Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities

Yogita Mantri and Kelly P. Williams*

Department of Biology, Indiana University, 1001 E. Third Street, Bloomington, IN 47405, USA

Received August 15, 2003; Revised September 16, 2003; Accepted September 24, 2003

ABSTRACT

Prokaryotic chromosomes often contain islands, such as temperate phages or pathogenicity islands, delivered by site-specific integrases. Integration usually occurs within a tRNA or tmRNA gene, splitting the gene, yet sequences within the island restore the disrupted gene. The regenerated RNA gene and the displaced fragment of that gene thus mark the endpoints of the island. We applied this principle to search for islands in genomic DNA sequences. Our algorithm generates a list of tRNA and tmRNA genes, uses each as the query for a BLAST search of the starting DNA and removes unlikely hits through a series of filters. A search for islands in 106 whole bacterial genomes produced 143 candidates, with the search itself providing an estimate of three false candidates among these. Preliminary phylogenetic analysis of the associated integrases reduced this set to 89 cases of independently evolved site specificity, which showed strong bias for the tmRNA gene. The website Islander (http://www.indiana.edu/~islander) presents the candidate islands in GenBank-style files and correlates integrase phylogeny with site specificity.

INTRODUCTION

The whole genome sequences for two closely related archaeal or bacterial strains can show near identity throughout most of the genome, yet show dramatic differences in the presence or absence of multigene DNA segments termed islands (1). In addition to sporadic occurrence, islands often exhibit other features of mobile genetic elements that indicate their acquisition by horizontal transfer; many contain a gene encoding an integrase of the tyrosine recombinase family that is responsible for the site-specific positioning of the island. By efficient horizontal delivery of a gene or set of genes that benefits the bacterial host, perhaps promoting pathogenicity, a catabolic pathway or other physiological process, islands are major agents of bacterial evolution. Temperate bacteriophages, which may likewise carry genes beneficial to their bacterial host, can also be considered integrative islands. Site specificity has changed frequently during integrase evolution, which allows diverse islands to accumulate combinatorially in a given host genome. Understanding how site specificity evolves among integrases is therefore a key question in bacterial evolution. We sought to expand the number of known pairs of integrases and the sites they specify, through a bioinformatic search among whole genomes, and in the process have found the endpoints of several previously unrecognized islands. The data from this search, and for other islands from the literature with a known integrase and integration site, are presented at the Islander website (www.indiana.edu/~islander).

SEARCH STRATEGY

By analogy with well-studied integration systems such as that of phage λ (2), it can be presumed that integrative islands exist in circular DNA form prior to integration, and that the integrase catalyzes recombination between a site (attP) in the circular pre-island and the target site (attB) in the chromosome. Most integrases specify an attB that lies within a tRNA or tmRNA gene (tDNA) (3,4). When the island integrates it splits the target tDNA, yet the gene is restored because identical sequence in the attP of the island replaces the fragment of the original gene that was displaced. Thus the regenerated tRNA gene and its displaced fragment mark the endpoints of the island. We used this principle in a bioinformatic search for genomic islands that coordinates several pre-existing computer programs. The algorithm proceeds for each genome as follows:

(i) tRNA and tmRNA genes are identified using tRNAscan-SE (5) and BRUCE (6), and the tDNAs with CAT anticodons are sorted into isoleucine, initiator and elongator methionine classes (7).

(ii) Genes for candidate integrases are identified using HMMER with the ‘phage integrase’ hidden Markov model from Pfam (8), but rejecting those identified as XerC or XerD (housekeeping tyrosine recombinases that do not function as integrases) using Reverse PSI-BLAST (9).

(iii) Each tDNA is used as a query in a BLAST (9) search of the starting DNA and removes unlikely hits through a series of filters. A search for islands in 106 whole bacterial genomes produced 143 candidates, with the search itself providing an estimate of three false candidates among these. Preliminary phylogenetic analysis of the associated integrases reduced this set to 89 cases of independently evolved site specificity, which showed strong bias for the tmRNA gene. The website Islander (http://www.indiana.edu/~islander) presents the candidate islands in GenBank-style files and correlates integrase phylogeny with site specificity.

*To whom correspondence should be addressed. Tel: +1 812 856 5697; Fax: +1 812 855 6705; Email: kellwill@indiana.edu
(iv) Candidate islands that do not contain or overlap an integrase open reading frame are rejected.

(v) Remaining candidate islands that entirely contain an integrase open reading frame are rejected if another protein gene of >300 bp overlaps the BLAST hit. This eliminates many false candidates in protein-coding regions, which account for the bulk of a prokaryotic genome. However, integrase genes themselves must be excepted, since some (from the viruses Mx8 and SSV, for example) are known to extend across attP (10,11). Currently protein gene coordinates are taken from data (.ptt files) at GenBank (12).

(vi) Remaining candidate islands with the BLAST hit corresponding to a known tRNA gene are rejected. However, it should be noted that a pre-island could in principle contain an entire tRNA gene overlapping the crossover site in attP, and therefore generate intact tRNA genes at both of its post-integration endpoints.

(vii) Remaining candidate islands longer than 200 kb are rejected. This cut-off was selected because it would allow detection of all previously known islands except for the 611 kb rejected. This cut-off was selected because it would allow post-integration endpoints.

(viii) The tDNA fragments split off by islands are from one end or the other of the original tDNA gene. Remaining candidate islands where the BLAST hit does not extend to one end of the tDNA query are therefore rejected. An exception was made at the 3′ end because certain islands appear to induce upon integration a small deletion at a position in the tDNA fragment 3 bp to the 5′ side of the discriminator position (14). To detect islands with such damaged tDNA fragments, we tolerated BLAST hits that extended only until this deletion site.

(ix) Although the displaced fragment can be from either the 5′ or (more usually) 3′ end of the tRNA gene (15), certain tDNA(fragment configurations are not allowed. Remaining candidate islands with a 5′ fragment downstream of the tDNA, or with a 3′ fragment upstream of the tDNA are rejected.

(x) Cases where multiple remaining candidate islands share the same integrase or the same gene fragment are resolved to single candidate islands. However, multiples sharing the same tDNA are allowed, as tandem arrays, as long as each member of the array would have its own integrase and endpoint.

(x) Remaining candidate islands with the tRNA gene in the opposite orientation from the BLAST hit are rejected. In principle, such a configuration could produce an invertible DNA segment, but inversion at a tRNA gene has not been reported. This rejection step comes last so that it can serve as a measure of false positives among the final candidates. In genomes without islands, false candidates with the BLAST hit in the opposite orientation to the tRNA gene should be as likely as those in the same orientation.

It should be noted that many integrative islands will be missed by this algorithm, mainly those with an integration site that is not in a tRNA or tmRNA gene. Additionally, several islands are known whose displaced tDNA fragments are too short to be detected by BLAST, at least in its default mode as we currently run it. Vestigial islands with missing or damaged integrase genes would be missed. A small number of temperate phages are known to use an integrase of the serine recombinase family.

### SEARCH RESULTS FOR WHOLE BACTERIAL GENOMES

The 106 whole bacterial genomes available at GenBank in July 2003 were searched using the above algorithm, producing 143 final candidates. In the last step of the search algorithm, three had been rejected because the BLAST hit was in the opposite orientation from the tRNA gene, which is an estimate of the number of false positives among the 143 final candidates, as described above. Preliminary inspection allowed the rejection of two of these final candidates, where neighboring islands not themselves integrated into tRNA genes contained clusters of low- or non-scoring tRNA genes. Table 1 shows the phylogenetic breakdown for the remaining 141 candidate islands. Three major groups of bacteria, the Firmicutes (Gram-positives), α-proteobacteria and γ-proteobacteria, account for 77% of the strains analyzed, yet contain 95% of the detected islands. The *Escherichia/Shigella/Salmonella* group of γ-proteobacteria average five islands detected per strain, reaching nine in *Escherichia coli* O157:H7 EDL933. For almost half of the bacteria examined, including all the obligate pathogens and endosymbionts, no islands were detected: many of these had no integrase gene.

Partly because some of the genomes analyzed were very closely related, and partly because some clades of closely related integrases have gained a wide host range, multiple islands may represent essentially the same integrase with the same site specificity. Preliminary phylogenetic analysis of the integrases of the 141 islands compressed them into 89 tribes of close relatives with the same site specificity. The catalytic domain sequences of the integrases were aligned using HMMALIGN (8), and pairwise BLOSUM62 distance scores were taken. Tribes were assembled by grouping the integrases.

<table>
<thead>
<tr>
<th>Bacterial group</th>
<th>Strains</th>
<th>Islands</th>
<th>Islands per strain</th>
<th>Strains with at least one island</th>
<th>Highest per-strain island count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firmicutes</td>
<td>42</td>
<td>37</td>
<td>0.9</td>
<td>21</td>
<td>5</td>
</tr>
<tr>
<td><em>Escherichia/Shigella/Salmonella</em></td>
<td>9</td>
<td>45</td>
<td>5.0</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Other γ-proteobacteria</td>
<td>20</td>
<td>28</td>
<td>1.4</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>α-Proteobacteria</td>
<td>11</td>
<td>24</td>
<td>2.2</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Other proteobacteria</td>
<td>6</td>
<td>0</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other bacteria</td>
<td>18</td>
<td>7</td>
<td>0.4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>106</td>
<td>141</td>
<td>1.3</td>
<td>54</td>
<td></td>
</tr>
</tbody>
</table>
Bias against the use of tDNA Ala was also highly significant; a found throughout the just four site types: tDNAArg, tDNASer, tDNALeu and tmRNA. 23 types of tDNA is shown in Table 2. Half of the tribes used those sites by islands. The distribution of the 89 tribes among integrases, as opposed to the current occupancy of genomes, provides our best view into the evolution of site specificity among integrases, and also serve as a useful resource for microbiologists studying bacteriophages or other genetic functions provided by islands.

**DISCUSSION**

As currently implemented, our algorithm does not detect islands integrated into sites outside tRNA genes, and even misses some known islands in tRNA genes if, for example, the displaced gene fragment is too small to be detected by BLAST. However, since the majority of integrative islands are in tRNA genes, we made a rough prior estimate that we would detect half of all islands encoding intact integrases of the tyrosine recombinase family. Among the four complete *E.coli* genomes, where this number has been determined by comparative genome analysis (1,16–18), our success rate was 43% (26 of 61: three of nine in K12, six of 13 in CFT073, eight of 19 in O157:H7 Sakai and nine of 20 in O157:H7 EDL933). In principle, the algorithm could be modified to detect more of the islands in tRNA genes that are now missed, or to search for islands in other types of integration site. Many genes for small RNAs unrelated to tRNA have recently been detected in other ways. Other pages present integrase alignments and phylogeny. The same information is also presented for several additional integrative islands known from the literature or detected in other ways.

**REFERENCES**


