has a much higher expression level than the downstream gene. For example, in the S. typhimurium hisJQMP and E. coli malEFG and deoCABD operons, the first gene, followed by one BIME-2 sequence, is expressed at up to 40 times the level of the distal genes (57, 142, 187, 200). The mRNA corresponding to the proximal gene is overrepresented compared with the full-size transcript of the operon, with a 3′ end located precisely in the BIME-2 region. Since deletions within the BIME-2 region decrease both the expression of the proximal gene and the amount of the transcript ending at BIME-2, it was proposed that BIMEs act as retroregulators by stabilizing the mRNA of the upstream gene. This stabilization, probably due to a protection of the RNA against a 3′-5′ exonuclease activity, can be explained by the ability of

BIME-2 RNA to form complex secondary structures. However, the extent of this BIME-2 effect cannot account for all of the differential expression observed in these operons. Indeed, a total deletion of the BIME-2 sequence between hisJ and hisQ leads to only a twofold decrease in the expression of hisJ (188). This effect on mRNA stability is not confined to BIMEs located between two genes of the same operon but holds for BIMEs located at the distal end of an operon (for example, glyA [152] and gdhA [11]).

In summary, both BIME-1 and BIME-2 sequences can participate in the differential expression within polycistronic operons by protecting the proximal part of the mRNA against exonucleolytic degradation. This effect depends on the number of PU motifs present within the BIME, since one PU does not stabilize mRNA (142), and on the structure of the transcript outside BIME, since the same BIME has different mRNA stabilization effects according to the operon into which it is inserted (219). Although the increase in the upstream gene expression is modest, it can be of biological importance in some operons. Indeed, the removal of part of the BIME-2 sequence in the *malEFG* operon leads to a decrease in *malE* expression and to a partial defect in maltose utilization (142).

We believe that this stabilization effect cannot account for the high level of BIME sequence conservation for the following reasons. (i) The presence or absence of BIME-2 in the *glpK-X* intergenic region has no effect on transport or on growth on glycerol (196). (ii) Insertion of the *hisJQ* BIME-2 downstream of the *atpH* gene does not affect the half-life of the corresponding mRNA (219). (iii) Since any stem-loop structure is sufficient to stabilize upstream RNA, for example as observed in the regulation of λ*int* expression by *sib* (171) or in the maturation of the 3′ end of the *trp* operon (139), it is difficult to involve RNA stabilization as an explanation for the sequence conservation of BIMEs, in particular for the nonpalindromic extra-PU motifs.

The BIME-2 of the *rplL-rpoB* region includes an RNase III processing site in one PU motif (57). The sequence of this PU is atypical: the upper part of the stem-loop is missing. Interestingly, some loose homology exists between the lower part and a known RNase III site in phage T7 (57). No other evidence exists for an association of a PU with RNase III processing. In particular, the BIME-2 present in the *hisJ-Q* region is not processed by RNase III in vitro (187).

*BIMEs and translation initiation.*The removal of the hisJ-Q BIME affects both the expression of the upstream hisJ gene (see above) and the translation of the downstream hisQ gene (188). The secondary structures that can be adopted in the RNA of the intergenic region could modify the accessibility of the ribosome binding site.

Functional organization of the bacterial chromatin. The observation that BIMEs specifically bind nucleoid-associated proteins (60) could provide a plausible cause for BIME sequence homogeneity. It has been shown that DNA gyrase (213), DNA Pol I (61), and IHF (18, 147) are able to specifically recognize BIME DNA. Gyrase and Pol I interact with the PU motif, while IHF binds in the center of the L motif (Fig. 8).

BIME-DNA Pol I complexes. Starting from a crude E. coli extract, two moieties which specifically protect a BIME-2 DNA against digestion with exonuclease III were purified. One of these involved Pol I. This interaction requires the presence of the PU motif. The other activity is less characterized but has been shown to be devoid of DNA gyrase (61). This finding was the first evidence that Pol I is able to bind intact duplex DNA. Whether BIME-1 DNA is also able to bind Pol I is not known.

```
A
    ... ........ ... ..CGC GA AGCTC TT ATCCGGC CTACG        - adk
    GGT GCCGGAT GGC GGCCT GA ACGCC TT ATCCGGC CTACT   adk  - visA
    TGT GCCGGAT GGC GACAT AA ATGCC TT ATTCGGC CTACA   araA - araD
    CAT GCCGGAT GGC GAGCT GC GCGTC TT ATCCGGC CTACC   cchA - cchB
    ATT GCCGGAT GGC GGTGC GA GCACC TT ATCCGGC CTACA   eutG - eutH
    ATT GCCGGAT GGC GACGT A- ACGCC TT ATCCGGC CTACA   lexA -
    ATT GCCGGAT GGC AGCGT AA ACGCC TT ATCCAGC CTACA        - livB
    TGT GCCGGAT GGC -GCAT AA ATGCC TT ATCCGGC CTACA   malE - malF
    ACA GCCGGAT ACG GCAGA AG TTGCG TT ATCCGGT CGATG   mviM - mviN
    ATT GCCGGAT GGC GACGT T- GCGCC TT -TCCGGC CTGAA   mviN - flgN
    AGC GCCGGAT GGC GGCGC AA GCGCC TT ATCCGGC CTACA   nrdA - nrdB
    TTT GCCGGAT GGC GGCGT GA ACGCC CG ATC.... .....        - nrdE
    AGC GCCGGAT GGC GGCTT AG ATGCC TT GTCCGGC CTACG   parC - parF
    AAA GCCGGAT AAT GGCGT GA ACGCC TT ATCCGGC CTACG   pepM - ORF1
    ATT GCCGGAT AGC GGCGC AA ACGCC TT ATCCGGC CTACA   ushA -
    ------------------------------------------------
      G               A
    A T GCCGGAT GGC GGCGT  A ACGCC TT ATCCGGC CTACA
      T               G

B
    ATT GCCTGAT GGC GT T-AC GC TT ATCAGGC TTGCG    adk  - visA
    ATT GCCCGGT GGC AC T-GC GT TT ACCGGGC CTACG    araA - araD
    CCT GCCGGAT GGC GC T-TC GC TT ATCCGGC CTACG    araA - araD
    GGG GCCTGAT GGC GC T-GC GC TT ATCAGGC CTGCG    argA - recD
    ATT GCCTGAT GGT GC G-CA GA TT -CAAGGC CTACG    cchA - cchB
    TTG ACCTGAT GGC GC T-GC GC TT ATCGGGT TCGAA    cheZ - flhB
    ACT GCCTGAT GGC GC T-AC GC TT ATCAGGC CTACA    crr  - orf287
    CTT GCCTGAT GGC GC T-GT GC TT ATCAGGC CTACA    cysJ - cysI
    ATT GCCTGAT GGC GT T-AC GC TT ATCAGGC CTGAA    hemA -
    GTT GCCCGGT GGC GC T-GC GC TT ACCAGGC CTACA    hemN -
    GAC GCCTGAT GGC GC T-GC GC TT ATCAGGC CTACG    hisG - hisD
    ATT GCCTGAT GGC GC T-GT GC GT GTCAGGC CTACG    hisJ - hisQ
    CCT GCCTGAT GGC GC T-AC GC TT ATCGGAC CTGGG    ilvH -
    GGA GCTTGAT GGC GC T-AC GC TT ATCAGGC CTACA    leuD -
    GTA GCCTGAT GGC GC T-GC GC TT ATCAGGC CTACG    lexA -
    TTT GCCCGAT GGC GC T-TC GC TT ATCGGGC CTACG         - livB
    AAA GCCGGAT GGC GC TGAC GC TT ATCCGGC CTACA    livB - livC
    TGT GCCTGAT GGC GC T-GC GC TT ATCAGGC CTACG    malE - malF
    ATT GCCGGAT GCG GC T-AC GC TT ATCCGGC TTATC    melB -
    AGT GCCTGAT AGC GC T-GT GC T- ATCAGGC CTACA    mviN - flgN
    ACT GCCTGAT GGC GC T-TC GC TT ATCAGGC CTACA    nadR -
    AGT GCCGGGT AGC GC T-AT GC TT ACCCGGC CTACG    nrdA - nrdB
    TCT GCCTGAT GGC GC --GC AA CC TTAAGGC CTACG    ushA -
    ------------------------------------------------
                    G               A
    ATT GCCTGAT GGC GC T  C GC TT ATCAGGC CTAC
                    A               G

C
    AAC GCCGGAT GAC GCCGT TCAA- AGCCG CC ATCCGGC CT-GTT    envZ -
    AAT GCCGGA- GGC GACG- TAAA- -CGTC TT ATCCGGT TT-ACA    eutG - eutH
    TTC GCCGGAT GAC GACGC GCGT- GCGTT TT ATCCGGT CT-ACT    ilvH -
    CGC GCCTGAC GGC GACGT TAAA- ACGTC TT ATCAGGC CTTACA    mviM - mviN
    TTT GCCGGAT GGC GATGC AAT-- GCGTC TT ATCCGGC CT-ACG    nadR -
    AAT GCCTGAT G-C GACGC TTACC GCGTC TT ATCAGGC CT-ACA         - nadR
    GTT GCCGGAT GAC GACGC AGTA- GCGTC TT ATCCGGC CT-ACG    parC - parF
    TTT GCCTGAT GGC GATGC CTTC- GCATC TT ATCAGGC CT-ACA    pepM -
    AGC GCCTGAT G-C GACGC TGCC- GCGTC TT ATCAGGC CT-ACA    trpR -
    ------------------------------------------------
    AAT     G       G         TA           C
        GCC GAT G C GACGC TA     GCGTC TT ATC GGC CT ACA
    TTC     T       A          AC           A

D
    AAT GCCGGAT TGC --C TGG CGC TT ATCCGGC CTACA         - adk
    CGG GCCGGGT GGC G-C TTG CGC TT ATCCGGC TTGTA    alr  -
    ATT GCCGGAT GGC GCG TTG CGC TT ACCGGGC ATAAT    cpsB - cpsG
    ATC GCCGGAT GGC GGC -GC -AA GT A-CAGGC AAAAA    crr  - orf287
    AAA ACCGGGT GGC GGC TTC GCC TT ACCCGGC CTACG    hemL -
    TGT GCCGGAT GGC G-C TAA TGC TT ATCCGGC CTACA    hemN -
    TTT ACTCCAT CACAGGG TTC CGCCTT TT ATCCGGC CTACA hisG - hisD
    CAC GCCGGAT GGC GGC TGT GCC TT GCCCGGC CTACG    hisJ - hisQ
    ATC GCCGAAT GGC GCT GGC GCC TT ATTCGGC TTACA    lamB - malM
    TCT GCCGGAT GGC GGC TCC GCC TT ATCCGGC CTACA         - murB
    ATC GCCGGAT GGC GGC TTC GCC TT ATCCGGC CTACA    uvrD - corA
    ------------------------------------------------
        C               T  GC
    AT  GCCGGAT GGC GGC T C    C TT ATCCGGC CTACA
        T               G  CG
```
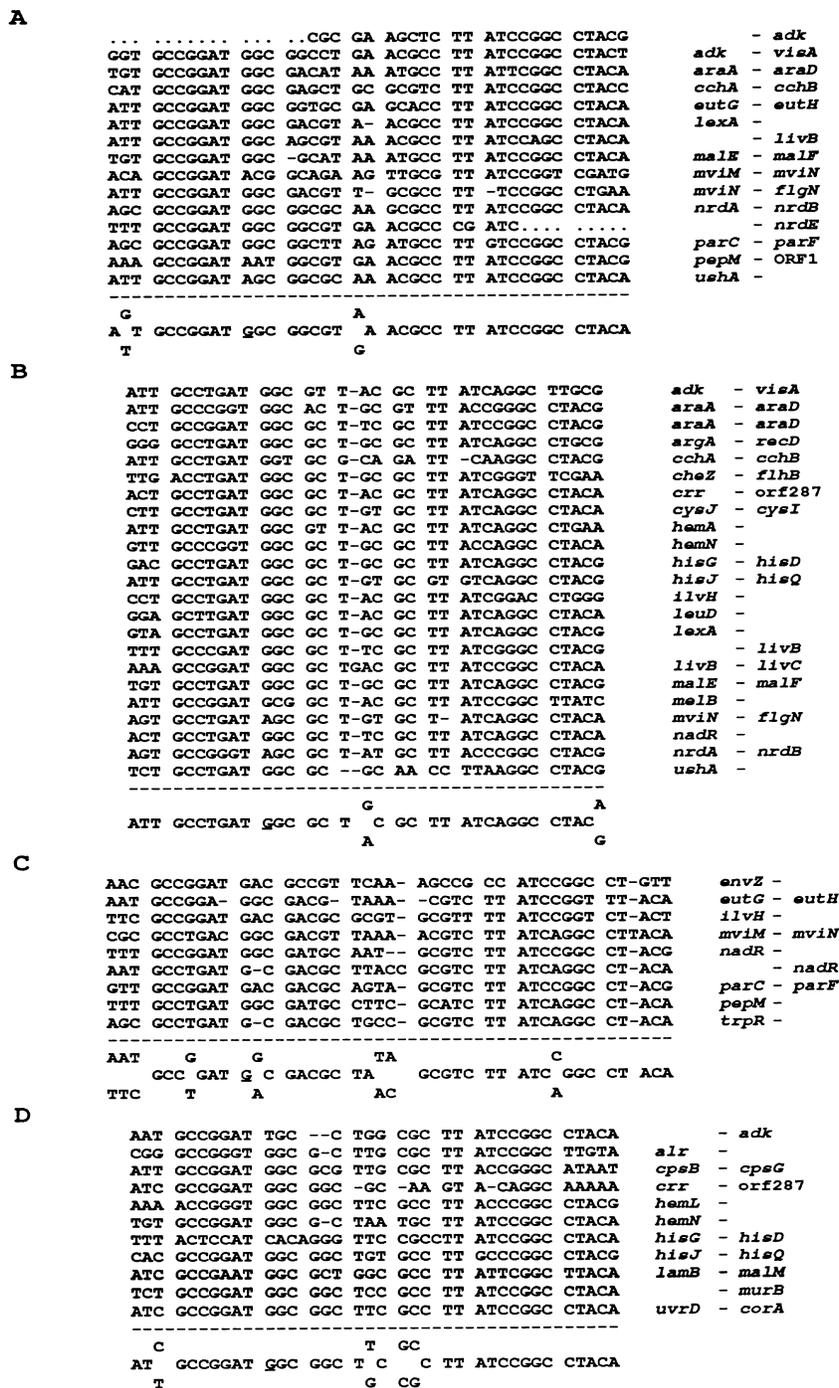
FIGURE 7 Four PU motifs in *S. typhimurium*. PU sequences of *S. typhimurium* were subdivided, according to size and sequence, into four groups, for which a consensus was found. Groups A, B, and C present sequence similarities with the *E. coli* Y, Z^1, and Z^2 motifs, respectively. Sequences of group D are not homologous to any *E. coli* motif and could constitute a new PU motif specific to the *S. typhimurium* genome. Positions of the sequences which differ from the consensus are in lowercase letters; dashes represent gaps that have been introduced to maximize the alignments; dots indicate when a sequence is not known; the positions underlined in the four consensus sequences are not present in the *E. Coli* PUs. Each sequence is identified by the names of the flanking gens (as in Table 7).
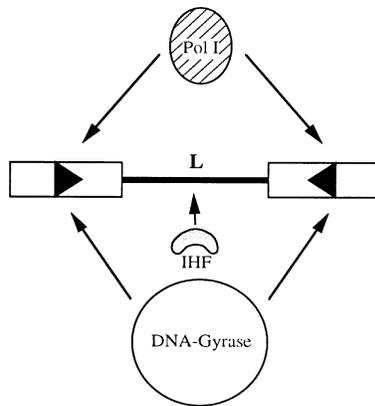
FIGURE 8 Summary of the known DNA-protein interactions occurring at BIME DNA. DNA Pol I and DNA gyrase both interact with the PU motifs, while IHF binds at the center of the L motif.

The functional significance of this interaction is still a matter of speculation. BIME DNA could serve as a preferred entry site for Pol I, providing a specific pausing site for the polymerization reaction or playing a role in replication fidelity. A possible effect of the BIME-Pol I interaction is the amplification of BIME regions (see above).

BIME-gyrase complexes. Purified gyrase binds specifically to the PU motif of BIMEs (213). Differences in binding affinity of up to threefold have been observed between the different PU motifs and between BIME-1 and BIME-2, with the order $Y > Z^2 > Z^1$ and BIME-2 > BIME-1 (7, 8). From these findings, the following determinants for an efficient BIME-gyrase interaction are revealed. (i) PU motif appears to be the basic sequence recognized by gyrase. Interestingly, a critical size of 7 to 9 nucleotides (nt) in the central part of the PU seems important for an efficient binding, and sequence variability in this part of the PU can account for the differences in affinity that we observed between the three PU motifs (8). This finding strongly suggests that the two external parts of the PU, which are highly conserved between the different motifs, are directly recognized by gyrase only if a proper spacing between them is respected (7 to 9 nt).

(ii) In BIME-1 DNA, it appears that a particular arrangement of PU sequences impairs an efficient PU DNA-gyrase interaction (7). The L sequence is slightly bent, as revealed by a circular permutation assay (18). This DNA curvature could be unfavorable for a proper wrapping of DNA around gyrase (10). In BIME-2 DNA, the spacing between two PUs seems optimum to allow a proper wrapping around gyrase. The high gyrase affinity for BIME-2 DNA could be due to a binding of one gyrase dimer on a Y and a Z^2 sequence. Gyrase DNase I footprinting on a BIME-2 sequence has revealed a protected region covering both PUs, substantiating this hypothesis (213).

The fact that PU and BIME DNAs are specific binding sites for gyrase does not imply that they represent sites of catalysis. No strong cleavage site induced by oxolinic acid has been mapped within a BIME-2 DNA (213). A binding site in pBR322 that is not a cleavage site has also been mapped (102). This shows that a binding site does not necessarily define a site of cleavage and catalysis. For example, BIMEs could be preferred sites of gyrase binding without catalysis: gyrase could enter into BIME DNA and then move along DNA or, alternatively, BIME DNA could be a stop sequence for gyrase movement. Evidence for such a linear diffusion of eukaryotic topoisomerase II has been reported *(89, 148)*.

BIME-IHF *complexes*. Purified IHF binds specifically to all L DNA sequences examined (18, 147). Indeed, the central part of the L motif includes an ihf consensus sequence (YAANNNNTTGATW) (55). For this reason, the L-containing BIMEs were called RIP, for repetitive IHF-binding palindromic elements (147), or RIB, for reiterative IHF-BIME (18). Since all of the L motifs are present in a BIME-1

structure and since the ihf consensus is highly conserved within the L sequences (18, 64), most, if not all, of the RIP/RIB elements belong to the BIME-1 family.

IHF is known to play a role in various processes, including site- specific recombination, transcriptional regulation, and replication (reviewed in reference 52). Since IHF binding induces a strong bend of 140° and since an IHF-binding site can be functionally replaced by an intrinsic curved sequence, it is believed that the main function of IHF is to facilitate the formation of higher-order protein-DNA complexes.

Higher-order nucleoprotein complexes at BIME DNA.At least three proteins have been shown to specifically bind BIME DNA in vitro: DNA gyrase, DNA Pol I, and IHF (Fig. 8). Evidence for the occurrence of these interactions in living cells is still lacking, except for IHF–BIME-1 complexes that have been shown to be formed in vivo (43a). A synergistic fixation of gyrase and IHF on BIME-1 has been observed in vitro (7). This cooperativity suggests that the inefficient binding between BIME-1 and gyrase is relieved by IHF. Changing the bend angle in the L motif can lead to a spatial closeness of both PUs, now in a favorable arrangement to wrap around gyrase. In this hypothesis, it is worth noting that the length of a BIME-1 DNA (the consensus has 144 nt) is within the range of the size of DNA wrapped around gyrase (120 to 140 nt) (156). This model is also in good agreement with a previously proposed role for IHF in the formation of higher-order specialized nucleoprotein structures (40) by the proper alignment of distant protein-binding sites. This model does not exclude a direct interaction between both proteins, but evidence for such an interaction is lacking (53). An increase in gyrase binding on BIME-2 DNA has been reported in the presence of HU, another histone-like protein that binds DNA with low specificity and that introduces a DNA curvature (214) (for a review on HU, see reference 38).

On the basis of the specific binding for gyrase, BIMEs have been proposed to be the counterpart of the eukaryotic scaffold-associated regions (58, 213), which are in vivo binding sites for topoisomerase II (98); scaffold-associated-region DNA is believed to be involved in the formation of chromatin loops and independent topological domains (for a recent review, see reference 162). Since BIMEs are almost exclusively located within transcriptional units, BIME-gyrase interactions could also be involved in the removal of positive supercoils generated ahead of RNA polymerase during transcription (18, 124). The presence of an IHF-binding site in some BIMEs suggests the formation of multiprotein-DNA complexes (reviewed in reference 52). The formation of such higher-order structures reinforces the idea that BIMEs play an important role in the architecture of the bacterial chromosome (18).

Redundancy and functional specialization. BIMEs are functionally diverse (see above), and a clear relationship can be drawn between the different BIME functions and variations in their own organization and in their flanking sequences. For example, (i) a subclass of PU (Y* [63]), and not the other PU sequences, is implicated in bidirectional transcription termination; (ii) only BIME-1 DNA binds specifically IHF (18, 147); and (iii) modulations in gyrase binding can be achieved by different combination of BIME motifs and by the binding of histone-like proteins, such as IHF and HU (7).

BIMEs are not the only case of noncoding repeated sequences that can be recruited to achieve different functions according to local sequence variation or genetic context. One can cite the following examples. (i) The repeated uptake sequences of *Haemophilus influenzae* and *Neisseria gonorrhoeae* are involved both in species-specific DNA transformation and in transcription termination when they are part of an inverted structure (66, 110, 195). (ii) Short eukaryotic interspersed repetitive elements, like the Ocr elements of *Xenopus laevis* and the rat ID, mouse B1 and B2, and human *Alu* sequences, appear to constitute modules in complex regulatory elements ensuring the coordinated expression of various genes (137, 167, 205). (iii) A class of repeated simple sequences acts as a functional telomere when located at the very ends of eukaryotic chromosomes or as a transcriptional regulatory element when located at internal sites (reviewed in reference 59).

Short interspersed repetitive sequences appear, thus, as building blocks of a variety of genetic elements. This extends the concept of gene duplication involved in genetic diversity (145) to noncoding repeated elements.

Other Short Repeats

IRUs. The IRUs (179) were initially identified in several members of the family *Enterobacteriaceae* during the analysis of the region adjacent to the 3′ end of the *tls* gene from *E. coli.* This region is homologous to 17 other bacterial regions. Another laboratory identified the same sequences and renamed them ERIC, for enterobacterial repetitive intergenic consensus (91).

TABLE 3 Distribution of IRUs (ERICs) in various bacterial species

Bacterium	No. of sequences	Comments
Citrobacter amalonaticus	1	See reference 203
C. diversus	1	See reference 203
C. freundii	1	See reference 203
Erwinia carotovora	1	1[a]
E. chrysanthemi	2	2[a]
Escherichia coli	14	4[a]
Klebsiella oxytoca	1	S. Bachellier, unpublished sequence[b]
K. pneumoniae	2	1[a]
Salmonella typhimurium	9	1,[b] 2[c]
Serratia marcescens	1	1[a]
Vibrio cholerae	4	3[a,d]
Xenorhabdus luminescens	6	See reference 133[d]
Yersinia pseudotuberculosis	2	

[a]Number of sequences which were not used to build up the first consensus.
[b]The IRU sequence lies at the same position as a BIME in *E. coli.*
[c]In the same region, there is one IRU and one BIME.
[d]A single region contains several IRUs, the numbers and the positions of which vary between two strains of the same species.

General characteristics. IRUs are imperfect palindromic sequences which are about 125 nt long; they may form a stem-loop structure and can be oriented. They have been detected simultaneously in several bacterial species, and their sequences are very homogeneous, which allows establishment of a consensus (91, 179). IRUs were found in a number of *Enterobacteriaceae* (Table 3) as well as in *V. cholerae,* a species not very distant from members of the *Enterobacteriaceae.* These sequences are always extragenic. The majority of these sequences are transcribed. In contrast to BIMEs, IRUs are sometimes associated with promoter regions (in at least six cases). They are dispersed on the bacterial chromosome and have not been detected in bacteriophages or in plasmids. Their total number in bacterial genomes has been estimated to be between 30 and 150 (91, 179), and their number is not necessarily identical in all species. For example, IRUs seem more abundant in *S. typhimurium* than in *E. coli* (Table 4). By screening sequence banks, we found several supplementary IRUs, sometimes in other bacterial species (Table 3). This led us to establish a new consensus of 127 nt for IRUs (Fig. 9). As in the case of a number of BIMEs, some IRUs are located in corresponding intergenic regions of bacteria belonging to different species: *metE-R* (91, 179) and *ahpC-F* in *S. typhimurium* and *E. coli;* and *rpsU-dnaG* in *S. typhimurium, Levinea malonatica, Citrobacter freundii,* and *C. amalonaticus* (203).

Variability of IRU sequences from different bacterial species. The sequences of IRUs do not vary much from one species to another, at least in comparison with BIMEs; for example, IRUs from *E. coli, Klebsiella* species, and even *Vibrio* species have homologous sequences, which can be detected by hybridization or by computer search in sequence banks (91, 179). Because of the small number of IRUs known for each bacterial genus, it is not possible to decide whether the variations with respect to the consensus from one species to another reflect a species specificity. It does seem clear, however, that if there is a species specificity for IRUs, it is much less apparent than in the case of BIMEs. Their sequences are indeed mostly conserved in distant species, which can be explained in several ways. They may have appeared recently and been dispersed, for example, via an association with a transposable element or by some other type of horizontal transfer. On the contrary, it is possible to assume that IRUs are ancient elements whose sequences have been conserved either because of gene

conversion or because these sequences have a function which is selected for. The fact that IRUs are dispersed on the chromosome suggests a transposition step, but since there is no ORF, they do not carry information for a protein involved in transposition. There are no known functions for IRUs. All of the data can be interpreted by saying that the IRUs are (or were) very mobile sequences, susceptible to local rearrangements (203).

The Box C Sequences.

Box C sequences as highly repetitive sequences on the *E. coli* genome.The box C sequences were initially described as sequences of 43 nt, located in extragenic positions, transcribed, and composed mainly of G and C; a consensus was deduced from the alignment of the first eight box C sequences identified (12). The 5′ end of the sequence is composed mainly of pyrimidines (mainly C), while the 3′ end presents a large proportion of purines (mainly G). The box C sequences are imperfect palindromes which can therefore be oriented.

Box C sequences were identified in a region which is partially homologous between *E. coli* and *S. typhimurium* (12). The *envM* gene is present in both bacterial species, but the adjacent nucleotide sequences are completely different. Upstream of the *envM* gene in *S. typhimurium,* there is a small ORF which is not present in *E. coli;* instead, there is a short DNA sequence, the box C sequence (12). The center of box C had been previously identified as a repetitive sequence present in five extragenic regions of *E. coli* K-12 (111). The existence of this sequence as a repetitive element was also reported following computer analysis of inverted repetitions on the *E. coli* genome (group IV in reference 16). Box C was used as a probe in hybridization experiments on the genomic DNAs from *S. typhimurium* and *E. coli* (12). The probe hybridizes with several genomic DNA fragments from *E. coli* but not with the genomic DNA from *S. typhimurium.*

Box C sequences exist in other bacterial species. Upon reexamination of the eight regions previously identified by Bergler and colleagues (12), we found four other box C sequences. The regions ahead of *fepB* and *envM* contain two box C sequences as well as the *mtlA-D* and *phnP-Q* intergenic regions. When two box C sequences are located in the same region, their orientation can be direct *(mtlA-D, phnP-Q)* or inverted *(fepB, envM).* The sequences located between the two box C sequences may vary substantially in length (from 8 bp in the *phnP-Q* region up to 131 bp ahead of *envM*). In a few cases, flanking regions exhibit sequence similarities (S. Bachellier, unpublished data). Searches in sequence banks led us to identify 28 regions containing box C sequences, including four in *Rhizobium* sp., a gram-negative bacterium quite distant from members of the *Enterobacteriaceae* (examples 24–27 in Fig. 10). By aligning all of the sequences (Fig. 10), we defined a new consensus which comprises 56 nt. In addition, two intergenic regions, one from a natural isolate of *E. coli* (ECOR 8 [144]) and the other from *Klebsiella pneumoniae,* which we sequenced in our laboratory, also contained a box C (Bachellier et al., unpublished data) (Table 5).

Like IRUs, the box C sequences seem to be restricted to genomes of bacteria which are very closely related to *E. coli,* but in contrast to IRUs, they are not present in genome of *S. typhimurium* (12). Their total number on the *E. coli* chromosome can be estimated to be between 40 and 45. Their sequences present several interesting features: a quasipalindromic structure and a separation into two regions, one purine rich and the other pyrimidine rich, including a mirror symmetry of 10 bp (see consensus in Fig. 10). It has been shown in vitro that DNA with similar characteristics may form a triple helical structure, also called form H (206). Like BIMEs and IRUs, the box C sequences have been found so far only in chromosomal DNA.

TABLE 4 Compilation of IRU (ERIC) sequences

Sequence	Flanking gene(s)			GenBank accession no.	Observations
1[a]	*aceF*	–	*lpd*	V01498	
2	*ahpC*	–	*ahpF*	D13187	
3		–	*aldH*	M38433	
4	*aspS*	–		X53863	
5	f83	–	f143	U00039	
6	*fdx*	–		U01827	
7	*hpaC*	–	*cts1*	Z37980	
8	*hsdR*	–	*hsdM*	X06545	
9	*metE*	–	*metR*	J04155	
10	*narK*	–	*narG*	X15996	
11	*pgk*	–	*fda*	X14436	
12	*pmbA*	–	*cybC*	U14003	
13	*rhlB*	–	*gppA*	X56310	
14	*rplA*	–	*rplJ*	V00339	
15	*rpsB*	–	*tsf*	V00343	
16	*yafP*	–		D38582	
17	*ahpC*	–	*ahpF*	J05478	
18		–	*cysJ*	M23007	
19	*glnA*	–		J01803	
20	*livB*	–	*livC*	D00478	A BIME is distant from 30 bp.
21	*metE*	–	*metR*	M17356	
22		–	*nirC*	M64606	
23	*pepM*	–		X55778	The IRU is inserted within a BIME.
24	*rpsU*	–	*dnaG*	M14427	
25	*tdk*	–	*osmZ*	–	IRU sequence is given in reference 91.
26	*topA*	–	*cysB*	M15040	

[a]The numbers identifying each sequence are the same as in Fig. 9. Sequences 1 through 6 are from *E. coli;* sequences 17 through 26 are from *S. typhimurium.*

As was noticed for BIMEs, box C sequences may be differently located in various strains of the same species (the region between *araA* and *araD* from *E. coli* K-12 and ECOR 8, the 3′ region of the *recA* gene from *Shigella flexneri* and *E. coli*). This finding suggests strongly that these sequences are (were) mobile.

Functions hypothesized for the box C sequences include an effect on the level of transcription on the gene located downstream and/or a role in the stabilization of the upstream mRNA. Because three of the five sequences identified by Kunisawa and Nakamura (111) are linked to genes involved in the transport of different substrates, these authors suggested that these sequences could play a role in regulating the expression of transport systems. This seems rather unlikely since most of the box C sequences identified later are located near genes whose products are not implicated in transport.

FIGURE 9 Alignment and consensus of IRU sequences. IRU sequences from *E. coli* (sequences 1 to 13) and *S. typhimurium* (sequences 14 to 23) are aligned. Below the sequences is given the consensus, which was deduced from the alignment of all available IRUs, including those from bacteria other than *E. coli* and *S. typhimurium*. The positions which are different from the general consensus are in lowercase letters; points indicate that the sequence is not known; dashes represent deletions introduced to optimize the alignment. The numbers following the sequences refer to the numbering in Table 4.

```
GGcCTGTTCCCCTCACCCTAACCCTCTCCCgA----Gg--GGG-CGAGGGGAC-tG-TCCG-TGC 1
GG-CTGTTCCCCTCACCCTAACCCTCTCCCCA----AA--GGGGCGAGGGGAC-gG-AttG-TGC 2
GaACaTTaaCCCTCACCCcAgCCCTCaCCCtG----GA--aGGGAGAGGGGgC-aGaACgG-cGC 2'
GcGCTTTTCCCCTCACCCTAACCCTCTCCCCA----GA--GGGGCGAGGGGAC-CGtAttG-TGC 3
ctGCaGgaaCtgaacCggTcACCCTCTCCCtG----AA--aGaGCGAGGGGgC-aG-ACCG-aGC 3'
cGGgTTTTCCCCTCACCCTAACCCTCTCCCCA----AA--GGGGCGAGGGGAC-CG-TCCGgTGC 4
GGtCTGTYCCCCTCACCCTAACCCTCTCCCCA---gGA--GGGGCGAGGGGAC-tG-TCCG-gGC 5
GGcCTGTTCCCCTCACCCTAACCCTCTCCCCA----AA--GGGGCGAGGGGAC-tG-TCtG-aGC 6
GaAtaTTTaCCCTCACCCcggCCCTCTCCCtG----AA--aGGGCGAGGGGga-aa-AgCG-TGC 6'
GGttTGcgCCCCTCACCCTAAtCCTCTCCCCA----tA--GGGGAGAGGGaAC-tG-cCaG-TGC 7
GtGaacgggCCCTCACCCTAACCCTCTCCCCA-----A--GGGGCGAGGGGAC-CG-TCCa-ctC 8
GtttgcgcCCCCTCACCCTAACCCTCTCCC-t----cA--GGG-AGAGGGGAC-CG-TtCG--GC 9
GGcCTTTTCCCCTCACCCcggCCCTCTCCCCG----GA--GGGGCGAGGG-Ag-aa-Aaaa-aGa 9'
cGAgTGggagCacggtttTcACCCTCTtCCCA----GA--GGGGCGAGGGGAC-tc-TCCG-aGt 10
GaGgcTTTtCCCTCACCCaAACCCTCTCC-gt----GA--GGaGAG-GGGcAg-CG-cgga-TGC 11
GcAtcaggCaatTaACCCc-ACCtTCTCC-at----Gt--GGaGAG-GGtGgg-at-T-gG-att 11'
caGggTTcCCtCTCACCCTAACCCTCTCCC-G----Gt--GGGGCGAGGGGAC-tG-ACCG-aGC 12
tattcGTaaggtTggtttTctCCCTCTCCCtG----t----GGGAGAGGG--C-CG--ggG-TGa 13
GcACgTTcaCCCTCACCCTAACCCTCTCCCtc----AA--GGG-AGAGGGGAC-CG-AtCG-aGC 13'
cGtacGTTCCCCTCACCCTAACCCTCTCCCCA----AA--GGGGCGAGGGGAC-CGttCaG-TaC 14
acAaTTTTCCCCTCACCCTAACCCTtTtCCCG----AA--GGGGCGAGGGGAC-tG-TCCG-gGC 15
aacggGTTCCCCTCACCCTAACCCTCaCCCA----AA--GGG-CGAGGGGAC-CG-TtCGtTGC 16
GaAaaaTTaCCCTCACCCcgACCCTCTCCCtG----GA--aGGa-GAGGaGAa-aa--CgG-TGC 17
GGcCTGTTCCCCTCACCCTAACCCTCTCCCCA----AAc-GGGGCGAGGGGAC-tG-ACCG-aGt 17'
cGcCTTTTCCCCTCACCCTAACCCTCTCCCCA----GA--GGGGCGAGGGGAC-CG-AtCG-cGC 18
aatCaGccCCtaTCAaCCgcctttaCgaatCA----AA---taaCGAtaaGgC-aG-TCCa-TcC 18'
GcGCTGgcCCCCTCACCCTAACCCTCTCCCCA----GA--GGGGCGAGGGGAC-CG-AttG-TGC 19
GaA-TaTTgCgCTCgtttTctCCCTCTCCCCA----tt--GGGGtGAGGGG-C-ga-TgCc-TGC 19'
acGgTTTTCCCCTCACCCTAACCCTCTCCCCA----GA--GGGGCGAGGGGAC-CG-ACCG-aGC 20
Gt-CgTTTCCCCTCACCCcggCCCTCTtCCCA----AA--GGGGCGAGGGGga-aa-ACaa-TGC 21
GGcCTGTTCCCCTCACCCTAACCCTCTCCCCA----AA--GGG-CGAGGGGAC-CG-TCCG-TGC 22
GaGgccgTCCCCTCACCCTtcCCCTCTCCCG----tt---GGGAGAGGGGAC-CG-ACCG-gGC 23
cGACTTTgCCCCTCACCCTAACCCTCTCCCAc-agGc--GGGGAGAGGGGAC-tt-T--G-gGC 24
GGAgaGccCCCCTCACCCc-ACCCTCTCCCGc-aaGc--GGGGCGAGGG-Ag-aG-cagGcaGC 25
acctaGcgCataTgACCCTAACCCTCTCCCCGtgaaGAacGGGGAGAGGGGAC--GtTCCa-TaC 26
GGG-TTaaaCCCgaggaCTAgCCCTCTCCCGc-aaGc--GGGGAGgGGGGACgtG-cCtG-atC 27
tGAaTGTcgatC-CgCCCTAgCCCTCTCCCCtc-tcGA--GGGGAGAGGGttg-gG-AgaG-gGg 28
-----------------------------------------------------------------
GGRCTKTTCCCCTCACCCTAACCCTCTCCCR      RA   GGGGMGAGGGGAC CG WCCG TGC
```

FIGURE 10 Box C sequences: alignment and consensus. Shown in an alignment of all known box C sequences as well as the consensus which we deduced from it according to the method described in reference 64. This consensus is slightly longer than that in reference 12 (56 bp instead of 43 bp). Sequences 1 to 20 are from the *E. coli* K-12 genome, 21 is from *E. coli* ECOR 8, 22 is from *Shigella flexneri*, 23 originates from the *K. pneumoniae araA-D* intergenic region (Bachellier, unpublished data), 24 to 27 are from *Rhizobium* sp., and sequence 28 is from *Pseudomonas paucimobilis*. The 5' end of the box C sequence 24 is included in the end of the *nolC* gene of *R. fredii*; the stop codon TAA is in italics. The *E. coli* sequences are numbered according to Table 5. Underlined sequences belong to PUs flanking the box C (see text). Positions differing from the consensus (deletions, additions, or different nucleotides) are in lowercase letters.

TABLE 5 Compilation of box C sequences

Sequence[a]	Flanking gene(s)			GenBank accession no.	Observations
1	*Adk*	–	*visA*	D90259	Presence of a BIME in the same intergenic region of *S. typhimurium*
2	*EntC*	–	*fepB*	M29730	2 box C sequences in inverted orientation, separated by 11 bp, within putative ORF1
3		–	*envM*	M97219	2 box C sequences in direct orientation separated by 131 bp
4	f308	–	f418	L19201	
5	*FimH*	–	f477	U14003	Presence of a BIME in the same intergenic region at a distance of 18 bp
6	*GvcP*	–		L20872	2 box C sequences in inverted orientation separated by 11 bp
7	*GlnH*	–	*glnP*	X14180	
8		–	*hemB*	M24488	
9	*LctD*	–	ORF	L13970	2 box C sequences in inverted orientation separated by 38 bp
10		–	*melA*	X04894	
11	*MtlA*	–	*mtlD*	X51359	2 box C sequences in direct repeat, bracketing a BIME
12	ORF2	–	ORF3	X52904	Presence of a BIME in the same intergenic region at a distance of 14 bp ORF2 is downstream from *cstA*
13	ORF	–	*ponB*	D26562	2 box C sequences in direct orientation separated by 12 bp; the first one is poorly conserved
14	o265	–	f361	U18997	Presence of a BIME in the same intergenic region, at a distance of 24 bp
15	o290	–	f90	U00006	f90 is downstream from *lysC*
16	o309	–	f481	U18997	
17	o321	–	o414	U18997	2 box C sequences in inverted orientation separated by 11 bp
18	o732	–	f81	U00039	
19	*PhnP*	–	*phnQ*	J05260	2 box C sequences in direct orientation separated by 8 bp within putative *phnQ* gene
20	*PstA*	–	*pstB*	K01992	Presence of a BIME in the same intergenic region at a distance of 29 bp
21	*AraA*	–	*araD*		From *E. coli* ECOR 8; the box C sequence flanks a BIME; unpublished
22		–	*recA*	X55561	From *Shigella flexneri*

[a] As in Fig. 10. Sequences 1 to 20 are from *E. coli* K-12.

RSA Sequences. RSA sequences were initially found in four extragenic regions of *E. coli* as well as in one region of *S. typhimurium* and one region of *Erwinia carotovora* (K. Mizobuchi, personal communication). In *E. coli,* the region upstream of gene *envM* contains two box C sequences (12) (see also above), whereas in *S. typhimurium,* there is an ORF which could encode 99 amino acid residues (197); the 3′ region of this ORF contains an RSA sequence.

An RSA sequence is also found upstream of the gene *araC* from *Erwinia carotovora,* and in *E. coli,* RSAs are located upstream of *rpsP,* in the intergenic region between *rplT* and *pheS,* and upstream of *cysP* (Table 6). The RSAs are thus in most instances located outside structural genes. We established for them a consensus of 152 bp (Fig. 11). The RSAs could form a large stem-loop structure (Fig. 11) which would be quite stable since their folding energy calculated with the program foldRNA (220) would vary from –39.6 kcal (1 kcal = 4.184 kJ)/mol for the sequence in *S. typhimurium* to –54.7 kcal/mol for the sequence of *Erwinia carotovora.* Since these elements are not perfect palindromes, they can be oriented. The proportion of G and C is about 50%, and the ratio of purines to pyrimidines is 40%. Two new RSA sequences were recently found. The first is located in an intergenic region from *E. coli* located between genes *smbA* and *frr;* the second is upstream of a short ORF located on a recombinant plasmid (pNM506). In both cases, the RSA sequence is extragenic.

TABLE 6 Compilation of RSA sequences

Flanking gene(s)[a]		GenBank accession no.	Observations
-	*cysP*	M32101	
-	ORF1	S67014	ORF1 is located upstream of *glnB*. The sequence in the database starts at position 38 of the RSA consensus.
ORF -	*rpsP*	X01818	The first 18 nt of the RSA are included within the ORF.
rplT -	*pheS*	K02844	
smbA -	*frr*	D13334	
ORF -	*envM*	M31806	The 5′ end of the RSA sequence is included within the putative ORF. A box C is present at the same place in *E. coli*.

[a] The first five sequences are from *E. coli*; the last one is from *S. typhimurium.*

RSAs constitute a distinct family of repeated sequence in the genomes of members of the *Enterobacteriaceae.* The sequences of RSAs from distant phylogenetic species (*E. coli* and *Erwinia carotovora,* for example) are homologous. RSA sequences are dispersed on the *E. coli* chromosome, and

their sequences do not contain an ORF which could play a role in their dispersion (data not shown). If one assumes a uniform distribution of RSAs on the *E. coli* chromosome, one may estimate that there are about 10 of them. There are no data pertaining to a possible function of RSAs in the bacterial genome.

iap **Sequences.** Between min 59 and 60 on the *E. coli* genetic map, there is a highly conserved sequence of 29 bp, containing an inverted repeat of 7 bp that appears 14 times, 32 or 33 bp apart, downstream of the *iap* gene coding region (Fig. 12). About 24 kb downstream of the 14 repeats, a second intergenic region containing similar 29-bp sequences occurring seven times with a spacing of 32 bp was first described (95, 141). The same intergenic region appears to contain two additional 29-bp repeats, at a distance of 500 bp from the other repeats (Bachellier, unpublished data). These sequences, called *iap* sequences, correspond to imperfect palindromes with the first half pyrimidine rich and the second half purine rich. This is reminiscent of the structure of box C sequences. The spacing sequences have a constant size but a variable nucleotide composition.

Nucleotide sequences hybridizing with the 29-bp fragment were not detected in other regions of the *E. coli* chromosome. Thus, the 23 repeats of these sequences are clustered in a 1-min region of the *E. coli* genetic map. In that sense, they are not dispersed throughout the genome, but they can be considered as highly repetitive sequences. Hybridizing sequences were also detected in *Shigella dysenteriae* and *S. typhimurium* but not in *K. pneumoniae* or *Pseudomonas aeruginosa.*

The sizes of these sequences are of the order of that of PUs. No function has been proposed for the *iap* repeats.

Ter Sequences. In a search for dispersed recurrent DNA sequences in the *E. coli* genome, Blaisdell and colleagues (16) found eight groups of structural repeat identities on a contig sequence which was 1.6×10^6 bp long. While other groups corresponded to known extragenic repeats (see above) or to coding sequences, group III corresponded to *rho*-independent terminators. The target sequence was a palindromic sequence of 30 bp in lsength. This group consisted of 22 members, 3 of which occurred in tandem on the same noncoding region following gene *rnpB* at successive displacement of 83 bp each. This means that the exact palindromic members occur in 20 compact regions. All of the members lie in noncoding regions that are highly variable in length. These sequences are examples of the *rho*-independent terminators reviewed by Rosenberg and Court (163): a C+G-rich region capable of forming a stable stem-loop structure followed closely by a T-rich region that does not bind strongly to the DNA template strand. In the case presented (16), the form has been specialized by the incorporation of an A-rich region at the beginning that is capable, with the following T-rich region, of extending the stem-loop structure. Such structures are capable of acting as bidirectional terminators (153) and bear also similarities to Y* (see above).

BIME, IRU, Box C, RSA, *iap*, **and Ter Sequences**

Similarities and Differences. The six small elements which we have described have a number of common properties: they are found only on the bacterial chromosome, dispersed and in extragenic locations, form imperfect palindromes, and are usually transcribed. In addition, four of them show some kind of tropism for location (see below) which may be interpreted as if they had a common mechanism for dispersion**.**
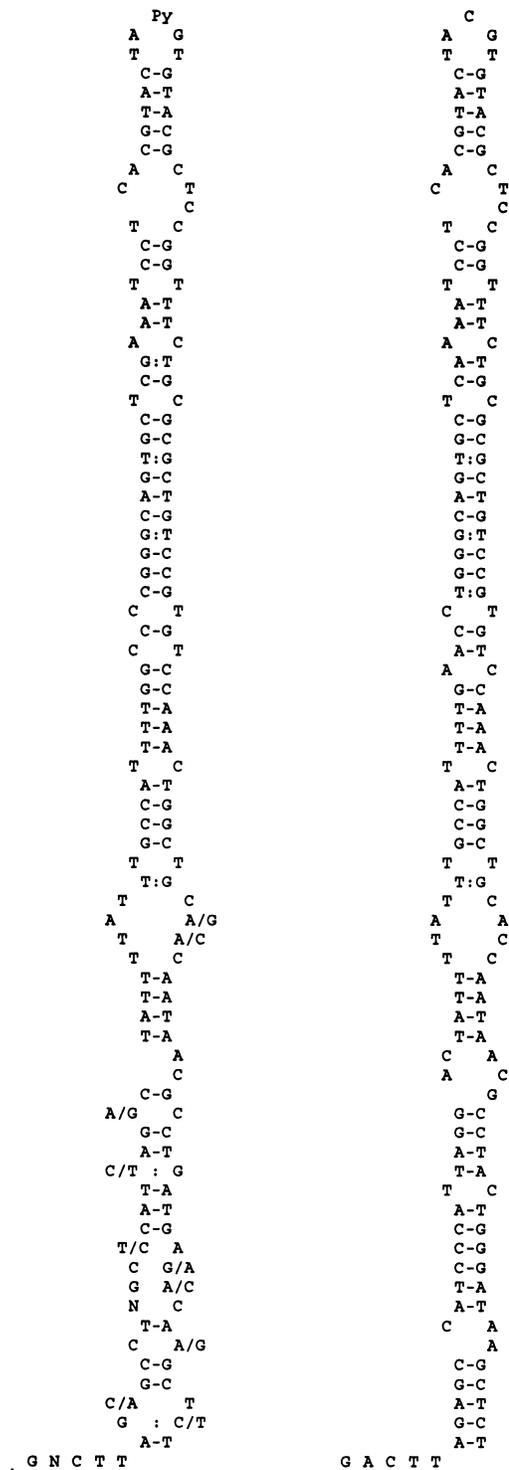
```
          Py                         C
        A    G                     A    G
        T    T                     T    T
         C-G                        C-G
         A-T                        A-T
         T-A                        T-A
         G-C                        G-C
         C-G                        C-G
        A    C                     A    C
       C       T                  C       T
               C                          C
       T    C                     T    C
         C-G                        C-G
         C-G                        C-G
        T    T                     T    T
         A-T                        A-T
         A-T                        A-T
        A    C                     A    C
         G:T                        A-T
         C-G                        C-G
        T    C                     T    C
         C-G                        C-G
         G-C                        G-C
         T:G                        T:G
         G-C                        G-C
         A-T                        A-T
         C-G                        C-G
         G:T                        G:T
         G-C                        G-C
         G-C                        G-C
         C-G                        T:G
        C    T                     C    T
         C-G                        C-G
        C    T                      A-T
         G-C                       A    C
         G-C                        G-C
         T-A                        T-A
         T-A                        T-A
         T-A                        T-A
        T    C                     T    C
         A-T                        A-T
         C-G                        C-G
         C-G                        C-G
         G-C                        G-C
        T    T                     T    T
         T:G                        T:G
       T       C                  T    C
      A          A/G             A       A
       T      A/C                 T      C
       T    C                     T    C
         T-A                        T-A
         T-A                        T-A
         A-T                        A-T
         T-A                        T-A
             A                     C    A
             C                    A       C
         C-G                            G
      A/G    C                      G-C
         G-C                        G-C
         A-T                        A-T
      C/T : G                       T-A
         T-A                       T    C
         A-T                        A-T
         C-G                        C-G
       T/C  A                       C-G
        C  G/A                      C-G
        G  A/C                      T-A
        N    C                      A-T
         T-A                       C    A
        C    A/G                         A
         C-G                        C-G
         G-C                        G-C
      C/A      T                    A-T
       G  : C/T                     G-C
         A-T                        A-T
     . G N C T T            G A C T T
```

FIGURE 11 RSA sequences: consensus and palindromic sequences. The left side represents the palindromic structure that the consensus of RSA sequences could adopt. N=A,C,G, or T: Py=C or T. At the other variable positions of the consensus, the two possible nucleotides are shown separated by a slant line. On the right side, we show the palindrome which could be formed by the RSA sequence located upstream of the araC gene from *Erwinia carotovora* (G=--54.7 kcal/mol).

```
GGAGTTCTACCGCAGAGGCGGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTGATGCGGGGAACAC < 32 bp >
CGGTTTATCCCTGCTGGCGCGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTAACGCGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACTC < 33 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACTC < 33 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACTC


ATGGTTATCCCCGCTGACGCGGGGAACAT < 32 bp >
CGGTTTATCCCCGCTGATGCAGGGAACAT <500 bp >

TGGTTTATCCCCGCTGGCGCGGGGAACTC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACAC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACAC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACAC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACAC < 32 bp >
AGGTTTATCCCCGCTGGCGCGGGGAACAC < 32 bp >
CGGTTTATCCCCGCTGGCGCGGGGAACAC
```

FIGURE 12 Structure of *iap* repeats. The 29-bp repeats located after the *iap* gene as well as in the second region (see text) have been aligned in two separate clusters. The sequences of the two first repeats of the second region appeared in GenBank U29580. Sizes of spacer sequences are given in base pairs (bp) on the right.

One of the main differences between these sequences is the number of repeats found for each family. It is quite variable from one species to another, and the four types of elements are not necessarily found in the genomes of all *Enterobacteriaceae;* in *E. coli,* BIMEs are present in hundreds of copies and Ter sequences are present in 60 to 100 copies, while there are about 50 IRUs and box C sequences and about 10 RSAs. In *S. typhimurium,* there is no box C, while the IRUs are more represented than in *E. coli* and the BIMEs are present in a lower number of copies. The differences in repartition of these elements could be due to differences in the mechanisms of dispersion between these elements from one bacterial species to another.

For these six families of sequences, no transposition mechanism has been demonstrated, and none of these elements encodes an ORF likely to direct the synthesis of proteins playing a role in dispersion as is found for ISs. One hypothesis could be that the sequences of these families are the "ghosts" of transposable elements or that they depend on *trans*-acting factors for their transposition. Since these six elements are generally transcribed, they are located on a number of mRNA molecules which could be used as intermediates in transposition (see above). Another large difference between these elements is the species specificity of the sequences between different phylogenetic groups of bacteria. This specificity is high for BIMEs and weak or absent for box C sequences, RSAs, and IRUs. The specificity has not been examined for *iap* and Ter sequences.

Tropism for Location. In a large number of cases, it is interesting that repetitive sequences belonging to two different families are located at the same site. In five intergenic regions of *E. coli* and two from *S. typhimurium,* one finds two repetitive elements, which are adjacent and separated by a few nucleotides (30 at most) or are inserted one into the other (Fig. 13). In *E. coli,* all of these regions carry box C and BIME sequences. The two sequences located between *pstA* and *pstB* are separated by 29 nt, those in the 3′ regions of *cstA* and *fimH* are separated by 14 and 18 nt, respectively, and the BIME and the box C located between *araA* and *araD* from ECOR 8 overlap by 5 nt. In the region between *mtlA* and *mtlD,* a BIME is flanked by two box C sequences in direct repetition (Fig. 13). The two box C sequences of this region are incomplete in that one of their extremities is quite different from the consensus: the box C located 5′ to the BIME is deleted in its 3′ part, while the box C located 3′ to the BIME is deleted in its 5′ region (underlined in Fig. 10). One interpretation of the structure of this region is that the insertion of a BIME led to a duplication of part of the box C sequence. In *S. typhimurium,* one finds IRUs and BIMEs

in the same regions. One IRU and one BIME are separated by 30 nt in the region located between *livB* and *livC,* and in the region located immediately after gene *pepM,* an IRU is inserted within a BIME. The frequency with which one finds an association between two of these elements is striking, especially for the box C sequences: in 5 of 18 regions known to carry box C sequences in *E. coli,* they are associated with BIMEs.
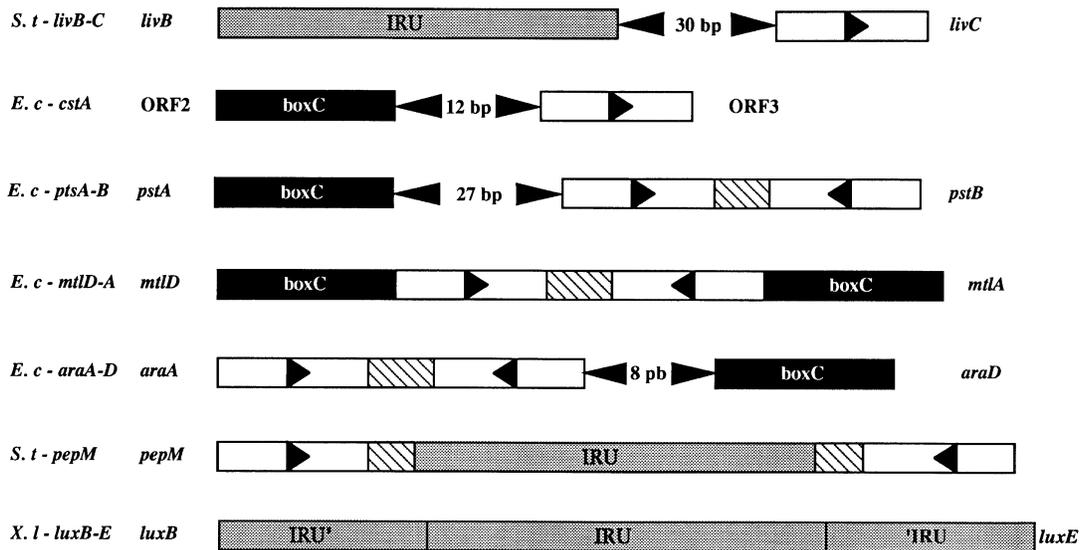
FIGURE 13 Location tropism of extragenic highly repetitive elements. The structures of extragenic regions including several repetitive elements are shown in a schematic form. On the left, the bacteria which contain these elements *(E. c, E. coli; S. t, S. typhimurium; X. l, Xenorhabdus luminescens)* and the names of the adacent genes are indicated. Distances separating the two elements are indicated between two arrows; the names of the elements are indicated except for BIMEs, which are shown in the same form as in other figures.

How can we explain this location tropism? One may suppose that the mechanisms which are responsible for their dispersion are similar or even identical and/or that the insertion sites have common characteristics. Examination of the nucleotide sequence of the intergenic region containing two repeated sequences did not show any obvious homology; this may indicate that it is the structure of the insertion site rather than its sequence which is important. Another possibility would be that the target of the transposition system is the repeated sequence itself. This would account also for the cases where repetitive sequences insert one into the other. At this stage, it is difficult to exclude that this tropism might just reflect the limited number of extragenic regions.

TABLE 7 Compilation of BIME sequences

Flanking gene(s)[a]			GenBank accession no.	BIME structure[b]	Observations
aceA	–	*aceK*	U00006	Y 1 Z^2 S Y	
acs	–	f104	U00006	Z^2 s Y	
aidB	–		L20915	Z S Y	
aldB	–	f107	U00039	Y S Z^2	
alkA	–		K02498	Y S Z^2	
alr	–	*tyrB*	U00006	Y L	
apaH	–	*folA*	X04711	Z^2	
appA	–		M58708	Y	
	–	*appA*	M58708	Z^2	
araA	–	*araD*	M62646	Z^2 S Y s Z^2 S Y s Z^2 S Y	
araJ	–	ORF	M64787	Y S Z^2 1 Y S Z^2	
argA	–	*recD*	U29581	PU	
argD	–	o696	U18997	Y	
argE	–	*ppc*	X55417	Y< 24 >Z^2> 21 <Y< 24 >Z^2> 21 <Y< 24 >Z^2	
aroG	–		J01591	Y	
aroG	–		J01591	Y S Z^2 A	
arsG	–	o260	U00039	Z^2 L Z^1	
artM	–	*artJ*	X67753	Z^2 1 Y	
artJ	–	–	X67753	A Z^1 L Y B	
aslB	–	*aslA*	M90498	Y L Z^2	
aspC	–		X03629	A Z^1 L Y′	The sequence in the database contains only A Z^1; for the end of the sequence see reference 112.
	–	*avtA*	Y00490	B Y L Z^1 A	
barA	–		D10888	B Y L Z^1 A	
bioF	–		J04423	Z^1	Included within the *bioF* gene
caiA	–	*caiB*	X73904	Z^2 S Y	
	–	*caiE*	X73904	Y S Z^2	
cbpA	–		D16500	Z^2 S Y 1 Z^2	
cca	–		M12788	Y S Z^2 s Y	
cdd	–		M60916	Y	
cheZ	–		M13463	Z^2	
chlN	–		M21151	Z^2	
chm	–		X53666	Z^1 S Y	From *E. coli* C
	–	*codB*	X63656	Y 1 Z^2 S Y 1 Z^2	
codA	–	*cynR*	X63656	Y S Z^1	
	–	*cynR*	M93053	Y S Z^1	This BIME is present in a duplicated segment between *codA* and *cynR*. This segment constitutes one end of a discontinuity between *E. coli* and *S. typhimurium* chromosomes (the *lac* loop) (20).
cpdB	–	o163	U14003	PU S Z^2 s Y	
cpxA	–	ORF	M36795	Y 1 Z^2	
cutE	–		X58070	Y	
	–	*cyd*	J03939	B Y L Z^1 A	
cysA	–	*cysM*	M32101	Z^2	
cysE	–		M34333	Y	
cysI	–	*cysH*	Y07525	Z^2	
cysM	–		M32101	A Z^1 L Y B	
dacB	–		X59460	B Y L Z^1 A	
dacC	–		X06480	Z^1	
dapX	–	*purC*	M33928	Y	
dbpA	–		X52647	Z^2	
dedD	–	*dedE*	M68934	Z^2 S Y	
deoC	–	*deoA*	X03224	Z^1 S Y	
dppA	–	*dppB*	L08399	A Z^1 L Y B	
ecoA	–		U00008	Y< 14 >Z^2> 21 <Y< 14 >Z^2> 21 <Y< 14 >Z^2> 21 <Y< 14 >Z^2> 21 <Y< 14 >Z^2> 21 <Y< 14 >Z^2	
exbD	–		M28819	Z^1 L Y 1 Z^1 L Y	
f104	–	f338	U18997	Z^2 S Y	
f159	–	f451	L10328	Y S Z′	
f170	–	*rrfA*	L19201	Z^2	
f199	–	*spf*	L19201	Z^1	
f230	–	*araE*	U29581	Y*	
f239	–	f118	U28377	B Y S Z^1 A	
f254	–	o207	U28379	Y	
f268	–	f432	U29581	A Z^1 L Y B	
f289	–	o128	U14003	B Y L Z^1 A	
f310	–	f848	U28377	Y L Z^1 r Y	
f311	–	f510	U14003	Z^1 S PU	

TABLE 7 *Continued*

Flanking gene(s)[a]			GenBank accession no.	BIME structure[b]	Observations
f342	–	f289	U14003	A Z^1 L Y B	
f481	–	f506	U18997	Y S Z^2 s Y S Z^2 s Y S Z^2 s Y	
f500	–	o361	U14003	Z^2	
f577	–	f283	U00006	Y> 9 <Z^2	
f589	–	dppA	U00039	Y 1 Z^2	
f675	–	f560	U28377	Y*	
fabB	–		M24427	Z^2	
fadA	–	ubiB	M59368	Y S Z^2 1 Y S Z^2 1 Y S Z^2	
fda		– ORF4	X14436	A Z^1 L Y B	
fdhE	–	o80	L19201	B Y L Z^1 A	
fepA	–	entD	X17426	B Y L Z^1 A	
fhuB	–	gsa	D26562	Z^2 S Y	
fimH	–	f477	U14003	Z^2	This intergenic region also contains a box C, at a distance of 18 bp from the PU.
fliC	–	fliA	Z36877	A Z^1 L Y′	From *E. coli* B38. BIMEs with the same structure (A Z^1 L Y B) are found in *Shigella dysenteriae* and *Shigella boydii*.
fpg	–	kdtB	X06036	Y*	
fucO	–		X15025	Y	
gabT	–	gabP	M88334	Y	
	–	galE	X51449	Y	
	–	gatY	X79837	Y B	
gcd	–	ORF	D26562	A Z^1 L Y B	
gdhA	–		X00988	Y L Z^1	
gldA	–	f205	U00006	Y L Z^1	
glnA	–	glnL	K02176	Z^2 S Y	
glnG	–	hemN	L19201	Z^2 1 Y	
glnQ	–		X14180	Z^1 L Y B	
glnS	–		M10187	Y r Z^1 L Y	
glu-tRNA–			J01713	Z^2	
glpD	–	ORFX	D00425	A Z^1 L Y B	
glpK	–	glpX	Z11767	Y	A BIME with the structure Z^2 S Y is present in the same intergenic region of *Shigella flexneri* (Z11766).
gltP	–	f229	U00006	Z^2 L Y s Z^2 L Y s Z^2 L Y s Z^2 L Y s Z^2	
gltS	–	recG	L10328	Z^1 L Y	
glyA	–		J01620	B Y L Z^1 A	
glyK	–		M18393	Y	
gsk	–		D00798	Y 1 Z^2	
gusC	–	ORF	M14641	B Y L Z^1 A	
gutD	–	gutM	X13463	Z^1	
gyrB	–	f135	L10328	Z^1 L Y	
hag	–		M14358	A Z^1 L′	
hpaF	–	hpaH	Z37980	Z^1 S Y	
hpaI	–	hpaX	Z37980	Z^1 S Y ; Y	The two Y are in the same orientation, separated by 14 bp.
htrB	–		X61000	Y*	
hybG	–		U09177	Z^2 1 Y′	
hycH	–		X17506	A Z^1 S Y	
hydA	–		D14422	Y S Z^2	
hydN	–	hydA	D14422	Z^2	
iciA	–		M62865	Y*	
ilvA	–	ilvY	M11689	Z^1	
ilvBN	–	o396	L10328	Z^2	
inaA	–	nrd	K02672	Z^2	See reference 209.
infB	–	P15B	X00153	Y s Z^2	
katE	–		M55161	B Y L Z^1 A	
lacA	–		J01636	Z^1	
lacY	–	lacA	J01636	Y	
lamB	–	malM	X04477	Y s Z^2 S Y	
	–	lctP	L13970	Y	
leuQ	–	f343b	U14003	Z^2 S Y	
livF	–	ugpB	J05516	A Z^1 L Y B	
livJ	–	livK	J05516	B Y L Z^1 A	
lpd	–		V01498	Y*	
lspA	–	ORF1	X54945	Z^2 S Y	
lysV	–		X17321	Z^1	
malE	–	malF	J01648	Y S Z^2 s Y	

TABLE 7   *Continued*

Flanking gene(s)[a]			GenBank accession no.	BIME structure[b]	Observations
malQ	–		M32793	Z^2 S Y	
malT	–		M13585	Y	
mdoH	–		X64197	Z^1 L Y	
melB	–		U14003	A Z^1 L Y B	
metJ	–	f202	L19201	Z^1 L Y B	
metK	–	*galP*	U28377	Y	
metL	–	*metF*	K01547	Y	
mglC	–		M59444	Y 1 Z^2 L Y 1 Z^2	
mgtA	–	f141	U14003	B Y L Z^1 A	
moaE	–		X07420	Y	
motB	–		J01652	Z^2	Included within *motB*
mrcA	–	f186	U18997	B Y L Z^1	
mrr	–	ORF	X54198	Y	
msbB	–		M77039	Y	
msyB	–		X59939	Y L Z^1	
mtlA	–	*mtlD*	V01503	Z^1 L Y	The BIME is flanked by two box C sequences in direct orientation.
mtrA	–	o335	M58338	Z^1	
mutT	–		D26562	Y 1 Z^2 L Y 1 Z^2	
nadR	–	f579	U14003	Y 1 Z^2 L Y	
nagD	–		X14135	A Z^1 L Y B	
narP	–	*yejP*	U00008	A Z^1 L Y B	
	–	*ndh*	V00306	Y 1 Z^1	
ndh	–		V00306	Y	
nlpE	–	*yeaF*	U18345	Y 1 Z^2	
nrdA	–	*nrdB*	K02672	Z^2 s Y S Z^2	
nrdD	–	*ndrG*	L06097	Y S Z^2	
ntr	–		M15328	B Y L Z^1 A	
o56	–	f188	U18997	Y L Z^2	
o128	–	f138	M87049	Y	
o138	–	f97	U14003	Y B	
o147	–	f268	U29581	Y	
o163	–	o384	U18997	Z^2 S Y	
o180	–	f341	U28377	Y S Z^2 ; Z^1	The Z^2 and the Z^1 are separated by 9 bp and are in direct orientation.
o186	–	f470	U14003	A Z^1 L Y B	
o191	–	f346	U18997	Z^2 S Y	
o191	–	o437	U18997	Z^2	
o191	–	f297	U29579	Y s Z^2	
o207	–	f506	U28379	Y S Z^2	
o211	–	f287	U14003	Y 1 Z^2 S Y 1 Z^2 S Y 1 Z^2 S Y 1 Z^2	
o265	–	f361	U18997	Z^2	Presence of a box C in the same intergenic region at a distance of 24 bp
o310	–	f541	U28375	B Y L Z^1 A	
o318	–	o417a	U14003	Y 1 Z^2	
o320	–	o361	L02123	Y	o320 is downstream from *dinG*.
o326	–	f219	U28375	Z^2 L Y 1 Z^2	
o328	–	o130	U18997	B Y L Z^1 A	
o337	–	o440	U00039	Y S Z^2	
o351	–	f582	L19201	Y*	
o361	–	f125	L02123	Z^2 S Y	o361 is downstream from *dinG*.
o378	–	f335	U28377	Y S Z^2 s Y	
o386	–	f271	U28377	Y*	
o417	–	o103	U00039	A Z^1 L Y B	
o421	–	f723	L19201	Z^1 S Y	
o440	–	f472	U00039	Z^1 L Y B-15- Y*	
o499	–	f111	U00039	B Y L Z^1 A	
o548	–	*thdF*	L10328	Z^1	
o569	–	f772	L10328	Y*	
o622	–	*feoA*	U18997	Y	
o698	–	*yjbA*	U00006	Y B	*yjbA* is upstream from *xylE*.
o783	–	o672	U18997	Y S Z^2< 11 >Y S Z^2< 11 >Y S Z^2	
ORF	–	ORF	D10483	Z^2 S Y	The first ORF is downstream from *araC*.
ORF1	–		M26934	S Z^2	ORF1 is located downstream from *ansA*.
ORF2	–	ORF3	X52904	Y	ORF2 is located downstream from *cstA*.
ORFX	–	ORFY	M96791	Y 1 Z	ORFY is located upstream from *g30k*.
ORF	–		K01197	Z^1 S Y	ORF is located downstream from *tyrT*.
ORF122	–	ORF153	L20897	Z^2	ORF122 is located downstream from *purT*.
pepD	–		M34034	Y*	
pepE	–	o543	U00006	Z^2< 11 >Y*	

TABLE 7 *Continued*

Flanking gene(s)[a]			GenBank accession no.	BIME structure[b]	Observations
pfkA	–	*sbp*	X02519	A Z^1 L Y B	
pgi	–	o80	U00006	Y s Z^2	
pgm	–		U08369	A Z^1 L Y B	
pheU	–		X00912	Z^1	
phnA	–	*phnB*	J05260	Y 1 Z^2 S Y 1 Z^2 S Y 1 Z^2 S Y 1 Z^2 S Y 1 Z^2	This sequence originates from *E. coli* B.
phnD	–	*phnE*	D90227	Z^2 1 Y	
phnK	–	*phnL*	D90227	Y S Z^2	
phoE	–		V00316	Y*	
phoH	–	ORF	D10391	Z^2 S Y	
pho	–	*phoW*	L10328	Y B	
phr	–		K01299	Y L'	
plsB	–	*ubiA*	K00127	Y L Z^1 1 Y*	
pmsR	–		M89992	B Y L Z^1 A	
polA	–	ORF	J01663	Z^1 L Y	
priC	–		D13958	A Z^1 L Y'	
proP	–	*basS*	U14003	Z^2	
proX	–		M24856	Y 1 Z^2	
pstA	–	*pstB*	L10328	Y Z^1	The BIME is flanked by a box C.
purK	–		M19657	B Y S Z^2 1 Y/2 S Z^2	
pyrE	–	ORF	L10328	Y*	
rcsC	–	*rcsB*	M28242	Y 1 Z^2 L Y 1 Z^2	
recD	–		X04582	Y	
rep	–		X04794	Z^2	
	–	*rf2*	M11520	Y	
rhaA	–	*rhaD*	X60472	Y S Z^2 s Y S Z^2 s Y S Z^2 s Y S Z^2 s Y S Z^2 s Y	
rhlE	–	f160	L02123	A Z^1 L Y B	
rnpB	–	ORFX	D90212	B Y L Z^1 A	
rol	–		M89934	Z^2 S Y	
rplY	–		D13326	Y	
rpoD	–		J01687	Y	
ruvB	–		M21298	Y 1 Z^2 S Y	
sdhB	–	*sucA*	J01619	Z^2 s Y S Z^1	
sodB	–		J03511	A Z^1 S Y B	
speB	–		M32363	r Y S Z^2 r Y	
srmB	–		X14152	A Z^1 L Y B	
sseB	–		D10496	Z^1 L Y	
sucB	–	*sucD*	J01619	Y s Z^2 S Y s Z^2	
thrC	–	ORF	J01706	Z^2 L Y	
tktB	–		S65463	Z' Z^2 'Z	The Z^2 PU is included within another PU.
torA	–		X73888	Z^2 1 Y S Z^2	
trmA	–	o119	U00006	Y*	
trpR	–	ORF	J01715	Z^1 1 Y L Z^2	
trpS	–	f228	U18997	A Z^1 > 25 < Y B	
tsx	–	ORF6	M57685	B Y L Z^1 A	ORF6 is downstream from *secF*.
tyrA	–	*pheA*	X02137	Y*	
ubiG	–	ORFX	Y00544	Z^2 S Y	This region (around 48 min) has been studied in several strains, where the sequences downstream of *ubiG* differ: in M87509, the BIME contains 3 PUs, and ORFX is *acs,* located around 90 min in *E. coli* MG1655.
ugpB	–	*ugpA*	X13141	Z^1 S Y	
ung	–		J03725	Y	
ushA	–		X03895	Y*	
uvrB	–		X03722	A Z^1 L Y B	
uvrD	–	ORF	L02122	Z^2 L Y	
uxuB	–	*uxuR*	U14003	Y	
vacB	–	o243	U14003	B Y L Z^1	
valS	–	o204b	U14003	Y*	
x96	–		M95096	Z^2 1 Y	
xerB	–	ORF13	X15130	Y S Z^2	
xylR	–	f274b	U00039	Z^2 1 Y, Z^2	The last Z^2 is separated from the Y PU by 49 bp and does not seem to belong to the same BIME.
yehZ	–	*yohA*	U00007	Y	
yohD	–	*yohE*	U00007	Y	
yohK	–		U00007	Y	
yeiI	–	*yeiJ*	U00007	Y > 24 < Z^1	
yeiL	–	*yeiM*	U00007	Z^2	
*rhsD**			X60997	Z^1 L Y 1 Z^2 S Y	Segment replacing *rhsD* element in *E. coli* ECOR 39 (166)

TABLE 7   *Continued*

Flanking gene(s)[a]	GenBank accession no.	BIME structure[b]	Observations
pSCIS1IN*	M25018	'Y S Z^2 s Y	Target region of IS1–mediated type I deletions (185)
	I01818	A Z^1 L Y B'	From "unclassified" organisms. The sequences of the BIME motifs are typical of E. coli or close relatives.
	I03295	A Z^1 L	
– adk	L26246	'Y> 12 <D	First PU incompletely sequenced
adk – visA	L26246	$Z^1$1 Y	In *E. coli,* box C at the same location
alr –	M12847	D	In *E. coli,* Y L
araA – araD	M11047	Z^1> 4 <Y< 8 <Z^1	In *E. coli,* BIME-2 with 6 PUs
argA – recD	–	Z^1> 14 <Y/2	See reference 182.
cchA – cchB	U18560	Y< 11 >Z^2	
cheW –	J02656	PU	
cheZ – flhB	M16691	Z^1> 7 <PU	In *E. coli,* Z^2
cpsB – cpsG	X59886	D	
crr – ORF287	U11243	D< 10 >Z^1	
cysI – cysH	M23007	Z^1	In *E. coli,* Z^2
envZ –	X12374	Z^2	
eutG – eutH	U18560	Y< 10 >Z^1	
– hemA	J04243	Z^1	
hemL –	M60064	D	
hemN –	U06779	Z^1> 13 <D	In *E. coli,* BIME-2 with 2 PUs
hisG – hisD	J01804	Z^1> 23 <D	
hisJ – hisQ	V01373	Z^1> 8 <D	
ilvH – fruR	X55456	Z^2> 13 <Z^1	
lamB – malM	X54292	D	In *E. coli,* BIME-2 with 3 PUs
leuD –	X02528	Z^1	
lexA –	X63002	Y/2<3>Z^1> 27 <Y	
– livB	D00478	Y<3>Z^1	
livB – livC	D00478	Z^1	An IRU is distant from 30 bp
malE – malF	X54292	Y> 6 <Z^1	In *E. coli,* BIME-2 with 3 PUs
melB –	S41630	Z^1	In *E. coli,* BIME-1
– murB	J14816	D	
mviM – mviN	Z26133	Z^2<5>Y	
mviN – flgN	Z26133	Z^1> 5 <Y	
– nadR	M85181	Z^2> 19 <Z^1	
nadR –	M85181	Z^2> 29 <PU	In *E. coli,* BIME with 3 PUs
nrdA – nrdB	X72948	Z^1> 11 <Y	In *E. coli,* BIME-2 with 3 PUs
– nrdE	X73226	'Y	
– ORF1	L13395	PU	ORF1 is in the 5' region of *pmrA*.
parC – parF	M68936	Y<5>Z^2	
pepM – ORF1	X55778	Z^2> 137 <Y	An IRU is inserted between the 2 PUs.
trpR –	L13768	Z^2	In *E. coli,* BIME with 3 PUs
ushA^0 –	X04651	Z^1> 12 <Y	In *E. coli,* Y*
uvrD – corA	L11043	PU> 20 <D	In *E. coli,* BIME with 2 PUs

[a]The first 256 BIMEs are from *E. coli* or close relatives, while the last 39 BIMEs are from *S. typhimurium.* Stars indicate two regions where BIMEs are not defined according to flanking genes.

[b]> indicates PU orientation, from tail to head (and inversely for <). *S. typhimurium* PUs are named according to Fig. 7: group A, Y; group B, Z^1; group C, Z^2; group D, D. When the sequence of the PU is not related to any of these groups, it is named PU.

Exchange of Repetitive Sequences. Another intriguing type of relations between four of the families of repetitive sequences has been found. Corresponding intergenic regions of two different bacteria are sometimes occupied by two different sequences. Sharples and Lloyd (179) noticed that one IRU of *S. typhimurium* is found in the intergenic region *glnA-L* instead of the BIME which is found in the same region in *E. coli.* Between the genes *malE* and *malF* from *E. coli, Enterobacter aerogenes, K. pneumoniae,* and *S. typhimurium,* there is a BIME sequence. However, the corresponding region of *Klebsiella oxytoca* contains an IRU (Bachellier et al., unpublished data). The region ahead of *envM* carries an RSA in *S. typhimurium* and two box C sequences in *E. coli.* The *araA-D* region of *K. pneumoniae* contains a box C (see above and Fig. 10), while *S. typhimurium* and *E. coli* carry a BIME at the same place; it is remarkable that in *E. coli* ECOR 8 there is at the same site one BIME and one box C. Finding one or the other type of these sequences at the same site could suggest that they have similar functions (interaction with proteins or need for a secondary structure at a precise site) or, as supposed

previously, that they insert at the same sites. In the case of the region located between *malE* and *malF,* one would have to think that there was independent insertion of the repetitive element in each bacterium or that there was a replacement of the BIME by the IRU in *K. oxytoca.* Two other examples of exchanges have been reported. (i) In *E. coli* K-12, a BIME (Y*) is located in the *pyrE-ttk* intergenic region (Table 7), while a retron is found at the same location in the *E. coli* ECOR 70, 71, and 72 (73). Retrons are not repetitive sequences but are chromosomal elements encoding a reverse transcriptase (reviewed in references 94 and 202). (ii) In *E. coli* ECOR 39, a BIME is located at the same place as the *RhsD* element of *E. coli* K-12 (166) (see above).

A Superfamily? BIMEs, box C sequences, RSAs, IRUs, *iap* sequences, and Ter sequences constitute six families of repetitive sequences which are found in the genomes of a number of gram-negative bacteria. It has been shown that at least some of these repetitive sequences can be useful for the typing of bacteria (35, 204). They share a number of characteristics: they are scattered over the bacterial chromosome (except *iap* sequences), all are transcribed, all are imperfectly palindromic, and four of them present some tropism in their location. Their sizes are less than the size of ISs, and no mechanism for their transposition has yet been demonstrated. It appears that they can be exchanged locally, which may reveal a function or a dispersion mechanism which would be common. All these similarities lead us to propose that these families define a superfamily of extragenic elements which may play a critical role in genome function and evolution.

PERSPECTIVES

Although our knowledge is yet severely limited, the distribution and interaction of both large and small repeated sequences in *E. coli* and *S. typhimurium* populations is an important source of genetic plasticity. More studies of repeated sequences may provide greater insights into the degree of natural diversity and into the selective pressures that maintain the profiles. Such studies might also give clues as to the natural mechanisms that rearrange and/or redistribute the diverse examples of repeated sequences that have been discovered.

ACKNOWLEDGMENTS

LITERATURE CITED

1. **Albertini, A. M., M. Hofer, M. P. Calos, and J. H. Miller.** 1982. On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. *Cell* **29:**319–328.
2. **An, G., and J. D. Frisen.** 1980. The nucleotide sequence of *tufB* and four nearby tRNA structural genes of *Escherichia coli. Gene* **12:**33–39.
3. **Anderson, R. P., and J. R. Roth.** 1977. Tandem genetic duplications in phage and bacteria. *Annu. Rev. Microbiol.* **31:**473–505.
4. **Anderson, R. P., and J. R. Roth.** 1978. Tandem chromosomal duplications in *Salmonella typhimurium:* fusion of histidine genes to novel promoters. *J. Mol. Biol.* **119:**147–166.
5. **Anderson, R. P., and J. R. Roth.** 1978. Gene duplication in bacteria: alteration of gene dosage by sister-chromosome exchanges. *Cold Spring Harbor Symp. Quant. Biol.* **119:**147–166.
6. **Bachellier, S., D. Perrin, M. Hofnung, and E. Gilson.** 1993. Bacterial interspersed mosaic elements (BIMEs) are present in the genome of *Klebsiella. Mol. Microbiol.* **7:**537–544.
7. **Bachellier, S., D. Perrin, M. Hofnung, and E. Gilson.** Two major bacterial interspersed mosaic

element (BIMEs) families with different gyrase and integration host factor binding abilities. Submitted for publication.

8. **Bachellier, S., W. Saurin, D. Perrin, M. Hofnung, and E. Gilson.** 1994. Structural and functional diversity among bacterial interspersed mosaic elements (BIMEs). *Mol. Microbiol.* **12:**61–70.

9. **Baron, L. S., Jr., P. Gemski, E. M. Johnson, and J. A. Wohlhieter.** 1968. Intergeneric bacterial matings. *Bacteriol. Rev.* **32:**362–369.

10. **Bates, A. D., and A. Maxwell.** 1989. DNA gyrase can supercoil DNA circles as small as 174 base pairs. *EMBO J.* **8:**1861–1866.

11. **Becerril, B., F. Valle, E. Merino, L. Riba, and F. Bolivar.** 1985. Repetitive extragenic palindromic (REP) sequences in the *Escherichia coli gdhA* gene. *Gene* **37:**53–62.

12. **Bergler, H., G. Högenauer, and F. Turnowski.** 1992. Sequences of the *envM* gene and of two mutated alleles in *Escherichia coli. J. Gen. Microbiol.* **138:**2093–2100.

13. **Bi, X., and L. F. Liu.** 1994. *recA*-independent and *recA*-dependent intramolecular plasmid recombination. Differential homology requirement and distance effect. *J. Mol. Biol.* **235:**414–423.

14. **Bisercic, M., and H. Ochman.** 1993. The ancestry of insertion sequences common to *Escherichia coli* and *Salmonella typhimurium. J. Bacteriol.* **175:**7863–7868.

15. **Bisercic, M., and H. Ochman.** 1993. Natural populations of *Escherichia coli* and *Salmonella typhimurium* harbor the same classes of insertion sequences. *Genetics* **133:**449–454.

16. **Blaisdell, B. E., K. E. Rudd, A. Matin, and S. Karlin.** 1993. Significant dispersed recurrent DNA sequences in the *Escherichia coli* genome: several new groups. *J. Mol. Biol.* **229:**833–848.

17. **Blattner, F. R., V. Burland, G. Plunkett, III, H. J. Sofia, and D. L. Daniels.** 1993. Analysis of the *Escherichia coli* genome. IV. DNA sequence of the region from 89.2 to 92.8 minutes. *Nucleic Acids Res.* **21:**5408–5417.

18. **Boccard, F., and P. Prentki.** 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.

19. **Britten, R. J., and D. E. Kohne.** 1968. Repeated sequences in DNA. *Science* **161:**529–540.

20. **Buvinger, W. E., K. A. Lampel, R. J. Bojanowski, and M. Riley.** 1984. Location and analysis of nucleotide sequences at one end of a putative *lac* transposon in the *Escherichia coli* chromosome. *J. Bacteriol.* **159:**618–623.

21. **Campbell, A.** 1963. Segregants from lysogenic heterogenotes carrying recombinant lambda prophages. *Virology* **20:**344–356.

22. **Campbell, A.** 1965. The steric effect in lysogenization by bacteriophage lambda I. Lysogenization of a partially diploid stain of *Escherichia coli* K12. *Virology* **27:**329–339.

23. **Campbell, A.** 1981. Evolutionary significance of accessory DNA elements in bacteria. *Annu. Rev. Microbiol.* **35:**55–83.

24. **Campbell, A. M.** 1992. Chromosomal insertion sites for phages and plasmids. *J. Bacteriol.* **174:**7495–7499.

25. **Capage, M., and C. W. Hill.** 1979. Preferential unequal recombination in the *glyS* region of the *Escherichia coli* chromosome. *J. Mol. Biol.* **127:**73–87.

26. **Chédin, F., E. Dervyn, R. Dervyn, S. D. Ehrlich, and P. Noirot.** 1994. Frequency of deletion formation decreases exponentially with distance between short direct repeats. *Mol. Microbiol.* **12:**561–569.

27. **Condon, C., S. French, C. Squires, and C. L. Squires.** 1993. Depletion of functional ribosomal RNA operons in *Escherichia coli* causes increased expression of the remaining intact copies. *EMBO J.* **12:**4305–4315.

28. **Condon, C., J. Philips, Z.-Y. Fu, C. Squires, and C. L. Squires.** 1992. Comparison of the expression of the seven ribosomal RNA operons in *Escherichia coli. EMBO J.* **11:**4175–4185.

29. **Curtiss, R., III.** 1964. A stable partial diploid strain of *Escherichia coli. Genetics* **50:**679–694.

30. **Davidson, N., R. C. Deonier, S. Hu, and E. Ohtsubo.** 1975. Electron microscope heteroduplex studies of sequence relations among plasmids of *Escherichia coli.* X. Deoxyribonucleic acid sequence organization of F and of F-primes, and the sequences involved in Hfr formation, p. 56–65. *In* D. Schlessinger (ed.), *Microbiology—1974.* American Society for Microbiology, Washington, D.C.

31. **de Massy, B., S. Bejar, J. Louarn, J.-M. Louarn, and J.-P. Bouche.** 1987. Inhibition of replication forks exiting the terminus region of the *Escherichia coli* chromosome occurs at two loci separated by 5 min. *Proc. Natl. Acad. Sci. USA* **84:**1759–1763.

32. **Deonier, R. C., E. Ohtsubo, H. J. Lee, and N. Davidson.** 1974. Electron microscope heteroduplex studies of sequence relations among plasmids of *Escherichia coli.* VII. Mapping the ribosomal RNA genes of plasmid F14. *J. Mol. Biol.* **89:**619–629.

33. **Dianov, G. L., A. V. Kuzminov, A. V. Mazin, and R. I. Salganik.** 1991. Molecular mechanisms of deletion formation in *Escherichia coli* plasmids I. Deletion formation mediated by long direct repeats. *Mol. Gen. Genet.* **228:**153–159.

34. **Dimpfl, J., and H. Echols.** 1989. Duplication mutation as an SOS response in *Escherichia coli:* enhanced duplication formation by a constitutively activated RecA. *Genetics* **123:**255–260.

35. **Dimri, G. P., K. E. Rudd, M. K. Morgan, H. Bayat, and G. Ferro-Luzzi Ames.** 1992. Physical mapping of repetitive extragenic palindromic sequences in *Escherichia coli* and phylogenetic distribution among *Escherichia coli* strains and other enteric bacteria. *J. Bacteriol.* **174:**4583–4593.

36. **Doolittle, W. F.** 1994. RNA-mediated gene conversion? *Trends Genet.* **1:**64–65.

37. **Dover, G., S. Brown, E. Coen, J. Dallas, T. Strachan, and M. Trick.** 1982. The dynamics of genome evolution and species differentiation, p. 343–372. *In* G. A. Dover and R. B. Flavell (ed.), *Genome Evolution.* Academic Press, London.

38. **Drlica, K., and J. Rouvière-Yaniv.** 1987. Histonelike proteins of bacteria. *Microbiol. Rev.* **51:**301–319.

39. **Duester, G., R. K. Campen, and W. M. Holmes.** 1981. Nucleotide sequence of an *Escherichia coli* tRNA^{Leu1} operon and identification of the transcription promoter signal. *Nucleic Acids Res.* **9:**2121–2139.

40. **Echols, H.** 1986. Multiple DNA-protein interactions governing high-precision DNA transactions. *Science* **233:**1050–1956.

41. **Edlund, T., and S. Normark.** 1981. Recombination between short DNA homologies causes tandem duplication. *Nature* (London) **292:**269–271.

42. **Ellwood, M., and M. Nomura.** 1980. Deletion of a ribosomal ribonucleic acid operon in *Escherichia coli. J. Bacteriol.* **143:**1077–1080.

43. **Ellwood, M., and M. Nomura.** 1982. Chromosomal locations of the genes for rRNA in *Escherichia coli* K-12. *J. Bacteriol.* **149:**458–468.

44. **Engelhorn, M., F. Boccard, C. Murtin, P. Prentki, and J. Geiselmann.** 1995. In vivo interaction of the *Escherichia coli* integration host factor with the specific binding sites. *Nucleic Acids Res.* **23:**2959–2965.

45. **Ennis, D. G., S. K. Amundsen, and G. R. Smith.** 1987. Genetic functions promoting homologous recombination in *Escherichia coli:* a study of inversions in phage λ. *Genetics* **115:**11–24.

46. **Fanning, T. G., and M. F. Singer.** 1987. LINE-1: a mammalian transposable element. *Biochim. Biophys. Acta* **910:**203–210.

47. **Fellner, P., C. Ehresmann, and J. P. Ebel.** 1970. Nucleotide sequences present within the 16S ribosomal RNA of *Escherichia coli. Nature* (London) **225:**26–29.

48. **Ferat, J.-L., M. Le Gouar, and F. Michel.** 1994. Multiple group II self-splicing introns in mobile DNA from *Escherichia coli. C. R. Acad. Sci.* **317:**141–148.

49. **Feulner, G., J. A. Gray, J. A. Kirschman, A. F. Lehner, A. B. Sadosky, D. A. Vlazny, J. Zhang, S. Zhao, and C. W. Hill.** 1990. Structure of the *rhsA* locus from *Escherichia coli* K-12 and comparison of *rhsA* with other members of the *rhs* multigene family. *J. Bacteriol.* **172:**446–456.

50. **Foster, S. J.** 1993. Molecular analysis of three major wall-associated proteins of *Bacillus subtilis* 168: evidence for processing of the product of a gene encoding a 258 kDa precursor two-domain ligand-binding protein. *Mol. Microbiol.* **8:**299–310.

51. **Fournier, M. J., and H. Ozeki.** 1985. Structure and organization of the transfer ribonucleic acid genes of *Escherichia coli* K-12. *Microbiol. Rev.* **49:**379–397.

52. **Friedman, D. I.** 1988. Integration host factor: a protein for all reasons. *Cell* **55:**545–554.

53. **Friedman, D. I., E. J. Olson, D. Carver, and M. Gellert.** 1984. Synergistic effect of *himA* and

gyrB mutations: evidence that Him functions control expression of *ilv* and *xyl* genes. *J. Bacteriol.* **157:**484–489.

54. **Galas, D. J., and M. Chandler.** 1989. Bacterial insertion sequences, p. 109–162. *In* D. E. Berg and M. M. Howe (ed.), *Mobile DNA.* American Society for Microbiology, Washington, D.C.

55. **Gamas, P., M. G. Chandler, P. Prentki, and D. J. Galas.** 1987. *Escherichia coli* integration host factor binds specifically to the ends of the insertion sequence IS*1* and to its major insertion hot-spot in pBR322. *J. Mol. Biol.* **195:**261–272.

56. **Gilson, E., S. Bachellier, S. Perrin, D. Perrin, P. A. D. Grimont, F. Grimont, and M. Hofnung.** 1990. Palindromic units highly repetitive DNA sequences exhibit species specificity within *Enterobacteriaceae. Res. Microbiol.* **141:**1103–1116.

57. **Gilson, E., J.-M. Clément, D. Brutlag, and M. Hofnung.** 1984. A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli. EMBO J.* **3:**1417–1421.

58. **Gilson, E., J.-M. Clément, D. Perrin, and M. Hofnung.** 1987. Palindromic units: a case of highly repetitive DNA sequences in bacteria. *Trends Genet.* **3:**226–230.

59. **Gilson, E., T. Laroche, and S. Gasser.** 1993. Telomeres and the functional architecture of the nucleus. *Trends Cell. Biol.* **3:**128–134.

60. **Gilson, E., D. Perrin, J.-M. Clément, S. Szmelcman, E. Dassa, and M. Hofnung.** 1986. Palindromic units from *E. coli* as binding sites for a chromoid-associated protein. *FEBS Lett.* **206:**323–328.

61. **Gilson, E., D. Perrin, and M. Hofnung.** 1990. DNA polymerase I and a protein complex bind specifically to *E. coli* palindromic units highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Res.* **18:**3941–3952.

62. **Gilson, E., D. Perrin, W. Saurin, and M. Hofnung.** 1987. Species specificity of bacterial palindromic units. *J. Mol. Evol.* **25:**371–373.

63. **Gilson, E., J.-P. Rousset, J.-M. Clément, and M. Hofnung.** 1986. A subfamily of *E. coli* palindromic units implicated in transcription termination? *Ann. Inst. Pasteur Microbiol.* **137 B:**259–270.

64. **Gilson, E., W. Saurin, D. Perrin, S. Bachellier, and M. Hofnung.** 1991. Palindromic units are part of a new bacterial interspersed mosaic element (BIME). *Nucleic Acids Res.* **19:**1375–1383.

65. **Gilson, E., W. Saurin, D. Perrin, S. Bachellier, and M. Hofnung.** 1991. The BIME family of bacterial highly repetitive sequences. *Res. Microbiol.* **142:**217–222.

66. **Goodman, S. D., and J. J. Scocca.** 1988. Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae. Proc. Natl. Acad. Sci. USA* **85:**6982–6986.

67. **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.

68. **Guest, J. R.** 1981. Hybrid plasmids containing the citrate synthase gene (*gltA*) of *Escherichia coli* K-12. *J. Gen. Microbiol.* **124:**17–23.

69. **Gustafson, C. E., S. Chu, and T. J. Trust.** 1994. Mutagenesis of the paracrystalline surface protein array of *Aeromonas salmonicida* by endogenous insertion elements. *J. Mol. Biol.* **237:**452–463.

70. **Harvey, S., and C. W. Hill.** 1990. Exchange of spacer regions between rRNA operons in *Escherichia coli. Genetics* **125:**683–690.

71. **Harvey, S., C. W. Hill, C. Squires, and C. L. Squires.** 1988. Loss of the spacer loop sequence from the *rrnB* operon in the *Escherichia coli* K-12 subline that bears the *relA1* mutation. *J. Bacteriol.* **170:**1235–1238.

72. **Heath, J. D., and G. M. Weinstock.** 1991. Tandem duplications of the *lac* region of the *Escherichia coli* chromosome. *Biochimie* **73:**343–352.

73. **Herzer, P. J., M. Inouye, and S. Inouye.** 1992. Retron-EC107 is inserted into the *Escherichia coli* genome by replacing a palindromic 34bp intergenic sequence. *Mol. Microbiol.* **6:**355–361.

74. **Herzer, P. J., S. Inouye, M. Inouye, and T. S. Whittam.** 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli. J.*

Bacteriol. **172:**6175–6181.

75. **Higgins, C. F., G. Ferro-Luzzi Ames, W. M. Barnes, J.-M. Clément, and M. Hofnung.** 1982. A novel intercistronic regulatory element of prokaryotic operons. *Nature* (London) **298:**760–762.

76. **Hill, C. W., and G. Combriato.** 1973. Genetic duplications induced at very high frequency by ultraviolet irradiation in *Escherichia coli. Mol. Gen. Genet.* **127:**197–214.

77. **Hill, C. W., J. Foulds, L. Soll, and P. Berg.** 1969. Instability of a missense suppressor resulting from a duplication of genetic material. *J. Mol. Biol.* **39:**563–581.

78. **Hill, C. W., R. H. Grafstrom, B. W. Harnish, and B. S. Hillman.** 1977. Tandem duplications resulting from recombination between ribosomal RNA genes in *Escherichia coli. J. Mol. Biol.* **116:**407–428.

79. **Hill, C. W., and J. A. Gray.** 1988. Effects of chromosomal inversion on cell fitness in *Escherichia coli* K-12. *Genetics* **119:**771–778.

80. **Hill, C. W., J. A. Gray, and H. Brody.** 1989. Use of the isocitrate dehydrogenase structural gene for attachment of e14 in *Escherichia coli* K-12. *J. Bacteriol.* **171:**4083–4084.

81. **Hill, C. W., and B. W. Harnish.** 1981. Inversions between ribosomal RNA genes of *Escherichia coli. Proc. Natl. Acad. Sci. USA* **78:**7069–7072.

82. **Hill, C. W., and B. W. Harnish.** 1982. Transposition of a chromosomal segment bounded by redundant rRNA genes into other rRNA genes in *Escherichia coli. J. Bacteriol.* **149:**449–457.

83. **Hill, C. W., S. Harvey, and J. A. Gray.** 1990. Recombination between rRNA genes in *Escherichia coli* and *Salmonella typhimurium,* p. 335–340. *In* K. Drlica and M. Riley (ed.), *The Bacterial Chromosome.* American Society for Microbiology, Washington, D.C.

84. **Hill, C. W., C. H. Sandt, and D. A. Vlazny.** 1994. *Rhs* elements of *Escherichia coli:* a family of genetic composites each encoding a large mosaic protein. *Mol. Microbiol.* **12:**865–871.

85. **Hill, C. W., C. Squires, and J. Carbon.** 1970. Glycine transfer RNA of *Escherichia coli.* I. Structural genes for two glycine tRNA species. *J. Mol. Biol.* **52:**557–569.

86. **Hoffmann, G. R., and R. W. Morgan.** 1976. The effect of ultraviolet light on the frequency of a genetic duplication in *Salmonella typhimurium. Radiat. Res.* **67:**114–119.

87. **Hoffmann, G. R., R. W. Morgan, and R. C. Harvey.** 1978. Effect of chemical and physical mutagens on the frequency of large genetic duplications in *Salmonella typhimurium:* induction of duplications. *Mutat. Res.* **52:**73–80.

88. **Horiuchi, T., S. Horiuchi, and A. Novick.** 1963. The genetic basis of hypersynthesis of β-galactosidase. *Genetics* **48:**157–169.

89. **Howard, M. T., M. P. Lee, T.-S. Hsieh, and J. D. Griffith.** 1991. *Drosophila* topoisomerase II-DNA interactions are affected by DNA structure. *J. Mol. Biol.* **217:**53–62.

90. **Hu, M., and R. C. Deonier.** 1981. Mapping of IS*1* elements flanking the *argF* gene region of the *Escherichia coli* K-12 chromosome. *Mol. Gen. Genet.* **181:**222–229.

91. **Hulton, C. S. J., C. F. Higgins, and P. M. Sharp.** 1991. ERIC sequences: a novel family of repetitive elements in the genomes of *Escherichia coli, Salmonella typhimurium* and other enterobacteria. *Mol. Microbiol.* **5:**825–834.

92. **Ilgen, C., L. L. Kirk, and J. Carbon.** 1976. Isolation and characterization of large transfer ribonucleic acid precursors from *Escherichia coli. J. Biol. Chem.* **251:**922–929.

93. **Ingram, V. M.** 1963. *The Hemoglobins in Genetics and Evolution.* Columbia University Press, New York.

94. **Inouye, M., and S. Inouye.** 1992. Retrons and multicopy single-stranded DNA. *J. Bacteriol.* **174:**2419–2424.

95. **Ishino, Y., W. Shinagawa, K. Makino, M. Amemura, and A. Nakata.** 1987. Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli,* and identification of the gene product. *J. Bacteriol.* **169:**5429–5433.

96. **Jinks-Robertson, S., R. L. Gourse, and M. Nomura.** 1983. Expression of rRNA and tRNA genes in *Escherichia coli:* evidence for feedback regulation by products of rRNA operons. *Cell* **33:**865–876.

97. **Jones, I. M., S. B. Primose, and S. D. Ehrlich.** 1982. Recombination between short direct repeats in a RecA host. *Mol. Gen. Genet.* **188:**486–489.

98. **Käs, E., and U. K. Laemmli.** 1992. *In vivo* topoisomerase II cleavage of the *Drosophila* histone and satellite III repeats: DNA sequence and structural characteristics. *EMBO J.* **11:**705–716.

99. **Kenri, T., F. Imamoto, and Y. Kano.** 1994. Three tandemly repeated structural genes encoding tRNA$_{f1}^{Met}$ in the *metZ* operon of *Escherichia coli* K-12. *Gene* **138:**261–262.

100. **King, S. R., M. A. Krolewski, S. L. Marvo, P. J. Lipson, K. L. Pogue-Geile, J. H. Chung, and S. R. Jaskunas.** 1982. Nucleotide sequence analysis of in vivo recombinants between bacteriophage λ DNA and pBR322. *Mol. Gen. Genet.* **186:**548–557.

101. **King, S. R., and J. P. Richardson.** 1986. Role of homology and pathway specificity for recombination between plasmids and bacteriophage λ. *Mol. Gen. Genet.* **204:**141–147.

102. **Kirkegaard, K., and J. C. Wang.** 1981. Mapping the topography of DNA wrapped around gyrase by nucleolytic and chemical probing of complexes of unique DNA sequences. *Cell* **23:**721–729.

103. **Komine, Y., T. Adachi, H. Inokuchi, and H. Ozeki.** 1990. Genomic organization and physical mapping of the transfer RNA genes in *Escherichia coli* K12. *J. Mol. Biol.* **212:**579–598.

104. **Komine, Y., and H. Inokuchi.** 1991. Physical map locations of the genes that encode small stable RNAs in *Escherichia coli. J. Bacteriol.* **173:**5252.

105. **Komine, Y., and H. Inokuchi.** 1991. Precise mapping of the *rnpB* gene encoding the RNA component of RNase P in *Escherichia coli* K-12. *J. Bacteriol.* **173:**1813–1816.

106. **Komoda, Y., M. Enomoto, and A. Tominaga.** 1991. Large inversion in *Escherichia coli* K-12 1485IN between inversely oriented IS*3* elements near *lac* and *cdd. Genetics* **129:**639–645.

107. **Konrad, E. B.** 1977. Method for the isolation of *Escherichia coli* mutants with enhanced recombination between chromosomal duplications. *J. Bacteriol.* **130:**167–172.

108. **Kornberg, A., L. L. Bertsch, J. F. Jackson, and H. G. Khorana.** 1964. Enzymatic synthesis of deoxyribonucleic acid. XVI. Oligonucleotides as templates and the mechanism of their replication. *Proc. Natl. Acad. Sci. USA* **51:**315–323.

109. **Krawiec, S., and M. Riley.** 1990. Organization of the bacterial chromosome. *Microbiol. Rev.* **54:**502–539.

110. **Kroll, J. S., B. M. Loynds, and P. Langford.** 1992. Palindromic *Haemophilus* DNA uptake sequences in presumed transcriptional terminators from *H. influenzae* and *H. parainfluenzae. Gene* **114:**151–152.

111. **Kunisawa, T., and M. Nakamura.** 1991. Identification of regulatory building blocks in the *Escherichia coli* genome. *Protein Sequence Data Anal.* **4:**43–47.

112. **Kuramitsu, S., S. Okuno, T. Ogawa, H. Ogawa, and H. Kagamiyama.** 1985. Aspartate aminotransferase of *Escherichia coli:* nucleotide sequence of the *aspC* gene. *J. Bacteriol.* **97:**1259–1262.

113. **Lampson, B. C., J. Sun, M.-Y. Hsu, J. Vallejo-Ramirez, S. Inouye, and M. Inouye.** 1989. Reverse transcriptase in a clinical strain of *Escherichia coli:* production of branched RNA-linked ms DNA. *Science* **243:**1033–1038.

114. **Lawrence, J. G., D. E. Dykhuizen, R. F. Dubose, and D. L. Hartl.** 1989. Phylogenetic analysis using insertion sequence fingerprinting in *Escherichia coli. Mol. Biol. Evol.* **6:**1–14.

115. **Lawrence, J. G., H. Ochman, and D. L. Hartl.** 1992. The evolution of insertion sequences within enteric bacteria. *Genetics* **131:**9–20.

116. **Legrain, C., V. Stalon, and N. Glansdorff.** 1976. *Escherichia coli* ornithine carbamoyltransferase isoenzymes: evolutionary significance and the isolation of λ*argF* and λ*argI* transducing bacteriophages. *J. Bacteriol.* **128:**35–38.

117. **Lehner, A. F., S. Harvey, and C. W. Hill.** 1984. Mapping and spacer identification of rRNA operons of *Salmonella typhimurium. J. Bacteriol.* **160:**682–686.

118. **Lehner, A. F., and C. W. Hill.** 1980. Involvement of ribosomal ribonucleic acid operons in *Salmonella typhimurium* chromosomal rearrangements. *J. Bacteriol.* **143:**492–498.

119. **Lehner, A. F., and C. W. Hill.** 1985. Merodiploidy in *Escherichia coli-Salmonella typhimurium* crosses: the role of unequal recombination between ribosomal RNA genes. *Genetics* **110:**365–380.

120. **Lewis, E. B.** 1951. Pseudoparallelism and gene evolution. *Cold Spring Harbor Symp. Quant. Biol.* **16:**159–174.

121. **Lim, D., and W. K. Maas.** 1989. Reverse transcriptase-dependent synthesis of a covalently linked, branched DNA-RNA compound in *E. coli* B. *Cell* **56:**891–904.

122. **Lin, R.-J., M. Capage, and C. W. Hill.** 1984. A repetitive DNA sequence, *rhs,* responsible for duplications within the *Escherichia coli* K-12 chromosome. *J. Mol. Biol.* **177:**1–18.

123. **Lindsey, D. F., D. A. Mullin, and J. R. Walker.** 1989. Characterization of the cryptic lambdoid prophage DLP12 of *Escherichia coli* and overlap of the DLP12 integrase gene with the tRNA gene *argU. J. Bacteriol.* **171:**6197–6205.

124. **Liu, L. F., and J. C. Wang.** 1987. Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci. USA* **84:**7024–7027.

125. **Louarn, J., J. Patte, and J.-M. Louarn.** 1982. Suppression of *Escherichia coli dnaA46* mutations by integration of plasmid R100.1 derivatives: constraints imposed by the replication terminus. *J. Bacteriol.* **151:**657–667.

126. **Louarn, J. M., J. P. Bouche, F. Lengendre, J. Louarn, and J. Patte.** 1985. Characterization and properties of very large inversions of the *E. coli* chromosome along the origin-to-terminus axis. *Mol. Gen. Genet.* **201:**467–476.

127. **Lovett, S. T., P. T. Drapkin, V. A. Sutera, Jr., and T. J. Gluckman-Peskind.** 1993. A sister-strand exchange mechanism for *recA*-independent deletion of repeated DNA sequences in *Escherichia coli. Genetics* **135:**631–642.

128. **MacFarlane, S. A., and M. Merrick.** 1985. The nucleotide sequence of the nitrogen regulation *ntrB* and the *glnA-ntrBC* intergenic region of *Klebsiella pneumoniae. Nucleic Acids Res.* **13:**7591–7606.

129. **Marvo, S. L., S. R. King, and S. R. Jaskunas.** 1983. Role of short regions of homology in intermolecular illegitimate recombination events. *Proc. Natl. Acad. Sci. USA* **80:**2452–2456.

130. **Matfield, M., R. Badawi, and W. J. Brammar.** 1985. Rec-dependent and Rec-independent recombination of plasmid-borne duplications in *Escherichia coli* K12. *Mol. Gen. Genet.* **199:**518–523.

131. **Matic, I., M. Radman, and C. Rayssiguier.** 1994. Structure of recombinants from conjugational crosses between *Escherichia coli* donor and mismatch-repair deficient *Salmonella typhimurium* recipients. *Genetics* **136:**17–26.

132. **Mazin, A. V., A. V. Kuzminov, G. L. Dianov, and R. I. Salganik.** 1991. Molecular mechanisms of deletion formation in *Escherichia coli* plasmids. II. Deletion formation mediated by short direct repeats. *Mol. Gen. Genet.* **228:**209–214.

133. **Meighen, E. A., and R. B. Szittner.** 1992. Multiple repetitive elements and organization of the *lux* operons of luminescent terrestrial bacteria. *J. Bacteriol.* **174:**5371–5381s.

134. **Milkman, R., and A. Stoltzfus.** 1988. Molecular evolution of the *Escherichia coli* chromosome. II. Clonal segments. *Genetics* **120:**359–366.

135. **Mizobuchi, K.** 1993. Personal communication. First International Symposium Mapping and Sequencing of Small Genomes, Institut Pasteur, Paris.

136. **Moralejo, P., S. M. Egan, E. Hidalgo, and J. Aguilar.** 1993. Sequencing and characterization of a gene cluster encoding the enzymes for L-rhamnose metabolism in *Escherichia coli. J. Bacteriol.* **175:**5585–5594.

137. **Morgan, G. T., and K. M. Middleton.** 1990. Short interspersed repeats from *Xenopus* that contain multiple octamer motifs are related to known transposable elements. *Nucleic Acids Res.* **18:**5781–5785.

138. **Morse, M., E. Lederberg, and J. Lederberg.** 1956. Transduction in *Escherichia coli* K-12. *Genetics* **41:**121.

139. **Mott, J. E., J. C. Galloway, and T. Platt.** 1985. Maturation of *Escherichia coli* tryptophan operon mRNA: evidence for 3′ exonucleolytic processing after rho-independent termination. *EMBO J.* **4:**1887–1891.

140. **Naas, T., M. Blot, W. M. Fitch, and W. Arber.** 1994. Insertion sequence-related genetic variation in resting *Escherichia coli* K-12. *Genetics* **136:**721–730.

141. **Nakata, A., M. Amemura, and K. Makino.** 1989. Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome. *J. Bacteriol.* **171:**3553–3556.

142. **Newbury, S. F., N. H. Smith, and C. F. Higgins.** 1987. Differential mRNA stability controls relative gene expression within a polycistronic operon. *Cell* **51:**1131–1143.

143. **Nichols, B. P., and C. Yanofsky.** 1979. Nucleotide sequences of *trpA* of *Salmonella typhimurium* and *Escherichia coli:* an evolutionary comparison. *Proc. Natl. Acad. Sci. USA* **76:**5244–5248.

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

145. **Ohno, S.** 1970. *Evolution by Gene Duplication.* Springer-Verlag, Berlin.

146. **Ohshima, A., S. Inouye, and M. Inouye.** 1992. *In vivo* duplication of genetic elements by the formation of stem-loop DNA without an RNA intermediate. *Proc. Natl. Acad. Sci. USA* **89:**1016–1020.

147. **Oppenheim, A. B., K. E. Rudd, I. Mendelson, and D. Teff.** 1993. Integration host factor binds to a unique class of complex repetitive extragenic DNA sequences in *Escherichia coli. Mol. Microbiol.* **10:**113–122.

148. **Osheroff, N.** 1986. Eukaryotic topoisomerase II. *J. Biol. Chem.* **261:**9944–9950.

149. **Petes, T. D., and C. W. Hill.** 1988. Recombination between repeated sequences in microorganisms. *Annu. Rev. Genet.* **22:**147–168.

150. **Petit, M.-A., J. Dimpfl, M. Radman, and H. Echols.** 1991. Control of large chromosomal duplications in *Escherichia coli* by the mismatch repair system. *Genetics* **129:**327–332.

151. **Pierson, L. S., III, and M. L. Kahn.** 1987. Integration of satellite bacteriophage P4 in *Escherichia coli:* DNA sequences of the phage and host regions. *J. Mol. Biol.* **196:**487–496.

152. **Plamann, M. D., and G. V. Stauffer.** 1985. Characterization of a *cis*-acting regulatory mutation that maps at the distal end of the *Escherichia coli glyA* gene. *J. Bacteriol.* **161:**650–654.

153. **Platt, T.** 1986. Transcription termination and regulation of gene expression. *Annu. Rev. Biochem.* **55:**339–372.

154. **Radman, M.** 1989. Mismatch repair and the fidelity of genetic recombination. *Genome* **31:**68–73.

155. **Radman, M.** 1991. Avoidance of inter-repeat recombination by sequence divergence and a mechanism of neutral evolution. *Biochimie* **73:**357–361.

156. **Rau, D. C., M. Gellert, F. Thoma, and A. Maxwell.** 1987. Structure of the DNA gyrase-DNA complex as revealed by transient electric dichroism. *J. Mol. Biol.* **193:**555–569.

157. **Rayssiguier, C., D. S. Thaler, and M. Radman.** 1989. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature* (London) **342:**396–401.

158. **Rebollo, J.-E., V. Francois, and J.-M. Louarn.** 1988. Detection and possible role of two large nondivisible zones on the *Escherichia coli* chromosome. *Proc. Natl. Acad. Sci. USA* **85:**9391–9395.

159. **Riley, M.** 1984. Arrangement and rearrangement of bacterial genomes, p. 285–315. *In* R. P. Morlock (ed.), *Microorganisms as Model Systems for Studying Evolution.* Plenum, New York.

160. **Riley, M., and A. Anilionis.** 1978. Evolution of the bacterial genome. *Annu. Rev. Microbiol.* **32:**519–560.

161. **Riley, M., and S. Krawiec.** 1987. Genome organization, p. 967–981. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology,* vol. 2. American Society for Microbiology, Washington, D.C.

162. **Roberge, M., and S. M. Gasser.** 1992. DNA loops: structural and functional properties of scaffold-attached regions. *Mol. Microbiol.* **6:**419–423.

163. **Rosenberg, M., and D. Court.** 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. *Annu. Rev. Genet.* **13:**319–353.

164. **Russell, R. L., J. N. Abelson, A. Landy, M. L. Gefter, S. Brenner, and J. D. Smith.** 1970. Duplicate genes for tyrosine transfer RNA in *Escherichia coli. J. Mol. Biol.* **47:**1–13.

165. **Sadosky, A. B., A. Davidson, R.-J. Lin, and C. W. Hill.** 1989. *rhs* gene family of *Escherichia coli* K-12. *J. Bacteriol.* **171:**636–642.

166. **Sadosky, A. B., J. A. Gray, and C. W. Hill.** 1991. The *RhsD-E* subfamily of *Escherichia coli* K-12. *Nucleic Acids Res.* **19:**7177–7183.

167. **Saksela, K., and D. Baltimore.** 1993. Negative regulation of immunoglobulin kappa light-chain gene transcription by a short sequence homologous to the murine B1 repetitive element. *Mol. Cell. Biol.* **13:**3698–3705.

168. **Sanderson, K. E.** 1976. Genetic relatedness in the family Enterobacteriaceae. *Annu. Rev. Microbiol.* **30:**327–349.

169. **Savic, D. J., S. P. Romac, and S. D. Ehrlich.** 1983. Inversion in the lactose region of *Escherichia coli* K-12: inversion termini map within IS3 elements $\alpha_{3}\beta_{3}$ and $\beta_{5}\alpha_{5}$. *J. Bacteriol.* **155:**943–946.

170. **Sawyer, S. A., D. E. Dykhuizen, R. F. Dubose, L. Green, T. Mutangadura-Mhlanga, D. F. Wolczyk, and D. L. Hartl.** 1987. Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli. Genetics* **115:**51–63.

171. **Schmeissner, U., K. MacKenney, M. Rosenberg, and D. Court.** 1984. Removal of a terminator structure by mRNA processing regulates *int* gene expression. *J. Mol. Biol.* **176:**39–53.

172. **Schmid, M. B., and J. R. Roth.** 1983. Selection and endpoint distribution of bacterial inversion mutations. *Genetics* **105:**539–557.

173. **Schmid, M. B., and J. R. Roth.** 1987. Gene location affects expression level in *Salmonella typhimurium. J. Bacteriol.* **169:**2872–2875.

174. **Sclafani, R. A., and J. A. Wechsler.** 1981. High frequency of genetic duplications in the *dnaB* region of the *Escherichia coli* K12 chromosome. *Genetics* **98:**677–689.

175. **Segall, A. M., M. J. Mahan, and J. R. Roth.** 1988. Rearrangement of the bacterial chromosome: forbidden inversions. *Science* **241:**1314–1318.

176. **Segall, A. M., and J. R. Roth.** 1989. Recombination between homologies in direct and inverse orientation in the chromosome of Salmonella: intervals which are nonpermissive for inversion formation. *Genetics* **122:**737–747.

177. **Segall, A. M., and J. R. Roth.** 1994. Approaches to half-tetrad analysis in bacteria: recombination between repeated, inverse-order chromosomal sequences. *Genetics* **136:**27–39.

178. **Sharp, P. M., and W.-H. Li.** 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15:**1281–1295.

179. **Sharples, G. J., and R. G. Lloyd.** 1990. A novel repeated DNA sequence located in the intergenic regions of bacterial chromosome. *Nucleic Acids Res.* **18:**6503–6508.

180. **Shen, P., and H. V. Huang.** 1986. Homologous recombination in *Escherichia coli:* dependence on substrate length and homology. *Genetics* **112:**441–457.

181. **Shen, P., and H. V. Huang.** 1989. Effect of base pair mismatches on recombination via the RecBCD pathway. *Mol. Gen. Genet.* **218:**358–360.

182. **Shyamala, V., E. Schneider, and G. Ferro-Luzzi Ames.** 1990. Tandem chromosomal duplications: role of REP sequences in the recombination event at the join-point. *EMBO J.* **9:**939–946.

183. **Smith, G. P.** 1973. Unequal crossover and the evolution of multigene families. *Cold Spring Harbor Symp. Quant. Biol.* **38:**507–513.

184. **Smith, G. R.** 1988. Homologous recombination sites and their recognition, p. 115–154. *In* K. B. Low (ed.), *The Recombination of Genetic Material.* Academic Press, San Diego, Calif.

185. **Sommer, H., B. Schumacher, and H. Saedler.** 1981. A new type of IS*1*-mediated deletion. *Mol. Gen. Genet.* **184:**300–307.

186. **Sonti, R. V., and J. R. Roth.** 1989. Role of gene duplications in the adaptation of *Salmonella typhimurium* to growth on limiting carbon sources. *Genetics* **123:**19–28.

187. **Stern, M. J., G. Ferro-Luzzi Ames, N. H. Smith, E. C. Robinson, and C. F. Higgins.** 1984. Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell* **37:**1015–1026.

188. **Stern, M. J., E. Prossnitz, and G. Ferro-Luzzi Ames.** 1988. Role of the intercistronic region in post-transcriptional control of gene expression in the histidine transport operon of *Salmonella typhimurium:* involvement of REP sequences. *Mol. Microbiol.* **2:**141–152.

189. **Straus, D. S.** 1974. Induction by mutagens of tandem gene duplications in the *glyS* region of the *Escherichia coli* chromosome. *Genetics* **78:**823–830.

190. **Straus, D. S., and G. R. Hoffmann.** 1975. Selection for a large genetic duplication in *Salmonella*

typhimurium. Genetics **80:**227–237.

191. **Stroeher, U. H., L. E. Karageorgos, R. Morona, and P. A. Manning.** 1992. Serotype conversion in *Vibrio cholerae* O1. *Proc. Natl. Acad. Sci. USA* **89:**2566–2570.

192. **Timmons, M. S., A. M. Bogardus, and R. C. Deonier.** 1983. Mapping of chromosomal IS*5* elements that mediate type II F-prime plasmid excision in *Escherichia coli* K-12. *J. Bacteriol.* **153:**395–407.

193. **Timmons, M. S., M. Lieb, and R. C. Deonier.** 1986. Recombination between IS*5* elements: requirement for homology and recombination functions. *Genetics* **113:**797–810.

194. **Tlsty, T. D., A. M. Albertini, and J. H. Miller.** 1984. Gene amplification in the *lac* region of *E. coli. Cell* **37:**217–224.

195. **Tomb, J.-F., H. El-Hajj, and H. O. Smith.** 1991. Nucleotide sequence of a cluster of genes involved in the transformation of *Haemophilus influenzae* Rd. *Gene* **104:**1–10.

196. **Truniger, V., W. Boos, and G. Sweet.** 1992. Molecular analysis of the *glpFKX* regions of *Escherichia coli* and *Shigella flexneri. J. Bacteriol.* **174:**6981–6991.

197. **Turnowsky, F., C. Fuchs, C. Jeschek, and G. Högenauer.** 1989. *envM* genes of *Salmonella typhimurium* and *Escherichia coli. J. Bacteriol.* **171:**6555–6565.

198. **Umeda, M., and E. Ohtsubo.** 1989. Mapping of insertion elements IS*1,* IS*2* and IS*3* on the *Escherichia coli* K-12 chromosome. *J. Mol. Biol.* **208:**601–614.

199. **Umeda, M., and E. Ohtsubo.** 1990. Mapping of insertion element IS*5* in the *Escherichia coli* K-12 chromosome. Chromosomal rearrangements mediated by IS*5. J. Mol. Biol.* **213:**229–237.

200. **Valentin-Hansen, P., K. Hammer-Jespersen, F. Boetius, and I. Svendsen.** 1984. Structure and function of the intercistronic regulatory *deoC-deoA* element of *Escherichia coli* K-12. *EMBO J.* **3:**179–183.

201. **Van Vliet, F., R. Cunin, A. Jacobs, J. Piette, D. Gigot, M. Lauwereys, A. Pierard, and N. Glansdorff.** 1984. Evolutionary divergence of genes for ornithine and aspartate carbamoyl-transferases—complete sequence and mode of regulation of the *Escherichia coli argF* gene; comparison of *argF* with *argI* and *pyrB. Nucleic Acids Res.* **12:**6277–6289.

202. **Varmus, H. E.** 1989. Reverse transcription in bacteria. *Cell* **56:**721–724.

203. **Versalovic, J., T. Koeuth, R. Britton, K. Geszvain, and J. R. Lupski.** 1993. Conservation and evolution of the *rpsU-dnaG-rpoD* macromolecular synthesis operon in bacteria. *Mol. Microbiol.* **8:**343–355.

204. **Versalovic, J., T. Koeuth, and J. R. Lupski.** 1991. Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Res.* **19:**6423–6831.

205. **Vidal, F., E. Mougneau, N. Glaichenhaus, P. Vaigot, M. Darmon, and F. Cuzin.** 1993. Coordinated posttranscriptional control of gene expression by modular elements including *Alu*-like repetitive sequences. *Proc. Natl. Acad. Sci. USA* **90:**208–212.

206. **Voloshin, O. N., S. M. Mirkin, V. I. Lyamichev, B. P. Belotserkovskii, and M. D. Frank-Kamenetski.** 1988. Chemical probing of homopurine-homopyrimidine mirror repeats in supercoiled DNA. *Nature* (London) **333:**475–476.

207. **von Eichel-Streiber, C., M. Sauerborn, and H. K. Kuramitsu.** 1992. Evidence for a modular structure of the homologous repetitive C-terminal carbohydrate-binding sites of *Clostridium difficile* toxins and *Streptococcus mutans* glucosyltransferases. *J. Bacteriol.* **174:**6707–6710.

208. **Watt, V. M., C. J. Ingles, M. S. Urdea, and W. J. Rutter.** 1985. Homology requirements for recombination in *Escherichia coli. Proc. Natl. Acad. Sci. USA* **82:**4768–4772.

209. **White, S., F. E. Tuttle, D. Blankenhorn, D. C. Dosch, and J. L. Slonczewski.** 1992. pH dependence and gene structure of *inaA* in *Escherichia coli. J. Bacteriol.* **174:**1537–1543.

210. **Whoriskey, S. K., V.-H. Nghiem, P.-M. Leong, J.-M. Masson, and J. H. Miller.** 1987. Genetic rearrangements and gene amplification in *Escherichia coli:* DNA sequences at the junctures of amplified gene fusions. *Genes Dev.* **1:**227–237.

211. **Wren, B. W.** 1991. A family of clostridial and streptococcal ligand-binding proteins with conserved C-terminal repeat sequences. *Mol. Microbiol.* **5:**797–803.

212. **Xiang, S.-H., M. Hobbs, and P. R. Reeves.** 1994. Molecular analysis of the *rfb* gene cluster of a

group D2 *Salmonella enterica* strain: evidence for intraspecific gene transfer in O antigen variation. *J. Bacteriol.* **176:**4357–4365.

213. **Yang, Y., and G. Ferro-Luzzi Ames.** 1988. DNA gyrase binds to the family of prokaryotic repetitive extragenic palindromic sequences. *Proc. Natl. Acad. Sci. USA* **85:**8850–8854.

214. **Yang, Y., and G. Ferro-Luzzi Ames.** 1990. The family of repetitive extragenic palindromic sequences: interaction with DNA gyrase and histonelike protein HU, p. 211–226. *In* K. Drlica and M. Riley (ed.), *The Bacterial Chromosome.* American Society for Microbiology, Washington, D.C.

215. **Yokota, T., H. Sugisaki, M. Takanami, and Y. Kaziro.** 1980. The nucleotide sequence of the cloned *tufA* gene of *Escherichia coli. Gene* **12:**25–31.

216. **York, M. K., and M. Stodolsky.** 1981. Characterization of P1*argF* derivatives from *Escherichia coli* K12 transduction I. IS*1* elements flank the *argF* gene segment. *Mol. Gen. Genet.* **181:**230–240.

217. **Zhao, S., C. H. Sandt, G. Feulner, D. A. Vlazny, J. A. Gray, and C. W. Hill.** 1993. *Rhs* elements of *Escherichia coli* K-12: complex composites of shared and unique components that have different evolutionary histories. *J. Bacteriol.* **175:**2799–2808.

218. **Zieg, J., and S. R. Kushner.** 1977. Analysis of genetic recombination between two partially deleted lactose operons of *Escherichia coli* K-12. *J. Bacteriol.* **131:**123–132.

219. **Ziemke, P., and J. E. G. McCarthy.** 1992. The control of mRNA stability in *Escherichia coli:* manipulation of the degradation pathway of the polycistronic *atp* mRNA. *Biochim. Biophys. Acta* **1130:**297–306.

220. **Zuker, M., and P. Stiegler.** 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9:**133–148.