Periodic Extinctions of Transposable Elements in Bacterial Lineages: Evidence from Intragenomic Variation in Multiple Genomes

Andreas Wagner

Department of Biology, The University of New Mexico

Most previous work on the evolution of mobile DNA was limited by incomplete sequence information. Whole genome sequences allow us to overcome this limitation. I study the nucleotide diversity of prominent members of five insertion sequence families whose transposition activity is encoded by a single transposase gene. Eighteen among 376 completely sequenced bacterial genomes and plasmids carry between 3 and 20 copies of a given insertion sequence. I show that these copies generally show very low DNA divergence. Specifically, more than 68% of the transposase genes are identical within a genome. The average number of amino acid replacement substitutions at amino acid replacement sites is $K_a = 0.013$, that at silent sites is $K_s = 0.1$. This low intragenomic diversity stands in stark contrast to a much higher divergence of the same insertion sequences among distantly related genomes. Gene conversion among protein-coding genes is unlikely to account for this lack of diversity. The relation between transposition frequencies and silent substitution rates suggests that most insertion sequences in a typical genome are evolutionarily young and have been recently acquired. They may undergo periodic extinction in bacterial lineages. By implication, they are detrimental to their host in the long run. This is also suggested by the highly skewed and patchy distribution of insertion sequences among genomes. In sum, one can think of insertion sequences as slow-acting infectious diseases of cell lineages.

Introduction

Why is mobile DNA maintained in a genome? On one hand, mobile DNA may be a very effective parasite, replicating itself at the expense of its host. (Doolittle and Sapienza 1980; Orgel and Crick 1980). On the other hand, mobile DNA can also have beneficial effects. Most importantly, it can serve to mobilize genes for transfer between bacterial strains or species (Bushman 2002). In addition, the presence of insertion sequences or their transposition may be beneficial to one host cell, for reasons that are not always clear. Such benefits may often be temporary, but even so, they could lead to the sustenance of insertion sequences in bacterial populations (Hartl et al. 1983; Blot 1994; Treves, Manning, and Adams 1998; Cooper et al. 2001; Edwards and Brookfield 2003; Schneider and Lenski 2004).

To find out whether mobile DNA persists because it benefits a host, one needs to understand the dynamics of mobile DNA on evolutionary (not laboratory) timescales. However, compared to a large body of work on the molecular biology of mobile DNA (Berg and Howe 1989; Craig et al. 2002), only a modest number of evolutionary studies have been carried out. This holds in particular for the focus of this contribution, prokaryotic insertion sequences, which are arguably the simplest kinds of prokaryotic mobile DNA. These transposable elements range in size from 700 to 2,000 bp (Mahillon and Chandler 1998). Typically, they consist of a short inverted repeat sequence that flanks one or more open reading frames, whose products encode transposase proteins necessary for transposition.

Evolutionary studies on prokaryotic insertion sequences fall into two broad classes. The first focuses on the mutational events caused by insertion sequences in evolving (laboratory) populations (Naas et al. 1994; Treves, Manning, and Adams 1998; Schneider et al. 2000; Cooper et al. 2001; Schneider and Lenski 2004). The second focuses on the number and distribution of insertion sequences in bacterial populations or closely related bacterial strains (Sawyer et al. 1987; Ajioka and Hartl 1989; Hall et al. 1989; Lawrence, Ochman, and Hartl 1992; Naas et al. 1994; Bisercic and Ochman 1995). Much of this work dates to the pregenome era and is based on the detection of variation in copy number and sequence through DNA hybridization and restriction fragment length polymorphisms. Despite the limitations of such data, the microevolutionary resolution of these older studies has not yet been surpassed by bacterial genome sequencing projects. The reason is that most completely sequenced bacterial genomes show much wider phylogenetic spacing than the strains used in these studies.

Among pertinent earlier work, two papers are of particular interest. The first characterized a sample of insertion sequences in *Escherichia coli* and related enteric bacteria (Lawrence, Ochman, and Hartl 1992). It indicated low nucleotide diversity for the sampled insertion sequences within a genome, a high turnover of insertion sequences within a genome, and ready horizontal transfer of transposable elements among *E. coli* lineages. The second study (Sawyer et al. 1987) analyzed the copy number of six different insertion sequences in 71 natural isolates of *E. coli*. It showed that the distribution of copy numbers is skewed. Specifically, the majority of isolates have no or few copies of a given insertion sequence, and few isolates have a large number of insertion sequences.

The data I present here, taken together with information from earlier work, suggest that insertion sequences may show extinction-reinfection cycles. If so, insertion sequences have deleterious effects on their host, at least in the long run, and horizontal gene transfer is crucial to sustain insertion sequences in a metapopulation of clonal lineages. My analysis takes advantage of more than 200 completely sequenced bacterial genomes (376 DNA molecules including sequenced plasmids). It focuses on those 18 genomes where

Key words: mobile DNA, transposon, infection, extinction.

E-mail: wagnera@unm.edu.

Mol. Biol. Evol. 23(4):723–733. 2006
doi:10.1093/molbev/msj085
Advance Access publication December 22, 2005

five major insertion sequences occur at between 3 copies—minimally necessary to detect gene conversion—and 20 copies per genome. The respective organisms cover wide phylogenetic distances and a broad spectrum of lifestyles. They include a member of the genus *Azoarcus*, a denitrifying bacterium that can degrade aromatic compounds and whose relatives are often plant associated, the deep-sea bacterium *Idiomarina loihiensis,* which occurs in the vicinity of hydrothermal vents, *Legionella pneumophila,* the bacterial agent of Legionnaire's disease, the highly virulent pathogen *Coxiella burnetii*, and the free-living marine planctomycete *Rhodopirellula baltica* (Welch et al. 2002; Glockner et al. 2003; Seshadri et al. 2003; Chien et al. 2004; Rabus et al. 2005).

Methods

I obtained all 376 GenBank files that were available for bacterial genomes from the National Center for Biotechnology Information in February 2005 (ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Bacteria/; RefSeq ID NC_000117–NC_007530). These files comprise the DNA sequence and annotation for 202 completely sequenced bacterial genomes as well as multiple extrachromosomal DNA molecules. IS4, IS5, IS6, IS30, and IS605/IS200 are the most prominent members of families of bacterial insertion sequences whose transposition activity is encoded by a single open reading frame. I identified genes annotated with these names in all of the above chromosomal and extrachromosomal sequences. IS605/IS200 often occur jointly, and it is not clear in this case whether the individual insertion sequences or the composite form the active insertion element. I thus eliminated such composites from further analysis. Of the remaining insertion sequences, I retained those that had between 3 and 20 copies per genome for further analysis. They are contained in 18 different genomes (table 1), whose RefSeq identification numbers (www.ncbi.nlm.nih.gov/RefSeq) are as follows: *Azoarcus* sp. EbN1 (NC_006513), *C. burnetii* RSA 493 (NC_002971), *E. coli* CFT073 (NC_004431), *E. coli K12* (NC_000913), *Francisella tularensis* subsp. *tularensis Schu 4* (NC_006570), *Idiomarina loihiensis* L2TR (NC_006512), *Lactococcus lactis* subsp. *lactis Il1403* (NC_002662), *L. pneumophila pneumophila str. Philadelphia 1* (NC_002942), *L. pneumophila str. Pris* (NC_006368), *Neisseria meningitidis* MC58(NC_003112), *Photorhabdus laumondii* subsp. *Laumondii* TT01 (NC_005126), *R. baltica SH 1* (NC_005027), *Shigella flexneri 2a str. 301* (NC_004337), *Streptococcus thermophilus CNRZ1066* (NC_006449), *S. thermophilus* LMG 18311 (NC_006448), *Streptomyces avermitilis* MA-4680 (NC_003155), *Vibrio vulnificus* CMCP6 chromosome I (NC_004459), *Wolbachia* endosymbiont of *Drosophila melanogaster* (NC_002978), and *Xanthomonas oryzae pv. oryzae* KACC10331 (NC_006834).

Using annotation-based information has two caveats. First, annotation algorithms may mistakenly lump functional but highly divergent insertion sequences into the same family. However, if such misannotation was pervasive in the data I analyze, the divergence estimates of insertion sequences would be much higher than they are. Conversely, annotation may miss divergent yet functional members of an insertion sequence family. To assess whether this might be a problem, I used TBlastX (Altschul et al. 1990) to screen all genomes for sequences similar to one of the members of this insertion sequence families that may have been missed in the annotation. This search reveals that the annotation missed fewer than 10% of insertion sequences and that most of these insertion sequences are highly similar (>99% nucleotide sequence identity) to the other sequences.

For each insertion sequence that fulfilled the above requirements, I aligned the coding DNA sequence using ClustalW (Thompson, Higgins, and Gibson 1994) and eliminated, based on this alignment, sequences that are most likely truncated fragments which sometimes arise as byproducts of the transposition processes. Such sequences have a length of less than 50% of the "minimal" length reported for the respective insertion sequence, as given in Mahillon and Chandler (1998), are highly divergent, and usually contain multiple insertions and deletions. This procedure does not truncate the distribution of sequence divergence, as Figure S1 (Supplementary Material online) shows.

For the remaining sequences, I determined the nucleotide identity of each sequence pair from the multiple alignment. I also recorded the distance of nearest neighbors in terms of the number of base pairs and the fractional genome length between the start of each sequence and determined the number of genes (coding regions) between each nearest neighbor pair of insertion sequences.

To estimate synonymous and nonsynonymous divergence, I used a previously published tool (Conant and Wagner 2002). Briefly, the tool uses information from both the DNA and amino acid sequences of coding regions of interest and proceeds in three steps. First, it identifies related genes by a prescreening step that uses BlastP (*E* value $<$ 0.01; Altschul et al. 1997). Secondly, it aligns the resulting subset of genes globally with the Needleman and Wunsch dynamic programing alignment algorithm (Thompson, Higgins, and Gibson 1994). Before entering the third step, gene pairs with fewer than 40 alignable amino acid residues and with less than 50% percent amino acid identity are eliminated. In the third step, the tool calculates the number of substitutions per synonymous site (K_s) and number of substitutions per nonsynonymous site (K_a) using the maximum likelihood models of Muse and Gaut (1994) and Goldman and Yang (1994) for the remaining gene pairs. It uses a simple heuristic test (Conant and Wagner 2003) to determine whether a gene pair has been saturated with synonymous substitutions.

To test for gene conversion, I used the program GENECONV (available at http://www.math.wustl.edu/~sawyer/geneconv/gconvdoc.html), with default parameters to carry out Sawyer's test (Sawyer 1989), as well as DRUID (http://bioweb.pasteur.fr/seqanal/interfaces/druid.html), with a window size of 100 nt to carry out Farris's incongruence length difference (ILD) test (Farris et al. 1995).

Results

Many Insertion Sequences Have Identical or Very Closely Related Coding Regions

I focus here on insertion sequences whose transposition activity is encoded by a single transposase open reading

Table 1
Low Synonymous and Nonsynonymous Divergence of Transposase Genes

Species[a]	Insertion Sequence	Number of Copies (K_s = 0)[b]	K_s (s)[c]	K_a (s)	K_s (s)	K_a (s)	K_a/K_s (s)
			All Copies		Copies with $K_s > 0$		
Azoarcus sp. *EbN1*	IS5	5 (5)	0 (0)	0 (0)	NA	NA	NA
Coxiella burnetii RSA 493	IS30	5 (5)	0 (0)	0 (0)	NA	NA	NA
Escherichia coli CFT073	IS200	18 (8 + 10)	0.038 (0.003)	0.003 ($<10^{-3}$)	0.073 (0)	0.005 (0)	0.07 (0.001)
Escherichia coli K12	IS30	3 (1)	0.004 (0.002)	0.001 (0)	0.006 (0)	0.001 (0)	0.187 (0)
Escherichia coli K12	IS5	11 (9)	0.044 (0.012)	0.002 (0.001)	0.128 (0.027)	0.007 (0.001)	0.23 (0.046)
Francisella tularensis[d]	IS5	16 (15)	0.008 (0.002)	0.002 (0)	0.034 (0.003)	0.008 (0.001)	0.149 (0.029)
Idiomarina loihiensis L2TR	IS4	5 (5)	0 (NA)	0 (NA)	NA	NA	NA
Lactococcus lactis subsp. *lactis Il1403*	IS30	15 (15)	0 (NA)	0.001 (0.001)	NA	NA	NA
Legionella pneumophila str. Philadelphia 1	IS4	4 (3)	0.122 (0.054)	0.018 (0.008)	0.243 (NA)	0.036 (NA)	0.149 (NA)
Legionella pneumophila str. Pris	IS4	9 (5)	0.003 (0.001)	0.003 (0.001)	0.005 (0.001)	0.005 (0.001)	0.833 (0.191)
Neisseria meningitidis MC58	IS30	14 (4 + 2 + 2)	0.008 (0.001)	0.004 ($<10^{-3}$)	0.008 (0.005)	0.004 ($<10^{-3}$)	0.486 (0.035)
Photorhabdus laumondii	IS30	6 (4)	0.02 (0.006)	0.002 (0.001)	0.033 (0.008)	0.007 ($<10^{-3}$)	0.075 (0.024)
Rhodopirellula baltica SH 1	IS4	10 (4 + 2)	0.029 (0.011)	0.015 (0.006)	0.062 (0.014)	0.033 (0.02)	0.732 (0.277)
Shigella flexneri 2a str. 301	IS4	16 (2 + 2 + 2)	0.003 ($<10^{-3}$)	0.004 ($<10^{-3}$)	0.004 ($<10^{-3}$)	0.005 ($<10^{-3}$)	1.325 (0.08)
Streptococcus thermophilus CNRZ1066	IS30	10 (9)	0.001 ($<10^{-3}$)	0 ($<10^{-3}$)	0.004 ($<10^{-3}$)	NA ($<10^{-3}$)	NA
Streptococcus thermophilus LMG 18311	IS30	8 (7)	0.003 (0.001)	0.001 ($<10^{-3}$)	0.01 ($<10^{-3}$)	0.003 ($<10^{-3}$)	0.268 ($<10^{-3}$)
Streptomyces avermitilis MA-4680	IS4	7 (2 + 3)	1.3 (0.565)	0.156 (0.061)	2.34 (0.735)	0.234 (0.072)	0.133 (0.01)
Streptomyces avermitilis MA-4680	IS6	3 (3)	0 (0)	0 (0)	NA	NA	NA
Vibrio vulnificus CMCP6 chromosome I	IS30	7 (4)	0.064 (0.016)	0.024 (0.008)	0.09 (0.019)	0.046 (0.012)	0.243 (0.069)
Wolbachia (endosymbiont of *Drosophila melanogaster*)	IS4	3 (3)	0 (0)	0 (0)	NA	NA	NA
Xanthomonas oryzae	IS30	11 (8)	0.481 (0.131)	0.035 (0.006)	0.979 (0.232)	0.043 (0.007)	0.534 (0.08)

NOTE.—"NA" indicates "not applicable" due to insufficient data.

[a] RefSeq accession numbers of GenBank files for each genome can be found in Methods. All chromosomes analyzed are circular except that of *S. avermitilis*.

[b] Numbers in parentheses indicate the number of copies identical at synonymous sites. Except for IS30 of *L. lactis*, which has one nonsynonymous but no silent substitutions, these are also the number of copies with complete nucleotide identity.

[c] In this and all columns to the right of it, "s" indicates the standard error (standard deviation divided by the square root of the number of observations) and is given in parentheses.

[d] *F. tularensis* subsp. *tularensis Schu 4*.

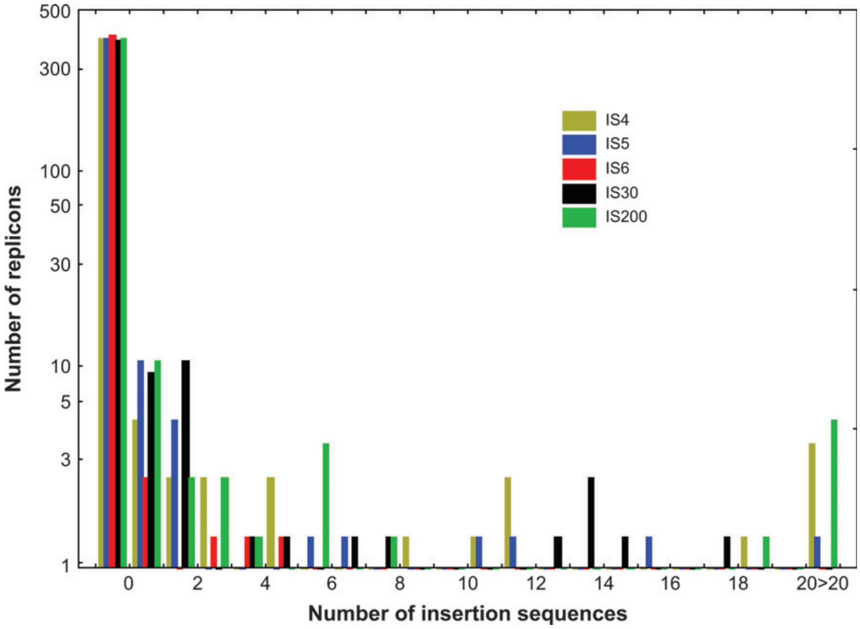

FIG. 1.—Sparse distribution of insertion sequences among genomes. The figure shows the number of genomes and plasmids (vertical axis, among 376 total) that carry a given number of insertion sequences (horizontal axis). Note the logarithmic scale of the vertical axis. For all five insertion sequences studied here, the distribution is dominated by genomes that carry zero copies. Very few genomes carry the minimal number of three insertion sequence copies necessary to detect gene conversion.

frame. I analyzed the main representative of each major family of such insertion elements (Mahillon and Chandler 1998), which are IS4, IS5, IS6, IS30, and IS200/IS605. I examined more than 376 completely sequenced bacterial chromosomes and plasmids (including more than 200 completely sequenced genomes) for the presence of these insertion sequences. Their distribution is very patchy: It is dominated by genomes that are devoid of insertion sequences (fig. 1). Overall, I identified a total of 21 incidences (in 18 different genomes) where one of the above insertion elements had between 3 and 20 copies per genome (table 1).

The high sequence homogeneity of these insertion sequences is already evident from the number of insertion sequences with completely identical transposase-coding regions. Specifically, in six of the 21 genomes from table 1 all transposase sequences show 100% nucleotide identity. On average, more than 68% (standard error s = 6.5%) of the examined insertion sequences within one genome are completely identical.

I next analyzed the nucleotide and amino acid divergence among the transposase-coding regions. Specifically, I estimated amino acid sequence divergence through the fraction K_a of nonsynonymous (amino acid replacement) substitutions per nonsynonymous site on DNA (Li 1997), a divergence measure that accounts for errors in divergence estimates arising through multiple nucleotide substitutions. The overall amino acid divergence among pairs of transposase genes within one genome is a very low $K_a = 0.013$ ($s = 0.007$, $n = 21$). For those 15 insertion sequences where not all members in the genome are identical (table 1), $K_a = 0.029$ ($s = 0.013$; $n = 15$). Silent (synonymous) divergence is also very low. I estimate it as the fraction K_s of synonymous substitutions per synonymous site on DNA (Li 1997). Divergence among synonymous sites is a crude indicator of relative times to common ancestry of repetitive coding sequences, partly because synonymous sites are under fewer evolutionary constraints than nonsynonymous sites. The overall synonymous divergence is again a low $K_s = 0.1$ ($s = 0.064$, $n = 21$). For those 15 insertion sequences where not all members in the genome are identical, $K_s = 0.27$ ($s = 0.14$, $n = 15$).

The ratio K_a/K_s of amino acid replacement to silent changes between two related genes is an indicator of the selective constraint acting on the genes' protein product. Specifically, if $K_a/K_s < 1$, fewer amino acid substitutions than neutral substitutions are preserved in the evolutionary record. This means that some amino acid substitutions have been detrimental to their carrier, which has therefore been eliminated. To estimate K_a/K_s is useful for my purpose because it indicates whether insertion sequences within a genome are functional, i.e., whether they have been under selection for transposition in the recent past or whether they evolve neutrally and may thus be inactive insertion sequences ($K_a = K_s$). K_a/K_s can only be estimated for the 15 cases of table 1 where not all transposases are identical. In one of these 15 cases (IS30 in *L. lactis*), one transposase gene shows a nonsynonymous substitution but none show any synonymous substitutions ($K_s = 0$), meaning that K_a/K_s cannot be calculated. A total of 78.6% (11/14) of the remainder show a ratio K_a/K_s significantly smaller than one, with an average of $K_a/K_s = 0.39$ ($s = 0.08$). I cannot strictly exclude that in a group of insertion sequences with this average K_a/K_s some insertion sequences have been inactivated, but the overall very high sequence similarity makes this unlikely because it indicates active transposition in the recent past. The three exceptions with higher K_a/K_s are IS4 sequences from *L. pneumophila*, *S. flexneri*, and

Pirellula sp.*1*. The divergence of the former two is extremely small ($K_s < 0.01$), such that it is difficult to confidently ascertain selective neutrality. In sum, the vast majority of the insertion sequences analyzed are likely to be under active selection and thus functional.

Are These Transposase Genes Simply Highly Constrained?

Do extremely high evolutionary constraints cause the low divergence of these insertion sequences? Their transposases might tolerate few mutations and therefore change very slowly on an evolutionary timescale. In addition, their mRNA sequence might also be highly constrained, explaining their low synonymous divergence. This is not far fetched, because mRNA secondary structures may play a role in the expression regulation of transposases (Kleckner 1989; Mahillon and Chandler 1998).

The variation of transposase sequences among genomes speaks against this possibility. First, it is worth pointing out that the insertion sequences examined here are merely the most prominent members of larger insertion sequence families. The IS5 family, for example, comprise more than a dozen subfamilies with varying degrees of amino acid similarity of which IS5 is only one element (Mahillon and Chandler 1998). An analysis of synonymous and nonsynonymous insertion sequence divergence among genomes also argues against strong evolutionary constraints. First, many pairs of insertion sequences in different genomes have been completely saturated with synonymous substitutions. Specifically, 14.5% of 2,232 pairs of IS4, 7.9% of 428 pairs of IS5, 16.6% of 24 pairs of IS6, 22.3% of 825 pairs of IS30, and 16.6% of 34,588 pairs of IS200 show saturation at synonymous sites. These saturated pairs were excluded from the analysis of figure 2*a*, which compares pairwise synonymous divergence (K_s) of insertion sequences in all the genomes studied with the intragenomic synonymous divergence from table 1. The figure shows that four out of the five insertion sequence families studied have a significantly higher synonymous divergence among genomes than within genomes. In some cases, this difference is quite striking. For example, the synonymous divergence of IS5 among genomes ($K_s = 0.593$) is more than a factor 30 higher than that within genomes ($K_s = 0.017$). Figure 2*b* shows an exactly analogous analysis but for nonsynonymous divergence K_a. Figure 2*c* and *d*, finally, show data analogous to that of figure 2*a* and *b* but where pairs of identical insertion sequences ($K_s = 0$ or $K_a = 0$) were excluded before the analysis. The results are again qualitatively identical: all but one of the insertion sequences are significantly more and often dramatically more divergent among genomes than within genomes. The one exception is IS4. However, IS4 shows generally an unusual pattern of sequence evolution in the genomes examined. For example, it is the only insertion sequence where the mean ratio K_a/K_s for all transposase pairs is close to one ($K_a/K_s = 0.848$, s = 0.015), suggesting that bacterial genomes contain many inactivated IS4 elements. Taken together, these results show that extreme evolutionary constraints cannot explain the low diversity of insertion sequences within a genome.

Transposase Genes Are Much More Homogeneous than Duplicate Genes Within Genomes

A further indication for the high homogeneity of insertion sequence comes from a larger class of repetitive coding sequences in bacteria: multicopy genes. Many multicopy genes are gene duplicates that may have arisen through gene duplication within a genome, but some may have been duplicated elsewhere and imported into the genome through horizontal transfer. I chose the three insertion sequences that are represented by the most members in my analysis, IS4, IS5, and IS30. I then chose three genomes in which these insertion sequences occur but where they are not all identical in nucleotide sequences (fig. 3). Subsequently, I identified all gene duplicates in these three genomes and determined their synonymous divergence K_s and their nonsynonymous divergence K_a. (I did all of this for only three genomes because this procedure is computationally costly.) Notably, many of the duplicate genes in these genomes are completely saturated with synonymous substitutions. Specifically, 30.7% of duplicate genes in *E. coli*, 24.8% of duplicate genes in *L. pneumophilae*, and 14.9% of duplicate genes in *V. vulnificus* are saturated in K_s. This stands in stark contrast to the insertion sequence transposase genes, none of which are saturated with synonymous substitutions. Secondly, multicopy genes that have not been saturated with synonymous substitutions show significantly greater synonymous divergence than insertion sequence. This difference can be quite dramatic, as for *E. coli*, where gene duplicates show more than 20-fold greater synonymous divergence than IS5 transposases (fig. 3*a*). Qualitatively the same holds for nonsynonymous divergence K_a: it is significantly higher for gene duplicates than for insertion sequences (fig. 3*b*).

No Signs of Concerted Evolution

Concerted evolution is the nonindependent evolution of repetitive DNA sequences. It leads to the homogenization of gene families within a genome and can be caused by reciprocal recombination of homologous sequences or gene conversion. The low diversity of the insertion sequences of table 1 naturally raises the question whether gene conversion is at work.

In eukaryotes, most well-documented cases of concerted evolution occur in closely linked genes. If the same held for prokaryotes, then it might be sufficient to show that insertion sequences are widely dispersed in the genome, which is the case. Insertion sequences do not show a highly clumped distribution in bacterial genomes. For most insertion sequences examined here, pairs of adjacent sequences are separated by at least two (and often many more) gene-coding sequences (table 2).

In contrast to eukaryotes, there is mounting evidence that in prokaryotes concerted evolution is abundant among unlinked sequence. (Santoyo and Romero 2005). Among the best characterized such genes are ribosomal DNA–coding genes and genes encoding the bacterial translation factor EF-Tu (Abdulkarim and Hughes 1996; Liao 2000). The underlying mechanism is probably gene conversion. It is thus necessary to test for concerted evolution in the insertion sequences studied here.

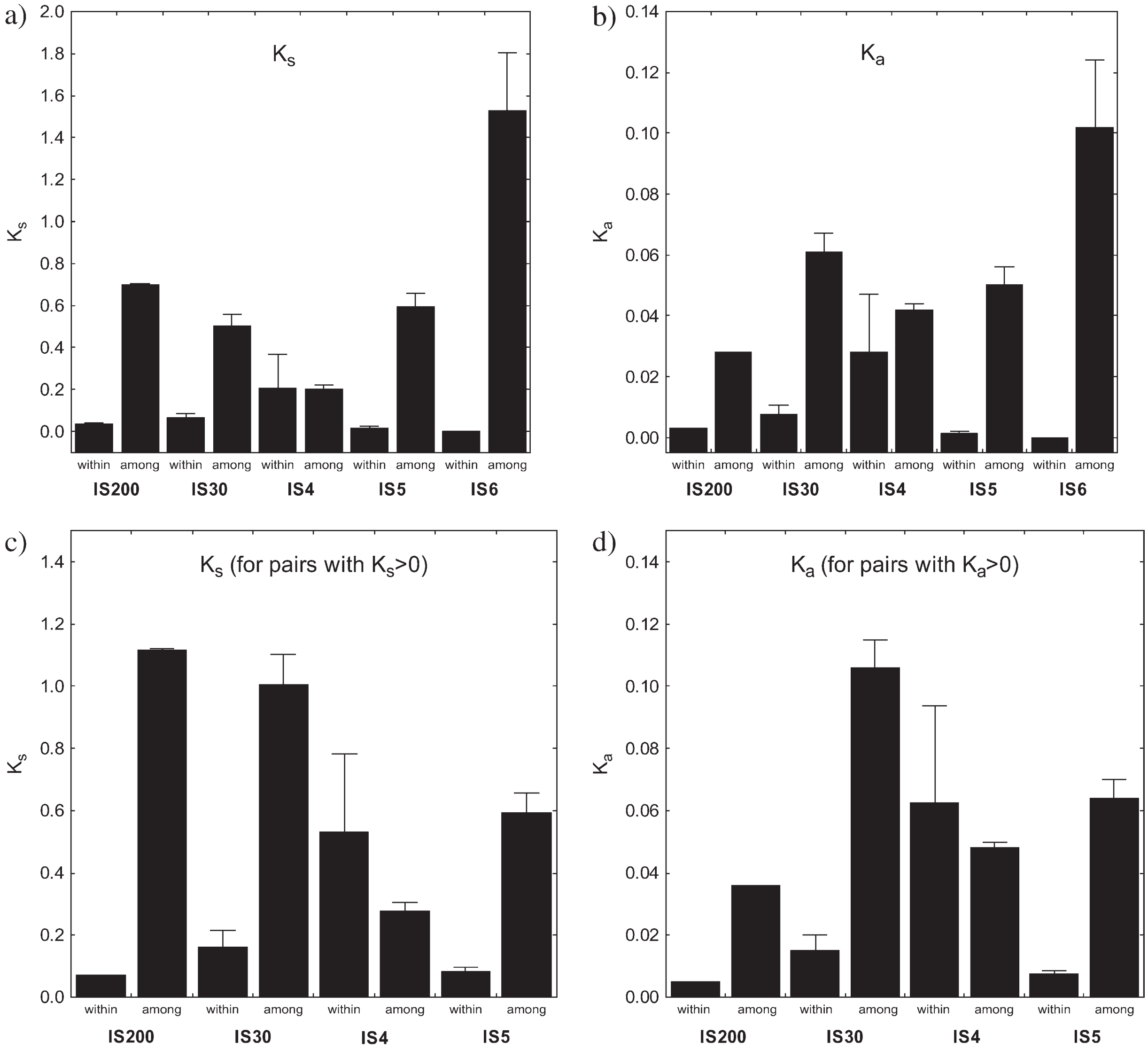

FIG. 2.—Greater sequence variation among genomes than within genomes. For each of the indicated insertion sequences, the bars labeled "within" correspond to the mean pairwise divergence of transposase genes within a genome (data from table 1). The bars labeled "among" correspond to the mean pairwise divergence of all transposase pairs, regardless of whether they occur in the same or in a different genome. (*a*) Mean synonymous divergence K_s. (*b*) Mean nonsynonymous divergence K_a; number of transposase pairs examined: n = 1,908 (IS4), n = 394 (IS5), n = 20 (IS6), n = 641 (IS30), and n = 34,588 (IS200). (*c*) Mean synonymous divergence K_s of all insertion sequence pairs with $K_s > 0$; n = 1,383 (IS4), n = 99 (IS5), n = 16 (IS6), n = 321 (IS30), and n = 21,748 (IS200). (*d*) Nonsynonymous divergence K_a of all insertion sequence pairs with $K_a > 0$; n = 1,653 (IS4), n = 99 (IS5), n = 17 (IS6), n = 373 (IS30), and n = 27,398 (IS200). Panels (*c*) and (*d*) do not contain information on IS6 because the within-genome divergence of the IS6 elements studied is zero. Error bars indicate standard errors, and their absence indicates lack of variation in divergence.

Gene conversion can be identified through the existence of conversion tracts, short stretches of nucleotides along which some members of a gene family are more similar to each other than to other members. Importantly, which members are highly similar to each other changes along the coding sequence, indicating that gene conversion homogenized short stretches of nucleotides among some members but not among others. This can be detected through sequence regions with varying degrees of sequence similarity in a multiple alignment or through statistical incongruities in phylogenetic trees generated from DNA fragments' multiple sequence alignments. I applied two commonly used tests for gene conversion to the insertion sequences from table 1. The first of them is the Sawyer test (Sawyer 1989), which detects pairs of sequence fragments of unusually high similarity in multiple sequence alignments. It then determines the statistical significance P of such pairs using two complementary methods, a method similar to scoring sequence similarity in the popular alignment tool Blast (Altschul et al. 1990), and a permutation test. The first scoring method yielded no significantly similar fragment pairs for the 21 entries of table 1. The second scoring method showed five of 21 entries of table 1 with $P < 0.05$. However, a Bonferroni correction for the 20 independent statistical tests renders none of these five P values statistically significant. I then also carried out the Farris ILD test (Farris et al. 1995), which determines likely recombination break points within a multiple sequence alignment through the incongruence of

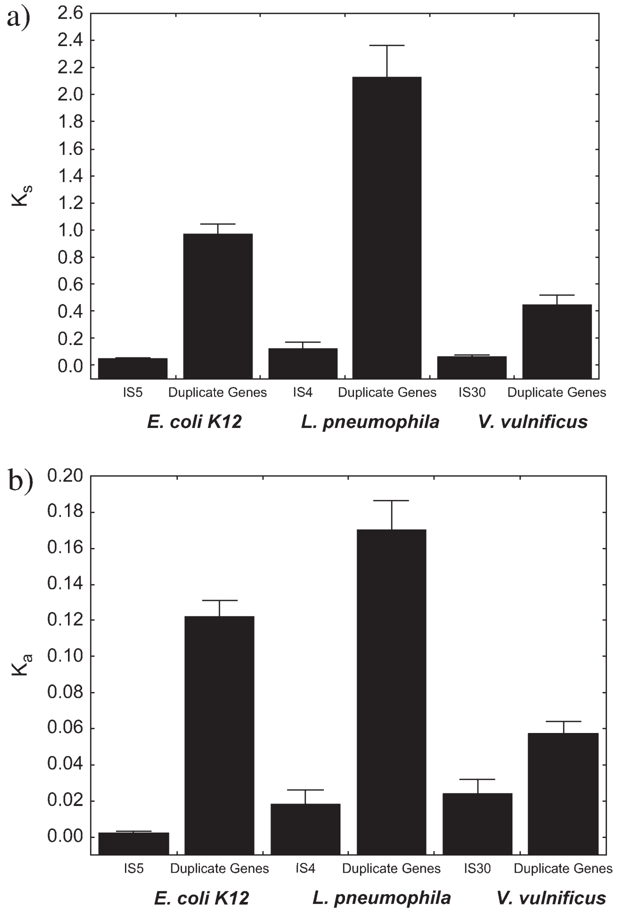

FIG. 3.—Gene duplicates within a genome are much more diverse than insertion sequences. The figure compares (*a*) synonymous divergence K_s and (*b*) nonsynonymous divergence K_a (vertical axes) of the insertion sequence transposase genes indicated on the horizontal axes (bars 1, 3, 5 from the left in both panels) within three indicated genomes to the same indicators of divergence for all duplicate genes in these genomes (bars 2, 4, 6). Divergence data on insertion sequences are taken from table 1. The comparison is shown for only three species because of the considerable computational cost of estimating K_a and K_s for all duplicates within a genome. The height of error bars corresponds to one standard error. For panel (*a*) gene duplicates that were saturated in K_s were excluded from the calculation.

phylogenetic trees derived from different subsets of the alignment (Farris et al. 1995). None of the sequences analyzed here showed evidence for gene conversion. In sum, two independent tests suggest that gene conversion is not a major factor in the high observed sequence homogeneity.

Finally, I note that insertion sequences are much more homogeneous than gene duplicates (fig. 3). Because both kinds of sequences would be subject to gene conversion to similar extents, they should show similar divergence, if gene conversion was solely responsible for the similarity of insertion sequences. However, this is not the case, suggesting that concerted evolution among gene-coding regions cannot be solely responsible for the homogeneity of insertion sequences.

Discussion

To summarize, the members of five prominent insertion sequence families examined here show very low nucleotide variation within bacterial genomes. Specifically, both synonymous and nonsynonymous variation in their transposase genes are significantly lower within genomes than among genomes. In addition, both kinds of variations are significantly lower for insertion sequences than for other duplicate genes within a genome. I emphasize that these results pertain strictly only to insertion sequences whose transposases are encoded by one open reading frame. However, the same may hold for other insertion sequences. For example, more limited sequence and taxonomic information suggests that IS1, where two open reading frames encode the transposase, also shows very low nucleotide variation (Sawyer et al. 1987).

Gene Conversion Is an Unlikely Sole Cause for the Low Divergence

Concerted evolution is an unlikely candidate explanation for this homogeneity. First, two different tests for gene conversion do not detect its signatures. A caveat is that the number of insertion sequence families polymorphic enough to examine for conversion tracts is small, and conversion tract lengths could be longer than the transposase open reading frames (≈700–1,400 bp). This, however, would not be consistent with observations suggesting that conversion tracts in dispersed bacterial genes are usually much shorter (Abdulkarim and Hughes 1996; Liao 2000). In addition, if gene conversion was a process of general importance for all dispersed coding sequences, one would expect gene duplicates also to be subject to it. Thus, transposase genes and duplicate genes should show comparable levels of divergence. However, transposase genes are much more homogeneous than gene duplicates. Relatedly, highly expressed genes are known to evolve slowly (Drummond et al. 2005). Because of the tight regulation of transposase activity in wild-type cells (Mahillon and Chandler 1998), transposase genes are among the most lowly expressed genes. If anything, they should thus evolve faster than other gene duplicates and be more diverse. The data show the opposite.

In addition, it may be relevant that multiple documented cases of gene conversion among dispersed genes (Santoyo and Romero 2005) affect highly expressed genes. Examples include genes encoding ribosomal DNA and elongation factor Tu (Abdulkarim and Hughes 1996; Liao 2000). The products of both are in very high demand in the cell and need to be expressed at stoichiometric levels with other proteins necessary for translation. For these proteins, gene conversion may serve adaptive roles, namely to slow the accumulation of deleterious mutations that may affect the concentration of biologically active gene product. In contrast, transposase genes are expressed at very low levels, sometimes at less than one copy per cell (Kleckner 1990), partly because very high transposition activities may be detrimental to the host and need to be avoided. Taken together, these observations argue against gene conversion as the sole explanation of insertion sequence diversity.

Transposition Rates Are Much Higher than Nucleotide Substitution Rates

If gene conversion does not cause the high homogeneity of these sequences, then what does? To get at the answer,

Table 2
Insertion Sequences Are Not Closely Linked

			Nearest Neighbor Distance[a] (number of coding sequences)		
Species[b]	Insertion Sequence	Copies ($K_s = 0$)[c]	Mean	σ[d]	Minimum
Azoarcus sp. *EbN1*	IS5	5 (5)	825.6	999.1	7
Coxiella burnetii RSA 493	IS30	5 (5)	400.8	497.3	6
Escherichia coli CFT073	IS200	18 (8 + 10)	297.8	456.5	13
Escherichia coli K12	IS30	3 (1)	1416.3	1230.6	363
Escherichia coli K12	IS5	11 (9)	385.6	346.7	35
Francisella tularensis[e]	IS5	16 (15)	127.7	94.02	21
Idiomarina loihiensis L2TR	IS4	5 (5)	524.6	726.9	32
Lactococcus lactis subsp. *Lactis Il1403*	IS30	15 (15)	153.7	202.1	6
Legionella pneumophila str. Philadelphia 1	IS4	4 (3)	734.5	489.3	3
Legionella pneumophila str. Pris	IS4	9 (5)	341.4	350.5	37
Neisseria meningitidis MC58	IS30	14 (4 + 2 + 2)	147.4	128.9	6
Photorhabdus laumondii	IS30	6 (4)	816.2	570.1	214
Rhodopirellula baltica SH 1	IS4	10 (4 + 2)	664.3	620.3	7
Shigella flexneri 2a str.301	IS4	16 (2 + 2 + 2)	276.1	269.4	16
Streptococcus thermophilus CNRZ1066	IS30	10 (9)	190.3	128.6	12
Streptococcus thermophilus LMG 18311	IS30	8 (7)	235.1	247.6	31
Streptomyces avermitilis MA-4680	IS4	7 (2 + 3)	571.7	1091.2	0
Streptomyces avermitilis MA-4680	IS6	3 (3)	3643.5	5149.9	2
Vibrio vulnificus CMCP6 chromosome I	IS30	7 (4)	420.7	1078.3	3
Wolbachia (endosymbiont of *Drosophila melanogaster*)	IS4	3 (3)	397.3	351.6	124
Xanthomonas oryzae	IS30	11 (8)	420.5	298	18

[a] Distances are given as the number of coding sequences between two neighboring insertion sequence copies. On a circular molecule, the mean distance of n genes is equal to a constant times $1/n$. Therefore, it is not the mean distance that is informative but the relationship between mean distance and its standard deviation σ. The standard deviation of nearest neighbor distance is smaller than the mean in 10 of 21 cases shown here and is greater than 150% of the mean in only three cases. Note that insertion sequences that transposed into other insertion sequences are not accessible through this analysis.

[b] RefSeq accession numbers of GenBank files for each genome can be found in Methods. All chromosomes analyzed are circular except that of *S. avermitilis*.

[c] Numbers in parentheses indicate the number of copies identical at synonymous sites. Except for IS30 of *L. lactis*, which has one nonsynonymous but no silent substitutions, these are also the number of copies with complete nucleotide identity.

[d] "σ" indicates the standard deviation of nearest neighbor distances.

[e] *F. tularensis* subsp. *tularensis Schu 4*.

it is useful to compare transposition rates to synonymous substitution rates. Estimates of transposition rates for bacterial insertion sequences and similar transposable elements in eukaryotes range over two orders of magnitude, between 10^{-3} and 10^{-5} per infected cell and generation, with excision rates (insertion sequence loss) typically one order of magnitude lower (Egner and Berg 1981; Hartl et al. 1983; Shen, Raleigh, and Kleckner 1987; Berg and Howe 1989; Charlesworth and Langley 1989; Kleckner 1990). Importantly, this rate of transposition refers to successful transposition events, events without strong deleterious effects that would kill the host. Such transposition events initially occur only in a single cell of a population. If most such transposition events are neutral, then the rate at which they arise and go to fixation is independent of population size and identical to the transposition rate. If a substantial proportion of transposition events is beneficial, this rate will be higher, and if a substantial proportion is weakly deleterious, it will be lower. Whether most successful transposition events are neutral, beneficial, or weakly deleterious is essentially unknown.

Transposition and excision rates are small on the timescale of laboratory biology, but they are very large on the timescale at which DNA sequences change. In wild *E. coli*, for example, it has been estimated that synonymous nucleotide substitutions accumulate at a rate of $K_s = 0.0045$ ($K_s = 0.009$ per gene pair) every 10^6 years (10^8–3×10^8 generations) (Ochman, Elwyn, and Moran 1999). Even if most transposition events are weakly deleterious, such that the rate at which they arise and go to fixation is only 10^{-4}–10^{-6} per generation (10% of the transposition rate), then between 100 and 30,000 ($3 \times 10^8/10^4$–$10^8/10^6$) transposition events can arise and go to fixation by the time two nucleotide sequences accumulate 1% of sequence divergence.

Insertion Sequences in a Genome Have Often Been Recently Acquired

These considerations explain quite naturally why insertion sequences are much more homogeneous than gene duplicates. If one thinks of transposase genes as promoting their own duplication, they should be more homogeneous than gene duplicates because they duplicate much faster. However, this explanation raises another question: why are there not many more insertion sequences in bacterial genomes if they arise so rapidly and if their excision rate is much smaller than the transposition rate? There are two possible answers. First, the examined genomes may be saturated with insertion sequences, such that all possible insertion sites are occupied. Second, the observed insertion sequences may have entered the genome very recently. The first possibility is unlikely on several grounds. First, many insertion sequences do not seem to be very selective in their target site (e.g., IS6; Mahillon and Chandler 1998). In addition, more selective insertion sequences often have very short target sites that would occur thousands of times in a genome (e.g., IS5, whose target sequence is YTAG; Mahillon and Chandler 1998). Relatedly, bacterial genomes show a great amount of variation in the number of insertion sequences per million base pairs. For example, in the 376 bacterial genomes and plasmids examined in this study, the number of IS30 elements per Mbp varies almost by a factor 50 between a low 0.019 IS30 Mbp^{-1} (*E. coli* CFT) and a high 10.02 IS30 Mbp^{-1} (*S. thermophilus CNRZ1066).* Further relevant observations come from work showing that the number of insertion sequences in 71 natural isolates of *E. coli* can vary widely (Sawyer et al. 1987). For example, for IS5 examined here, it ranged from 0 to 21 copies for IS5. Taken together, this evidence suggests that saturation with insertion sequences cannot account for the low number of insertion sequences per genome.

Having eliminated saturation as a possible explanation, one is left with the second explanation, namely that the insertion sequences analyzed here have been acquired in the (very) recent evolutionary past. For example, insertion sequences that are completely identical to each other (six cases in table 1) have been in the genome for less than the smallest time unit of molecular evolution, the time it takes to acquire one synonymous nucleotide change. If one saw this pattern in only one species, or with one kind of insertion sequence, it might well be coincidental. However, the diversity of genomes and insertion sequences examined here suggests otherwise and means that this pattern has general implications for the persistence of insertion sequences in bacterial genomes.

The main implication of this pattern is that insertion sequences may go periodically extinct in bacterial populations and become reintroduced by horizontal transfer. This is because several other scenarios are not consistent with the data of table 1, neither for *E. coli* nor for the other 16 species. If insertion sequences did not go periodically extinct, they would show higher divergence within a genome. If natural selection resulted in a net increase of insertion sequence copy number in the long run, then insertion sequences should be much more diverse within a genome because they would remain part of the genome indefinitely. The same should hold if not selection but the downregulation of transposition activity with increasing copy number reduced the number of insertion sequences within a genome. Finally, if they were not reintroduced by horizontal transfer, bacterial genomes would be devoid of insertion sequences.

The Evolutionary Forces Determining Insertion Sequence Copy Number

Earlier work on *E. coli*, one of the taxa of table 1, indicates which factors might be important in influencing insertion sequence copy numbers (Sawyer et al. 1987). This work examined among genomes of 71 *E. coli* strains the copy number distribution of six different insertion sequences, three of which (IS4, IS5, and IS30) I also examine here. This distribution is qualitatively similar to the distribution of the insertion sequences in the wide spectrum of bacterial species examined here (fig. 1). Specifically, this distribution is dominated by strains with no or few copies of any one insertion sequence. To use but the example of IS5, 46 strains have zero copies, 12 have one copy, 3 have two copies, 2 have three copies, 2 have four copies, and 6 have more than five copies of IS5. The strain with the maximal number of copies has 21 copies. Similar distributions are observed for IS4 and IS30.

Several evolutionary forces may shape this distribution. First, transposition will increase copy number, whereas excision will reduce it. Left unchecked, these two processes would lead to a net increase in copy number over time because excision is much rarer than insertion. For instance, the insertion sequence IS10, where pertinent measurements are available, transposition and excision occur at respective rates of 10^{-3} and $<10^{-9}$ per cell generation (Shen, Raleigh, and Kleckner 1987; Kleckner 1989). Second, natural selection may increase copy number further or decrease it. Because transposition, excision, and cell death due to a high transposable element load are stochastic events, one best thinks of the resulting distribution of copy number as a "stochastic" equilibrium. The observation alone that *E. coli* strains with many insertion sequence copies are rare indicates that the net effect of natural selection is a decrease in copy number. This is further supported by fitting the copy number distribution with an explicit quantitative model of the joint action of transposition and natural selection, which also takes into account the downregulation of transposition activity with increasing copy number (Sawyer and Hartl 1986; Sawyer et al. 1987).

These results, in conjunction with the recent acquisition of *E. coli* IS4 and IS30 (table 1), and the low sequence diversity of other insertion sequences within enteric bacteria (Lawrence, Ochman, and Hartl 1992) are best explained by the following evolutionary scenario. A virgin genome is infected by a piece of mobile DNA that carries one or more copies of an insertion sequence. This infection may be facilitated by genes carried on the mobile DNA that are beneficial to the host. Similarly, establishment of the insertion sequence may be favored by temporary benefits it provides (Blot 1994; Naas et al. 1994; Schneider et al. 2000; Edwards and Brookfield 2003). The insertion sequence's copy number then expands rapidly through transposition (hence the low sequence diversity). Through the action of natural selection, perhaps aided by the downregulation of transposition activity and the occasional excision event, the insertion sequence becomes extinct again from the lineage. Some time thereafter, it may become reintroduced through horizontal gene transfer. In this scenario, if one were to follow a genome infected by transposable elements forward in time, it would either lose its insertion sequences through excision (less likely) or become obliterated through natural selection. If one would follow it backward in time, it would be devoid of insertion sequences until the infection events whose descendants now occupy the genome. I emphasize again that several other scenarios, such as an indefinite persistence of insertion sequences in a genome or a net beneficial effect of increasing insertion sequence copy numbers are not consistent with the data of table 1.

Horizontal transfer, as opposed to positive natural selection, has also been emphasized as an important reason for the maintenance of "mariner"-like transposable elements in eukaryotes (Capy et al. 1994; Lohe et al. 1995; Lampe et al. 2003). An important difference to the eukaryotic case is that transposition is much more tightly regulated in prokaryotes and usually restricted to *cis*-activity of the insertion sequence from which a transposase molecule is expressed (Mahillon and Chandler 1998). In contrast, eukaryotic tranposases can readily act in *trans* on unlinked insertion sequence copies. This means that deactivated or truncated eukaryotic transposable elements can undergo transposition via functional "helper transposons." In consequence, eukaryotic genomes are littered with deactivated transposon copies that continue to proliferate passively. This fact, together with the greater tolerance of eukaryotic genomes for repetitive DNA, means that insertion sequences can persist in a eukaryotic genome for much longer times.

Horizontal Transfer of Insertion Sequences

An important ingredient of the above explanation is that insertion sequences for horizontal transfer from uninfected to infected organisms must be available, preferentially in organisms that are closely related and in close proximity to the organism of interest. This is an uncontroversial proposition. Past work shows that it is the case for *E. coli* and closely related bacteria (Lawrence, Ochman, and Hartl 1992). For most genomes in table 1, the necessary taxonomic sampling of closely related lineages is sparse or nonexistent. Exceptions include the two *S. thermophilus* strains of table 1. The IS30 elements in each of them have multiple identical counterparts in the other strain. Similarly, the IS30 element of the *V. vulnificus* strain CMCP6 of table 1 has a counterpart with synonymous divergence $K_s = 0.11$ in *V. vulnificus* YJ016. Overall, insertion sequences in four (10) out of the 18 genomes of table 1 have related insertion sequences in some other bacterial genome with $K_s < 0.12$ ($K_a < 0.5$). There is little doubt that this number will increase with sufficiently fine taxonomic sampling.

In sum, the empirical data presented here suggest that whatever short-term benefits insertion sequences may provide, they are deleterious in the long run and will cause the extinction of the host genome. They are maintained in bacterial lineages through horizontal transfer. They are thus akin to infectious diseases, with two important differences: they kill entire host lineages and they act slowly on the timescale of hundreds of thousands of generations.

Supplementary Material

Figure S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

I would like to acknowledge support through National Institutes of Health grant GM63882. I am also indebted to two anonymous reviewers for their valuable suggestions.

Literature Cited

Abdulkarim, F., and D. Hughes. 1996. Homologous recombination between the tuf genes of *Salmonella typhimurium*. J. Mol. Biol. **260**:506–522.

Ajioka, J., and D. Hartl. 1989. Population dynamics of transposable elements. Pp. 185–210 *in* D. Berg and M. Howe, eds. Mobile DNA. American Society for Microbiology Press, Washington, D.C.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped Blast and Psi-Blast: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Berg, D. E., and M. M. Howe. 1989. Mobile DNA. ASM Press, Washington, D.C.

Bisercic, M., and H. Ochman. 1995. Natural populations of *Escherichia coli* and *Salmonella typhimurium* harbor the same classes of insertion sequences. Genetics **133**:449–454.

Blot, M. 1994. Transposable elements and adaptation of host bacteria. Genetica **93**:5–12.

Bushman, F. 2002. Lateral DNA transfer: mechanisms and consequences. Cold Spring Harbor University Press, Cold Spring Harbor, N.Y.

Capy, P., T. Langin, Y. Bigot, F. Brunet, M. J. Daboussi, G. Periquet, J. R. David, and D. L. Hartl. 1994. Horizontal transmission versus ancient origin—mariner in the witness box. Genetica **93**:161–170.

Charlesworth, B., and C. H. Langley. 1989. The population genetics of Drosophila transposable elements. Annu. Rev. Genet. **23**:251–287.

Chien, M. C., I. Morozova, S. D. Shi et al. (37 co-authors). 2004. The genomic sequence of the accidental pathogen Legionella pneumophila. Science **305**:1966–1968.

Conant, G. C., and A. Wagner. 2002. GenomeHistory: a software tool and its applications to fully sequenced genomes. Nucleic Acids Res. **30**:1–10.

———. 2003. Asymmetric sequence divergence of duplicate genes. Genome Res. **13**:2052–2058.

Cooper, V. S., M. Schneider, M. Blot, and R. E. Lenski. 2001. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli*. J. Bacteriol. **183**:2834–2841.

Craig, N., R. Craigie, M. Gellert, and A. L. Lambowitz. 2002. Mobile DNA II. ASM Press, Washington, D.C.

Doolittle, W. F., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm, and genome evolution. Nature **284**:601–607.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA **102**:14338–14343.

Edwards, R. J., and J. F. Y. Brookfield. 2003. Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. Mol. Biol. Evol. **20**:30–37.

Egner, C., and D. E. Berg. 1981. Excision of transposon Tn5. Proc. Natl. Acad. Sci. USA **78**:459–463.

Farris, J. S., M. Kallersjo, A. G. Kluge, and C. Bult. 1995. Constructing a significance test for incongruence. Syst. Biol. **44**: 570–572.

Glockner, F. O., M. Kube, M. Bauer et al. (14 co-authors). 2003. Complete genome sequence of the marine planctomycete Pirellula sp strain 1. Proc. Natl. Acad. Sci. USA **100**:8298–8303.

Goldman, N., and Z. H. Yang. 1994. Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Hall, B. G., L. L. Parker, P. W. Betts, R. F. DuBose, S. A. Sawyer, and D. L. Hartl. 1989. IS103, a new insertion element in Escherichia coli: characterization and distribution in natural populations. Genetics **121**:423–431.

Hartl, D. L., D. E. Dykhuizen, R. D. Miller, J. Green, and J. de Framond. 1983. Transposable element IS50 improves growth rate of *E. coli* cells without transposition. Cell **35**:503–510.

Kleckner, N. 1989. Transposon Tn10. Pp. 211–226 *in* D. Berg and M. Howe, eds. Mobile DNA. American Society for Microbiology Press, Washington, D.C.

———. 1990. Regulating Tn10 and IS10 transposition. Genetics **124**:449–454.

Lampe, D. J., D. J. Witherspoon, F. N. Soto-Adames, and H. M. Robertson. 2003. Recent horizontal transfer of mellifera subfamily mariner transposons into insect lineages representing four different orders shows that selection acts only during horizontal transfer. Mol. Biol. Evol. **20**:554–562.

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

Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Liao, D. 2000. Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. J. Mol. Evol. **51**:305–317.

Lohe, A. R., E. N. Moriyama, D. A. Lidholm, and D. L. Hartl. 1995. Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. Mol. Biol. Evol. **12**:62–72.

Mahillon, J., and M. Chandler. 1998. Insertion sequences. Microbiol. Mol. Biol. Rev. **62**:725–774.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutiuon rates, with application to the chloroplast genome. Mol. Biol. Evol. **11**:715–724.

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.

Ochman, H., S. Elwyn, and N. A. Moran. 1999. Calibrating bacterial evolution. Proc. Natl. Acad. Sci. USA **96**:12638–12643.

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

Rabus, R., M. Kube, J. Heider, A. Beck, K. Heitmann, F. Widdel, and R. Reinhardt. 2005. The genome sequence of an anaerobic aromatic-degrading denitrifying bacterium, strain EbN1. Arch. Microbiol. **183**:27–36.

Santoyo, G., and D. Romero. 2005. Gene conversion and concerted evolution in bacterial genomes. FEMS Microbiol. Rev. **29**:169–183.

Sawyer, S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6**:526–538.

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

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.

Schneider, D., E. Duperchy, E. Coursange, R. E. Lenski, and M. Blot. 2000. Long-term experimental evolution in *Escherichia coli.* IX. Characterization of insertion sequence-mediated mutation and rearrangements. Genetics **156**:477–488.

Schneider, D., and R. E. Lenski. 2004. Dynamics of insertion sequences elements during experimental evolution of bacteria. Res. Microbiol. **155**:319–327.

Seshadri, R., I. T. Paulsen, J. A. Eisen et al. (24 co-authors). 2003. Complete genome sequence of the Q-fever pathogen Coxiella burnetii. Proc. Natl. Acad. Sci. USA **100**:5455–5460.

Shen, M. M., E. A. Raleigh, and N. Kleckner. 1987. Physical analysis of Tn10 and IS10-promoted transpositions and rearrangements. Genetics **116**:359–369.

Thompson, J., D. Higgins, and T. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Treves, D. S., S. Manning, and J. Adams. 1998. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli.* Mol. Biol. Evol. **15**:789–797.

Welch, R. A., V. Burland, G. Plunkett et al. (19 co-authors). 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. USA **99**:17020–17024.

Herve Philippe, Associate Editor

Accepted December 12, 2005