Identification of Small Regulatory RNA within Miniature Inverse-Repeat Elements of *E. Coli K-12.*

Introduction

Within the genome of *Escherichia coli*, short non-coding sequences (about 127 base pairs), called ERICs are transcribed and reinserted into the genome. ERICs are enterobacterial repetitive intergenic consensus sequences. ERICs are a type of sequence that is repeated many times in the genome and are part of a class of small repetitive sequences known as mobile inverted repeat transposable element. Figure 1 shows the general structure and possible activity of a mobile inverted repeat transposable element (Delhias *et al* 2008).

Noncoding small regulatory RNA can influence gene expression in animals and bacteria. They achieve this by affecting the stability of mRNA transcripts. In eukaryotes, miRNA are responsible for this action. In bacteria, sRNA are known to have this affect, but are usually 50-200nt in length. Kang *et al* found that miRNA-sized molecules exist in *E. coli* and have the ability to alter gene expression like miRNA in eukaryotes (Kang *et al* 2013).

Since ERICs are noncoding sequences that have the ability to be transcribed, it is possible that miRNA-sized sequences can exist within them, considering that *E. coli* has only 15% noncoding DNA. The purpose of this experiment is to locate msRNA within ERICs.

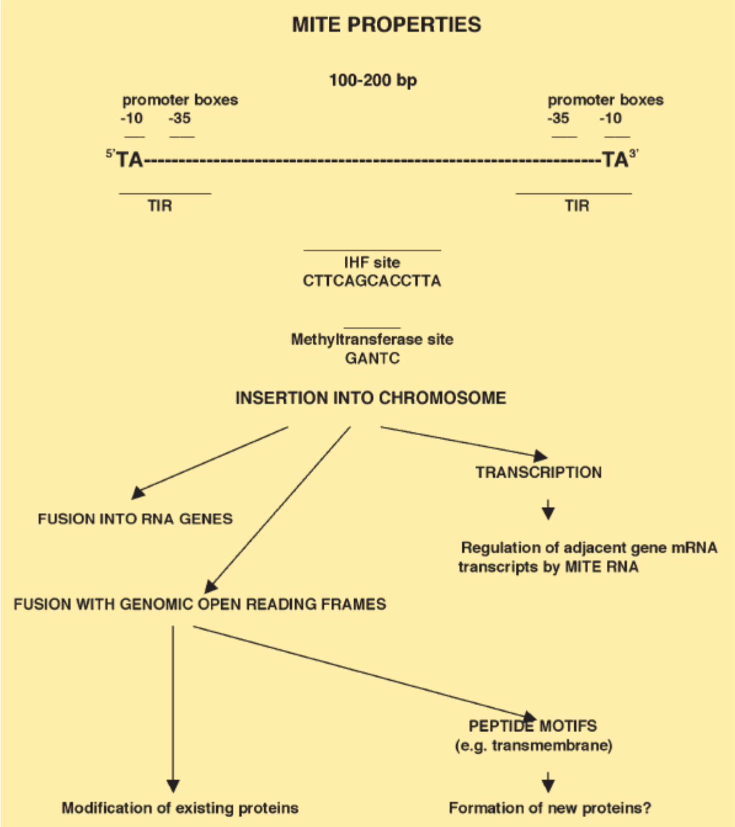

MITE PROPERTIES
100-200 bp
promoter boxes
-10 -35
promoter boxes
-35 -10
5'TA--TA3'
TIR
TIR
IHF site
CTTCAGCACCTTA
Methyltransferase site
GANTC
INSERTION INTO CHROMOSOME
FUSION INTO RNA GENES
TRANSCRIPTION
Regulation of adjacent gene mRNA transcripts by MITE RNA
FUSION WITH GENOMIC OPEN READING FRAMES
PEPTIDE MOTIFS
(e.g. transmembrane)
Modification of existing proteins
Formation of new proteins?

Figure 1. Shows the structure and possible activity of a MITE. TIR represents the terminal inverted repeated sequences that are characteristic of ERICs. (Taken from Delihas *et al* 2008)

Methods

The PhAnToMe version of BioBIKE was used to analyze all sequences described (http://biobike.csbc.vcu.edu)(Elhai *et al* 2009). The genome sequence used to locate ERICs corresponds to the Escherichia coli str. K-12 sub-strain MG1655 assembly ASM584v2 on National Center for Biotechnology Information (NCBI). The coordinates of the ERIC sequences used in this analysis correspond to those found using the method described by Wilson *et al (*2006). Genome sequences were searched exhaustively for matches above a certain threshold to ERIC sequences containing a single deletion at any position between sites 7 and 120. To evaluate the significance of the matches found, the same search was conducted against random sequences with the same length and dinucleotide content as the intergenic sequences of *E. coli* K-12 (Wilson *et al* 2006).

The genome sequence used to locate the msRNA corresponds to the ASM1942v1 assembly of *Escherichia coli* substrain DH10B (NCBI) using methods described by Kang *et al* 2013. Total RNA of *E. coli* was extracted and sequenced. Micro-RNA-sized sequence reads were collected and the relative abundance of each read was calculated. Some sequences were verified using qRT-PCR and northern blot (Kang *et al* 2013).

The sequences of ERICs were extracted from the *E. coli* genome using the coordinates provided by Wison *et al (*2006). ERIC sequences 127nt in length, which is the length of the consensus sequence of ERICs, were selected for analysis. The msRNA sequences provided by Kang *et al* (2013) were co-located with the ERIC sequences. Co-location was achieved by searching each ERIC sequence for each msRNA sequences. msRNA sequences with 100% match with one or more ERIC sequences were collected for analysis. To confirm the significance of the matches, random DNA with the same length and GC content as the ERIC sequences were searched for each msRNA sequence.

127nt ERIC sequences
AATTTCCTTCGTCTTTCACGCCATAGCGGCGTTGGCGTCGCCCGCTCACCCCGGTCACTTACTTGTGTAAGCTCCCG GGGATTCACAGGCTAGCCGCCTTGCTCTGACGCGAAATACTTCGGAAATT (127755 – 127881)
CATACCCTATGGATTTCTGGGTGCAGCAAGGTAGCAAGCGCCAGAATCCCCAGGAGCTTACATAAGTAAGTGACTGG GGTGAGGGCGTGAAGCTAACGCCGCTGCGGCCTGAAAGACGACGGGTATG (190613 – 190739)
CTCCCCCAAAATAGTTCGAGTTGCAGAAAGGCGGCAAGCTCGAGAATTCCCGGGAGCTTACATCAGTAAGTGACCGG GATGAGCGAGCGAAGATAACGCATCTGCGGCGCGAAATATGAAGGGGGAG (253339 – 253465)
TATACTCTAAATAATTCGAGTTGCAGGAAGGCGACAAGCGAGTGAATCGCCAGGAGCTTACATAAGTAAGTGACTGG GGTGAACGAACGCAGTCGCAGTACATGCAACTTGAAGTATGACGAGTATA (437374 – 437500)
TATACTCGTCATACTTCAAGTTGCATGTGCTGCGGCTGCATTCGTTCACCCCAGTCACTTACTTATGTAAGCTCCTG GGGCTTCACTCGTTTGCCGCCTTCCTGCAACTCGAATTATTTAGAGTCTA (596203 – 596329)
TATACTCGTCATACTTCAAGTTGCATGTGCTGCGTCTGCGTTCGCTCACCCCAGTCACTTACTTATGTAAGCTCCTG GGGATTCACTCGCTTGTCGCCTTCCTGCAACTCGAATTATTTAGAGTATG (638731 – 638857)
TATTCTCGTCATACTTCAAGTTGCATGTGCTGCGTCTGCGTTCGCTCACCCCAGTCACTTACTTATGTAAGCTCCTG GGGATTCACTCGCTTGTCGCCTTCCTGCAACTCGAATTATTTAGAGTATA (802545 – 802671)
TATACACAAAATCATTCAAGTTGCATCAAGGCGGCAAGTGAGCGAATCCCGATGAGCTTACTCAGGTAAGTGATTCG GGGGAGCGAACGCAGCCAAGGCAGAGGCGGCTTGAAGGATGAAGTGTATA (1360538 – 1360664)
TATACACTTTATCCTTCACGCTGCCTCTTCGTTGACTGCCTTCGCTCATCCCATTCACATAGTTATCTATGCTCATG GGAGTTCACTCAGTTGCCGCCTCGATGCAACGCGAATGATTTCGTGTATT (1946638 – 1946764)
TACTCGTCATACTTCAAGTTGCATGTGCTGCGTCTGCGTTCGCTCACCCCAGTCACTTACTTATGTAAGCTCCTGGG GATTCACTCTCTTGTCGCCTTCCTGCAACTCGAATTATTTAGAGTATGAA (3069293 – 3069419)
CACCAGCTGTTTGCCCTGTACGGCATCGAAGCGACGCTGTTCATAACGCGGCGTAATACCGTTTTCTTCAGGCATGA TCCAGATCTGATACAGATGCAGACGCTCGGTGCTGCTTGGGTTGTACTCT (3576845 – 3576971)
ATCGTAGTTAAAGACGTGCGTCACTGCCGGAATATGCAAACCACGCGCGGCAACGTCGGTGGCAACCAGAATATCCA GATCGCCACGGGTAAATTCATCAAGAATACGCAGACGTTTTTTCTGCGCG (3962250 – 3962376)
CCTGTTCCGTATTGGTCGTGGACGTGCGCCGACTGGCGAACCTGCGGCGGCAGCGGAAATGACCAAATGGTTTAACA

CCAACTATCACTACATGGTGCCGGAGTTCGTTAAAGGCCAACAGTTCAAA (4010927 – 4011053)
GTCTCTTTCCATGCTTTGCGCAGGGAAGATTCCTCAAAGTGCTGGCGGTCAAACCACTCCTGTAGCTCGACCAGCCCTTTACGGGTGAGATCGCGCGGGCGATTAATAACTGCCTGCAATGCCGGTT (4581141 – 4581267)

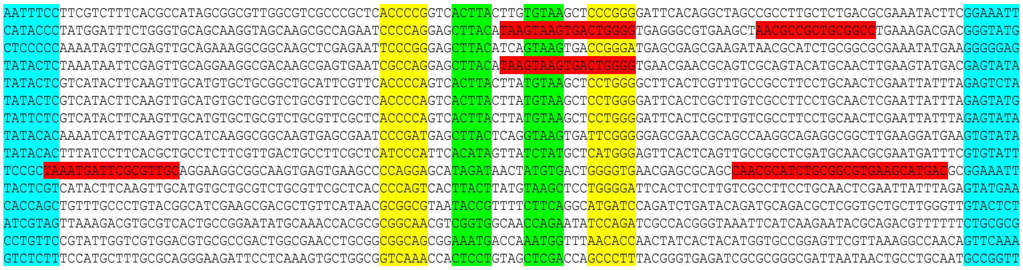
Figure 1. List of ERIC sequences 127nt in length. Coordinates are given after the sequence.

Example of msRNAs
TGTGGGCACTCGAAGATACGGAT
TGTGGGCACTCGAAGATACGGAT
TGTGGGCACTCGAAGATACGGAT
TGTGGGCACTCGAAGATACGGAT
TGTGGGCACTCGAAGATACGGAT
TGTGGGCACTCGAAGATACGGAT
TGTGGGCACTCGAAGATACGGAT
GTTGTGAGGTTAAGCGACT
TTTGCTCTTTAAAAATC
CTCGAAGATACGGATTCTTAAC
TCAAGACGATCCGGTACGCGTGAT
CTTAAGACCGCCGGTCTTGTC
ACAATCTGTGTGGGCACTCGA
TTTGTAGGCCTGATAAG
TTTGTAGGCCTGATAAG
TTTGTAGGCCTGATAAG
TTTGTAGGCCTGATAAG
TTTGTAGGCCTGATAAG
TTTCTCCCTCTCCCTG

Table 2. List of 10 highly expressed msRNA in *E. coli.* (of 503 total).

Results and Discussion

Fifteen known ERICs sequences 127nt in length were searched for 503 msRNA sequences that were experimentally determined (Wilson *et al* 2006)(Kang *et al* 2013). Matches of msRNA sequences within ERIC sequences found to be a 100% match were collected. Sequences less than 100% matched or overlapped were not counted. Two ERIC sequences were found to contain two msRNA sequences. One ERIC sequence was found to contain one msRNA sequence (Figure 2)



```
AATTTCCTTCGTCTTTCACGCCATAGCGGCGTTGGCGTCGCCCGCTCACCCCGGTCACTTACTTGTGTAAGCTCCCGGGGATTCACAGGCTAGCCGCCTTGCTCTGACGCGAAATACTTCGGAAATT
CATACCCTATGGATTTCTGGGTGCAGCAAGGTAGCAAGCGCCAGAATCCCCAGGAGCTTACATAAGTAAGTGACTGGGGTGAGGGCGTGAAGCTAACGCCGCTGCGGCCTGAAAGACGACGGGTATG
CTCCCCCAAAATAGTTCGAGTTGCAGAAAGGCGGCAAGCTCGAGAATTCCCGGGAGCTTACATCAGTAAGTGACCGGGATGAGCGAGCGAAGATAACGCATCTGCGGCGCGAAATATGAAGGGGGAG
TATACTCTAAATAATTCGAGTTGCAGGAAGGCGACAAGCGAGTGAATCGCCAGGAGCTTACATAAGTAAGTGACTGGGGTGAACGAACGCAGTCGCAGTACATGCAACTTGAAGTATGACGAGTATA
TATACTCGTCATACTTCAAGTTGCATGTGCTGCGGCTGCATTCGTTCACCCCAGTCACTTACTTATGTAAGCTCCTGGGGCTTCACTCGTTTGCCGCCTTCCTGCAACTCGAATTATTTAGAGTCTA
TATACTCGTCATACTTCAAGTTGCATGTGCTGCGTCTGCGTTCGCTCACCCCAGTCACTTACTTATGTAAGCTCCTGGGGATTCACTCGCTTGTCGCCTTCCTGCAACTCGAATTATTTAGAGTATG
TATTCTCGTCATACTTCAAGTTGCATGTGCTGCGTCTGCGTTCGCTCACCCCAGTCACTTACTTATGTAAGCTCCTGGGGATTCACTCGCTTGTCGCCTTCCTGCAACTCGAATTATTTAGAGTATA
TATACACAAAATCATTCAAGTTGCATCAAGGCGGCAAGTGAGCGAATCCCGATGAGCTTACTCAGGTAAGTGATTCGGGGGAGCGAACGCAGCCAAGGCAGAGGCGGCTTGAAGGATGAAGTGTATA
TATACACTTTATCCTTCACGCTGCCTCTTCGTTGACTGCCTTCGCTCATCCCATTCACATAGTTATCTATGCTCATGGGAGTTCACTCAGTTGCCGCCTCGATGCAACGCGAATGATTTCGTGTATT
TCCGCTAAATGATTCGCGTTGCAGGAAGGCGGCAAGTGAGTGAAGCCCCAGGAGCATAGATAACTATGTGACTGGGGTGAACGAGCGCAGCCAACGCATCTGCGGCGTGAAGCATGACGCGGAAATT
TACTCGTCATACTTCAAGTTGCATGTGCTGCGTCTGCGTTCGCTCACCCCAGTCACTTACTTATGTAAGCTCCTGGGGATTCACTCTCTTGTCGCCTTCCTGCAACTCGAATTATTTAGAGTATGAA
CACCAGCTGTTTGCCCTGTACGGCATCGAAGCGACGCTGTTCATAACGCGGCGTAATACCGTTTTCTTCAGGCATGATCCAGATCTGATACAGATGCAGACGCTCGGTGCTGCTTGGGTTGTACTCT
ATCGTAGTTAAAGACGTGCGTCACTGCCGGAATATGCAAACCACGCGCGGCAACGTCGGTGGCAACCAGAATATCCAGATCGCCACGGGTAAATTCATCAAGAATACGCAGACGTTTTTTCTGCGCG
CCTGTTCCGTATTGGTCGTGGACGTGCGCCGACTGGCGAACCTGCGGCGGCAGCGGAAATGACCAAATGGTTTAACACCAACTATCACTACATGGTGCCGGAGTTCGTTAAAGGCCAACAGTTCAAA
GTCTCTTTCCATGCTTTGCGCAGGGAAGATTCCTCAAAGTGCTGGCGGTCAAACCACTCCTGTAGCTCGACCAGCCCTTTACGGGTGAGATCGCGCGGGCGATTAATAACTGCCTGCAATGCCGGTT
```

Figure 2. Shows the msRNA sequences found within the fifteen ERIC sequences. The red sequences are the msRNA sequences. The blue, yellow, and green sequences are characteristic structural elements of ERICs.

To test for significance, the number of expected matches were calculated and compared to the observed matches using the Poisson expression.

By definition, ERICs are found exclusively within the intergenic sequences of *E. coli* (with the exception of rRNA sequences, because an insertion of a sequence within a rRNA gene or its promoter would be harmful to the cell).

The length of the total intergenic sequence in *E. coli* minus the lengths of the rRNA genes and -35nt) is 571410nt.

The total length of ALL fifteen ERIC sequences (127nt each) is 15 × 127nt = 1835nt. The chances of a hit within the total ERIC length is 1835/571410.

The expected number of hits (λ) for 503 msRNA = 503 × 1835/571410 = 1.6

The observed number of hits (k) for 503 msRNA = 5

The probability of k number of hits for P(k| λ) = $\frac{\lambda^k}{k!}e^{-\lambda}$

$P(5|1.6) = (1.6^5 \div 5!) \times e^{-1.6} = 0.01764$

$P(4|1.6) = (1.6^4 \div 4!) \times e^{-1.6} = 0.05513$

$P(3|1.6) = (1.6^3 \div 3!) \times e^{-1.6} = 0.13782$

$P(2|1.6) = (1.6^2 \div 2!) \times e^{-1.6} = 0.25842$

$P(1|1.6) = (1.6^1 \div 1!) \times e^{-1.6} = 0.32303$

$P(0|1.6) = (1.6^0 \div 0!) \times e^{-1.6} = 0.20189$

The expected number of msRNA hits (λ_2) for a SINGLE ERIC is 0.111

k_2 for two of the ERICs is 2

k_2 for one of the ERICs is 1

k_2 for the rest is 0

$P(2|0.111) = 0.00551$

$P(1|0.111) = 0.09933$

$P(0|0.111) = 0.89493$

Works Cited

Delihas, N. (2008). Small mobile sequences in bacteria display diverse structure/function motifs. *Molecular microbiology*, *67*(3), 475–81. doi:10.1111/j.1365-2958.2007.06068.x

Elhai, J., Taton, A., Massar, J. P., Myers, J. K., Travers, M., Casey, J., … Shrager, J. (2009). BioBIKE: a Web-based, programmable, integrated biological knowledge base. *Nucleic acids research*, *37*(Web Server issue), W28–32. doi:10.1093/nar/gkp354

Kang, S.-M., Choi, J.-W., Lee, Y., Hong, S.-H., & Lee, H.-J. (2013). Identification of microRNA-size, small RNAs in Escherichia coli. *Current microbiology*, *67*(5), 609–13. doi:10.1007/s00284-013-0411-9

Wilson, L. A., & Sharp, P. M. (2006). Enterobacterial repetitive intergenic consensus (ERIC) sequences in Escherichia coli: Evolution and implications for ERIC-PCR. *Molecular biology and evolution*, *23*(6), 1156–68. doi:10.1093/molbev/msj125