OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences

Citation
Robinson, David G., Ming-Chun Lee, and Christopher J. Marx. 2012. Oasis: an automated program for global investigation of bacterial and archaeal insertion sequences. Nucleic Acids Research 40(22): e174.

Published Version
doi:10.1093/nar/gks778

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:11729522

Published online 16 August 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 22* ***e174***
doi:10.1093/nar/gks778

OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences

David G. Robinson[1], Ming-Chun Lee[1] and Christopher J. Marx[1,2,*]

[1]Department of Organismic and Evolutionary Biology and [2]Faculty of Arts and Sciences Center for Systems Biology, Harvard University, Cambridge, MA 02138, USA

Received March 31, 2012; Revised July 6, 2012; Accepted July 24, 2012

ABSTRACT

Insertion sequences (ISs) are simple transposable elements present in most bacterial and archaeal genomes and play an important role in genomic evolution. The recent expansion of sequenced genomes offers the opportunity to study ISs comprehensively, but this requires efficient and accurate tools for IS annotation. We have developed an open-source program called OASIS, or Optimized Annotation System for Insertion Sequences, which automatically annotates ISs within sequenced genomes. OASIS annotations of 1737 bacterial and archaeal genomes offered an unprecedented opportunity to examine IS evolution. At a broad scale, we found that most IS families are quite widespread; however, they are not present randomly across taxa. This may indicate differential loss, barriers to exchange and/or insufficient time to equilibrate across clades. The number of ISs increases with genome length, but there is both tremendous variation and no increase in IS density for genomes >2 Mb. At the finer scale of recently diverged genomes, the proportion of shared IS content falls sharply, suggesting loss and/or emergence of barriers to successful cross-infection occurs rapidly. Surprisingly, even after controlling for 16S rRNA sequence divergence, the same ISs were more likely to be shared between genomes labeled as the same species rather than as different species.

INTRODUCTION

The ever-increasing number of sequenced bacterial and archaeal genomes provides a valuable opportunity to understand genome architecture and evolution. However, as new high-throughput sequencing methods are developed, genome annotation quickly becomes the bottleneck for genomic research. Despite the development of various annotation programs for particular genomic features, some important features such as insertion sequences (ISs), the smallest and simplest autonomous mobile genetic elements, remain poorly annotated.

ISs are short regions of DNA, usually between 700 and 3000 bp long that can move or copy themselves within a genome through self-transposition. The majority of ISs possess one or two open reading frames (ORFs) that encode a transposase. These ORFs are surrounded by linker regions that frequently end with short-terminal inverted repeats (IRs) ranging from 7 to 20 bp in length. Upon insertion, ISs often generate short directed repeats from 2 to 14 bp immediately outside the IRs (1). Despite considerable sequence divergence, ISs can be classified into 26 families based on transposase homology and overall organization, with some families divided further into groups (2).

Due to their movement within and across genomes, ISs not only represent an important source of genetic variation within genomes but also mediate horizontal gene transfer (HGT) among organisms and thus play a key role in genome evolution. Through transposition, ISs can interrupt the coding region of a gene, or disrupt promoter regions and alter gene expression. Given that there can be hundreds of copies of the same IS in a genome, they can also serve as sites of rearrangements

*To whom correspondence should be addressed. Tel: +1 617 496 8103; Fax: +1 617 495 8848; Email: cmarx@oeb.harvard.edu
Present addresses:
David G. Robinson, Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA.
Ming-Chun Lee, Department of Biochemistry, The University of Hong Kong, Pok Fu Lam, Hong Kong.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

such as deletions, duplications and inversions through homologous recombination. On a practical level, transposition has been a classic genetic technique to generate mutant alleles, generally loss-of-function insertions into gene products. ISs are thus often regarded as 'selfish' genomic parasites proliferating at the cost of their host and surviving only through horizontal transfer (3). In contrast, experimental evolution in the laboratory has demonstrated that both transpositions (4,5) and rearrangements (6–10) can also generate beneficial mutations and commonly represent a large portion of the mutations identified in genomes following a period of adaptation (11,12).

The most comprehensive resource for ISs is the ISfinder database (13). ISfinder annotations have been submitted by users and manually verified by the curators. Therefore, we assumed this database (as of 17 September 2011) to be an accurate set of all ISs, but incomplete due to the fact that genomes are being sequenced faster than they are annotated to this extent. The most common practice in genome annotation has been to stop at the point of labeling ORFs as 'transposase' or 'integrase' where sufficient homology was observed. Without classification of ISs into families and enumeration within genomes, neither broad-scale studies across taxa nor dynamics within closely related strains are possible.

Previous approaches to annotate ISs across a wide range of genomes have been either involved internal pipelines, require manual annotation, or have identified very few elements. An annotation program was used for an analysis of 19 cyanobacterial and 31 archaeal genomes, but this has yet to be made publicly available as an automated pipeline (2). ISSaga is a web application pipeline that allows semi-automated annotation based on BLAST against the ISfinder database (14). While ISSaga provides useful tools both for recognizing transposase ORFs and for manually curating their edges, it cannot automatically identify novel ISs not already present in ISfinder. Due to the requirements for manual annotation of novel elements and individual submission of genomes online, ISSaga is impractical for comprehensive evolutionary analyses involving a large number of sequenced genomes. To our knowledge, the only publically available program that has been developed to identify new ISs is IScan (15). This system utilizes BLAST with a single reference transposase sequence per IS family to locate novel transposases. An investigation of ISs in 438 prokaryotic genomes concluded that, given the limited number of ISs identified by this approach in most taxa, most IS families are quite limited in their phylogenetic distribution (16). Whole taxa, such as the α-proteobacteria, had almost no identifiable ISs via this method, whereas manual annotation has found genomes of many of these organisms, for example *Methylobacterium extorquens* (17), are rather densely populated with ISs.

To perform a global investigation of ISs in bacterial and archaeal genomes, we developed OASIS, or Optimized Annotation System for Insertion Sequences, a computational tool for automated annotation of ISs. OASIS takes advantage of widely available transposase annotations to identify candidate ISs and then uses a computationally efficient maximum likelihood method of multiple sequence alignment to identify the edges of each element. Thanks to its speed and flexibility, OASIS is capable not only of providing detailed IS information for a single genome but also of annotating thousands of genomes within hours, making it a valuable high-throughput tool for a global investigation of IS distribution across diverse taxa. We applied OASIS to 1737 sequenced bacterial and archaeal genomes. Through comparisons across 1319 genomes to a benchmark of ISfinder annotations, OASIS performed approximately an order of magnitude better than IScan. With a more comprehensive and accurate overall picture of IS distribution across genomes, we were able to address the pattern of ISs evolution at scales ranging from all sequenced Bacteria and Archaea to diversity between sets of extremely closely related lineages. Broadly, it is clear that IS families are quite widespread; however, they are not present randomly across phyla suggesting either borders to exchange or insufficient time to equilibrate. Considering IS number and density with regard to genome length, we find tremendous variation, and with the exception of genomes <2 Mb which have considerably fewer ISs, there is no positive correlation of IS density and genome length. Finally, we find that the proportion of shared IS content between recently diverged genomes drops quite quickly. Quite surprisingly, even after 16S divergence is taken into account, strains labeled as the same species share more IS elements than those termed as being different, suggesting that IS exchange/survival is correlated with current taxonomy.

MATERIALS AND METHODS

OASIS algorithm

OASIS is a free open source program implemented in Python. It is available at https://github.com/dgrtwo/OASIS. The input data in this study consisted of the 1737 curated microbial genomes available in NCBI as of 17 September 2011. These genomes were downloaded from NCBI in GenBank format from the publicly available server. OASIS uses a library of 3703 ISs from ISfinder (13), in the form of an amino acid FASTA file, which are used to identify the family and group of each IS. These transposase sequences were automatically downloaded from the ISfinder database on 17 September 2011.

ISs may occur once in a genome or may consist of a set of almost identical copies (2). ISs in a particular genome can therefore be classified into two major groups in terms of copy number: multicopy and single-copy ISs. As there are distinct levels of information available in each of these cases, different algorithms perform better with each class. As such, we have designed OASIS to find these two groups of ISs in two separate steps: first finding multicopy ISs and then single-copy ISs. The overall schematic pipeline is shown in Figure 1.

OASIS identifies multicopy ISs in each genome by finding conserved regions surrounding already-annotated transposase genes, which are identified by the word 'transposase' in the 'product' field of GenBank files.

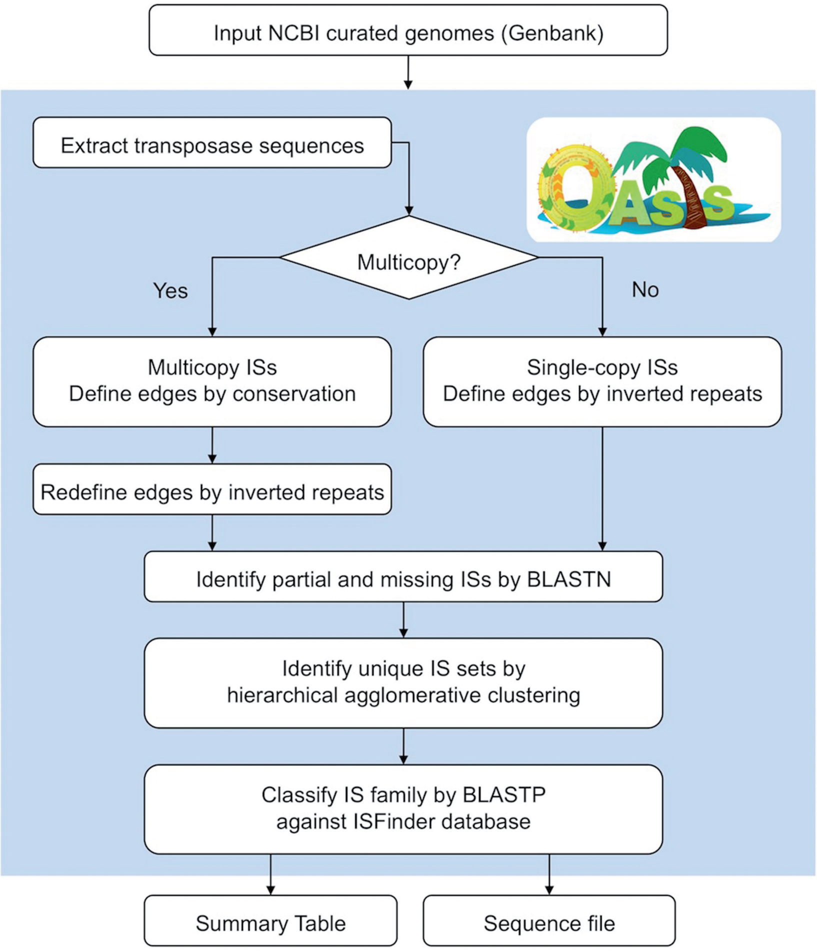
Input NCBI curated genomes (Genbank)
Extract transposase sequences
OASIS
Multicopy?
Yes
No
Multicopy ISs
Define edges by conservation
Single-copy ISs
Define edges by inverted repeats
Redefine edges by inverted repeats
Identify partial and missing ISs by BLASTN
Identify unique IS sets by
hierarchical agglomerative clustering
Classify IS family by BLASTP
against ISFinder database
Summary Table
Sequence file

Figure 1. Flowchart portraying the full workflow of OASIS.

The choice to require prior annotation limits use on unannotated genomic data; however, as published complete genomes are nearly always annotated to this level of detail via other pipelines, we feel that the computational speed gained makes this decision quite worthwhile. One additional issue is that transposases are occasionally misannotated as integrases, a term more commonly found for prophage. In order to avoid false positives from prophage, which are generally present at single copies in a genome, genes containing the word 'integrase' are considered only when present in multiple copies. Groups of transposase genes that compose multiple copies of an IS are identified by length and sequence similarity. Two ORFs are considered similar if the identity between the sequences, assuming no gaps, is >95%, This threshold allows for small differences between the transposase sequences in multiple copies of the same ISs, though it should be noted that the great majority of cases of multiple copies are within 2% of each other. Genes that fit this similarity threshold are combined into the same IS set. OASIS then uses a maximum likelihood algorithm, described in Supplementary Methods, to determine the edges of multicopy ISs based on conservation between their surrounding regions.

To define the edges of single-copy ISs, we use an approach first developed by IScan to find IRs around the transposases, which are present for the majority of ISs (15). Briefly, a Smith–Waterman alignment, with a match score of 1, a mismatch penalty of −3 and a gap penalty of −4, is performed comparing the region upstream of the transposase (500 bp) with the reverse complement of the downstream region (500 bp) and the highest match with a score >10 is assumed to be the pair of terminal IRs.

Multicopy ISs are also checked for IRs using a Smith–Waterman alignment with the same parameters, comparing the regions within 100 bp of each edge. If the IRs disagree with the edges determined by the maximum likelihood algorithm, the edges are changed to match the IRs. If none are found, the region immediately inside the IS is checked with a mismatch penalty of only −1, in case OASIS had already identified the edges correctly.

As a result, each full-length IS is composed of a transposase, a protein of one or more ORFs, and upstream and downstream sequences defined as linker regions, typically ranging from 0 to 500 bp. The extreme edges of the IS can include a partially conserved IR on each end ranging from 8 to 20 bp in length.

Once groups of ISs are identified, BLASTN (NCBI) is used with one example from each set of ISs, selected based on the presence of inferred IRs and the mode length of the ISs, against the genome sequence to identify missing and partial copies of the IS (for which there is often no transposase annotated). Thus, when present in multiple copies OASIS finds partial ISs; it is not capable of finding these small IS fragments when no intact copy with an annotated transposase is present. Redundant BLAST results within a set are filtered out. OASIS then uses hierarchical agglomerative clustering to identify groups of IS lengths, clustering together groups whose mean lengths are closer than 100 bp apart. The mode cluster is then assumed to be the true size of the IS and any fragments that are shorter than that threshold or 600 bp are classified as partials. One intact ORF was selected from each set of ISs and BLASTP was used against the ISfinder database. If there were matches with an *e*-value $<10^{-12}$, the IS set is classified according to the family and group of the best match, otherwise the IS set is classified as 'None'.

The final output of OASIS includes two files for each annotated genome: a file in GFF format of the ISs in each genome and each ISs characteristics, including the chromosome ID, start and end positions, direction, family and group, IRs (if found) and whether the element is a partial element, and a file containing the nucleotide sequence of each identified IS and the amino acid sequence of each transposase in FASTA format.

Evaluation

The performance of OASIS was compared to IScan, using ISfinder database as the benchmark annotation. Each genome was then annotated using three methods: OASIS, IScan and a BLASTN against ISfinder database. OASIS was performed using its default settings and IScan annotations were obtained by running IScan with its default settings on each genome. IScan uses BLAST with reference transposase sequences to find transposases. In this run, we extracted the same reference ISs as in the original IScan analysis (16). Benchmark annotations were obtained by mapping ISfinder sequences onto genomes. BLASTN was used with ISfinder nucleotide sequences to search genomes in the same genus as the query sequence's origin. BLAST hits with identity <90% were removed, as were hits that were redundant (the starts and ends are within 100 bp of others in the same genome). Only 1311 of the genomes shared a genus with any ISfinder element and were included in the evaluation analysis.

Sensitivity and specificity were assessed by matching elements between OASIS or IScan against the benchmark data set of ISfinder. Partial and truncated ISs are usually not annotated in ISfinder and thus were excluded from the evaluation analysis. As edges could be mis-annotated either by the BLAST from ISfinder or by one of the automated annotation methods, any elements that overlapped were considered matches, though the effect of requiring accurate matches is shown in Supplementary Figure S1.

OASIS$^+$ data set

While OASIS found two-thirds of the ISs obtained by mapping ISfinder elements back to genomes and found nearly half as many new elements, there are many it did not find (see 'Discussion' section). In order to make the subsequent biological analysis as comprehensive as possible, we combined the OASIS annotations with the ISfinder annotations to form the OASIS$^+$ data set. To remove any redundancy between the annotation sets, any copy in ISfinder that overlapped a copy in OASIS was not included.

Phylogenetic analysis

Aligned 16S rRNA sequences were extracted from Ribosomal Database Project Release 10 on 14 November 2011, by matching the genbank accession number of each genome (18). For genomes that have multiple annotated 16S sequences, the 16S sequence from each genome with the highest average similarity to all other 16S sequences was selected. In total, 1502 16S sequences were extracted from RDP database and an extra 195 sequences were aligned first and added to the RDP alignment by ClustalW2 profile alignment (19). Fourteen sequences were deleted due to poor quality of the 16S sequences (e.g. multiple Ns) after manual verification. An overall phylogeny of the 1682 genomes was constructed using Neighbor-joining method (20) and displayed by Interactive Tree of Life (21). The complete tabulation of IS family content per genome is available in Supplementary Table S3.

We used SourceCluster, a pairwise distance test (22), to test the non-randomness of IS distribution across genomes for each family. The sum of all pairwise distances between 16S sequences of the genomes with at least one IS of that family was calculated to represent the degree of clustering effect. To generate a corresponding null distribution for each family, 1000 Monte Carlo simulations were performed by sampling the same number of pairs of genomes randomly and the position of the test statistic within the null distribution was used as the *p*-value.

IS content comparison for closely related genomes

To investigate shared IS content between closely related genomes, we computed the numbers of matching ISs between each pair of genomes whose 16S divergence is <10%. Since the maximum 16S divergence within bacterial species is commonly cited to be 3%, we conservatively excluded genomes with 16S sequences that, on an average, diverged >3% with other genomes in their species to prevent cases of misannotation or poor alignments from skewing our results (23). We used an alternative index rather than the actual genomic position due to the lack of consistency in nucleotide coordinates between even recently diverged genomes. We considered ISs between two genomes to match if their families are identical, their lengths are within 200 bp of each other and if the ratio of the Levenshtein edit distance (the number of insertions, substitutions and deletions necessary to transform one string into another (24) to the average length is <0.2.

For each genome in the pair, we computed the number of elements that match an element in the other genome. We then fit the following logistic model:

$$p_{i,j} = \frac{1}{1+e^{-z_{i,j}}}$$

$$z_{i,j} = \beta_0+\beta_1 \log_{10}(D_{i,j})+\beta_2 S_{i,j}+B_F$$

where $p_{i,j}$ is the probability an IS in one of the two genomes i and j appears in the other, $D_{i,j}$ is the 16S divergence between the genomes, $S_{i,j}$ is whether the two genomes are the same species and β_F is a coefficient specific to that IS family. We added 0.0001 to the divergence so that the logarithm could be taken, as the divergence can be zero. We performed logistic regression with the number of shared and lost elements in each pair as the binomial response variables. A higher coefficient indicates that a factor is associated with a higher probability of IS sharing between two genomes.

RESULTS AND DISCUSSION

OASIS performance

A total number of 41 821 copies representing 6829 unique IS sets were identified by OASIS in 1240 genomes out of the 1737 analyzed (Table 1 and Figure 2). Among those, 16.4% of the unique IS sets are single-copy ISs. The remaining sets belong to multicopy ISs and comprise 97.3% of the total copies, with the largest multicopy set in a single genome containing 232 elements. OASIS took a total of 9 h and 40 min to annotate all 1737 genomes on a 4-core 2.8-Ghz processor, with a maximum per-genome running time of 6 min.

To evaluate the performance of OASIS, we compared the program's annotation to those of IScan by testing each against benchmark annotations obtained from the ISfinder database. Given that many genomes are not represented in ISfinder, we considered sensitivity as a measure of annotation accuracy. Within the 1311 benchmarked genomes, OASIS found ~66.0% of the 40 078 ISfinder annotations, whereas IScan found only 8.5%. OASIS and IScan frequently identify ISs in ISfinder but disagree about the boundaries; OASIS annotated 42.3% of the ISfinder annotations to within 10 bp of each edge, while IScan found 3.1% at that error tolerance (the effect of error tolerance on sensitivity is shown in Supplementary Figure S1). The low sensitivity of IScan was not an artifact of our execution of the software, as the number and distribution of ISs qualitatively matches the previously reported results (16,25). In addition, OASIS identified 14 264 ISs that were not present in ISfinder, which is over half of the size of the database (Figure 3). We manually validated the novel ISs in a subset of data (1012 genomes, downloaded on February 2010) and found that at least 85% of the 2625 unique IS types contain real, intact transposes that had not submitted to the ISfinder database, confirming the incompleteness of ISfinder database. Interestingly, despite the great difference in numbers and sensitivity between OASIS and IScan, IScan found 1146 ISfinder elements that OASIS did not.

The sensitivities of both OASIS and IScan depend very strongly on the genus of the annotated genome. Supplementary Table S1 shows the individual sensitivities for genera that include 10 or more genomes. IScan annotations were most sensitive in *Escherichia* genomes, finding 25.1% of *E. coli* ISs, while finding no ISfinder elements at all in several other large genera that contain ISs, such as *Streptococcus*, *Bacillus*, *Clostridium*, *Mycoplasma* and *Mycobacterium*. This specificity to particular taxa is likely

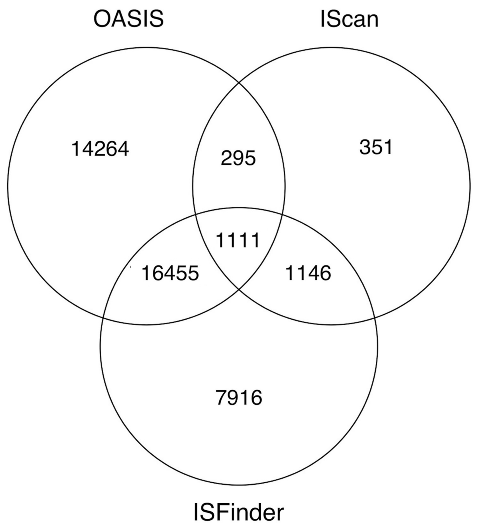

Figure 2. Venn diagram illustrating the identified number of ISs in ISfinder, OASIS and IScan. OASIS found a total of 37 427 ISs (in copy numbers) in the 1319 benchmarked genomes while IScan only found 2902 ISs, demonstrating a better performance of OASIS over IScan. In addition to identifying 18 112 ISfinder elements, OASIS found 19 365 new ISs, indicating the advantage of OASIS in finding novel ISs.

Table 1. The size and characteristics of the four sets of IS annotations in this analysis

Annotation set	Number of copies	Number of sets	Number of genomes	Number of partial elements	% of partial elements	Sensitivity (%)	Specificity (%)
OASIS	41 821	6829	1240	5655	13.5	66	54.7
IScan	3252	428	257	335	10.3	8.5	77.5
ISfinder BLAST	40 078	6012	932	10 584	26.4	–	–
OASIS+	56 786	10 529	1347	11 022	19.4	–	–

OASIS$^+$ is the combination of the OASIS and ISfinder data sets. ISfinder and OASIS$^+$ have no sensitivity or specificity since ISfinder is our benchmark and OASIS+ incorporates ISfinder.

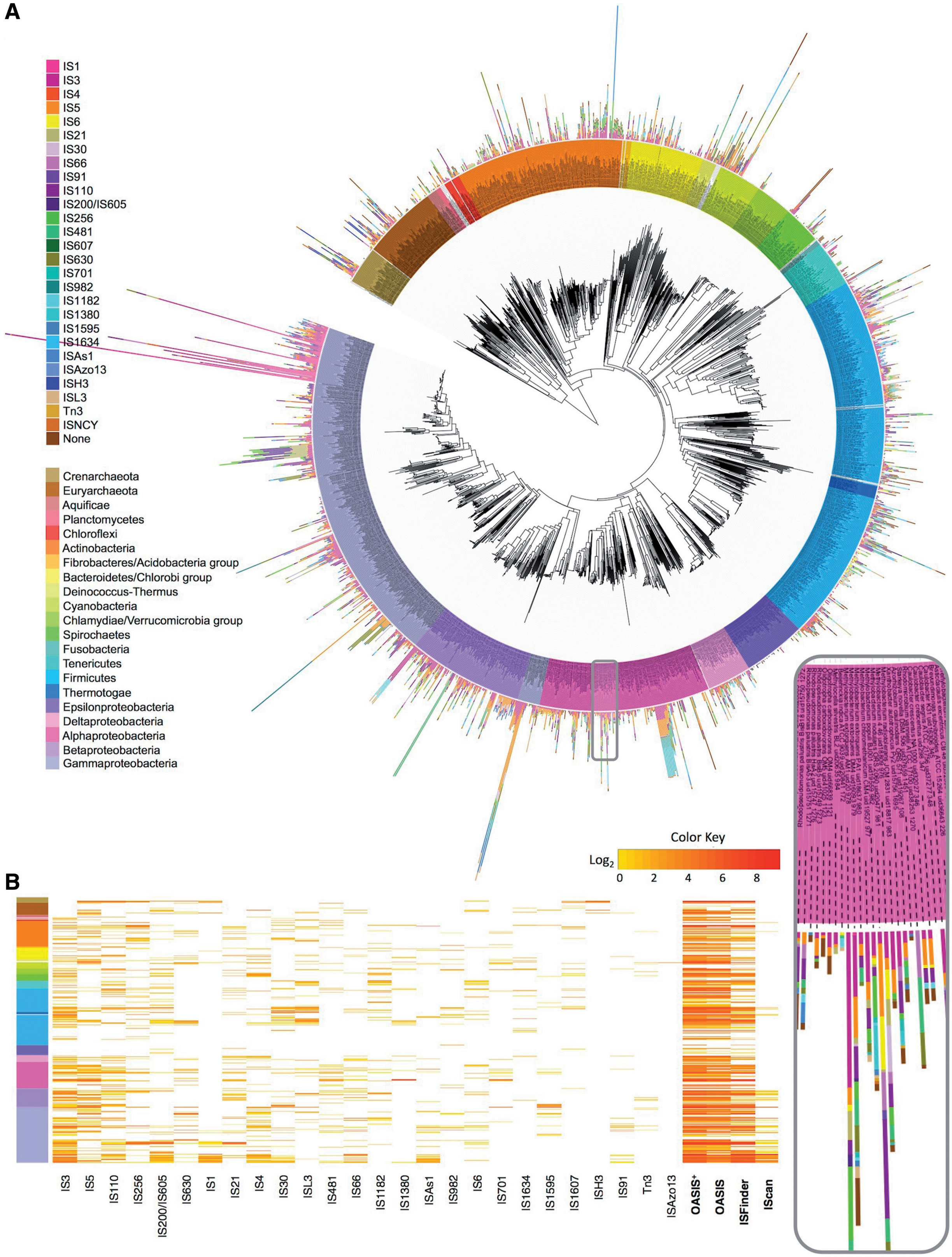

Figure 3. IS distribution across bacterial and archaeal genomes. (**A**) A neighbor-joining tree of 16S rDNA from 1682 completely sequenced bacterial and archaeal genomes, colored by phylum (proteobacteria are separated by class). Length of colored bars outside the tree is proportional to the numbers of ISs identified by OASIS^{+}, colored by 26 IS families (ISNCY: not classified yet; None: no hit in ISfinder database). The OASIS^{+} data set consists of the combined annotations of OASIS and a BLASTN from ISfinder (see 'Materials and Methods' section, complete data in Supplementary Table S3). (**B**) A heatmap showing the non-random occurrence of each IS family across genomes. Genomes are ordered according to the 16S tree in (A). The final four columns indicate the distribution of ISs present in the various data sets we compare.

Table 2. The size and prevalence of each family in OASIS^{+}, sorted by the number of IS copies in each

Family	Number of copies	Number of IS sets	Number of genomes	Number of genera	Number of classes	Number of phyla	Average copy number
IS3	9630	1902	799	246	28	18	5.06
IS5	6692	1073	521	213	35	20	6.24
None	3951	742	483	237	40	22	5.32
IS110	3535	829	488	177	26	16	4.26
IS256	2890	512	337	131	25	16	5.64
IS200/IS605	2770	499	334	107	25	13	5.55
IS630	2721	424	252	126	23	11	6.42
ISL3	2262	426	268	85	19	12	5.31
IS1	2247	152	110	27	11	7	14.78
IS4	2245	521	295	118	21	12	4.31
IS21	2225	386	251	115	18	12	5.76
IS481	1984	270	207	96	17	10	7.35
IS30	1916	355	259	76	15	8	5.4
IS66	1654	360	204	87	17	9	4.59
IS1182	1564	292	224	95	16	9	5.36
ISAs1	1143	259	136	47	12	7	4.41
IS1380	1075	133	106	52	13	6	8.08
IS982	907	133	100	45	16	11	6.82
IS701	882	126	91	57	18	11	7
IS1634	862	140	89	54	20	11	6.16
IS6	725	219	177	60	18	12	3.31
IS1595	674	162	117	52	11	7	4.16
ISNCY	660	182	148	80	22	13	3.63
IS607	498	178	94	42	18	12	2.8
ISH3	491	62	25	10	4	3	7.92
IS91	314	94	72	35	13	6	3.34
Tn3	213	85	72	47	8	4	2.51
ISAzo13	56	13	13	9	7	5	4.31
OASIS+	56 786	10 529	1347	448	50	26	5.39
All	–	–	1737	563	58	30	–

Note that many ISs had families that could not be identified and were recorded as None. The OASIS^{+} row describes the total number of taxa that contain ISs, while the 'All' row represents the total number of taxa used in the analysis.

an artifact of IScan for identifying transposases, which depends upon BLAST homology with selected reference sequences (mostly from *E. coli* and other Proteobacteria). In terms of the total number of annotations and sensitivity to ISfinder, we found OASIS has better performance than IScan overall and also within every genus with more than 10 genomes.

Both OASIS and IScan have a much lower rate of detecting single-copy ISs than multicopy ISs, particularly at annotating their edges accurately. They find 10.2 and 3.9% of single-copy ISs, respectively, whereas at an error tolerance of 10 bp, these fall to only 6.0 and 1.3%. OASIS outperforms IScan by using sequence conservation between multiple copies to identify the edge of the element, which gives a much higher signal. Among elements that overlapped a benchmark element, OASIS missed the edges of the benchmark by a median of 2 bp, while IScan missed the edges by a median of 51 bp.

Despite the significant improvement of OASIS on IS annotation, the major limitation of our program is that it depends on the quality of NCBI annotation. Through manual inspection, we found many elements do not have annotated transposases in the NCBI genomes, which makes OASIS incapable of identifying them as candidates. OASIS is thus limited by the efficacy of transposase annotation and identification algorithms, which can be inaccurate. Furthermore, that the sensitivity of OASIS increases as the error tolerance increases indicates that OASIS finds some elements but misannotates their edges. It is also possible in some cases that the edges are incorrectly annotated in the benchmark data set.

IS element abundance and distribution

For further analysis of broad-scale patterns of ISs across bacteria and archaea we combined our OASIS annotations of ISs with those already in ISfinder to generate a collection, 'OASIS^{+}', of 10 529 ISs (due to multicopy ISs, 56 786 elements in total) across 26 IS families (Table 2). ISs appeared in 77.5% of the genomes, indicating a global IS distribution across archaeal and bacterial domains (Figure 3 and Supplementary Table S2). This result concurs with the prevailing view of ISs being widespread but absent from a moderate number of microbial genomes (26); however, it is in contrast to the results of IScan (Supplementary Figure S2). The proportion of ISs identified across taxa varied substantially. Largely due to the large number of sequenced genomes from these groups, >60% of the ISs we found are from Proteobacteria and ~20% are from Firmicutes. For phyla that have more than 10 genomes, the average number per genome ranges from 15 to 50 except the Chlamydiae/Verrucomicrobia and the Epsilonproteobacteria groups, which have fewer than 3 ISs per genome on average.

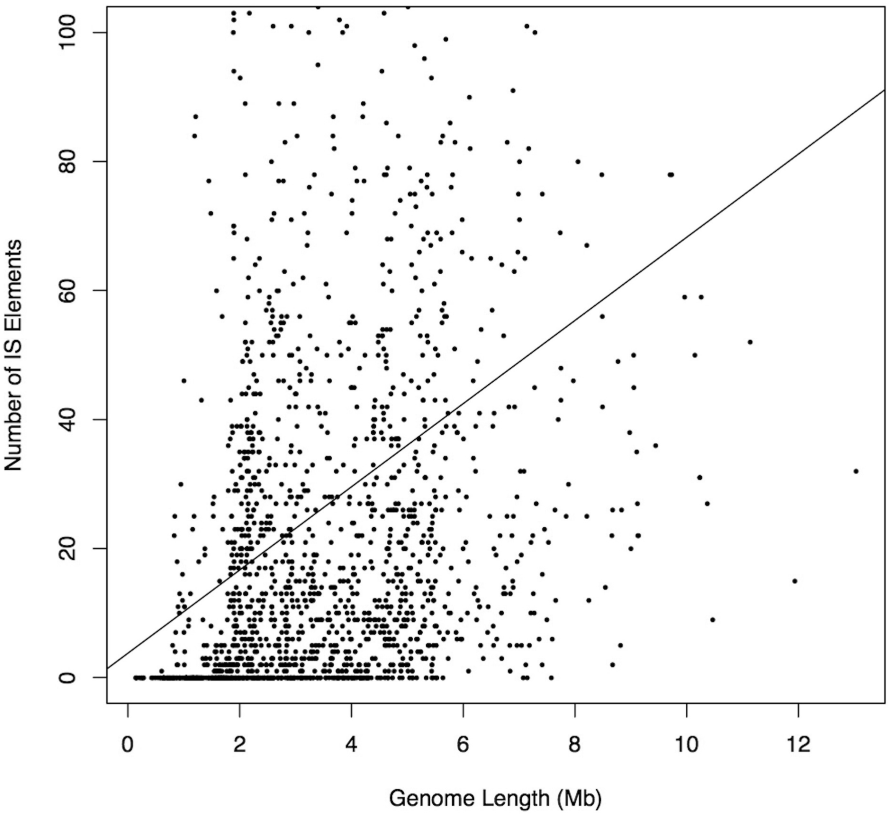

Figure 4. Plot of genome length versus number of ISs per genome, (Spearman's $\rho = 0.459$, $p \approx 0$). A best-fit line from a linear regression is shown, but it captures relatively little of the variation present (slope 8.42, $R^2 = 0.074$). Note that the y axis is restricted to 100 for clarity, though some points lie above the plot's range.

In addition, our results show no ISs in the 13 genomes in *Buchnera* or the 12 genomes in *Prochlorococcus*, which is consistent with previous reports (27). This indicates the absence of ISs in reduced genomes, which is consistent with previous observations (28–32).

Given that certain phyla, particularly those with smaller genomes, have few IS elements we directly explored the effect of genome size on IS number and density in the OASIS$^+$ data set. In a previous study of ISs in 262 genomes, a positive correlation was found between IS content and genome size, which they attributed to the lower density of essential genes and thus of deleterious insertion sites (26). In the OASIS$^+$ data, we found a positive Spearman's coefficient of 0.46 ($p \approx 0$) between each genome's length and its number of ISs, though this is weaker than the correlation found earlier using fewer genomes (Figure 4). Much of this correlation is driven by the rarity of ISs in small genomes: the correlation among genomes of length >2 Mb drops to just 0.21. The previous study also discovered a correlation between IS density and genome length. While we discovered a weak correlation between density and length (Spearman's $\rho = 0.25$, $p = 0$), it was entirely due to the rarity of ISs in small genomes. The correlation drops to −0.040 when only genomes >2 Mb are considered. This suggests that ISs have difficulty surviving in genomes <2 Mb, but that the density of ISs permitted does not further increase for genomes >2 MB.

Beyond considering IS prevalence as a whole, we also determined that there is some degree of clade specificity and a non-random distribution for nearly every individual IS family (Figure 3B and Supplementary Table S3). IS families differ greatly in both their frequency and diversity of hosts, appearing in as few as 13 genomes (ISAzo13) to as many as 799 genomes (IS3) (Table 2). Each IS family also differs in its copy number per IS set (unique IS type). For example, IS1 has an average of 14.8 copies per set while Tn3 (technically a transposon that has additional passenger genes, but it is treated as an IS family by ISfinder) has an average of only 2.5 copies per set, suggesting different replicative transposition rates across different types of ISs. Interestingly, several families were limited to only a single domain. IS1380, Tn3, ISAs1 and IS30 were present only in bacterial genomes, and IS3, despite being the most abundant IS family, was absent in all archaeal genomes except two. On the other hand, ISH3 was found only in Archaea. To statistically examine the clustering effects of IS families, we applied phylogenetic clustering analysis to compare the average pairwise distance of 16S sequences within each IS family and test it against a null distribution (see 'Materials and Methods' section). Consistent with the above observations, the results showed 22 out of 26 families are significantly phylogenetically clustered ($p < 0.05$) (Supplementary Table S4), indicating a non-random distribution of ISs across genomes. This non-random distribution of ISs

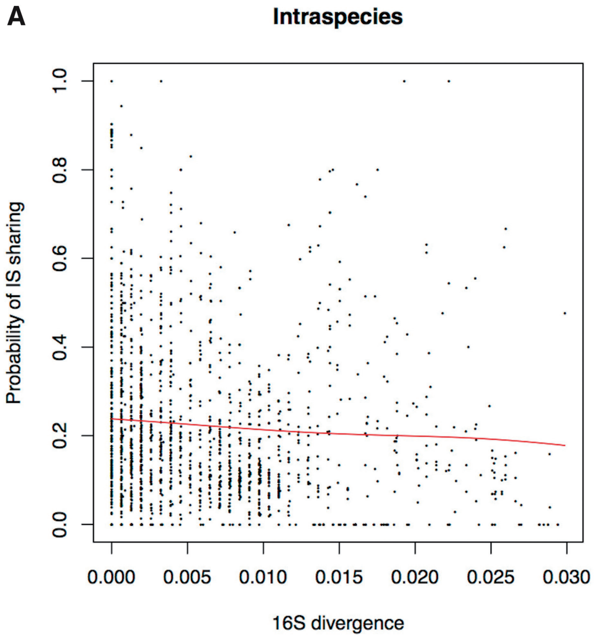

Figure 5. A plot of the probability of two genomes sharing an IS copy compared to their 16S distance, for: (**A**) intraspecies and (**B**) interspecies pairs. An average using a Nadaraya–Watson kernel smoother with a bandwidth of 1% is shown in red. Note that at all 16S sequence distances the intraspecies value remains at least 2-fold higher than interspecies.

across genomes might be due to either boundaries to successful horizontal transfer, distinct selective forces faced by different IS families in different taxa, or recent emergence and insufficient time to equilibrate across taxa.

IS element survival in closely related genomes

The large number of taxa in our OASIS$^+$ data set that represent closely related genomes (strains within the same species or genus) allowed us to address the extent to which IS turnover is seen at a finer scale. Previous studies have determined that IS counts can vary widely even between closely related genomes (26,33,34). Most of these studies concluded that ISs are short lived in natural lineages and that rapid HGT is required for them to persist. To investigate the short-term proliferation and survival of ISs, we compared the fraction of ISs shared in each pair of closely related genomes with their 16S divergence. The results show that the probability of an IS being shared between genomes decreases dramatically with increasing 16S distance: an IS has a 35.1% chance of being shared between two genomes that have identical 16S sequences, while it has only a 0.13% chance of being shared between two genomes with 9–10% 16S sequence divergence. Also notably, the probability of an IS being shared between two genomes in the same species is 24.7%, while two genomes of different species within 10% 16S divergence have only a 2.2% chance of being shared. Figure 5 compares the divergence between each genomic pair to the percentage of ISs shared between them, for both intraspecies and interspecies pairs and shows that probability of sharing decreases very quickly with increasing divergence. This rapid decrease in shared IS content during the early divergence of genomes could be due to a decline in the probability of vertical inheritance or in the chance of being acquired horizontally. Figure 5 also shows for a given degree of 16S sequence divergence, pairs of taxa that are defined as the same species have more similar IS content than if they are defined as different species. As an example, the probability of an IS being shared between two *Escherichia* genomes is 17.4% and the probability of being shared between two *Shigella* genomes is 41.1%, while the probability of being shared between an *Escherichia* genome and a *Shigella* genome is only 4.7%. *Shigella* is a polyphyletic genus completely within the *Escherichia* genus (35) (the average 16S divergence between the two genera is 1.6%), so the difference in IS content might reflect biological differences rather than just the time since divergence.

In order to quantify the effect of these factors on the probability of shared IS content, we performed logistic regression to predict the proportion of matched IS copies between two genomes, using the log of the 16S distance, IS family and whether a pair was within a single species. Different IS families have very different probabilities of being shared between closely related genomes (Supplementary Table S5). Families such as IS1595 and IS1380 were particularly likely to appear in multiple closely related genomes, while families such as IS91, Tn3 and IS66 were unlikely to be shared. While the log genomic distance was found to have a very significant effect, particularly important was whether the two genomes were considered to be part of the same species ($p \approx 0$). The intraspecies factor had a logistic regression coefficient of 1.85 ± 0.01, which can be interpreted as an odds ratio of $e^{1.85} = 6.36$, indicating that there is a strong effect of being defined as part of the same species on top of the effect of the pairwise 16S distance. The wide range of coefficients among IS families shows that family does have a large effect on the probability of being shared, suggesting that elements in different families have different probabilities of net gain or loss.

Part of the difference causing certain IS families to be shared across closely related genomes can be attributed to their differing copy numbers, For example, the IS family with the lowest average copy number, Tn3, has one of the lowest coefficients in the regression and therefore the lowest probability of IS sharing. Also important is the effect of sequencing bias, as a family's presence in highly sequenced taxa allows its shared presence in closely related pairs to be observed. While the effects of distance and species are controlled for in the logistic regression, the highly sequenced taxa can still inflate the estimates. For example, IS1380 has one of the highest coefficients in the regression, but >80% of the pairs that share IS1380 elements are within closely related genomes in the *Acetobacter pasteurianus* or *Streptococcus pneumonia* species. If those species had not been highly sequenced the tendency would not have been discovered. However, other families, such as IS256 and IS200/605, are found to have shared pairs in a wider variety of taxa. OASIS might also be capable of annotating some families more accurately than others, as its accuracy is affected by the properties such as the presence of IRs and the level of conservation between copies. The various reasons that IS families have different rates of being shared across closely related genomes deserves further study.

CONCLUSIONS

As the sequencing technology progresses, the need for user-friendly, high-throughput annotation systems continues to grow. We developed OASIS, an automated annotation system for ISs, which is capable not only of providing detailed IS information for a single genome, but also of annotating thousands of genomes within hours. A tradeoff inherent to computationally efficient automated annotation at this scale is that, although OASIS fares better than previous software platforms, some ISs present in the manually curated ISfinder database were missed. In developing our algorithms we erred on the side of caution, trying to minimize false positives so that these would not be further propagated.

By applying OASIS for high-throughput IS annotation across genomes, this study examined IS evolution at both broad and narrow phylogenetic scales. Looking across all taxa, we revealed a nearly global distribution of ISs across both Bacteria and Archaea and a non-randomness of IS family distribution across taxa. Interestingly, with the fuller OASIS^{+} data set it becomes clear that, although small genomes <2 Mb have relatively few ISs, beyond this size the density of ISs per genome size is relatively constant. The clearest finding of ISs versus genome size; however, is just how large the variation is for any given genome size, in accord with the variance previously noted within a single species (31). At a finer scale of closely related genomes, we found that the probability of an IS being present drops precipitously, indicating either rapid loss and/or emergence of increasing barriers to successful exchange across diverging lineages. We also found, to our surprise, that for a given 16S divergence, strains considered to be of the same species have greater IS content similarity than strains that are named as different species. This perhaps suggests that our current bacterial taxonomy manages to capture some real differences, at least as far as they relate to differential exchange and survival of ISs conditions and thresholds relevant for IS survival and horizontal transfer. The availability of the OASIS platform will hopefully aid in future work to tease apart the individual contributions of IS loss and exchange that give rise to these global patterns and to explore other biological and evolutionary characteristics of ISs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5, Supplementary Figures 1 and 2 and Supplementary Methods.

FUNDING

Taiwan Ministry of Education [Study Abroad Scholarship to M.-C.L.]; Harvard University Microbial Sciences Initiative [fellowship to D.G.R.]; Harvard University [Harvard Merit Fellowship to M.-C.L. and set-up funds to C.J.M.]. Funding for open access charge: Harvard University.

Conflict of interest statement. None declared.

REFERENCES

1. Chandler,M. (2002) Insertion sequences revisited. *Mobile DNA II.* In: Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (eds), *Mobile DNA II.* American Society for Microbiology, Washington D.C., pp. 305–366.
2. Zhou,F. and Olman,V. (2008) Insertion Sequences show diverse recent activities in Cyanobacteria and Archaea. *BMC Genomics*, **9**, 36.
3. Schaack,S., Gilbert,C. and Feschotte,C. (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.*, **25**, 537–546.
4. Chou,H.-H., Berthet,J. and Marx,C.J. (2009) Fast growth increases the selective advantage of a mutation arising recurrently during evolution under metal limitation. *PLoS Genet.*, **5**, e1000652.
5. Schneider,D., Duperchy,E. and Coursange,E. (2000) Long-term experimental evolution in *Escherichia coli.* IX. Characterization of insertion sequence-mediated mutations and rearrangements. *Genetics*, **156**, 477–488.
6. Chou,H.-H. and Marx,C.J. (2012) Optimization of gene expression through divergent mutational paths. *Cell Rep.*, **1**, 133–140.
7. Lee,M.-C. and Marx,C.J. (2012) Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet.*, **8**, e1002651.
8. Cooper,V.S., Schneider,D., Blot,M. and Lenski,R.E. (2001) Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J. Bacteriol.*, **183**, 2834–2841.
9. 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.

10. Zhong,S., Khodursky,A., Dykhuizen,D.E. and Dean,A.M. (2004) Evolutionary genomics of ecological specialization. *Proc. Natl Acad. Sci. USA*, **101**, 11719–11724.
11. Barrick,J.E., Yu,D.S., Yoon,S.H., Jeong,H., Oh,T.K., Schneider,D., Lenski,R.E. and Kim,J.F. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature*, **461**, 1243–1247.
12. Chou,H.-H., Chiu,H.-C., Delaney,N.F., Segrè,D. and Marx,C.J. (2011) Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science*, **332**, 1190–1192.
13. Siguier,P., Perochon,J., Lestrade,L., Mahillon,J. and Chandler,M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.*, **34**, D32–D36.
14. Varani,A.M., Siguier,P., Gourbeyre,E., Charneau,V. and Chandler,M. (2011) ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol.*, **12**, R30.
15. Wagner,A. and Lewis,C. (2007) A survey of bacterial insertion sequences using IScan. *Nucleic Acids Res.*, **35**, 5284–5293.
16. Wagner,A. (2008) Distant horizontal gene transfer is rare for multiple families of prokaryotic insertion sequences. *Mol. Genet. Genomics*, **280**, 397–408.
17. Vuilleumier,S., Chistoserdova,L., Lee,M.-C., Bringel,F., Lajus,A., Zhou,Y., Gourion,B., Barbe,V., Chang,J., Cruveiller,S. *et al.* (2009) *Methylobacterium* genome sequences: a reference blueprint to investigate microbial metabolism of C1 compounds from natural and industrial sources. *PloS One*, **4**, e5584.
18. Cole,J.R., Wang,Q., Cardenas,E., Fish,J., Chai,B., Farris,R.J., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Marsh,T., Garrity,G.M. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.
19. Larkin,M., Blackshields,G. and Brown,N. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
20. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
21. Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, **39**, W475–W478.
22. Nightingale,K.K., Lyles,K., Ayodele,M., Jalan,P., Nielsen,R. and Wiedmann,M. (2006) Novel method to identify source-associated phylogenetic clustering shows that *Listeria monocytogenes* includes niche-adapted clonal groups with distinct ecological preferences. *J. Clin. Microbiol.*, **44**, 3742–3751.
23. Stackebrandt,E. (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Bacteriol.*, **44**, 846–849.
24. Navarro,G. (2001) A guided tour to approximate string matching. *ACM Comput. Surv.*, **33**, 31–88.
25. Wagner,A. (2009) Transposable elements as genomic diseases. *Mol. BioSystems*, **5**, 32–35.
26. Touchon,M. and Rocha,E.P.C. (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.*, **24**, 969–981.
27. Sakaki,Y., Shigenobu,S., Watanabe,H., Hattori,M. and Ishikawa,H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids: *Buchnera* sp. APS. *Nature*, **407**, 81–86.
28. Andersson,S.G. and Kurland,C.G. (1998) Reductive evolution of resident genomes. *Trends Microbiol.*, **6**, 263–268.
29. Luo,H., Friedman,R., Tang,J. and Hughes,A.L. (2011) Genome reduction by deletion of paralogs in the marine cyanobacterium *Prochlorococcus*. *Mol. Biol. Evol.*, **28**, 2751–2760.
30. van Ham,R.C.H.J., Kamerbeek,J., Palacios,C., Rausell,C., Abascal,F., Bastolla,U., Fernández,J.M., Jiménez,L., Postigo,M., Silva,F.J. *et al.* (2003) Reductive genome evolution in *Buchnera aphidicola*. *Proc. Natl Acad. Sci. USA*, **100**, 581–586.
31. Moran,N. and Plague,G. (2004) Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.*, **14**, 627–633.
32. Ochman,H. and Davalos,L.M. (2006) The nature and dynamics of bacterial genomes. *Science*, **311**, 1730–1733.
33. Wagner,A. (2006) Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Mol. Biol. Evol.*, **23**, 723–733.
34. Lawrence,J.G., Ochman,H. and Hartl,D.L. (1992) The evolution of insertion sequences within enteric bacteria. *Genetics*, **131**, 9–20.
35. Pupo,G.M., Lan,R. and Reeves,P.R. (2000) Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc. Natl Acad. Sci. USA*, **97**, 10567–10572.