6-2015

ISQuest: Finding Insertion Sequences in Prokaryotic Sequence Fragment Data

Abhishek Biswas
Old Dominion University

David T. Gauthier
Old Dominion University

Desh Ranjan
Old Dominion University

Mohammad Zubair
Old Dominion University

Original Publication Citation

Biswas, A., Gauthier, D. T., Ranjan, D., & Zubair, M. (2015). ISQuest: Finding insertion sequences in prokaryotic sequence fragment data. *Bioinformatics, 31*(21), 3406-3412. doi:10.1093/bioinformatics/btv388

Bioinformatics, 31(21), 2015, 3406–3412
doi: 10.1093/bioinformatics/btv388
Advance Access Publication Date: 27 June 2015
Original Paper

OXFORD

Genome analysis

ISQuest: finding insertion sequences in prokaryotic sequence fragment data

Abhishek Biswas[1,*], David T. Gauthier[2], Desh Ranjan[1] and Mohammad Zubair[1]

[1]Department of Computer Science and [2]Department of Biological Sciences, Old Dominion University, Norfolk, Virginia, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on April 13, 2015; revised on June 12, 2015; accepted on June 20, 2015

Abstract

Motivation: Insertion sequences (ISs) are transposable elements present in most bacterial and archaeal genomes that play an important role in genomic evolution. The increasing availability of sequenced prokaryotic genomes offers the opportunity to study ISs comprehensively, but development of efficient and accurate tools is required for discovery and annotation. Additionally, prokaryotic genomes are frequently deposited as incomplete, or draft stage because of the substantial cost and effort required to finish genome assembly projects. Development of methods to identify IS directly from raw sequence reads or draft genomes are therefore desirable. Software tools such as Optimized Annotation System for Insertion Sequences and IScan currently identify IS elements in completely assembled and annotated genomes; however, to our knowledge no methods have been developed to identify ISs from raw fragment data or partially assembled genomes. We have developed novel methods to solve this computationally challenging problem, and implemented these methods in the software package ISQuest. This software identifies bacterial ISs and their sequence elements—inverted and direct repeats—in raw read data or contigs using flexible search parameters. ISQuest is capable of finding ISs in hundreds of partially assembled genomes within hours, making it a valuable high-throughput tool for a global search of IS elements. We tested ISQuest on simulated read libraries of 3810 complete bacterial genomes and plasmids in GenBank and were capable of detecting 82% of the ISs and transposases annotated in GenBank with 80% sequence identity.

Contact: abiswas@cs.odu.edu

1 Introduction

The ever-increasing number of sequenced bacterial and archaeal genomes provides an opportunity to understand their architecture and evolution. However, as new high-throughput sequencing methods are developed, annotation quickly becomes the bottleneck for genomic research. In addition to open reading frames (ORFs) and regulatory elements, correct annotation of other features such as mobile genetic elements (MGEs) is also essential. These MGEs include bacteriophages, conjugative transposons, integrons, unit transposons, composite transposons and insertion sequences (ISs). Such transposable elements are defined as specific DNA segments that can repeatedly insert into one or more sites in one or more genomes. ISs are transposable elements that are regarded as genomic parasites proliferating in their host and surviving only through horizontal gene transfer (Schaack *et al.*, 2010). ISs play a major role in genome evolution and plasticity, mediating gene transfers and promoting genome duplication, deletion and rearrangement (Frost *et al.*, 2005). Insertion sequences may be abundant in host genomes and are intimately involved in mediating horizontal gene transfer, generation of pseudogenes, genomic rearrangement and alteration of regulatory elements (Frost *et al.*, 2005; Schaack *et al.*, 2010). Experimental evolution in the laboratory has demonstrated that both

transpositions (Chou *et al.*, 2009; Schneider *et al.*, 2000) and rearrangements (Chou and Marx, 2012; Cooper *et al.*, 2001; Dunham *et al.*, 2002; Lee and Marx, 2012; Zhong *et al.*, 2004) can generate beneficial mutations. Prokaryotic DDE transposons (mainly ISs) can move in two different ways, depending on the donor site. Replicative transposons copy their DNA, leaving the parent site intact, while conservative transposons cut themselves out of the donor molecule in order to paste their DNA into the target.

Despite the development of various annotation programs for particular genomic features, some important features such as insertion sequences (ISs), the smallest and simplest autonomous MGEs, remain poorly annotated. In many cases, annotations of these elements include only ORFs and ignore terminal inverted repeats (TIRs), which are an essential feature of their activity in mediating gene rearrangements. Moreover, partial ISs are rarely annotated, leading to the loss of potentially valuable evolutionary information. Another major limitation of current tools is the requirement of a complete annotated genome sequence for IS identification and analysis.

The majority of ISs are between 700 and 3000 bp and possess one or two ORFs that encode transposases or helper proteins. For an IS element with more than one ORF, the first (upstream) ORF encodes a DNA recognition domain, while the second overlapping ORF encodes the catalytic domain. There are two types of IS: ISs carrying TIRs elements; and ISs not carrying TIR elements. A TIR IS element carries a pair of partially conserved 7 to 20 bp inverted repeats at its terminus for cleavage and binding of the transposase. Upon insertion, ISs often generate short directed repeats from 2 to 14 bp immediately outside the IRs (Mahillon and Chandler, 1998). ISs of the non-TIR type do not have discernible conserved inverted repeats.

Metagenomic analysis has revealed that IS transposases are among the most abundant and ubiquitous genes in nature (Aziz *et al.*, 2010). Based on transposase sequence similarities, ISs have been classified in 25 different families that belong to three main classes of enzymes: DDE transposases; serine recombinases and tyrosine recombinases (Mahillon and Chandler, 1998). Another recent classification of ISs categorizes them into 26 families based on transposase homology and overall organization, with some families divided further into groups (Zhou *et al.*, 2008). An IS family can be defined as a collection of elements sharing conserved spacers between key residues, identical genetic organization, similar terminal sequence arrangements and uniform target insertion behavior. However, not all families are so coherent. Consequently, some (e.g. families IS4 and IS5) are divided into subgroups composed of a core of closely related elements that can be linked to other members of the family by weaker but still significant similarities. The naming convention of transposable elements (insertion sequences, transposons, etc.) generally follows the recommendations of Campbell *et al.* (Chumley *et al.*, 1979). However, in some cases a revised system of IS naming is used based on a registry where researchers can request for a new sequence number to define novel mobile elements (Roberts *et al.*, 2008). IS and transposable element abundance in prokaryotes is highly variable (Touchon and Rocha, 2007) but they occupy a substantial fraction of some genomes. For example, 11 and 25% of the genome in *Clostridium difficile* and *Enterococcus faecalis* is composed of mobile elements (Paulsen *et al.*, 2003; Sebaihia *et al.*, 2006). Therefore, it is estimated that an average of up to 10% of bacterial (Mahillon and Chandler, 1998) and archaeal (Filée *et al.*, 2007) genomes are comprised of MGEs.

Current IS-related software tools such as IScan and Optimized Annotation System for Insertion Sequences (OASIS) operate only on complete genomes with fully annotated ORFs. Complete genome assembly of a single strain of bacteria can be time-consuming and costly, due in large part to ambiguities introduced by repetitive elements themselves. Consequently, most publicly available prokaryotic genomes are deposited as incomplete, contig- or scaffold-level assemblies, and IS and other repetitive elements may or may not be present in the deposited sequence. For example, Celera WGS (Myers *et al.*, 2000), a widely used assembly software, commonly moves full or partial IS elements to a 'degenerates' folder that is not frequently deposited as part of the draft genome. Therefore, to perform a global investigation of ISs in unassembled prokaryote genomes, we developed ISQuest, or Insertion Sequence Quest, a computational tool for automated detection of ISs in unassembled or partially assembled genomes. ISQuest takes advantage of widely available transposase annotations to identify candidate IS seed regions and then uses a computationally efficient extension method based on BLAST (Altschul *et al.*, 1990) to grow the seed regions and identify the edges of each IS element. ISQuest is capable of finding MGEs in hundreds of genomes within hours, making it a valuable high-throughput tool for a global search of IS elements. We applied ISQuest to 3810 sequenced bacterial genome and plasmid sequences. Compared with the benchmark of GenBank annotations, ISQuest identified 82% successfully with 80% sequence identity.

2 Related work

The abundance and diversity of MGE elements in prokaryotic genomes poses significant challenges in automated identification and annotation using computational methods. The ISFinder database is currently the most comprehensive dedicated resource for high-quality, manually curated ISs annotations (ISFinder at https://www-is.biotoul.fr/). Therefore, we assume this database to be an accurate set of ISs, but incomplete because genomes are being sequenced faster than they are annotated to this extent. However, several studies have used the referenced sequences in the ISFinder database to mine various collections of genomic data using BLAST-based software (Cerveau, *et al.*, 2011; Filée *et al.*, 2007; Leclercq and Cordaux, 2011; Mahillon and Chandler, 1998; Wagner, 2006).

The development of high-throughput sequencing techniques has led to the availability of thousands of sequenced genomes and metagenomes that require automated identification of ISs. Genome annotation pipelines such as Prokka (Seemann, 2014) and Institute for Genome Sciences (2015) stop at the point of labeling ORFs as 'transposase' or 'integrase' where sufficient homology is observed. Without classification of ISs into families and enumeration within genomes, broad-scale comparison studies across closely related strains are not possible. The first automated approach to annotate ISs 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 (Zhou *et al.*, 2008). ISSaga is a web application pipeline that allows semi-automated IS annotation in complete genomes (Varani *et al.*, 2011). ISSaga employs a library-based method using BLAST seeded with the ISFinder sequences to classify ORFs into IS families. Although ISSaga represents significant progress in automated IS annotation, the efficiency of this approach in identifying transposable elements is questionable due to its dependency on the ISFinder database; ISSaga cannot automatically identify novel ISs not already present in ISFinder. IScan is a publicly available application that makes use of BLAST with a single reference transposase sequence per IS family to scan whole genomes for ISs, and includes in its prediction pipeline searches for transposases and inverted and direct repeats (Wagner *et al.*, 2007). IScan was used to investigate

ISs in 438 prokaryotic genomes and found a limited number of ISs in most taxa (Wagner and de la Chaux, 2008). OASIS is another publicly available computational tool for automated annotation of ISs (Robinson *et al.*, 2012) in whole genomes. 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. Although OASIS is capable of predicting IS families, this functionality seems to be deprecated in the current version of the software. Through comparisons across 1319 genomes to a benchmark of ISFinder annotations, OASIS detected 37 427 ISs while IScan (Wagner *et al.*, 2007) detected only 2902 ISs.

Software tools have also been developed to predict IS sequences and families based on profile-sequence comparisons. These tools employ Hidden Markov Models (HMMs) based on transposases of characterized IS families. HMMs have been generated for transposases belonging to 19 characterized families of ISs in the PFAM database (Finn *et al.*, 2014). The Superfamily database of structural and functional annotation of genomes currently hosts six HMM profiles from domains belonging to two prokaryotic families of transposases: mu bacteriophage transposase and IS*200* (Gough and Chothia, 2002). The TnpPred web service provides profile HMMs for the remaining IS families and improves on the accuracy of the HMMs in the PFAM database (Riadi *et al.*, 2012). Effective prediction of ISs and Miniature Inverted repeat Transposable elements (MITEs) using HMMs has been shown for 30 archaeal genomes (Kamoun *et al.*, 2013), demonstrating that HMM-based predictions can augment BLAST-based sequence–sequence IS search methods to improve accuracy and find novel ISs.

The current software tools described earlier operate only on complete genomes with fully annotated ORFs. Complete genome assembly of a single strain of bacteria can be time-consuming and costly, and draft genomes or raw read sets are increasingly used for comparative genomics studies of prokaryotes. Here, we present the ISQuest tool for global investigation of ISs in unassembled or partially assembled prokaryote genomes.

3 Methods

ISQuest is a computationally efficient algorithm designed to find and annotate Insertion Sequences (IS) and transposases in fully assembled, partially assembled or unassembled genomes. The algorithm uses BLAST (Altschul *et al.*, 1990) to determine potential IS locations by searching against an automatically curated database of IS and transposase sequences derived from GenBank. The potential locations are further extended by Smith–Waterman alignment extension. The IS elements may occur once in a genome (single-copy) or may consist of a set of almost identical copies (multicopy). 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 ISQuest 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. The pipeline has been specially modeled to identify ISs but the algorithm is capable of detecting other MGEs and the generic steps are described later with IS elements as special cases.

3.1 Search terms and transposaseDB

ISQuest identifies single-copy and multicopy ISs and transposases a each genome by finding conserved regions of already-annotated transposase elements, which are identified by the word

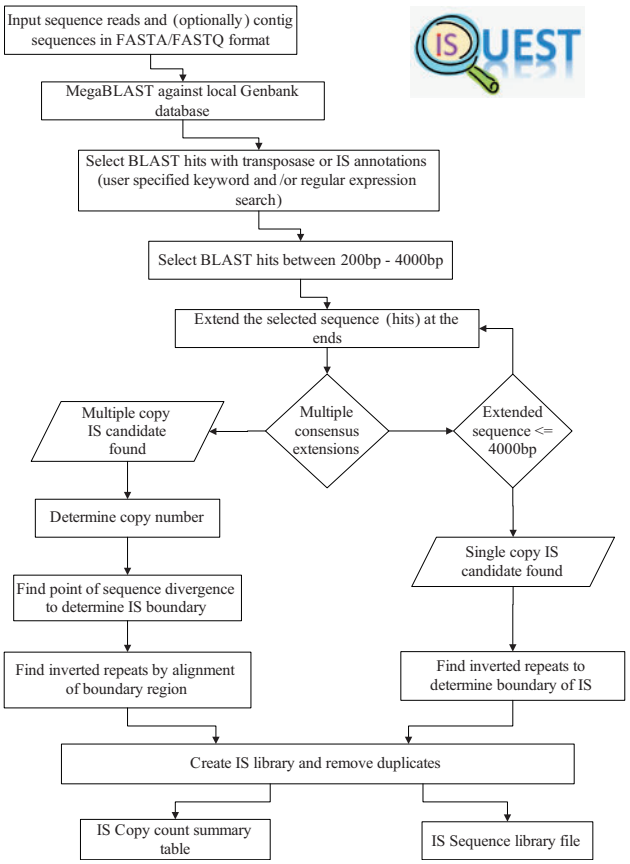

Fig. 1. Flowchart of the full workflow of ISQuest

'transposase', or 'insertion sequence' in the 'product' field of GenBank files. The search keywords may be extended by user-provided regular expressions since there is a significant amount of inconsistently annotated data in GenBank. For example, transposases are frequently mis-annotated as integrases. Generating the database of known MGEs is done once as a preprocessing step during the first run of ISQuest which generates a BLAST database called TransposaseDB. This database is stored for subsequent use by future executions. The user can force updates of the database when new versions of the GenBank files are available.

3.2 BLAST searching parameters

A candidate sequence for extension is determined by a BLAST search against TransposaseDB. ISQuest can operate directly on raw reads provided in FASTA/FASTQ format. Efficiency can be significantly improved by assembling the reads and providing a set of assembled contigs in FASTA format. This assembly can be performed using an appropriate assembler for the input reads. The assembled contigs are BLAST-searched against the TransposaseDB database to find potential seed locations for ISs and transposases. These seed locations represent all possible MGE locations that must be searched and analyzed. Therefore, we use MegaBLAST for finding matches with higher sequence similarity and better performance. Because we further extend these seed sequences to find the boundaries of the MGEs, we can tolerate partial or inexact matches.

3.3 Extending potential IS matches

Once the possible MGE seed locations have been identified, raw reads are used to extend the seed sequences to determine boundaries.

The extension is done by pairwise alignment of the raw reads to the ends of the seed sequence. This alignment algorithm is implemented using BLAST allowing 5 bit score errors. This parameter is configurable by the user depending on the sequencing technology used and the expected error profile of the reads. For Illumina reads we allowed a bit score error of 5, which corresponds to 98% sequence similarity using 250 bp reads.

The extension step aligns all reads to the end of a seed sequence then executes the boundary detection step (next Section 3.4). The extension step does not align reads that do not have at least a partial overlap with the core seed sequence as we do not want to miss the boundary of the MGE by large extensions. Therefore, each extension step builds no more than twice the input read length. The seed sequence is expanded to include the aligned reads and the larger consensus sequence is used as the new seed. Therefore, the extension step is iteratively executed for the remaining sequences for which the boundary cannot be found until the seed sequence becomes too long. The termination length of the seed sequence is user configurable and defaults to 4 kb.

3.4 Determining IS boundary

We apply different approaches to find the boundary of single- and multi-copy MGE elements. In the case of a single-copy, we can only find the boundary in cases where there are flanking inverted repeats. 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 (Wagner *et al.*, 2007). 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.

Because the various copies of a multi-copy ISs are from different genomic loci, they have different unique sequence beyond the boundaries of the IS. Therefore, if the consensus of the aligned reads disagrees with the end of the seed sequence, this indicates that the boundaries of the IS have been reached. Based on the number of possible disagreements we calculate the number of possible sequence groups. If each group has coverage within a specified range we can be certain that we have reached the final boundary for all the sequence groups and have run into the flanking unique sequence. However, if a sequence group has coverage several times that of the expected coverage, we know that there exist longer MGEs the form of tandem repeats which will require further extension. These sequence groups are separated out for extension in the next iteration.

The sequence groups with appropriate coverage are processed to determine the IRs using a Smith–Waterman sequence alignment. The alignment parameters are the same as those described for the single-copy IS case. In some cases, the boundary defined by the IRs may disagree with the boundary defined by the synteny of the aligned reads due to nested repeats, flanking direct repeats at the ends, or inaccurate IR identification. ISQuest addresses this ambiguity by prioritizing the IR edges and changing the boundary to match the IRs. If IRs are found, a direct repeat finding subroutine attempts to align 10 bp fragments on either side of the IRs to identify direct repeats. If no IRs are found, the edges of the MGE are solely determined by the alignment of the reads. This allows annotation of partial MGEs as many of these sequences do not have IRs. Thus, when present in multiple copies, ISquest finds partial ISs; it is not capable of finding these IS fragments when no intact copy with an annotated transposase is present in GenBank.

The same MGE element may result in one or more BLAST seeds and may cause redundant copies of the same IS to be generated. Therefore, the redundant results within the final set are filtered out using a pairwise global alignment to identify groups of IS lengths, which are clustered together. The clustering algorithm groups sequences such that the mean lengths are within 100 bp of each other. The cluster is then assumed to be the true copy size of the IS and any fragments that are shorter than that threshold are classified as partials.

3.5 Iterative extension and boundary finding

Sequences with known boundaries are removed from the extension set and all remaining sequences are expanded based on the consensus of the reads aligned to the boundaries. Extension and boundary finding are performed iteratively until all seed sequences have been processed. The end of each boundary finding step generates a new set of seed sequences. The new seed sequences are generated from the alignments that have no disagreement in the aligned reads, signifying that the boundary has not been reached. The consensus sequences generated from all these alignments is used as the fresh set of seeds in the extension step. Some new seed sequences may be derived from alignments with disagreements as well. In such cases the alignment disagreements can be grouped such that some groups have a very large coverage. The consensus sequences generated from these large coverage groups are separated and treated as new seed sequences.

3.6 ISQuest output

The output of the pipeline is a library of full and partial MGEs. IS elements in particular are composed of a transposase with one or more ORFs and appropriate upstream and downstream sequences. The extreme edges are annotated in GenBank format for IS elements and may include a partially conserved inverted repeat on each end ranging from 8 to 40 bp in length with direct repeats ranging from 4 to 8 bp in length. Partial IS elements and other MGEs such as transposases do not have special annotations defining the boundary.

The final output of ISQuest includes two files for the given input of raw reads and contig(s): (i) a file in GenBank format listing each MGE and its characteristics, including the chromosome ID, start and end positions, direction, family and group, IRs (if found), DRs (if found) and whether the element is a partial element; and (ii) a file containing the copy number of each identified IS in .csv format.

3.7 Using the ISQuest tool

ISQuest is a free open source program implemented in C++. It is available at http://sourceforge.net/projects/isquest. ISQuest requires the read library of input reads in FASTA/FASTQ format and can be optionally provided with an assembly of the reads. The program accepts four command line parameters (a) the configuration file, (b) the raw reads, (c) the prefix of the output files and (d) the optional set of assembled contigs. The configuration file contains the required file paths to the local BLAST database and other configurable parameters such as the maximum number of iterations ISQuest performs, the maximum length of the MGEs to be built and the search terms for MGE's in GenBank. A complete wiki with required documentation is provided on the forge.

3.8 Preparation for ISQuest tool evaluation

To evaluate ISQuest we used 3810 microbial genomes and plasmid sequences >100 kb available in GenBank as of 15th October 2014. The ART tool was used to generate synthetic Illumina paired-end

fragment libraries with read length of 250 bp and 50 × coverage. The read length of 250 bp is typical of Illumina sequencing machines and was selected for experimentation. ART simulates sequencing reads by mimicking real sequencing process with empirical error models or quality profiles summarized from large recalibrated sequencing data. ART can also simulate reads using a user specified error profile which requires the user to specify probability of sequencing errors at each base position of the read. ART was used as a primary tool for the simulation study of the 1000 Genomes Project (Huang *et al.*, 2012). ISQuest performance was evaluated by first fragmenting each genome using the simulation process described earlier. We then used the Celera WGS assembler to assemble these simulated reads into contigs. The ISQuest algorithm was operated on these contig sequences to generate a set of candidate MGEs. This run can be performed using the raw reads but will significantly slow down the execution. Also, we ensure that the ISQuest testing algorithm does not include the genomes being processed in the search database to ensure that the test and training sets are disjoint.

4 Results

We performed two experiments to show the MGE detection capability of ISQuest and present a summary of IS sequences found by ISQuest classified by IS family. The performance of the ISQuest tool was compared with that of OASIS using annotated transposases in GenBank as a benchmark. This first experiment compared the accuracy of ISQuest and OASIS by measuring the percentage of GenBank annotated ISs found by each tool. Unlike ISQuest, OASIS operates on completely assembled and annotated genomes and uses only the annotation information available in the genome. ISQuest operates on partially assembled contigs or directly on the raw reads and does not require annotation to identify the ORFs. This experiment shows the predictive capability of ISQuest to find ISs from a draft and un-annotated assembly and compares it to the predictive capability of OASIS using completely annotated sequences. The capability of ISQuest to find other repetitive elements (e.g. rRNA operons) is not measured in this experiment.

As ISQuest uses an un-annotated draft genome, ORFs are not clearly defined and finding the exact lengths of the MGEs is difficult using the seed extension algorithm. Therefore, due to these inaccuracies, the testing result in Figure 2A considers 70% sequence length match as a true positive; if ISQuest returns a sequence that matches a 70% of the length of an annotated sequence in GenBank with 95% sequence similarity we consider it a true positive. The count numbers in the figure represent IS counts in single-copy; multiple copies of a particular IS are not included. Within the 3810 benchmarked genomes and plasmids, ISQuest found 84.5% of the 9422 unique GenBank annotations, whereas OASIS found 58.9%. The 5346 GenBank ISs found both by ISQuest and OASIS represent insertion sequences with well-defined inverted repeats. The 2558 sequences found by ISQuest and also present in GenBank are full and partial transposase elements that do not contain completely defined inverted repeats and therefore cannot be identified by OASIS. The 1350 annotations found only by ISQuest include partially assembled insertion sequences and partial MGEs found by ISQuest that have not been annotated in deposited genomes. These sequences may also include potential sets of new insertion sequence and transposase elements identified by ISQuest based on sequence similarity to other ISs in GenBank. The intersection of ISQuest and OASIS is zero as ISQuest cannot identify any sequence that has not been annotated in more than one GenBank submission using the keywords 'transposase', or 'insertion sequence' in the 'product' field. ISQuest does not take the annotated genome as input and therefore requires similar annotation to be present in other submissions.

We further evaluated ISQuest under increasingly strict constraints by increasing the length match threshold which we accept as a true positive to 80 and 90% of the sequence length (Fig. 2). Figure 2B shows the results of considering only sequences with greater than or equal to 80% length matches with 95% sequence similarity with GenBank sequences as valid true positives of ISQuest. We notice a slight reduction in the number of insertion sequences detected by ISQuest to 82.2% of the 9422 unique GenBank annotations. Increasing the length match threshold to 90% (Fig. 2C) shows significant reduction in the number of insertion sequences detected by ISQuest to 65.7%. However, this shows that ISQuest is able to reproduce 90% of the actual IS sequence using the fast seed extension algorithm in the majority of cases.

4.1 MGE detection using ISQuest

In order to study the overall sensitivity and specificity of ISQuest we directly compared its output to GenBank. Comparison to OASIS is problematic as OASIS only identifies insertion sequences with clearly defined inverted repeats. ISQuest can identify full ISs, partial ISs and other MGEs such as transposases. Table 1 shows the IS sequences found by ISQuest grouped by phylum. The numbers in the

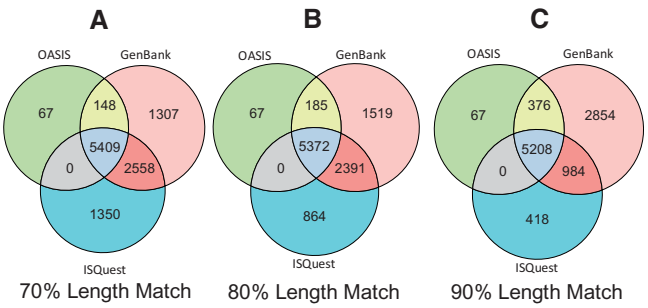

Fig. 2. Venn diagram illustrating the number of IS annotations identified by ISQuest and OASIS compared with GenBank at three length match thresholds. (**A**) ISQuest and OASIS both found a total of 5409 ISs (in single copies) in the 3810 GenBank benchmarked genomes and plasmids. Additionally, ISQuest identified 2558 ISs that OASIS did not annotate and OASIS found 148 ISs that ISQuest failed to detect. OASIS found 67 insertion sequences that were not correctly annotated in GenBank as IS. ISQuest generated 1350 partial IS sequences that have not been annotated in GenBank. The intersection of ISQuest and OASIS is 0 as ISQuest cannot identify any sequence that has not been annotated in more than one GenBank submission using the keywords 'transposase', or 'insertion sequence' in the 'product' field. ISQuest does not take the annotated genome as input and therefore requires similar annotation to be present in other submissions. (**B**) same as (**A**) but only allowing 80% length matches as true positives. (**C**) same as (**A**) but only allowing 90% length matches as true positives

table represent ISs in multiple copies, i.e. the multiple copies of the IS are included (collapsed). Likely because of the number of sequenced genomes from Proteobacteria and Firmicutes, >50% of the ISs we found are from Proteobacteria and an additional ∼16% are from Firmicutes (Table 1, Column 3). ISQuest detected 82.2% of the Proteobacteria ISs and 81.1% on average from GenBank (Table 1, column 3, 5). The prediction capability of ISQuest is limited by the assumption that a similar annotation of the IS element is present in other genomes. So, in some cases we cannot identify certain ISs correctly due to sequence divergence or absence of annotation. Also, the copy number computation based on the number of possible flanking unique sequence regions is conservative in estimating the number of copies and reduces the copy count to the least possible value.

ISQuest was also used to identify transposase elements and the sequences generated by ISQuest without clearly defined inverted repeats were compared with transposase annotations in GenBank. Similar to IS elements, Proteobacteria and Firmicutes account for majority of the transposase annotation is GenBank (52.3 and 18.3%, respectively). ISQuest detected 57.7% of the Proteobacteria transposases and 44.4% of transposases from GenBank (Table 1, column 4, 6). The significantly lower detection accuracy relative to ISs is due to the presence of single-copy transposases. These elements do not possess inverted repeats, and in single-copy cases, do not possess multiple unique flanking sequences; therefore, their length cannot be estimated by ISQuest. Such single-copy elements with no discernable end regions are extended to the default maximum length and often include unique sequence that does not match an existing transposase element from GenBank.

4.2 IS Family detection using ISQuest

It was also interesting to study the performance of ISQuest in terms of the IS families discovered. This provided insight into the annotations and predictive capability of ISQuest for mining ISs from families with high divergence. Table 2 shows the top 20 IS families detected, some of which are predicted better than others due to the inherent divergence in the IS families and inaccurate annotations from GenBank. IS4 family is the most annotated IS family in GenBank with a total of 5521 annotations. ISQuest identified the IS elements in IS4 family with ∼ 60% accuracy which is significantly less that overall accuracy of ISQuest. This is due to the high internal divergence of IS4 elements which makes classification and identification challenging.

Overall, a total of 60 502 MGE elements representing 9317 unique IS sets and 26 767 transposase annotations were identified by ISQuest in 3810 genomes and plasmids. ISQuest took a total of 23 h and 44 min to annotate all 3810 genomes on a 4 x Intel Xenon

Table 1. ISQuest annotations compared with GenBank annotations grouped by Phylum at 80% length match threshold

Phylum	Number of Genomes[a]	Number of GB IS[b]	Number of GB TP[c]	Number of ISQ IS[d]	ISQ TP[e]
Proteobacteria	1 810	22 375	31 918	18 412	14 164
Firmicutes	794	7 906	11 029	6 297	4 962
Actinobacteria	520	4 029	7 970	3 416	3 513
Cyanobacteria	128	1 590	3 674	1 267	1 534
Bacteroidetes	92	1 016	1 342	858	582
Tenericutes	53	434	468	321	226
Spirochaetes	48	357	569	264	253
Deinococcus-Thermus	47	283	323	188	160
Others	318	3 754	3 097	2 712	1 373
Total	3 810	41 564	60 309	33 735	26 767

[a]The number of genomes under each phylum.
[b]The number of IS annotations(multiple copies) in GenBank.
[c]The number of Transposase annotations in (multiple copies) GenBank.
[d]The number of IS detected (multiple copies) detected by ISQuest.
[e]The number of Transposase detected (multiple copies) detected by ISQuest.

Table 2. ISQuest annotations compared with GenBank annotations group by IS Type

IS Fam.[a]	Number of GB[b]	Number of ISQ[c]	Percentage[d]	IS Fam.[e]	Number of GB[b]	Number of ISQ[c]	Percentage[d]
IS4	5 521	3 340	60.5	**IS110**	308	308	100
IS911	2 496	1 872	75	**ISL3**	308	298	96.8
IS902	1 738	1 603	92.2	**IS21**	233	232	99.6
IS3	1 061	1 060	99.9	**IS982**	229	171	74.7
IS5	772	679	88	**IS256**	223	222	99.6
IS66	568	426	75	**IS200**	190	190	100
IS1165	491	367	74.7	**IS1341**	146	146	100
IS605	377	376	99.7	**IS6**	98	98	100
IS30	362	361	99.7	**IS1182**	75	55	73.3
IS630	337	252	74.8	**IS1595**	55	54	98.2

[a]The top 10 IS families annotated in GenBank.
[b]The number of IS annotations (single-copy) in GenBank.
[c]The number of IS detected (single-copy) by ISQuest.
[d]The percentage IS detected (single-copy) by ISQuest.
[e]The top 11–20 IS families annotated in GenBank.

X7550, 2.0-Ghz processor using partially assembled contigs. The maximum per-genome running time was 8 min.

5 Conclusion and future work

As sequencing technology progresses, the need for user-friendly, high-throughput annotation systems continues to grow. We developed ISQuest, an automated annotation system for insertion sequences, which is capable not only of providing detailed IS information for a single genome, but also of processing thousands of genomes within a few hours. The major feature implemented in ISQuest is the capability of detecting ISs and other MGEs using partially assembled sequences or even raw reads. This makes ISQuest a more usable tool over previous such implementations which require a fully assembled and annotated genome. The design of ISQuest can also identify various types of MGEs other than ISs and therefore can be used for many purposes, such as mapping the evolutionary history and analyzing horizontal gene transfer patterns.

We tested ISQuest on simulated read libraries of 3810 complete bacterial genomes and plasmids in GenBank. Of 101 954 IS and transposable elements annotated for these sequences in GenBank, we identified 82 with 80% sequence length match. ISQuest is capable of identifying a large number of MGE elements from unassembled genomes with acceptable sequence accuracy to be used in comparative genomics and assembly verification. ISQuest can be used for many purposes, such as mapping the evolutionary history, comparing IS structure among divergent strains and horizontal gene transfer patterns of different ISs. The ISQuest tool can also have interesting application is metagenomics analysis. Therefore, the future versions of ISQuest tool will be extended to handle metagenomic datasets and tested with metagenomic raw reads.

Conflict of Interest: The work in this paper is supported by the Jeffress Trust Awards Program in Interdisciplinary Research and the M&S fellowship fund of the Old Dominion University.

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

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

Cerveau,N. *et al.* (2011) Short- and long-term evolutionary dynamics of bacterial insertion sequences: insights from wolbachia endosymbionts. *Genome Biol. Evol.*, **3**, 1175–1186.

Chou,H.-H. and Marx,C.J. (2012) Optimization of gene expression through divergent mutational paths. *Cell Rep.*, **1**, 133–140.

Chou,H.-H. *et al.* (2009) Fast growth increases the selective advantage of a mutation arising recurrently during evolution under metal limitation. *PLoS Genet*, **5**, e1000652.

Chumley,F.G. *et al.* (1979) Hfr formation directed by Tn10. *Genetics*, **91**, 639–655.

Cooper,V.S. *et al.* (2001) Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of Escherichia coli B. *J. Bacteriol.*, **183**, 2834–2841.

Dunham,M.J. *et al.* (2002) Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. USA*, **99**, 16144–16149.

Filée,J. *et al.* (2007) Insertion sequence diversity in archaea. *Microbiol. Mol. Biol. Rev.*, **71**, 121–157.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Frost,L.S. *et al.* (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Micro.*, **3**, 722–732.

Gough,J. and Chothia,C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, **30**, 268–272.

Huang,W. *et al.* (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Institute for Genome Sciences (2015) Manatee: Web-based tool used to perform manual functional annotation.

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

Leclercq,S. and Cordaux,R. (2011) Do phages efficiently shuttle transposable elements among prokaryotes? *Evolution*, **65**, 3327–3331.

Lee,M.-C. and Marx,C.J. (2012) Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet*, **8**, e1002651.

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

Myers,E.W. *et al.* (2000) A whole-genome assembly of drosophila. *Science*, **287**, 2196–2204.

Paulsen,I.T. *et al.* (2003) Role of mobile DNA in the evolution of vancomycin-resistant enterococcus faecalis. *Science*, **299**, 2071–2074.

Riadi,G. *et al.* (2012) TnpPred: a web service for the robust prediction of prokaryotic transposases. *Comp. Funct. Genomics*, **2012**, 5.

Roberts,A. *et al.* (2008) Revised nomenclature for transposable genetic elements. *Plasmid*, **60**, 167–173.

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

Schaack,S. *et al.* (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.*, **25**, 537–546.

Schneider,D. *et al.* (2000) Long-term experimental evolution in Escherichia coli. IX. Characterization of insertion sequence-mediated mutations and rearrangements. *Genetics*, **156**, 477–488.

Sebaihia,M. *et al.* (2006) The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nat. Genet.*, **38**, 779–786.

Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

Touchon,M. and Rocha,E.P.C. (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.*, **24**, 969–981.

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

Wagner,A. and de la Chaux,N. (2008) Distant horizontal gene transfer is rare for multiple families of prokaryotic insertion sequences. *Mol. Genet. Genomics*, **280**, 397–408.

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

Zhong,S. *et al.* (2004) Evolutionary genomics of ecological specialization. *Proc. Natl. Acad. Sci. USA*, **101**, 11719–11724.

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