Islander: A tool for precisely mapping genomic islands in tRNA and tmRNA genes

Corey M. Hudson[1] and Kelly P. Williams[1]

[1]Sandia National Laboratories, 7011 East Avenue, Livermore, CA 94550

SAND2015-6887 P

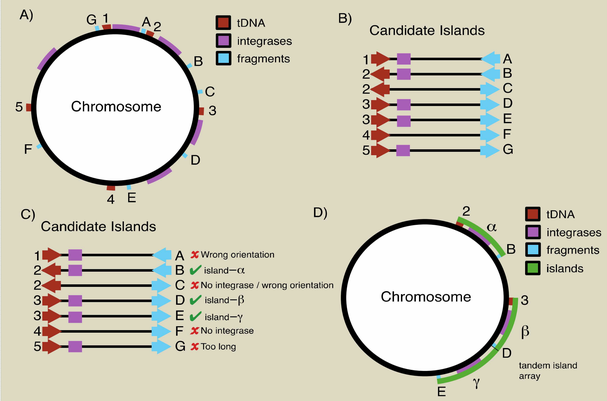
Abstract

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 4065 candidates. The website Islander [1](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. The development of the Islander algorithm and software has also led to a number of supplementary software packages including tFind, a tool for accurately calling and identifying tDNA (tRNAs and tmRNA) including the very challenging to identify two-piece tmRNA and intron interrupted one-piece tmRNA [2]. We have also developed an integrase finder, capable of identifying integrases and distinguishing them from the closely related Xer family of proteins, as well as integron integrases. These softwares are available as part of the Islander Genomic Island Identification Suite of tools http://bioinformatics.sandia.gov/software/.

Mechanism for the integration of mobile DNA

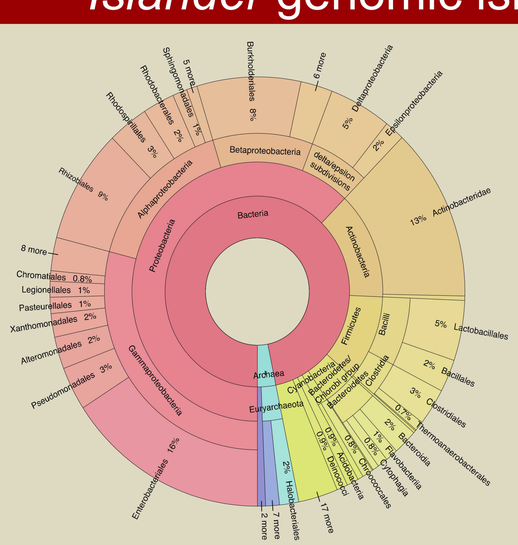
tDNA integration of mobile DNA into prokaryotic genomes. One common path of mobile DNA (often either phage of plasmid in origin) integration into the genome is through integration into tDNAs (either tRNA or tmRNA). These mobile circular DNAs integrate at the *attB* site of the chromosome, replacing it with *attL*, *attR* site, a replacement fragment that restores tDNA function after the integration of the island, and the DNA of the mobile circle. This integration of these segments is mediated by the activity of the integrase (a tyrosine recombinase in tDNA integrations). The result of this activity is the incorporation of the pre-island into the genome.

Islander algorithm



Description of the *Islander* algorithm for genomic island prediction. At a fundamental level, the *Islander* algorithm attempts to directly identify genomic islands in a mechanistic manner, identifying tRNA genes, displaced fragments and integrases. This makes the strategy highly conservative, but also calls genomic island endpoints to single nucleotide precision. *Islander* works by **A**) identifying tRNA genes (tDNAs) and their replicon-wide fragments, integrase genes, and functional CDSs, **B-C**) passing cognate tDNA/fragment pairs (candidate islands) through a series of filters, and **D**) resolving cases where multiple candidates share a tDNA, occasionally yielding a tandem island array and deduplicating these.

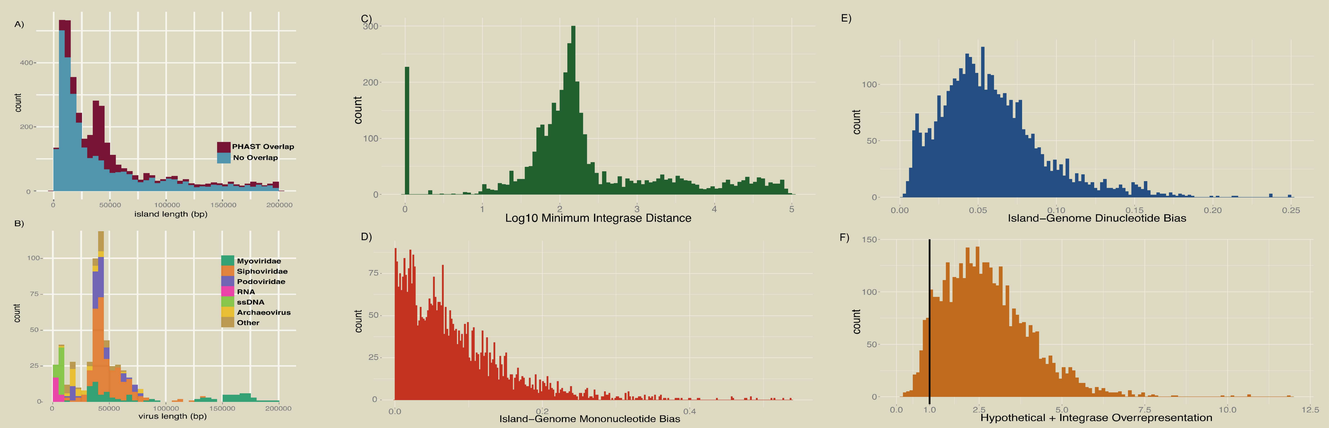
Islander genomic islands in prokaryotes



Distribution of genomic islands among prokaryotic genomes. The Islander Website was generated from 2031 bacterial and 137 archaeal whole genomes, and from 571 virus-only and 958 plasmid-only genomes, downloaded in November 2012, rejecting eukaryotic viruses and plasmids. This yielded 4065 unique islands. In most genomes no islands were found; the maximum number found in any genome was 19 (*Desulfovibrio magneticus* RS-1). The tmRNA gene was more highly enriched among integration targets than any tRNA isoacceptor type.

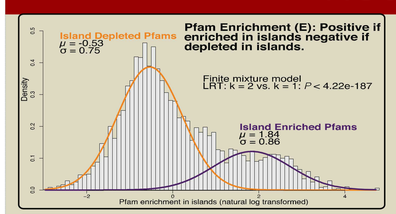
Islander islands	4065
Islands that overlap RefSeq CDS	626
Overlapping RefSeq CDS called "hypothetical"	276
Tandems >= 2	490
Tandems >= 3	38
Whole genomes	3697
Genomes with at least one island	1302
Islands per genome with at least one (mean)	3.01
Islands with damage	152
Islands with 3' tDNA fragment	3884
Islands with 5' tDNA fragment	175
Islands with T-stem region att site*	2133
Islands with anticodon centered att site	1932

Features of Genomic Islands



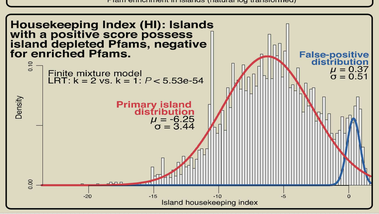
Features of Genomic Islands. Several major features of genomic islands emerge from our analysis of the 4065 tDNA integrated islands. **A)** Islands are constrained to be 2-200kbp long, but show distinct peaks at 13.6 kbp and 39.9 kbp. Islands overlapping PHAST [4], phage-like islands, calls are in dark red. **B)** This pattern in A is generally recapitulated in viruses, in particular ssDNA viruses around 13.6kbp and Sophoviridae around 39.9kbp. **C)** Distance between the integrase and the island endpoints shows that integrases frequently abut the ends of islands and nearly always occur between 1 and 1000 bp. **D-E)** Mononucleotide bias falls dramatically the further it gets from genome, whereas the bias in dinucleotide bias peaks between 0.043 and 0.053. **F)** As has been reported previously [5] islands are distinctly enriched in hypothetical genes, values greater than 1.0 are overenriiched.

Housekeeping Index



Panel 1. Protein families can be distinctly labeled as 'Island Enriched' and 'Island Depleted'. We 6-frame translated and annotated 2168 bacterial and archaeal. Using our 3927 genomic islands, we calculated island enrichments (*E*) = $\frac{Pfam_{island}}{\sum_{i=1}^{n} Pfam_{island}} \Big/ \frac{Pfam_{genome}}{\sum_{i=1}^{n} Pfam_{genome}}$

Using a Finite Mixed Gaussian Model, we find statistical support for a two Gaussian model (Likelihood Ratio Test k = 2 Gaussians vs. k = 1 Gaussian $P < 4.22e-187$). These families are labeled 'Island Depleted Pfams' and 'Island Enriched Pfams' and have estimated means of E = 0.53 and 1.84 respectively.

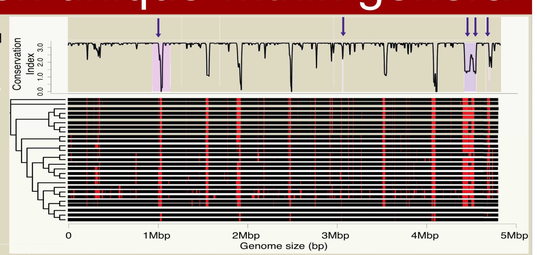


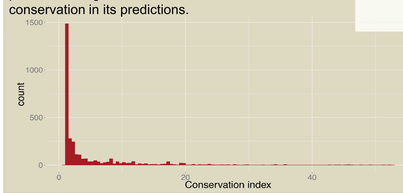
Panel 2. A small group of islands have a disproportionately large number of Depleted Pfams, which points to potential false-positives. For each of the genomic islands we calculated the Housekeeping Index (HI) as: $\sum_{i=1}^{n} \frac{1}{E_i} - \sum_{i=1}^{n} E_i$

The HI distribution goes from a high density of Island Enriched proteins (negative) to a high density of Island Depleted proteins (positive). Using a Finite Mixed Gaussian Model, we find statistical support for a two Gaussian model (LRT k = 2 Gaussians vs k = 1 Gaussian $P < 5.53e-54$). These two distributions are labeled 'Primary Island' and 'False-positive'.

Islands are often unique within genera

Calculation of conservation index for *Salmonella enterica* Typhi str. CT18. We collected whole completed genomes for 27 *Salmonella* strains. These were aligned using Mugsy. The phylogenetc tree was calculated using RAxML-HPC-IV with a GTR + Gamma mode after a 50% filter using Gblocks. The (black) illustrates shared sequence between *Salmonella enterica* Typhi str. CT18 and the 26 other strains. Red blocks illustrate missing sequence. The conservation index is calculated individually for each strain using the natural log transformed count of strains sharing sequence across genera. Arrows and shaded regions illustrate islands predicted using *Islander*, which makes no use of conservation in its predictions.





Conservation index across all islands. Across genera, islands are very rarely conserved. The most common number of shared species/strains with a given island is 1. This may point to the reason why islands are enriched in hypotheticals, given that, at the current sampling of genomes within NCBI, islands occur fairly idiosyncratically, rarely occurring within genera and almost never occurring between genera (*data not shown*). It is unclear if future sequencing efforts will further support this conclusion, or if we have simply reached nothing close to sample saturation.

Conclusions

Since tDNA islands were last thoroughly surveyed [3], numbers have increased from 143 to 4065 islands, and from 106 to 2168 whole genomes treated. We have developed and sophisticated a number of our software tools. Ultimately, we intend for Islander to be a gold standard tools for accurately mapping genomic islands. Using this tools, we have identified several key features of genomic islands, including their occurrence phylogenetically , their length, integrase distance, nucleotide bias and gene class and hypothetical gene overrepresentation. It appears that islands are very rarely shared and even ephemeral in genomes, illustrated by the low conservation. Our intention is for these observation to further winnow the potential spurious islands and in the future point out new islands.

References:
[1] Hudson, Lau and Williams (2015) *NAR* 43: D48-D53
[2] Hudson and Williams (2015) *NAR* 43: D138-D140
[3] Mantri and Williams. (2004) *NAR* 32:D55-D58
[4] Zhou et al. (2011) *NAR* 39:W347-W352
[5] Langillea et al. (2010) *Nat. Rev. Microbiol* 8:373-382

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.