5284–5293 Nucleic Acids Research, 2007, Vol. 35, No. 16 *Published online 7 August 2007*
doi:10.1093/nar/gkm597

A survey of bacterial insertion sequences using IScan

Andreas Wagner[1,*], Christopher Lewis[2] and Manuel Bichsel[1]

[1]Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, 27-J-54, CH-8057 Zurich, Switzerland and [2]Department of Biology, MSC03 2020,1 University of New Mexico, Albuquerque, NM, 87131-0001

Received May 25, 2007; Revised and Accepted July 19, 2007

ABSTRACT

Bacterial insertion sequences (ISs) are the simplest kinds of bacterial mobile DNA. Evolutionary studies need consistent IS annotation across many different genomes. We have developed an open-source software package, IScan, to identify bacterial ISs and their sequence elements—inverted and target direct repeats—in multiple genomes using multiple flexible search parameters. We applied IScan to 438 completely sequenced bacterial genomes and 20 IS families. The resulting data show that ISs within a genome are extremely similar, with a mean synonymous divergence of $K_s = 0.033$. Our analysis substantially extends previously available information, and suggests that most ISs have entered bacterial genomes recently. By implication, their population persistence may depend on horizontal transfer. We also used IScan's ability to analyze the statistical significance of sequence similarity among many IS inverted repeats. Although the inverted repeats of insertion sequences are evolutionarily highly flexible parts of ISs, we show that this ability can be used to enrich a dataset for ISs that are likely to be functional. Applied to the thousands of genomes that will soon be available, IScan could be used for many purposes, such as mapping the evolutionary history and horizontal transfer patterns of different ISs.

INTRODUCTION

Transposable elements occur in many bacterial genomes. We can thus not fully understand bacterial genome evolution, unless we understand how such mobile DNA is maintained, and how it spreads among bacterial genomes. Because transposable elements also cause an important public health threat, the spreading of drug-resistance genes among pathogenic bacteria, such understanding may ultimately also shed light on the epidemiology of drug-resistant pathogens.

Insertion sequences (ISs) are among the simplest kinds of bacterial mobile DNA. They range in size from 600 to more than 3000 bp and fall into 20 major families (1–3). Most ISs consist of short inverted repeat sequences that flank one or more open reading frames (ORFs), whose products encode the transposase proteins necessary for their transposition. Some but not all ISs transpose into specific target sites. ISs typically generate a direct repeat of their target site after transposition. Transposition is often associated with an increase in IS copy number in a genome. In eukaryotes many non-functional copies of transposable elements can often be passively proliferated using transposase from functional copies (4–6). In contrast, bacterial IS transposition is often tightly regulated, occurs at a very low level, and is often restricted to *cis* activity, where a transposase promotes only the transposition of the IS from which it is expressed (7). Exceptions exist, for example in the form of very short miniature inverted repeat elements that may only proliferate passively (8–10).

In the near future a flood of bacterial genome data will become available. Such data will see many uses in studying IS families, including the identification of functionally important sequences from hundreds of family members, and the reconstruction of the evolutionary history of individual ISs. Efforts like these require a comparison of ISs across (many) different genomes. Such a comparison is hindered by existing IS annotations which may differ greatly among genomes, because they have been produced by different research groups using different tools. In addition, existing annotations provide limited information about sequence elements such as inverted repeats, or about the structure of ISs where the transposase is encoded by more than one open reading frame. With these limitations in mind, we have developed IScan, a software tool that allows a user to identify ISs and their associated direct and inverted repeats automatically, flexibly and in multiple genomes, using a curated reference IS from a database such as ISfinder (11). The consistent annotation provided by IScan will greatly aid evolutionary studies.

In two analyses that address two different classes of questions, we applied IScan to 438 completely sequenced bacterial genomes and all 20 major IS families. The first

*To whom correspondence should be addressed. Tel: +41 446356141; Fax: +41 44635 6144; Email: aw@bioc.uzh.ch set of analyses addresses the biological question: Why is mobile DNA maintained in bacterial genomes? Mobile DNA might be a very effective parasite, a prototypical example of selfish DNA (12,13), or it might confer benefits to its host. [For example, mobile DNA can mobilize genes for transfer between bacterial strains or species (14)]. Despite its long history, this question has not been completely resolved.

To find out whether mobile DNA persists because it benefits a host, one needs to understand the dynamics of mobile DNA on evolutionary time scales. Laboratory evolution experiments (15–21) are of limited use here. The reason is that the rates at which ISs transpose, are transferred horizontally, and can cause recombinational and other instabilities are so small (22,23) that even long laboratory evolution experiments may detect IS copy number and position variation, but may not be sufficient to determine whether ISs have net deleterious or beneficial effects.

A different approach to understanding the evolutionary dynamics of ISs focuses on the number and distribution of ISs in bacterial populations or closely related bacterial strains (20,24–28). Most pertinent studies were carried out before large-scale genome sequence data became available, and are thus very limited. In a recent paper, we overcame some of the limitations of pre-genome work by analyzing the distribution of five major IS families in 202 complete genomes (29). This analysis suggested that ISs within a genome have very low nucleotide diversity, cause their host to go extinct on evolutionary time scales, and can only be sustained by horizontal transfer. In other words, ISs are likely to be detrimental to their host in the long run. However, this earlier analysis was also hampered by our reliance on available genome sequence annotations to identify ISs. We here overcome this limitation by our use of IScan to study the distribution and sequence similarity of ISs in more than twice as many genomes and four times as many IS families than in earlier work.

The second of our two applications of IScan addresses a methodological rather than a biological question: Is it possible to distinguish functional from non-functional (especially truncated) ISs computationally—without time-consuming experiments—and for hundreds or thousands of ISs? We suggest an approach based on the similarity of IS inverted repeats. IScan is ideal for this approach, because it can calculate various statistical significance measures for inverted repeat similarity. We show that our approach, while certainly not allowing for perfect discrimination, may enrich a dataset for ISs that are likely to be functional.

METHODS

Details on information produced by IScan

Via the procedure outlined in the Results, IScan produces a file in FastA format. This file contains the following parts for each IS:

IS section. The section contains a unique identifier, the accession number and version of the sequence in which the IS was found, start and end coordinates on the queried DNA molecule, length, *P*-value of the inverted repeat alignments and nucleotide sequence.

ORF section. The section contains one entry for each ORF query (and, therefore, BLAST hit), start and end coordinates on the queried DNA molecule, length, strand and nucleotide sequence of the BLAST hit.

BLAST hit section. One entry for each ORF query (and, therefore, BLAST hit), containing the start and end coordinates of the BLAST hit on the queried DNA molecule, the number and proportion of amino acid identities, the expect (E) value and the amino acid sequence of the BLAST hit.

Inverted repeat section. One entry for each of the upstream and downstream inverted repeats, containing the start and end coordinates on the queried DNA molecule; repeat unit length; alignment score; number of matches, mismatches and gaps; *P*-value of the inverted repeat alignment; and nucleotide alignment for the inverted repeat.

Direct repeat section. One entry for each of the upstream and downstream direct repeats, containing the start and end coordinates on the queried DNA molecule; repeat unit length; alignment score; number of matches, mismatches and gaps; and nucleotide alignment for the direct repeat.

A survey of ISs using IScan

We used IScan to search for ISs belonging to the 20 major IS families listed in Table 1 (1–3) in 438 curated bacterial genomes (consisting of 790 sequenced DNA molecules)

Table 1. Number of ISs in different families found for the analysis carried out here

Family	Reference IS	Number of ISs
IS1	IS1A	863
IS481	IS481	259
IS3	IS2	242
IS5	IS5	239
IS4	IS4	171
IS110	IS110	88
IS982	IS982	57
IS630	IS630	55
IS256	IS256	29
IS21	IS21	25
IS91	IS91	19
Tn3	IS1071	18
IS30	IS30	13
ISL3	ISL3	7
IS66	ISRm14	3
ISCR	ISCR1	2
IS6	IS15	1
ISAs1	ISAs1	0
IS1380	IS380A	0
IS605	IS605	0
TOTAL		2091

According to [(1), Table 2, (2)], the IS6 family has one ORF, but the curated reference sequence IS15 (11) has two ORFs.

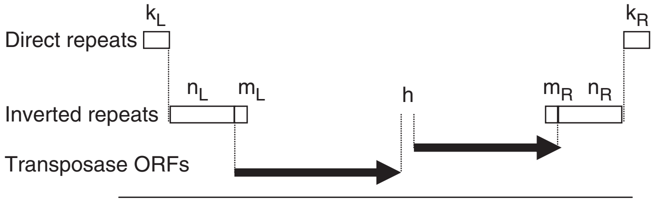

Figure 1. Illustration of the different parameters IScan uses to identify ISs. The thin horizontal line at the bottom of the panel indicates the queried DNA sequence (genome). Thick arrows indicate matches to IS ORFs. Open bars indicate inverted repeats (middle) and direct target repeats (upper). See main text for explanation of parameters.

available from GenBank (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). The curated query ISs we used were obtained from the IS repository IS Finder (http://www-is.biotoul.fr 11). We retained BLAST hits to IS ORFs with an E-value of $E \leqslant 1$ and at least 35% amino acid identity to the query sequence. For ISs with more than one ORF, we used the parameter $h = 50$ (Figure 1) to assign ORFs to the same IS. For identification of target direct repeats we used curated data on the length of direct repeats from ISfinder [(2), Table 1, (11)] to define k_R and k_L for each IS family analyzed. To identify inverted and direct repeats we set $m_L = m_R = 0$, and $n_L = n_R = 1.1 \times$ (total IS length – length of IS coding region), where the total and coding region lengths are again derived from information curated for each IS family's reference sequence (11). Inverted and direct repeat alignments were performed with the same scoring matrix of 1 for matches, −2 for mismatches and −5 for gaps and gap extensions.

Analysis of alignment scores for inverted repeat *P*-values

We determined various P-values (P_{LR}, P_3, P_L, P_R, see Results) that indicate whether the candidate inverted repeats of an IS are statistically significantly similar. We did so by aligning 10^5 randomly chosen sequence fragment pairs of length $n_L + m_L + 1 = n_R + m_R + 1$ from the DNA molecule in which the IS was found. Specifically, for P_{LR}, two randomly chosen fragments are aligned against each other, for P_3, two randomly chosen fragments are aligned against the left reference inverted repeat and for P_L (P_R), one randomly chosen fragment is aligned against the left (right) reference inverted repeat. The fraction of these alignments whose score is greater (indicating greater similarity) than the alignment score of the candidate inverted repeat corresponds to the desired P-value. For P_{LR} we used Smith–Waterman local alignment, for P_3, P_L and P_R we used clustalw, which implements a global dynamic programming alignment algorithm (30).

Determination of K_a and K_s for pairs of ISs within the same family

To estimate synonymous and non-synonymous divergence among IS coding regions, we used a previously published tool (31). Briefly, the tool uses information from both the DNA and amino acid sequences, and proceeds in three steps. First, it pre-screens related gene pairs using BLASTP (32) and the Needleman and Wunsch dynamic programming alignment algorithm [Thompson *et al.* (30)]. Then, it eliminates gene pairs with fewer than 50 alignable amino acid residues and with <50% amino acid identity from further analysis. In the third step, the tool calculates the number of substitutions per synonymous site (K_s) and the number of substitutions per non-synonymous site (K_a) using the maximum likelihood models of Muse and Gaut (33) and Goldman and Yang (34) for the remaining pairs. It uses a simple heuristic test (35) to determine whether a gene pair has been saturated with synonymous substitutions.

For ISs with overlapping ORFs, we merged, for reasons of computational tractability, the overlapping ORFs into one ORF for the calculation of K_a and K_s. (The short overlapping regions are subject to different evolutionary constraints than the non-overlapping regions). Specifically, we calculated the number of nucleotides that overlap in the two ORFs, and eliminated from a sequence containing both ORFs the segment containing the overlap and any additional nucleotides upstream or downstream of the overlapping segment required to retain the reading frames of the two ORFs. On average, IS ORFs were shortened by four nucleotides through this procedure.

RESULTS

IScan, a tool to identify ISs

IScan identifies transposase sequences, inverted repeats and candidate target direct repeats of ISs in complete genomes. IScan is a free open source package developed on a Linux platform and implemented in perl. It is available from the website: http://www.bioc.uzh.ch/wagner/publications-software.html. IScan uses a curated reference or query IS (which, in our case, is a representative member of a major IS family; Table 1) to identify other ISs in one or more completely sequenced genomes, or any other DNA molecules. This query sequence contains (i) the amino acid sequences encoded by one or more transposase ORFs, and (ii) the nucleotide sequence of the upstream (IR_L) and downstream inverted repeat (IR_R). We note that ISs with two or more transposase ORFs frequently express a single functional transposase through a translational frameshifting mechanism. The extent and length of required sequence similarity to the reference IS are user-specifiable, such that arbitrarily weakly similar ISs or short IS fragments can be identified if needed. IScan identifies ISs in three major steps.

(i) Identification of transposase ORFs. IScan identifies the ORFs of an IS through a tblastn search [using WUBLAST, (32), http://blast.wustl.edu/] which matches the query amino acid sequence(s) to the translation products of the genomic sequence in all six possible reading frames. For ISs consisting of more than one ORF, hits to different ORFs of the query are identified as belonging to the same IS if the ORFs fall within a user-specified distance of each other (Figure 1).

(ii) Identification of (candidate) inverted repeats. IScan applies a user-specifiable alignment algorithm, such as the Miller–Myers version (36) of the Smith–Waterman local alignment algorithm (37) to (a) a window of DNA comprising n_L nucleotides upstream of the upstream-most ORF's start, and m_L nucleotides downstream of the upstream-most ORF's start (thus comprising a total of $n_L + m_L + 1$ nucleotides; Figure 1), and (b) the reverse complement of a window of DNA m_R nucleotides upstream of the downstream-most ORF's end, and n_R nucleotides downstream of the downstream-most ORF's end. If the IS has only one ORF, the upstream-most and downstream-most ORFs are the same ORF. The local alignment of the upstream and downstream windows is used to identify the candidate inverted repeats of this IS. The parameters n_L, m_L, n_R and m_R are user-specifiable parameters, as are the match, mismatch and gap penalties used by the alignment algorithm.

(iii) Identification of (candidate) direct repeats. Many ISs generate short direct repeats upon transposition into a target site. IScan first identifies a window of k_L nucleotides upstream of the upstream inverted repeat, and a window of k_R nucleotides downstream of the downstream inverted repeat (as identified in step 2; Figure 1). Alignment of these two windows then yields candidate direct repeats.

A patchy distribution of ISs among bacterial genomes

We applied IScan to the complete DNA sequences (chromosomes and plasmids) of 438 bacterial genomes, to identify all candidate ISs in the 20 major IS families whose ORFs had at least 35% amino acid sequence identity to a family prototype sequence (Table 1) over the length of the prototype sequence. This approach yielded 2091 ISs. 95% (1987) of them occurred on bacterial chromosomes, and the remainder occurred on plasmids. The length distribution of the ISs we identified is shown in Figure 2a. In the literature, the lower end of the typical length range for all IS families considered is 540 bp (1,11). Only 1.2% of the ISs we identified were shorter than 540 bp. The conspicuous peaks in the length distribution come from individual highly abundant ISs, such as IS1, with a highly abundant member of length 695 bp that causes the highest peak in Figure 2a.

Among all ISs we identified, the most abundant IS families are IS1, IS3, IS481 and IS5 (Table 1). The distribution of any one IS family is extremely patchy and highly skewed: The vast majority of genomes contains no member of the family; most genomes that contain one member of the family contain only one member; and typically only very few genomes contain a large number of members. We illustrate this distribution in Figure 2b (note the logarithmic scale on the y-axis). The numbers of IS copies within a genome show no strong statistical associations among different IS families. Specifically, the copy numbers of only 10 among 136 possible IS family pairs, show a statistically significant (Bonferroni-corrected $P = 0.05$) Spearman rank order correlation coefficient. All but three of these statistically significant associations vanish, however, if one eliminates genomes from the analysis in which one or both ISs have no copies. The remaining significant associations involve IS1 and IS3 (Spearman's $r = 0.93$; $n = 16$), IS110 and IS4 ($r = 0.87$; $n = 13$), as well as IS110 and IS3 (Spearman's $r = 0.83$; $n = 15$). However, the genomes that account for this correlation are from the extremely closely related species of *Escherichia coli* and *Shigella*. This indicates that a common evolutionary history rather than similar host preferences is responsible for the co-occurrence of these ISs. It also illustrates that the shared evolutionary history of many sequenced genomes introduces a bias into the data that has to be taken into account when testing certain hypotheses about IS evolution.

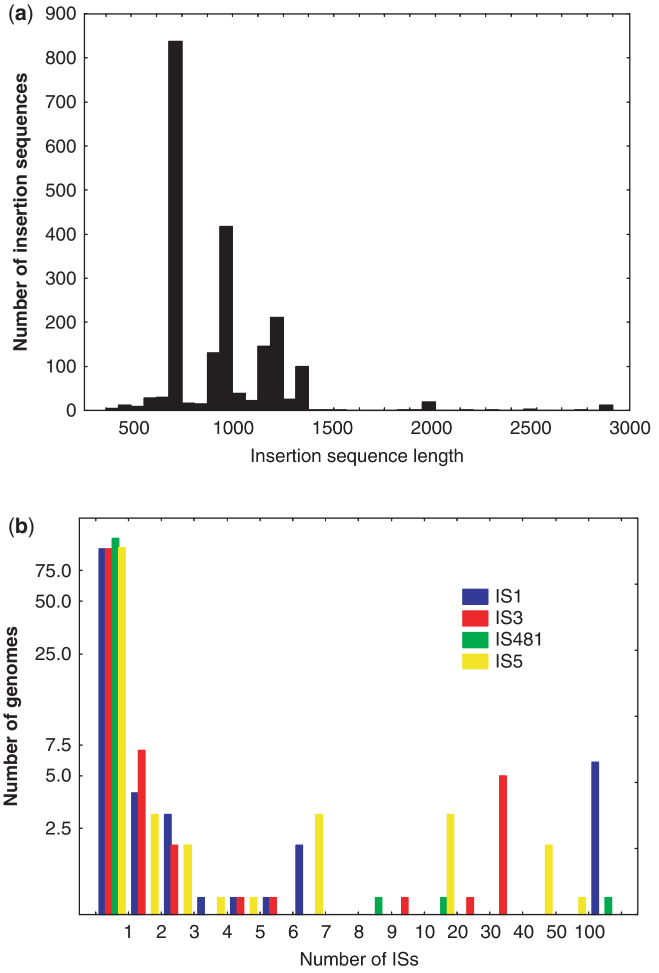

Figure 2. (**a**) Length distribution of ISs with more than 35% amino acid sequence identity relative to a curated reference sequence. (**b**) Distribution of the number of IS copies per genome for four abundant IS families studied here. Note the logarithmic scale on the vertical axis, and the skewed distribution.

High similarity of ISs within a genome

We next examined the sequence divergence of ISs within and among genomes. Beyond simple nucleotide divergence (Figure 3a), we determined K_a, the fraction of amino acid replacement substitutions at amino acid replacement sites, and K_s, the fraction of synonymous substitutions at synonymous sites, an indicator of the synonymous divergence within an IS's transposase-coding genes. Synonymous sites are generally under weaker selection than amino acid replacement sites. In addition, because of the low expression level of transposases, the synonymous sites we study are not subject to selection for translation efficiency. These two observations render K_s a better (albeit crude) indicator of the IS's age than K_a (38). To estimate K_a and K_s, we used GenomeHistory, a software tool that estimates K_a and K_s using a maximum likelihood method (31).

We first focused on the sequence divergence of ISs within a family and within genomes. It is very low. Figure 3b shows a histogram of the distribution of amino acid sequence divergence K_a for pairs of ISs within the same genome. Note the logarithmic scale on the y-axis, indicating a very large number of sequences at low divergence. The mean (median) K_a is 0.012 (0.0019), and its 90th percentile is $K_a = 0.0067$, even though our approach would readily detect ISs with amino acid sequence divergence up to K_a greater than 1.33.3% of sequence pairs within a genome are completely identical in their amino acid sequence. Synonymous divergence K_s is similarly low (Figure 3c). Excluding IS pairs with saturated synonymous divergence (1.01% of all pairs), the mean synonymous divergence is $K_s = 0.033$ with its 90th percentile at $K_s = 0.013$. More than 60% of all IS pairs within a genome are identical at their synonymous sites, such that the median $K_s = 0$.

Figure 4 shows the mean and SEs of K_s and K_a separately for each IS family (see also Table 2). While amino acid divergence is relatively homogeneous among families, synonymous divergence K_s varies to a greater extent, and it is particularly high for IS5. Most variable is the ratio K_a/K_s, which is normally taken as an indicator of selective constraint on a protein. The mean ratio is $K_a/K_s = 0.39$ and it varies between $K_a/K_s = 0.13$ (IS66) and $K_a/K_s = 0.86$ (IS4). A high ratio K_a/K_s might be taken to indicate the presence of many inactive ISs whose coding region are pseudogenes and thus evolve effectively neutrally ($K_a = K_s$). However, this interpretation would be misleading, because ISs with high K_a/K_s generally show extremely low intragenomic synonymous and non-synonymous divergence. For example, for IS4, where the mean $K_a/K_s = 0.86$, the mean values of K_a and K_s are an extremely low $K_a = 6.3 \times 10^{-3}$ and $K_s = 7 \times 10^{-3}$. Similarly low divergences also hold for other ISs with high K_a/K_s ratios (Table 2). Such small values mean that virtually all IS pairs of a family within a genome differ by one or very few nucleotides. At such low divergence, the interpretation of the ratio K_a/K_s as an indicator of selective constraint is not appropriate, because there may be a large amount of stochastic variation in the number of synonymous and non-synonymous nucleotide changes. In this regard, it is also worth mentioning that IS5, which is an extreme outlier with its high mean synonymous divergence of $K_s = 0.58$ shows a $K_a/K_s = 0.22$, close to the lower end of the observed range across families.

Not unexpectedly, the sequence divergence among pairs of ISs in different genomes is substantially greater. At a mean $K_s = 0.3$ synonymous divergence of ISs among

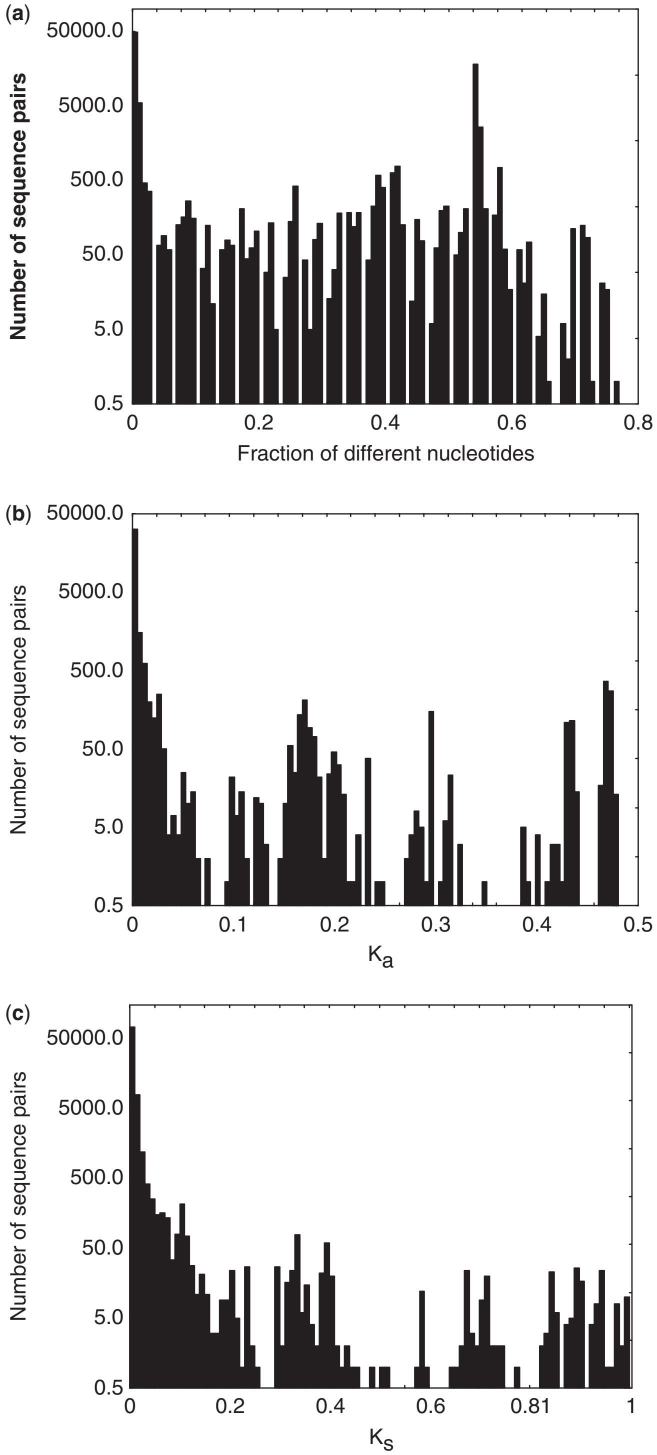

Figure 3. Within-genome distribution of (**a**) overall nucleotide divergence (in fraction of pairwise nucleotide differences), (**b**) non-synonymous divergence per non-synonymous site K_a (38), (**c**) synonymous divergence at synonymous sites K_s (38), of IS pairs in the same family, for all ISs where there exist genomes with more than one IS of the same family. Note the logarithmic scale and the peak at identical ISs, which shows that many ISs from the same family within a genome are identical to each other. Here, 1.42% of IS pairs in a genome have a $K_s > 1$ and are not shown in (b).

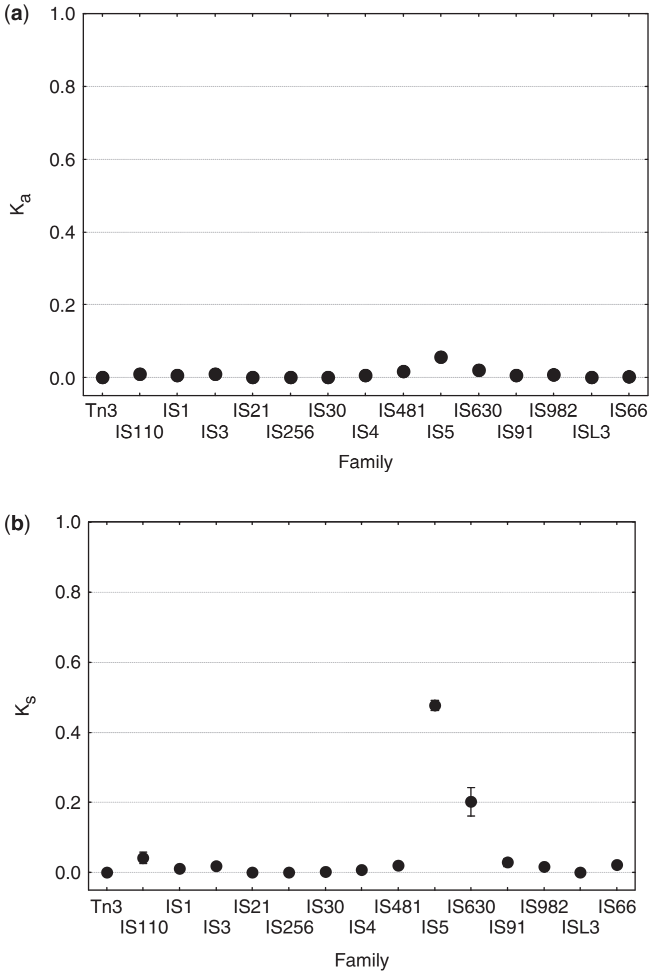

Figure 4. Means and SEs of (**a**) within-genome K_a and (**b**) within-genome K_s for those IS families where more than one family member occurred in at least one genome.

genomes is almost 10 times greater than within genomes. Ten times more IS pairs in different genomes have saturated synonymous divergence (10.5% as opposed to 1.01% within genomes). The mean non-synonymous divergence $K_a = 0.064$ (90th percentile $K_a = 0.29$) is a factor five higher amongst genomes than within genomes. It is nonetheless very low compared to the maximum $K_a = 1$ our approach could have revealed. This suggests that the IS families we study are well defined on the sequence level.

The presence of closely related ISs of the same family in different genomes indicates the importance of horizontal gene transfer in their spreading. For instance, we find a large number (1108) of ISs with identical transposase coding regions in different organisms. Many of these pairs stem from species closely related in evolutionary history or life style, such as ISs in the closely associated genera *Escherichia*/*Shigella*/*Salmonella* and *Staphylococcus*/*Enterococcus*. However, some of these pairs involve more distantly related organisms, such as the psychrophilic (cold-loving) arctic bacterium *Desulfotalea psychrophila* on one hand, and the human *E. coli* and *Haemophilus ducreyi* on the other hand, which all share identical IS1 elements.

Table 2. Mean and SEs for K_a, K_s and K_a/K_s within genomes for those IS families where there exist genomes that contain more than one IS of the same family

IS Family	Mean K_a	SE	Mean K_s	SE	Mean K_a/K_s	SE
IS1	0.0057	0.00015	0.011	0.00031	0.396	0.0029
IS4	0.0063	0.00026	0.007	0.00035	0.86	0.02
ISL3	0	0	0	0	NA	NA
IS5	0.043	0.0013	0.48	0.015	0.22	0.0048
IS110	0.0039	0.003	0.043	0.016	0.7	0.15
IS21	0.00076	0.0002	0.00053	0.00031	0.41	0.15
IS3	0.0098	0.00084	0.018	0.0011	0.41	0.0087
IS256	0	0	0	0	NA	NA
IS91	0.0065	0.0012	0.029	0.011	0.57	0.078
IS30	0.0006	0.00023	0.0032	0.0012	0.19	0
IS630	0.021	0.0041	0.2	0.041	0.17	0.019
IS481	0.0057	0.00025	0.021	0.0013	0.33	0.009
IS982	0.0078	0.00011	0.016	0.00039	0.62	0.017
IS66	0.003	NA	0.022	NA	0.13	NA
Tn3	0.00057	0.00044	0	0	NA	NA
All Families	0.012	0.0002	0.033	0.00081	0.39	0.003

NA (not applicable) indicates insufficient data.

Inverted repeats are flexible in sequence but provide a signal to enrich for functional ISs

Very few among the many known ISs have been subject to experimental tests for their ability to transpose. For evolutionary studies, it is useful to distinguish such functional from non-functional ISs. Given the flood of new ISs that bacterial genome sequencing is producing, time-consuming experimental approaches are not suitable to make this distinction. We thus suggest a computational strategy that may enrich for non-truncated and likely functional ISs.

In principle, two strategies are conceivable to identify putatively functional ISs computationally. The first takes advantage of coding sequence similarity of a candidate IS to a reference IS. The limitation of this strategy is that functional transposases may be very divergent from any one reference sequence. An alternative strategy focuses on the second major sequence feature of ISs, their inverted repeats. Inverted repeats with significant sequence similarity may indicate functionality of an IS, or at least that an IS has not been truncated.

We used several different approaches to estimate the 'quality' of an IS's inverted repeats, as indicated by their similarity to each other and to a reference sequence. (These approaches are also implemented in IScan and are available to IScan users.) The first approach involves a local dynamic programming alignment (39) of the sequences immediately upstream and downstream of the coding sequences that contain the inverted repeats. We compared the score of this alignment with that of a large number (10^5) of alignments of sequence fragment pairs of the same length but chosen from random positions within the same DNA molecule. This allows us to assign a significance threshold P_{LR} ('left-right', Figure 5) of observing an alignment score as high as that observed

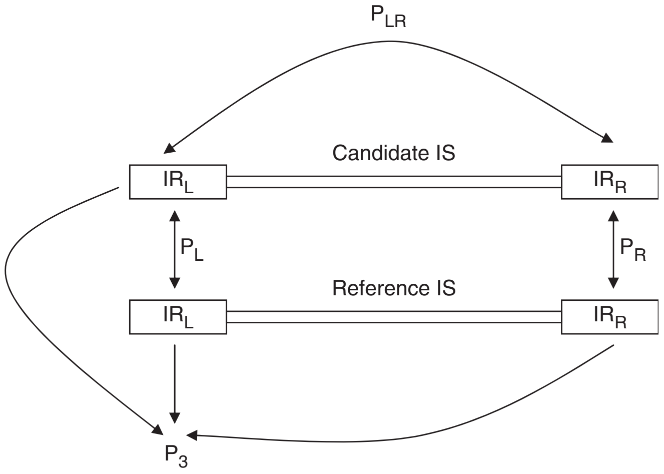

Figure 5. Illustration of the different alignment strategies pursued to assess statistical significance of inverted repeat alignments. The candidate IS is the sequence match produced by IScan to a reference IS from a given family. See text for details.

between putative IR_L and IR_R by chance alone. Figure 6a shows a histogram of the distribution of these *P*-values together with an indication of the significance threshold $P = 0.05$, as well as the significance threshold $P = 0.05/2091 = 2.3 \times 10^{-5}$. This lower value corresponds to a Bonferroni-corrected $P = 0.05$. It takes into account that we carry out multiple independent tests, but it is excessively conservative. Although a large fraction (55%) of the *P*-value we determined is significant at $P = 0.05$, not one of them is smaller than $P = 2.3 \times 10^{-5}$.

In the second approach, we aligned the left inverted repeat of the 'reference' sequence we used to identify ISs of a given family with both the left and right (candidate) inverted repeats of the ISs we identified (Figure 5). For each IS, we evaluated the statistical significance P_3 ('3-way') of each alignment score by the same randomization approach as above, using a large number (10^5) of random DNA fragment pairs from the same molecule, and aligning them to the left inverted repeat of the reference sequence. In this analysis, 79.4% (1660) of ISs showed $P_3 < 0.05$, and two showed $P_3 < 2.3 \times 10^{-5}$ (Figure 6b).

In a third analysis, we aligned the putative IR_L of each IS candidate with the IR_L of the reference IS that we used to identify the IS in the first place. We then carried out the same randomization approach as above to identify the likelihood P_L ('left') to observe such an alignment by chance alone. We repeated this approach for the putative IR_R to obtain P_R ('right'). To obtain a joint significance score for both IR_L and IR_R of an IS, we simply calculated the product $P_L \times P_R$. Figure 6c shows a histogram of ($P_L \times P_R$). In this analysis, 90.3% (1890) of $P_L \times P_R$ -values are smaller than 0.05, and 55.1% (1152) values are smaller than the Bonferroni-corrected $P = 2.3 \times 10^{-5}$. Thus, this alignment strategy significantly enriches for ISs with highly similar inverted repeat units. To be sure, this does not demonstrate that ISs with highly similar repeat units are more likely to be functional. The following analysis suggests, however, that this is the case.

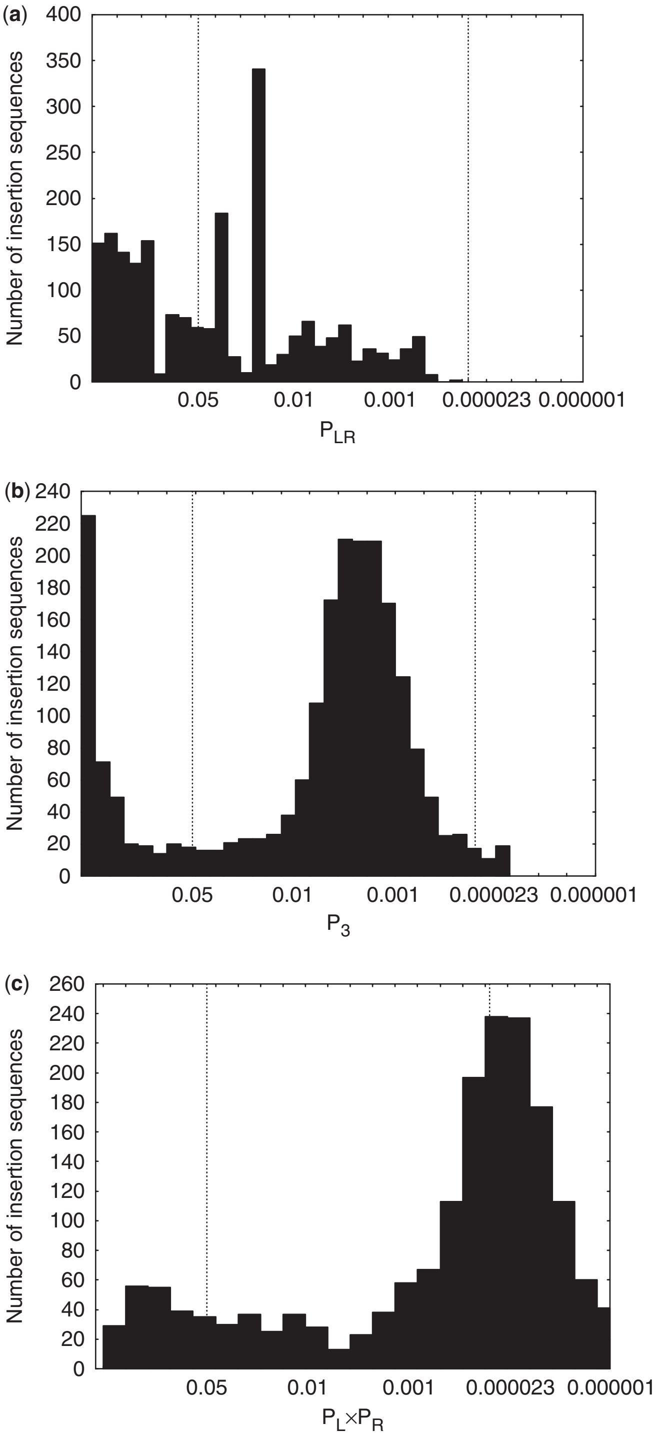

Figure 6. Distribution of five test statistics (see text) indicating inverted repeat quality for all ISs studied here. (**a**) P_{LR}; (**b**) P_3; (**c**) $P_L \times P_R$. The value 0.000023 is the Bonferroni-corrected *P*-value of 0.05.

ISs that have a family member with identical DNA sequence in the same genome are more likely to be functional than ISs for which this does not hold. The reason is that the two identical family members have most likely arisen through transposition, because gene duplication, the other prominent process that could account for the two IS copies, is much less frequent than transposition (29). In addition, bacterial IS transposase activity is usually tightly regulated, and often restricted to the IS from which transposase is expressed, such that passive transposition of a defective IS with the aid of an intact 'helper' IS may not occur often (1,7). This means that many ISs in our dataset with an identical IS in the same genome will be functional. If the alignment strategy we pursued above enriched for functional ISs, we would predict that the $P_L \times P_R$ values would be significantly lower for ISs with an identical family member in the same genome, than for other ISs. This is exactly what we observe. For example, for ISs with an identical IS in the same genome, the mean $(P_L \times P_R) = 0.019$, whereas for other ISs, the mean $(P_L \times P_R) = 0.046$. This difference is highly statistically significant as assessed by either a Mann–Whitney U-test ($n_1 = 1478$; $n_2 = 613$; $P = 1.2 \times 10^{-15}$) or a t-test ($P < 10^{-17}$).

In sum, although the inverted repeat units of an IS have limited sequence similarity, it is possible to enrich a data set for likely functional ISs by considering ISs with high $P_L \times P_R$ values. We note parenthetically that we also carried out a sequence similarity analysis among transposition target 'direct' repeat units generated by those ISs that are known to generate long direct repeats. This analysis showed that the direct target repeat units have too limited sequence similarity to be useful for this or other purposes (data not shown).

DISCUSSION

We have developed IScan, a publicly available tool to identify IS coding regions, and associated sequence elements (direct/inverted repeats). IScan is able to identify ISs with an arbitrary number of ORFs, including ISs with ORFs encoded on both strands. IS annotation in existing genomes may be highly heterogeneous, because different researchers may use different annotation methods. A tool like IScan thus allows the user to create consistent IS annotation with multiple user-specified parameters (repeat length, sequence similarity to a reference family member, etc.) across multiple genomes. This consistency and flexibility is essential for detailed analyses of IS evolution across multiple genomes.

Using IScan, we have surveyed 438 bacterial genomes for members of 20 IS families, and studied the similarity in their coding sequences as well as their inverted repeats. Recent other surveys of IS families focused on different aspects of IS biology and studied fewer or different genomes. Specifically, an intriguing analysis (40) studied ISs in 262 genomes and focused on the question: What determines IS copy numbers in a genome? (Briefly, the answer is genome size.) A short review by Siguier and collaborators (2) surveys different IS families and examples of IS evolution based on individual case studies. Yet another recent analysis focuses on archaeal ISs (8). In contrast to these papers, our work focuses on the sequence divergence of coding regions and inverted repeats in the largest number of genomes analyzed to date. Earlier work on the molecular evolution of ISs dates to the pre-genome era, restricted itself to narrow categories of ISs, or relied on previously available (and heterogeneous) genome sequence annotation for a smaller number of IS families (25,29). Our current analysis overcomes these limitations by analyzing members of all 20 IS families in an unprecedented (>400) number of genomes.

We find that the IS families we analyzed are well defined on the transposase sequence level. Specifically, although our approach would have admitted ISs with as little as 35% amino acid sequence identity to curated reference sequences, the mean amino acid divergence is lower than 7%, and more than 90% of all ISs have more than 70% amino acid identity to the reference. The different IS families show a skewed and patchy distribution, where most genomes carry no members of any given IS family, and a very small number of genomes carry many members [Figure 2b, see also (40)]. This distribution of IS occurrence resembles a similarly skewed distribution of IS occurrence on a much smaller taxonomic resolution, namely for 71 *E. coli* strains (25), where many strains carry few or no IS copies. At least two explanations might account for this skewed distribution. One of them involves selection against genomes with high IS copy numbers (see also below), another is an unidentified transposition immunity mechanism that suppresses IS copy number. These causes are non-exclusive and might operate jointly.

A strong or highly significant statistical association among IS families of IS copy numbers per genome might indicate that some bacteria are more susceptible to 'infection' by ISs in general. However, we do not find strong evidence of such an association for any pair of ISs beyond what would be expected from the shared evolutionary history of many bacterial species. Conversely, some ISs might be more successful than others, in that they can more easily 'infect' a larger number of genomes. Different ISs clearly show very different abundances. However, a thorough recent analysis suggests that among many possible factors determining IS abundance, genome size has by far the most important influence (40).

One biologically significant finding of our survey is the extreme sequence homogeneity of ISs within genomes. This substantially extends earlier, more limited work (25,29) and demonstrates a consistent pattern across IS families and vast taxonomic scales. This high sequence homogeneity stands in stark contrast to the greater sequence diversity among duplicate genes, another prominent class of repetitive DNA (29). Our earlier work suggests that gene conversion is not a likely sole cause of this high homogeneity, because common signatures of gene conversion are absent within IS families (29). Instead, this high homogeneity is readily explained by the rapid spreading of ISs within a genome. Consistent with this hypothesis is the observation that transposition and excision rates are very high on the time scale at which DNA sequences change. The high-sequence homogeneity of ISs might be explained by the following evolutionary scenario. After an IS enters a genome, its copy number expands rapidly through transposition (hence the low sequence diversity). Eventually, the IS becomes extinct again from the lineage, mostly due to natural selection, but perhaps aided by excision events. Some time thereafter, it may become reintroduced through horizontal gene transfer. Several other scenarios are not consistent with the data (29): If ISs did not go periodically extinct, they would show higher divergence within a genome; if they were not reintroduced by horizontal transfer, bacterial genomes would be devoid of them; and if the net effect of natural selection was an increase in IS copy number, then ISs should be much more diverse within a genome, because they would remain part of the genome indefinitely.

The only requirement for this evolutionary scenario is the frequent horizontal transfer of ISs. This is not a problematic requirement, as horizontal gene transfer may account for more than 10% of a bacterial genome's gene content (14,41). Its likely importance has been noted in an earlier, sequence-limited study on IS evolution in enteric bacteria (24). In addition, the mere fact that highly similar ISs of the same family occur in multiple distantly related genomes speaks to the importance of horizontal gene transfer for IS maintenance.

Among the methodological questions that a tool like IScan can address is whether rapid, computational, and automatic identification of functional or non-truncated ISs is possible. Aside from coding transposase-coding regions, ISs are typically associated with two sequence elements (direct target repeats and inverted repeats) that might be usable in such an identification. Direct target repeats, however, are of limited use for this purpose. High similarity of direct repeat units might indicate whether the transposition event that produced them occurred recently or a long time ago. However, this criterion only tells us whether the IS in question transposed or inserted successfully, not that it could again do so. In addition, target repeats for members of one IS family are very short, rendering their unambiguous identification difficult. Furthermore, their sequence similarity among different IS family members is highly limited (data not shown).

The second class of potential diagnostic sequence features are inverted repeats, except for IS families ISCR and IS91 that do not harbor such repeats. Even though a truncated IS may be only slightly shorter than its intact counterparts, one would expect that truncation is often associated with deletion of one inverted repeat unit. Also, it is reasonable to expect that the inverted repeat units of a functional IS show greater sequence similarity to each other than DNA fragments of the same length but randomly sampled from the same genome. We implemented a statistical test in IScan that relies on a large number n of such random fragments, to ask whether IS inverted repeats show such significant similarity. We applied this test ($n = 10^5$) to all the ISs we identified in the 438 genomes. When sequence similarity of inverted repeats is assessed through comparison with a reference IS (Figure 6c), then 90.3% of ISs have inverted repeats with significant similarity at $P = 0.05$. Of these, one would expect a fraction of 0.05 to be false positives, leading to an expected number of ISs with significantly similar inverted repeats of $0.95 \times 90.3\% = 85.8\%$ (1794) ISs. The more stringent Bonferroni-corrected $P = 2.3 \times 10^{-5}$ would yield 55.1% (1152) ISs with significantly similar inverted repeats. The inverted repeat P-values are significantly greater for ISs with an identical member of the same IS family in the same genome, many of which would be functional. Taken together, these observations suggest that, on one hand, inverted repeats may be highly flexible and cannot always be unambiguously identified. On the other hand, identification of ISs with highly similar inverted repeats may allow enrichment of a dataset with functional ISs, which may facilitate subsequent analyses. The high sequence similarity of ISs within genomes thus has not only biological implications. It also aids in defining a heuristic criterion—perhaps the only one—to identify functional ISs based on sequence data alone.

In closing, we note that the applications we illustrated here are only two among many uses that IScan might find. These uses will only increase as more genomes become available, and will include mapping of horizontal transfer histories, as well as transposition sequences within a genome.

ACKNOWLEDGEMENTS

Support through SNF grant 315200-116814 and through the Santa Fe Institute is gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by The University of Zurich.

Conflict of interest statement. None declared.

REFERENCES

1. Mahillon,J. and Chandler,M. (1998) Insertion sequences. *Microbiol. Mol. Biol. Rev.*, **62**, 725–774.
2. Siguier,P., Filee,J. and Chandler,M. (2006) Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.*, **9**, 526–531.
3. Toleman,M.A., Bennett,P.M. and Walsh,T.R. (2006) ISCR elements: Novel gene-capturing systems of the 21st century? *Microbiol. Mol. Biol. Rev.*, **70**, 296–316.
4. Lohe,A.R., Moriyama,E.N., Lidholm,D.A. and Hartl,D.L. (1995) Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. *Mol. Biol. Evol.*, **12**, 62–72.
5. Capy,P., Langin,T., Bigot,Y., Brunet,F., Daboussi,M.J., Periquet,G., David,J.R. and Hartl,D.L. (1994) Horizontal transmission versus ancient origin - mariner in the witness box. *Genetica*, **93**, 161–170.
6. Lampe,D.J., Witherspoon,D.J., Soto-Adames,F.N. and Robertson,H.M. (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.
7. Nagy,Z. and Chandler,M. (2004) Regulation of transposition in bacteria. *Res. Microbiol.*, **155**, 387–398.
8. Filee,J., Siguier,P. and Chandler,M. (2007) Insertion sequence diversity in Archaea. *Microbiol. Mol. Biol. Rev.*, **71**, 121–157.
9. Brugger,K., Redder,P., She,Q.X., Confalonieri,F., Zivanovic,Y. and Garrett,R.A. (2002) Mobile elements in archaeal genomes. *FEMS Microbiol. Lett.*, **206**, 131–141.
10. Buisine,N., Tang,C.M. and Chalmers,R. (2002) Transposon-like Correia elements: structure, distribution and genetic exchange between pathogenic Neisseria sp. *FEBS Lett.*, **522**, 52–58.
11. Siguier,P., Perochon,J., Lestrade,L., Mahillon,J. and Chandler,M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res. (Database Issue)*, **34**, D34–D36.
12. Orgel,L.E. and Crick,F.H.C. (1980) Selfish DNA: the ultimate parasite. *Nature*, **284**, 604–607.
13. Doolittle,W.F. and Sapienza,C. (1980) Selfish genes, the phenotype paradigm, and genome evolution. *Nature*, **284**, 601–607.

14. Bushman,F. (2002) *Lateral DNA Transfer: Mechanisms and Consequences*. Cold Spring Harbor University Press, Cold Spring Harbor, NY, USA.
15. Condit,R., Stewart,F. and Levin,B. (1988) The population biology of bacterial transposons - a priori conditions for maintenance as parasitic DNA. *Am. Naturalist*, **132**, 129–147.
16. Cooper,V.S., Schneider,M., Blot,M. and Lenski,R.E. (2001) Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli*. *J. Bacteriol.*, **183**, 2834–2841.
17. Schneider,D. and Lenski,R.E. (2004) Dynamics of insertion sequences elements during experimental evolution of bacteria. *Res. Microbiol.*, **155**, 319–327.
18. Schneider,D., Duperchy,E., Coursange,E., Lenski,R.E. and Blot,M. (2000) Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertionsequence-mediated mutation and rearrangements. *Genetics*, **156**, 477–488.
19. Treves,D.S., Manning,S. and Adams,J. (1998) Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*. *Mol. Biol. Evol.*, **15**, 789–797.
20. Naas,T., Blot,M., Fitch,W.M. and Arber,W. (1994) Insertion sequence-related genetic variation in resting *Escherichia coli* K-12. *Genetics* **136**, 721–730.
21. Dunham,M.J., Badrane,H., Ferea,T., Adams,J., Brown,P.O., Rosenzweig,F. and Botstein,D. (2002) Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proc. Natl Acad. Sci. USA*, **99**, 16144–16149.
22. Kleckner,N. (1989) In Berg,D. and Howe,M. (eds), *Mobile DNA*. American Society for Microbiology Press, Washington, DC, pp. 211–226.
23. Shen,M.M., Raleigh,E.A. and Kleckner,N. (1987) Physical analysis of Tn10 and IS10-promoted transpositions and rearrangements. *Genetics*, **116**, 359–369.
24. Lawrence,J.G., Ochman,H. and Hartl,D.L. (1992) The evolution of insertion sequences within enteric bacteria. *Genetics*, **131**, 9–20.
25. Sawyer,S.A., Dykhuizen,D.E., DuBose,R.F., Green,L., Mutangadura-Mhlanga,T., Wolczyk,D.F. and Hartl,D.L. (1987) Distribution and abundance of insertion sequences among natural isolates of Escherichia coli. *Genetics*, **115**, 51–63.
26. Hall,B.G., Parker,L.L., Betts,P.W., DuBose,R.F., Sawyer,S.A. and Hartl,D.L. (1989) IS103, a new insertion element in Escherichia coli: Characterization and distribution in natural populations. *Genetics*, **121**, 423–431.
27. Bisercic,M. and Ochman,H. (1995) Natural populations of *Escherichia coli* and *Salmonella typhimurium* harbor the same classes of insertion sequences. *Genetics*, **133**, 449–454.
28. Ajioka,J. and Hartl,D. (1989) In Berg,D. and Howe,M. (eds), *Mobile DNA*. American Society for Microbiology Press, Washington, DC, pp. 185–210.
29. Wagner,A. (2006) Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Mol. Biol. Evol.*, **23**, 723–733.
30. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (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.
31. Conant,G.C. and Wagner,A. (2002) GenomeHistory: a software tool and its applications to fully sequenced genomes. *Nucleic Acids Res.*, **30**, 1–10.
32. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped Blast and Psi-Blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
33. Muse,S.V. and Gaut,B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutiuon rates, with application to the chloroplast genome. *Mol.r Biol. Evo.*, **11**, 715–724.
34. Goldman,N. and Yang,Z.H. (1994) Codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.
35. Conant,G.C. and Wagner,A. (2003) Asymmetric sequence divergence of duplicate genes. *Genome Res.*, **13**, 2052–2058.
36. Myers,E. and Miller,W. (1988) Optimal alignments in linear space. *Comput. Appl. Biosci.*, **4**, 11–17.
37. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
38. Li,W.-H. (1997) *Molecular Evolution*. Chapter 7 Sinauer, MA.
39. Smith,T.F., Waterman,M.S. and Fitch,W.M. (1981) Comparative biosequence metrics. *J. Mol. Evol.*, **18**, 38–46.
40. Touchon,M. and Rocha,E.P.C. (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.*, **24**, 969–981.
41. Ochman,H., Lawrence,J. and Groisman,E. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.