D48–D53 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 05 November 2014*
doi: 10.1093/nar/gku1072

Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes

Corey M. Hudson, Britney Y. Lau and Kelly P. Williams*

Sandia National Laboratories, Department of Systems Biology, Livermore, CA 94550, USA

Received September 15, 2014; Revised October 13, 2014; Accepted October 16, 2014

ABSTRACT

Genomic islands are mobile DNAs that are major agents of bacterial and archaeal evolution. Integration into prokaryotic chromosomes usually occurs site-specifically at tRNA or tmRNA gene (together, tDNA) targets, catalyzed by tyrosine integrases. This splits the target gene, yet sequences within the island restore the disrupted gene; the regenerated target and its displaced fragment precisely mark the endpoints of the island. We applied this principle to search for islands in genomic DNA sequences. Our algorithm identifies tDNAs, finds fragments of those tDNAs in the same replicon and removes unlikely candidate islands through a series of filters. A search for islands in 2168 whole prokaryotic genomes produced 3919 candidates. The website Islander (recently moved to http://bioinformatics.sandia.gov/islander/) presents these precisely mapped candidate islands, the gene content and the island sequence. The algorithm further insists that each island encode an integrase, and attachment site sequence identity is carefully noted; therefore, the database also serves in the study of integrase site-specificity and its evolution.

INTRODUCTION

Genomic islands are horizontally transferred DNA segments integrated into prokaryotic chromosomes. They came into especially sharp view when the second *Escherichia coli* genome (O157:H7) was sequenced and could be compared to the *E. coli* K-12 genome, revealing many large islands unique to O157:H7, some unique to K-12 and even alternative strain-specific islands found at the same chromosomal locus (1,2). Many such islands bear clues to their own mechanisms for their two most basic properties: integration and mobility. Like the prototypical genomic island, prophage lambda, many islands contain (i) an integration module including an integrase gene and (ii) the phage structural and regulatory genes that provide a ready hypothesis for the cross-bacterial mobility of the island, through phage particles. One major class of genomic islands can thus be identified as site-specifically integrated prophages. A second major class of islands displays an integration module together with genes for a Type IV secretion system (T4SS), suggesting that they move through conjugation pili like many plasmids; these are termed integrative conjugative elements (3,4).

Islands need not encode their own mobility or integration functions to be mobile and integrative. Islands with neither phage nor T4SS genes may nonetheless contain a mobilization signal, as do satellite phages and mobilizable plasmids, that accesses the mobility vehicles encoded by helper elements. Not all islands encode an integrase; some rely on the integrase-family host enzyme Xer, which is normally responsible for resolving dimeric chromosomes (5).

In addition to the island-selfish functions of mobility and integration, islands can also carry cargo genes benefitting the host bacterium, promoting for example virulence (as conveyed by the term pathogenicity island (6)), symbiosis or catabolic pathways, (7,8). Cargo-bearing islands are major agents of bacterial evolution.

A hypothetical pre-integration or post-excision circular form for genomic islands can be reconstructed from the genome sequence if the precise endpoints of the island can be determined. These circles have a ∼250 bp DNA segment termed *attP*, where integrase acts to promote recombination with a specific target site *attB* in the chromosome. For a large fraction of integrative islands (∼30–50%) (9,10), *attB* lies within a tRNA or tmRNA gene (together termed tDNAs). In these cases the island *attP* contains a fragment of the tDNA target, such that integration restores a functional tDNA despite disrupting the original tDNA. This leaves a bioinformatically detectable island signature in chromosomes: a tDNA and a tDNA fragment, with the island in between. Although some islands use integrases of the serine recombinase family, only those of the tyrosine recombinase family are known to target tDNAs; in what follows, we restrict the term integrase to members of the tyrosine recombinase family only.

Our algorithm Islander (11) searches for the above island signature, to find genomic islands that contain an integrase gene and target a tDNA *attB* and our website (http://bioinformatics.sandia.gov/islander) presents its results. Al-

*To whom correspondence should be addressed. Tel: +1 925 294 4730; Fax: +1 925 294 3020; Email: kpwilli@sandia.gov

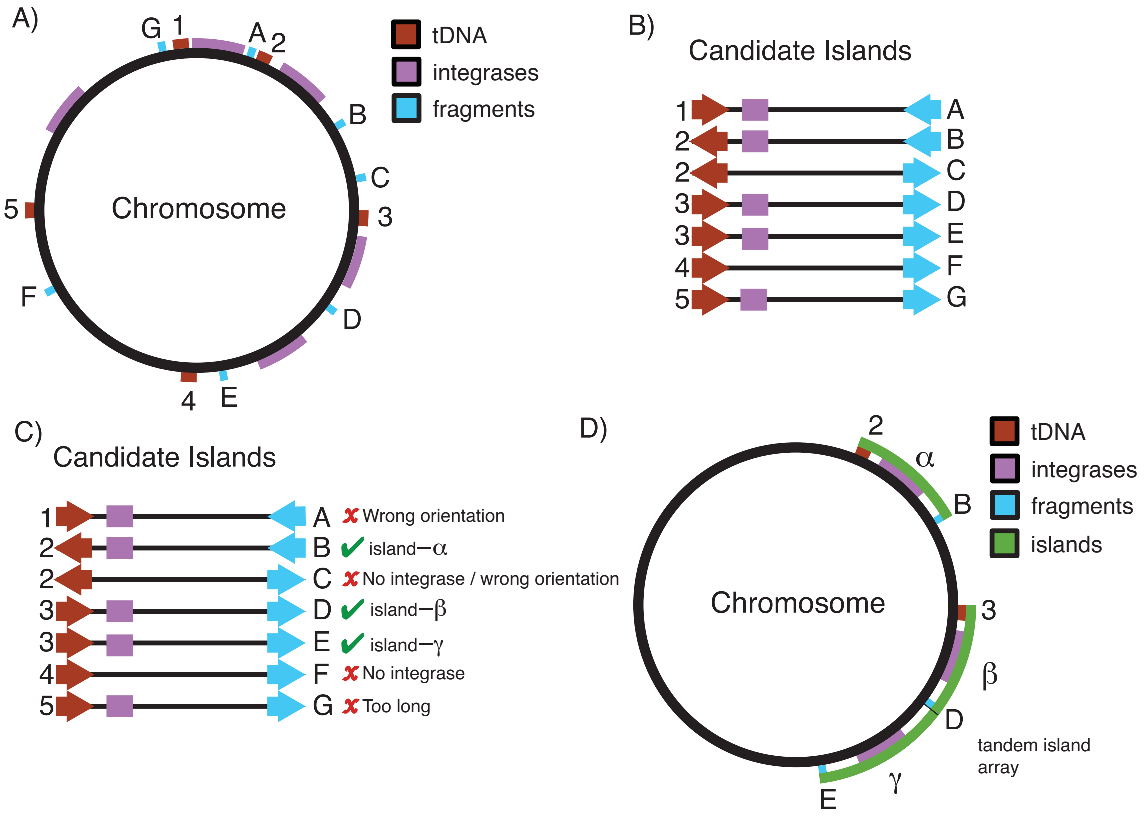

Figure 1. Islander algorithm. (**A** and **B**) Population Phase: tRNA and tmRNA genes (tDNAs), tDNA fragments and integrase genes are placed on the chromosome, and each interval between a tDNA and its cognate fragments is considered a candidate island. (**C**) Filtering Phase: Candidates pass through several filters, including tests for an integrase gene, correct fragment/tDNA orientation and length. (**D**) Resolution Phase: Multiple candidates at the same tDNA are resolved, identifying tandem arrays when each island in the array has its own tDNA fragment and integrase gene.

though this misses some islands, those that it finds are mapped with single-nucleotide precision. They are moreover putatively active, with an integrase gene and both *att* sites intact. Archaea, despite their many differences from bacteria, also carry genomic islands, many meeting our criteria.

Genomic islands are multigene mobile DNA units found in prokaryotic chromosomes (12). Some contain viral structural gene sets or conjugation gene sets that indicate their mode of mobility between cells. Chromosome site-specificity is determined by an integrase, overwhelmingly from the tyrosine recombinase family, whose gene is usually found on board the island. For approximately half of islands the integration target site is within a tRNA gene (9,10). Integrases of the serine recombinase family are not known to specify tRNA genes, and so we restrict the term ‘integrase’ in what follows to members of the tyrosine recombinase family only.

In addition to the selfish genes of mobility and integration, islands can carry cargo genes that benefit the host, contributing to phenotypes such as pathogenicity and metabolic repertoire (8). The target DNA genes specified by integrases have switched frequently during integrase evolution (10), facilitating the combinatorial accumulation of diverse islands in a given host genome. We describe our current algorithm for identifying tDNA-integrated, integrase-encoding genomic islands and present its results along with other such islands from the literature at the Islander Website.

SEARCH STRATEGY

The Islander website collects t(m)RNA-targeted genomic islands, identified using the Islander.pl software package (http://bioinformatics.sandia.gov/software/). Changes from the previously described algorithm (13) include: six-frame translation and Pfam domain annotation; more precise identification of island integrases; and changing the Coding DNA Sequence (CDS) filter from using all proteins called at RefSeq to only those encoding Pfam domains (Figure 1).

The algorithm proceeds through each replicon as follows:

(i) *Find tDNAs*. tRNA and tmRNA genes are identified using tRNAscan-SE (tRNA), BRUCE (tmRNA), ARAGORN (both) and rFind.pl (two-piece tmRNA) (14–17). tFind.pl orchestrates implementation of these tools, corrects endpoints as necessary and sorts tDNAs with CAT anticodons into isoleucine, initiator and elongator methionine classes (18).

(ii) *Find integrases, excluding Xer and integron subclasses*. The replicon is six-frame translated, with amino acid

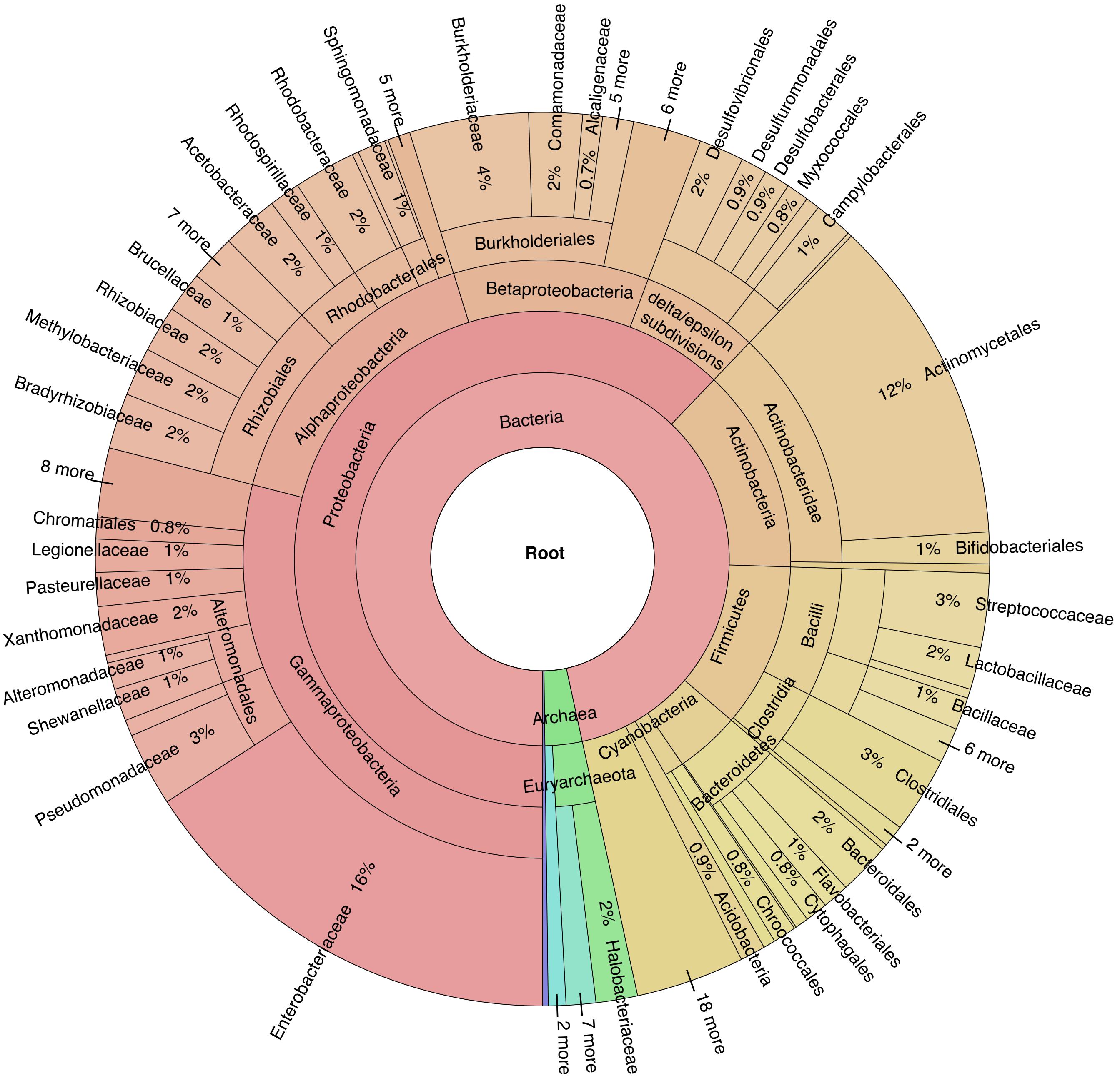

Figure 2. Taxonomic distribution of island prokaryotic hosts. The Krona metagenome visualizer (28) is used to display Islander islands by host, showing the preponderance of proteobacterial islands in the database. It has been modified to also serve as a navigation tool for the website.

sequences extending from stop-codon to stop-codon. Genes for candidate integrases are identified using HMMER with the hidden Markov model (HMM) PF00589 from Pfam (19). The Xer proteins act at *dif* sites to resolve dimeric chromosomes. They can also act as integrases for islands but in these cases are not encoded within the island and specify integration into *dif* sites not tDNA sites. Xer proteins are excluded, using pfscan (20) with strict cutoffs on HAMAP (21) profiles for XerC, XerD, XerS and XerD-like subfamilies. Integron integrases mobilize cassettes within integrons, but not canonical islands. They were excluded using an HMM prepared from a MUSCLE (22) alignment of sequences from the ACLAME family Famint8 (23). Testing with all integrases from the ACLAME database, it was found that a threshold of 1.2e-24 was sufficient to segregate integron integrases from other integrases.

(iii) *Find tDNA fragments*. Each tDNA is used as a query in a BLASTN search of its source replicon (parameters: task, blastn; gapopen, 0; gapextend, 2.5; word_size, 7). The hit and query gene define the endpoints of a candidate island, which is passed through several filters sequentially.

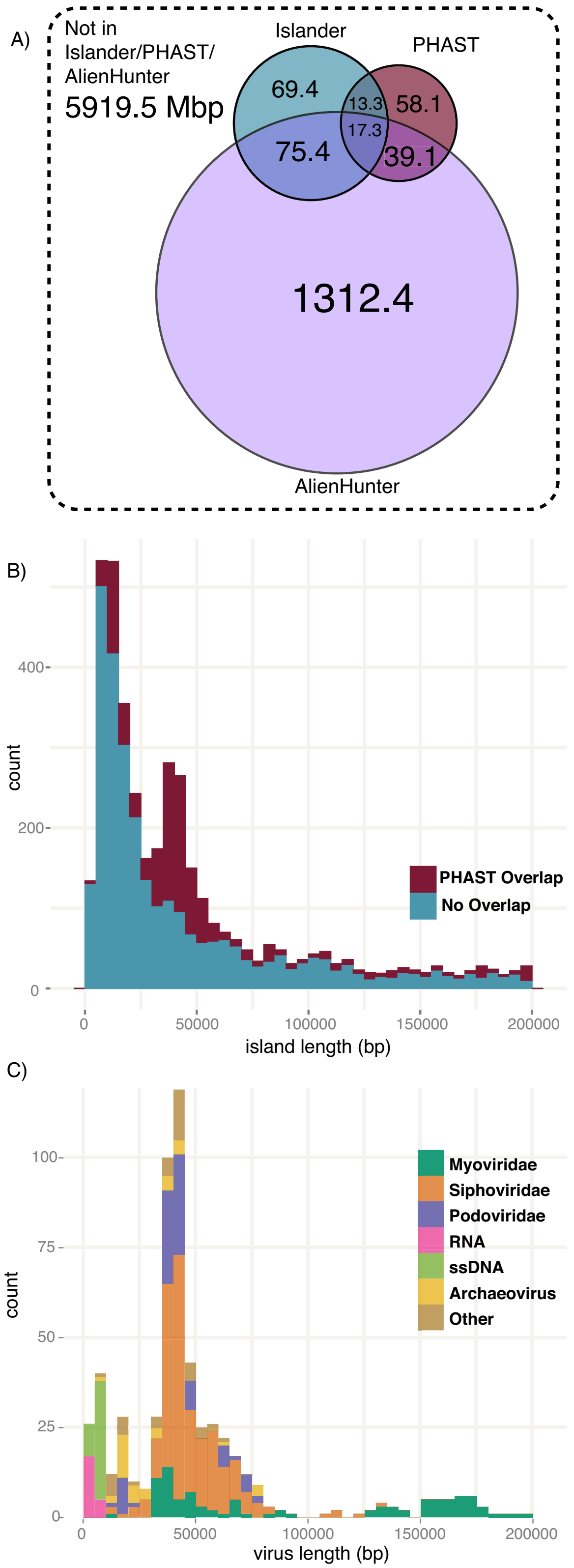

Figure 3. Island overlap. (**A**) DNA length overlap for the outputs from Islander, PHAST (prophage finder) and Alien_hunter (anomalous composition finder). (**B**) Size profile of Islander islands. For each island, overlap by at least 10% with PHAST regions was determined. The peak at 13.6 kbp has both PHAST and non-PHAST components, while that at 39.9 kbp is primarily from the PHAST-labeled islands. (**C**) Size profile of RefSeq prokaryotic viruses, broken down by high-level taxonomic group, matching both peaks of panel (B).

(iv) *Integrase filter*. Candidate islands that do not contain or overlap an integrase open reading frame are rejected.

(v) *CDS filter*. Since true tDNAs and their island-split fragments should not overlap conserved portions of protein coding genes, Pfam-A domains are found among six-frame translated sequences and candidate tDNA fragments overlapping the domain portions of CDSs are excluded. As an exception, overlap of integrase domains is allowed, since some integrases (e.g. those from the viruses Mx8 and SSV) are known to extend across *attP* (24,25).

(vi) *tDNA filter*. BLAST hits that fall within a known full-length tRNA gene are rejected.

(vii) *Length filter*. Candidates shorter than 2 kb or longer than 200 kb are rejected. Some legitimate islands, e.g. the 611 kb symbiosis island of *Mesorhizobium loti* (26), are lost through this filter.

(viii) *Internal tDNA fragment filter*. Integration splits off tDNA end fragments. Candidate islands where the tDNA fragment is internal to the full tDNA are therefore rejected. An exception was made at the 3′ end because certain islands have a small deletion in the tDNA fragment, 3 bp upstream of the discriminator position (27). To detect such damaged tDNA fragments, we allowed BLAST hits that extended only until this deletion site.

(ix) *Configuration filter*. Integration displaces the tDNA fragment to one side. Candidate islands with a 5′ fragment downstream or with a 3′ fragment upstream of the tDNA are rejected.

(x) *Orientation filter*. Candidate islands with the tRNA gene in the opposite orientation from the fragment are rejected. This rejection step, coming late in the analysis, serves as a measure of the false positives among the candidates who should appear with equal frequency in each orientation.

(xi) *Resolve*. Finally, cases where multiple remaining candidate islands share the same tDNA are resolved to single candidate islands, except when tandem arrays can be discerned where each member of the array has its own integrase and tDNA fragment.

ISLANDER WEBSITE DESCRIPTION

The Islander Website was generated from 2031 whole bacterial genomes (that included 1640 plasmids and six phages) and 137 whole archaeal genomes (that included 75 plasmids), and from 1711 additional bacterial plasmids and 543 phages, and 44 additional archaeal plasmids and 38 archaeal viruses, downloaded from four directories at RefSeq (archaea, bacteria, plasmid, and viruses) in November 2012, rejecting eukaryotic viruses and plasmids. The Islander algorithm yielded 3927 unique islands, although all seven that were found in viruses and one in the smallest island-flagged plasmid were excluded after finding evidence that each was a false positive. Some statistics on the 3919 final genomic islands are shown in Table 1. In the majority of genomes at least one island was found; the highest number found in any genome was 19 (*Desulfovibrio magneticus* RS-1). As noted before (11), the tmRNA gene was more highly en-

Table 1. Statistics for Islander islands

Group	No.
Islander islands	3919
Islands that overlap RefSeq CDS	626
Overlapping RefSeq CDS called ‘hypothetical’	276
Tandems > = 2	125
Tandems > = 3	4
Whole genomes	2168
Genomes with at least one island	1302
Islands per genome with at least one (mean)	3.01
Islands with damage	148
Islands with 3′ tDNA fragment	3760
Islands with 5′ tDNA fragment	159
Islands with T-stem region *attP* site J[a]	1836
Islands with anticodon centered *attP* site A[a]	2083

[a] Attachment site crossover sites in tRNA genes appear to fall into three subsites, encoding the anticodon loop (A), the T loop, or the junction (J) between T and acceptor stems (). The latter two can be difficult to distinguish so we combine them here in J.

riched among integration targets than any tRNA isoacceptor type.

Drop down menus become awkward with thousands of genomes and islands. We use a search text box and have also modified the javascript for an interactive metagenome taxonomy viewer, Krona (28), to enable navigation to individual island pages, while also providing a visual depiction of the phylogenetic distribution of islands in the database (Figure 2).

COMPARISON WITH ORTHOGONAL METHODS

Our algorithm detects islands primarily by their target site, while other orthogonal methods look for phage-like genome content or for anomalous nucleotide composition. We evaluated the whole genomes in our study with PHAST (29) to identify prophage-like regions and with Alien_hunter (30) to find regions with biased composition. Figure 3A shows base-pair coverage of the genomes by Islander (2.34% of the total genome length), PHAST (1.70%) and Alien_hunter (19.3%), and their considerable overlaps. A key overlap is the 17.6% of the DNA of Islander islands that is covered by PHAST calls; the remaining PHAST regions may indicate prophages in sites other than tDNAs, or having lost an integrase gene or the tDNA fragment. Alien_hunter regions, despite their high genomic coverage, are enriched in Islander islands.

Figure 3B illustrates the island length distribution among Islander islands. There are distinct peaks centered at 13.6 kbp and 39.9 kbp, the latter mostly due to islands with PHAST overlap. These peaks are both found in the size profile of RefSeq prokaryotic viruses, which we show broken down by virus phylogeny in Figure 3C, with interesting trends. (Note that neither size profile necessarily represents the frequencies of islands or viruses in the natural world; each is biased by the selections researchers have made for genome sequencing.) The RefSeq profile matches best to the portion of the island profile with PHAST overlap (red segments in Figure 3B). The 40 kbp peak is enriched in *Podoviridae* and *Siphoviridae*. That the 40 kbp peak is negligible among non-PHAST islands (blue segments in Figure 3B) indicates that PHAST finds most of the Islander islands that are prophages, at least at this size range. Extrapolating, if nearly all self-mobilizing prophages among Islander islands are among the 1119 found by PHAST, that leaves ∼71% of Islander islands whose mobility must be explained by other means, perhaps either as PHAST-undetectable satellite phages or by conjugation. PHAST-overlapping islands are also enriched in the 13.6 kbp peak, but less so than in 40 kbp peak, and there may be a relative shift between PHAST and non-PHAST components within this peak. This peak is not explained by the single stranded DNA viruses populating that size range of Figure 3C, since none of them encode integrases.

WEBSITE UPDATE

Since the website was last published (13), numbers have increased from 143 to 3919 islands, and from 106 to 2168 whole genomes treated. Our algorithm has changed as described above. We intend for Islander to be a gold standard repository of accurately mapped genomic islands, and are therefore currently combatting the few false positives that are mainly due to relaxing our CDS filter.

Our Islander naming convention takes the first letter of the genus name (excluding *Candidatus*) and the first two letters of the species name, adding a serial number to distinguish strains with the same three-letter nickname, then adds the island length in kbp and a single letter name for the integration site. As an example the 49 591 bp island in *E. coli* O157:H7 str. Sakai (Eco661) integrated into a tRNA-Ser gene is named Eco661_50S.

Additionally the updated website marks islands that are putative prophages overlapping with PHAST calls, reports all the integrases in the replicon, additional matching tDNA fragments, a gene list for the island and the island sequence.

ACKNOWLEDGEMENTS

This research was fully supported by the Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-AC04–94AL85000.

FUNDING

Funding for open access charge: Laboratory Directed Research and Development program at Sandia National Laboratories, grant number 14-1155.

Conflict of interest statement. None declared.

REFERENCES

1. Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K., Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K., Murata,T. *et al.* (2001) Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.
2. Perna,N.T., Plunkett,G. 3rd, Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A. *et al.* (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature*, **409**, 529–533.

3. Burrus,V., Pavlovic,G., Decaris,B. and Guedon,G. (2002) Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.*, **46**, 601–610.
4. Guglielmini,J., Quintais,L., Garcillan-Barcia,M.P., de la Cruz,F. and Rocha,E.P. (2011) The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.*, **7**, e1002222.
5. Huber,K.E. and Waldor,M.K. (2002) Filamentous phage integration requires the host recombinases XerC and XerD. *Nature*, **417**, 656–659.
6. Hacker,J., Bender,L., Ott,M., Wingender,J., Lund,B., Marre,R. and Goebel,W. (1990) Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates. *Microb. Pathog.*, **8**, 213–225.
7. Barondess,J.J. and Beckwith,J. (1990) A bacterial virulence determinant encoded by lysogenic coliphage lambda. *Nature*, **346**, 871–874.
8. Dobrindt,U., Hochhut,B., Hentschel,U. and Hacker,J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.*, **2**, 414–424.
9. Reiter,W.D., Palm,P. and Yeats,S. (1989) Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Res.*, **17**, 1907–1914.
10. Williams,K.P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.*, **30**, 866–875.
11. Mantri,Y. and Williams,K.P. (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.*, **32**, D55–D58.
12. Boyd,E.F., Almagro-Moreno,S. and Parent,M.A. (2009) Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends Microbiol.*, **17**, 47–53.
13. Mantri,Y. and Williams,K.P. (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.*, **32**, D55–D58.
14. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 0955–0964.
15. Laslett,D. and Canback,B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.*, **32**, 11–16.
16. Laslett,D., Canback,B. and Andersson,S. (2002) BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3449–3453.
17. Hudson,C.M., Lau,B.Y. and Williams,K.P. (2014) Ends of the line for tmRNA-SmpB. *Front. Microbiol.*, **5**, 421.
18. Marck,C. and Grosjean,H. (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, **8**, 1189–1232.
19. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29-W37.
20. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.*, **3**, 265–274.
21. Pedruzzi,I., Rivoire,C., Auchincloss,A.H., Coudert,E., Keller,G., de Castro,E., Baratin,D., Cuche,B.A., Bougueleret,L., Poux,S. *et al.* (2013) HAMAP in 2013, new developments in the protein family classification and annotation system. *Nucleic Acids Res.*, **41**, D584–D589.
22. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
23. Van Houdt,R., Leplae,R., Lima-Mendez,G., Mergeay,M. and Toussaint,A. (2012) Towards a more accurate annotation of tyrosine-based site-specific recombinases in bacterial genomes. *Mobile DNA*, **3**, 6.
24. Tojo,N., Sanmiya,K., Sugawara,H., Inouye,S. and Komano,T. (1996) Integration of bacteriophage Mx8 into the *Myxococcus xanthus* chromosome causes a structural alteration at the C-terminal region of the IntP protein. *J. Bacteriol.*, **178**, 4004–4011.
25. Peng,X. (2008) Evidence for the horizontal transfer of an integrase gene from a fusellovirus to a pRN-like plasmid within a single strain of *Sulfolobus* and the implications for plasmid survival. *Microbiology*, **154**, 383–391.
26. Sullivan,J.T. and Ronson,C.W. (1998) Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 5145–5149.
27. Williams,K.P. (2003) Traffic at the tmRNA gene. *J. Bacteriol.*, **185**, 1059–1070.
28. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
29. Zhou,Y., Liang,Y., Lynch,K.H., Dennis,J.J. and Wishart,D.S. (2011) PHAST: a fast phage search tool. *Nucleic Acids Res.*, **39**, W347–W352.
30. Vernikos,G.S. and Parkhill,J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics*, **22**, 2196–2203.