*Bioinformatics*, 2017, 1–8
doi: 10.1093/bioinformatics/btx433
Advance Access Publication Date: 6 July 2017
Original Paper

OXFORD

Genome analysis

# ISEScan: automated identification of insertion sequence elements in prokaryotic genomes

**Zhiqun Xie and Haixu Tang***

School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on November 1, 2016; revised on June 20, 2017; editorial decision on July 3, 2017; accepted on July 4, 2017

## Abstract

**Motivation:** The insertion sequence (IS) elements are the smallest but most abundant autonomous transposable elements in prokaryotic genomes, which play a key role in prokaryotic genome organization and evolution. With the fast growing genomic data, it is becoming increasingly critical for biology researchers to be able to accurately and automatically annotate ISs in prokaryotic genome sequences. The available automatic IS annotation systems are either providing only incomplete IS annotation or relying on the availability of existing genome annotations. Here, we present a new IS elements annotation pipeline to address these issues.
**Results:** ISEScan is a highly sensitive software pipeline based on profile hidden Markov models constructed from manually curated IS elements. ISEScan performs better than existing IS annotation systems when tested on prokaryotic genomes with curated annotations of IS elements. Applying it to 2784 prokaryotic genomes, we report the global distribution of IS families across taxonomic clades in Archaea and Bacteria.
**Availability and implementation:** ISEScan is implemented in Python and released as an open source software at https://github.com/xiezhq/ISEScan.
**Contact:** hatang@indiana.edu
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Insertion sequence (IS) elements are the smallest and most abundant autonomous transposable elements (TEs). ISs are short DNA segments ranging from 400 to 10 000 bps, which generally encode a transposase (Tpase) that catalyzes the intra-genome or inter-genomes mobility of the IS elements. The short imperfect terminal inverted repeat (IR) sequences are carried by many ISs at their ends. ISs are distributed in wide taxa and can occur in very high numbers in some prokaryotic genomes. Tpases (or proteins with related functions) are annotated as the most abundant functional class of proteins in both prokaryotic and eukaryotic genomic/metagenomic sequences in public databases (Aziz *et al.*, 2010). ISs play a key role in prokaryotic genome organization and evolution (Siguier *et al.*, 2014). The IS-mediated genome rearrangement and reduction through DNA fragment insertion and deletion can promote pathogen evolution (Larsson *et al.*, 2009; Salzberg *et al.*, 2008; Song *et al.*, 2010). More interestingly, IS elements may contribute to virulence loss in bacteria pathogens. For instances, Katherine et al. reported that IS-mediated genome rearrangement could lead to the loss of virulence in a bacteria fish pathogen (Tanaka *et al.*, 2013). Kearney and Staskawicz (1990) reported that IS element insertion could disrupt the virulence genes in pathogen and then increased the host range by abolishing virulence gene triggered immunity. More details about features of ISs can be found in recent reviews (Mahillon and Chandler, 1998; Siguier *et al.*, 2014, 2015). With the fast growing genomic data, it is becoming increasingly critical for research community to be able to accurately and automatically annotate ISs in genomic sequences.

ISfinder is a human curated database for prokaryotic IS elements (Siguier *et al.*, 2006), representing the most comprehensive resource for ISs up to today, which can be searched by Blast tools on their web site (https://www-is.biotoul.fr). ACLAME is a database

dedicated to the collection and classification of TEs from various sources (Leplae *et al.*, 2010), which can be used as Tpase template database with the availability of the freely downloaded Tpase sequences. This database can be used as a starting point to annotate similar or novel IS elements by using homology search-based computational methods.

The increasing availability of complete prokaryotic genomes offers the opportunity to study ISs comprehensively, but the DNA features of IS are often not annotated or incorrectly annotated in publicly available genome sequences. Therefore, efficient and accurate pipelines are in high demand to automatically identify and annotate massive genome sequences accumulated from high throughput genome sequencing platforms. Several computational methods have been developed to annotate IS elements in prokaryotic genomes *in silico*. One of the first publicly available tools to annotate IS elements is IScan (Wagner *et al.*, 2007). It uses tblastn to align one representative member of each IS family in ISfinder to an input genome sequence, in an attempt to identify the open reading frames (ORFs) encoding Tpases in the genome and then extend the identified ORFs to obtain the full-length IS elements with IRs. Zhou et al. used a similar BLAST search-based method to conduct a genome-scale annotation of recently active IS elements in complete cyanobacterial and archaeal genomes (Zhou *et al.*, 2008). Their method focused on the multiple-copy IS elements, which may lead to incomplete IS annotations because IS elements with low activity may occur in single-copies in prokaryotic genomes. ISsaga is a BLAST search-based web application for semi-automatic annotation of ISs in prokaryotic genomes (Varani *et al.*, 2011). Although ISsaga can accurately determine the boundaries of IS elements through manual inspection, it is not practical for large-scale genome analysis because it requires manual submission of genome sequences to a web site, followed by manual inspection of each putative novel IS element (that are not similar to any element in ISfinder). TnpPred (Riadi *et al.*, 2012) is another web service for IS annotation. The advantage of TnpPred is the use of profile Hidden Markov Models (pHMMs) instead of BLAST to search for remote homologs of known IS elements as it is known that pHMM searching is more sensitive than pairwise sequence alignment (e.g. by using BLAST), and was adopted for the identification of eukaryotic TEs (Lee *et al.*, 2016; Rho and Tang, 2009). However, the software is built upon only 19 pHMMs, each representing Tpase sequences from one major IS family in ISfinder. We observed the IS elements from the same family demonstrated high nucleotide sequence diversity. The pHMMs built from diverse protein sequences may decrease the sensitivity of Tpases detection, leading to missing IS elements in annotation. Furthermore, TnpPred reports only the annotated ORFs of Tpases, and does not output the full-length IS elements. A recent paper presented a pipeline combining the de novo and pHMM searching methods for annotating IS elements (Kamoun *et al.*, 2013). The pipeline was applied to 30 archaeal and 30 bacterial genomes and it was shown that the sensitivity of pHMM searching allowed a more reliable detection of ISs in prokaryotic genomes. The pipeline uses the similar strategy as TnpPred, based on 28 pHMMs constructed from major IS families. Unfortunately, the pipeline also stops at the stage of Tpase annotation, and the software is not publically available for large-scale prokaryotic genome annotation. To the best of our knowledge, among all published tools for automated identification of IS elements in prokaryotic genomes, only two recently released tools are publicly available for local installation and ready for use in large-scale genome annotation: OASIS (Robinson *et al.*, 2012) and ISQuest (Biswas *et al.*, 2015), even though OASIS has not been updated since 2012 while ISQuest focuses on the annotation of fragmental IS elements in unassembled next-generation sequencing data. Both OASIS and ISQuest require the annotated prokaryotic genome sequences downloaded from GenBank, relying on the gene (of Tpases) annotation in GenBank, which may lead to missing ISs in prokaryotic genome annotation. Among all these tools, it appears OASIS is the most popular tool for IS element identification in prokaryotic genomes, and thus was used in benchmarking comparison in this paper.

To improve the performance of the currently available tools in public domain and explore the global distribution of IS families across taxonomic clades in Archaea and Bacteria, we developed the new software pipeline, ISEScan for highly sensitive and automated annotation of full-length IS elements in prokaryotic genome sequences. ISEScan is capable of identifying novel IS element with sequence divergent from the known ones in the database and achieves higher sensitivity comparing existing tools (e.g. OASIS), because of several strategies it uses: (i) it benefits from the 621 pHMMs for searching for remote homology Tpases, which are built from phylogenetically classified Tpase sequences, much more comprehensive than those (<30 models) adopted by other pHMM-based annotation tools; (ii) it is not constrained by the existing annotation of Tpase genes in public genome database, enabling it to identify novel IS elements with divergent sequences from those annotated by using sequence similarity searching methods; and (iii) it searches for IRs by comparing the upstream and downstream complementary sequences flanking the ORFs of Tpases in the input genome sequence without relying on the sequence similarity between the putative IR sequences and the known IR sequences included in ISfinder. ISEScan is implemented in Python and released as an open source software at https://github.com/xiezhq/ISEScan.

## 2 Materials and methods

Before searching for IS elements in prokaryotic genomes, we first build a library of pHMMs based on the Tpase sequences in the curated ACLAME dataset, which is assembled by combining Tpase sequences in ACLAME and an additional 66 Tpase sequences collected from the NCBI protein database. It is not necessary to rebuild the pHMMs when performing IS element annotation in prokaryotic genomes each time though they can be rebuilt and updated if needed. Our ISEScan pipeline consists of the following steps: (i) predicting protein coding genes in the input genome sequence and translating them into protein sequences, (ii) identifying putative Tpases by searching the predicted proteins against the library of pHMMs, (iii) extending the putative Tpase genes into full-length IS elements by locating IRs in the upstream and downstream regions and (iv) refining and reporting the final set of annotated IS elements in the input genome. The overall schematic pipeline workflow is shown in Figure 1. Below we will describe each of these steps in details.

### 2.1 Building the library of pHMMs from Tpases in curated ACLAME dataset

We collected 122 154 protein sequences from ACLAME (version 0.4) and retrieved 3776 Tpases from them. We then collected additional 66 Tpases (see the sheet of 'Tpases_from_NCBI' in the supplementary excel file) from NCBI protein database and combined them with the 3776 Tpases from ACLAME to form the curated ACLAME dataset (see Curated ACLAME dataset in Supplementary Material for details). Based on the protein sequence identity, the Tpases in the curated ACLAME dataset were first hierarchically

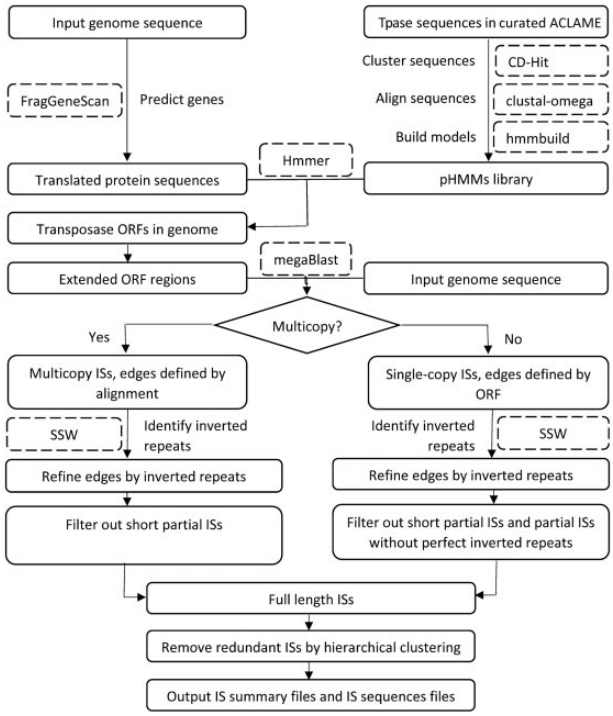


**Fig. 1.** Flowchart illustrating the workflow of ISEScan

clustered with the incremental sequence identity thresholds of 90, 60 and 30%, respectively, using CD-hit (Fu *et al.*, 2012; Li and Godzik, 2006) after the redundant Tpases with sequence identity of 100% were removed from the dataset. From the whole set of Tpases in the curated ACLAME dataset, a total of 621 clusters were obtained. One representative Tpase was selected from each cluster and Blastp was used to search against the ISfinder database. If there were matches with an *E*-value < e-10, the Tpase cluster is classified according to the IS family of the best hit, otherwise the Tpase cluster is classified as 'new'. The protein sequences in each cluster were then aligned by using clustal-omega-1.2.1 (Sievers *et al.*, 2011), which is the improved version of the ClustalW (Larkin *et al.*, 2007), and the resulting multiple alignments were provided as input to hmmbuild in hmmer-3.1b2 package (Eddy, 2011) to construct a pHMM for each cluster. For 355 clusters containing only one Tpase sequence, multiple alignment was not performed and the single protein sequence was simply retained in phmmer search.

## 2.2 Searching for ORFs of Tpases in the input genome sequence

The protein sequence comparison is more sensitive than the nucleic acid sequence comparison when searching for remote homologs. ISEScan uses FragGeneScan (Rho *et al.*, 2010) to first predict protein-coding ORFs in the input genome sequence and translate all predicted ORFs into protein sequences, which are assembled into a protein sequence database subject to the homology search of Tpases. Notably, ISEScan utilizes the frameshift option in FragGeneScan that allows for the detection of IS elements containing recently deactivated Tpase pseudogenes in the input genomes.

ISEScan then uses hmmsearch in the hmmer3 package to query pHMMs built from protein clusters against the protein sequence database assembled from input genome sequence. For the clusters containing only one protein after CD-hit clustering, ISEScan uses phmmer (in hmmer3 package) instead of hmmsearch to query the sole protein sequence in the cluster against the same protein database. ISEScan searches all 621 Tpase clusters (including 355 clusters containing only one protein) against the protein database and reports the hits (putative Tpase) with *E*-values ≤ e-10. If multiple clusters hit the same protein sequence (the same ORF in the input genome), only the most significant hit (with the lowest *E*-value) is retained as the ORFs of the Tpases in the genome. Finally, ISEScan reports a list of ORFs of the Tpases with their coordinates in the genome sequence, and each Tpase will be extended to full-length IS elements by identifying either the multiple copies of the same IS elements in the input genome sequence or the IRs flanking the identified ORFs of Tpases in the genome sequence.

Based on the observation that IS200/IS605 family members in ISfinder include either only a short Tpase or only a long accessory gene or both and the length of the accessory gene is usually two times greater than the length of the Tpase gene, two neighboring Tpase ORFs with distance less than 100 bp are assigned to the same IS200/IS605 element if both are annotated as the IS200/IS605 family and one ORF is two times longer than the other one, before ORFs of Tpases are extended to full-length IS elements.

## 2.3 Extending ORFs of Tpase to full-length IS elements

ISEScan uses different strategies to identify full-length IS elements with multiple copies and those with only a single copy in the input genome. It first searches for Tpase ORFs and then determine the boundaries of each IS element depending on whether multiple copies of the same elements are present in the input genome. If multiple copies are found, the boundaries of the full-length IS elements are determined by IRs and the aligned regions among the multiple copies. If only one copy is found, the boundaries of the full-length IS element is determined solely based on IRs. To obtain the full-length IS elements with multiple-copies, ISEScan uses the workflow as following: (i) to determine genomic locations of each Tpase ORF through profile HMM search; (ii) to extend each Tpase ORF to its upstream and downstream sequences with a fixed window size (referred to here as the extended ORF region), where the window size is set to be the maximum length of the IS in the corresponding IS family (Supplementary Table S1) subtracting the length of the Tpase ORF (if the ORF is longer than the maximum length, no extension is performed); (iii) to determine the copy number of each putative IS element in the input genome sequence by searching each extended ORF region against whole input genome sequence by using megaBlast (Morgulis *et al.*, 2008) with disabled low complexity repeat filter, and to retain the conserved alignments (see details described in Conserved alignment definition in Section 2); (iv) depending on the copy number of each IS element, to determine the boundaries of the IS elements through the locations of their IRs: (a) for putative IS element with multiple copies, to identify IRs in the aligned regions within m bp of each edge where m is the length of the longest IR of the known elements in the IS family in ISfinder (see column maxIR in Supplementary Table S1); and (b) for putative IS element with single copy, to identify IRs using double search method in the regions surrounding the Tpase ORF (see details described in Double search method in Section 2); (v) to remove redundant ISs by hierarchical clustering (Müllner, 2013) where ISs overlapped by a fraction (default 50%) of the length of the shorter one with at least one other IS are clustered and then one representative IS with multiple copies or the least *E*-value is selected and retained for each cluster; and (vi) to report all putative full-length ISs after filtering out potential false predictions in new IS elements and defining partial IS elements (see Section 2.6).

To accurately identify IRs for putative IS elements, ISEScan uses an efficient implementation (SSW) of Smith-Waterman (SW) local alignment algorithm (Zhao *et al.*, 2013) to align the terminal sequences of each putative IS element. ISEScan uses the same scoring scheme: 2 for matches, −2 for mismatches, −6 for gap initiation and −2 for gap extensions in the SW alignment for the IRs. The family-specific shortest length thresholds (instead of a uniform length threshold) between 2 and 18 bp (see column minIR in Supplementary Table S1) were also selected for each distinct IS family in ISEScan. The Supplementary Table S1 shows that the lengths of IRs are between 2 and 144 bp.

## 2.4 Characterizing multiple copies of IS elements

To determine the copy number of each putative IS element in the input genome sequence, ISEScan searches each extended ORF region against whole input genome sequence by using megaBlast with disabled low complexity repeat filter, retains the conserved alignments, and then counts the number of conserved alignments, namely the number of IS copies. An alignment is considered as the conserved alignment if it meets the following criteria:

$$identity \geq 90\%$$

$$length \geq L$$

$$coverage \geq \begin{cases} minPep \ \ if \ \ or\ f\ \ Len \geq minPep \\ or\ f\ \ Len \ \ if \ \ or\ f\ \ Len < minPep \end{cases}$$

where *identity* is the sequence identity of the two aligned sequences, *length* is the alignment length, $L$ is the family-specific value which is the length of the shortest one of known elements in the IS family in ISfinder (see column minIS in Supplementary Table S1), *coverage* is the length of overlap between ORF and alignment, *orfLen* is the length of the predicted Tpase ORF, and *minPep* is the length of the shortest Tpase ORF of known elements in the family in ISfinder (see the column minPEP in the Supplementary Table S1). In this study, unless otherwise stated, multiple copies of IS element include both partial and full-length copies of the IS element containing the same Tpase. ISEScan uses the above criteria to identify both partial and full-length IS copies in eliminating potential false predictions and defining partial IS elements before reporting the final results.

## 2.5 Determination of IRs

The double search method is used to determine the IRs of single-copy IS elements. It uses SSW with match score of 2, mismatch penalty of −2, gap penalty of −2 and gap extension penalty of −6 to search for IRs in the *near* and *far* regions, respectively (Fig. 2), and then the two IRs identified in those two regions are compared with each other and the IRs with higher IR score is preferred and retained as the IRs of the IS element. The *near* and *far* regions are defined as in Figure 2. The IR score is calculated empirically by

$$Score = 2 * (irIdCore + irId - nGaps) - irLen$$

where, *irIdCore* is the number of matches in core regions (the region containing at least three consecutive matches in alignment) in the pairwise alignment between two IR sequences, *irId* is the number of matches in the alignment, *nGaps* is the number of gaps and *irLen* is the length of alignment.

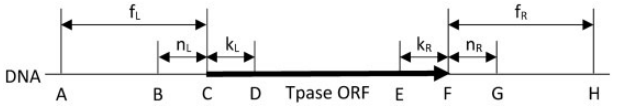


**Fig. 2.** Searching of IRs for single-copy IS elements. The horizontal line indicates the input DNA sequence (genome). The thick arrow indicates the transposase ORF predicated by FragGeneScan. The near regions are defined as from B to D and E to G, and the far regions are defined as from A to D and E to H. Both near and far regions are searched for IRs. The identified pairs of IRs in each of these two regions are compared with each other and the pair with higher score (see text for details) is reported as the IRs of the IS element. The lengths of $n_L$ (or $n_R$), $f_L$ (or $f_R$) and $k_L$ (or $k_R$) are set to 150, 500 and 150 bp, respectively, in ISEScan by default

## 2.6 Post-processing to eliminate potential false predictions in new IS elements and to define partial IS elements

ISEScan eliminates the potential false predictions in 'new' IS elements if they satisfy one of the two criteria: (i) single-copy IS elements with *E*-value < threshold4evalue (default e-50) or without IRs or with IRs containing gaps or *irId* < threshold4irId1 (default 20) or *irId/irLen* < threshold4sim (default 75%); (ii) multi-copy IS elements with *E*-value < threshold4evalue if either *irId* < threshold4irId2 (default 13) or *irId* < threshold4irId1 but with gapped IRs. Here *irId* and *irLen* are described in Section 2.5.

ISEScan defines the partial IS elements by the following criteria: (i) short IS elements with length shorter than the shortest length of known elements in the IS family (as collected in ISfinder); (ii) single-copy IS elements satisfying one of the following four criteria: (i) *E*-value < threshold4evalue, (ii) irId < threshold4irId2 except the IS200/IS605 family and (iii) gapped IRs with *irId* < threshold4irId1 except the IS200/IS605 family; (3) multi-copy IS elements with E-value < threshold4evalue for IS200/IS605 family or with *irId* < threshold4irId3 (default 10) for families except the families of IS110, IS4, IS5, IS6, ISAS1, ISH3 and ISNCY.

## 2.7 Final output of ISEScan

ISEScan outputs five files for an input prokaryotic genome sequence, including (i) a file in GFF3 format containing the characteristics of all annotated ISs in the input genome, such as the genome sequence ID, the genomic locations of identified IS elements, the IS family and the subgroup (cluster) ID, and the genomic locations of all identified IRs; (ii) a file containing the summary statistics about the distribution of IS families in genome sequence, including the number of IS elements in each family across the whole genome, the fraction of the genome covered by IS elements in each family etc; (iii) a file in FASTA format containing the nucleotide sequences of all identified IS elements; (iv) a file in FASTA format containing the nucleotide sequences of all annotated Tpase ORFs; and (v) a file in FASTA format containing the amino acid sequences of all annotated Tpases.

## 2.8 Phylogenetic classification

SILVA is a comprehensive resource for up-to-date quality-controlled databases of aligned ribosomal RNA (rRNA) gene sequences from the Bacteria, Archaea and Eukaryota domains (Quast *et al.*, 2013). SILVA provides a manually curated taxonomy for all three domains of life, based on representative phylogenetic trees for the small- and large-subunit rRNA genes (Yilmaz *et al.*, 2014). To confidently classify each of 2784 genomes into the specific taxonomic path (composed of six ranks: Domain, Phylum, Class, Order, Family and Genus), the small subunit (SSU) reference database (SSURef)

consisting only high quality, full-length SSU rRNA sequences was used. Specifically, a FASTA file (SILVA_123.1_SSURef_tax_silva_trunc.fasta) and a taxonomy file (taxmap_slv_ssu_ref_123.1.txt) of SSURef database from SILVA Release 123 were downloaded from SILVA online archive (https://www.arb-silva.de/download/archive/) for the phylogenetic analysis. The chromosomal DNA sequences in each genome were BLASTed against the rRNA gene sequence database with both rRNA gene coverage and alignment identity set as 100% to find the exactly matched rRNA gene sequence in SSURef for each genome sequence. The BLAST search produced a list of identifiers of the matched rRNA genes, each of which was mapped to one or more genomes. And then the specific taxonomic path was assigned to each genome by searching for the specific rRNA gene identifier in taxmap_slv_ssu_ref_123.1.txt file from SSURef database, which maps each rRNA gene sequence to a taxonomic path. There is no taxonomic path assigned to a genome if (i) BLAST search did not find any matched rRNA gene in SSURef for the genome; or (ii) the multiple chromosomal DNAs in the genome were found and mapped to multiple taxonomic paths.

## 3 Results

### 3.1 Benchmark testing of ISEScan versus OASIS

The performance of ISEScan was evaluated on two benchmark datasets in comparison with OASIS: (i) an expert curated benchmark dataset from the ISbrowser database (Kichenaradja *et al.*, 2010) released by ISfinder; (ii) the IS element annotations of *Escherichia coli* str. K-12 substr. MG1655 genome (Zhou and Rudd, 2013) as MG1655 is the most well studied model bacterial organism and thus its genome annotation is considered to be accurate.

The ISbrowser benchmark dataset contains the annotation of 73 prokaryotic genomes (as of April 2015) available at http://www-genome.biotoul.fr/ISbrowser.php, which is manually curated by IS experts. Among them, 36 genomes contain full length ISs and were used in benchmark testing, including 53 sequences of which 22 are plasmid sequences). We use IS annotations from the ISbrowser instead of GenBank genome annotations as the benchmark data because GenBank annotation is at the genome scale and IS elements are sometimes incorrectly or incompletely annotated in GenBank (Kichenaradja *et al.*, 2010; Robinson *et al.*, 2012). As our main purpose is to identify full-length IS elements, we removed the partial IS elements from annotations. Note that the boundaries of IS elements may not be accurately determined by either ISEScan or OASIS. We consider an IS element is correctly identified (by ISEScan or OASIS) if the overlap between it and an annotated element in ISbrowser is greater than 50% of the full length of the longer one between the two annotations. The performance may vary slightly when different overlap tolerances are used (see Supplementary Tables S2 and S3 for details), but the general trends remain the same. In the evaluation, both ISEScan and OASIS were executed with their default parameters.

We evaluate the performance of ISEScan and OASIS based on their sensitivity (Sn) and false discovery rate (fdr) using the manually annotated IS elements in ISbrowser dataset as the true elements: $Sn = N_{tp}/N$ and $fdr = (N_h - N_{tp})/N_h$, where $N$ represents the total number of true IS elements (annotated in ISbrowser), and $N_{tp}$ represents the number of true positives, i.e. the true IS elements that are identified by ISEScan or OASIS, and $N_h$ represents the total number of IS elements identified by ISEScan or OASIS.

A total of 858 full-length IS elements were manually annotated in the 36 genomes from the ISbrowser dataset, and was considered to be the true elements in the comparison (i.e. $N = 858$). ISEScan identified a total of 988 IS elements ($N_h = 988$) in these genomes. Among them, 790 are true elements ($N_{tp} = 790$), and 198 are not, resulting in the sensitivity of 92% and the fdr of 20%. In contrast, the sensitivity of OASIS is 66%, while its fdr is 25% (Table 1; see the detailed comparison in Supplementary Table S4). We note that the false positive identification here may not be all false (see the fdr_isbrowser sheet in Supplementary excel file), as some of them may represent novel IS elements or new IS copies that are missed by manual annotation in the genomes by ISbrowser.

To further characterize the performance of ISEScan, we collected all IS elements annotated in MG1655 genome (accession.version number: NC_000913.3) from NCBI reference genome annotation database (Tatusova *et al.*, 2014), which is the high-quality genome annotation database, and EcoGene (Zhou and Rudd, 2013) which is a manually curated resource containing genomic and proteomic information about *E.coli* str. K-12 substr. MG1655. The two datasets are combined into one IS element annotation list, which is used as our second benchmark dataset (see Supplementary Table S5). It contains 49 ISs but only 40 full-length ISs are retained for performance evaluation.

We ran ISEScan and OASIS as described earlier on the MG1655 genome. OASIS obtains the high sensitivity of 95% with a low fdr of 12% because it requires the pre-annotated GenBank files for the input genome and therefore can benefit from the high quality of NCBI genome annotation for *E.coli* MG1655. In contrast, ISEScan reached a sensitivity of 100% while the fdr of 20% did not change comparing with the rate of 20% on the first benchmarking dataset from ISbrowser. ISEScan reported 50 IS elements in MG1655 genome and 10 of them are not annotated as full-length IS elements from NCBI reference genome database and EcoGene database, as indicated in Table 1. To find out whether those 10 ISs denoted as the false discovery in our evaluation are IS element candidates or inaccurate annotations, we manually examined MG1655 genome for annotated genes overlapping with each of the 10 false discovery ISs, and found that all these putative ISs overlap with either unknown Tpase genes or incomplete IS elements or pseudo Tpase genes (see Supplementary Table S6).

### 3.2 Predicted IS elements in prokaryotic genomes by ISEScan

To present the practical applications of ISEScan to prokaryotic genome analysis, we applied it to a large set of 2784 genomes (referred to as the NCBI prokaryotic genomes) available at NCBI (as of September 2014), in attempt to give an overview of IS distribution in prokaryotes at genomic scale. Note that one genome (Vibrio parahaemolyticus O1 K33 CDC K4557) was excluded from our analysis as it contains sequences from multiple species. We also downloaded

**Table 1.** Performance of ISEScan and OASIS on two benchmark datasets, ISbrowser and the *E.coli* genome

| Dataset | Method | Sn (%) | $N_{tp}$ | $N$ | fdr (%) | $N_h$–$N_{tp}$ | $N_h$ |
|---|---|---|---|---|---|---|---|
| ISbrowser | OASIS | 66 | 568 | 858 | 25 | 185 | 753 |
| | ISEScan | 92 | 790 | 858 | 20 | 198 | 988 |
| *E.coli* | OASIS | 95 | 38 | 40 | 12 | 5 | 43 |
| | ISEScan | 100 | 40 | 40 | 20 | 10 | 50 |

*Note*: Sn and fdr are sensitivity and false discovery rate, respectively; $N$ and $N_h$ are the total numbers of ISs annotated by ISbrowser and ISEScan (or OASIS), respectively; $N_{tp}$ is the number of the matched ISs in ISbrowser; $N_h$–$N_{tp}$ is the number of the ISs falsely annotated by ISESCAN (or OASIS).

1305 human microbiome reference genomes from http://hmpdacc.org/catalog/ and applied ISEScan to this set, in attempt to compare the IS distributions in these two groups of genomes.

#### 3.2.1 Distribution of IS elements in NCBI prokaryotic genomes

A total number of 78 991 IS elements representing 28 IS families (27 ISfinder families and the non-classified group denoted by new, see Section 2.1 in Section 2 for details) were identified by ISEScan in 2259 out of 2784 genomes analyzed in this study. The abundances of the identified IS elements in these genomes vary substantially across the 28 families (Supplementary Fig. S1). There are ∼44% more elements identified in the most abundant family (IS3) as compared with the second most abundant family (IS5). As the least abundant family, ISKRA4 has 47 identified elements (i.e. ∼0.4% of that of IS3).

To explore the distribution pattern of ISs across all taxonomic clades, we performed the taxonomic analysis on 2784 genomes based on SILVA rRNA gene database (Quast *et al.*, 2013). The taxonomic paths were successfully assigned to 2736 out of 2784 genomes (see Section 2 for details). In total, 2736 genomes were classified into 723 genera from Archaea and Bacteria domains, including 158 assigned as Archaea and 2578 assigned as Bacteria. The complete tabuilation of IS families and the assigned taxonomic path per genome is available in the sheet of 'complete tabulation of IS, taxa' in the supplementary excel file. Out of 78 991 IS element copies, 78 597 (99%) appeared in 2227 out of 2736 genomes across Archaea and Bacteria (Table 2) while only 394 (0.50%) appeared in 32 genomes without taxonomic assignment. The IS elements appeared in 2227 genomes and 623 genera in the CORE dataset (Table 2), which indicated a pervasive IS distribution across archaeal and bacterial domains. However, IS families differ greatly in both their frequency and diversity of hosts. For example, the ISKRA4 elements appear in as few as 19 (0.68%) genomes and 16 genera while IS3 appears in as many as 1540 (55%) genomes and 426 genera.

To explore the distribution of each IS family in bacterial and archaeal domains, we examined the number of IS copies of each family appearing in different clades and the number of genomes where each family appeared (Table 2). Interestingly, we found that there is some degree of domain specificity and non-random distribution of some IS families. Four families only appeared in bacteria genomes: IS1380, IS30, ISAS1 and ISKRA4 (see highlighted in Table 2). And the most abundant family IS3 was absent in all archaeal genomes except seven while ISH3 was absent in all bacteria genomes except three (see highlighted in Table 2).

#### 3.2.2 Distribution of IS elements in human microbiome reference genomes

A total number of 11 226 IS elements representing 25 IS families were identified by ISEScan in 1112 out of 1305 human microbiome reference genomes analyzed in this study. The abundances of the identified IS elements in these genomes vary more substantially across families than those in NCBI prokaryotic genomes (Table 2). There are ∼58% more elements identified in the most abundant family (IS3) as compared with the second most abundant family (IS200/IS605). As the least abundant family, IS701 has 19 identified elements (i.e. ∼0.95% of that of IS3).

The taxonomic paths were successfully assigned to 1090 out of 1305 genomes. In total, 1090 genomes were classified into 191 genera from Archaea and Bacteria domains, including 2 assigned as Archaea and 1088 assigned as Bacteria. The complete tabulation of IS families and the assigned taxonomic path per genome is reported in the sheet of 'complete tabulation of HMP' in the supplementary excel file. Note that the human microbiome reference genomes are almost exclusively Bacterial except two archaeal species, which implies the absence of some IS families (e.g. ISH3, see highlighted in Table 2) that were observed almost only in archaeal genomes.

Out of 11 226 IS element copies identified in human microbiome reference genomes, 9543 (85%) are identified in 919 out of 1090 genomes (Table 2), while 1683 (15%) are identified in 193 genomes without taxonomic assignment. Overall, the IS elements are identified in 9543 genomes and 182 genera in the CORE dataset (Table 2), indicating a pervasive IS distribution across bacterial species in human microbiome. Comparing with the NCBI prokaryotic genomes, IS families in human microbiome reference genomes also differ greatly in both their frequency and diversity of hosts. For example, the IS701 elements are present in as few as 9 (0.98%) genomes and 7 genera, while IS3 elements are present in as many as 599 (65%) genomes and 143 genera.

## 4 Discussions

With the fast growing genomic data, it is becoming increasingly critical for research community to be able to accurately and automatically annotate ISs in genomic sequences. We developed ISEScan, a pipeline for automated identification of IS in prokaryotic genomes, which is capable of not only providing the detailed IS information without requiring the availability of the pre-annotated genome data like GenBank genome annotation and but also of performing better than previous automatic IS annotation system.

As indicated by both benchmark testing of IS annotations on ISbrowser and *E.coli*, ISEScan demonstrated a significant improvement on sensitivity. In the matter of fdr, it should not be simply considered as annotation errors in many cases as indicated in the fdr_isbrowser sheet (see Supplementary excel file) for ISbrowser annotation and the Supplementary Table S6 for *E.coli* annotation. Due to the requirement of already-annotated Tpase genes identified by the term 'transposase' in the 'product' field of GenBank file (Robinson *et al.*, 2012), OASIS suffers two major disadvantages: (i) it is not able to be used on unannotated genomic sequences, and in particular metagenomic sequences; and (ii) it may miss some IS elements because both IS elements and the Tpase genes are sometimes incorrectly or incompletely annotated in GenBank annotations (Biswas *et al.*, 2015; Robinson *et al.*, 2012). As indicated in Table 1, OASIS's performance heavily relied on the specific input genome, namely, the availability and the quality of GenBank annotations for the input genome. On the other hand, ISEScan is independent of the availability of the third-party annotations such as the GenBank annotation of the input genome, and can reach higher sensitivity than OASIS while retaining similar or better fdrs. As the completed genome sequences grew quickly in recent years, the high-quality GenBank annotations for most genomes, particularly the newly sequenced genomes are not available. Thus ISEScan is more suitable for automated identification of IS elements in prokaryotic genomes.

Despite the significant improvement on sensitivity and the ability to find novel or new IS elements that are not currently annotated, the major limitation of ISEScan is that it depends on the quality of gene prediction of the individual gene prediction program and the completeness of Tpase database that were used to build profile HMMs. Based on our observation, we found many IS elements missed by ISEScan in the benchmark datasets were due to either the inaccurate gene prediction or the incomplete profile HMMs missing

**Table 2.** The distribution of IS families in NCBI prokaryotic genomes and human microbiome reference genomes

| | NCBI prokaryotic genomes | | | | | | | Human microbiome reference genomes | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | nIS | | ngm | | np | ngr | avg | nIS | | ngm | | np | ngr | avg |
| Family | Archaea | Bacteria | Archaea | Bacteria | | | | Archaea | Bacteria | Archaea | Bacteria | | | |
| IS1 | 63 | 2720 | 11 | 150 | 9 | 51 | 17.29 | 0 | 104 | 0 | 29 | 2 | 8 | 3.59 |
| IS110 | 255 | 6418 | 24 | 895 | 23 | 329 | 7.26 | 0 | 492 | 0 | 178 | 6 | 76 | 2.76 |
| IS1182 | 49 | 1951 | 6 | 436 | 15 | 184 | 4.52 | 1 | 460 | 1 | 209 | 7 | 50 | 2.20 |
| **IS1380** | **0** | **1287** | 0 | 174 | 8 | 82 | 7.40 | 0 | 93 | 0 | 53 | 4 | 18 | 1.75 |
| IS1595 | 72 | 942 | 15 | 146 | 10 | 80 | 6.30 | 0 | 54 | 0 | 38 | 3 | 7 | 1.42 |
| IS1634 | 93 | 937 | 13 | 157 | 16 | 102 | 6.06 | 0 | 29 | 0 | 18 | 5 | 9 | 1.61 |
| IS200/IS605 | 748 | 5323 | 89 | 761 | 24 | 244 | 7.14 | 0 | 1261 | 0 | 414 | 6 | 92 | 3.05 |
| IS21 | 21 | 2296 | 2 | 535 | 17 | 231 | 4.31 | 0 | 312 | 0 | 172 | 5 | 58 | 1.81 |
| IS256 | 95 | 4554 | 22 | 723 | 21 | 251 | 6.24 | 0 | 914 | 0 | 359 | 6 | 102 | 2.55 |
| **IS3** | 29 | 13085 | **7** | **1533** | 24 | 426 | 8.52 | 0 | 1996 | 0 | 599 | 7 | 143 | 3.33 |
| **IS30** | **0** | **3059** | 0 | 570 | 13 | 166 | 5.37 | 0 | 872 | 0 | 305 | 7 | 73 | 2.86 |
| IS4 | 182 | 3713 | 22 | 466 | 13 | 192 | 7.98 | 0 | 350 | 0 | 130 | 5 | 33 | 2.69 |
| IS481 | 28 | 2543 | 5 | 484 | 13 | 172 | 5.26 | 0 | 100 | 0 | 58 | 5 | 35 | 1.72 |
| IS5 | 447 | 8794 | 58 | 854 | 22 | 316 | 10.13 | 0 | 457 | 0 | 140 | 5 | 52 | 3.26 |
| IS6 | 177 | 1359 | 33 | 338 | 11 | 104 | 4.14 | 0 | 276 | 0 | 129 | 3 | 22 | 2.14 |
| IS607 | 36 | 155 | 14 | 70 | 5 | 35 | 2.27 | 0 | 47 | 0 | 32 | 2 | 19 | 1.47 |
| IS630 | 215 | 3925 | 39 | 491 | 18 | 221 | 7.81 | 0 | 115 | 0 | 52 | 3 | 27 | 2.21 |
| IS66 | 98 | 2473 | 19 | 444 | 12 | 177 | 5.55 | 0 | 422 | 0 | 198 | 4 | 47 | 2.13 |
| IS701 | 34 | 1246 | 11 | 212 | 12 | 118 | 5.74 | 0 | 19 | 0 | 9 | 3 | 7 | 2.11 |
| IS91 | 1 | 188 | 1 | 130 | 7 | 68 | 1.44 | 0 | 60 | 0 | 48 | 3 | 21 | 1.25 |
| IS982 | 10 | 1271 | 2 | 163 | 11 | 74 | 7.76 | 0 | 230 | 0 | 88 | 4 | 22 | 2.61 |
| **ISAS1** | **0** | **1637** | 0 | 245 | 10 | 91 | 6.68 | 0 | 147 | 0 | 64 | 6 | 16 | 2.30 |
| ISAZO13 | 4 | 51 | 1 | 21 | 7 | 18 | 2.50 | | | | | | | |
| **ISH3** | 453 | 3 | **34** | **3** | 5 | 20 | 12.32 | | | | | | | |
| **ISKRA4** | **0** | **47** | 0 | 19 | 5 | 16 | 2.47 | | | | | | | |
| ISL3 | 16 | 2720 | 7 | 561 | 17 | 178 | 4.82 | 0 | 432 | 0 | 215 | 6 | 56 | 2.01 |
| ISNCY | 144 | 1799 | 30 | 469 | 16 | 148 | 3.89 | 0 | 230 | 0 | 74 | 5 | 23 | 3.11 |
| new | 106 | 725 | 22 | 356 | 19 | 156 | 2.20 | 0 | 70 | 0 | 56 | 5 | 25 | 1.25 |
| CORE | 3376 | 75221 | 116 | 2111 | 33 | 623 | 35.29 | 1 | 9542 | 1 | 918 | 8 | 182 | 10.38 |
| All | | | | | 39 | 723 | | | | | | 8 | 191 | |

*Note*: nIS is the number of IS copies; ngm is the number of genomes; nd is the number of domains; np is the number of phyla; ngr is the number of genera; avg is the average copy number per genome. The 'CORE' row indicates the total number of taxa that contain ISs, while the 'All' row represents the total number of taxa used in the analysis.

new Tpases in the Tpase database. The sensitivity would not improve if the genes currently annotated by FragGeneScan in ISEScan were simply replaced by the genes annotated by other annotation systems. As an experiment, we replaced the genes predicted by FragGeneScan with the genes annotated by NCBI reference genome to evaluate the performance of ISEScan on the ISbrowser dataset, the sensitivity dropped from 92 to 87%. Such performance degradation came from the incomplete gene annotation in the current NCBI gene annotation. Although it is difficult to improve the sensitivity by simply replacing the individual gene annotation tool due to the lack of perfect gene annotation, a practical proposal is to combine multiple complementary gene annotation tools. On the other hand, updating the profile HMMs built from Tpase clusters in Tpase database can also further improve the sensitivity of ISEScan.

By applying ISEScan to the high-throughput IS annotation in 2784 genomes across Archaea and Bacteria, we found the nearly pervasive distribution of IS elements across 623 out of the 723 analyzed genera. However, the occurrences of IS do not appear to be random across taxa. Interestingly, some IS elements accumulated in genomes more quickly than the others (Table 2). It might be driven by the different transposing activity of the different IS families. As indicated by the 'Average copy number per genome' in Table 2, IS1 has ∼17 copies per genome while IS91 and IS607 have only ∼1 and 2 copies per genome, respectively. By examining the existence of each IS family in different taxonomic clades, we found that four families (IS1380, IS30, ISAS1, ISKRA4) only appeared in bacteria genomes and the most abundant family IS3 was absent in most archaeal genomes (Table 2). In a previous study using OASIS on 1737 genomes (Robinson *et al.*, 2012), the same observations were reported for IS1380, IS30 and ISAS1, and for IS3 that was absent in all archaeal genomes except two. But our analysis on ISH3 highlighted a difference. We found ISH3 could spread in bacterial genomes though much preferentially in archaeal genomes while the previous paper reported ISH3 was found only in Archaea (Robinson *et al.*, 2012). To exclude the possibility that ISH3 appearing in bacterial genomes was due to the false positive identifications by ISEScan, we manually examined each ISH3 copy in bacterial genomes and ensured that the ISH3 existence in bacterial genomes was not false (see Discussion in Supplementary Material). Therefore, our new results were due to the improved sensitivity of ISEScan over OASIS, as well as the larger set of genomes used in our analysis. The results in Table 2 also show that each IS family preferentially appears in different clades, suggesting that each IS family has the taxonomic clade specificity in their evolution and survival in the genomes.

Applying ISEScan to the human microbiome reference genomes, we found the IS elements are less abundant in these genomes than in the NCBI prokaryotic genomes as indicated by the average copy number per genome in Table 2. This implies that the IS elements are perhaps less active in human microbiome than in the NCBI microbial genomes (in that many are cultured in labs). The insertion of IS elements often disrupts the integrity of the host genome and thus are likely selective disadvantageous, in particular in relatively isolated community such as human microbiome. Therefore, we hypothesize the low transposition activity of IS elements might be a result of selection to sustain the relative stability of the composition of the species in human microbiome in healthy human host. This hypothesis needs to be tested by the future analysis of IS elements in the metagenomic sequences of human microbial communities sampled from healthy human subjects. One alternative hypothesis to explain the lower abundance of ISs in the human microbiota is the possible reduction of horizontal transfer of IS elements because living in a relatively isolated microbial community may significantly reduce their ability to gain new elements. We will test this hypothesis by comparing the IS content in the free-living and endosymbiont/parasitic species in future studies.

The other global and/or specific patterns can be explored and inferred from the research on IS elements in prokaryotic genomes. Both the biological characteristics of IS elements and the evolution of genomes can be hinted from the results of ISEScan, such as the distribution pattern of IS elements in the specific taxonomic clades and the potential horizontal IS element transfer between two genomes in different taxonomic clades. The ISEScan pipeline offers the community a powerful tool to discover the roles of IS elements on the evolution of prokaryotic genomes.

## Acknowledgements

The authors thank Etienne Nzabarushimana, Sujun Li and Wazim MohammedIsmail for helpful discussions.

## Funding

This research was supported by the National Science Foundation under [grant number DBI-1262588]. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, and in part by the Indiana METACyt Initiative. The Indiana METACyt Initiative at IU was also supported in part by Lilly Endowment, Inc.

*Conflict of Interest*: none declared.

## References

Aziz,R.K. *et al.* (2010) Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.*, **38**, 4207–4217.

Biswas,A. *et al.* (2015) ISQuest: finding insertion sequences in prokaryotic sequence fragment data. *Bioinformatics*, **31**, 3406–3412.

Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol.*, **7**, e1002195.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Kamoun,C. *et al.* (2013) Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. *BMC Genomics*, **14**, 700.

Kearney,B. and Staskawicz,B.J. (1990) Characterization of IS476 and its role in bacterial spot disease of tomato and pepper. *J Bacteriol.*, **172**, 143–148.

Kichenaradja,P. *et al.* (2010) ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. *Nucleic Acids Res.*, **38**(Database issue), D62–D68.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Larsson,P. *et al.* (2009) Molecular evolutionary consequences of niche restriction in Francisella tularensis, a facultative intracellular pathogen. *PLoS Pathog.*, **5**, e1000472.

Lee,H. *et al.* (2016) MGEScan: a Galaxy-based system for identifying retrotransposons in genomes. *Bioinformatics*, **32**, 2502–2504.

Leplae,R. *et al.* (2010) ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.*, **38**(Database issue), D57–D61.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

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

Morgulis,A. *et al.* (2008) Database indexing for production MegaBLAST searches. *Bioinformatics*, **24**, 1757–1764.

Müllner,D. (2013) fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *J. Stat. Softw.*, **53**, 1–18.

Quast,C. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**(Database issue), D590–D596.

Rho,M. and Tang,H. (2009) MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. *Nucleic Acids Res.*, **37**, e143.

Rho,M. *et al.* (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.*, **38**, e191.

Riadi,G. *et al.* (2012) TnpPred: a web service for the robust prediction of prokaryotic transposases. *Comp. Funct. Genomics*, **2012**, 678761.

Robinson,D.G. *et al.* (2012) OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences. *Nucleic Acids Res.*, **40**, e174.

Salzberg,S.L. *et al.* (2008) Genome sequence and rapid evolution of the rice pathogen Xanthomonas oryzae pv. oryzae PXO99A. *BMC Genomics*, **9**, 1–16.

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

Siguier,P. *et al.* (2014) Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.*, **38**, 865–891.

Siguier,P. *et al.* (2015) Everyman's Guide to Bacterial Insertion Sequences. *Microbiol. Spectr.*, **3**, MDNA3-0030-2014.

Siguier,P. *et al.* (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.*, **34**(Database issue), D32–D36.

Song,H. *et al.* (2010) The early stage of bacterial genome-reductive evolution in the host. *PLoS Pathog.*, **6**, e1000922.

Tanaka,K.H. *et al.* (2013) IS-mediated loss of virulence by Aeromonas salmonicida: A tangible piece of an evolutionary puzzle. *Mob. Genet. Elements*, **3**, e23498.

Tatusova,T. *et al.* (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**(Database issue), D553–D559.

Varani,A.M. *et al.* (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.

Wagner,A. *et al.* (2007) A survey of bacterial insertion sequences using IScan. *Nucleic Acids Res*, **35**, 5284–5293.

Yilmaz,P. *et al.* (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.*, **42**(Database issue), D643–D648.

Zhao,M. *et al.* (2013) SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS One*, **8**, e82138.

Zhou,F. *et al.* (2008) Insertion sequences show diverse recent activities in Cyanobacteria and Archaea. *BMC Genomics*, **9**, 36.

Zhou,J. and Rudd,K.E. (2013) EcoGene 3.0. *Nucleic Acids Res.*, **41**(Database issue), D613–D624.