SAND2017-7650C

Short genomic sequences conserved in Bacteria or Archaea

Michelle Hughes, Student Intern, California High School | Kelly P. Williams, Systems Biology, 8623 | July 2017

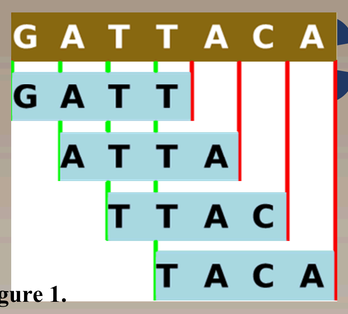
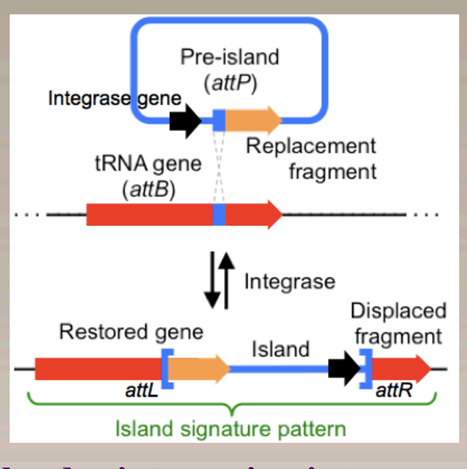


Figure 1.

What does a *K*-mer look like?

INTRODUCTION:

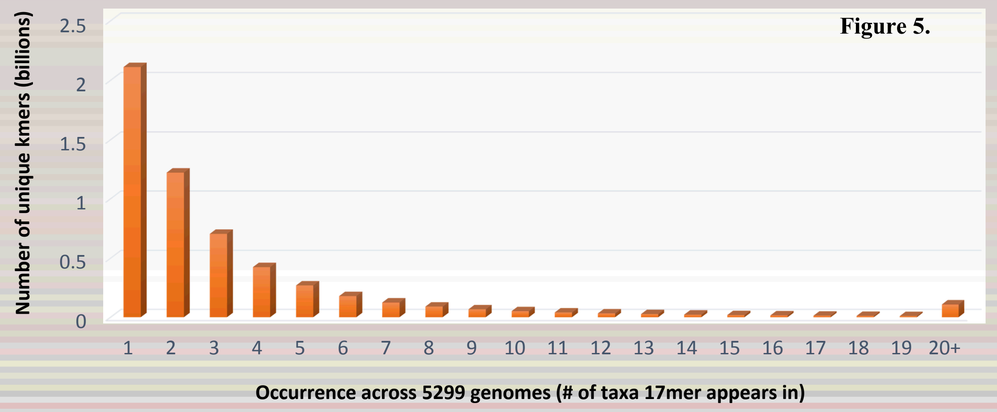
Genomic islands are prokaryote-specific mobile elements that code for pathogenic factors such as antibiotic resistance. They carry integrase genes and integrate into chromosomes at specific sites. It has been observed that these islands often integrate into tRNA and tmRNA genes. This project sought to answer the question of why islands prefer these sites. To do this, we looked at *k*-mers where *k* equals 17 nucleotides, or 17mers, across a wide phylogenetic range of bacteria and archaea (5299 genomes). The length 17 was chosen because the integrase sites that genomic islands typically match to are 17 base pairs long (figure 2). It is thought that because tRNA and tmRNA have such a specific 3 dimensional molecular shape that those genes are highly conserved in DNA. Genomic islands would therefore want to place themselves in these genes to ensure a better chance of long-term survival in the genome.



Islander integration into a gene Figure 3.

Pattern matching for genomic island integration Figure 2.

Frequency of 17mers across genomes studied Figure 5.



17mers mapped to most common gene, position, and orientation Figure 4.

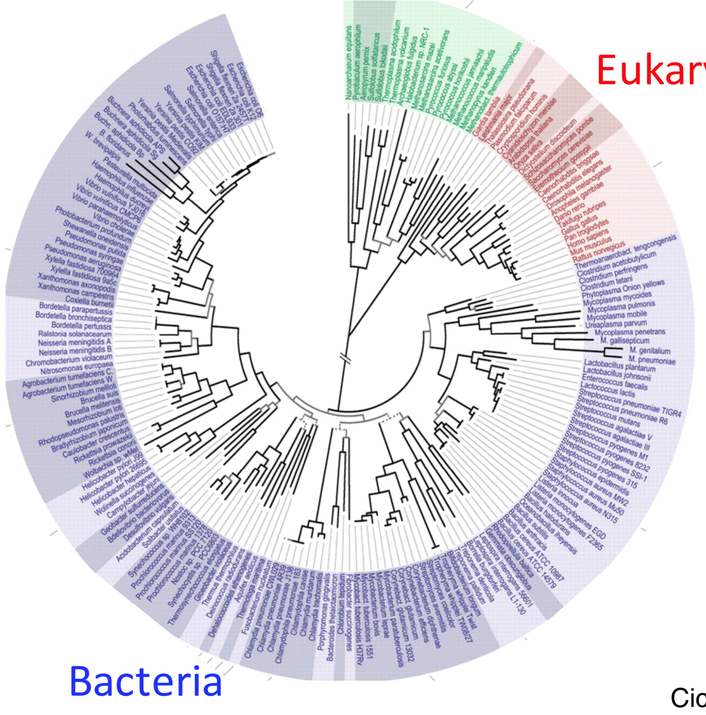
17mer	# taxa found in	Gene	Type	Position	Ori
AAAGGAATTGGCGGGGG	246	16S ribosomal RNA	rRNA	854	1
ACTTAAAGGAATTGGCG	246	16S ribosomal RNA	rRNA	850	1
CCGCCAATTCCTTTAAG	246	16S ribosomal RNA	rRNA	851	-1
CCCGCCAATTCCTTTAA	246	16S ribosomal RNA	rRNA	852	-1
CCCCGCCAATTCCTTTA	244	16S ribosomal RNA	rRNA	853	-1
AACTTAAAGGAATTGGC	244	16S ribosomal RNA	rRNA	849	1
AAACTTAAAGGAATTGG	244	16S ribosomal RNA	rRNA	848	1
GACGGGCGGTGTGTGCA	243	16S ribosomal RNA	rRNA	1337	-1
ACGGGCGGTGTGTGCAA	243	16S ribosomal RNA	rRNA	1336	-1
CAATTCCTTTAAGTTTC	243	16S ribosomal RNA	rRNA	847	-1
CCTTGCACACACCGCCC	243	16S ribosomal RNA	rRNA	1334	1
CGGGCGGTGTGTGCAAG	242	16S ribosomal RNA	rRNA	1335	-1
AATTCCTTTAAGTTTCA	242	16S ribosomal RNA	rRNA	846	-1
AATTGGCGGGGGAGCAC	234	16S ribosomal RNA	rRNA	859	1
AGGAATTGGCGGGGGAG	233	16S ribosomal RNA	rRNA	856	1
AAGGAATTGGCGGGGGA	233	16S ribosomal RNA	rRNA	855	1
GCTCCCCCGCCAATTCC	233	16S ribosomal RNA	rRNA	857	-1
GAATTGGCGGGGGAGCA	233	16S ribosomal RNA	rRNA	858	1
CCTGCTCCTTGCACACA	232	16S ribosomal RNA	rRNA	1328	1
CTGCTCCTTGCACACAC	232	16S ribosomal RNA	rRNA	1329	1
GGCGGTGTGTGCAAGGA	232	16S ribosomal RNA	rRNA	1333	-1
CCCTGCTCCTTGCACAC	232	16S ribosomal RNA	rRNA	1327	1
CGGTGTGTGCAAGGAGC	231	16S ribosomal RNA	rRNA	1331	-1
GGTGTGTGCAAGGAGCA	231	16S ribosomal RNA	rRNA	1330	-1
CTCCTTGCACACACCGC	231	16S ribosomal RNA	rRNA	1332	1
CCCGGGTTCAAATCCCG	227	tRNA-Gly(TCC) tRNA	46	1	
GTCCCTGCTCCTTGCAC	226	16S ribosomal RNA	rRNA	1325	1
CCGGGATTTGAACCCGG	226	tRNA-Gly(TCC) tRNA	47	-1	
TCCCTGCTCCTTGCACA	224	16S ribosomal RNA	rRNA	1326	1
AAGTCGTAACAAGGTAG	220	16S ribosomal RNA	rRNA	1422	1
CCGGATTAGATACCCGG	217	16S ribosomal RNA	rRNA	720	1
CGGCTACCTTGTTACGA	217	16S ribosomal RNA	rRNA	1425	-1
GTAACAAGGTAGCCGTA	217	16S ribosomal RNA	rRNA	1427	1
ACGGCTACCTTGTTACG	217	16S ribosomal RNA	rRNA	1426	-1
AGTCGTAACAAGGTAGC	217	16S ribosomal RNA	rRNA	1423	1
GGCTACCTTGTTACGAC	217	16S ribosomal RNA	rRNA	1424	-1
AACCGGATTAGATACCC	216	16S ribosomal RNA	rRNA	718	1
CCCGGGTATCTAATCCG	216	16S ribosomal RNA	rRNA	721	-1
CGCGGGTTCAAATCCCG	216	tRNA-Asp(GTC) tRNA	48	1	
ACCGGATTAGATACCCG	216	16S ribosomal RNA	rRNA	719	1
ATTAGATACCCGGGTAG	215	16S ribosomal RNA	rRNA	724	1

METHOD/SOFTWARE:

This project involves creating and running Perl code and troubleshooting problems as they arise. It began with 5299 genome sequences from the Progenomes projects, that corrects for overrepresentation. From there a program called Jellyfish was run on each genome's fata files which contained the DNA sequence for that taxon's genome. The program made an output file with the unique 17mers and their frequencies for each taxon. Compiling a single file of all the 17mers found in the genomes, Jellyfish was rerun to find the 17mer distribution across the taxa, as opposed to within each taxon. From there a Perl script was made that put together a matrix file for the bacterial and archaeal genomes that can answer questions such as what is the most widely distributed 17mer (CTCTACCAACTGAGCTA). Afterwards, the project involved mapping the most conserved 17mers to the genes they are in by way of a Perl script that incorporates a software called Bedtools. The 17mers that map to a tRNA gene are then modelled onto a tRNA gene. Another way we are trying to analyze the data is by creating a phylogenetic tree for the 5,299 genomes to compare the 17mer distribution relative to how distanct the taxa are who share the 17mer.

An example phylogenetic tree Archaea

Figure 6.



Ciccarelli, ..., Bork, 2006

PRELIMINARY RESULTS:

We found large numbers of conserved kmers; out of approximately 5 billion individual 17mers in the genomes studied, 107,853,427 were found in over 20 genomes, with 2,559 found in over 1,000 genomes. After mapping all the 17mers from archaeal genomes to the genes and positions they are most commonly found at, I found that the top 71 conserved 17mers in archaea map to tRNA or rRNA genes. The most widespread 17mers in the archaeal genomes mapped more often to rRNA than tRNA genes, which suggests that genomic islands may also have a mechanistic preference for choosing tRNA genes instead of a purely evolutionary approach. Future steps will be to reapply the gene mapping programs to the much larger set of bacterial data and analyze the results. What can then be done is to compare the results for bacteria and archaea against the phylogenetic tree of our genomes, enabling us to see the kmer distribution in relation to the relatedness of the taxa.

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND No. 2011-XXXXP

Photos placed in horizontal position with even amount of white space between photos and header

Box size can be altered to fit your images. Keep even white space between photos.

Poster Title

Area for content

All components of this template are movable and can be resized if necessary to fit your content. However in order to maintain brand strength, please do not change the color pallet.

Please refer to the Corporate Graphic Style Guide for questions regarding design attributes.
http://scg.sandia.gov/resources/common-look-and-feel

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND No. 2011-XXXXP

Photos placed in horizontal position with even amount of white space between photos and header

Photos placed in horizontal position with even amount of white space between photos and header

Poster Title

Area for content

All components of this template are movable and can be resized if necessary to fit your content. However in order to maintain brand strength, please do not change the color pallet.

Please refer to the Corporate Graphic Style Guide for questions regarding design attributes.
http://scg.sandia.gov/resources/common-look-and-feel

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND No. 2011-XXXXP

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration