Causes of Insertion Sequences Abundance in Prokaryotic Genomes

Marie Touchon† and Eduardo P. C. Rocha*†*

*Génétique des Génomes Bactériens, CNRS URA2171, Institut Pasteur, Paris, France; and †Atelier de Bioinformatique, Université Pierre et Marie Curie-Paris6, Paris, France

Insertion sequences (ISs) are the smallest and most frequent transposable elements in prokaryotes where they play an important evolutionary role by promoting gene inactivation and genome plasticity. Their genomic abundance varies by several orders of magnitude for reasons largely unknown and widely speculated. The current availability of hundreds of genomes renders testable many of these hypotheses, notably that IS abundance correlates positively with the frequency of horizontal gene transfer (HGT), genome size, pathogenicity, nonobligatory ecological associations, and human association. We thus reannotated ISs in 262 prokaryotic genomes and tested these hypotheses showing that when using appropriate controls, there is no empirical basis for IS family specificity, pathogenicity, or human association to influence IS abundance or density. HGT seems necessary for the presence of ISs, but cannot alone explain the absence of ISs in more than 20% of the organisms, some of which showing high rates of HGT. Gene transfer is also not a significant determinant of the abundance of IS elements in genomes, suggesting that IS abundance is controlled at the level of transposition and ensuing natural selection and not at the level of infection. Prokaryotes engaging in obligatory associations have fewer ISs when controlled for genome size, but this may be caused by some being sexually isolated. Surprisingly, genome size is the only significant predictor of IS numbers and density. Alone, it explains over 40% of the variance of IS abundance. Because we find that genome size and IS abundance correlate negatively with minimal doubling times, we conclude that selection for rapid replication cannot account for the few ISs found in small genomes. Instead, we show evidence that IS numbers are controlled by the frequency of highly deleterious insertion targets. Indeed, IS abundance increases quickly with genome size, which is the exact inverse trend found for the density of genes under strong selection such as essential genes. Hence, for ISs, the bigger the genome the better.

Introduction

Insertion sequences (IS) are the simplest form of transposable elements, coding only the information allowing their mobility (Mahillon and Chandler 1998). They belong to the numerous group of type I transposable elements, widespread in eukaryotes, prokaryotes, viruses, and plasmids and are themselves the most widely distributed transposable elements in the replicons of bacteria and archaea (Mahillon and Chandler 1998; Syvanen 1998; Wagner 2006). Through a succession of enzymatic reactions catalyzed by the transposase, IS can spread within a genome. IS elements are transferred between genomes by all the classical mechanisms of horizontal gene transfer (HGT) (Frost et al. 2005). In association with other elements, they can mediate the transfer of genetic information between genomes or between replicons of the same genome. They have thus been found to shuttle the transfer of adaptive traits, such as antibiotic resistance (Boutoille et al. 2004), virulence (Kunze et al. 1991; Lichter et al. 1996), and new metabolic capabilities (Schmid-Appert et al. 1997). This is just one of the many effects these elements have on genomes, where they can also induce duplications, deletions, and rearrangements (Naas et al. 1995; Papadopoulos et al. 1999; Alokam et al. 2002; Kothapalli et al. 2005; Redder and Garrett 2006). Because of their effects, ISs are regarded as key determinants of genome plasticity (Kolsto 1997; Casjens 1998; Syvanen 1998; Bentley and Parkhill 2004; Rocha 2004; Schneider and Lenski 2004) and have been suggested to provide significant adaptive changes to genomes (Shapiro 1999b; Capy et al. 2000). Yet, very few demonstrated cases have been found for which the action of ISs led to the direct acquisition of an advantageous trait, contrary to the situation of some transposable elements in eukaryotes (Volff 2006). Hence, ISs, as other transposable elements, are frequently thought to persist in natural populations because of their infective character irrespective of the fitness loss or gain they lead to (Doolittle and Sapienza 1980; Orgel and Crick 1980; Charlesworth B and Charlesworth D 1983; Condit et al. 1988; Nuzhdin 1999; Kidwell and Lisch 2001). According to this view, if the deleterious effects of transposition rates are very high, only highly infective ISs will persist. The selfish hypothesis is corroborated by the extensive degree of HGT of ISs (Hartl and Sawyer 1988), their highly incongruent phylogenetic trees (Daniels et al. 1990), and by the diversity of ISs found within strains or closely related species (Lawrence et al. 1992; Wagner 2006). This hypothesis is also frequently challenged, usually on the basis of their theoretically important role in the acquisition of adaptive changes (Chao et al. 1983; Blot 1994; Hall 1999; Shapiro 1999b).

The frequent reports indicating the variability in the type and frequency of transposable elements in different organisms led many authors to regard them as nearly ubiquitous (Mahillon and Chandler 1998; Capy et al. 2000; Kidwell and Lisch 2001; Petrov et al. 2003; Cordaux et al. 2006). Yet, the sequencing of the first bacterial genomes confirmed that *Bacillus subtilis*, a widely known bacterial model organism, lacks IS or any other transposable element (Kunst et al. 1997). This observation is especially striking at the light of the biology of *B. subtilis*: it is a frequently recombining, naturally competent bacterium (Istock et al. 1992), whose genome contains several bacteriophages and other remnants of HGT events (Moszer et al. 1999). One IS has been reported in 2 other strains of *B. subtilis* (Nagai et al. 2000), suggesting that the rarity of IS elements in this species is not caused by any special mechanism precluding their insertion in the genome. Subsequent genome sequencing showed that other genomes, mostly small in size, also lacked IS elements (Siguier et al. 2006). Hence, absence of

Key words: Transposable elements, virulence, genome dynamics, genome size, growth.

E-mail: marie@abi.snv.jussieu.fr.

Mol. Biol. Evol. 24(4):969–981. 2007
doi:10.1093/molbev/msm014
Advance Access publication January 23, 2007

ISs may be more widespread than usually acknowledged. On the other hand, some genomes contain many IS elements. Previous analyses of IS distribution have detailed the distribution of either IS families among many strains of a species (Lawrence et al. 1992) or some IS families throughout all bacterial genomes (Wagner 2006). Both reports showed a very large range for the number of IS elements in prokaryotic genomes, even when comparing closely related ones. The same picture emerges from the analysis of closely related strains of genomes particularly enriched with IS elements such as *Shigella* (Buchrieser et al. 2000; Jin et al. 2002; Wei et al. 2003; Yang et al. 2005), *Yersinia* (Parkhill et al. 2001; Deng et al. 2002; Chain et al. 2004), and *Bordetella* (Parkhill et al. 2003).

Several hypotheses have been put forward to explain the wide variability of IS elements among the genomes of bacteria and archaea, most of which have finally become testable using available genome sequences. 1) Several IS families have been proposed to be of very narrow phylogenetic range, and by historical or mechanistic reasons some lineages might then lack and/or be immune to many IS families (Mahillon and Chandler 1998). 2) Many authors hold that ISs only persist by periodic invasion of naif genomes (Wagner 2006). This should compensate for losses caused by their deleterious fitness effects. If so, one might expect the number of IS elements to be positively correlated with the rate of HGT. 3) Smaller genomes might have lower IS density: (i) if genome size is under selection and smaller genomes reflect stronger selection for this trait, (ii) because transposition increases genome size, and (iii) because transposition is expected to be more deleterious when the fraction of highly deleterious insertion sites (e.g., essential genes) in the genome increases. 4) In the literature, one also finds pervasive associations between the abundance of ISs and pathogenicity, especially among emergent or facultative pathogens. This association has been explained in 2 different ways. Within the selfish gene hypothesis, smaller population sizes, frequently associated with changes in ecological niches or selective sweeps in pathogens, would involve relaxed selection against the presence of ISs (Parkhill et al. 2003; Moran and Plague 2004). Within the paradigm that IS elements are advantageous for genomes, ISs are thought as partly responsible for the evolutionary breakthroughs leading to the invasion of a new niche (Shapiro 1999a; Chain et al. 2004). 5) A variant of the latter ideas suggests that facultative ecological associations are correlated with higher IS abundance. 6) Finally, a recent work suggested that the changes associated with the Neolithic revolution, some 12,000 years ago, led human-associated prokaryotes to be invaded by IS elements while adapting to the growing population of human hosts (Mira et al. 2006). This would affect the prokaryotes in intimate association with man or domesticated animals and plants and provide the genome fluidity necessary for the ecological adaptation. A common theme to all of these hypotheses is that they have never been tested in a large-scale comparative genomics framework.

Given the large number of hypotheses put forward to explain the variance in the distribution of IS elements among genomes, we aimed at unraveling which of them significantly fit the available data. A major difficulty with such an analysis is that annotation of IS elements in genomes is heterogeneous in quality and that different nomenclatures have been used for different and sometimes within genome annotations. This has led most genome analyses of ISs to rely on the available annotations or in using the small set of IS families that are easier to annotate, for example, because they have no multiple open reading frames. IS elements are classified into several families based on the transposase protein sequence, their small inverted repeats, and the length of their target site sequence. Although several classification schemes have been proposed and are currently used to annotate genomes, the scheme of Mahillon and Chandler (1998) has become over the years the main reference. Here, we reannotated the ISs of 262 bacterial/archaeal genomes and classed them according to this scheme. We then used this curated data set to assess the distribution of ISs and some of the variables that may be important to understand their dynamics.

Methods and Data

Data

We analyzed the genomes of 262 different bacterial and archaeal genomes, taken from GenBank genomes (ftp://ftp.ncbi.nih.gov/genomes/). Data on the genes putatively horizontally transferred was taken from the Horizontal Gene Transfer Database (HGT-DB) (http://www.tinet.org/~debb/HGT/) (Garcia-Vallve et al. 2003). To analyze prophages, we used the prophage prediction data of the Brussow group (Canchaya et al. 2004), refining and systematizing the list previously reported by Casjens (2003), which covers a total of 115 genomes. Genomes not analyzed in the previous works and presenting less than 20 putative phage annotated genes were considered as lacking prophages. This threshold is justified by the fact that many genes with weak similarities to databank proteins come out as having weak similarities to prophage-related genes even though they are not functionally characterized and have no contiguous phage-related genes, that is, they show no evidence of belonging to prophages.

Reannotation–Reassessment of IS Data

We reassessed the annotation of transposases by making BlastP analyses of all coding sequence (CDS) identified in GenBank files against the ISfinder database (www.is.biotoul.fr; Siguier et al. 2006). This database is a collection and classification of around 1500 IS proteins from bacterial and archaeal IS elements which are grouped into 21 distinct families (Mahillon and Chandler 1998). A CDS was considered as belonging to an IS element if its BlastP best hit had an E value $<10^{-10}$. To class IS elements, we assigned each putative IS CDS to the family to which it shows the best BlastP hit in the ISfinder database.

Having identified the CDSs corresponding to transposases and classed them into families, we proceeded to the reconstruction of the ISs. Most IS elements have a single CDS that encodes a transposase, in which case we assimilate the IS to its transposase. Note that terminal repeats of ISs are small and poorly characterized and in many cases do not allow a precise definition of the edges

of the elements. For example, many of these repeats are less than 15 bp, but there is a significant chance to find by chance alone 2 copies of a 15 bp generic repeat within a 2 kb DNA sequence (Rocha 2003a). Because, we are only interested in IS sequences that are putatively autonomous and not interested in elements such as miniature inverted-repeats transposable elements, this is not a serious inconvenience.

Some elements, such as the ones of IS1 and IS3 families, have 2 consecutive and partially overlapping CDSs, from which the transposase is produced by translational frameshift. In the elements of IS1 family, the absence of frameshift leads to a protein that instead of catalyzing transposition inhibits it (Machida C and Machida Y 1989; Zerbib et al. 1990). To handle multigenic IS elements, we reconstituted the integrity of the transposase by identifying consecutive CDSs, having the same orientation and respecting the order of the expected CDSs (e.g., IS1A and IS1B). Elements of the IS66 family are an exception in that they contain 3 successive CDSs in a strand and 1 or 2 in the other (Machida et al. 1984; Bonnard et al. 1989). This family has not been extensively studied, and although the 3 cooriented CDSs seem necessary for transposition, it is unclear if they code for one single protein (Han et al. 2001).

An IS element can be interrupted by the insertion of an IS or another genetic element. Thus, elements that fulfilled the above requirements but are stopped by the insertion of ISs or of other elements were considered as a single IS, but marked as pseudogenes or partial. The same was done for IS elements showing signs of large deletions (>20% of difference in protein length). We made, however, an exception to these rules if a smaller IS was present with the same size in at least 3 copies in the genome. In this case, we considered there was substantial evidence of an element engaging actively on transposition, and hence, we considered it as complete and a new member of the family. By using these criteria, we identified a total of 10,938 putative CDSs corresponding to 8,123 IS elements (of which we estimate 6,742 to be complete). It is important to note that the majority of these complete IS elements have not been tested for transposition activity and some may therefore carry mutations, which render them inactive. Hence, they should be regarded as putatively active transposases.

Annotation

To compare our annotation with the one of the genomes in GenBank, we identified all genes encoding putative IS elements with a broad panel of keywords in the RefSeq files (e.g., ISRm1, RSalpha-9, IS[0–9], transposase, IS, etc.). We identified a total of 10,017 putative CDSs of ISs in the GenBank files. We then defined a putative false negative annotation of IS as a CDS annotated as such in GenBank but not detected by our method and that presents at least 3 homologs (>90% protein similarity and <20% difference in length) in the genome. This is an element that does not have significant homologies in ISfinder but its multiplicity suggests, although does not demonstrate, its transposable character. Naturally, if an element has no homology with any known ISs and does not show signs of transposition we cannot identify it.

Evolutionary Distances

The evolutionary distances between prokaryotes were computed from the multiple alignment of 16S rDNA subunits, with Tree-Puzzle (Schmidt et al. 2002) using the HKY + Γ model.

Definition of Strain-Specific Regions

To check if ISs are overrepresented within laterally transferred regions, we used the data of HGT-DB and prophagic regions (see above) but also defined strain-specific regions. This was done by comparing very closely related pairs of genomes, that is, the ones for which the median similarity between the orthologs was between 97% and 99.9%. For this, we first identified the putative orthologs common to each pair of genomes. Orthologs were identified as unique pairwise reciprocal best hits with at least 90% similarity in amino acid sequence and less than 20% difference in protein length. Then, we made the union of pairwise lists of orthologs of each set of close genomes. Finally, we used this latter list of orthologs to detect specific regions in a given strain. A specific region was defined by identifying regions with at least 10 consecutive genes without an ortholog in the other close genomes.

Statistical Analysis

The statistical analyses were performed using R (http://www.r-project.org/). Most statistical results were obtained through nonparametric tests, notably Wilcoxon two-sample and Kruskal–Wallis multiple tests. Significance was usually defined as $P < 0.05$. For the parametric analyses, we transformed the data using a Box–Cox transformation on the number of IS elements +1 (because Box–Cox transformations require $X > 0$) (Draper and Gober 2002). The log transformation of the number of IS elements (+1) was found to be optimal and thus used throughout this article. We also tested several transformations on the number of IS elements directly (i.e., without adding +1 to the number of ISs). We found that the square root transformation was the best in this case, and showed results qualitatively similar to the ones of the logarithmic transformation (data not shown). The regression results are given as R^2 values, the coefficients of determination. They represent the fraction of the variation in the dependent variable explained by the independent variables in the model. In the stepwise regression the discrete binary variables were coded as −1/+1.

Box plot Representation

Box plot graphs summarize distributions of a set of data values divided in different groups. The upper and lower edges ("hinges") of the box are located at the first quartile (25th percentile of the data) and the third quartile (75th percentile). The central line indicates the sample median (50th percentile). The central vertical lines, called "whiskers" extend from the box as far as the data extend, to a distance of at most 1.5 interquartile ranges. An interquartile range refers to the distance between the first quartile and the third quartile sample. Any values more extreme than this are identified as outliers.

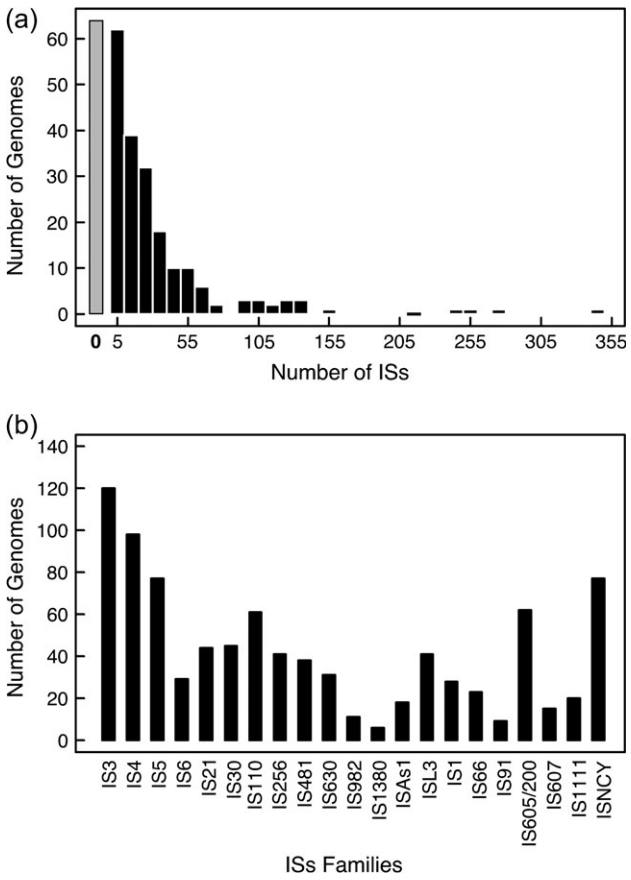

FIG. 1.—(*a*) Distribution of the number of complete IS elements in the 262 genomes. The first gray bar corresponds to the number of genome having no complete IS elements. The black bars correspond to the histogram of prokaryotic genomes (archaea and bacteria) harboring at least one IS, with class interval sizes of 10 ISs. (*b*) Distribution of the number of genomes containing at least one IS, for each IS family. We show the data for each of the 20 distinct ISs families, as well as for the group of unclassified elements (ISNCY). For more details, see Supplementary table 2 (Supplementary Material online).

Results

The Reassessment of ISs Confirms Their Rapid Dynamic of Gain and Loss

Because ISs elements are heterogeneously and incompletely annotated in the genomic databanks, we started by reassessing their annotations and classification. We examined 262 completely sequenced bacterial and archaeal genomes as described in Materials and Methods. We were able 1) to detect the transposases of the IS elements, 2) to assign a IS family for each of them, 3) to reconstitute the multigenic ISs, and 4) to characterize the CDSs of complete and truncated transposases. This allowed us to detect 88% (8,823) of the CDSs annotated as IS related in GenBank and to identify 20% (2,115) of new transposases (fig. 1*a*). Hence, we increased by about 10% the number of available annotated ISs (supplementary figure 1, Supplementary Material online). Classifying the transposases as hypothetical proteins was the major cause of underestimation of the number of ISs in GenBank (supplementary table 1, Supplementary Material online). On the other hand, most elements that indicated a functional link with transposable elements in GenBank that we missed show no significant similarity to known IS elements, that is, with the elements of ISfinder (Siguier et al. 2006). They are thus likely to be misannotated. To estimate how many IS elements we failed to identify, we analyzed the annotated putative transposases that we did not recognize as ISs and that showed signs of actively transposing in genomes (see Materials and Methods). We only found 227 such elements, which suggests a rate of false negatives around 2.3%. We were also able to consistently class the CDSs of ISs, when only 20% of GenBank ISs had a consistent classification. Hence, our reannotation shows that most transposases are annotated in current genome sequences but that the vast majority is not classed according to one coherent classification scheme and that the available annotation in the great majority of the cases, does not allow reconstituting the multigenic transposases.

We then analyzed the distribution of ISs in the different genomes, separating the complete elements, which may be active, from the ones showing signs of pseudogenisation, which probably are not autonomously transposing (see Materials and Methods). About 24% of the genomes have no complete IS element (fig. 1*a*), and 21% of the genomes also lack remnants of transposases among annotated CDSs. The reports of the lack of transposable elements in genomes arose quickly with genome sequencing (Kunst et al. 1997). However, their near ubiquity is still frequently claimed in the literature. Although a more detailed analysis of intergenic regions might reveal very small putative fragments of transposases or flanking repeats, this result clearly shows that the absence of functional ISs is widespread among prokaryotes. Genomes lacking ISs are not only abundant but also phylogenetically diverse. They include: 8 archaea, 9 chlamydia, 5 cyanobacteria, 5 actinobacteria, 3 firmicutes, 5 mollicutes, 12 α-proteobacteria, 4 ε-proteobacteria, 10 γ-proteobacteria, and 3 spirochaetes (table 1). They also cover a broad spectrum of life styles: 27% are free living and commensal, 11% facultative pathogens, 51% are obligatory pathogens, and 11% are obligatory mutualists. Although they tend to be small (see below), there are 4 genomes with more than 3 Mb in the set: *B. subtilis*, *Listeria innocua*, *Mycobacterium leprae*, and *Vibrio fischeri*. Furthermore, 48% of the remaining genomes have very few ISs (less than 10), indicating that ISs are absent or present at very moderate numbers in the majority of genomes (fig. 1*a*).

On the other hand, some genomes show remarkably high numbers of ISs and they also occur in very different clades (table 1). For example, among genomes with more than 60 complete ISs, we found the archaeon *Sulfolobus solfataricus* (123), the firmicute *Bacillus halodurans* (91), the α-proteobacteria *Nitrobacter winogradskyi* (105), the β-proteobacteria *Bordetella pertussis* (246), the γ-proteobacteria *Shigella sonnei* (342), and even the small genome of the mollicute *Mycoplasma mycoides* (63). We also computed the distribution of each IS family using the ISfinder classification (Mahillon and Chandler 1998; Siguier et al. 2006; fig. 1*b*). Some families are more frequent and diverse than others, but some families are more loosely defined than others and this contributes to explain their different abundance (Mahillon and Chandler 1998).

It has been abundantly demonstrated that IS elements have the ability to quickly multiply in genomes, resulting in

Table 1
Distribution of IS Elements among the 262 Genomes

Group	Number genomes	Number genomes IS = 0	% Genomes IS = 0	Median number IS	Min:Maximal number IS	Median genome size (Mb)	Median IS density (Mb)	Min:Maximal IS density (Mb)
Archaea	23	8	34.8	4	0:123	2.01	1.91	0:41.1
Firmicutes	52	3	5.8	12	0:91	2.81	4.29	0:24.2
Actinobacteria	20	5	25	23	0:77	0.30	5.50	0:31.3
Mollicutes	12	5	41.7	2	0:63	0.84	2.43	0:52
Cyanobacteria	11	5	45.4	1	0:53	2.59	0.37	0:20
Spirochaetes	5	3	60	0	0:51	1.14	0	0:10.9
α-Proteobacteria	27	12	44.4	5	0:105	2.06	1.46	0:30.9
β-Proteobacteria	19	0	0	32	2:246	4.30	11.03	0:60.2
δ-Proteobacteria	5	0	0	13	1:33	3.66	3.64	0.3:9.0
ε-Proteobacteria	6	4	66.6	0	0:14	1.72	0	0:6.6
γ-Proteobacteria	59	10	17	24	0:342	4.61	6.68	0:70.9
Chlamydiae	10	9	90	0	0:19	1.20	0	0:7.9
Cytophaga	3	0	0	39	8:42	5.28	6.24	1.5:17.9
Others	10	0	0	7	0:8	2.02	3.85	0:4.3
Total	262	63	24	12	0:342	2.75	3.52	0:70.9

that closely related species have very different numbers and families of IS elements (Lawrence et al. 1992; Wagner 2006). We therefore started by evaluating the rapidity with which IS elements evolve in genomes by plotting for each pair of genomes the difference in the number of ISs they contain versus the distance in their respective 16S rDNA (see Materials and Methods), which we used as a surrogate of time since the divergence of lineages (fig. 2). The results clearly indicate that the number of IS elements changes so fast that any correlation between time since the divergence and the number of ISs present in genomes quickly vanishes, that is, the phylogenetic signal is not persistent. Even the comparisons with genomes differing by less than 0.1 changes/nt in 16S rDNA (the 9.5% closest comparisons), show no significant correlation (fig. 2). This means that for practical purposes there is no phylogenetic inertia in the distribution of the number of IS elements in prokaryotes.

After reassessing and classifying the IS elements, we were able to test for the first time the theories that have been put forward to explain the differential abundance of IS elements in prokaryotic genomes. Throughout the remaining text when we mention IS, we denote an IS element that we could detect and that showed no signs of being nonfunctional. We found the same qualitative results when we included the few partial IS elements (data not shown). We also found the same qualitative results when we benchmarked our results with a small group of genomes recently reannotated by Chandler et al. (see Discussion).

The Effect of IS Families Specificity

If IS families are clade specific, one could suppose that genomes lacking IS elements either never had any specific IS family or that we could not identify it because it might lack sufficient resemblance with the other known IS families. Several lines of evidence argue against this hypothesis. Firstly, we found no clade-specific IS family (supplementary table 2, Supplementary Material online). Even families that were originally thought to be clade specific, such as IS1, in enterobacteria and IS66, in rhizobacteria, can be found in phylogenetically distant genomes (Mahillon and Chandler 1998). Furthermore, IS elements are very diverse within families, and most frequently the phylogeny of the elements is at odds with the expected one, as extensively discussed in previous works (see Lawrence et al. 1992; Wagner 2006 and references therein). For example, we found that several elements of the IS1 family are nearly identical between enterobacteria, *Bacillus cereus*, *Desulfotalea psychrophila*, and *Haemophilus ducreyi*. There are also similarities between IS elements of bacteria and archaea, which, given the rate of evolution of transposases, can hardly be explained by simple common descent. All these arguments strongly support the idea that ISs are highly adaptable and have the potential to invade genomes

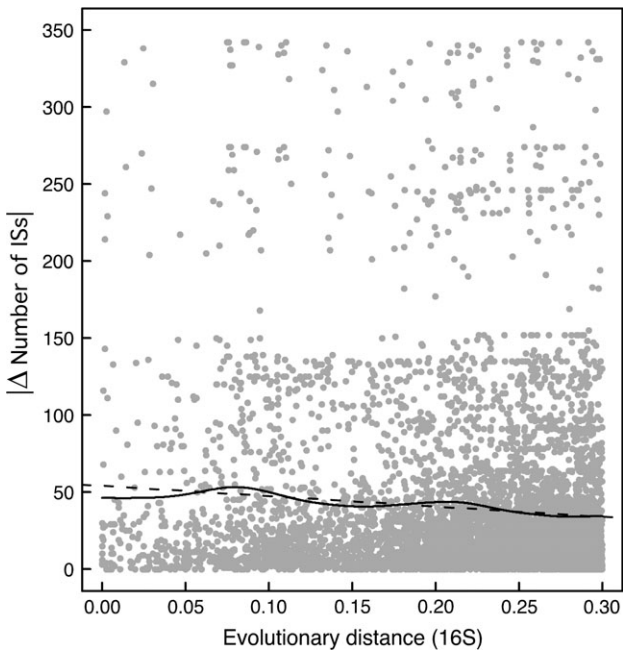

FIG. 2.—Lack of association between the absolute difference in the number of IS elements between 2 genomes and the time since their last common ancestor (see Materials and Methods). For the values in the range [0, 0.1], where the correlation should be stronger if there was phylogenetic inertia, the Spearman's correlation coefficient is 0.09, which is not significant ($P > 0.1$). The thick line is a spline interpolation showing that the lack of correlation spans the entire range of the graph, whereas the dashed regression line shows that the association between the 2 variables using the whole data is even slightly negative ($R^2 = 0.007$).

Table 2
Multiple Stepwise Regression for Number of ISs (log transformation)

Model	Covariate	Cumulative R^2	R^2
Number of ISs	Genome size	0.398***	0.398***
	Obligatory association	0.467***	0.237***
	G + C content	0.469	0.086**
	HGT abundance	0.469	0.173**

NOTE.—Significant codes: '***'0.0001; '**'0.001; '*'0.01; '.'0.05.

from all prokaryotic clades. Strains carrying one type of insertion element also tend to carry other types. The distribution of copy numbers is skewed such that the majority of isolates has no or few copies of a given sequence, and few isolates have a large number of ISs elements and a large diversity of IS families (fig. 3). The seemingly inescapable conclusion is that IS elements are continually being acquired by lateral transfer, which spans such large phylogenetic ranges that its absence in a genome can hardly be due to a lineage never infected. Hence, the widespread presence of many IS families, their rampant HGT, and the phylogenetic diversity of genomes lacking ISs suggests that this effect is unlikely to explain the variability of IS abundance among genomes.

The Effect of Genome Size

Because many of the genomes without IS elements are small, we analyzed the association between genome size and the number of IS elements. We started by comparing the 64 genomes with no IS elements (median genome length of 1.31 Mb), with the 198 genomes having at least one IS (median length of 3.42 Mb). The genome sizes of these groups are significantly different ($P < 0.0001$, Wilcoxon test). The overall correlation between genome size and IS number is also highly significant (Spearman's $\rho = 0.63$, $P < 0.0001$; fig. 4). To precisely estimate the contribution of genome size to IS abundance, we made a linear regression of the appropriately transformed variables (see Materials and Methods) finding that genome size accounts for ~40% of the overall variance ($P \ll 0.001$; table 2). This confirms the association between genome size and IS number, the larger the genome, the more IS elements it contains. The association between genome size and IS *density* is also highly significant (Spearman's $\rho = 0.43$, $P < 0.001$). This shows that larger genomes also have higher densities of IS elements.

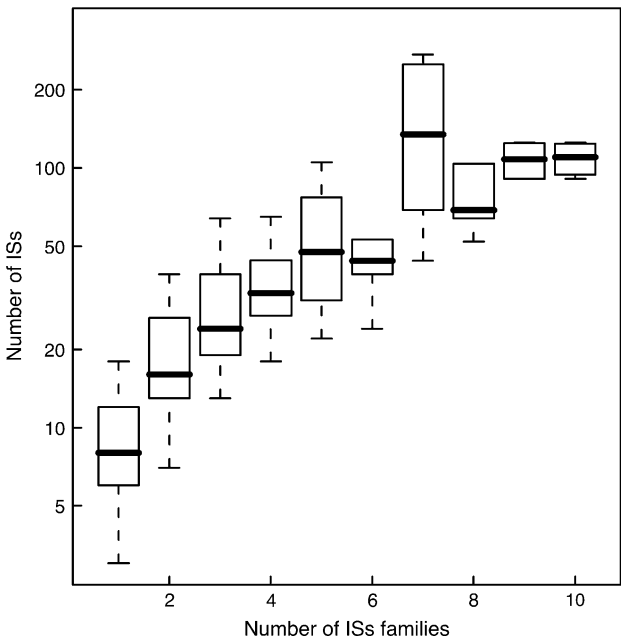

FIG. 3.—Box plot of the number of ISs elements among prokaryotes, when these are divided according to the number of ISs families. The correlation between the 2 variables is highly significant (Spearman's $\rho = 0.83$, $P < 0.0001$).

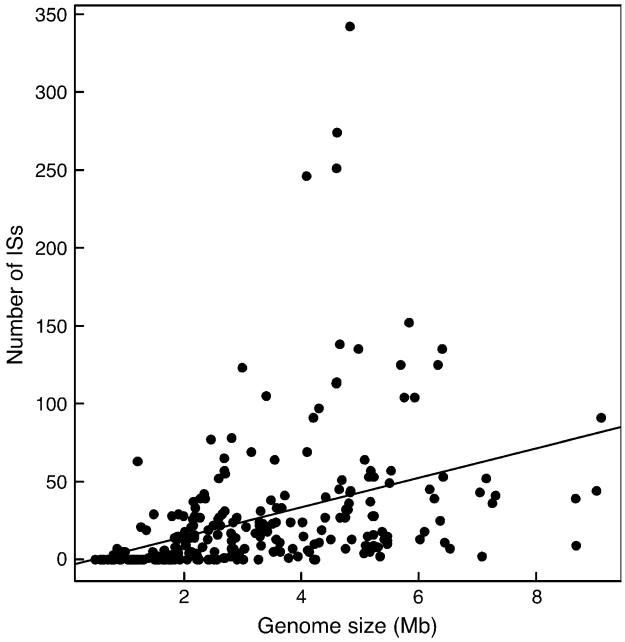

FIG. 4.—Scatter plot of the number of ISs against genome size (Mb). (Spearman's $\rho = 0.63$, $P < 0.0001$).

The association between genome size and IS abundance could result trivially from the ISs themselves occupying space in the chromosome: the more ISs, the larger the genome. For example, transposable elements represent about half of the human genome (Lander et al. 2001) and are thus a major cause of its size. We found no significant difference between the correlation of the number of ISs with the total number of genes in the genome (Spearman's $\rho = 0.62$, $P < 0.001$) and with the number of genes not contained in the ISs (Spearman's $\rho = 0.61$, $P < 0.001$). Thus, the control shows that although some genomes do have a significant fraction of their genome covered with IS (up to 8%), this is not affecting the correlation between genome size and IS abundance. Unsurprisingly, most genome size variation in prokaryotes is not caused by transposable elements.

The Effect of Horizontal Gene Transfer

The rapid dynamics of IS gain, expansion, and loss in the genomes of prokaryotes strongly depends on these elements being imported by HGT. ISs spread by all the known

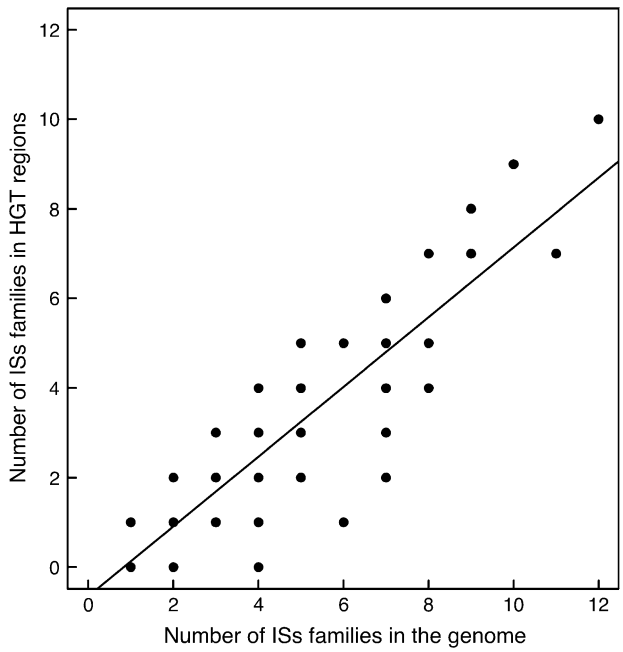

FIG. 5.—Scatter plot of the number of ISs families in HGT regions against the number of ISs families in the genome. (Spearman's $\rho = 0.84$, $P < 0.0001$).

types of HGT (Frost et al. 2005). One would thus expect a strong correlation between the abundance of IS elements and the frequency in the genomes of other genes arising from HGT. In the absence of data for the rates of HGT, we used the number of horizontally transferred genes over the genome size as a surrogate of HGT frequency. Three sources of data were considered: horizontally transferred regions given in HGT-DB (Garcia-Vallve et al. 2003), prophagic regions given in (Canchaya et al. 2004), and our own computed strain-specific regions (see Materials and Methods). We made the union of the 3 sets, which we will use henceforth. The controls using each data source separately gave similar qualitative results (see below). Genomes lacking ISs have fewer horizontally transferred genes than genomes with ISs (medians 5.4% and 11.1%, $P = 0.002$, Wilcoxon test). Both the number and density of ISs observed in HGT regions are correlated with the number and density of ISs in genomes (Spearman's $\rho = 0.75$ and 0.74, $P < 0.001$). This further confirms that IS rich genomes have IS rich HGT regions, which results both from ISs being horizontally transferred and from these regions being more tolerant to gene disruption by transposition. HGT regions contain ~4 times the density of ISs of the rest of the genome ($P < 0.001$, t-test). This is in accordance with the idea that ISs arise in genomes by HGT. As a confirmation of this scenario, we found that IS families diversity in HGT regions is almost as high as in the entire genome, in spite of them covering only around 10% of it (fig. 5).

The overrepresentation of ISs in the HGT regions can lead to 2 distinct effects. If HGT regions contain a very large fraction of the ISs present in genomes, then the frequency of HGT is a major determinant of IS abundance. On the other hand, if HGT regions have higher density of IS than the rest of the genome but a low fraction of the overall number of IS elements, then the abundance of ISs is determined by their rate of transposition and the ensuing natural selection: HGT regions being more tolerant to transposition, because they provide more neutral insertion targets, will enjoy both effects. We tried to put these 2 hypotheses to test. The median relative frequencies of IS elements among prophage and among specific regions are of 0% and 19%. The IS in HGT-DB have a tendency to be automatically included in the database, and therefore the number of ISs in this set is higher (31% of median). Yet, in the vast majority of genomes, ISs are found much more frequently outside of non-HGT regions (81% of the times). This suggests that the frequency of HGT may be a determinant of the presence of ISs, but not of its abundance. To test this hypothesis, we computed the correlations between IS number and density and HGT number and density (supplementary table 3, Supplementary Material online). In both cases, the correlation is not significant. To verify that these results are not dependent on our HGT data set, we also made these analyses with the 3 data sets separately: HGT-DB, prophagic regions, and strain-specific regions. We found in all cases the same qualitative results (supplementary table 3, Supplementary Material online). Hence, the conclusion that HGT and IS abundance are not correlated is robust to the choice of the data set.

Overall, these results suggest a role for HGT in the spread of IS elements, but they also suggest that the intensity of HGT has a strikingly minor role in determining the abundance of ISs in genomes. In fact, both the number and the density of IS elements fail to correlate significantly with the size and density of transferred regions in genomes, even before controlling for the effect of genome size and in spite of some methodological bias toward classifying ISs as HGT elements. Interestingly, although phages are now regarded as major shuttles of horizontal transfer (Lawrence et al. 2001; Weinbauer 2004), several genomes without IS elements contain large prophagic regions, notably, *B. subtilis*, *L. innocua*, and *Xylella fastidiosa* Temecula1. Several of these also contain plasmids and all have other genes putatively horizontally transferred. Hence, we can also rule out the hypothesis that ISs are only absent from a genome if it is sexually isolated.

The Effect of Pathogenicity

Pathogens are under periodic stress and their population structure shows frequent selective sweeps, which has been suggested to strongly select for genome plasticity and thus favor the presence of ISs in their genomes (Finlay and Falkow 1997; Baquero 2004). Yet, as far as we know the association between IS abundance and pathogenicity has not been put to test. Hence, we classed genomes in 2 classes: pathogens and nonpathogens. Naturally, such a classification is subject to some error, because the distinction between commensalism and pathogenicity is not always straightforward, and the lifestyle of some prokaryotes is not well known. Yet, changes in the classification scheme, for example, by eliminating borderline cases, resulted in the

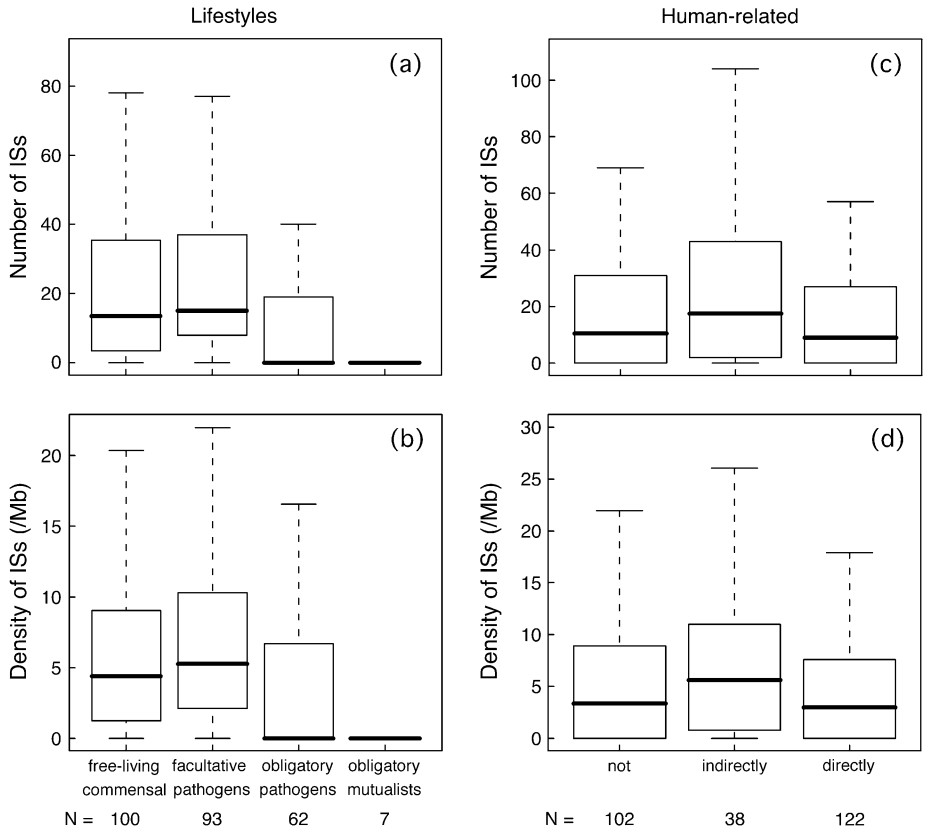

FIG. 6.—Box plot of the number (*a*) and density (*b*) of IS elements among prokaryotes, when these are divided according to lifestyles. IS abundance is significantly lower in obligatory symbionts (pathogens or mutualists) (Wilcoxon test [$P > 0.001$]). Box plot of the number (*c*) and density (*d*) of IS elements among prokaryotes, when these are divided according to their relation to humans. There is no significant difference in none of the Kruskal–Wallis tests ($P > 0.1$). The number of genomes of each group is indicated in bottom of the figure.

same qualitative results (data not shown). Surprisingly, we found no significant difference in the number of ISs between pathogens and nonpathogens (medians 14 and 13.5, respectively, $P > 0.2$, Wilcoxon test). Comparing the density of IS elements with the 2 groups also showed that they are not different (medians 5.2 and 4.5 IS/Mb, $P > 0.15$, Wilcoxon test). Hence, there is no empirical evidence that pathogens have a particularly high frequency of IS elements.

To reach a better understanding of the link between lifestyle and the abundance of ISs, we updated a previous classification of genomes (Rocha 2006) in: free living and commensal, facultative pathogens, obligatory pathogens, and obligatory mutualists. IS abundance is significantly lower in obligatory symbionts (either mutualists or pathogens). We found no IS element in 55% of the genomes of the obligatory pathogens and none among the obligatory mutualist genomes. This should be compared with 17% genomes lacking ISs among free living/commensal and 8% among facultative pathogens (fig. 6*a* and *b*). In fact, free-living and facultative pathogens show no significant difference in the number of IS elements ($P = 0.15$, Wilcoxon test). Altogether, these results indicate no association between the presence of IS and pathogenicity, but a strong association between the frequency of IS elements and the facultative character of the ecological associations (*P* value <0.001, Wilcoxon test).

These results do not necessarily weaken the idea that IS elements contribute to the innate ability of an organism to acquire accessory functions (e.g., drug resistance) by increasing genome plasticity. They just show that this is not specifically associated with pathogenicity. IS-associated genome plasticity could be more useful in prokaryotes oscillating between environments and ecological associations such as facultative pathogens but also free-living and commensal ones. However, these prokaryotes have larger genomes and that could explain the higher IS abundance. To control for genome length, we made a stepwise multiple regression of the appropriate transformation of the number of ISs as a dependent variable (*Y*) and genome size and the facultative character of ecological associations as independent variables (*X*) (see Materials and Methods and table 2). Genome size is the most important variable, but the posterior inclusion of the second variable significantly increased the explained variance from 40% to 47% ($P < 0.001$). This control shows that the type of ecological association is by itself a significant determinant of IS abundance, possibly because obligatory pathogens and symbionts show little selection for genome plasticity. The effect could also be explained by some of these genomes being under sexual isolation, that is, not receiving any genetic information horizontally, thus inaccessible to ISs. Because genomes with at least one functional IS should not be under total sexual isolation, we removed the 64 genomes lacking ISs and the

effect became nonsignificant ($P > 0.2$ when using either IS number or IS abundance as dependent variables, Wilcoxon tests). Hence, when genomes are not under sexual isolation, lifestyle comes out as a nonsignificant determinant of IS abundance in all tested categorizations.

The Effect of Human Sedentarization

Recently, Mira et al. (2006) proposed that the Neolithic revolution had a deep impact in the number of IS elements present in prokaryotic genomes. This proposal was based on 2 observations. Firstly, most very large observable intragenomic IS expansions are recent. However, theoretical population genetics predicts exactly this outcome for selfish elements whose expansion leads to deleterious polymorphisms. A rate of transposition leading to the fixation of dozens of transpositions per century, as is the case in *Yersinia pestis* (Chain et al. 2004) simply cannot last in genomes with ~90% of CDSs. Secondly, data showed that the majority of genomes with many ISs are from specialist prokaryotes associated either with humans or with domesticated animals and plants. Hence, the theory that the Neolithic revolution led to IS expansion in human-related prokaryotes depends heavily on the specific observation that such prokaryotes have more IS elements than prokaryotic generalists. We have previously shown that obligatory pathogens do not significantly have more IS than free-living bacteria. They also do not have more ISs than obligatory pathogens when controlled for genome size and sexual isolation. Hence, it remains to be tested if bacteria interacting with men or domesticated eukaryotes have more ISs.

We classed all genomes into 3 classes. Prokaryotes that are known to interact with man, independently of the duration (pervasive or sporadic) and character (commensals or pathogens) of the interaction were classed as directly human related. Prokaryotes with similar interactions with domesticated animals or plants or deeply involved in the processing of human food such as dairy products, were regarded as indirectly human related. The remaining prokaryotes were regarded as not human related. Naturally, the planet is an integrated ecosystem and every species is directly or indirectly related with every other. However, one does expect marine cyanobacteria or hyperthermophile archaea to be much less affected by the increase in human population size since the Neolithic revolution than a human pathogen. Once this classification was done, it revealed that there is no difference in IS numbers ($P = 0.24$, Kruskal–Wallis test) nor IS density ($P = 0.36$, Kruskal–Wallis test) between the 3 classes (fig. 6*c* and *d*). Hence, there is no evidence that human-related prokaryotes have more IS elements. This of course does not preclude some genomes that recently became human specialists of having large IS expansions because of lower effective population sizes (see Discussion). But it does shows that there is no evidence for a dramatic generic change in the number of IS elements caused by recent expansions in human populations.

Discussion

A reviewer of this article pointed out that restricting the analyses to putatively functional transposases oversimplifies the problem because many small IS vestiges would be missed. One must carry in mind that our goal is to analyze the distribution of IS elements among genomes and what constrains their abundance. As in any functional genomics analysis, it seems more meaningful to consider only elements that are likely to be active. It also seems fairer to test theories in the context where they have been proposed, and all theories regarding IS abundance have been proposed based on the abundance of putatively functional IS elements. Finally, although one does expect the transposition activity of 2 ISs to be larger than that of 1 IS, the same does not hold for vestiges of ISs. Indeed, the presence of IS vestiges may either enhance or inhibit the transposition of the autonomously transposing ISs (Gueguen et al. 2006). It is thus impossible using current knowledge on the role of IS vestiges to put together in a meaningful way the number of IS elements and their vestiges. The absence of published methodology to find the very small vestiges of IS elements that result when fragments are too small to be identified using standard methods of annotation also complicates this analysis. However, the Chandler's group has compiled a detailed list of putative complete and vestigial IS elements in archaeal genomes (personal communication). We have thus confronted our 2 analyses and found an excellent agreement (Spearman's $\rho = 0.94$) between the number of complete ISs that we found and Chandler's fraction of genome occupied by complete ISs (supplementary figure 2, Supplementary Material online). Secondly, the total genomic length occupied by ISs and their vestiges in Chandler's analysis correlates strongly with genome size, as in our analysis (Spearman's $\rho = 0.76$ for Chandler's list and 0.79 for ours). Hence, including IS fragments for all genomes, even if it was possible and meaningful, would most likely not change any of the conclusions of our analysis.

Transposable elements have been considered under many different regards. These range from taking them as the minimal parasite to taking them as cell's genetic engineering tools. After decades of study their exact nature is still debated. Here, we have tried to unravel the reasons why some genomes have so many more elements than others. This is a promising approach because understanding what makes them more or less abundant will certainly shed some light on their intrinsic nature and their impact on genome structure. Our study suggests that HGT is a necessary but not sufficient condition to the presence of ISs. Interestingly, the frequency of HGT is not a determinant of the differential abundance of IS elements in genomes. This suggests that the abundance of ISs is not controlled at the level at which genomes are invaded by transposable elements. Instead, selection resulting from their presence is more likely to determine how many ISs a genome contains. We then tested which of the proposed selective factors was more determinant of IS abundance. We found no significant role in the distribution of IS elements in prokaryotic genomes for variables such as the clade specificity of IS families, the effect of human association nor pathogenicity. Hence, these variables should not be invoked to explain the abundance of IS in a particular genome. For example, pathogenicity has often been said to be associated with higher IS frequencies but shows no explanatory role in our regression analyses. This does not exclude ISs from having adaptive roles in some pathogens.

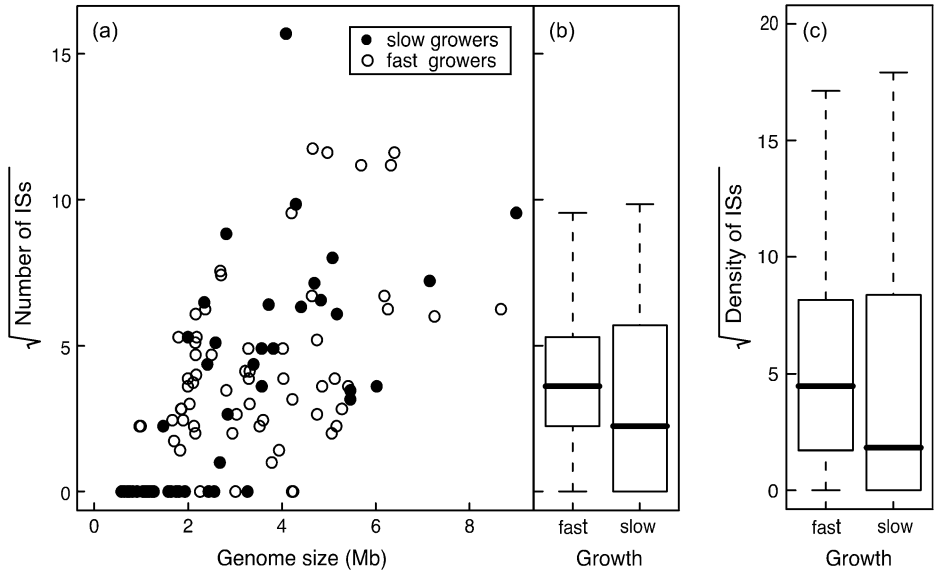

FIG. 7.—The distribution of the number of IS elements in function of genome size when prokaryotes are separated in fast and slow growing (*a*) and box plot of the number (*b*) and density (*c*) of IS elements among fast and slow growers. The number of genomes in fast and slow growers is 64 and 47, respectively, which correspond to the previous published list of growth rates (Rocha 2006). Prokaryotes with a minimum doubling time higher than 2.5 h were classed as slow growers, the others as fast growers. The Spearman's rank association between genome size and optimal doubling time is highly significant (-0.32, $P < 0.001$), as is the association between IS number and density with doubling time (both -0.19, $P < 0.05$).

It just shows that the likeness of this occurring in pathogens is the same as in nonpathogens. It may just be a frame of mind in the genomics discipline that pathogens are different regarding effective population sizes and selection for genomic plasticity from the other bacteria. The facultative character of ecological associations was the only significant determinant of the number of IS elements in genomes apart from genome size. Although this variable is a statistical determinant of IS abundance, the underlying mechanistic causes may not be strictly related with the ecological association but more likely with its consequences in terms of genome size and sexual isolation.

Although rarely considered, genome size explains ~40% of the variance in IS abundance and thus is the most relevant of the tested hypotheses. The smaller the genome, the lower the number but also the lower the density of ISs. There are at least 2 ways of explaining this result. 1) Selection could favor small genomes to allow the optimal use of limited resources especially because gene expression is expensive and larger chromosomes take longer to replicate (e.g., Dufresne et al. 2005). Hence, an increase in genome size caused by ISs could be counterselected. We find this hypothesis unlikely on several grounds. First, many transposases seem to be expressed at very low levels, thus their weight on the cell's energetics is extremely small. In addition, ISs are very small and usually occupy a very small fraction of the genome, thus their contribution to genome size variation is very small. Finally, the implication that selection for efficient replication would suffice to eliminate repeated DNA is severely weakened by the following observations. It has been suggested that there is no correlation between minimal doubling growth time and genome size (Mira et al. 2001). With significantly more data, we now observe a small but significant negative correlation between the 2 (Spearman's $\rho = -0.32$, $P < 0.001$), which indicates that smaller genomes are slower growers. Similar correlations are found when the analysis is restricted to the 2 largest clades: firmicutes ($\rho = -0.27$) and γ-proteobacteria ($\rho = -0.31$). As a consequence, genomes with fewer ISs and smaller IS density, correspond to the slowest growing prokaryotes (fig. 7; $P < 0.05$, Kruskal–Wallis tests). Taken together, these evidences suggest that IS abundance should not correlate with genome size because of selection for fast growth or selection for small chromosomes. 2) It has also been argued that ISs are selected to generate genetic variation, and that such selection should be stronger in larger genomes, which are more complex and are present in a wider range of ecological niches. We are not inclined to favor this hypothesis either, although it is notoriously difficult to test because selection for evolvability is such an elusive concept. It is worth pointing out that this hypothesis is the basis of many of the theories that we have shown in this work not to fit the available data.

We suggest that the association between genome size and IS abundance can be explained more effectively by the change in the density of highly deleterious insertion sites with genome size. Transposable elements have 3 major consequences in cell's fitness (Nuzhdin 1999). Firstly, there is a cost for the expression of transposases, but this should be very small given their low expression levels (Nagy and Chandler 2004). Secondly, transposition directly inactivates genes with high probability because ~90% of prokaryotic genomes are coding for genes and much of the rest contains regulatory sequences. Thirdly, the presence of multiple copies of repeated elements increases genome rearrangements by homologous recombination (Rocha 2003b), and transposition itself may reshape the chromosome (Gray 2000). Could the indirect effects of ISs account for their counterselection in smaller genomes? We have previously failed to find a significant association between

genome size and genome stability (Rocha 2006), which suggests that indirect effects have little role in explaining the dependence of IS abundance on genome size. This does not necessarily weaken the hypothesis that ISs are counterselected because of the rearrangements they promote. It just indicates that such effect, if it exists, is probably not associated with genome size. Could the direct effects of IS transposition explain the association between IS abundance and genome size? The total number of essential genes in the 2 model bacteria *B. subtilis* (Kobayashi et al. 2003) and *E. coli* (Baba et al. 2006) is around 300. To these one must sum more 200–300 genes that are not essential but are nearly ubiquitous among genomes (Koonin 2003; Fang et al. 2005), which suggest that they typically cannot be lost. This is why one rarely finds in nature genomes with much less than 500 genes. These nearly essential genes constitute a large fraction of the smaller genomes. As genomes get larger the fraction of genes whose inactivation leads to a very high fitness cost quickly decreases and the maintenance of the vast majority of genes is under weak selection. As a consequence, transpositions are more tolerated in larger genomes and the number of ISs may increase. Interruption of some parts of the larger genomes may even be positively selected, for example, if ISs inactivate other ISs or prophages. In fact, some ISs have developed the specificity of interrupting other transposable elements (Hallet et al. 1994). In our multiple regressions, we found that the number of IS decreases quicker than linearly with the decrease in genome size, which required a procedure of linearization. This is just what one would expect given that the fraction of genes under strong selection is expected to increase quickly with decreasing genome size. As a result, the abundance of IS elements in the genomes of prokaryotes could largely depend on the density of neutral of slightly deleterious transposition targets. Although prokaryotes and eukaryotes differ significantly in genome structure, the available eukaryote genomes suggest that these results may be extended to them. Indeed, small eukaryote genomes have few or no transposable elements, whereas large genomes contain many.

One should note that even if genome size and the facultative character of ecological associations explain almost half of the variance of IS abundance in genomes, some other hypothesis have been put forward that are impossible to test using available data. These include the variations on genetic deletional bias, intrinsically different IS transposition frequency and effective population sizes. Firstly, mutational events leading to the deletion of genetic information are dominant over the events leading to its amplification (Lawrence et al. 2001; Mira et al. 2001). If this effect is highly variable between genomes, a higher deletion bias could lead to fewer ISs. Unfortunately, the deletion bias has not been quantified in the vast majority of genomes. Secondly, high rates of transposition could lead to higher IS densities. One cannot predict transposition rates in silico. Yet, the observations indicating that transposable elements tend to evolve towards low transposition rates raise some doubts on the explanatory power of this hypothesis (Nuzhdin 1999). Finally, if IS abundance is mostly determined by selection, as our data suggests, different effective population sizes should strongly influence IS abundance. The explosion in the number of ISs in some bacterial pathogens would thus result from recent contractions of effective population sizes not by the fact they are pathogens, nor because they are human related and the human population increased after the Neolithic revolution. Yet, the association of effective population sizes with IS abundance must be a complex one because the smallest (<1 Mb) prokaryotic genomes have the lowest IS densities, as predicted by our analysis but not by its effective population sizes which are thought to be very small. Interestingly, *Prochlorococcus* species have relatively small genomes (<2 Mb) and contain no ISs, as predicted both by our analysis and by their huge population sizes. The seemingly single way of reconciling these results is that even though selection is less efficient for the very small genomes, this is compensated either by sexual isolation in some extreme cases or more commonly by the much higher counterselection of gene inactivation in these already highly reduced genomes.

Supplementary Material

Supplementary tables 1–4 and figures 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Guillaume Achaz, Isabelle Gonçalves, and the reviewers for comments and criticisms on previous versions of this manuscript. M.T. is funded by the Conseil Régional de l'Ile de France.

Literature Cited

Alokam S, Liu SL, Said K, Sanderson KE. 2002. Inversions over the terminus region in *Salmonella* and *Escherichia coli*: IS200s as the sites of homologous recombination inverting the chromosome of *Salmonella enterica* serovar typhi. J Bacteriol. 184:6190–6197.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2:0008.

Baquero F. 2004. From pieces to patterns: evolutionary engineering in bacterial pathogens. Nat Rev Microbiol. 2:510–518.

Bentley SD, Parkhill J. 2004. Comparative genomic structure of prokaryotes. Annu Rev Genet. 38:771–791.

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

Bonnard G, Vincent F, Otten L. 1989. Sequence and distribution of IS866, a novel T region-associated insertion sequence from Agrobacterium tumefaciens. Plasmid. 22:70–81.

Boutoille D, Corvec S, Caroff N, Giraudeau C, Espaze E, Caillon J, Plesiat P, Reynaud A. 2004. Detection of an IS21 insertion sequence in the mexR gene of Pseudomonas aeruginosa increasing beta-lactam resistance. FEMS Microbiol Lett. 230:143–146.

Buchrieser C, Glaser P, Rusniok C, Nedjari H, D'Hauteville H, Kunst F, Sansonetti P, Parsot C. 2000. The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of Shigella flexneri. Mol Microbiol. 38:760–771.

Canchaya C, Fournous G, Brussow H. 2004. The impact of prophages on bacterial chromosomes. Mol Microbiol. 53:9–18.

Capy P, Gasperi G, Biémont C, Bazin C. 2000. Stress and transposable elements: co-evolution of useful parasites? Heredity. 85:101–106.
Casjens S. 1998. The diverse and dynamic structure of bacterial genomes. Annu Rev Genet. 32:339–377.
Casjens S. 2003. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol. 49:277–300.
Chain PS, Carniel E, Larimer FW, et al. 2004. Insights into the evolution of Yersinia pestis through whole-genome comparison with *Yersinia pseudotuberculosis*. Proc Natl Acad Sci USA. 101:13826–13831.
Chao L, Vargas C, Spear BB, Cox EC. 1983. Transposable elements as mutator genes in evolution. Nature. 303:633–635.
Charlesworth B, Charlesworth D. 1983. The population dynamics of transposable elements. Genet Res. 42:1–27.
Condit R, Stewart FM, Levin BR. 1988. The population biology of bacterial transposons: a priori conditions for maintenance as parasitic DNA. Am Nat. 132:129–147.
Cordaux R, Udit S, Batzer MA, Feschotte C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA. 103: 8101–8106.
Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A. 1990. Evidence for horizontal transmission of the P transposable element between *Drosophila* species. Genetics. 124:339–355.
Deng W, Burland V, Plunkett G 3rd, et al. 2002. Genome sequence of *Yersinia pestis* KIM. J Bacteriol. 184:4601–4611.
Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature. 284:601–603.
Draper GC, Gober JW. 2002. Bacterial chromosome segregation. Annu Rev Microbiol. 56:567–597.
Dufresne A, Garczarek L, Partensky F. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol. 6:R14.
Fang G, Rocha EPC, Danchin A. 2005. How essential are nonessential genes? Mol Biol Evol. 22:2147–2156.
Finlay BB, Falkow S. 1997. Common themes in microbial pathogenicity revisited. Microbiol Mol Biol Rev. 61:136–169.
Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 3:722–732.
Garcia-Vallve S, Guzman E, Montero MA, Romeu A. 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. 31:187–189.
Gray YHM. 2000. It takes two transposons to tango. Trends Genet. 16:461–468.
Gueguen E, Rousseau P, Duval-Valentin G, Chandler M. 2006. Truncated forms of IS911 transposase downregulate transposition. Mol Microbiol. 62:1102–1116.
Hall BG. 1999. Transposable elements as activators of cryptic genes in E. coli. Genetica. 107:181–187.
Hallet B, Rezsohazy R, Mahillon J, Delcour J. 1994. IS231A insertion specificity: consensus sequence and DNA bending at the target site. Mol Microbiol. 14:131–139.
Han CG, Shiga Y, Tobe T, Sasakawa C, Ohtsubo E. 2001. Structural and functional characterization of IS679 and IS66-family elements. J Bacteriol. 183:4296–4304.
Hartl DL, Sawyer SA. 1988. Why do unrelated insertion sequences occur together in the genome of Escherichia coli? Genetics. 118:537–541.
Istock CA, Duncan KE, Ferguson N, Zhou X. 1992. Sexuality in a natural population of bacteria-*Bacillus subtilis* challenges the clonal paradigm. Mol Ecol. 1:95–103.
Jin Q, Yuan Z, Xu J, et al. 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. Nucleic Acids Res. 30:4432–4441.
Kidwell MG, Lisch DR. 2001. Transposable elements, parasitic DNA and genome evolution. Evolution. 55:1–24.
Kobayashi K, Ehrlich SD, Albertini A, et al. 2003. Essential *Bacillus subtilis* genes. Proc Natl Acad Sci USA. 100:4678–4683.
Kolsto A-B. 1997. Dynamic bacterial genome organization. Mol Microbiol. 24:241–248.
Koonin EV. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol. 1:127–136.
Kothapalli S, Nair S, Alokam S, Pang T, Khakhria R, Woodward D, Johnson W, Stocker BA, Sanderson KE, Liu SL. 2005. Diversity of genome structure in *Salmonella enterica* serovar Typhi populations. J Bacteriol. 187:2638–2650.
Kunst F, Ogasawara N, Moszer I, et al. 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature. 390:249–256.
Kunze ZM, Wall S, Appelberg R, Silva MT, Portaels F, McFadden JJ. 1991. IS901, a new member of a widespread class of atypical insertion sequences, is associated with pathogenicity in Mycobacterium avium. Mol Microbiol. 5:2265–2272.
Lander ES, Linton LM, Birren B, et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.
Lawrence JG, Hendrix RW, Casjens S. 2001. Where are the pseudogenes in bacterial genomes? Trends Microbiol. 9: 535–540.
Lawrence JG, Ochman H, Hartl DL. 1992. The evolution of insertion sequences within enteric bacteria. Genetics. 131: 9–20.
Lichter A, Manulis S, Valinsky L, Karniol B, Barash I. 1996. IS1327, a new insertion-like element in the pathogenicity-associated plasmid of Erwinia herbicola pv. gypsophilae. Mol Plant Microbe Interact. 9:98–104.
Machida C, Machida Y. 1989. Regulation of IS1 transposition by the insA gene product. J Mol Biol. 208:567–574.
Machida Y, Sakurai M, Kiyokawa S, Ubasawa A, Suzuki Y, Ikeda JE. 1984. Nucleotide sequence of the insertion sequence found in the T-DNA region of mutant Ti plasmid pTiA66 and distribution of its homologues in octopine Ti plasmid. Proc Natl Acad Sci USA. 81:7495–7499.
Mahillon J, Chandler M. 1998. Insertion sequences. Microbiol Mol Biol Rev. 62:725–774.
Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589–596.
Mira A, Pushker R, Rodriguez-Valera F. 2006. The Neolithic revolution of bacterial genomes. Trends Microbiol. 14:200–206.
Moran NA, Plague GR. 2004. Genomic changes following host restriction in bacteria. Curr Opin Genet Dev. 14:627–633.
Moszer I, Rocha EPC, Danchin A. 1999. Codon usage and lateral gene transfer in *Bacillus subtilis*. Curr Opin Microbiol. 2:524–528.
Naas T, Blot M, Fitch WM, Arber W. 1995. Dynamics of IS-related genetic rearrangements in resting *Escherichia coli* K-12. Mol Biol Evol. 12:198–207.
Nagai T, Phan Tran LS, Inatsu Y, Itoh Y. 2000. A new IS4 family insertion sequence, IS4Bsu1, responsible for genetic instability of poly-gamma-glutamic acid production in Bacillus subtilis. J Bacteriol. 182:2387–2392.
Nagy Z, Chandler M. 2004. Regulation of transposition in bacteria. Res Microbiol. 155:387–398.
Nuzhdin SV. 1999. Sure facts, speculations, and open questions about the evolution of transposable element copy number. Genetica. 107:129–137.
Orgel LE, Crick FHC. 1980. Selfish DNA: the ultimate parasite. Nature. 284:604–607.

Papadopoulos D, Schneider D, Meier-Eiss J, Arber W, Lenski RE, Blot M. 1999. Genomic evolution during a 10,000-generation experiment with bacteria. Proc Natl Acad Sci USA. 96:3807–3812.

Parkhill J, Sebaihia M, Preston A, et al. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet. 35:32–40.

Parkhill J, Wren BW, Thomson NR, et al. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature. 413:523–527.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol Biol Evol. 20:880–892.

Redder P, Garrett RA. 2006. Mutations and rearrangements in the genome of Sulfolobus solfataricus P2. J Bacteriol. 188:4198–4206.

Rocha EPC. 2003a. An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. Genome Res. 13:1123–1132.

Rocha EPC. 2003b. DNA repeats lead to the accelerated loss of gene order in Bacteria. Trends Genet. 19:600–604.

Rocha EPC. 2004. Order and disorder in bacterial genomes. Curr Opin Microbiol. 7:519–527.

Rocha EPC. 2006. Inference and analysis of the relative stability of bacterial chromosomes. Mol Biol Evol. 23:513–522.

Schmid-Appert M, Zoller K, Traber H, Vuilleumier S, Leisinger T. 1997. Association of newly discovered IS elements with the dichloromethane utilization genes of methylotrophic bacteria. Microbiology. 143:2557–2567.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18: 502–504.

Schneider D, Lenski RE. 2004. Dynamics of insertion sequence elements during experimental evolution of bacteria. Res Microbiol. 155:319–327.

Shapiro JA. 1999a. Genome system architecture and natural genetic engineering in evolution. Ann N Y Acad Sci. 870:23–35.

Shapiro JA. 1999b. Transposable elements as the key to a 21st century view of evolution. Genetica. 107:171–179.

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34:D32–D36.

Syvanen M. 1998. Insertion sequences and their evolutionary role. In: Weinstock GM, editor. Bacterial genomes. New York: Chapman & Hall. p. 213–220.

Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays. 28:913–922.

Wagner A. 2006. Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. Mol Biol Evol. 23:723–733.

Wei J, Goldberg MB, Burland V, et al. 2003. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. Infect Immun. 71:2775–2786.

Weinbauer MG. 2004. Ecology of prokaryotic viruses. FEMS Microbiol Rev. 28:127–181.

Yang F, Yang J, Zhang X, et al. 2005. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res. 33:6445–6458.

Zerbib D, Prentki P, Gamas P, Freund E, Galas DJ, Chandler M. 1990. Functional organization of the ends of IS1: specific binding site for an IS 1-encoded protein. Mol Microbiol. 4:1477–1486.

Jennifer Wernegreen, Associate Editor

Accepted January 18, 2007